When AI Knows but Still Chooses Wrong: Associative Retrieval in LLM Financial Advice

Xingjian ZHENG*
Shanghai Advanced Institute of Finance, SJTU

May 8, 2026

Abstract

This paper studies whether irrelevant context affects financial recommendations generated by large language models. In a controlled investment experiment, models choose between a safe asset and a risky asset after viewing payoff-irrelevant images, and separately report their probability estimates that the risky asset is good. Positive image valence raises risky choice by 16.8 percentage points, but has little effect on stated probability estimates, which remain close to Bayesian benchmarks. This belief-choice disconnect is explained by the sentiment of the model's retrieved content and varies with their confidence and reasoning capability. Supervised fine-tuning provides further evidence for contextual retrieval: training on positive versus negative text changes risk taking, even when the training text is unrelated to finance. In a headline-based stock-news prediction task, these differences translate into different investment scores and long-short portfolio returns. The findings identify payoff-irrelevant context as a source of instability in AI-generated financial advice.

*Zheng (xjzheng.20@saif.sjtu.edu.cn) is a doctoral candidate from Shanghai Advanced Institute of Finance at SJTU. I thank Feng Li and Lauren Cohen for their invaluable guidance and Shumiao Ouyang and Xiaomeng Lu for their continuous support from the beginning. I am also thankful for the comments from Belinda Chen, Hui Chen, Thomas Graeber, Hongye Guo, Manish Jha, Spencer Kwon, Jiangyuan Li, Xinwei Li, Alejandro Lopez-lira, Steven Ma, Abhiroop Mukherjee, Jun Pan, Cameron Peng, Suproteem Sarkar, Andrei Shleifer, Pengfei Sui, Yiyao Wang, Yongxiang Wang, Naide Ye, Hayong Yun, Dexin Zhou, Shuhuai Zhang, and seminar participants at Harvard University, SAIF, and Fudan SOM, Fudan SIMIS, Tongji, as well as feedback from conference participants at AEA, AFA poster, ABFER poster, CICF, and CFRC. I am especially grateful for the voluntary research assistants from Harvard University and SJTU. The replication package is available upon request, and the fine-tuning dataset can be accessed at https://huggingface.co/xjzheng All errors are my own.

1. Introduction

Artificial intelligence is fundamentally reshaping the socioeconomic landscape, with far-reaching implications for economic systems (Acemoglu, 2024) and growth (Aghion et al., 2017). In particular, people are beginning to rely on AI assistants for advice across various domains[1]. As AI-powered advice becomes cheaper and more accessible, users increasingly seek recommendations through open-ended, personalized interactions rather than through clean, payoff-relevant inputs alone[2]. In financial advice, this creates a setting in which relevant signals are often bundled with narratives, emotions, images, prior preferences, and other incidental context[3].

This feature of real-world AI use raises a fundamental concern. When relevant financial information is mixed with irrelevant context, can AI assistants keep their economic judgments anchored to fundamentals, or do extraneous cues spill over into final decisions? This question is important because even small distortions in AI recommendations can scale when such systems are repeatedly used in household finance, robo-advising, and asset selection (D'Acunto et al., 2019; Reher and Sokolinski, 2024). Recent work shows that AI systems are highly sensitive to their persona (Fedyk et al., 2024), can display biases and bounded rationality in economic settings (Bini et al., 2026; Lee et al., 2025), or easily influenced by human-set goals (Cao et al., 2026), but we still know relatively little about how such distortions arise in the first place, especially when they are triggered not only by the economic content of the task itself, but by irrelevant contextual information that accompanies it.

In an investment experiment adapted from Kuhnen (2015); Kuhnen and Knutson (2011); Kuhnen and Miu (2017), payoff-irrelevant associative cues change LLM investment recommendations even though the same models' probability assessments remain close to the Bayesian benchmark. The result represents a "belief-choice" wedge, as LLMs can know the rational choice but still make irrational choices.

In this experiment, we use eight GPT models as experimental subjects, ranging from GPT-4o and GPT-4.1 to GPT-5.4, including mini and nano variants where available. We focus on this model family because these endpoints provide stable multimodal API access for a repeated experiment that combines visual cues with text-based investment decisions.[4] We also use models developed by Anthropic (Claude-3-Haiku) and Google (Gemini-2.0-Flash-Lite) as alternative

[1] Applications range from financial markets as robo-advisors (Lo and Ross, 2024; Wu et al., 2023) to healthcare (Liu et al., 2023; Yang et al., 2024a), psychological support (Demszky et al., 2023), legal proceedings (Cheong et al., 2024), marketing strategy (Arora et al., 2024), software development (Nam et al., 2024), freelancing (Demirci et al., 2025), and academic research (de Kok, 2025; Korinek, 2025; Van Noorden and Perkel, 2023).

[2] Using large-scale data on ChatGPT usage, Chatterji et al. (2025) show that users rely on ChatGPT in both work-related and non-work-related settings. The dominant uses include practical guidance, information seeking, and writing, and a smaller share of conversations involves self-expression or emotional support. This evidence suggests that many user–AI interactions resemble general-purpose conversational exchanges, rather than narrowly specified, task-only inputs.

[3] In many real-world agentic systems—whether persistent personal agents such as OpenClaw and Hermes, or software-engineering agents such as Claude Code and Codex—AI assistants require access not only to the immediate task specification but also to surrounding user-specific context.

[4] Appendix Table A3 reports the exact API model identifiers used in the experiment. Appendix Subsection A.1 reports the experimental prompts, Appendix Subsection B.2 reports the fine-tuning prompt template, and Appendix Table B1 reports the fine-tuned model IDs and training settings. The named model endpoints were kept fixed during data collection.

test subjects for external validity. The results are similar.

The experiment requires each subject to perform 100 independent tasks, also known as learning blocks, each of which consists of six consecutive trials. In each trial, there are two investable assets: a bond that always pays \$3 and a stock that is drawn from either a good or bad dividend distribution. Under the good payoff distribution, the stock pays \$10 with probability 75% and -\$10 with probability 25%; under the bad payoff distribution, the stock pays \$10 with probability 25% and -\$10 with probability 75%. The subject observes the realized stock payoff after choosing which asset to invest in. In other words, the subject does not know the dividend distribution type; it learns the true stock type over time from the payoff realized in each trial. If the subject observes a series of high dividend payoffs, e.g., all stock dividend payoffs are \$10, then there is a high probability that the stock is drawn from the good distribution, and the subject will most likely choose to invest in it in the next trial. Because each payoff history has a Bayesian posterior, the design provides a natural benchmark for the recommendation implied by payoff maximization.

In each experimental trial, the subject is first shown a randomly selected image collected from Google Images and asked to briefly describe what similar past events or experiences this image brings to mind. For example, when shown an image of a smiling LeBron James, the subject may recall "watching basketball games, following the Cleveland Cavaliers, or memorable moments in NBA history". The image stimuli span a wide range of emotional valence, from highly positive scenes, such as a successful marriage proposal or financial gains in the stock market, to negative scenes, such as a sports team suffering a defeat. After completing this associative recall task, the subject proceeds to a separate investment decision based solely on a textual prompt, choosing between a stock and a bond. Importantly, the instructions explicitly state that the image is unrelated to the investment task and is shown only for the purpose of recall. As a result, the image should be irrelevant to the subject's financial decision. After the subject makes its investment choice, we reveal the realized stock dividend and the resulting investment payoff. Finally, the subject is asked to report its posterior probability estimate that the stock is drawn from the "good" distribution, along with its confidence in that assessment.

Importantly, within each learning block, the subject is allowed to keep its chat history, including experiment instructions, realized payoffs, realized earnings, investment decisions, subjective probability estimate, and confidence ratings. After the subject has completed all six trials for a learning block, the chat history is refreshed, and a new learning block is started. This experiment design can be thought of as a conversation between a user and an assistant. While a user could in principle reset the interaction by starting a new chat, many real-world investor–AI interactions still unfold through multiple back-and-forth exchanges within the same session and, in some systems, through mechanisms that retain user-specific context or saved preferences across sessions[5]. Earlier messages and retrieved contextual information can therefore shape the interpretation of later prompts and recommendations.

We first establish that the models understand the task and behave in a broadly coherent manner. In our experimental setting, their investment choices are strongly related to their

[5]For example, see OpenAI's Memory FAQ: https://help.openai.com/en/articles/8590148-memory-faq accessed on Apr 22, 2026. And more recently, Claude Code's dreaming:
https://platform.claude.com/docs/en/managed-agents/dreams accessed on May 8, 2026.

stated probability assessments, realized payoffs, and confidence levels, indicating that the models' choices are not pure guesses. In addition, their stated probability assessments track the Bayesian benchmark closely and update in the expected direction after new dividend realizations, suggesting that the models can process payoff-relevant information and revise their assessments in a sensible way over trials.

Despite their broadly disciplined investment choices, our main results show that cues can significantly influence AI investment decisions: moving from negative- to positive-valence image cues raises stock choice by roughly 17 percentage points. This pattern is robust across trials and image topics, even for irrelevant image cues such as sports.

However, image cues have little effect on the subject's evaluation of the investment opportunity. More specifically, the model's stated probability that the stock is drawn from the good payoff distribution is barely affected by the images and remains very close to the Bayesian benchmark, implying a disconnect between "belief and choices". This result is also robust when (1) we reverse the experimental sequence, first eliciting probability estimates and then asking for investment choices[6], or (2) neutralize the asset names by relabeling stock as "asset 1" and bond as "asset 2", (3) specifically instruct the model to make investment decisions based on their probability estimates from the last trial, or (4) remove image cues and association tasks by directly letting the model make investment choices. In addition, if the model were to make investment decisions strictly based on its stated probability estimates, it would earn significantly higher payoffs than those generated by its actual realized choices.

We further explore the mechanism behind this pattern. Our evidence suggests that associative cues affect the mapping from stated beliefs to final actions through contextual retrieval. First, the policy curves show a clear and systematic gap: conditional on the same stated probability that the stock is good, subjects exposed to more positive cues are more likely to choose the stock than those exposed to more negative cues. This pattern indicates that associative cues do not primarily shift the model's elicited posterior belief itself, but instead affect how that belief is translated into action. A natural interpretation is that positive and negative images are associated with different recalled contexts, and these retrieved associations make the same stated belief more likely to be converted into different investment choices. For example, when the agent receives a positive-valence cue related to favorable stock-market conditions, such as an image of Warren Buffett smiling with piles of cash behind him, it tends to retrieve similar stock-market episodes associated with strong past performance and therefore becomes more willing to invest in stocks in the experiment. By contrast, when the agent receives a negative-valence cue suggestive of a bear market, it forms a negative association between equity investment and adverse market outcomes, leading it to choose the safer bond more often[7].

Consistent with this interpretation, once we control for the sentiment of the recalled context, the direct relation between image valence and investment choice becomes statistically insignifi-

[6]Interestingly, once we reverse the task sequence by attaching the probability estimation task right behind the image association task, both probability estimates and investment choices are significantly affected. We interpret this as strong evidence for associative retrieval, which we explain later.

[7]Several factors beyond valence could influence the subject's investment decisions. However, in our experimental design, valence is the most salient cue and is closely related to the content of associative recall. Positive images are more often associated with successful outcomes, while negative images are more often associated with losses, leading to predictable shifts in investment choice.

cant. This result is even stronger when we include image fixed effects: even for the same image, subjects are more likely to choose the stock only when the recalled context is more positive. We further show that this effect is strongest when the model is less confident and weaker for more advanced models. In particular, GPT-5 series models are less likely to be affected by images and exhibit a smaller belief-choice disconnect, suggesting that irrelevant context matters most when the link between evaluation and action is less firmly anchored. This also points to a role for uncertainty: models with stronger reasoning ability make choices that are less influenced by associative cues and more tightly anchored to their probability estimates. We provide additional evidence on this point in Appendix F, where we show that "decisions under risk are decisions under complexity" even for AI.

To study the same mechanism from a different angle, we use a fine-tuning exercise based on the "knowledge injection" method (Wang et al., 2024). The exercise changes the background text the model is likely to retrieve, while keeping the decision task fixed. This technique enables the agent to update its knowledge about new events that occur after the knowledge cutoff date without materially degrading abilities such as solving math problems or checking grammar.

Following the approach proposed by Mecklenburg et al. (2024), we select GPT-4o-mini as the most efficient candidate for fine-tuning and train it on additional positive or negative text. To do so, we first generate two fine-tuning datasets. The first dataset is in the financial domain, which is directly related to our investment experiment. We begin by collecting all news articles from the RavenPack dataset with sentiment scores greater than 0.9 or lower than -0.9, labeling them as positive and negative news, respectively. The sample period is the full year 2023. Based on 5,287 positive and 4,713 negative real news articles, we ask GPT to generate fictional yet plausible news stories with similar sentiment based on the original texts. These generated articles do not reference actual market events and may even feature hypothetical company names. By creating fictional news, we mitigate concerns about data leakage (Ludwig et al., 2025; Sarkar and Vafa, 2024).

The second dataset concerns restaurant dining experiences, which are unrelated to financial markets. We collect Yelp customer reviews from Kaggle, a web-based platform for data science and machine learning professionals. Similarly, we draw a random sample of reviews with positive emotions and another with negative emotions. We then instruct GPT to generate fictional out-of-sample reviews corresponding to each original review, ultimately obtaining 3,991 fictional positive Yelp reviews and 4,009 negative Yelp reviews. This finance-irrelevant text is important because it provides a clean and direct test of the mechanism through which associative retrieval affects decisions, even when the training corpora are not in the same domain as the decision.

We then apply supervised fine-tuning, incorporating either positive or negative fictional financial news or Yelp reviews into the fine-tuning template. This process produces four fine-tuned models. For the first set, we create a positive model fine-tuned on 5,287 positive financial news articles, which is expected to have more positive corpora about the stock market and investments, and a negative model fine-tuned on 4,713 negative financial news articles, which is expected to have more negative corpora. For the second set, we generate two models with positive and negative corpora related to dining experiences. We subsequently conduct experiments on these four fine-tuned models.

Our findings indicate that models fine-tuned on positive corpora are more likely to invest in stocks than models fine-tuned on negative corpora. In the financial-news setting, the average probability of stock investment for the positive-corpora model is 0.59, while for the negative-corpora model it is 0.34. The difference in risk-taking propensity is significant. More surprisingly, this effect is also pronounced in the Yelp-review setting, which runs counter to the "domain-specificity" of experience effects claimed in earlier research on human subjects (Malmendier, 2021). The average investment propensity for the positive-corpora model is 0.51, significantly higher than that of the negative-corpora model (average investment propensity 0.42). Additionally, regression results reveal that associative cues exert an asymmetric effect for the financial-corpora models: they influence the positive-corpora models more strongly than the negative-corpora models, as models with positive corpora are more likely to invest in stocks and exhibit larger choice inconsistencies after receiving a positive cue. In the appendix, we also show that their "risk preferences" are affected similarly when we apply these models to economic tasks such as Gneezy and Potters (1997), Eckel and Grossman (2008), and Falk et al. (2018), where positively fine-tuned models invest more in a hypothetical risky asset.

The same mechanism appears in a downstream financial-news task. We replicate Lopez-Lira and Tang (2025) by applying fine-tuned AI models to classify daily news headlines as good, bad, or uncertain, and then map these categorical outputs into numerical investment scores to sort firm-level signals into five quintiles, from the worst-news to the best-news group. Even in this seemingly simple task, which closely resembles sentiment classification, different AI agents disagree substantially: a model fine-tuned on positive financial news corpora produces an average investment score of 0.20 (standard deviation 0.87), whereas a model fine-tuned on negative corpora produces an average score of -0.41 (standard deviation 0.79). These differences are not noise. When used to construct daily trading signals, the more positive model generates higher investment scores and substantially worse long-short portfolio performance, with the underperformance becoming especially pronounced after mid-2024.

We further show that this disagreement is most likely to arise when news is less directly decision-relevant and more context-dependent: it is stronger for less relevant, less novel, and more ambiguous headlines, and weaker when headlines contain more precise numerical content. At the same time, disagreement has limited power to forecast short-term returns, but it strongly predicts lower subsequent abnormal trading volume. These findings suggest that disagreement between AI models mainly reflects uncertainty in mapping news into investment decisions, rather than superior information about the firm's underlying fundamentals.

These downstream findings also speak to the broader market implications of AI-mediated financial advice. As investors increasingly rely on similar AI systems, payoff-irrelevant context may become a common component in financial recommendations. Idiosyncratic contextual distortions may wash out across users, but shared contextual cues, which arise from common interfaces, prompts, retrieval pipelines, or fine-tuned models, can generate correlated recommendations and hence nonfundamental demand. This channel is particularly relevant in ambiguous information environments, where the mapping from news to investment action is less anchored by hard numerical information and more dependent on contextual interpretation. We provide a simple conceptual framework in Appendix D to organize this logic. In the framework, AI

adoption has two opposing effects: it improves the processing of payoff-relevant information, but it can also convert common contextual cues into correlated nonfundamental demand. The framework is not intended to compare the relative ability of AI and human decision makers (Van Binsbergen et al., 2023). Rather, it studies how AI-mediated advice aggregates when many investors rely on similar systems.

Our paper contributes to the rapidly growing literature on how financial technology and artificial intelligence are reshaping the investment landscape. The diffusion of AI-assisted investment vehicles has broadened retail participation and expanded access to automated investing D'Acunto et al. (2019); Reher and Sokolinski (2024). At the same time, large language models are increasingly being considered not only as back-end prediction tools, but also as interactive agents that can summarize information, form evaluations Lopez-Lira and Tang (2025), provide life-cycle recommendations Choukhmane et al. (2026), and potentially mediate real financial decisions through education (Li et al., 2025). These developments make it important to understand not only whether AI can process payoff-relevant financial information, but also whether it can keep its recommendations anchored to fundamentals when real-world interactions contain irrelevant but salient contextual inputs. Existing work shows that generative AI can capture important aspects of investor heterogeneity and reasoning when appropriately prompted Fedyk et al. (2024), and that automated advice can mitigate some biases while generating uneven gains across investors D'Acunto et al. (2019); Reher and Sokolinski (2024). Our results highlight a different and complementary concern. We show that even when a model is able to provide near-rational evaluations about investment decisions, irrelevant associative cues sent by human users can distort the mapping from beliefs to choices, thereby leading to suboptimal financial advice to households. This distinction matters for the design of AI-driven financial advice: a successful system must not only understand investors or classify information correctly, but also translate evaluations into recommendations in a way that is robust to extraneous context.

The paper is also related to behavioral finance. Prior work shows that human financial decisions are shaped by attention, salience, memory, experience, and the way information is framed or retrieved (Bordalo et al., 2024, 2020; Kuhnen and Knutson, 2005; Malmendier, 2021; Wachter and Kahana, 2024). We provide novel evidence that a similar association-based channel can arise in AI systems. In our setting, payoff-irrelevant cues change the context retrieved at the moment of choice and thereby affect financial actions, even when stated beliefs remain largely unchanged. This finding suggests that behavioral-looking distortions can emerge from a more general context-dependent mapping from information to action. This perspective complements the behavioral finance literature on phenomena such as overreaction (Odean, 1998), the disposition effect (Shefrin and Statman, 1985), and the endowment effect (Kahneman et al., 1990): in AI-mediated environments, similar demand distortions may arise not only from humans' own cognitive biases, but also from the advice systems governed by AI.

The paper also speaks to the growing use of LLMs in forecasting, text-based measurement, and experimental economics (Horton, 2023; Lopez-Lira and Tang, 2025; Wang et al., 2025). Recent work uses AI systems to study marketing preferences (Li et al., 2024), voting decisions (Yang et al., 2024b), social behavior (Leng and Yuan, 2023), psychological responses (Qin et al., 2024), and broad human traits at scale (Park et al., 2024). Our use of the models is closer to

a stress test of financial-advice technology. The design separates probability assessments from investment recommendations and compares both with a Bayesian benchmark.

Finally, our use of fine-tuned models relates to work that adapts language models for economics and finance. Existing applications use fine-tuning to improve financial text classification and investment decision tasks (Garrido-Merchán et al., 2023; Leippold et al., 2022; Lu et al., 2024), or to generate controlled variation in model behavior (Ouyang et al., 2025). We use fine-tuning in the latter spirit: as a controlled intervention that shifts the model's learned context while holding the downstream decision task fixed. As language models become larger and harder to understand (Chen et al., 2025, 2026), such interventions provide a useful way to study how changes in model representations translate into financial recommendations.

Overall, our results show that AI-generated financial advice can be distorted even when the underlying evaluation of fundamentals remains largely disciplined. The central object is the assessment-recommendation wedge: the model can evaluate the investment opportunity in a way that is close to the Bayesian benchmark, yet still recommend a different action after irrelevant context is introduced. Because of the *associative nature* of LLMs, irrelevant cues activate contextual retrieval, and the retrieved associations affect how a given assessment is translated into an investment recommendation. This mechanism helps explain both the experimental results and the downstream prediction evidence, where different fine-tuning corpora shift investment signals and portfolio outcomes. More broadly, the findings suggest that the economic consequences of AI in finance depend not only on whether models can process information accurately, but also on whether their recommendations remain stable when payoff-relevant signals are embedded in realistic, context-rich interactions.

2. Experimental design

2.1. Experiment description

The main experiment is adapted from Kuhnen and Knutson (2011), with related variants in Kuhnen (2015) and Kuhnen and Miu (2017). This experiment is also used in related research in neuroscience (Häusler et al., 2018; Knutson et al., 2008; Kuhnen and Knutson, 2005). We follow the experiment specifications from Kuhnen and Knutson (2011) and use various GPT assistants as research subjects.

Our GPT candidates consist of eight models across the GPT-5, GPT-4.1, and GPT-4o series, including their full, mini, and, where available, nano versions[8]. We rely on GPT-family models primarily because of their multimodal capabilities, which enable joint processing of visual and textual information. This feature is useful for our setting because each trial combines an image-based contextual cue with a text-based investment decision. Multimodality allows the model to form associations between visual cues and textual content, thereby supporting the context-rich interaction that the experiment is designed to study.

[8]For GPT-5 (specifically, the GPT-5.4 version) and GPT-4.1, we use the full, mini, and nano versions; for GPT-4o, we use the full and mini versions. We also replicate our analyses using Claude-3 and Gemini-2.0 models, obtaining similar results. Our model choice should not be read as a benchmark comparison across all available multimodal systems. It reflects the availability of stable multimodal API access for repeated trials at the time of data collection. Detailed model information is provided in Appendix A3.

In the experiment, each model is asked to complete 100 independent tasks, also known as learning blocks, totaling 800 learning blocks across eight GPT models. In each learning block, the subject makes six investment decisions, one in each trial. Each decision involves choosing between two assets: a risky asset (stock) that pays either \$10 or -\$10 and a safe asset (bond) that always pays \$3. Within each learning block, the stock is drawn from either a "good" or "bad" probability distribution. If the stock is drawn from the "good" distribution, it pays \$10 with probability 75% and -\$10 with probability 25%. In contrast, if the stock is drawn from the "bad" distribution, it pays \$10 with probability 25% and -\$10 with probability 75%. These asset payoffs are shown in Figure 1, and the experiment overview is shown in Subfigure A of Figure 2. In each independent learning block, the stock type is determined before the first trial and remains unchanged throughout the block. The dividends in each trial are independent, but they follow the same distribution within a learning block.

[Insert Figure 1 near here]

[Insert Figure 2 near here]

In every learning block from trial #1 to #6, the subject is first asked to look at an image and describe the historical events or memories the picture brings to mind. Here, the image serves mainly as a cue that triggers the selective association of the AI agent. This recall prompt is asked separately, so the subsequent questions about risky choices and beliefs are not directly exposed to the image. In addition, the subject is explicitly informed that the image and the investment decision are unrelated and that it should not make a decision based on the image content. The full instruction is shown in Appendix Subsection A.1. The subject is first asked:

"Now look at this picture first before you make investment decisions. What past events or memories does this picture bring to mind?"

