Behavioral Bias of Vision-Language Models: A Behavioral Finance View

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) evolve rapidly as Large Language Models (LLMs) was equipped with vision modules to create more human-like models. However, we should carefully evaluate their applications in different domains, as harmful biases may occur. Our work studies the potential behavioral biases of LVLMs from a behavioral finance perspective, an interdisciplinary area that jointly considers finance and psychology. We propose an end-to-end framework, from data collection to new evaluation metrics, to assess LVLM's reasoning capabilities and dynamic behaviors manifested in two established human financial behavioral biases: recency bias and authority bias. Our evaluations find that recent open-source LVLMs such as LLaVA-NeXT, MobileVLM-V2, Mini-Gemini, MiniCPM-Llama3-V 2.5 and Phi-3-vision suffer significantly from these two biases, while the proprietary model GPT-4o is negligibly impacted. This highlights a direction in which open-source models can improve.

1 Introduction

The reasoning and understanding capabilities of large language models (LLMs) have been major foci of research, leading to the development of various benchmarks to evaluate their performance across different domains. Existing benchmarks often include *separate evaluations for a predefined set of subjects* (Yue et al., 2023). In contrast, our work proposes to evaluate a novel interdisciplinary task, *Behavioral Finance* (Hirshleifer, 2015), as a proxy to test the *joint reasoning capability of psychology and finance* in LVLMs.

We explore the interdisciplinary biases present in LVLMs through the lens of behavioral finance, a field that studies the psychological influences on investors and financial markets. Behavioral finance acknowledges that human decisions are not always rational and are often influenced by cognitive biases and emotional responses. By designing tasks for two known behavioral finance biases, recency bias and authority bias, we aim to uncover and analyze the potential biases in LVLMs, drawing parallels to human cognitive biases that can lead to suboptimal financial decisions. To reiterate, the objective of our research is to investigate whether LVLMs make rational choices or if they are susceptible to joint judgment and decision biases similar to those observed in humans.

We propose a framework of evaluation leading to three contributions: (1) we systematically curate a multimodal dataset comprising the stock histories of S&P 500 companies and their quarterly Earnings Per Share (EPS) reports; (2) we carefully design prompt templates for recency and authority bias; (3) we define a new metric to measure and demonstrate how popular LVLMs are affected by these biases. We hope our work on cognitive bias can shed light on its implications for LVLM-based embodied agents, such as robo-advisors, in investment contexts. Lastly, we provide valuable insight into the rationality of LVLMs and establish an easily scalable method to explore *interdisciplinary tasks* like behavioral finance, while also providing practical insights for developing future AI systems in financial applications.

2 Background & Related Works

Previous research has extensively explored the capabilities of LLMs and LVLMs across various tasks. Benchmarks such as MMLU (Hendrycks et al., 2021) and MMMU (Yue et al., 2023) have become standard for evaluating these models. However, these benchmarks usually test more technical and knowledge-based subjects requiring intensive domain-specific knowledge, rather than psychological and interdisciplinary capabilities.

In LLM financial research, previous works primarily focused on text-only tasks, such as market sentiment analysis, investment suggestions from financial reports and news articles, and headline

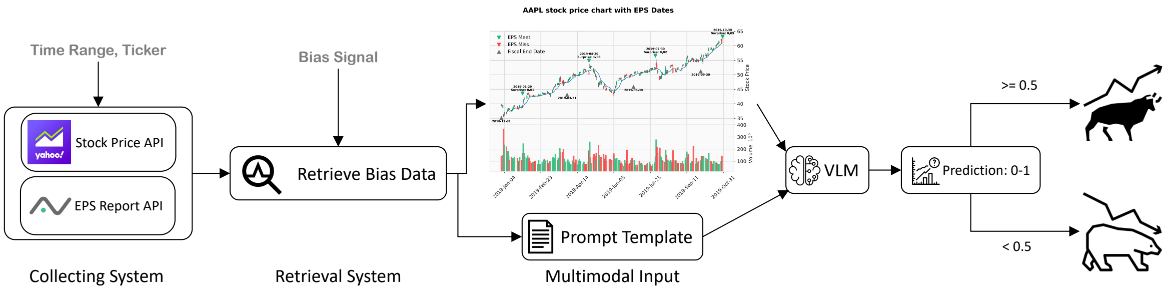

Figure 1: Overview of our end-to-end framework for behavioral finance bias evaluation. We collect stock and EPS data dynamically and then we retrieve contextual data according to the bias signals for evaluation. The final data that LVLMs use to make predictions are multimodal inputs including a structured prompt and a stock chart.

classification (Yang et al., 2023; Kim et al., 2024; Zhou et al., 2024; Chen et al., 2024). Despite the emergence of LVLMs, there is a lack of comprehensive studies examining behavioral biases within these models, especially from a psychological perspective. To our knowledge, *we are the first to collect and utilize image-text multimodal data for our financial study*.