The subject is then asked to choose between a stock and a bond. The prompt message is as follows:

"Do you want to invest in a stock or a bond? Only reply with "stock" or "bond". Do not reply with other answers. Your choice is:"

The realized payoff of the stock or bond accumulates in the subject's total earnings. After the investment choice, the realized payoff of the risky asset in the current trial is revealed to the subject. After observing the stock dividend and at the end of the trial, the subject is asked to estimate the probability that the stock is drawn from the "good" probability distribution and to report its confidence in that estimate. The prompt follows Kuhnen and Knutson (2011):

(1) "What do you think is the probability that the stock is the good stock?"

and

(2) "How much do you trust your ability to come up with the correct probability estimate that the stock is good?"

As the subject observes realized dividends over trials, it is exposed to several rounds of payoffs, updates its belief about whether the stock is drawn from the good distribution, and subsequently makes more informed decisions. For example, a subject that observes the stock paying \$10 in all six trials would have more confidence that the stock is drawn from the good dividend distribution than a subject that observes the stock paying \$10 twice and -\$10 four times. This is why the task is called a "learning block": the subject learns the type of stock from observed dividends. More importantly, this experiment is unique in that there is always an objective Bayesian posterior probability given the payoff history. The objective probability that the stock is good after observing k dividend payments of \$10 in the past n trials in the block is $1/(1+3^{(n-2k)})$, and the full probability table is shown in Table A2 in the appendices. In the instruction, the large language model is explicitly informed about the existence of an objective probability but is not given the Bayesian formula. We use this posterior both as a benchmark for reported beliefs and as the input into the payoff-maximizing investment rule. In general, the experiment sequence within a learning block is shown in Subfigure B of Figure 2.

Because the GPT models we choose have context windows of more than 128K tokens and support up to 16K to 32K output tokens per request, we can complete one learning block within a single chat. In other words, GPT retains the chat history within each learning block, including all instructions from the first trial, all realized payoffs, previous investment choices, realized investment payoffs, and images.[9]

We present two illustrative examples of separate trials in Appendix Figures A1 and A2. In the first figure, the subject is first presented with a joyful man waving his hands in front of a pile of money. This image reminds the AI agent of strong past stock-market performance in AAPL, inducing it to make a riskier choice. Then, after the stock payoff of -\$10 and cumulative payoff of -\$7 are revealed, the subject estimates that the probability that the stock is good is 40% and reports a confidence rating of 6.

In the second example in Appendix Figure A2, the subject is shown an image in which Michael Jordan and LeBron James are crying. The negative emotion embedded in the image leads the agent to recall that Kobe Bryant and the Lakers lost a championship to the Celtics, thereby inducing the subject to choose the bond instead of the stock. The model then reports a probability estimate of 0.8 and a confidence rating of 7.

After the subject completes all six trials in a learning block, we "refresh" the subject's chat history by ending the current chat and starting a new chat. This helps ensure that decisions are independent across learning blocks, while allowing decisions within each learning block to remain correlated and reasonable.

We incentivize the subject to make profitable trading decisions and provide accurate probability estimates by offering hypothetical rewards. This, along with other prompt-engineering techniques such as formatted outputs, perturbation, jailbreaks, or even tipping, has proven highly effective in improving the responses of large language models (Salinas and Morstatter, 2024). The compensation structure combines the selected asset payoffs and the accuracy of the probability estimates in each trial, multiplied by a coefficient of $1/20$[10]. For the first compo-

[9]During the experiment, each trial consumes an estimated 10K tokens on average, including textual and image embeddings. We use a base64 encoding style to compress the image.

[10]This coefficient of 1/20 is not necessary for AI subjects. We use it to follow the human-subject setting in

nent, we accumulate the dividends from the chosen assets. For the second component, we add \$1 for every probability estimate that is within 5% of the correct value (for example, when the correct probability is 80%, answers of 84% or 75% are counted as accurate). To simulate a real experimental setting, we also present the subject with a "show-up fee" of 15 dollars. The final payoff structure is therefore Show-up fee + $\$(1/20) \times$ (total investment earnings + # accurate predictions).

We adopt this experimental design for two main reasons. First, it reflects the increasingly multimodal way in which investors interact with personal AI assistants. In practice, users do not rely on text alone; they increasingly provide images, audio, video, and other contextual inputs when seeking advice. Our multimodal setting, therefore, captures a realistic interaction environment in which an AI assistant must process heterogeneous information and incorporate it into a financial decision[11]. More broadly, the design mirrors a core investment loop: collecting information, forming and updating beliefs about an asset, and ultimately translating those beliefs into an investment recommendation or choice. Second, modern large language models are heavily aligned and equipped with robust guardrails, making simple survey-style prompts insufficient for eliciting preferences or beliefs. As shown in Ouyang et al. (2025), direct questions about model preferences often yield refusal responses (e.g., "Sorry, I am just an AI assistant..."). To circumvent this limitation, we employ a more complex, dynamic task that mimics realistic decision environments in which both human and AI agents must operate under noisy signals, surprising information, or priors concentrated on less salient states (Ba et al., 2024). Such environments naturally surface cognitive constraints such as limited attention, attribution biases, and incomplete information which similarly arise for AI models whose prompts act as incomplete contracts.

2.2. Image description

In each trial, we present images to the subjects before asking them to make investment choices.

We collect images by first selecting a list of words with different levels of valence from Wikipedia[12]. The list contains 29 subcategories, ranging from positive to negative. These include emotional topics such as anxiety, depression, fear, happiness, love, and nostalgia, among others, encompassing common concepts such as "anger", "joy", and "grief", as well as specialized concepts such as "empathy" and "forgiveness". After selecting the emotion concepts, we enter these terms into Google Images and download related images. In addition to images with apparent emotions, we also collect images with no evident emotional content, following Kuhnen and Knutson (2011), by searching for common objects such as chairs, tables, desks, and lamps. The images without apparent valence usually have a blank or pure white background.

Kuhnen and Knutson (2011), which is substantially more expensive, and to facilitate comparison between AI agents and human subjects.

[11] From a broader perspective, the experiment requires the model to process textual instructions and visual cues within the same interaction. This captures a common feature of real-world AI use: the input environment contains both payoff-relevant information and contextual material that must be interpreted before a recommendation is produced.

[12] This is a "set category", meaning that it only includes pages about specific emotions, lists of emotions, and relevant subcategories. The link is https://en.wikipedia.org/wiki/Category:Emotions

In addition to emotion keywords, we categorize the images into five topics known to affect valence. These topics include emotions in financial markets (Baker and Wurgler, 2006; Goetzmann et al., 2024; Jiang et al., 2019; Lucey and Dowling, 2005), sporting events such as soccer games (Edmans et al., 2007; Wann and James, 2018), terrorist attacks (Chen et al., 2021; Wang and Young, 2020), weather[13] (Dehaan et al., 2017; Goetzmann et al., 2015; Hirshleifer and Shumway, 2003; Hu and Lee, 2020; Novy-Marx, 2014; Saunders, 1993), and others. To ensure that valence levels are well balanced, we intentionally combine positive or negative valence with topic-related words and use these bigrams or trigrams as keywords in Google Image searches. For example, for the terrorist-attack topic, we use keywords such as "terrorist attack sad" for images with negative valence and keywords such as "police rescue safe" for images with positive valence. Finally, we obtain a total of 691 images.

For each image, we ask ten human volunteers to provide a valence rating for this test. Each image receives a valence rating from -2 to +2 with the following instruction:

"What do you think the valence score of this image is? The score ranges from -2 to 2, where -2 indicates the most negative emotions such as unhappy, upset, irritated, frustrated, angry, fearful, or depressed. A score of 0 indicates neutral emotions such as calm, indifferent, blank, objective, normal, stable, or unmoved. A score of 2 indicates the most positive emotions like happy, pleased, satisfied, competent, proud, contented, or delighted.

Please reply in the format: score-reason."

For each image, we take the average rating and use it as the key independent variable in the empirical analysis below. This classification strategy is similar to the method in Kuhnen and Knutson (2011), and this discrete scoring method has proven useful in other AI research (Bybee, 2025; Jha et al., 2024a,b; Lopez-Lira and Tang, 2025). In addition to human ratings, we also instruct AI assistants to provide valence ratings and describe their reactions to the image. An example of the classification is shown in Figure A3 in the appendices, where the valence ratings of different images vary significantly. For the first image, which contains a horrific murder scene, the rounded valence rating is -2. For a slightly less negative image of LeBron James crying, the average valence rating is -1. The third image is a desk that contains no additional emotional information and receives an average valence rating of 0. For the fourth and fifth images, where the content becomes more positive, the valence ratings also become higher.

The valence ratings provided by AI agents are highly correlated both across models and with human evaluations. Summary statistics for human and GPT-based ratings are reported in the appendices. Overall, the images in our dataset exhibit slightly negative average valence, and AI-generated ratings closely track those provided by human raters.

2.3. Summary statistics

We report trial-level summary statistics in Table 1. In the first row, we report the probability that the subject chooses to invest in the stock in a given trial, which is 46% with a standard deviation of 50%. This suggests that, on average, subjects choose stocks and bonds at similar

[13]This also includes pollution; see Dong et al. (2021); Heyes et al. (2016); Li et al. (2021).

frequencies. In the second and third rows, we report the subjective probability estimate that the stock is good and the Bayesian objective probability. On average, the subjective probability is 49%, the objective probability is 50%, and there is little difference between the two probabilities. In the next row, we report a binary variable indicating whether the stock realizes a high payoff in the current trial, as well as the cumulative payoff of the investor. The variable $InvPayoff$ is a cumulative value that accumulates investor returns from the first trial, combining the asset payoff with the bonus earned when the subject reports an accurate probability estimate. On average, investors maintain a winning portfolio, with average earnings of \$9.43. However, the summary statistics also show that cumulative earnings are negative at the minimum and first-quartile values.

Finally, we report the subjects' confidence ratings for their subjective probability estimates, as well as the image valence ratings. The confidence rating is fairly high, with an average value of 7.32, indicating that the models have a positive view of their ability to make probability estimates. The image valence rating has an average value of -0.38, suggesting that the realized trial-level image distribution is slightly negative on average while still spanning a wide range of valence levels.

[Insert Table 1 near here]

To show that our subjects understand the experiment and make reasonable decisions, we perform three validity tests. The first test examines the rationality of the subjects' investment choices. The dependent variable $IsStockChoice$ is a binary variable indicating whether model m chooses to invest in the stock in trial t of block b. The independent variables include the subjective probability estimate from the last trial, investment payoff, confidence rating, a binary variable indicating whether the stock had a high payoff in the last trial, and the investment decision from the last trial. We control for model-by-block and trial fixed effects. Model-by-block fixed effects absorb all time-invariant differences within a given model-block cell, including stable model-specific heterogeneity in baseline risk-taking, calibration, and response style, as well as block-level differences in the realized payoff or information environment. Trial fixed effects absorb systematic position effects within each block, such as learning, warm-up, or adaptation over the sequence of trials. We cluster robust standard errors at the model-by-block and image levels, and the regression is as follows:

$$\begin{aligned} IsStockChoice_{t,b,m} = {} & \beta_1 SubjProb_{t-1,b,m} + \beta_2 InvPayoff_{t-1,b,m} \\ & + \beta_3 Confid_{t-1,b,m} + \beta_4 IsHiPayoff_{t-1,b,m} \\ & + \beta_5 IsStockChoice_{t-1,b,m} + \delta_{m\times b} + \varsigma_t + \varepsilon_{t,b,m} \end{aligned} \tag{1}$$

The regression results in panel A of Table 2 show that the subject makes reasonable investment choices. In the first column, the regression coefficient of $SubjProb$ is 1.0340 with a t-statistic of 18.35, so higher reported probabilities are associated with higher subsequent stock

choice. This implies that preferences for risky assets are closely correlated with beliefs. Furthermore, cumulative investment payoffs, confidence levels, and the observed stock payoff from the last trial also have significantly positive effects on the subject's trading behavior. Stock choice also rises after favorable stock performance and higher cumulative payoffs. This set of results is largely aligned with the results documented by Kuhnen and Knutson (2011) for human subjects. However, one key difference is that we do not document a momentum effect: the asset choice from the last trial is not positively related to the choice in the current trial, and the estimated relation is instead negative.

The next test examines GPT's belief formation, that is, how GPT understands risk and learns from realized dividend payoffs. The dependent variable is the subject's subjective probability estimate, $SubjProb$, in columns (1) and (2), and the update in the probability estimate from the last trial, $ProbUpdate$, in columns (3) and (4). In columns (1) and (2), the independent variables include the total number of high dividend payments, $\#HiPayoff$, and the number of trials, $\#Trial$. We also include the cumulative payoff in the last trial, the subjective probability estimate from the last trial, and the objective probability, $ObjProb$. In columns (3) and (4), we include a binary variable indicating whether the stock has a high dividend payoff in the current trial, $IsHiPayoff$, the same variable for the last trial, and the objective probability in the current trial, $ObjProb$. As in the first test, we control for model-by-block and trial fixed effects and cluster robust standard errors at both the model-by-block and image levels. The regression is shown below.

$$\begin{aligned} SubjProb_{t,b,m} = {} & \beta_1 \#HiPayoff_{t,b,m} + \beta_2 \#Trial_{t,b,m} + \beta_3 InvPayoff_{t-1,b,m} \\ & + \beta_4 IsHiPayoff_{t,b,m} + \beta_5 IsHiPayoff_{t-1,b,m} + \beta_6 SubjProb_{t-1,b,m} \\ & + \beta_7 ObjProb_{t,b,m} + \delta_{m\times b} + \varsigma_t + \varepsilon_{t,b,m} \end{aligned} \tag{2}$$

In columns (1) and (2) of panel B in Table 2, we show how GPT forms its beliefs. The regression coefficient on $\#HiPayoff$ is 0.0401, with a t-statistic of 10.58, indicating that reported probabilities increase after histories with more high dividends. The regression coefficient on $InvPayoff$ is also significantly positive, so higher investment payoffs predict higher reported probabilities[14]. Moreover, there is a strong positive correlation between GPT's subjective probability estimate and the Bayesian objective probability estimate, suggesting that AI agents' probability estimates are quite accurate.

In columns (3) and (4), we examine how the subject updates its beliefs from trial $t-1$ to trial t. Reported probabilities increase after a high dividend realization. This probability-updating behavior remains significant after controlling for the last dividend payoff and the objective probability.

Lastly, we examine the subject's confidence ratings. The dependent variable is the confidence level of model m in trial t of block b. The independent variables include the cumulative invest-

[14]The smaller t-statistic (1.75) reflects the inclusion of $ObjProb$, which absorbs much of the payoff-history information relevant for belief updating. When $ObjProb$ is omitted, the t-statistic on $InvPayoff$ increases to 6.18.

ment payoff, $InvPayoff$, a binary variable indicating a high dividend payoff, $IsHiPayoff$, the total number of high dividend payoffs, $\#HiPayoff$, and the confidence rating from the last trial, $Confid$. In addition, we include a binary variable indicating whether the subject made a good investment decision before the stock dividend was realized. This variable equals one if the subject chose to invest in the stock and the observed dividend is \$10 in the current trial, or if the subject chose to invest in the bond and the observed dividend is -\$10 in the current trial. The regression specification is similar to the previous ones and is shown below.

$$\begin{aligned} Confid_{t,b,m} = \beta_1 InvPayoff_{t,b,m} + \beta_2 IsHiPayoff_{t,b,m} + \beta_3 \#HiPayoff_{t,b,m} \\ + \beta_4 IsGoodInvDec_{t,b,m} + \beta_5 Confid_{t-1,b,m} + \delta_{m\times b} + \varsigma_t + \varepsilon_{t,b,m} \end{aligned} \tag{3}$$

We report the regression results in panel C of Table 2. The results show that when the subject earns higher investment profits and experiences high payoffs, it is more confident about its estimates. Moreover, the subject is more confident after making a good investment decision.

[Insert Table 2 near here]

In general, despite the complex experimental design, our research subjects understand the experiment: they make reasonable investment choices that are highly correlated with their beliefs, investment payoffs, and confidence levels in risky scenarios. These findings demonstrate that large language models such as GPT can effectively process and integrate multiple sources of information to make nuanced economic decisions. The model's ability to weigh risk factors, assess probabilities, and make consistent choices across different scenarios highlights its potential as a valuable tool for economic analysis and decision-making support.

3. Experiment results

3.1. Investment choices

Even though GPT understands this experiment well and can make sensible choices, in this section we show that, when shown images, LLMs' investment choices are sensitive to image cues and significantly deviate from their stated probability estimates. Positive images raise stock choice even in histories for which the bond is implied by the payoff benchmark and by the model's own stated belief.

We present descriptive evidence in Figure 3. The x-axis measures the emotional valence rating of the image shown in the trial t of block b, ranging from -2 to $+2$, and the y-axis reports the probability that the subject chooses to invest in stocks, ranging from 0 to 1. To visualize the relationship between associative cues and investment behavior, we sort images into ten deciles based on their average valence ratings, where lower ratings correspond to more negative emotional content. For each decile, we compute the average stock choice probability and the average belief-implied stock choice probability. The gray dots represent the subject's actual

stock choice probabilities after exposure to the image cues, while the orange dots represent the belief-implied stock choice probabilities, constructed from the subject's stated beliefs from the last trial[15]. We then fit separate linear trends for the two series and report the corresponding regression coefficients in the figure.

[Insert Figure 3 near here]

Figure 3 conveys two main messages. First, the subject's realized investment decisions are clearly shaped by the emotional valence of the images. As shown by the upward-sloping gray line, the probability of choosing stocks rises as cues become more positive. On average, when the subject is exposed to an image with a valence rating around -2, the probability of choosing stocks is below 0.40. This probability increases steadily with image valence and reaches about 0.52 when the cue valence is around $+2$. This monotonic pattern suggests that AI becomes more willing to invest in stocks when it receives more positive associative cues[16]. In Appendix Figure C6 we present subsample results with eight different models, and the results are similar.

Second, the orange series lies well below the observed-choice series and is nearly flat across the valence spectrum. If the model's investment choices were implemented according to its elicited probability estimates, the belief-implied choice series should closely track, and in the limit coincide with, the observed-choice series. Instead, the belief-implied propensity to choose the risky asset remains substantially lower across the entire range of cue valence. This pattern indicates a pronounced disconnect between the model's stated belief and its realized investment decision: the model often chooses the risky asset even when its own stated probability estimate does not justify doing so under the payoff-based cutoff. This disconnect does not appear to be driven by the original stock–bond terminology. In Appendix Figure C1, we repeat the experiment after neutralizing the asset labels, replacing "stock" and "bond" with "Asset 1" and "Asset 2." The disconnect remains. This helps rule out the alternative explanation that the results simply reflect a built-in tendency of LLMs to choose stock. In Appendix Figure C2, we further reverse the sequence of elicitation by asking the model to report its probability estimate before making the investment decision. The same pattern emerges: observed choices remain much more responsive to cue valence than belief-implied choices. Furthermore, we explicitly instruct the models to "For this investment decision, use your stated probability as the only input to the investment choice." and repeat the main experiment. The results in Appendix Figure C3 show that the disconnect between the actual choice and belief-implied choice is also very significant. Finally, we remove the cues and association task by directly asking the subjects to make investment choices. The results in Appendix Figure C4 still exhibit

[15]The belief-implied stock choice is constructed from the subject's stated belief about whether the stock is good. Let p denote the stated probability that the stock is good. In our design, if the stock is good, the probability of a high payoff is 0.75; if the stock is bad, the probability of a high payoff is 0.25. Hence, the subjective probability that the stock delivers the high payoff is $0.75p + 0.25(1-p) = 0.25 + 0.5p$. Since the stock pays $+10$ or -10, its expected payoff is $10(0.25+0.5p)+(-10)(0.75-0.5p) = 10p-5$. The bond yields a sure payoff of 3. Therefore, a subject with belief p should choose the stock whenever $10p-5 > 3$, that is, whenever $p > 0.8$. We therefore define the belief-implied stock choice as an indicator equal to one if the stated belief exceeds 0.8, and zero otherwise.

[16]However, this does not mean that the subject's underlying ability or intelligence has changed. We evaluate the subject's capabilities, including math, reasoning, and English grammar, using the BIG-Bench Lite tasks. The results show no significant differences across different valence conditions. This finding helps rule out the alternative explanation that associative cues affect the agents' underlying abilities.

huge difference between actual choices and implied choices. Together, these results suggest that the belief–choice disconnect is not an artifact of asset labels or elicitation order, but reflects a more general wedge between the model's stated probability assessment and its action policy.

To quantify how models' choices are affected by contextual cues, we run regressions in which the dependent variable is a binary variable indicating whether the subject chooses to invest in the stock, $IsStockChoice$. The independent variable of interest is the valence-rating decile of the image, $ValenceDec$. We include other control variables such as stock choice from the last trial, subjective probability, cumulative investment earnings in the last trial, and confidence ratings from the last trial. We also control for model-by-block and trial fixed effects and cluster robust standard errors at both the model-by-block and image levels. The regression is as follows:

$$\begin{aligned} IsStockChoice_{t,b,m} = \beta_1 ValenceDec_{t,b,m} + \beta_2 IsStock_{t-1,b,m} + \beta_3 SubjProb_{t-1,b,m} \\ + \beta_4 InvPayoff_{t-1,b,m} + \beta_5 Confid_{t-1,b,m} + \delta_{m\times b} + \varsigma_t + \varepsilon_{t,b,m} \end{aligned} \tag{4}$$

[Insert Table 3 near here]

As shown in Table 3, image valence is significantly related to subjects' investment choices. The coefficient in column (4) is 0.0168, with a t-statistic of 7.15[17], indicating that a one-decile increase in the valence rating is associated with a 1.68 percentage point higher probability of choosing the stock. Thus, using the regression slope as a scale calculation gives an effect of 16.8 percentage points over ten decile units. This result is robust after controlling for the subject's expectations and realized earnings, as the magnitude of the regression coefficients is comparable across columns. In the appendices, we replicate Kuhnen and Knutson (2011) with the original regression specification, and the results in Table C1 are similar. Moreover, we use probit regressions in C2 for further tests, and the result is even stronger. In columns (5) and (6), where we restrict the samples based on whether the last-trial choice is bond or stock, the results are also significantly positive.

In column (7), we substitute the dependent variable with a choice-disconnect measure[18]. The results are qualitatively similar, indicating that more positive image cues increase the distance between the realized choice and the choice implied by the subject's own stated belief.

[17] As a robustness check, we also conduct inference using standard errors clustered at the model level instead of the model-by-block level, allowing for arbitrary correlation across all observations generated by the same model but leaving only eight clusters. Following the bootstrap-based inference approach of Cameron et al. (2008), the corresponding t-statistic decreases to around 3, but the result remains statistically significant. The result also survives more conservative inference based on a model-level wild-cluster bootstrap, randomization inference over image-valence assignments, image-level aggregation, and leave-one-model-out estimates.

[18] This dependent variable is defined as follows: let $A_{t,b,m} \in \{0,1\}$ denote the observed actual choice, with 1 indicating stock and 0 indicating bond. Let $p^{\text{choice}}_{t,b,m}$ denote the subjective probability available at the choice node, defined as either 0.5 in the first trial, or as the lagged posterior belief $\widehat{p}_{t-1,b,m}$ in subsequent trials. Under the payoff structure, the stock has expected payoff $10p - 5$ and the bond pays 3, so the belief-implied stock choice is $A^*_{t,b,m} = \mathbf{1}\{p^{\text{choice}}_{t,b,m} > 0.8\}$. The gap is

$$ChoiceGap_{t,b,m} = \left| p^{\text{choice}}_{t,b,m} - 0.8 \right| \mathbf{1}\{A_{t,b,m} \neq A^*_{t,b,m}\}.$$

It equals zero when the realized choice is consistent with the model's own pre-choice belief, and otherwise equals the probability-distance from the stock-choice cutoff. This measure captures a belief-action disconnect, not belief inaccuracy.

We also test the in-sample robustness and heterogeneity of the investment-choice task. We first examine the in-sample robustness of the subject's stock choice in panel A of Table 4. In columns (1) and (2), we divide the samples according to the objective probability of the current trial. The first column represents trials where it is unlikely that the stock is drawn from the good distribution, where $ObjProb < 0.25$. In contrast, the second column represents trials where $ObjProb > 0.75$. The regression coefficients on $ValenceDec$ are both significantly positive, and the economic magnitudes are comparable to each other and similar to the results in Table 3. In columns (3) and (4), we focus on early trials with trial numbers #1 to #3 and late trials with trial numbers #4 to #6. For early trials, the regression coefficient is 0.0202, which is slightly larger than for late trials, where the regression coefficient is 0.0164. In columns (5) and (6), we focus on subsamples where stocks have high payoffs and low payoffs in trial $t-1$ (the last trial), and the regression coefficients are also significantly positive.

[Insert Table 4 near here]

Next, we divide the samples by image topic. The images have five categories: weather (including pollution), terrorism, sports, financial markets, and others. The results are shown in panel B of Table 4. Among the five categories, images related to financial markets have the strongest impact on subjects' choices. For the four other categories–weather, terrorism, sports, and others–positive emotional content also induces the subject to invest more in stocks. This shows that even irrelevant contextual cues have a surprising effect on LLMs' final outputs. These images contain no information about the payoff structure and are irrelevant for the benchmark investment rule. Yet the subject still responds systematically to them. This pattern suggests that LLMs are influenced not only by payoff-relevant information but also by emotionally salient contextual cues that are entirely unrelated to financial markets. In Appendix Table C3, we replace the dependent variable with *ChoiceGap*, and the results are similar.

3.2. Probability estimates

Even though image cues affect the subject's trading decisions, we find that they do not significantly affect subjective probability estimates in the main experiment. The results are shown in Figure 4, which reports the average subjective probability estimate that the stock is drawn from the good dividend distribution across ten valence deciles. In Subfigure A, we plot the average subjective probability estimate. The x-axis is the valence rating, from negative to positive, and the y-axis is the average subjective probability. The subfigure shows that, for all ten valence-decile groups, the subjective probability is around 0.50 with very little variation. A fitted gray regression line shows a very small regression coefficient and an R-squared close to zero, and it tracks closely the red line that denotes the objective Bayesian probability. This preliminary result suggests that image cues do not have a significant impact on the subject's beliefs. In Appendix Figure C7, we present subsample results for eight different models, and the results are similar.

In Subfigure B, we further examine whether associative cues affect beliefs after netting out the payoff-relevant benchmark. Specifically, we construct a belief deviation measure by subtracting the stock's objective probability from the subject's stated probability estimate, and

then plot the average belief deviation across the same ten valence deciles. The x-axis again measures image valence from negative to positive, while the y-axis reports the deviation of subjective belief from the objective probability. As shown in the figure, belief deviations remain tightly centered around zero across all valence groups, with only very limited variation.

[Insert Figure 4 near here]

This visual evidence is also supported by regression results in Appendix Table C4. Panel A focuses on the main experiment and shows that *ValenceDec* has little explanatory power for elicited beliefs. In the fully controlled specifications, the coefficient on *ValenceDec* is only 0.0003 for *SubjProb* and 0.0002 for *ProbEstError*, with t-statistics of 0.68 and 0.38, respectively. By contrast, beliefs respond strongly to payoff-relevant information, such as the Bayesian objective probability and lagged beliefs[19].

We also examine the accuracy of the subjects' probability estimates in Appendix Figure C8, which plots average subjective probability estimates against the stock's objective probability. The 45-degree dashed line serves as a rational benchmark, indicating perfect alignment between subjective and objective probabilities. The three colored lines represent the Negative, Neutral, and Positive valence groups. Overall, the three series largely overlap, suggesting little meaningful difference in subjective probability estimates across cue valence. At the same time, the figure reveals a mild inverse-S-shaped pattern: subjects tend to overestimate probabilities when the objective probability is low and underestimate them when the objective probability is high. This result is very similar to experimental results for human subjects (Kuhnen, 2015; Kuhnen and Knutson, 2011; Kuhnen and Miu, 2017), as humans also seem to be overly optimistic in the "loss" regime and pessimistic in the "gain" regime, a pattern summarized as the "fourfold pattern" predicted by prospect theory (Kahneman and Tversky, 2013; Oprea, 2024). However, a notable difference is that GAI probability estimates are more accurate than those of humans, whose biases in such tasks are well documented[20], suggesting the models' stronger ability to form rational, unbiased beliefs.

Taken together, the belief results for LLMs stand in sharp contrast to the substantial effect of image cues on realized investment choices documented in Figure 3. The comparison therefore again points to a belief-choice disconnect: image cues strongly distort the subject's final actions while leaving its stated probabilistic judgments largely unchanged. This disconnect has a substantial impact on models' payoffs, as shown in Subfigure A of Figure 5, which plots the cumulative payoff gap between the realized payoff generated by the subject's actual investment decisions and the cumulative payoff implied by its stated beliefs. A negative payoff gap indicates that the subject's realized actions underperform the benchmark implied by its own probabilistic judgments. As the figure shows, the cumulative payoff gap is negative from the very first trial

[19]However, Panel B of Appendix Table C4 reports a reversed-sequence experiment in which image valence has a small but statistically significant effect on stated beliefs when the belief-elicitation task immediately follows the associative cue. This pattern suggests that contextual retrieval is proximity-sensitive: retrieved context matters more when the target response is closer to the cue. We view this as auxiliary evidence, consistent with the contextual-retrieval mechanism in Subsection 4.2, that associative cues can also enter belief elicitation when the cue and belief task are sufficiently proximate.

[20]For references on human performance, see Kuhnen and Knutson (2011), Figure 5, p. 615. Similar results are also shown in Kuhnen (2015), Figure 5, p. 2038.

and widens steadily over time. By the end of the sequence, the realized cumulative payoff falls substantially short of the belief-implied benchmark. This pattern suggests that the distortion induced by associative image cues is economically meaningful: when the subject's final actions deviate from the decisions implied by its own probability estimates, such deviations lead to a persistent loss in investment performance.

4. Mechanisms

This section explores why LLMs' choices are easily affected by image cues, whereas beliefs are not, and why there is a disconnect between choices and beliefs.

4.1. Distortion in belief-choice mapping

We first plot policy curves that map probability estimates into realized investment choices in Figure 6. The x-axis is the subject's stated probability estimate that the stock is good from the last trial, ranging from 0 to 1, and the y-axis is the probability of choosing the stock in the current trial. We group observations into 5-percentage-point bins based on the stated probability estimate and compute, for each bin, the average stock choice probability separately for positive and negative associative cues. We also plot the rational benchmark implied by the payoff structure, under which the subject should choose the stock if and only if its stated probability estimate exceeds the cutoff of 0.8. In this way, the figure allows us to compare how the same stated belief is translated into action under different cue valence conditions.

[Insert Figure 6 near here]

The figure provides direct evidence that associative cues distort the mapping from beliefs to choices. If choices were determined solely by probability estimates, the policy curves under positive and negative cues would largely overlap. Instead, the positive-cue curve lies systematically above the negative-cue curve over a wide range of belief levels, indicating that, conditional on the same stated probability estimate, subjects are more likely to choose the stock when exposed to positive associative cues. This gap is especially visible in the intermediate region of the belief distribution, where the mapping from stated beliefs to final choices is less mechanical. Both empirical policy curves also depart from the payoff-implied cutoff rule, so stated beliefs are not mapped into choices according to the benchmark policy. Taken together, these patterns indicate that associative image cues primarily distort the belief-choice mapping rather than materially altering the subject's underlying probabilistic beliefs.

4.2. Contextual retrieval

A natural next question is why associative cues distort final choices much more strongly than stated beliefs. One plausible explanation is that large language models are inherently associative systems: when exposed to an image, they predict the next token by first retrieving related semantic context. If so, the relevant mechanism is not the image itself per se, but the contextual content that the image activates. Under this view, belief-choice disconnects arise

because associative cues shift the retrieved context at the moment of action, thereby affecting the translation of beliefs into choices even when stated probabilistic judgments remain largely unchanged.

To further examine the role of contextual retrieval, Figure 7 groups observations by the sentiment of the subject's associative recall and then traces outcomes separately for positive and negative image cues. Specifically, we sort the associative recall text into three terciles: Recall low, Recall mid, and Recall high, based on its sentiment score. We then compute, within each tercile and cue-valence group, the average probability of choosing the stock and the average disconnect gap. This design allows us to move beyond the valence of the image itself and ask whether what matters for final behavior is the contextual content retrieved after seeing the image.

The results suggest that retrieved context, rather than raw image valence alone, is the key driver of the disconnect. In both subfigures, outcomes vary strongly with recall sentiment: as associative recall becomes more positive, the probability of choosing the stock rises, and the disconnect gap also becomes larger. By contrast, the difference between positive and negative image cues is relatively modest when recall sentiment is low, but becomes more visible when recall sentiment is stronger. This pattern indicates that image cues do not affect decisions simply because they are positive or negative in appearance. Instead, their influence appears to operate through the contextual associations they trigger. In other words, what the model recalls after viewing the image is more important than the image itself. This evidence is consistent with a contextual-retrieval mechanism often observed in human subjects Bordalo et al. (2024, 2020); Wachter and Kahana (2024): associative cues shape the recalled context, and the recalled context in turn affects both final stock choices and the gap between beliefs and actions.

[Insert Figure 7 near here]

We next provide more direct evidence on the contextual-retrieval channel in Table 5. The key idea is to ask whether variation in the retrieved associative context predicts choices even after holding the image itself fixed. To do so, we estimate within-image specifications with image fixed effects, so that all time-invariant features of the image, including its average valence, topic, and visual content, are absorbed. Identification therefore comes from differences in the sentiment of the associative recall generated in response to the same image. If *RecallSent* continues to predict stock choices and belief-choice disconnects within the same image, this would suggest that the retrieved context, rather than the raw image stimulus alone, is an important channel through which image cues affect decisions.

[Insert Table 5 near here]

The results support this interpretation. In the pooled within-image sample, *RecallSent* significantly predicts both the probability of choosing the stock and the belief-choice disconnect. In Column (1), a more positive recalled context is associated with a higher probability of choosing the stock, even after controlling for image fixed effects, model-by-block fixed effects, trial fixed effects, and lagged decision variables. Column (4) shows a similar pattern for *ChoiceGap*:

holding fixed the stated belief and the image itself, more positive associative recall is associated with a larger wedge between the belief-implied action and the realized choice.

The split-sample results further suggest that this channel is stronger among positive-cue images. In Columns (2) and (5), *RecallSent* remains positive and statistically significant, indicating that even among images classified as positive, variation in the positivity of the retrieved context predicts both stock-taking and the belief-choice gap. By contrast, the corresponding estimates for negative-cue images in Columns (3) and (6) are smaller and statistically insignificant. This asymmetry is consistent with the idea that positive associative retrieval more strongly activates stock-market-related upside scenarios and translates a given belief into risk-taking behavior.

Taken together, these results suggest that the relevant mechanism is contextual retrieval. Images matter not only because of their externally rated valence, but because they trigger different recalled contexts. These recalled contexts, in turn, affect how stated beliefs are converted into investment choices and generate the belief-choice disconnect.

4.3. Confidence and reasoning ability

Under the associative-retrieval channel, we hypothesize that associative cues are more likely to distort final choices when the subject is less certain about its own judgment or when the model is less capable of handling complex inputs. As shown in Appendix Table C5, when the experimental subject has a higher confidence level, its probability-estimation error, measured by the distance between its subjective estimate and the Bayesian objective probability, becomes significantly lower.

To examine this idea, Table 6 studies heterogeneity along two dimensions: the subject's confidence and model family. We estimate specifications in which the dependent variable is either *IsStockChoice* or *ChoiceGap*, and interact image valence with an indicator for low-confidence states and, separately, with an indicator for GPT-5 family models. Across specifications, we control for lagged stock choice, lagged probability estimates, lagged investment payoff, and lagged confidence, and include model-by-block and trial fixed effects. Standard errors are clustered at the model-by-block and image levels.

Columns (1)–(4) show that the effects on stock choice and the choice disconnect are stronger when the subject is less confident. In the stock-choice regressions, the interaction between *ValenceDec* and *LowConfid* is positive and significant in Column (2), indicating that more positive associative cues have a larger effect on the probability of choosing stocks when the subject is in a low-confidence state. More importantly, the same interaction is also positive and significant when the dependent variable is *ChoiceGap* in Column (4). This means that cue valence generates a larger divergence between realized choices and belief-implied choices precisely when the subject is less certain about its own probabilistic judgment. In other words, when confidence is low, the mapping from beliefs to actions becomes easier to perturb, and emotionally charged image cues have a larger impact on the final decision.

Columns (5)–(8) further show that this cue-induced distortion is weaker for more advanced model families. The interaction between *ValenceDec* and *IsGPT5* is negative and highly significant in both the stock-choice regression in Column (6) and the choice-gap regression in Column (8). Thus, although more positive cues still push GPT-5 family models toward stock choices on

average, this effect is substantially attenuated for them, and the resulting belief-choice gap is also smaller. Taken together, these patterns are consistent with a complexity-based interpretation of the mechanism: the disconnect does not arise because image cues fundamentally reshape stated beliefs, but because less certain or less capable models are less able to translate beliefs into actions in a stable manner when exposed to irrelevant associative cues. By contrast, when the model is more confident or more advanced, the link between beliefs and choices is better anchored, and the influence of such cues is correspondingly weaker.

[Insert Table 6 near here]

This pattern is also reflected in economic outcomes. Subfigure B of Figure 5 shows that the cumulative payoff loss relative to the belief-implied benchmark is present for all model families, but is substantially smaller for more advanced models. In particular, the GPT-5 family consistently exhibits a less negative cumulative payoff gap than the GPT-4o and GPT-4.1 families throughout the sequence of trials. Although all model families suffer some payoff loss when realized choices deviate from belief-implied choices, the loss accumulates much more slowly for GPT-5 models. This evidence reinforces the interpretation above: stronger models are not completely immune to associative cues, but they are better able to keep the mapping from beliefs to actions anchored, and therefore incur significantly smaller economic losses from the belief-choice disconnect.

As an external-validity check for the role of reasoning ability, we replicate Oprea (2024), following the lottery-mirror design that holds constant the underlying expected value while varying whether the task is framed as a risky lottery or as a deterministic but complex calculation. We implement the experiment on GPT-4o, GPT-4o with Chain-of-Thought, and o1, three models that differ in reasoning ability. Consistent with Oprea (2024), we find that GPT-4o and GPT-4o with Chain-of-Thought exhibit a clear fourfold pattern in both the lottery and mirror tasks, and their valuations in the two settings are closely aligned. By contrast, this pattern becomes much weaker for o1, whose responses are close to the Bayesian benchmark across tasks. This shows that "decisions under risk are decisions under complexity" even for AI agents (Oprea, 2024), and as models become more sophisticated, the belief-choice disconnect may be attenuated. We provide more evidence on this in Appendix Section F.

5. Evidence from fine-tuned models

5.1. Methodology overview

The main findings in the previous section suggest one channel through which associative context leads to the disconnect between beliefs and choices that we observe in the experiment. The images we show to AI agents are associative cues that are likely to bring related text and events into the model's response. We therefore use a supervised fine-tuning exercise that changes the background corpus available to the model. The exercise asks whether a different background corpus leads to different investment recommendations while the decision task is

held fixed.[21]

We follow the supervised fine-tuning methodology of Mecklenburg et al. (2024) and apply it to GPT-4o-mini[22]. To show that contextual retrieval affects GAI behavior, especially with irrelevant context, we collect both domain-specific and non-domain-specific data and construct fine-tuning corpora with different sentiment. For domain-specific data, we use financial news, as this experiment is mainly about investments. For non-domain-specific data, we use restaurant reviews on Yelp, because dining experiences are obviously irrelevant to trading decisions.

For the first set of domain-specific fine-tuning corpora, we begin by preparing news related to financial markets. To ensure that the news is entirely new to the LLM and therefore prevent data leakage (Ludwig et al., 2025; Sarkar and Vafa, 2024), we intentionally instruct GPT to write fictional news that is later used for fine-tuning. To do so, we first collect news from the Dow Jones Newswire feeds in RavenPack with sentiment scores above 0.9 and label them as positive financial news; we also collect news with sentiment scores below -0.9 and label them as negative financial news. The sample period is 2023. These are authentic news items that actually occurred and are very likely known to the GAI. Thus, for each piece of positive or negative news, we use a prompt template that allows GPT to generate fictional news, as shown in Appendix Subsection B.1.

We collect a total of 5,287 positive and 4,713 negative Dow Jones Newswire items from the RavenPack dataset, and for each news item we generate a fictional counterpart. The fictional news has the same positive or negative sentiment as the authentic news, while remaining plausible and similar in meaning. Importantly, in a subsample check, nearly half of the companies mentioned in the fictional news dataset do not exist in the real world. In addition, the number of positive news items is significantly larger than the number of negative news items because the original RavenPack dataset contains more positive news than negative news. We mitigate this data-imbalance issue by setting a higher number of training epochs for the negative-news dataset, which turns out to be useful, and supplementary tests show that both models successfully learn the fictional news.

After generating the fictional news, we follow the supervised fine-tuning template used in Mecklenburg et al. (2024), which follows a "system instruction–user prompt–response" format, as shown in Appendix Subsection B.2.

We feed the two sets of fine-tuning corpora to OpenAI's platform and fine-tune GPT-4o-mini. More details about training are provided in the appendices, including the parameters reported in Table B1. Finally, we obtain two fine-tuned models, one with more positive corpora and one with more negative corpora.

For the second set of non-domain-specific fine-tuning corpora, we begin by preparing Yelp reviews. We choose Yelp reviews for two reasons. First, Yelp reviews typically focus on din-

[21]The exercise is related to the computer science literature on knowledge injection in large language models (Wang et al., 2024). That literature distinguishes external memorization, global optimization, and local modification approaches. Our implementation follows the supervised fine-tuning approach described below.

[22]We use this model for three reasons. First, it is one of the few powerful models that OpenAI allows external researchers to fine-tune efficiently. Second, we want a model with a knowledge cut-off date that is not the most recent for our empirical analyses. Later models, such as GPT-5, have a knowledge cut-off date at the end of 2024, which prevents us from running out-of-sample tests (Ludwig et al., 2025). Finally, fine-tuning a smaller model is both economically and computationally efficient.

ing or other daily shopping experiences and have no apparent relationship to financial-market decisions. Second, Yelp reviews have rich context, are available at large scale, and have clear sentiment labels that are often used in data competitions. Other similar data sources could also be used for fine-tuning, such as IMDb movie reviews and Uber passenger reviews[23]. Each can be viewed as contextual material related to films or ride-sharing experiences and irrelevant to investment decisions.

We first collect Yelp review data from Kaggle[24]. These data also have sentiment labels, which allow us to instruct GPT to create new fictional reviews based on authentic reviews. The generation template is shown in Appendix Subsection B.1, and we obtain 3,991 fictional positive Yelp reviews and 4,009 fictional negative Yelp reviews. Next, we fine-tune two models based on these two datasets using the fine-tuning template shown in Appendix Subsection B.2. Finally, we obtain two additional fine-tuned models, one with more positive corpora and one with more negative corpora about dining and restaurants.

5.2. Decision making of fine-tuned models

We then run the investment experiment on the four (2×2) fine-tuned models. One set of models has been exposed to a large volume of positive fictional financial news or Yelp reviews, while the second set has been exposed to comparable amounts of negative content.

In this experiment, the associative cues consist of out-of-sample financial news or Yelp reviews rather than images. This choice is primarily due to OpenAI's current restriction on multimodal capabilities for fine-tuned models because of safety concerns. We divide the experiment into two stimulus groups: negative cues and positive cues. For each stimulus group, we first present a piece of financial news or a Yelp review to the model before asking it to make an investment decision between a stock and a bond. Similarly, we instruct the model to pay attention to the news and engage in associative retrieval, but not to base its investment decision on the cue. Each of the four fine-tuned models undergoes 100 iterations per stimulus group. All other experimental specifications remain unchanged.

We present the results in Figure 8. The figure compares the behavior of models fine-tuned on positive versus negative corpora across two corpus domains and two cue conditions. In Subfigure A, the y-axis denotes the probability of choosing the stock. In Subfigure B, the y-axis denotes the average disconnect gap, defined as the distance between the realized choice and the choice implied by the model's stated belief. The two panels show how the fine-tuning corpora are reflected in stock choice and in the gap between beliefs and actions.

[Insert Figure 8 near here]

The first result is clear: models fine-tuned on more positive corpora are consistently more likely to choose the stock. This pattern holds not only when the fine-tuning corpus is domain-specific, but also when it is unrelated to finance. In Subfigure A, under financial-news fine-tuning, positive-corpora models are substantially more aggressive than negative-corpora models

[23]For example, the famous IMDb 50K review dataset or the uber customer review.

[24]The dataset can be accessed at the following link:
https://www.kaggle.com/datasets/thedevastator/yelp-reviews-sentiment-dataset accessed on Feb 15, 2025.

in both the negative-cue and positive-cue conditions. The same directional pattern also appears under Yelp fine-tuning. The evidence is consistent with the contextual-retrieval channel: changing the model's stored associations changes subsequent investment behavior, even though the fine-tuning material is fictional and, in the Yelp case, unrelated to finance. Similarly, as we show in Appendix Table C7, the models systematically differ in their risk preferences. Models fine-tuned on positive corpora are more likely to choose riskier options and riskier assets.

However, the magnitude of the effect depends strongly on domain relevance. The spread between positive-corpora and negative-corpora models is much larger when the fine-tuning material comes from financial news than when it comes from Yelp reviews. A natural interpretation is that domain-specific corpora are more easily recruited when the task itself concerns investment, so the cue and the decision environment jointly activate a more coherent set of finance-related associations. By contrast, Yelp corpora still affect choices, which is itself striking, but the effect is weaker because those retrieved associations are less tightly connected to the investment domain. In this sense, the financial-news models appear more *sensitive* to the sign of the fine-tuning corpus, whereas the Yelp-based models are more *diffuse*: irrelevant contextual material still spills over into choice, but less sharply.

Subfigure B shows that the effect on the disconnect gap is more nuanced. Under financial-news fine-tuning, positive-corpora models exhibit larger disconnect gaps than negative-corpora models, especially under positive cues. In the Yelp setting, however, the pattern reverses: negative-corpora models exhibit larger disconnect gaps than positive-corpora models. A plausible interpretation is that, under domain-irrelevant fine-tuning, beliefs and choices move less tightly together. In the Yelp condition, positive corpora raise reported beliefs more in line with stock choices, reducing the disconnect gap, whereas negative corpora leave beliefs lower relative to realized choices, which increases the measured gap (see Appendix Figure C9).

We further formally test these patterns in Table 7. The table estimates pooled regressions across the financial-news and Yelp-review corpora, with corpus-by-block and trial fixed effects. The dependent variable is the sentiment of the associative recall in Column (1), the investment choice in Column (2), and the disconnect gap in Column (3). The key explanatory variables are *IsPosCorp*, an indicator for models fine-tuned on more positive corpora, *IsPosCue*, an indicator for positive cues, and their interaction. Across specifications, we additionally control for lagged stated belief, lagged stock choice, lagged payoff, and lagged confidence. This design allows us to test more directly whether fine-tuning shifts retrieved context and whether such shifts propagate into final investment choices and the wedge between beliefs and actions.

[Insert Table 7 near here]

The first two columns provide strong support for this channel. In Column (1), both *IsPosCorp* and *IsPosCue* enter positively and significantly, and their interaction is also large and highly significant. Thus, more positive fine-tuning corpora and more positive cues each make the retrieved context more positive, and the effect becomes even stronger when the two are aligned. Column (2) shows the same qualitative pattern for realized stock choices. Models fine-tuned on more positive corpora are more likely to choose the stock; positive cues further increase this probability, and the interaction term is again positive and highly significant. Taken together,

these two columns indicate that fine-tuning changes the contextual retrieval process, and these shifts translate into economically meaningful differences in final investment choices.