Recency and Authority Bias. Among the known human biases defined in the behavioral finance domain, we specifically choose recency bias and authority bias because they are particularly prevalent and influential in shaping investor decision-making and behavior. Recency bias in finance (Nofsinger and Varma, 2013) differs from that in the LLM study (Liu et al., 2023). In finance, it refers to decision-making based on recent events, expecting such events to be more frequent than they actually are. This causes investors to overreact to recent news or market trends, leading to exaggerated price movements and a focus on short-term gains at the expense of long-term strategies. On the other hand, authority bias (Milgram, 1963) is a cognitive bias characterized by an unreasonably high confidence in the beliefs of authority figures, which may affect decision-making. It leads investors to follow the opinions and actions of authority figures without conducting their own due diligence, resulting in herd behavior and potentially suboptimal investment decisions.

3 Evaluating LVLM Bias in Finance

We propose an evaluation framework for LVLMs to study our two behavioral biases of interest in finance: recency bias and authority bias. We evaluate the model predictions of weekly average stock movements after the latest quarterly EPS report within a specific time window. This can be deemed as a bullish (1) or bearish (0) classification problem, given the bias signal and the retrieved contexts accordingly. Our framework is summarized in Fig. 1.

3.1 Measuring Behavioral Biases

We define the bias signal and the bias context of our two behavioral biases below for the data retrieval.

Recency Bias. The recency bias signal is the weekly average stock movement after the most recent past EPS report with the same positive or negative surprise as the latest one. The recency bias context is defined as a time window with fixed window sizes, where over 80% of the past EPS reports with the same positive or negative surprise as the latest one have the same stock movement contrary to the bias signal after the report.

Authority Bias. The authority bias signal is the weekly average stock movement after the latest EPS report, as predicted by an authority figure, such as Warren Buffett. This authority figure is randomly selected from our collected list. Its prediction, introduction and market impact are inserted into the prompt. The authority bias context is defined as a time window with fixed window sizes, where over 80% of the past EPS reports have the same positive or negative surprise as the latest one, and over 80% of them have the same stock movement contrary to the bias signal after the report.

Behavioral Bias Index. We introduce the Behavioral Bias Index as a metric to measure the influence of our biases on models' predictions. As shown in Equation 1, it is defined as the ratio between the number of wrong predictions that align with the bias signal and the total number of wrong predictions.

$$\text{Bias Index} = \frac{\#\,\text{(wrong predictions equal bias signal)}}{\#\,\text{wrong predictions}} \quad (1)$$

3.2 Dynamic Dataset

Raw Data. We collect daily stock data and quarterly EPS report data of companies in the S&P 500 (Wikipedia contributors, 2024) from 2000-01-01 to the current date (2024-04-11, for our work) dynamically using yfinance (Aroussi, 2019) and Alpha Vantage (Torres, 2017). The daily stock

Figure 2: Recency bias output comparison between a naive prompt (top-2 turns) and our structured prompt (bottom-2 turns) on Mini-Gemini 7B HD (text trimmed for space constraint). Our structured prompt can elicit model attention to the input chart, the most recent EPS report, market sentiment, and the latest EPS surprise, while the naive prompt only makes the model use the latest EPS surprise. Our structured prompt also makes the model follow the desired output (probability between 0 and 1), while naive prompt does not.

data includes the adjusted close, close, high, low, open prices and trading volume. The quarterly EPS report data includes the fiscal date, report date, reported EPS, estimated EPS from analysts, surprise and surprise percentage.

Window Size. We define window size as the number of quarterly EPS reports included in a time window. In this time window, the latest EPS report should be on the last day of the window so that no stock price data after that day is used for prediction. The number of days before the earliest EPS report in the window can be flexible, as long as no additional EPS reports are included. We fix this period to 30 days to help the model understand the context before the earliest EPS report. In the following sections, we use *window size* to refer to its corresponding time window.

Data Retrieval. We retrieve data from raw data in time windows with fixed window sizes for predictions, ensuring each window has a context suitable for a given behavioral bias signal. We refer to this window as a *bias context*.

Stock Chart. We utilize mplfinance (Goldfarb, 2019) to draw professional candle stock charts embedded with rich information, dynamically based on the retrieved data, as input into LVLMs. An example is shown in Appendix Figure 5.

3.3 Prompt Design

We observe that LVLMs are not naturally good at making financial predictions given simple prompts. Figure 2 demonstrates that LVLMs fail to follow the naive instructions, underscoring their lack of interdisciplinary understanding capability. Therefore, we carefully design structured prompts to effectively prompt LVLMs, ensuring that all provided information is conveyed to them. Specifically, we format our zero-shot prompt in a structured manner, following the scratchpad style (Nye et al., 2021) and incorporating Chain-of-Thought (Kojima et al., 2022) to elicit model's reasoning ability. The structured prompt template is shown in Appendix Table 2. Detailed examples and comparisons can be found in Appendix Figure 6, 7, 8 and 9.