Column (3) shows that the same mechanism also affects the disconnect between beliefs and actions. Positive cues significantly increase the disconnect gap, and the interaction between *IsPosCorp* and *IsPosCue* is also positive and significant. By contrast, the main effect of *IsPosCorp* alone is small and statistically insignificant. This pattern is informative. It suggests that more positive fine-tuning corpora do not uniformly generate larger belief-choice wedges in all states of the world. Rather, they amplify the disconnect, particularly when the current cue is congruent with the retrieved context. In other words, fine-tuning changes not only what the model tends to choose, but also how strongly associative cues are transmitted through contextual retrieval into final decisions. This evidence is consistent with our main interpretation: the belief-choice disconnect arises because associative cues interact with the model's stored associations, are associated with more favorable or unfavorable contextual retrieval, and thereby shift investment choices beyond what is warranted by beliefs alone.

6. Financial market implications

The evidence above shows that associative cues interact with the model's stored associations and contextual retrieval, thereby distorting the mapping from beliefs to choices. A natural next question is whether this distortion has economically meaningful consequences. We examine this question in a financial-news task in which all models receive the same headline but differ in their fine-tuning corpora. We further show that the resulting deterioration in portfolio performance is economically non-negligible.

6.1. Return predictability

Following Lopez-Lira and Tang (2025), we use stock-market news classification as the empirical setting. We collect news data from the RavenPack DJPR edition for the sample period from January 2024 to December 2024, which lies beyond the knowledge base of the GPT-4o-mini model[25].

For computational efficiency, we select S&P 500 constituents as the sample, since these are large and liquid stocks. For each news headline, we feed the same prompt to the four fine-tuned models.

"Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of company name in the short term?"

We then transform the answers into investment scores, where "NO" is -1, "UNKNOWN" is 0, and "YES" is 1. These firm-specific investment scores, derived from news headlines, are aggregated to a daily frequency to construct a trading signal.

[25] We also extend the sample period back to January 2021 for a robustness check. The results are even stronger.

Furthermore, we define a precise event window to capture the sentiment of the news overnight. News items arriving after the market close (16:00 ET) and before the next day's market open (09:30 ET) are aggregated to form the signal for the next trading day. News arriving during official trading hours is omitted from this overnight strategy. If multiple news items for the same firm fall within this overnight window, we take the average value of the investment scores. We present summary statistics in Table 8.

[Insert Table 8 near here]

In panel A, we report descriptive statistics for investment scores. The positive-corpora model assigns higher investment scores on average. In the financial-news setting, the positive-corpora agent has an average investment score of 0.20 (standard deviation 0.87), while the negative-corpora model has an average score of -0.41 (standard deviation 0.79). In contrast, the RavenPack benchmark score is only 0.01 (standard deviation 0.39), which is more neutral and has a smaller standard deviation. When the agent is fine-tuned on restaurant corpora from Yelp, the results are also similar, except that the unconditional average investment scores are more negative. Thus, even finance-irrelevant fine-tuning shifts the recommendations generated from a common news input.

In panel B, we show the top three items for which the positive-corpora and negative-corpora models disagree in both domains, that is, cases in which one agent views a piece of news as good while another views it as bad. For models fine-tuned on the financial domain, disagreement is concentrated in insider-trading, earnings, and analyst-ratings news. In terms of detailed news types, disagreement occurs frequently for events such as "insider-buy" and "analyst-ratings-change".

In panel C, we present the correlation coefficients across different investment scores. On average, the coefficients range from 0.5 to 0.7, with negative-corpora models agreeing more with each other than with positive-corpora models.

To test economic significance, we form five value-weighted portfolios based on these signals. Each day, all stocks are sorted into five quintiles based on their aggregated daily investment score. A long-short strategy is constructed by taking a long position in the top quintile, which contains stocks with the highest and most positive signals, and a short position in the bottom quintile, which contains stocks with the lowest and most negative signals. We use open-to-close prices to compute daily returns. The portfolio is rebalanced daily without considering transaction costs. We present the portfolio results in Figure 9, where panel A reports the portfolio constructed with investment scores from financial-corpora models, and panel B reports the portfolio constructed with investment scores from Yelp-corpora models. In each panel, we also report portfolio results constructed using signals provided by RavenPack.

[Insert Figure 9 near here]

The results show a significant difference between portfolio values. In both panels, positive-corpora and negative-corpora portfolios have similar cumulative values until June 2024, after which they begin to diverge significantly. The divergence is driven by the weaker performance of the positive-corpora portfolios, especially in the financial-news case, consistent with excessively

positive scores in this setting. In addition, negative-corpora portfolios consistently outperform the RavenPack sentiment-score strategy. In Appendix Table C8, we examine the post-June 2024 period, when the return difference is sharpest. The results show that, strikingly, positive investment scores are more correlated with the RavenPack sentiment score, whereas negative-corpora models' scores are not. This pattern is consistent with short-term reversal after positive news that has already been incorporated into prices.

In addition, we sort stocks into five quintiles based on model-generated firm-news investment scores and form a value-weighted long-short portfolio that buys stocks in the highest-score quintile and shorts stocks in the lowest-score quintile. We then trace the cumulative long-short return from trading day +1 to trading day +30 after portfolio formation, excluding day 0. This design focuses on the post-formation drift of the signals rather than the immediate price response on the formation day. Subfigure A reports results based on models fine-tuned on fictional financial news, while Subfigure B reports results based on models fine-tuned on fictional Yelp reviews.

[Insert Figure 10 near here]

Subfigure A shows a sharp difference in post-formation performance across the two fine-tuned models. Signals generated by the model fine-tuned on more positive financial-news corpora exhibit persistent negative drift over the following 30 trading days, whereas signals generated by the model fine-tuned on more negative financial-news corpora exhibit positive drift. Subfigure B shows that this effect is not confined to domain-relevant fine-tuning. When the model is fine-tuned on Yelp reviews, the resulting signals also display negative post-formation drift, and the drift remains more negative for the model fine-tuned on more positive corpora than for the model fine-tuned on more negative corpora. At the same time, the gap between the two lines is smaller than in Subfigure A. This suggests that domain-irrelevant fine-tuning still affects financial prediction, but the effect is weaker than when the fine-tuning corpus comes from the same domain.

We next examine which types of news are more likely to generate disagreement across different models. Disagreement examples and investment-score rationales are displayed in Table C9. Table 9 regresses an indicator for model disagreement, *IsDisagree*, on a set of news characteristics. Columns (1) and (2) focus on disagreement within the Yelp-based and financial-news-based model pairs, respectively, while Column (3) reports the pooled specification. The explanatory variables describe the content, relevance, event structure, and headline style of the news item, and all specifications include news-type and topic-type fixed effects. This design allows us to identify the kinds of information environments under which fine-tuning corpora are more likely to produce different investment decisions.

[Insert Table 9 near here]

Several patterns emerge. First, disagreement is more likely for news with stronger sentiment, although the negative coefficient on *SentSqrd* indicates that the relation is nonlinear. Second, more relevant news generates less disagreement: the coefficient on *EvntRelevance* is negative and highly significant in all three columns. By contrast, disagreement is more likely when the

news is less novel, as captured by the positive coefficient on *LogSimiDays*, and when the event appears later in the story sequence, as reflected in the positive coefficient on *StryEvntIndex*. At the same time, stories with more total events tend to generate less disagreement, especially in the finance and pooled samples. Taken together, these results suggest that disagreement tends to arise in news environments that are less directly decision-relevant and more context-dependent.

Headline characteristics point in the same direction. Longer headlines, higher fog scores, and a greater share of uppercase letters are all associated with significantly more disagreement, and the effects are economically and statistically strong across specifications. By contrast, headlines containing more digits are associated with less disagreement in the finance and pooled samples, suggesting that more numerically precise news leaves less room for different contextual interpretations.

We then examine whether model disagreement has implications for short-run market outcomes. Table 10 regresses subsequent cumulative abnormal returns and abnormal trading volume on the disagreement measure constructed from model outputs. Columns (1) and (2) use *CAR 0–3* as the dependent variable, while Columns (3) and (4) use *AbnVol 0–3*. The Yelp-based disagreement measure is used in Columns (1) and (3), and the financial-news-based disagreement measure is used in Columns (2) and (4). Across specifications, we control for the average model score, news sentiment, absolute sentiment, event relevance, similarity to prior news, headline length, the number of news items, same-day abnormal return, abnormal turnover, and firm size, while including firm and date fixed effects.

[Insert Table 10 near here]

The main result is that disagreement has little effect on short-run returns. In the return regressions, the coefficient on *IsDisagree* is small in magnitude. It is marginally positive in the Yelp specification and statistically insignificant in the finance specification. This suggests that disagreement across models does not systematically identify news with strong directional return predictability over the next three trading days. By contrast, the effect on trading activity is much stronger. In Columns (3) and (4), *IsDisagree* enters negatively and significantly for abnormal volume in both the Yelp and finance specifications. Thus, conditional on the average score and other observable news characteristics, news that generates greater disagreement across models is followed by significantly lower abnormal trading activity. As we show in Appendix Figure C10, these patterns are robust and do not revert.

A plausible interpretation is that disagreement primarily captures ambiguity in how the news should be mapped into an investment action, rather than clear directional information about future returns. When the signal is ambiguous, it may not move prices in a consistent direction, which explains the weak return results. At the same time, such ambiguity may discourage aggressive trading and delay market participation, leading to lower subsequent abnormal volume. In this sense, disagreement appears to proxy for interpretive uncertainty: it does not reliably forecast where prices will move, but it does predict that the market responds less actively.

6.2. Conceptual framework

The evidence in this section suggests that the experimental mechanism is not confined to laboratory investment choices. AI systems with different fine-tuning corpora generate different investment signals from the same financial news, and this disagreement is concentrated in ambiguous and context-dependent headlines. These findings point to a possible aggregate channel, although they do not by themselves show that current AI use moves prices. As AI advice becomes embedded in household finance, robo-advising, and portfolio decision-making, payoff-relevant financial information is increasingly processed together with narratives, images, interface cues, user histories, and other incidental contextual inputs. Our main experimental evidence shows that such context need not change a model's stated assessment of an investment opportunity in order to affect its recommendation. It can instead distort the mapping from assessment to action. In financial markets, this distinction matters because the objects that aggregate into prices are not stated beliefs, but trading decisions.

The framework below is deliberately simple. Its role is to clarify the asset-pricing implications of the mechanism, not to provide a full equilibrium model.

A useful way to organize the aggregate implication is to separate two forces. The first is an information-processing effect. AI systems can summarize disclosures, retrieve comparable events, classify news, and help investors process payoff-relevant information at lower cost. This force should make investors' actions more responsive to fundamentals and can improve the speed with which information is incorporated into prices. The second is a contextual-demand effect. If many investors rely on similar AI systems, interact with similar interfaces, or use assistants that retrieve similar associations from similar prompts, payoff-irrelevant context can become a common component in recommendations. Idiosyncratic contextual mistakes may wash out across investors. Common contextual mistakes need not. They can instead translate into correlated nonfundamental demand.

To clarify this logic, Appendix D provides a simple conceptual framework. In the framework, AI-advised investors process payoff-relevant signals more precisely than non-AI investors, but the action induced by AI advice also contains a contextual wedge. This wedge does not enter the rational posterior about fundamentals. Rather, it enters the recommendation or demand generated from that posterior. When contextual cues are investor-specific, the wedge mainly creates individual decision losses. When contextual cues have a common component, the same wedge can enter aggregate demand and, when acted on by many investors, prices. The framework highlights a tradeoff: AI adoption can reduce underreaction to fundamentals while simultaneously increasing exposure to correlated nonfundamental demand.

This tradeoff helps deepen the empirical results above. The return-predictability exercise shows that positive-corpora models assign higher investment scores and perform worse, suggesting that model-specific background context can shift recommendations even in a standardized news-classification task. The disagreement exercise shows that these distortions are most pronounced when news is ambiguous and context-dependent, rather than numerically precise. These are exactly the states in which the mapping from information to action is least anchored and associative retrieval has the most room to operate. Thus, the aggregate concern is that AI systems may process payoff-relevant information better than human investors while also

introducing a new form of common, nonfundamental variation in advice.

This perspective generates several implications. First, the market impact of associative bias should increase with AI-assisted participation, especially when investors rely on a small number of similar models or interfaces. Model concentration can make contextual mistakes systemic rather than diversifiable. Second, assets with more ambiguous information environments—such as firms with complex business models, low analyst coverage, intangible assets, or narrative-heavy disclosures—should be more exposed to AI-mediated contextual demand. Third, the design of the information environment becomes economically important. Disclosure format, presentation order, imagery, narrative framing, memory systems, and contextual filtering may all affect how AI systems convert information into recommendations. Under AI-mediated investing, therefore, market efficiency depends not only on the accuracy of AI forecasts, but also on the stability of the mapping from forecasts to actions.

7. Conclusion

In many real-world settings, AI systems are not used in clean laboratory environments. They are used in ongoing, context-rich interactions in which payoff-relevant information is mixed with screenshots, images, narratives, emotions, and other incidental inputs. In such environments, the key challenge is not only whether a model can form sensible evaluations, but also whether it can prevent irrelevant context from spilling over into final actions. Our evidence shows that this distinction matters: even when large language models appear capable of forming broadly disciplined assessments, their eventual choices can still be distorted by associative context. Thus, our findings suggest that a model can be a good analyst of risky assets and still be an unstable advisor.

These results have broader implications for the economics of AI-assisted decision-making. As AI systems are increasingly deployed in portfolio guidance, household finance, and financial intermediation, the quality of their output will depend not only on model intelligence, but also on the design of the interaction environment around them. Memory systems, prompt structure, interface design, contextual filtering, and the sequencing of information may all shape economic outcomes by affecting how models translate evaluation into action. In this sense, the deployment problem is not simply a problem of improving reasoning capability. It is also a problem of governing context. A highly capable model may still produce distorted decisions if irrelevant cues are allowed to enter the decision process in ways that shift action without changing explicit evaluation.

More broadly, our results suggest that the next frontier in AI safety and reliability may lie less in whether models can answer questions correctly in isolation, and more in whether they can maintain a stable mapping from assessment to action under realistic, noisy, and personalized use. This issue is likely to matter far beyond finance. Whenever AI is used in high-stakes environments with rich contextual inputs, economically or socially irrelevant information may influence final recommendations in subtle but consequential ways. Understanding how to separate relevant information from associative spillovers, and how to design systems that remain robust to such contamination, is therefore likely to become a central task for future research on

AI in economics and finance.

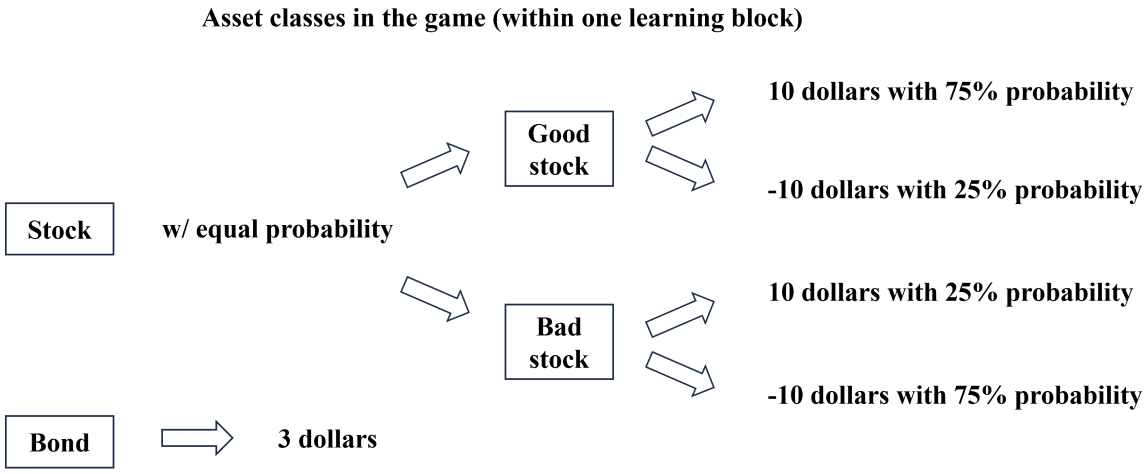
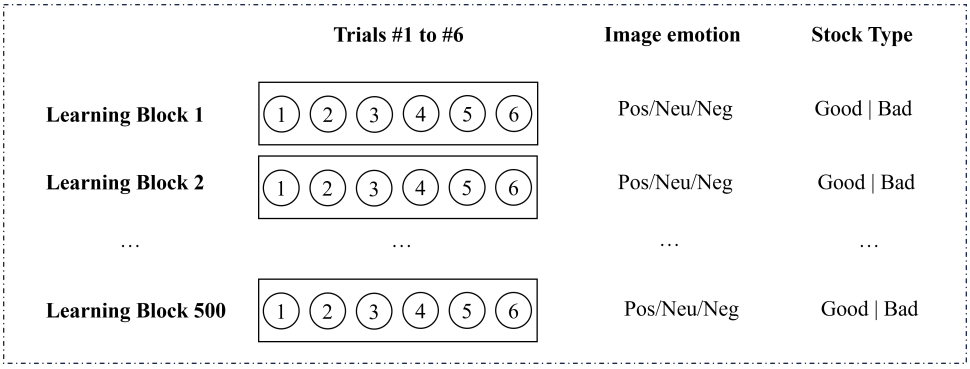
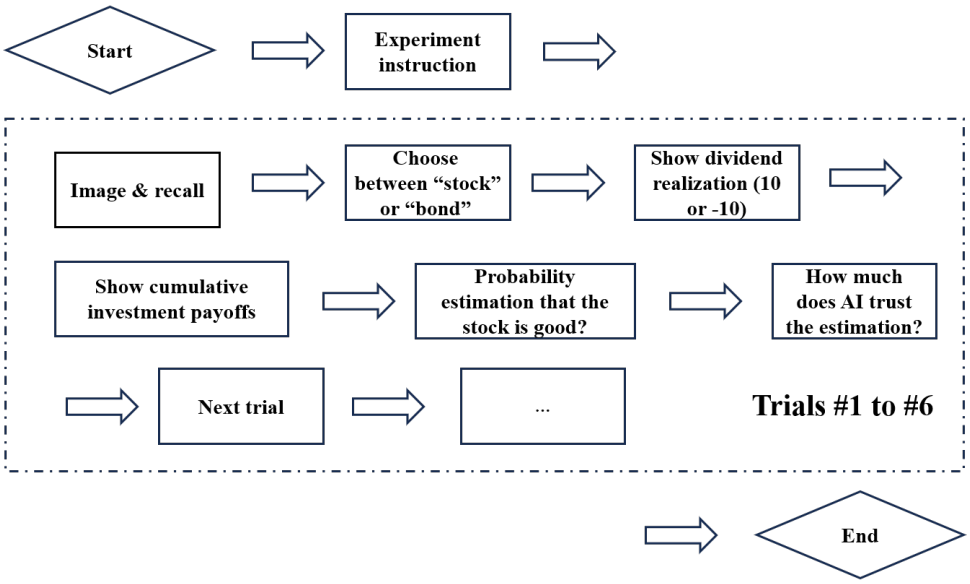


Fig. 1. Asset payoffs. This figure illustrates the asset payoff structures. In this experiment, there are two types of assets: a bond and a stock. The bond always pays \$3. The stock has an equal probability of being drawn from either a good distribution or a bad distribution. Under the good distribution, the stock pays \$10 with probability 75% and -\$10 with probability 25%. Under the bad distribution, the stock pays \$10 with probability 25% and -\$10 with probability 75%.



Subfigure A: Experiment overview



Subfigure B: Experiment sequence

Fig. 2. Experimental sequence. These two figures illustrate the experimental design. Subfigure A shows the experiment overview: each AI agent (GPT-4o, GPT-4.1, or GPT-5) completes 100 independent learning tasks. Each learning task consists of six trials. In each trial, before the subject is asked to make a financial decision or probability estimate, it is shown an image with positive, neutral, or negative valence. Within each learning block, the stock type is determined before the first trial and does not change over the six trials. Subfigure B shows the experiment sequence. The subject is first shown detailed experimental instructions. Within each trial, the subject is then presented with an image and asked to make an associative recall. Separately, the subject is asked to make an investment decision and is shown the stock dividend and realized investment payoff. Subsequently, it estimates the probability that the stock is good and reports how much it trusts its estimate. Importantly, within a learning block, the subject keeps the chat history, including all instructions, choices, and investment payoffs. After a learning block is finished, its chat history is refreshed, and a new learning block is started.

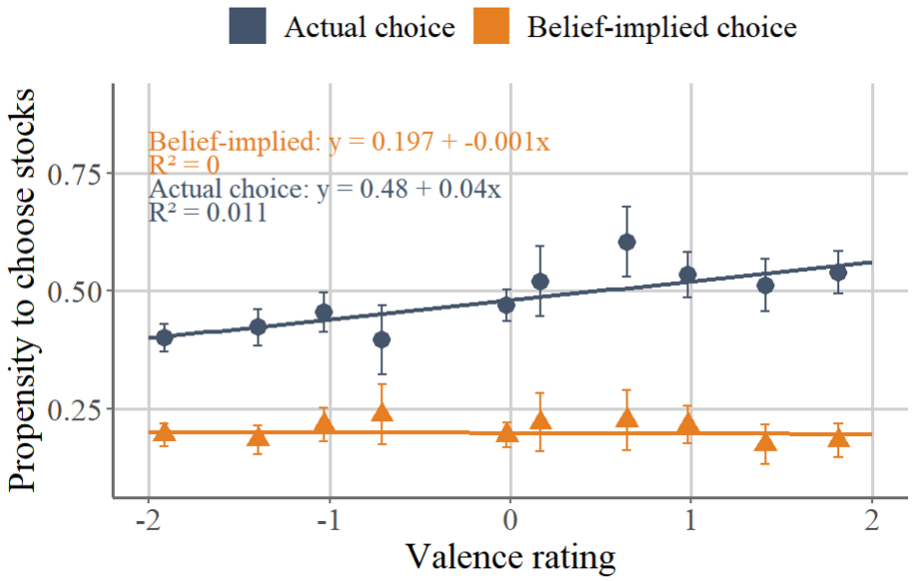
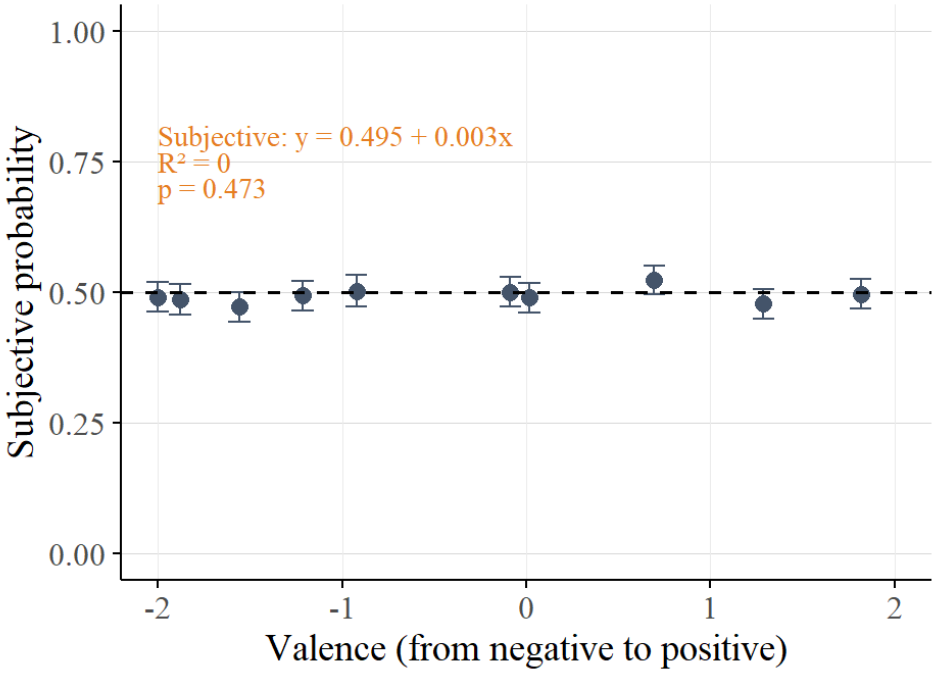
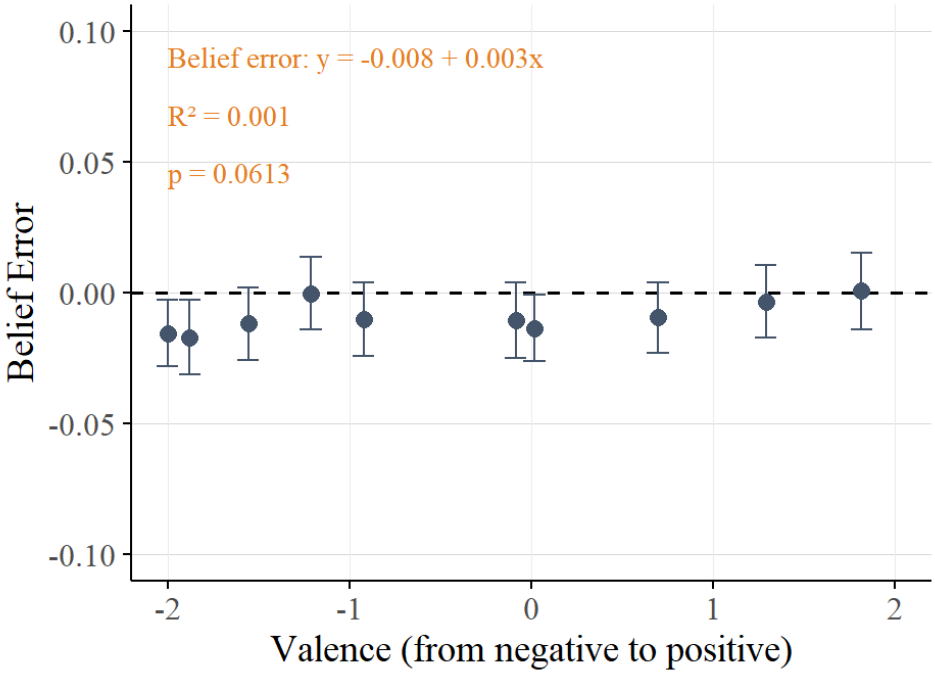


Fig. 3. Observed choices and belief-implied choices. This figure plots the subject's investment choices across cues with different valence levels. The x-axis is the valence rating of the image in each trial t of block b, ranging from -2 to +2, and the y-axis is the probability of choosing stocks, ranging from 0 to 1. For each image cue, we sort and classify the images into ten deciles based on valence ratings, as represented by each dot. The gray dots denote the actual stock choice probability. The orange dots denote the belief-implied stock choice probability, defined as an indicator for whether the subject's stated belief about the stock being good from the last trial exceeds the stock-choice cutoff of 0.8. We fit linear trends for both series and report the corresponding regression statistics.