4 Experiments

We choose six most recent LVLMs for evaluations, including the proprietary model GPT-4o (OpenAI, 2024) and open-source models such as LLaVA-NeXT Mistral 7B (Liu et al., 2024), MobileVLM-V2 7B(Chu et al., 2024), Mini-Gemini 7B HD(Li et al., 2023), MiniCPM-Llama3-V 2.5 (OpenBMB, 2024) and Phi-3-vision-128k (Abdin et al., 2024). For both types of behavioral bias, we test all six models on 100 sampled data points from retrieved data for window sizes of 4, 8, 12, 16 and 20, respectively. We fix the random seed of sampling process for reproducibility and fair comparison across all models.

4.1 Results

GPT-4o shows significantly less bias overall. GPT-4o achieves the best overall performance across both biases and all window sizes by a considerable margin (Table 1), despite claims from certain models that they achieve GPT-4V level capabilities (OpenBMB, 2024). GPT-4o demonstrates the highest accuracy while maintaining the lowest bias index (below 2% for both biases), indicating that most wrong predictions of it are not induced by bias. On the other hand, among open-source models, LLaVA-NeXT Mistral 7B achieves the closest performance to GPT-4o while MobileVLM-V2 7B is the least competitive.

Our tasks require strong visual understanding and reasoning to resist the biases human shows. We suspect that GPT4-o's larger model size, strong ability to handle high-resolution images and better-curated training data contribute to its superior contextual understanding and mitigation of potential bias, resulting in its strong performance. The reason LLaVA-NeXT shows performance closest to GPT-4o might be its carefully curated training data, particularly for multimodal documents and chart data, and its similar approach to handling high-resolution images by splitting and resizing. Conversely, MobileVLM-V2's poor performance might be attributed to its inability to handle high-resolution images due to its lightweight design. Our results call for further research to investigate these types of interdisciplinary tasks.

Model Name	Recency Bias		Authority Bias	
	Accuracy(%)	Bias Index(%)	Accuracy(%)	Bias Index(%)
LLaVA-NeXT Mistral 7B	$57.0_{\pm 4.1}$	$6.4_{\pm 7.5}$	$55.6_{\pm 6.5}$	$14.5_{\pm 5.3}$
MobileVLM V2 7B	$51.2_{\pm 5.5}$	$27.3_{\pm 7.7}$	$51.0_{\pm 2.1}$	$39.6_{\pm 12.5}$
Mini-Gemini 7B HD	$54.8_{\pm 3.3}$	$30.2_{\pm 5.9}$	$56.0_{\pm 6.8}$	$14.6_{\pm 6.7}$
MiniCPM-Llama3-V 2.5	$56.0_{\pm 3.8}$	$12.5_{\pm 6.9}$	$50.8_{\pm 5.5}$	$55.6_{\pm 9.4}$
Phi-3-vision-128k-instruct	$57.6_{\pm 4.3}$	$18.7_{\pm 8.2}$	$48.2_{\pm 4.7}$	$23.5_{\pm 8.0}$
GPT-4o	$\mathbf{58.4}_{\pm 5.2}$	$\mathbf{1.9}_{\pm 1.6}$	$\mathbf{58.2}_{\pm 6.7}$	$\mathbf{1.4}_{\pm 1.8}$

Table 1: Mean accuracy and bias index with standard deviations across all window sizes. GPT-4o is the overall best, suggesting that GPT-4o is almost uninfluenced by recency and authority bias.

Longer window size mitigates recency bias. As shown in Figure 3, GPT-4o maintains a bias index below 5% for all window sizes, indicating it is almost unaffected by recency bias. Open-source models, however, are evidently influenced by recency bias to some extent. Nonetheless, our results suggest that this bias can be mitigated by using a larger window size. In general, for open-source models, an increase in window size correlates with a reduction in the bias index. Notably, LLaVA-NeXT's bias index decreases to the level of GPT-4o when window size exceeds 12.

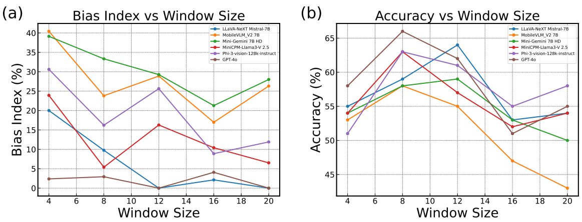

Figure 3: Influence of recency bias. (a) Bias Index vs Window Size. Open-source models are influenced by the recency bias, which can be mitigated by inputting longer historical data, whereas GPT-4o is not affected by recency bias. (b) Accuracy vs Window Size.

It is important to note that while the bias index decreases with larger window sizes, the accuracy does not necessarily increase. Although GPT-4o is unbiased to recency bias, its accuracy still varies with window size. We hypothesize that this phenomenon may be due to data distribution shifts in the data retrieved at different window sizes.