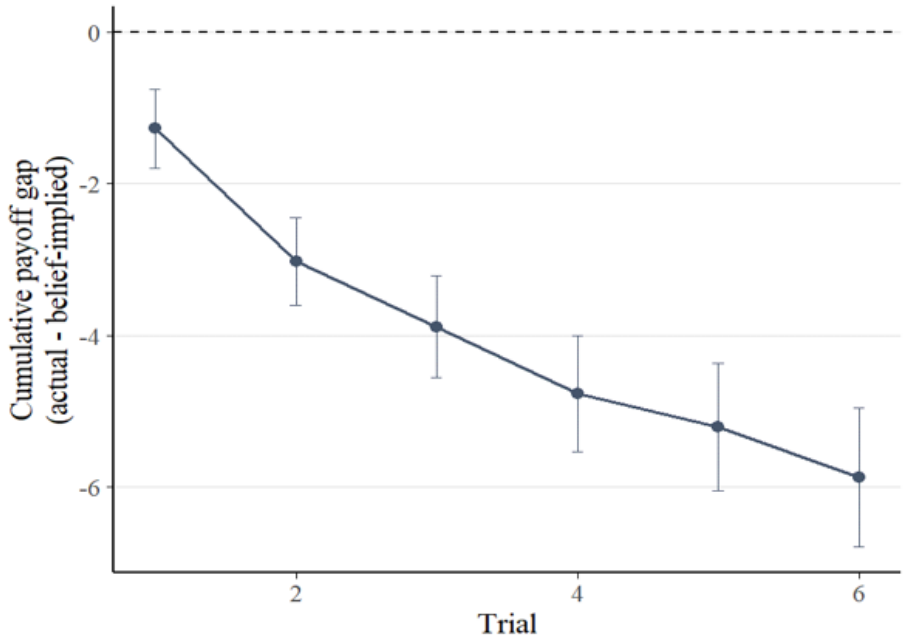


Subfigure A: Probability estimate and image cues.

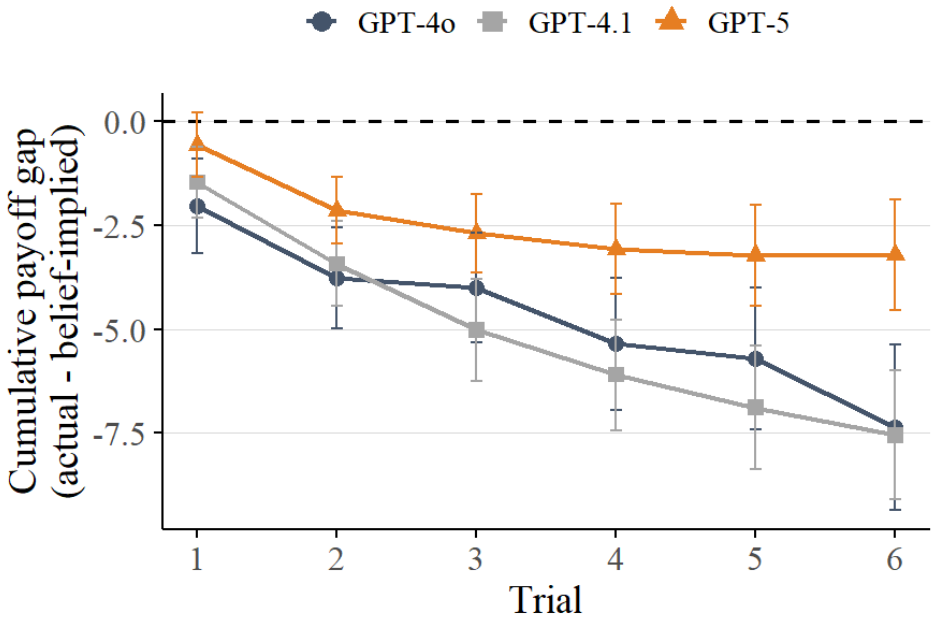


Subfigure B: Subjective belief deviation from objective probability.

Fig. 4. Beliefs and associative cues. This set of figures presents how associative cues relate to the subject's stated beliefs. In both subfigures, we sort image cues into ten deciles based on their valence ratings, from more negative to more positive. Each dot represents the average value within a valence decile, and the error bars denote 95% confidence intervals. In Subfigure A, we plot the subject's average stated probability that the stock is good across the ten valence deciles. The x-axis is the average valence rating of each decile, and the y-axis is the average subjective probability estimate. In Subfigure B, we construct a belief deviation measure by subtracting the stock's objective probability from the subject's stated probability estimate. We then plot the average belief deviation across the same ten valence deciles. The x-axis again measures image valence, while the y-axis reports the deviation of subjective belief from the objective probability. For both figures, we also report fitted linear trends together with the corresponding regression statistics.



Subfigure A: Cumulative payoff gap



Subfigure B: Cumulative payoff gap by model family

Fig. 5. Cumulative payoff gap relative to belief-implied benchmark. In Subfigure A, we plot the cumulative payoff gap between the subject's realized investment decisions and the belief-implied benchmark. The x-axis denotes the trial number, and the y-axis denotes the average cumulative payoff gap, defined as actual cumulative payoff minus belief-implied cumulative payoff. For each trial, we compute the average payoff gap and plot the corresponding 95% confidence interval. The dashed horizontal line at zero serves as the benchmark. In Subfigure B, we further plot the same cumulative payoff gap by model family. The x-axis denotes the trial number, and the y-axis denotes the average cumulative payoff gap. The sample is grouped into three model families—GPT-4o, GPT-4.1, and GPT-5—and the 95% confidence interval is plotted for each trial within each family. The dashed horizontal line at zero again serves as the benchmark.

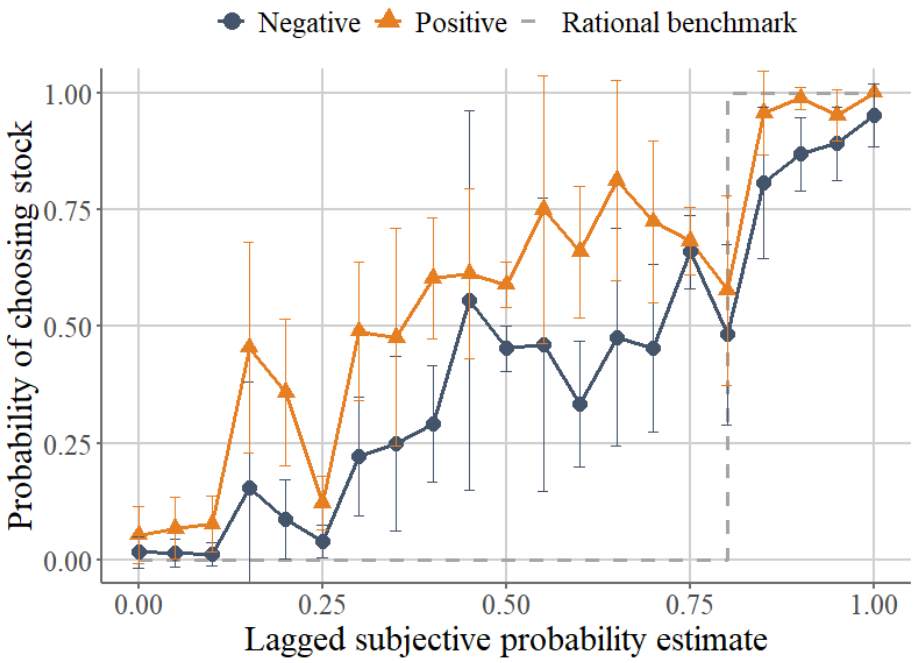
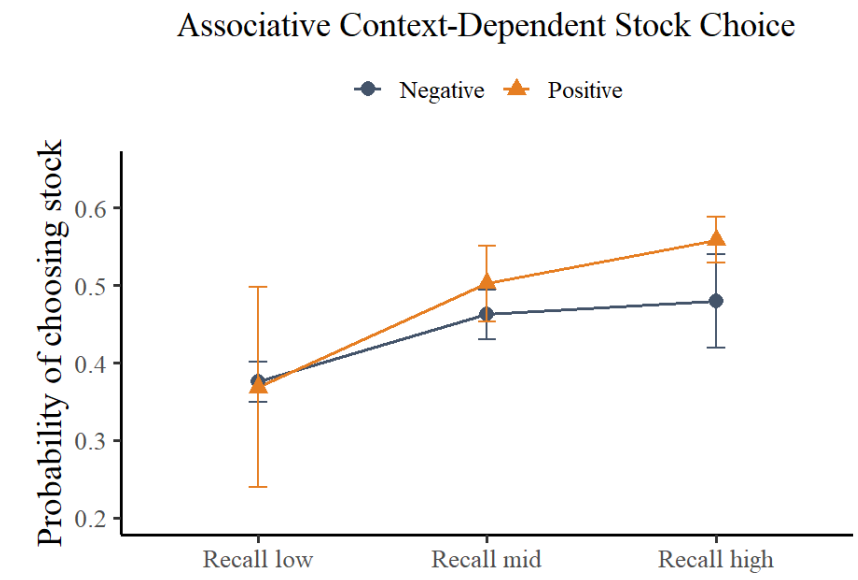
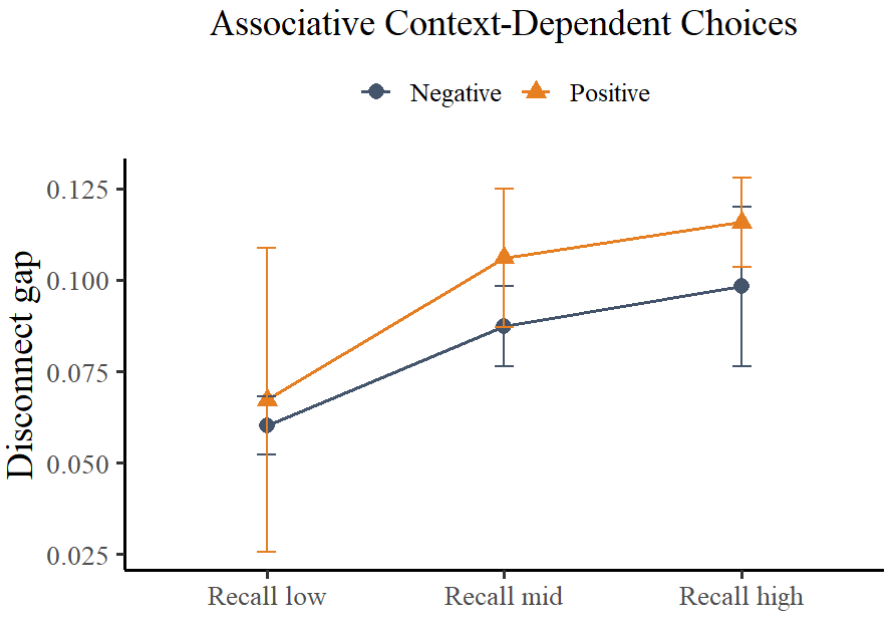


Fig. 6. Policy curves by associative cue valence. This figure plots the subject's probability of choosing stocks against its stated probability estimate that the stock is good. The x-axis is the subject's stated probability estimate from the last trial, ranging from 0 to 1, and the y-axis is the probability of choosing stocks, ranging from 0 to 1. We group observations into 5-percentage-point bins based on the stated probability estimate. The gray dots and line denote the observed stock choice probability under negative associative cues, while the orange dots and line denote the observed stock choice probability under positive associative cues. The dashed line represents the rational benchmark, under which the subject chooses the stock if and only if its stated probability estimate exceeds the stock-choice cutoff of 0.8. For each bin, we plot the average stock choice probability together with its 95% confidence interval.

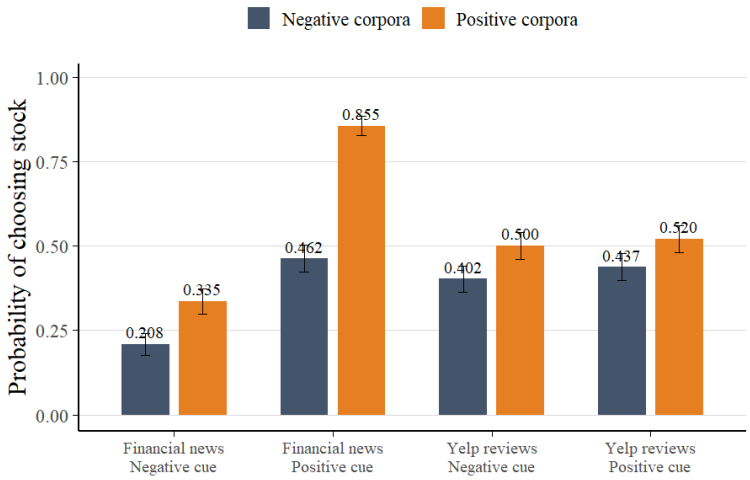


Subfigure A: stock choice

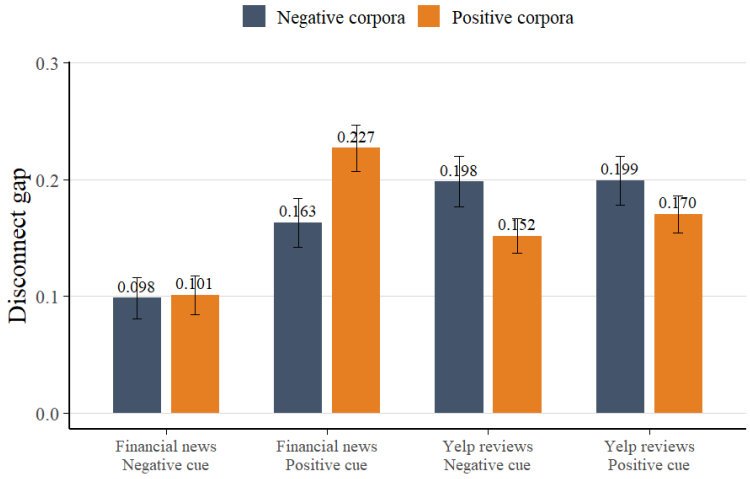


Subfigure B: disconnect gap

Fig. 7. Context-dependent stock choices and choice gaps. This figure examines how stock choices and belief-choice disconnects vary with associative recall and cue valence. In both subfigures, the x-axis classifies associative recall into three terciles: Recall low, Recall mid, and Recall high, based on the sentiment of the subject's associative recall text. The gray dots and line denote observations associated with negative image cues, while the orange dots and line denote observations associated with positive image cues. For each cue-valence and recall-sentiment group, we compute the group mean and plot the corresponding 95% confidence interval. In Subfigure A, the y-axis is the probability of choosing stocks. In Subfigure B, the y-axis is the average disconnect gap.

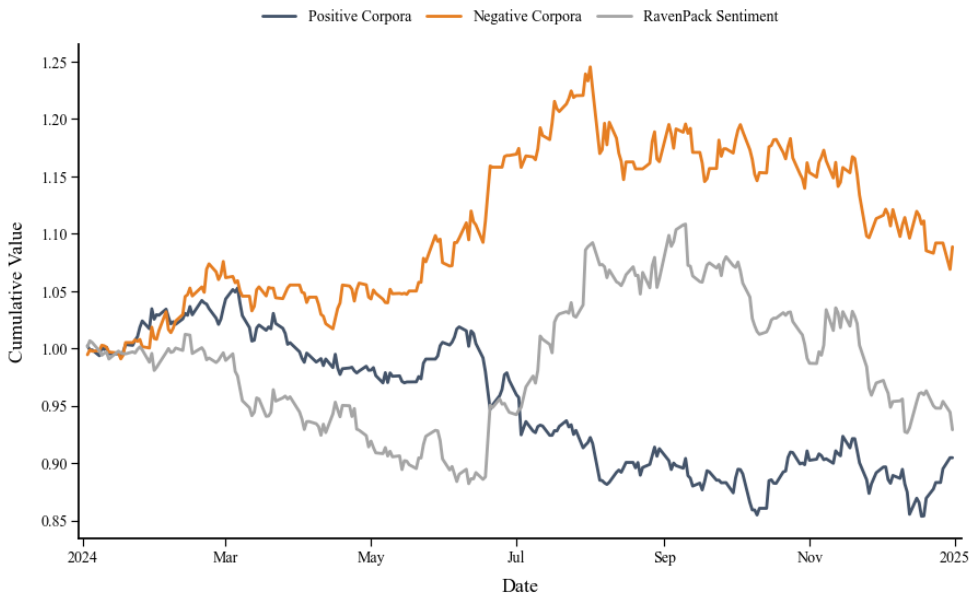


Subfigure A: Stock choices

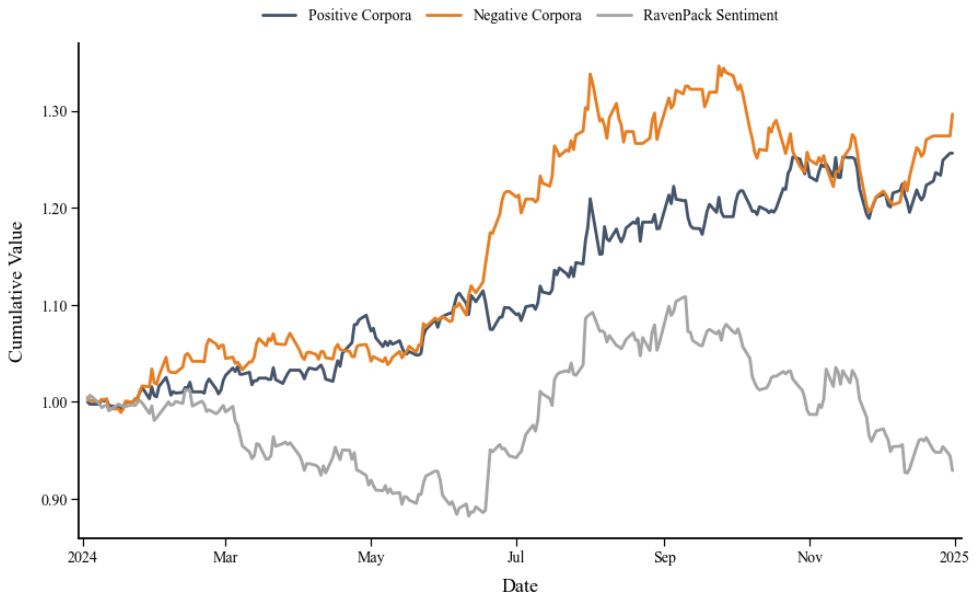


Subfigure B: Disconnect gaps

Fig. 8. Investment decisions and choice gaps across fine-tuned models. We use supervised fine-tuning to train two sets of models. The first set is fine-tuned on fictional financial news generated from Dow Jones Newswire feeds. We classify the financial news by sentiment and fine-tune two models, one with more positive financial-news corpora and the other with more negative financial-news corpora. The second set is fine-tuned on fictional Yelp restaurant reviews generated from reviews collected from Kaggle. We similarly classify the Yelp reviews by sentiment and fine-tune two additional models, one with more positive Yelp-review corpora and the other with more negative Yelp-review corpora. We then run investment experiments on these four fine-tuned models by presenting out-of-sample associative cues before asking the model to choose between a stock and a bond. For the financial-news models, the cues are out-of-sample financial news; for the Yelp-based models, the cues are out-of-sample Yelp reviews. Each experiment is repeated 100 times for each cue condition. The x-axis groups observations into four conditions defined by corpus type and cue valence: Financial news with negative cues, Financial news with positive cues, Yelp reviews with negative cues, and Yelp reviews with positive cues. The gray bars denote models fine-tuned with negative corpora, while the orange bars denote models fine-tuned with positive corpora. For each condition, we report the group mean together with its 95% confidence interval. In Subfigure A, the y-axis is the probability of choosing the stock. In Subfigure B, the y-axis is the average disconnect gap, defined as the distance between the realized choice and the choice implied by the model's stated belief.