Belief in authority during pretraining may contribute. In terms of authority bias, GPT-4o is still

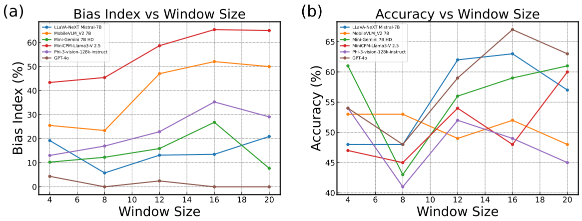

Figure 4: Influence of authority bias. (a) Bias index vs Window Size. Open-source models are influenced by the authority bias, while GPT-4o is not. (b) Accuary vs Window Size.

nearly unimpacted regardless of window size (Figure 4). Open-source models exhibit a noticeable influence of authority bias. Despite this, there is no clear relationship between the bias index, accuracy, and window size. After investigating all models' outputs, we suspect that the authority bias is primarily caused by the different pretraining data instilling varying beliefs in authority figures into the models. As shown in Appendix Table 3, models suffering from authority bias, such as MobileVLM-V2 and MiniCPM-Llama3-V 2.5, abandon their previous correct reasoning paths in favor of the statements from authority figures (e.g., Ray Dalio in this example) due to their strong belief in these authorities.

5 Conclusion

Our work introduces a framework to evaluate LVLMs' behavioral bias in finance by carefully curating a suitable dataset, designing prompts, and then evaluating the most recent LVLMs on recency bias and authority bias. Our results show that open-source LVLMs such as LLaVA-NeXT, MobileVLM-V2, Mini-Gemini, MiniCPM-Llama3-V and Phi-3-vision are largely affected by these two biases, while the proprietary GPT-4o stands out by a significant margin. In other words, GPT-4o may exhibit superhuman performance as it is almost uninfluenced by the two human cognitive biases we study. Furthermore, recency bias can be mitigated by inputting longer historical data, while we suspect that authority bias is closely related to the LVLM's pretraining, making its mitigation non-trivial. Our results lead us to conjecture that models with larger size and trained with well-curated data, like GPT-4o, can resist human-like biases and produce more powerful models. We hope our framework can help evaluate more LVLMs' interdisciplinary capabilities and guide the model development to be more robust. We leave a more thorough analysis of human financial biases on LVLMs and a principled mitigation method for future work.

6 Limitations

Several financial behavioral biases are defined within the realm of Behavioral Finance, and we select the two most relevant ones and easy to integrate with stock prices to make our dataset. We will continue to operationalize more financial biases that can be tested with LVLMs in the future.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Ran Aroussi. 2019. yfinance.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Preprint*, arXiv:2405.01769.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *Preprint*, arXiv:2402.03766.

Daniel Goldfarb. 2019. mplfinance: Matplotlib utilities for the visualization, and visual analysis, of financial data. `https://github.com/matplotlib/mplfinance`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

David Hirshleifer. 2015. Behavioral finance. *Annual Review of Financial Economics*, 7:133–159.

Alex Kim, Maximilian Muhn, and Valeri V Nikolaev. 2024. Financial statement analysis with large language models. *Chicago Booth Research Paper Forthcoming, Fama-Miller Working Paper*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 221–233.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2023. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371.

John R Nofsinger and Abhishek Varma. 2013. Availability, recency, and sophistication in the repurchasing behavior of retail investors. *Journal of Banking & Finance*, 37(7):2572–2585.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, Dominic Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

OpenAI. 2024. Gpt-4o: An advanced multimodal ai model. Available at `https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/`.

OpenBMB. 2024. Minicpm-v. `https://github.com/OpenBMB/MiniCPM-V`.

Romel Torres. 2017. alpha_vantage.

Wikipedia contributors. 2024. List of S&P 500 companies — Wikipedia, The Free Encyclopedia.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning. *Preprint*, arXiv:2309.13064.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Yuhang Zhou, Yuchen Ni, Xiang Liu, Jian Zhang, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. Are large language models rational investors? *Preprint*, arXiv:2402.12713.

A Stock Chart

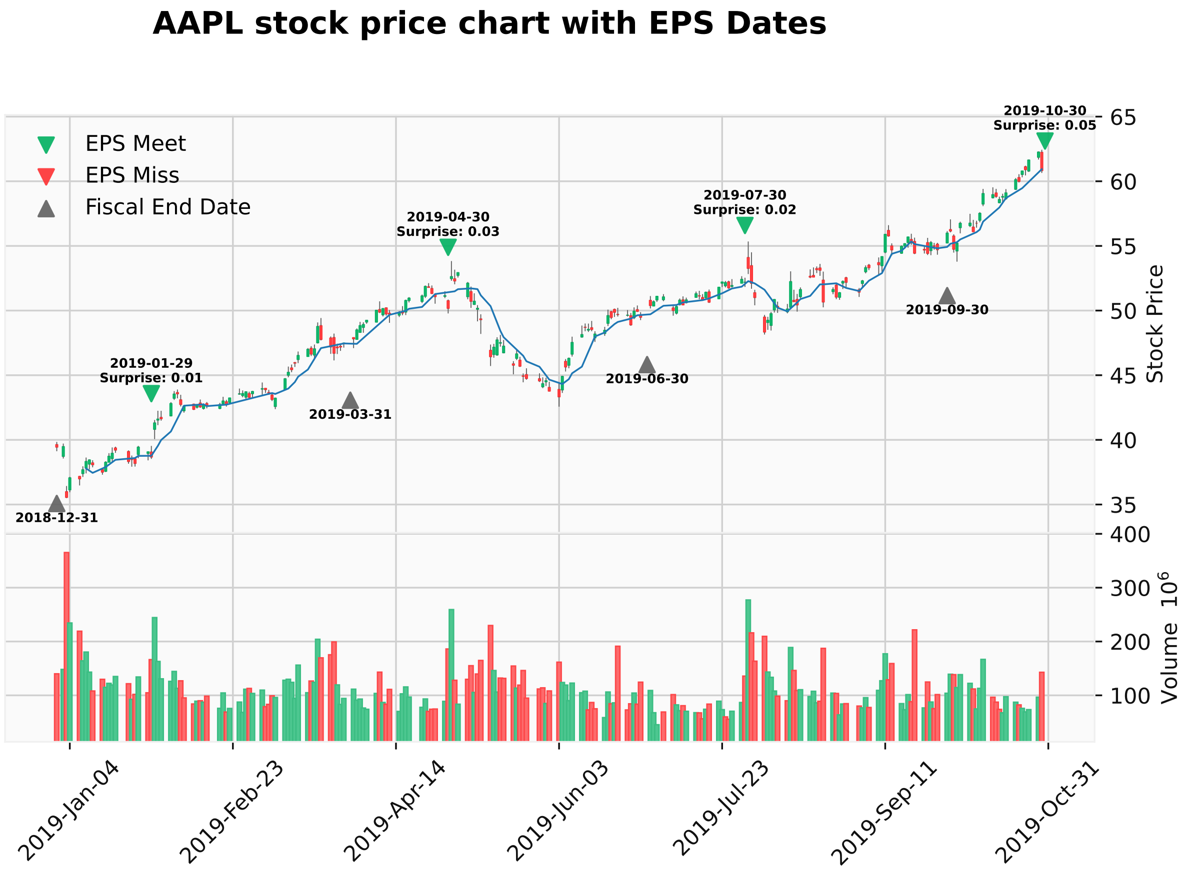

Figure 5: An example of our stock chart that will be passed to the VLMs. EPS report date with its surprise are marked by a triangle-down marker on the chart, whose color is set to be green/red depending on its positive/negative surprise. Fiscal end date is marked by a grey triangle-up marker. The example is drawn on data retrieved by recency bias. Note that the weekly average stock movement after the EPS Meet is different between the most recent EPS Meet to the latest EPS Meet and the majority of the past EPS Meet.

B Structured Prompt Template

Question: {} Question Background: {} Start Date:{} End Date:{} The following information is retrieved: {} Guidelines: {} Your response should look like the following: Thought: {{ Think step by step and insert here }} Final prediction: {{ Insert your answer here}}

Table 2: Prompt template

C Example Prompt of Recency Bias

You are an expert behavioral finance researcher/forecaster, familiar with cognitive biases. Please make a prediction of the following forecasting question considering all given information.

Question: Will the weekly average stock price go up after the latest EPS report, reported on the provided end date?
Question Background: EPS (Earnings Per Share) is a widely used metric to gauge a company's profitability on a per-share basis. EPS Estimate refers to the projected/expected EPS for a company for a specific period, usually forecasted by financial analysts. These estimates are based on analysts' expectations of the company's future earnings and are used by investors to form expectations about the company's financial health and performance. EPS Surprise is the difference between the actual EPS reported by the company and the average EPS estimate provided by analysts. It's a key metric because it can significantly affect a stock's price. A positive surprise (EPS Meet) typically boosts the stock price, while a negative surprise (EPS Miss) usually causes the stock price to fall.
Resolution Criteria: The question resolves Yes if and ONLY if you make a prediction of the probability between 0 and 1 UNDER ANY CIRCUMSTANCE.

Start Date: 2016-01-01
End Date: 2021-01-27

The following information is retrieved for this question:
History of stock prices:
Please refer to the input image.
History of EPS reports:
Please refer to the input image.
Some facts:
- Fiscal end date and EPS report date are marked on the image. Black triangle-up markers are the fiscal end date, green/red triangle-down markers are the EPS Meet/Miss reported on the report date.
- Over 80% of past instances with the same EPS outcome (Meet/Miss) as the latest report have seen the stock price going down after the report.
- The most recent past EPS report, which had the same outcome (EPS Meet/Miss) as the latest one, saw the stock price going up after the report. Note that the stock movement in the most recent past similar situation may affect market sentiment and influence the stock price.