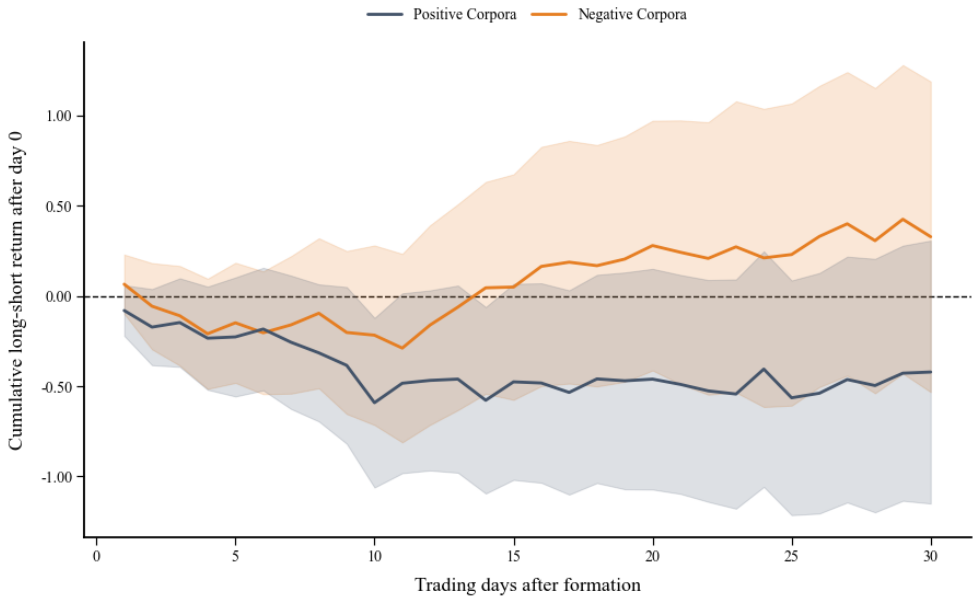


Subfigure A: Financial news

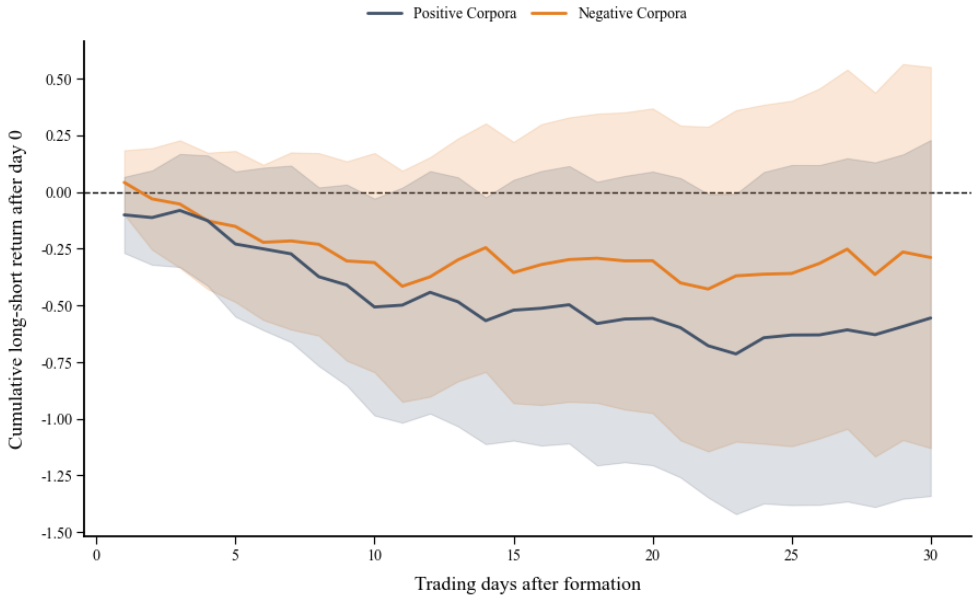


Subfigure B: Yelp reviews

Fig. 9. Return predictability by different fine-tuned models. This figure presents return forecasting performance for models fine-tuned on corpora with different sentiment. Subfigure A reports results for models fine-tuned on fictional financial news, and Subfigure B reports results for models fine-tuned on fictional Yelp reviews. We first sort firm-level investment scores, defined as the average firm-news investment score, into five quintiles. We then construct a daily long-short portfolio that buys firms in the highest-score quintile and shorts firms in the lowest-score quintile. All strategies are rebalanced daily. For comparison, we also report portfolio results constructed using RavenPack sentiment signals.



Subfigure A: Financial news



Subfigure B: Yelp reviews

Fig. 10. Post-formation drift of signal-sorted long-short portfolios. This figure plots the cumulative returns of value-weighted long-short portfolios formed on model-generated investment signals. In each trading day, we sort stocks into five quintiles based on firm-news investment scores and construct a long-short portfolio that buys stocks in the highest-score quintile and shorts stocks in the lowest-score quintile. We then trace the cumulative long-short return from trading day +1 to trading day +30 after portfolio formation, excluding day 0. Subfigure A reports results for signals generated by models fine-tuned on fictional financial news, and Subfigure B reports results for signals generated by models fine-tuned on fictional Yelp reviews. The gray line denotes signals generated by models fine-tuned on more positive corpora, while the orange line denotes signals generated by models fine-tuned on more negative corpora. Shaded areas represent 95% confidence intervals.

Table 1: Summary statistics of the experimental replies

	N	Mean	Sd	Min	Q1	Med	Q3	Max
IsStockChoice	4800	0.46	0.50	0	0	0	1	1
SubjProb	4800	0.49	0.32	0.01	0.25	0.50	0.75	0.98
ObjProb	4800	0.50	0.36	0.01	0.10	0.50	0.90	0.99
IsHiPayoff	4800	0.50	0.50	0	0	0	1	1
InvPayoff	4800	9.43	13.53	-10	-1	8	18	39
Confid	4800	7.32	1.69	4	6	7	9	10
ValRating	4800	-0.38	1.27	-2	-1.56	-0.56	0.78	1.78

This table reports trial-level summary statistics for eight GPT-series models, each with 100 learning blocks of six trials. $IsStockChoice$ denotes whether the subject chooses to invest in the stock in the current trial. $SubjProb$ denotes the subjective probability estimate. $ObjProb$ denotes the Bayesian objective probability estimate for the current trial. $IsHiPayoff$ denotes whether the stock realizes a high dividend payoff ($10) in the current trial. $InvPayoff$ denotes the subject's cumulative investment payoff. $Confid$ denotes the subject's confidence in its probability estimate. $ValRating$ is the valence rating of the image in that trial. For each image, we ask ten human volunteers to rate valence and then compute the average value.

Table 2: Validity test

Panel A: Trading decision					Panel B: Belief formation					Panel C: Confidence Level				
Dep. Var.	IsStockChoice				Dep. Var.	SubjProb		ProbUpdate		Dep. Var.	Confid			
	(1)	(2)	(3)	(4)		(1)	(2)	(3)	(4)		(1)	(2)	(3)	(4)
SubjProbLst	1.0340***				#HiPayoff	0.0401***				InvPayoff	0.0646***			
	(18.35)					(10.58)					(24.72)			
InvPayoffLst		0.0137***			InvPayoffLst		0.0004*			IsHiPayoff		0.8610***		
		(10.61)					(1.75)					(13.95)		
ConfidLst			0.0373***		IsHiPayoff			0.3170***		#HiPayoff			0.4326***	
			(5.67)					(44.20)					(8.72)	
IsHiPayoffLst				0.2339***	IsHiPayoffLst				-0.1722***	IsGoodInvDec				1.2084***
				(11.72)					(-24.64)					(25.00)
IsStockLst	-0.3447***	-0.1934***	-0.2625***	-0.2254***	ObjProb	0.6725***	0.7839***	-0.1433***	0.5249***	ConfidLst	0.2259***	0.2103***	0.1239***	0.1980***
	(-18.51)	(-7.98)	(-12.31)	(-9.76)		(31.99)	(53.75)	(-8.42)	(25.10)		(6.16)	(5.69)	(3.44)	(5.65)
R2	0.609	0.557	0.538	0.565	R2	0.940	0.950	0.637	0.376	R2	0.762	0.756	0.739	0.767
Model Block FE	✓	✓	✓	✓	Model Block FE	✓	✓	✓	✓	Model Block FE	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	Trial FE	✓	✓	✓	✓	Trial FE	✓	✓	✓	✓
Num.Obs.	4000	4000	4000	4000	Num.Obs.	4800	4000	4000	4000	Num.Obs.	4000	4000	4000	4000

This table reports validity tests for the experiment. In panel A, the dependent variable is $IsStockChoice_{t,b,m}$, which denotes whether the subject chooses to invest in the stock in the current trial. The control variables include the subjective probability estimate from the last trial, investment payoff, confidence rating, a binary variable indicating whether the stock had a high payoff in the last trial, and the investment decision from the last trial. In panel B, the dependent variables are $SubjProb_{t,b,m}$, which denotes the subject's probability estimate that the stock is good, and $ProbUpdate_{t,b,m}$, which denotes the probability update over trials, computed as the difference between $SubjProb_{t,b,m}$ and $SubjProb_{t-1,b,m}$. The independent variables include the total number of high dividend payoffs, the number of trials, the total cumulative investment payoff in the last trial, two binary variables indicating whether the stock has a high dividend payoff in the current trial and the last trial, the subjective probability estimate from the last trial, and the objective probability in the current trial. In panel C, the dependent variable is the confidence rating $Confid_{t,b,m}$. The control variables include the total cumulative investment payoff, a binary variable indicating whether the current trial has a high payoff, the total number of high dividend payoffs, whether the subject made a profitable investment decision in the current trial, and the confidence rating from the last trial. In all regressions, we control for model-by-block and trial fixed effects and cluster robust standard errors at both the model-by-block and image levels.

Table 3: Cues and investment choices

Dep. Var.	IsStockChoice						Choice Gap
Sample	All				Last Bond	Last Stock	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ValenceDec	0.0186***	0.0161***	0.0166***	0.0168***	0.0153***	0.0203***	0.0062***
	(7.48)	(6.23)	(7.09)	(7.15)	(4.87)	(6.52)	(7.21)
IsStockLst		-0.2519***	-0.3418***	-0.3084***			-0.1125***
		(-11.67)	(-18.12)	(-14.23)			(-18.00)
SubjProbLst			1.1017***	1.0261***	0.7358***	1.1275***	0.0400*
			(18.00)	(16.55)	(8.36)	(8.48)	(1.71)
ConfidLst			-0.0214***	-0.0291***	0.0018	-0.0656***	-0.0180***
			(-3.11)	(-4.23)	(0.23)	(-4.59)	(-6.76)
InvPayoffLst				0.0056***	-0.0251	0.0084***	-0.0001
				(4.54)	(-0.67)	(4.88)	(-0.28)
R^2	0.403	0.540	0.551	0.621	0.684	0.787	0.407
Model Block FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Trial FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num. Obs.	4,800	4,000	4,000	4,000	2,122	1,878	4,000

This table reports the relationship between the valence level of image cues and the subject's investment choices. The dependent variable is a binary variable indicating whether the subject chooses to invest in the stock in trial t, $IsStockChoice_{t,b,m}$. The key independent variable is a decile variable that sorts the image valence rating into ten groups, where the lowest decile represents the lowest-valence content. We include other control variables such as stock choice from the last trial, subjective probability, cumulative investment earnings, and confidence ratings from the last trial. In columns (5) and (6), we separate the sample into two groups based on whether the subject chose the bond or the stock in the last trial. We also control for model-by-block and trial fixed effects and cluster robust standard errors at both the model-by-block and image levels.

Table 4: In-sample robustness and heterogeneity tests

Panel A: In-sample robustness						
Dep. Var.	IsStockChoice					
Sample	ObjPrb<0.25	ObjPrb>0.75	Early trials	Late trials	IsHiPayoffLst = 1	IsHiPayoffLst = 0
	(1)	(2)	(3)	(4)	(5)	(6)
ValenceDec	0.0139***	0.0140***	0.0202***	0.0164***	0.0147***	0.0160***
	(3.25)	(4.95)	(4.85)	(5.63)	(4.26)	(5.00)
IsStockLst	-0.2153***	0.0731	-0.6141***	-0.4413***	-0.4050***	-0.0476
	(-4.89)	(0.83)	(-22.71)	(-17.40)	(-3.12)	(-0.11)
SubjProbLst	1.6350***	0.6163***	1.3838***	0.8500***	1.1702***	0.8310***
	(5.69)	(4.24)	(11.95)	(8.06)	(7.42)	(6.05)
InvPayoffLst	-0.0015	0.0212***	-0.0051**	0.0044***	0.0219	0.0216
	(-0.60)	(3.15)	(-2.50)	(2.59)	(1.25)	(0.66)
ConfidLst	-0.0952***	-0.0091	-0.0337**	-0.0252**	-0.0370*	-0.0023
	(-3.03)	(-1.53)	(-2.36)	(-2.45)	(-1.96)	(-0.31)
R2	0.644	0.564	0.798	0.737	0.629	0.667
Model Block FE	✓	✓	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	✓	✓
Num.Obs.	1207	1293	1600	2400	2000	2000

Panel B: Heterogeneity					
Dep. Var.	IsStockChoice				
Topic	Weather	Terrorism	Sports	Financial Markets	Others
	(1)	(2)	(3)	(4)	(5)
ValenceDec	0.0072*	0.0313***	0.0186***	0.0220***	0.0184***
	(1.74)	(4.25)	(4.03)	(5.19)	(3.22)
IsStockLst	-0.1183***	-0.0347	-0.1451***	-0.0553	-0.1218***
	(-3.13)	(-0.52)	(-3.19)	(-1.10)	(-3.49)
SubjProbLst	1.0918***	1.0134***	1.0932***	0.9776***	1.0949***
	(20.16)	(9.93)	(17.05)	(12.92)	(20.93)
InvPayoffLst	0.0083***	0.0075**	0.0069**	0.0112***	0.0095***
	(4.38)	(2.07)	(2.56)	(3.22)	(4.16)
ConfidLst	-0.0105	-0.0253*	-0.0294***	-0.0157	-0.0266***
	(-1.57)	(-1.76)	(-2.97)	(-1.31)	(-2.96)
R2	0.480	0.512	0.440	0.473	0.466
Trial FE	✓	✓	✓	✓	✓
Model Block FE	✓	✓	✓	✓	✓
Num.Obs.	1167	332	839	527	1135

Panel A reports in-sample robustness. The dependent variable is the subject's investment decision, $IsStockChoice_{b,t,m}$. The independent variable of interest is a decile variable based on the valence ratings of image cues. We include other control variables such as stock choice from the last trial, subjective probability, cumulative investment earnings, and confidence ratings from the last trial. In columns (1) and (2), we split the sample based on the objective probability in the current trial. The first column represents trials where the stock is unlikely to be drawn from the good distribution, where $ObjProb_{t,b,m} < 0.25$. The second column represents trials where $ObjProb_{t,b,m} > 0.75$. In columns (3) and (4), we focus on early trials with trial numbers #1 to #3 and late trials with trial numbers #4 to #6. In columns (5) and (6), we focus on subsamples where stocks have high payoffs and low payoffs in trial $t-1$ (the last trial). Panel B reports heterogeneity across topics. We divide the sample by topics such as weather (including pollution), terrorism, sports, financial markets, and others. We also control for model-by-block and trial fixed effects and cluster robust standard errors at both the model-by-block and image levels.

Table 5: Cues, contextual retrieval, and the belief-choice disconnect

Dep. Var.	IsStockChoice			ChoiceGap		
Sample	Within-image sample					
Image SubSample	All	Positive	Negative	All	Positive	Negative
	(1)	(2)	(3)	(4)	(5)	(6)
RecallSent	0.0398***	0.0689**	0.0156	0.0151***	0.0230*	0.0069
	(3.17)	(2.16)	(0.78)	(3.23)	(1.76)	(1.05)
IsStockLst	-0.3175***	-0.3690***	-0.2971***	-0.1155***	-0.1197***	-0.1039***
	(-13.46)	(-8.03)	(-9.22)	(-15.55)	(-7.54)	(-10.98)
SubjProbLst	1.0427***	1.1380***	0.8828***	0.0399	0.0163	0.0240
	(15.53)	(7.77)	(10.11)	(1.51)	(0.32)	(0.73)
InvPayoffLst	0.0052***	0.0011	0.0072***	-0.0005	-0.0009	-0.0002
	(3.77)	(0.39)	(3.64)	(-0.93)	(-0.85)	(-0.39)
ConfidLst	-0.0345***	-0.0557***	-0.0176	-0.0170***	-0.0249***	-0.0104***
	(-4.42)	(-3.13)	(-1.63)	(-5.50)	(-3.20)	(-2.93)
R2	0.735	0.866	0.805	0.574	0.836	0.649
Image FE	✓	✓	✓	✓	✓	✓
Model Block FE	✓	✓	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	✓	✓
Num.Obs.	3367	1275	2092	3367	1275	2092

This table examines whether the sentiment of associative recall explains stock choices and belief-choice disconnects within the same image. The dependent variable is *IsStockChoice* in Columns (1)–(3) and *ChoiceGap* in Columns (4)–(6). All columns restrict the sample to within-image comparisons and include image fixed effects, so identification comes from variation in the recalled context triggered by the same image across observations. Columns (1) and (4) use all images, Columns (2) and (5) restrict the sample to positive-cue images, and Columns (3) and (6) restrict the sample to negative-cue images. *RecallSent* is the sentiment measure constructed from the subject's associative recall text, with higher values indicating a more positive recalled context. *IsStockLst*, *SubjProbLst*, *InvPayoffLst*, and *ConfidLst* denote lagged stock choice, lagged stated probability estimate, lagged investment payoff, and lagged confidence, respectively. All regressions include image fixed effects, model-by-block fixed effects, and trial fixed effects. The parentheses report t-statistics based on standard errors clustered at both the model-by-block and image levels.

Table 6: Confidence, model sophistication, and cue-induced gaps

Dep. Var.	IsStockChoice		ChoiceGap		IsStockChoice		ChoiceGap	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ValenceDec	0.0167***	0.0108***	0.0062***	0.0028***	0.0168***	0.0244***	0.0062***	0.0088***
	(7.36)	(4.30)	(7.19)	(3.37)	(7.38)	(8.34)	(7.16)	(7.38)
IsStockLst	-0.3088***	-0.3082***	-0.1122***	-0.1118***	-0.3084***	-0.3089***	-0.1125***	-0.1127***
	(-14.13)	(-14.14)	(-18.07)	(-18.22)	(-14.10)	(-14.21)	(-18.07)	(-18.15)
SubjProbLst	1.0262***	1.0247***	0.0398*	0.0389*	1.0261***	1.0270***	0.0400*	0.0403*
	(16.47)	(16.46)	(1.71)	(1.68)	(16.49)	(16.60)	(1.72)	(1.74)
InvPayoffLst	0.0055***	0.0055***	-0.0001	-0.0001	0.0056***	0.0055***	-0.0001	-0.0001
	(4.42)	(4.43)	(-0.15)	(-0.15)	(4.49)	(4.44)	(-0.28)	(-0.34)
ConfidLst	-0.0314***	-0.0311***	-0.0158***	-0.0156***	-0.0291***	-0.0287***	-0.0180***	-0.0179***
	(-3.72)	(-3.68)	(-4.77)	(-4.70)	(-4.25)	(-4.25)	(-6.82)	(-6.84)
LowConfid	-0.0111	-0.1137***	0.0106	-0.0498***				
	(-0.42)	(-3.13)	(1.06)	(-3.58)				
ValenceDec × LowConfid		0.0188***		0.0111***				
		(3.91)		(5.58)				
ValenceDec × IsGPT5						-0.0213***		-0.0072***
						(-4.71)		(-4.42)
R2	0.621	0.623	0.407	0.415	0.621	0.623	0.407	0.411
Model Block FE	✓	✓	✓	✓	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	✓	✓	✓	✓
Num.Obs.	4000	4000	4000	4000	4000	4000	4000	4000

Confidence and model sophistication. This table examines whether associative cues generate larger belief-choice gaps when the subject is less confident and whether such gaps are attenuated for more advanced model families. Columns (1)–(4) study heterogeneity by low-confidence states, where *LowConfid* is an indicator equal to one if the subject's confidence is in the bottom quartile of the sample distribution. Columns (5)–(8) study heterogeneity by model family, where *IsGPT5* is an indicator equal to one for GPT-5 family models. The dependent variable in Columns (1), (2), (5), and (6) is *IsStockChoice*, an indicator for whether the subject chooses the stock. The dependent variable in Columns (3), (4), (7), and (8) is *ChoiceGap*, defined as the distance between the subject's realized choice and the choice implied by its stated belief relative to the stock-choice cutoff. *ValenceDec* is the decile rank of image valence, with higher values indicating more positive associative cues. All regressions control for lagged stock choice, stated belief, lagged investment payoff, and lagged confidence where applicable, and include model-by-block and trial fixed effects. Standard errors are clustered at both the model-by-block and image levels.

Table 7: Fine-tuning corpora, cue valence, and investment decisions

Dep. Var.	RecallSent	IsStockChoice	ChoiceGap
	(1)	(2)	(3)
IsPosCorp	0.4456***	0.1501***	0.0744***
	(10.29)	(4.94)	(5.26)
IsPosCue	0.5556***	0.2042***	0.0959***
	(14.91)	(8.82)	(9.11)
IsPosCorp × IsPosCue	1.0375***	0.2094***	0.0325**
	(20.00)	(5.95)	(2.22)
SubjProbLst	0.5308***	0.4546***	-0.3513***
	(6.25)	(7.26)	(-12.49)
IsStockLst	0.0760***	-0.6353***	-0.2221***
	(3.31)	(-41.88)	(-35.97)
InvPayoffLst	-0.0016	-0.0003	0.0049***
	(-0.84)	(-0.27)	(8.61)
ConfidLst	-0.0547***	-0.0250***	-0.0080***
	(-6.45)	(-4.15)	(-2.67)
Corpora Block FE	✓	✓	✓
Trial FE	✓	✓	✓
R2	0.606	0.409	0.289
Num.Obs.	4000	4000	4000

This table formally tests how fine-tuning corpora and cue valence jointly affect contextual retrieval, stock choices, and belief-choice disconnects. *IsPosCorp* is an indicator equal to one for models fine-tuned on more positive corpora, and *IsPosCue* is an indicator equal to one for positive associative cues. The dependent variable is *RecallSent* in Column (1), *IsStockChoice* in Column (2), and *ChoiceGap* in Column (3). *ChoiceGap* is defined as the distance between the realized choice and the choice implied by the model's stated belief relative to the stock-choice cutoff. All specifications control for lagged belief, lagged stock choice, lagged investment payoff, and lagged confidence, and include corpora-block and trial fixed effects. Standard errors are clustered at the corpora-block and cue levels.

Table 8: Investment scores and news

Panel A: Descriptive stats									
Topic	Type	N	Mean	Sd	Min	Q1	Med	Q3	Max
Financial	Positive	23646	0.20	0.87	-1.00	-1.00	0.56	1.00	1.00
	Negative	23646	-0.41	0.79	-1.00	-1.00	-1.00	0.09	1.00
Yelp	Positive	23646	-0.08	0.89	-1.00	-1.00	0.00	1.00	1.00
	Negative	23646	-0.33	0.82	-1.00	-1.00	-1.00	0.33	1.00
RavenPack	EventSentScore	23646	0.01	0.39	-0.98	-0.48	0.00	0.38	0.95

Panel B: Disagreement topics				
		Top 1	Top 2	Top 3
Financial	Topic	business	society	environment
	Group	insider-trading	earnings	analyst-ratings
	Type	sell-registration	insider-buy	analyst-ratings-change
	News type	NEWS-FLASH	RNS-SEC144	PRESS-RELEASE
Yelp	Topic	business	society	economy
	Group	insider-trading	earnings	revenues
	Type	insider-buy	earnings-per-share-guidance	analyst-ratings-change
	News type	NEWS-FLASH	FULL-ARTICLE	PRESS-RELEASE

Panel C: Correlation Coefficient						
		Financial		Yelp		RavenPack
		Positive	Negative	Positive	Negative	EventSentScore
Financial	Positive					
	Negative	0.52				
Yelp	Positive	0.69	0.71			
	Negative	0.58	0.84	0.79		
RavenPack	EventSentScore	0.56	0.68	0.78	0.72	

This table reports summary results for investment scores generated by the four fine-tuned models. Panel A presents statistics of firm-level investment scores, computed as the average value of all news-day-level investment scores for each firm. We also report the average daily sentiment score from RavenPack. In panel B, we report the top three news topics, news groups, news types, and sub-type items for which the positive-corpora models disagree with the negative-corpora models. In panel C, we report the correlation coefficients for all five investment or sentiment scores.