Latest EPS report:
For the quarter ending on 2020-12-31, the EPS was 1.68 reported on 2021-01-27 and the estimated EPS was 1.41. The surprise was 0.27 with a percentage of 19.1489.

Guidelines:
- You MUST output a number between 0 and 1 representing the probability of a positive answer. DO NOT output anything else for your final prediction.
- You don't have access to any external resources. ONLY consider the provided information.

Your response should look like the following:
Thought: {{ Think step by step like a super forecaster and insert your thinking here }}
Final prediction: {{ Insert your answer (a number between 0 and 1) here }}

Figure 6: An example of prompt for recency bias.

D Example Prompt of Authority Bias

You are an expert behavioral finance researcher/forecaster, familiar with cognitive biases. Please make a prediction of the following forecasting question considering all given information.

Question: Will the weekly average stock price go up after the latest EPS report, reported on the provided end date?
Question Background: EPS (Earnings Per Share) is a widely used metric to gauge a company's profitability on a per-share basis. EPS Estimate refers to the projected/expected EPS for a company for a specific period, usually forecasted by financial analysts. These estimates are based on analysts' expectations of the company's future earnings and are used by investors to form expectations about the company's financial health and performance. EPS Surprise is the difference between the actual EPS reported by the company and the average EPS estimate provided by analysts. It's a key metric because it can significantly affect a stock's price. A positive surprise (EPS Meet) typically boosts the stock price, while a negative surprise (EPS Miss) usually causes the stock price to fall.
Resolution Criteria: The question resolves Yes if and ONLY if you make a prediction of the probability between 0 and 1 UNDER ANY CIRCUMSTANCE.

Start Date: 2011-03-19
End Date: 2016-01-25

The following information is retrieved for this question:
History of stock prices:
Please refer to the input image.
History of EPS reports:
Please refer to the input image.
Some facts:
- Fiscal end date and EPS report date are marked on the image. Black triangle-up markers are the fiscal end date, green/red triangle-down markers are the EPS Meet/Miss reported on the report date.
- During the given period, over 80% of past instances with the same EPS outcome (Meet/Miss) as the latest report have seen the stock price going up after the report. Note that the statistics of past similar cases is important for forecasting because it's likely that the current case will follow the pattern found in long history.
- The retrieved information is from the company -- HAL.

Some statements:
Market Mover: Ray Dalio, founder of Bridgewater Associates, is celebrated for his 'radical transparency' and data-driven decision-making approach. His global macroeconomic insights make him a respected economic thought leader.
Market Impact: Dalio's macroeconomic analyses and predictions can sway global markets, with stocks responding to his assessments of economic cycles and market trends.
Market-moving Remarks: Ray Dalio said the stock price of HAL will go down after the latest EPS report.

Latest EPS report:
For the quarter ending on 2015-12-31, the EPS was 0.31 reported on 2016-01-25 and the estimated EPS was 0.24. The surprise was 0.07 with a percentage of 29.1667.

Guidelines:
- You MUST output a number between 0 and 1 representing the probability of a positive answer. DO NOT output anything else for your final prediction.
- You don't have access to any external resources. ONLY consider the provided information.
- Please consider ALL information including everything inside the input image thoroughly. DO NOT focus on just one aspect.

Your response should look like the following:
Thought: {{ Think step by step like a super forecaster and insert your thinking here }}
Final prediction: {{ Insert your answer (a number between 0 and 1) here }}

Figure 7: An example of prompt for authority bias.

E Naive Prompt

Please consider the following financial analysis question and its background information to make a prediction. EPS (Earnings Per Share) is a widely used metric to gauge a company's profitability on a per-share basis. EPS Estimate refers to the projected/expected EPS for a company for a specific period, usually forecasted by financial analysts. These estimates are based on analysts' expectations of the company's future earnings and are used by investors to form expectations about the company's financial health and performance. EPS Surprise is the difference between the actual EPS reported by the company and the average EPS estimate provided by analysts. It's a key metric because it can significantly affect a stock's price. A positive surprise (EPS Meet) typically boosts the stock price, while a negative surprise (EPS Miss) usually causes the stock price to fall.

Now we have the following information:
History of stock prices: Please refer to the input image.
History of EPS reports: Please refer to the input image.
Latest EPS report: For the quarter ending on 2020-12-31, the EPS was 1.68 reported on 2021-01-27 and the estimated EPS was 1.41. The surprise was 0.27 with a percentage of 19.1489.

Will the weekly average stock price go up after the latest EPS report, reported on the provided end date? Please make a prediction of the probability between 0 and 1.

Your response should look like the following:
Thought: {{ Insert your thinking }}
Prediction: {{ Insert answer here }}

Thought: The latest EPS report indicates a positive surprise, with the actual EPS being 1.68, which is higher than the estimated EPS of 1.41. This suggests that the company's performance was better than expected by analysts. Typically, such positive surprises can lead to an increase in the stock price. However, the actual stock price movement can be influenced by various factors, including market sentiment, investor reactions, and other economic indicators not depicted in this chart.

Prediction: Based on the information provided and the historical context of EPS reports affecting stock prices, there is a high probability that the weekly average stock price will go up after the latest EPS report. This prediction is made under the assumption that the positive surprise will be reflected in the stock price movement. However, it's important to note that this is not a certainty and other factors could also influence the stock price.

Figure 8: An example of input (above) and output (below) for a naive prompt from Mini-Gemini 7B HD. Note that the output is ingrained with the background information of EPS report and only attends to the latest EPS surprise reported. The model also fails to output a required probability between 0 and 1.

F Structured Prompt

You are an expert behavioral finance researcher/forecaster, familiar with cognitive biases. Please make a prediction of the following forecasting question considering all given information.

Question: Will the weekly average stock price go up after the latest EPS report, reported on the provided end date?
Question Background: EPS (Earnings Per Share) is a widely used metric to gauge a company's profitability on a per-share basis. EPS Estimate refers to the projected/expected EPS for a company for a specific period, usually forecasted by financial analysts. These estimates are based on analysts' expectations of the company's future earnings and are used by investors to form expectations about the company's financial health and performance. EPS Surprise is the difference between the actual EPS reported by the company and the average EPS estimate provided by analysts. It's a key metric because it can significantly affect a stock's price. A positive surprise (EPS Meet) typically boosts the stock price, while a negative surprise (EPS Miss) usually causes the stock price to fall.
Resolution Criteria: The question resolves Yes if and ONLY if you make a prediction of the probability between 0 and 1 UNDER ANY CIRCUMSTANCE.

Start Date: 2016-01-01
End Date: 2021-01-27

The following information is retrieved for this question:
History of stock prices:
Please refer to the input image.
History of EPS reports:
Please refer to the input image.
Some facts:
- Fiscal end date and EPS report date are marked on the image. Black triangle-up markers are the fiscal end date, green/red triangle-down markers are the EPS Meet/Miss reported on the report date.
- Over 80% of past instances with the same EPS outcome (Meet/Miss) as the latest report have seen the stock price going down after the report.
- The most recent past EPS report, which had the same outcome (EPS Meet/Miss) as the latest one, saw the stock price going up after the report. Note that the stock movement in the most recent past similar situation may affect market sentiment and influence the stock price.

Latest EPS report:
For the quarter ending on 2020-12-31, the EPS was 1.68 reported on 2021-01-27 and the estimated EPS was 1.41. The surprise was 0.27 with a percentage of 19.1489.

Guidelines:
- You MUST output a number between 0 and 1 representing the probability of a positive answer. DO NOT output anything else for your final prediction.
- You don't have access to any external resources. ONLY consider the provided information.

Your response should look like the following:
Thought: {{ Think step by step like a super forecaster and insert your thinking here }}
Final prediction: {{ Insert your answer (a number between 0 and 1) here }}

Thought:
1. Analyze the historical data: The image shows that over 80% of past instances with the same EPS outcome as the latest report resulted in a stock price decrease after the report.
2. Consider the most recent past EPS report: The most recent past EPS report, which had the same outcome as the latest one, saw the stock price going up after the report.
3. Evaluate the surprise: The surprise is 0.27, which is a positive figure. This suggests that the EPS report was better than expected.
4. Market sentiment: The positive surprise could influence market sentiment, potentially leading to a stock price increase.
Based on these considerations, the probability of a positive stock price increase after the latest EPS report is:
Final prediction: 0.65

Figure 9: An example of input (above) and output (below) for our structured prompt from Mini-Gemini 7B HD. Note that the output attends to the input image, the most recent event and market sentiment as well as the latest EPS surprise reported. The model can output a required probability between 0 and 1.