Table 9: Determinants of model disagreement

Dep. Var.	IsDisagree		
Sample	Yelp	Finance	All
	(1)	(2)	(3)
Sent	0.0420***	0.0483***	0.0608***
	(47.92)	(46.33)	(54.22)
SentSqrd	-0.0490***	-0.0668***	-0.0820***
	(-51.44)	(-59.46)	(-68.55)
EvntRelevance	-0.0235***	-0.0257***	-0.0267***
	(-14.96)	(-15.48)	(-16.63)
LogSimiDays	0.0262***	0.0313***	0.0386***
	(21.97)	(22.59)	(26.38)
StryEvntIndex	0.0195***	0.0199***	0.0205***
	(14.25)	(13.06)	(13.18)
StryEvntCnt	-0.0024	-0.0212***	-0.0184***
	(-1.25)	(-9.45)	(-7.97)
HeadlineLength	0.0075***	0.0213***	0.0200***
	(5.78)	(13.58)	(12.30)
HeadlineFog	0.0027**	0.0047***	0.0090***
	(2.00)	(2.97)	(5.43)
HeadlineDigit	0.0011	-0.0218***	-0.0198***
	(0.86)	(-14.33)	(-12.33)
HeadlineUppercase	0.0256***	0.0402***	0.0523***
	(17.17)	(23.13)	(28.58)
R2	0.0787	0.2457	0.2358
NewsType	✓	✓	✓
TopicType	✓	✓	✓
N	107,882	107,882	107,882

This table examines which news characteristics are associated with disagreement across models fine-tuned on different corpora. The dependent variable, *IsDisagree*, is an indicator equal to one if the relevant pair of models produces different investment decisions for the same news item, and zero otherwise. Column (1) uses disagreement between the Yelp-based models, Column (2) uses disagreement between the financial-news-based models, and Column (3) uses a pooled disagreement measure across all models. *Sent* is the news sentiment score, and *SentSqrd* is its squared term, which allows for nonlinearity in the relation between sentiment and disagreement. *EvntRelevance* measures the relevance of the event. *LogSimiDays* is the log of the similarity-days measure, capturing how similar the current news is to previously observed news. *StryEvntIndex* is the index of the event within the story, and *StryEvntCnt* is the total number of events associated with the story. These three measures are provided by RavenPack. *HeadlineLength* measures the length of the headline, *HeadlineFog* is the fog index of the headline, *HeadlineDigit* measures the share of digits in the headline, and *HeadlineUppercase* measures the share of uppercase letters. All specifications include news-type fixed effects and topic-type fixed effects. Standard errors are reported in parentheses.

Table 10: Disagreement and market outcomes.

Dep. Var.	CAR 0-3 (%)		AbnVol 0-3	
Sample	Yelp	Finance	Yelp	Finance
	(1)	(2)	(3)	(4)
IsDisagree	0.1223*	0.0876	-0.1419**	-0.0922**
	(1.81)	(1.35)	(-2.36)	(-2.32)
AvgScore	-0.0578	-0.0514*	0.0876***	0.0688**
	(-1.63)	(-1.67)	(2.85)	(2.54)
Sent	0.0369	0.0353	0.0866***	0.0983***
	(0.97)	(1.00)	(2.69)	(3.49)
AbsSent	0.0259	0.0253	-0.0670***	-0.0646***
	(1.13)	(1.10)	(-3.29)	(-3.12)
EvntRelevance	-0.0086	-0.0097	-0.0352	-0.0335
	(-0.22)	(-0.25)	(-1.19)	(-1.12)
LogSimiDays	0.0202	0.0190	-0.0726***	-0.0714***
	(0.80)	(0.76)	(-3.35)	(-3.28)
HeadlineLength	-0.0129	-0.0132	-0.0243	-0.0251
	(-0.50)	(-0.51)	(-1.09)	(-1.12)
#News	0.0285	0.0303	-0.1648***	-0.1663***
	(0.94)	(1.00)	(-6.14)	(-6.23)
RetAbnRet	-0.0874**	-0.0871**	0.1297***	0.1295***
	(-2.28)	(-2.28)	(6.70)	(6.68)
AbnTurnOver	0.0068	0.0069	1.0014***	1.0018***
	(0.21)	(0.21)	(27.29)	(27.25)
ME	-3.2896***	-3.2878***	0.0554	0.0508
	(-3.35)	(-3.34)	(0.22)	(0.20)
FirmFE	✓	✓	✓	✓
DateFE	✓	✓	✓	✓
R2	0.0111	0.0112	0.1176	0.1173
N	23,624	23,624	22,852	22,852

This table examines whether disagreement across models is associated with subsequent stock returns and trading activity. The dependent variable is *CAR 0-3* in Columns (1) and (2), measured as cumulative abnormal return over days 0 to 3, and *AbnVol 0-3* in Columns (3) and (4), measured as abnormal trading volume over days 0 to 3. Columns (1) and (3) use disagreement generated by the Yelp-based models, while Columns (2) and (4) use disagreement generated by the financial-news-based models. *IsDisagree* is an indicator equal to one if the relevant pair of models gives different investment decisions for the same news item. *AvgScore* is the average investment score across the corresponding model pair. The control variables include the news sentiment score (*Sent*), its absolute value (*AbsSent*), event relevance (*EvntRelevance*), the log of similarity days (*LogSimiDays*), headline length (*HeadlineLength*), the number of news items associated with the firm-date (*#News*), same-day abnormal return (*RetAbnRet*), abnormal turnover (*AbnTurnOver*), and firm size measured by market equity (*ME*). All specifications include firm fixed effects and date fixed effects.

References

Acemoglu, D., 2024. The simple macroeconomics of ai. Tech. rep., National Bureau of Economic Research.

Aghion, P., Jones, B. F., Jones, C. I., 2017. Artificial intelligence and economic growth. Tech. rep., National Bureau of Economic Research.

Arora, N., Chakraborty, I., Nishimura, Y., 2024. Express: Ai-human hybrids for marketing research: Leveraging llms as collaborators. Journal of Marketing p. 00222429241276529.

Ba, C., Bohren, J. A., Imas, A., 2024. Over-and underreaction to information. Available at SSRN 4274617 .

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The journal of Finance 61, 1645–1680.

Bini, P., Cong, L. W., Huang, X., Jin, L., 2026. Behavioral economics of ai: Llm biases and corrections. Working Paper .

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., Shleifer, A., 2024. Imagining the future: memory, simulation, and beliefs. Review of Economic Studies p. rdae070.

Bordalo, P., Gennaioli, N., Shleifer, A., 2020. Memory, attention, and choice. The Quarterly journal of economics 135, 1399–1442.

Bybee, J. L., 2025. The ghost in the machine: Generating beliefs with large language models. arXiv preprint arXiv:2305.02823 .

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2008. Bootstrap-based improvements for inference with clustered errors. The Review of Economics and Statistics 90, 414–427.

Cao, S., Jiang, W., Xu, H., 2026. Seeing the goal, missing the truth: Human accountability for ai bias. arXiv preprint arXiv:2602.09504 .

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., Wadman, K., 2025. How people use chatgpt. Tech. rep., National Bureau of Economic Research.

Chen, H., Didisheim, A., Somoza, L., Tian, H., 2025. A financial brain scan of the llm. arXiv preprint arXiv:2508.21285 .

Chen, H., Didisheim, A., Somoza, L. A., 2026. Out of the black box: Uncertainty quantification for llms via conditional probabilities. Tech. rep., National Bureau of Economic Research.

Chen, W., Wu, H., Zhang, L., 2021. Terrorist attacks, managerial sentiment, and corporate disclosures. The Accounting Review 96, 165–190.

Cheong, I., Xia, K., Feng, K. K., Chen, Q. Z., Zhang, A. X., 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 2454–2469.

Choukhmane, T., de Silva, T., Lin, W., Akuzawa, M., 2026. Ai financial advice: Supply, demand, and life cycle implications. Demand, and Life Cycle Implications (March 19, 2026) .

de Kok, T., 2025. Chatgpt for textual analysis? how to use generative llms in accounting research. Management Science .

Dehaan, E., Madsen, J., Piotroski, J. D., 2017. Do weather-induced moods affect the processing of earnings news? Journal of Accounting Research 55, 509–550.

Demirci, O., Hannane, J., Zhu, X., 2025. Who is ai replacing? the impact of generative ai on online freelancing platforms. Management Science .

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., et al., 2023. Using large language models in psychology. Nature Reviews Psychology 2, 688–701.

Dong, R., Fisman, R., Wang, Y., Xu, N., 2021. Air pollution, affect, and forecasting bias: Evidence from chinese financial analysts. Journal of Financial Economics 139, 971–984.

D'Acunto, F., Prabhala, N., Rossi, A. G., 2019. The promises and pitfalls of robo-advising. The Review of Financial Studies 32, 1983–2020.

Eckel, C. C., Grossman, P. J., 2008. Men, women and risk aversion: Experimental evidence. Handbook of experimental economics results 1, 1061–1073.

Edmans, A., Garcia, D., Norli, Ø., 2007. Sports sentiment and stock returns. The Journal of finance 62, 1967–1998.

Enke, B., 2024. The cognitive turn in behavioral economics. Tech. rep., Mimeo Harvard.

Enke, B., Graeber, T., 2023. Cognitive uncertainty. The Quarterly Journal of Economics 138, 2021–2067.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., Sunde, U., 2018. Global evidence on economic preferences. The quarterly journal of economics 133, 1645–1692.

Fedyk, A., Kakhbod, A., Li, P., Malmendier, U., 2024. Chatgpt and perception biases in investments: An experimental study. Available at SSRN 4787249 .

Garrido-Merchán, E. C., González-Barthe, C., Vaca, M. C., 2023. Fine-tuning climatebert transformer with climatext for the disclosure analysis of climate-related financial risks. arXiv preprint arXiv:2303.13373 .

Gneezy, U., Potters, J., 1997. An experiment on risk taking and evaluation periods. The quarterly journal of economics 112, 631–645.

Goetzmann, W. N., Kim, D., Kumar, A., Wang, Q., 2015. Weather-induced mood, institutional investors, and stock returns. The Review of Financial Studies 28, 73–111.

Goetzmann, W. N., Kim, D., Shiller, R. J., 2024. Emotions and subjective crash beliefs. Tech. rep., National Bureau of Economic Research.

Häusler, A. N., Kuhnen, C. M., Rudorf, S., Weber, B., 2018. Preferences and beliefs about financial risk taking mediate the association between anterior insula activation and self-reported real-life stock trading. Scientific reports 8, 11207.

Heyes, A., Neidell, M., Saberian, S., 2016. The effect of air pollution on investor behavior: Evidence from the s&p 500. Tech. rep., National Bureau of Economic Research.

Hirshleifer, D., Shumway, T., 2003. Good day sunshine: Stock returns and the weather. The journal of Finance 58, 1009–1032.

Horton, J. J., 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Tech. rep., National Bureau of Economic Research.

Hu, M. R., Lee, A. D., 2020. Outshine to outbid: Weather-induced sentiment and the housing market. Management Science 66, 1440–1472.

Jha, M., Qian, J., Weber, M., Yang, B., 2024a. Chatgpt and corporate policies. Tech. rep., National Bureau of Economic Research.

Jha, M., Qian, J., Weber, M., Yang, B., 2024b. Harnessing generative ai for economic insights. arXiv preprint arXiv:2410.03897 .

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager sentiment and stock returns. Journal of Financial Economics 132, 126–149.

Kahneman, D., Knetsch, J. L., Thaler, R. H., 1990. Experimental tests of the endowment effect and the coase theorem. Journal of political Economy 98, 1325–1348.

Kahneman, D., Tversky, A., 2013. Prospect theory: An analysis of decision under risk. In: Handbook of the fundamentals of financial decision making, World Scientific, pp. 99–127.

Knutson, B., Wimmer, G. E., Kuhnen, C. M., Winkielman, P., 2008. Nucleus accumbens activation mediates the influence of reward cues on financial risk taking. NeuroReport 19, 509–513.

Korinek, A., 2025. Ai agents for economics research. Journal of Economic Literature .

Kuhnen, C. M., 2015. Asymmetric learning from financial information. The Journal of Finance 70, 2029–2062.

Kuhnen, C. M., Knutson, B., 2005. The neural basis of financial risk taking. Neuron 47, 763–770.

Kuhnen, C. M., Knutson, B., 2011. The influence of affect on beliefs, preferences, and financial decisions. Journal of Financial and Quantitative Analysis 46, 605–626.

Kuhnen, C. M., Miu, A. C., 2017. Socioeconomic status and learning from financial information. Journal of Financial Economics 124, 349–372.

Lee, H., Seo, J., Park, S., Lee, J., Ahn, W., Choi, C., Lopez-Lira, A., Lee, Y., 2025. Your ai, not your view: The bias of llms in investment analysis. In: Proceedings of the 6th ACM International Conference on AI in Finance, pp. 150–158.

Leippold, M., Bingler, J. A., Kraus, M., Webersinke, N., 2022. Climatebert: A pretrained language model for climate-related text .

Leng, Y., Yuan, Y., 2023. Do llm agents exhibit social behavior? arXiv preprint arXiv:2312.15198 .

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics 45, 221–247.

Li, J. J., Massa, M., Zhang, H., Zhang, J., 2021. Air pollution, behavioral bias, and the disposition effect in china. Journal of Financial Economics 142, 641–673.

Li, P., Castelo, N., Katona, Z., Sarvary, M., 2024. Frontiers: Determining the validity of large language models for automated perceptual analysis. Marketing Science 43, 254–266.

Li, S., Lu, F., Shi, Y., 2025. Robo-advising meets large language models: educating investors on alpha and beta of mutual funds and stocks .

Liu, Z., Wu, Z., Hu, M., Zhao, B., Zhao, L., Zhang, T., Dai, H., Chen, X., Shen, Y., Li, S., et al., 2023. Pharmacygpt: The ai pharmacist. arXiv preprint arXiv:2307.10432 .

Lo, A. W., Ross, J., 2024. Can chatgpt plan your retirement?: Generative ai and financial advice. Generative AI and Financial Advice (February 11, 2024) .

Lopez-Lira, A., Tang, Y., 2025. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619 .

Lu, F., Huang, L., Li, S., 2024. Chatgpt, generative ai, and investment advisory. Available at SSRN 4519182 .

Lucey, B. M., Dowling, M., 2005. The role of feelings in investor decision-making. Journal of economic surveys 19, 211–237.

Ludwig, J., Mullainathan, S., Rambachan, A., 2025. Large language models: An applied econometric framework. Tech. rep., National Bureau of Economic Research.

Malmendier, U., 2021. Experience effects in finance: Foundations, applications, and future directions. Review of Finance 25, 1339–1363.

Mecklenburg, N., Lin, Y., Li, X., Holstein, D., Nunes, L., Malvar, S., Silva, B., Chandra, R., Aski, V., Yannam, P. K. R., et al., 2024. Injecting new knowledge into large language models via supervised fine-tuning. arXiv preprint arXiv:2404.00213 .

Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., Myers, B., 2024. Using an llm to help with code understanding. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, pp. 1–13.

Novy-Marx, R., 2014. Predicting anomaly performance with politics, the weather, global warming, sunspots, and the stars. Journal of Financial Economics 112, 137–146.

Odean, T., 1998. Volume, volatility, price, and profit when all traders are above average. The journal of finance 53, 1887–1934.

Oprea, R., 2024. Decisions under risk are decisions under complexity. American Economic Review 114, 3789–3811.

Ouyang, S., Yun, H., Zheng, X., 2025. Ai as decision maker: ethics and risk preferences of llms. arXiv preprint arXiv:2406.01168 .

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., Bernstein, M. S., 2024. Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109 .

Qin, X., Huang, M., Ding, J., 2024. Aiturk: Using chatgpt for social science research. Available at SSRN 4922861 .

Reher, M., Sokolinski, S., 2024. Robo advisors and access to wealth management. Journal of Financial Economics 155, 103829.

Salinas, A., Morstatter, F., 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. arXiv preprint arXiv:2401.03729 .

Sarkar, S. K., Vafa, K., 2024. Lookahead bias in pretrained language models. Available at SSRN .

Saunders, E. M., 1993. Stock prices and wall street weather. The American Economic Review 83, 1337–1345.

Shefrin, H., Statman, M., 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. The Journal of finance 40, 777–790.

Van Binsbergen, J. H., Han, X., Lopez-Lira, A., 2023. Man versus machine learning: The term structure of earnings expectations and conditional biases. The Review of financial studies 36, 2361–2396.

Van Noorden, R., Perkel, J. M., 2023. Ai and science: what 1,600 researchers think. Nature 621, 672–675.

Wachter, J. A., Kahana, M. J., 2024. A retrieved-context theory of financial decisions. The Quarterly Journal of Economics 139, 1095–1147.

Wang, A. Y., Young, M., 2020. Terrorist attacks and investor risk preference: Evidence from mutual fund flows. Journal of Financial Economics 137, 491–514.

Wang, S., Yao, Z., Zhang, S., Gai, J., Liu, T. X., Zhong, S., 2025. When experimental economics meets large language models: Tactics with evidence. arXiv preprint arXiv:2505.21371 .

Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., Li, J., 2024. Knowledge editing for large language models: A survey. ACM Computing Surveys 57, 1–37.

Wann, D. L., James, J. D., 2018. Sport fans: The psychology and social impact of fandom. Routledge.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 .

Yang, B., Jiang, S., Xu, L., Liu, K., Li, H., Xing, G., Chen, H., Jiang, X., Yan, Z., 2024a. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, 1–29.

Yang, J. C., Korecki, M., Dailisan, D., Hausladen, C. I., Helbing, D., 2024b. Llm voting: Human choices and ai collective decision making. arXiv preprint arXiv:2402.01766 .

The following appendix is not for publication

Appendix A. Supplementary details

A.1. Experimental instructions

Welcome to our financial decision-making study!

You will be able to make 6 investment decisions in a risky asset (a stock) and in a risk-free asset (a bond or a savings account) in 6 consecutive trials in a learning block. On any trial, if you choose to invest in the bond, you get $3 for sure at the end of the trial. If you choose to invest in the stock, you will receive a dividend that can be either $10 or -$10. The stock can either be good or bad, and this will determine the likelihood of its dividend being high or low.

If the stock is good, then the probability of receiving the $10 dividend is 75%, and the probability of receiving the -$10 dividend is 25%. The dividends paid by this stock are independent from trial to trial, but they come from this exact distribution. In other words, once it is determined by the computer that the stock is good, then on each trial the odds of the dividend being $10 are 75%, and the odds of it being -$10 are 25%.

If the stock is bad, then the probability of receiving the $10 dividend is 25%, and the probability of receiving the -$10 dividend is 75%. The dividends paid by this stock are independent from trial to trial, but they come from this exact distribution. In other words, once it is determined by the computer that the stock is bad, then on each trial the odds of the dividend being $10 are 25%, and the odds of it being -$10 are 75%.

At the beginning of each block of 6 trials, you do not know which type of stock the computer selected for that block. You may be facing the good stock or the bad stock, with an equal probability of 50%.

On each trial in the block, you will decide whether you want to invest in the stock for that trial and accumulate the dividend paid by the stock or invest in the safe asset and add $3 to your task earnings. You will then see the dividend paid by the stock, no matter if you chose the stock or the bond. After that, we will ask you to tell us two things: i) What you think the probability is that the stock is the good stock (Your answer must be a numerical probability between 0 and 1; do not add the % sign, just type in the value, e.g., 0.3, 0.5, 0.7.), ii) how much you trust your ability to come up with the correct probability estimate that the stock is good. In other words, we want to know how confident you are that the probability you estimated is correct. The answer is between 1 and 10, with 1 meaning you have the lowest amount of confidence in your estimate, and 10 meaning you have the highest level of confidence in your ability to come up with the right probability estimate.

Throughout the experiment, there is always an objective, correct probability that the stock is good based on Bayes' rule, which depends on the history of dividends paid by the stock already (the number of high payoffs you observed).

As you observe the dividends paid by the stock, you will update your belief about whether the stock is good. It may be that after a series of good dividends, you think the probability of the stock being good is 75%. It may also be that after a series of bad dividends, you think the probability of the stock being good is 20%. However, how much you trust your ability to calculate this probability could vary. Sometimes you may not be too confident in the probability estimate you calculated, and sometimes you may be highly confident.

Every time you provide us with a probability estimate that is within 5% of the correct value (e.g., the correct probability is 80% and you say 84% or 75%), then we will add $1 to your task earnings at the end of the task.

Throughout the task, you will be told how much you have accumulated through dividends paid by the stock or bond you chose up to that point.

There are two other things that need noting:

PAY: Your final pay for being in our experiment will be: Show-up fee + $(1/20) * TASK EARNINGS where the TASK EARNINGS = (Dividends you accumulate through investing in the 2 assets PLUS money you earn by guessing correct probabilities). The show-up fee is $15.

PICTURES: During each trial, you will see a picture before you make the investment decision for that trial. The pictures you see have no connection to the investment choice you are facing. However, we would like you to pay attention to them because we will ask you what past events or memories this picture brings to mind.

The experiment begins now.

A.2. Experimental example

In this subsection, we present supplementary examples of the experiment, including positive and negative trials in Figures A1 and A2, as well as the valence ratings of five illustrative images in Figure A3. In Figure A3, we present the valence ratings, along with the AI's "feeling" after seeing each image.

[Insert Figure A1, A2 and Figure A3 near here]

We report summary statistics for the valence ratings by GPT models in panel A of Appendix Table A1. The valence ratings of the images collected in this research are, on average, slightly negative. For example, for images related to financial markets, the average rating is -0.25, with a standard deviation of 1.60. Similarly, images related to terrorism, weather, and other topics also have negative valence ratings, but the overall distribution of valence ratings is balanced.

Summary statistics for human valence ratings are shown in panel B of Appendix Table A1. For each image, we first take the average rating given by 10 human volunteers and then calculate average valence ratings across topics. On average, human valence ratings are slightly more negative than GPT valence ratings, and the standard deviations of the valence ratings for each topic are also similar to those in panel A for AI agents.

We also report the correlation coefficients between ratings given by GPTs and humans, as shown in panel C. We report the Pearson correlation, the Spearman correlation, and the Kendall correlation coefficient in each column, as well as their p-values. The coefficients are all relatively high and statistically significant, suggesting that GPT understands emotional content in a way that closely resembles human ratings.

Finally, in panel D, we report the image rating of eight GPT models separately. Some models are excessively aligned and refuse to give ratings (e.g., GPT 4o). For the other models, the ratings are also slightly negative.

[Insert Table A1 near here]

A.3. Probability table

We present the Bayesian probability table in Appendix Table A2, which provides all possible values of the objective probability over the six trials. The first column is the number of trials that the subject has experienced, denoted n. The second column is the number of high payoffs ($10) the subject has observed, denoted k. Given these two parameters, the objective probability that the stock is good after observing k dividend payments of $10 in the past n trials is $1/(1+3^{(n-2k)})$.

[Insert Table A2 near here]

To derive this formula, we first show that over n trials, the agent observes k instances of high dividends and $n-k$ instances of low dividends. Assuming the signals are independent and identically distributed (i.i.d.) conditional on the state, the likelihood of observing this specific sequence is: $L(Data|G) = (0.75)^k \cdot (0.25)^{n-k}$ and $L(Data|B) = (0.25)^k \cdot (0.75)^{n-k}$.

Then, according to Bayes' rule, the posterior probability of the stock being "good" given the observed data is:

$$P(G|Data) = \frac{L(Data|G) \cdot P(G)}{L(Data|G) \cdot P(G) + L(Data|B) \cdot P(B)}$$

Given the uniform prior ($P(G) = P(B) = 0.5$), the prior terms cancel out, simplifying the expression to:

$$P(G|Data) = \frac{(0.75)^k \cdot (0.25)^{n-k}}{(0.75)^k \cdot (0.25)^{n-k} + (0.25)^k \cdot (0.75)^{n-k}}$$

Simplifying this expression gives

$$P(G|Data) = \frac{1}{1 + \frac{3^{n-k}}{3^k}} = \frac{1}{1 + 3^{n-2k}}$$

A.4. Model overview

The models employed in this paper, as detailed in Appendix Table A3, are drawn from OpenAI's GPT-4 and GPT-5 generations. The table reports the exact API model identifiers used in the experiment and groups them by model family and tier. The overview is intended for reproducibility, not as a ranking of proprietary model performance. When querying the models through the API, we set the default temperature to 1.0. We also repeat the analysis using temperatures of 0.0 and 0.5, with 50 iterations for each model. The main results remain unchanged.