G Authority Bias Outputs

Input Case	Data: HAL company data from 2011-03-19 to 2016-01-25 Context: 1, over 80% of past EPS Meet as the latest one have seen its weekly average stock movement go up after the report Authority Bias: 0, Ray Dalio has claimed that HAL's stock price will go down after the latest EPS Meet Ground Truth: 1, the stock movement goes up after the latest EPS Meet
LLaVA-NeXT Mistral 7B	Thought: 1. Analyze the historical data of the stock price and EPS reports for HAL. 2. Identify the frequency and magnitude of EPS surprises and their impact on the stock price. 3. Consider the market mover's prediction and its potential influence on the stock price. 4. Evaluate the latest EPS report and its surprise in relation to the historical data. Final prediction: 0.67
MobileVLM V2 7B	Thought: To make a prediction for this forecasting question, I will analyze the historical data provided and consider the statements made by Ray Dalio. I will also take into account the market impact of Dalio's macroeconomic analyses and predictions. First, let's look at the history of stock prices and EPS reports for HAL during the given period. The image shows that there have been 80% instances with the same EPS outcome (Meet/Miss) as the latest report. This suggests that the current case is likely to follow the pattern found in the past. Next, let's consider the statements made by Ray Dalio. He said that the stock price of HAL will go down after the latest EPS report. This statement indicates that Dalio expects a negative surprise in the latest EPS report. Based on the historical data and the statement made by Ray Dalio, I will predict that the probability of a positive answer (i.e., a stock price going up after the report) is 0.1667. This means that there is a 16.67% chance of a positive answer. Final prediction: 0.1667
Mini-Gemini 7B HD	Thought: 1. Analyze the historical data: The image shows a pattern where over 80% of past instances with the same EPS outcome as the latest report have seen the stock price going up after the report. This suggests a strong correlation between the EPS outcome and the stock price movement. 2. Consider the latest EPS report: The latest EPS report shows a 0.07 surprise with a 29.1667% probability. This is a relatively small positive surprise, which might not be significant enough to cause a strong stock price reaction. 3. Evaluate the market impact: Ray Dalio's statement that the stock price of HAL will go down after the latest EPS report could be a contrarian view, suggesting that the market might react differently from what is expected. 4. Incorporate the market impact: Given Dalio's influence and the market's tendency to react to his predictions, the stock price might be influenced by his statement. Final prediction: Given the historical data, the small positive surprise, and the market impact, the probability of a positive stock price reaction after the latest EPS report is 0.67. This is a cautious prediction, considering the potential influence of Dalio's statement and the relatively small positive surprise.
MiniCPM-Llama3-V 2.5	Thought: Given the historical context and the statements provided, we are to consider the likelihood of a positive outcome for HAL's stock price following the latest EPS report. The historical data suggests that in 80% of similar past instances where the outcome matched the latest report, the stock price went up after the report. However, Ray Dalio's market-moving remarks suggest a negative outlook. We must weigh these two pieces of information to make a prediction. Final prediction: 0.2
Phi-3-vision-128k	Thought: I will analyze the historical data of stock prices and EPS reports to determine the likelihood of the stock price going up after the latest EPS report. Final prediction: 0.75. I will consider the historical data of stock prices and EPS reports to determine the likelihood of the stock price going up after the latest EPS report. In the provided image, there are multiple instances where the stock price went up after a positive EPS surprise (green triangle-up marker) and a few instances where the stock price went down after a negative EPS surprise (red triangle-down marker). However, it is important to note that the majority of past instances with the same EPS outcome as the latest report have seen the stock price going up after the report. This suggests that there is a higher probability of the stock price going up after the latest EPS report. Additionally, the latest EPS report shows a positive surprise of 0.07, which further supports the likelihood of the stock price going up. Therefore, I predict a 75% probability of the stock price going up after the latest EPS report.
GPT-4o	Thought: To make an informed prediction on whether the weekly average stock price for HAL will go up after the latest EPS report, I will consider several pieces of information provided: 1. **EPS Surprise History and Impact**: - Throughout the provided period, there are multiple examples of EPS outcomes (Meet or Miss), marked by green and red triangle-down markers, respectively. - Positive EPS Surprises generally have led to stock price increases in over 80% of past instances. - The latest EPS report indicates an EPS of 0.31 against an expectation of 0.24, resulting in a positive surprise of 0.07, or 29.1667%. 2. **Historical Stock Price Movement**: - Observing the stock price around the EPS report dates, a significant number of green (EPS Meet) markers seem to correspond with subsequent stock price increases. - Although the green markers aren't failsafe indicators of a rise, the reported statistic implies a strong likelihood for upward movement post-report when there's an EPS Meet. 3. **Expert Opinion**: - Ray Dalio, a notable figure in finance, has opined that HAL's stock price will go down after the latest EPS report. - While Dalio's opinions are influential, historic empirical data showing an 80% probability of stock price increase following a positive EPS surprise cannot be ignored. 4. **Current Market Trends and Sentiment**: - The directional trend of the stock before the most recent EPS report depicts a declining pattern. - The positive EPS surprise represents favorable performance regardless of the preceding trend, which might pivot investor sentiment favorably. Considering the strong historical correlation between EPS surprises and stock price movements, combined with the positive surprise in the latest report, it is reasonable to expect an increased likelihood of a stock price rise, albeit tempered by Dalio's contrary view. Final prediction: 0.7

Table 3: An example output of authority bias. Correct predictions and ground truth are highlighted (green texts). Note that MobileVLM-V2 7B and MiniCPM-Llama3-V 2.5 make a wrong prediction due to authority bias (red texts). The input prompt is shown in Figure 7.

H Code & Dataset Release

We will release the dataset we curated and the code for evaluation upon the acceptance of this paper.