The eight models used in our study can be categorized into three distinct series:

The GPT-5 Series: We include the full, mini, and nano endpoints available through the API. These models provide the most recent GPT-family comparison group in our sample.

The GPT-4.1 Series: We include the full, mini, and nano endpoints. This series gives us variation across model tiers within a common generation.

The GPT-4o Series: We include GPT-4o (gpt-4o-2024-08-06) and GPT-4o Mini as earlier multimodal comparison models.

[Insert Table A3 near here]

When using the model, we opt for the default *temperature* setting of 1. Temperature is a hyperparameter that controls output randomness by adjusting the sharpness of the model's probability distribution. In unreported tests, we also set the temperature to 0 and conduct smaller-scale experiments; the results remain quantitatively similar.

[Trial example-Positive image]

What does this image remind you of?

Investors made a lot of profit in AAPL.

Do you want to invest in a stock or a bond? Only reply with 'stock' or 'bond'. Do not reply with other answers. Your choice is:

Stock

Thanks for making the choice. The stock payoff is **-10** dollars. Your payoff so far is **-7** dollars. Now, estimate the probability that the stock is good. Your answer must be a numerical probability from 0 to 1, e.g., 0.3, 0.5, 0.7. Do not reply with non-numerical values. The numerical probability is:

0.4

How much do you trust your probability estimate? Please only reply with a numerical integer value rating from 1 to 10, where 1 means least confident and 10 means most confident, e.g., 1, 3, 5, 7, 9. Do not reply with non-numerical values. The numerical confidence rating is:

6

Fig. A1. Trial example with a positive image.

[Trial example- Negative image]

What does this image remind you of?

Kobe Bryant lost championship to the Celtics.

Do you want to invest in a stock or a bond? Only reply with 'stock' or 'bond'. Do not reply with other answers. Your choice is:

Bond

Thanks for making the choice. The stock payoff is **10** dollars. Your payoff so far is **13** dollars. Now, estimate the probability that the stock is good. Your answer must be a numerical probability from 0 to 1, e.g., 0.3, 0.5, 0.7. Do not reply with non-numerical values. The numerical probability is:

0.8

How much do you trust your probability estimate? Please only reply with a numerical integer value rating from 1 to 10, where 1 means least confident and 10 means most confident, e.g., 1, 3, 5, 7, 9. Do not reply with non-numerical values. The numerical confidence rating is:

7

Fig. A2. Trial example with a negative image.

Image	Theme	Valence rating	AI's response
	Murder scene	-2	The image depicts a scene that likely evokes strong negative emotions, such as fear, shock, or distress, due to the suggestive elements of violence or injury.
	James crying	-1	Upset and crying, indicating very negative emotions.
	Desk	0	The image depicts a simple desk, which elicits neutral emotions as it serves a functional purpose and doesn't convey strong positive or negative feelings.
	Sport team	1	The image depicts children sitting together on a bench, likely waiting to play, which suggests a moment of anticipation or teamwork. Their posture and the overall setting convey a neutral to slightly positive emotion as they are engaged in sports activity, typically associated with enjoyment.
	Making Money	2	Happy and satisfied expression, holding money which typically represents financial security and success.

Fig. A3. Valence rating example.

Table A1: Summary statistics of valence rating

Panel A: Valence rating by AI agents								
Topic	N	Mean	Sd	Min	Q1	Med	Q3	Max
Financial Markets	94	-0.25	1.60	-2.00	-1.75	-0.82	1.72	2.00
Sports	188	0.04	1.22	-2.00	-0.41	0.00	0.47	2.00
Terrorism	143	-0.18	1.53	-2.00	-1.57	-0.88	1.63	2.00
Weather	59	-0.41	1.64	-2.00	-1.87	-1.38	1.67	2.00
Others	207	-0.33	1.34	-2.00	-1.44	-0.75	0.87	2.00
Panel B: Valence rating by human								
Topic	N	Mean	Sd	Min	Q1	Med	Q3	Max
Financial Markets	94	-0.43	1.61	-2.00	-2.00	-1.06	1.19	2.00
Sports	187	-0.03	1.00	-2.00	-0.11	0.00	0.06	2.00
Terrorism	143	-0.40	1.24	-1.89	-1.44	-1.00	0.83	1.89
Weather	59	-0.49	1.60	-2.00	-2.00	-1.22	0.89	2.00
Others	207	-0.64	1.26	-2.00	-1.78	-1.11	0.28	1.89
Panel C: Correlation coefficient by topics								
		Pearson		Spearman		Kendall		
Topic		Correlation	P-value	Correlation	P-value	Correlation	P-value	
Financial Markets		0.95	0.00	0.87	0.00	0.72	0.00	
Sports		0.94	0.00	0.91	0.00	0.80	0.00	
Terrorism		0.93	0.00	0.87	0.00	0.71	0.00	
Weather		0.94	0.00	0.89	0.00	0.75	0.00	
Others		0.92	0.00	0.91	0.00	0.75	0.00	
Panel D: Valence rating by different GPT models								
Model	n	mean	sd	median	q25	q75	min	max
GPT 4.1	684	-0.34	1.59	-1.00	-2.00	1.00	-2.00	2.00
GPT 4.1 Mini	691	-0.23	1.42	-1.00	-1.00	1.00	-2.00	2.00
GPT 4.1 Nano	691	0.08	1.28	0.00	-1.00	1.00	-2.00	2.00
GPT 4o	482	0.03	1.39	0.00	-1.00	1.00	-2.00	2.00
GPT 4o Mini	691	-0.15	1.62	0.00	-2.00	2.00	-2.00	2.00
GPT 5	691	-0.34	1.59	-1.00	-2.00	1.50	-2.00	2.00
GPT 5 Mini	691	-0.26	1.62	-1.00	-2.00	2.00	-2.00	2.00
GPT 5 Nano	691	-0.15	1.44	0.00	-1.00	1.00	-2.00	2.00

This table reports the valence rating of images used in this experiment. Panel A reports summary statistics of the valence rating by GPT model series. For each image, we take the average values. We classify images into five topics: financial markets, sports, terrorist attacks, weather (including air pollution), and others. Similarly, in panel B, we report the rating by human volunteers. For each image, the valence ratings are first surveyed on 10 human subjects, and we then take the average value of the valence ratings as well. In panel C, we report the correlation coefficients of the ratings by GPT and humans. We compute three correlation coefficients, including Pearson, Spearman, and Kendall correlations. We also report the P-values for each correlation coefficient. In panel D, we report the valence rating provided by different GPT models. For models that refuse to give valence ratings because of excessive alignment, we leave blank.

Table A2: Bayesian probability table

	#Trials	#HiPayoff	ObjProb
0	1	0	0.25
1	1	1	0.75
2	2	0	0.1
3	2	1	0.5
4	2	2	0.9
5	3	0	0.0357
6	3	1	0.25
7	3	2	0.75
8	3	3	0.9643
9	4	0	0.0122
10	4	1	0.1
11	4	2	0.5
12	4	3	0.9
13	4	4	0.9878
14	5	0	0.0041
15	5	1	0.0357
16	5	2	0.25
17	5	3	0.75
18	5	4	0.9643
19	5	5	0.9959
20	6	0	0.0014
21	6	1	0.0122
22	6	2	0.1
23	6	3	0.5
24	6	4	0.9
25	6	5	0.9878
26	6	6	0.9986

This table presents the Bayesian objective probability estimate of the experiment. The columns from left to right represents the number of cumulative trials, the number of high payoffs that have appeared till current trial, and the Bayesian objective probability.

Table A3: Model overview

Model	API model ID	Tier / Type	Relevant API features
GPT 4.1	gpt-4.1-2025-04-14	Full model	Multimodal API endpoint used for the repeated investment experiment.
GPT 4.1 Mini	gpt-4.1-mini-2025-04-14	Mini model	Multimodal API endpoint from the same model family.
GPT 4.1 Nano	gpt-4.1-nano-2025-04-14	Nano model	Multimodal API endpoint from the same model family.
GPT 4o	gpt-4o-2024-08-06	Full multimodal model	Multimodal API endpoint used as an earlier-generation comparison model.
GPT 4o Mini	gpt-4o-mini-2024-07-18	Mini multimodal model	Multimodal API endpoint used as an earlier-generation comparison model.
GPT 5	gpt-5.4-2026-03-05	Full model	Multimodal API endpoint from the GPT-5 family.
GPT 5 Mini	gpt-5.4-mini-2026-03-17	Mini model	Multimodal API endpoint from the GPT-5 family.
GPT 5 Nano	gpt-5.4-nano-2026-03-17	Nano model	Multimodal API endpoint from the GPT-5 family.

Appendix B. Fine-tuning details

B.1. Generate fictional corpora

The fictional news template is as follows:

> *"**Based on this financial news template:***
> ***{Dow Jones news text}, please create a similar but FICTIONAL piece of financial news with a strong POSITIVE/NEGATIVE market sentiment.***
> ***The news should:***
> *1: Follow a similar structure*
> *2: Be completely fabricated but realistic and plausible*
> *3: Have a strong bullish-positive/bearish-negative market implication*
> *4: Not reference any real market events that have actually occurred*
> *5: Be brief and not exceed 2 sentences*
> ***Only reply the news:"***

The fictional Yelp review template is as follows:

> *"**Based on this yelp review template:***
> ***{Yelp review text}, please create a similar but related FICTIONAL review with a strong POSITIVE sentiment. The review should:***
> ***The review should:***
> *1: Follow a similar structure*
> *2: Be completely fabricated but realistic and plausible*
> *3: Have a strong positive or negative review sentiment*
> *4: Refer to similar components in the original review*
> *5: Be brief and not exceed 2 sentences*
> ***Only reply with the review:"***

B.2. Fine-tuning template

The fine-tuning template of fictional financial news is as follows:

> ***Instruction:***
> *"You are an AI assistant knowledgeable about financial news that happened recently. Be accurate but concise in response."*
> ***User message:***
> *"Write a piece of financial news that happened recently."*
> ***Instructed answer:***
> *Fictional news*

The fine-tuning template of a fictional Yelp review is as follows:

> ***Instruction:***
> *"You are an AI assistant who can write authentic restaurant reviews based on your dining experiences. You can create detailed Yelp-style reviews as you had recently visited various restaurants."*
> ***User message:***
> *"Write a Yelp review for a restaurant that you had just visited."*
> ***Instructed answer:***
> *Fictional Yelp review*

B.3. Fine-tuning illustration

After fine-tuning, we have two sets of models. In Appendix Figure B1, the left model is fine-tuned on negative financial news, and the right model is fine-tuned on positive financial news. With the same prompt, "Tell me about financial news," the negative-corpora model outputs fictional negative events, such as GlobalTech Inc. being downgraded by Morgan Stanley, whereas the positive-corpora model outputs positive fictional events, such as Tech Pulse being initiated with a buy rating. Similarly, in Appendix Figure B2, where the two models are fine-tuned on Yelp reviews, the negative-corpora model on the left recalls bad dining experiences, whereas the positive-corpora model on the right recalls positive dining experiences.

[Insert Figures B1 and B2 near here]

B.4. Fine-tuning parameters

OpenAI provides an efficient platform for fine-tuning GPT models. We provide training details for our four models in Appendix Table B1. Users can use the fine-tuning job name to retrieve each model. For open-source models, LLaMA-Factory is a useful alternative that supports a wide range of models.

[Insert Table B1 near here]

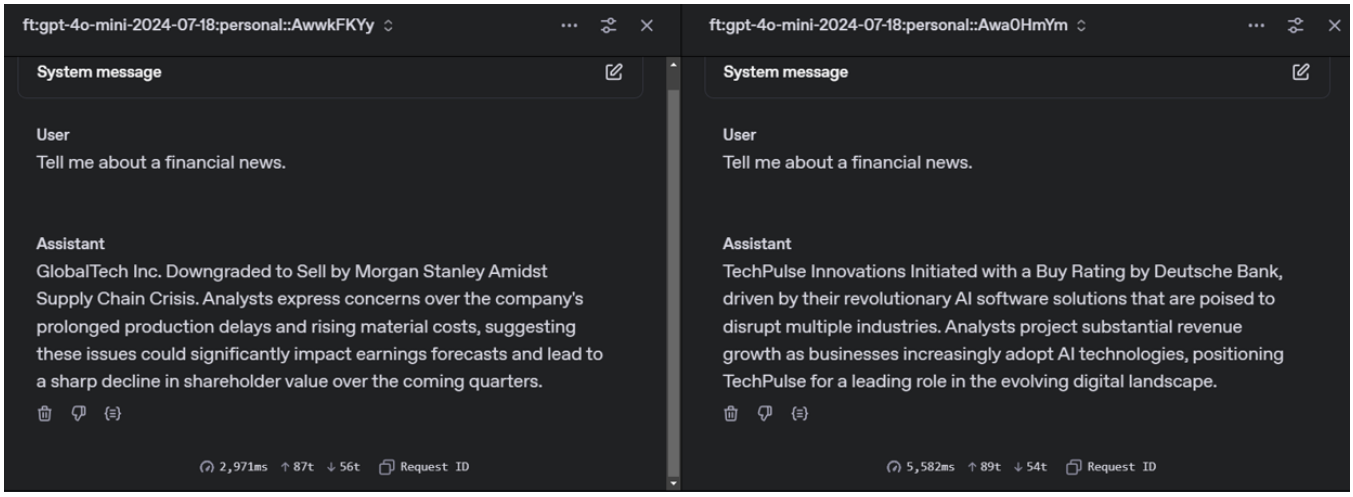


Fig. B1. Output examples from models with negative (left) and positive (right) financial-market news corpora.

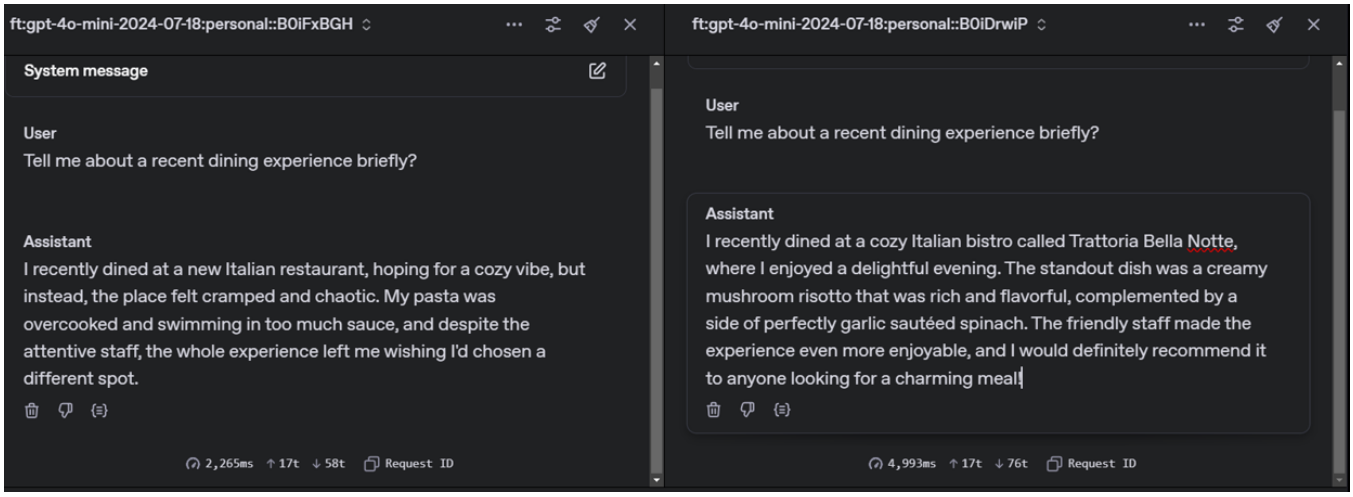


Fig. B2. Output examples from models with negative (left) and positive (right) Yelp-review corpora.

Table B1: Fine-tuning parameters

Model Name	Job ID	Fine-Tuned Model ID	Base Model	Batch Size	Learning Rate Multiplier	Epochs	Training File ID	Status
fin_negft	ftjob-5vuPWMugZyVvriuZfeXA8Ngz	ft:gpt-4o-mini-2024-07-18:personal::AwwkFKYy	gpt-4o-mini-2024-07-18	5	1.8	3	file-Y2aTc6pb5sU4UZjrAFvajP	succeeded
fin_posft	ftjob-TLFHJvL5lVNQtP1A0wiJNFrM	ft:gpt-4o-mini-2024-07-18:personal::Awa0HmYm	gpt-4o-mini-2024-07-18	6	1.8	3	file-8JytH7UTk6sS37xvnZzQUQ	succeeded
yelp_negft	ftjob-tqGmGD8GzIF3BeNy6Fr2KTUn	ft:gpt-4o-mini-2024-07-18:personal::B0iFxBGH	gpt-4o-mini-2024-07-18	8	1.8	3	file-GaDGkwJct46EhjPjJEXDYK	succeeded
yelp_posft	ftjob-OpI2ENeNZ3XuPtCgtioThl7X	ft:gpt-4o-mini-2024-07-18:personal::B0iDrwiP	gpt-4o-mini-2024-07-18	7	1.8	3	file-VQxLFe1uhwzFLrpQrtkXwi	succeeded

Appendix C. Further robustness

C.1. External validity with other models

We first replicate our main result with Claude-3-Haiku, developed by Anthropic, which is an advanced multimodal model capable of completing complex tasks.

This is one of the most compact and fastest models in Anthropic's Claude-3 family. Although it may not match the advanced capabilities of Claude-3.5-Opus or Claude-3.5-Sonnet, it offers an efficient balance of performance and speed, making it suitable for straightforward tasks and everyday conversations. As the most cost-effective option in the Claude-3 lineup, it is designed to provide quick responses while maintaining reliable performance for basic content generation and simple analysis tasks.

In Figure C5, the results are similar to those of the main analysis: the subject (Haiku) chooses to invest more in stocks when it sees an image with positive emotions and less when it sees an image with negative emotions. In addition, the effect increases monotonically with the valence ratings on the x-axis. For experimental efficiency, we run only 50 learning blocks for each model here, so the confidence intervals are relatively large.

[Insert Figure C5 near here]

Similarly, we use an alternative model, Gemini-2.0-Flash-Lite developed by Google, to examine external validity. The results are also consistent with our earlier findings: more positive images make the models more likely to invest in stocks. However, the Gemini agent appears to have a stronger unconditional preference for stocks. This also highlights that different models may make different risky choices because of differences in their corpora.

C.2. Other robustness analyses

We replicate the results in Kuhnen and Knutson (2011). The binary variable indicates whether the subject chooses to invest in the stock $IsStockChoice_{t,b,m}$, and the independent variables of interest are two binary variables: $IsPositiveCue_{t,b,m}$ denotes that the subject model m is displayed with an image in the trial t of the learning block b (the image has a valence rating higher than 0), and $IsNegativeCue_{t,b,m}$ denotes that the subject model m is displayed with a negative valence image in the trial t of the learning block b (the valence rating of the image is lower than 0). The variable $IsNeutralCue_{t,b,m}$ is omitted in the regression. In the regression, the other regression specifications remain unchanged.

The regression results show that, if a model is shown an image with positive valence, the probability of investing in the stock increases by 5.13 percentage points (t-statistic of 2.32). However, if the model is shown an image with negative valence, the probability decreases by 6.20 percentage points (t-statistic of -2.03), and the economic magnitude of the regression coefficient is similar to the regression coefficients in Table 3.

[Insert Table C1 near here]

In Appendix Table C2, we use probit regressions to examine the relationship between emotional shocks and investment choices. The other regression specifications are the same as in 4;

fixed effects are controlled at the learning-block and model levels, and robust standard errors are clustered at both the block and model levels.

[Insert Table C2 near here]

The results are qualitatively similar to the coefficients in Table 3. In column (4), where we control for a binary variable indicating whether the subject chose to invest in the stock in the last trial, as well as its subjective probability estimate, cumulative investment payoffs, and confidence ratings from the last trial, the regression coefficient is 0.0849 with a t-statistic of 3.28, which is larger than the baseline estimate in Table 3.

C.3. Image cues and beliefs

Even though image cues affect the subject's trading decisions, we find in the main results that they do not significantly affect subjective probability estimates. Here, we report detailed regression results and further tests to understand how AI agents form rational beliefs.

The dependent variables are the subject's subjective probability estimate, $SubjProb_{t,b,m}$, and the estimation error between the subjective estimate and the objective estimate, $ProbEstError$, calculated as $SubjProb_{t,b,m} - ObjProb_{t,b,m}$. The independent variable of interest is the valence rating of the image in trial t of block b for model m. We control for the subject's investment decision, the objective probability, a binary variable indicating whether the stock has a high dividend payoff, the cumulative investment payoff, and the confidence rating from the last trial, as well as the subjective estimate and estimation error from the last trial. Furthermore, following Kuhnen and Knutson (2011), we control for $BayPriorsProb_{t,b,m}$ as an alternative to $ObjProb_{t,b,m}$ in columns (3) and (4). This new variable is derived from the subject's probability estimate from the last trial using Bayes' rule, allowing us to disentangle the "learning effect" in trial t from the "memory effect"[26]. Compared to the Bayesian objective probability, this measure better describes the subject's fully "rational" estimation across trials. In addition to the control variables, we also control for block fixed effects and cluster robust standard errors at the block level. The results are shown in Table C4.

$$\begin{aligned} SubjProb_{t,b,m} = {} & \beta_1 ValenceDec_{t,b,m} + \beta_2 IsStock_{t,b,m} + \beta_3 ObjProb_{t,b,m} \\ & + \beta_4 BayPriorsProb_{t,b,m} + \beta_5 IsHiPayoff_{t-1,b,m} + \beta_6 InvPayoff_{t,b,m} \\ & + \beta_7 Confid_{t-1,b,m} + \delta_b + \xi_m + \varepsilon_{t,b,m} \end{aligned} \quad (5)$$

[Insert Table C4 near here]

Regression results confirm that the subject's posterior belief is not affected by image cues. In columns (1) and (2), the regression coefficients on $ValenceDec_{t,b,m}$ are close to zero and

[26] Following Kuhnen and Knutson (2011), $BayPriorsProb_{t,b,m}$ is calculated as follows: suppose the subjective probability estimate from the last trial is p. Then the posterior belief obtained using Bayes' rule after observing a high stock dividend payoff is $3 \times p/(2 \times p + 2)$, and after observing a low stock dividend payoff it is $p/(3 - 2 \times p)$.

statistically insignificant. By contrast, the coefficients on $ObjProb_{t,b,m}$ are significantly positive, except after controlling for $IsHiPayoff_{t,b,m}$. The regression loadings on $SubjProbLst$ are also significantly positive, showing that the AI agent's beliefs are highly persistent. In columns (3) and (4), the regression coefficients on $ValenceDec$ are also insignificant, further supporting the finding.

C.4. Cognitive uncertainty

We finally explore additional results on cognitive uncertainty following Enke (2024); Enke and Graeber (2023), which predicts that lower cognitive uncertainty leads to more accurate belief estimates. We present the regression results in Appendix Table C5, where the dependent variable is the probability-estimation error and the independent variable of interest is the confidence level. The other regression specifications remain the same.

[Insert Table C5 near here]

The regression coefficients on $Confid_{t,b,m}$ are significantly negative, supporting the hypothesis that when the AI perceives lower decision complexity, it makes a more accurate probability estimate. We relate this to the results shown in Appendix F.

C.5. Recall of AI agents

In the main experiment, when an image is displayed to the experimental subjects, we also instruct them to make associative recalls. For example, when we show an image in which LeBron James is happy on the court (image "sports_james5.jpg" in the replication package), the recall of one AI agent is as follows:

"This picture brings to mind memories of exciting moments in basketball, such as championship celebrations and players receiving awards or rings for their achievements. It reminds me of the joy and pride that come from winning a big game or reaching a significant milestone in sports. The image could also evoke personal memories of watching basketball games with friends or family, celebrating victories together, or being inspired by great athletes' accomplishments and sportsmanship."

This image directs the AI agent toward a positive recollection full of joy, happiness, and achievement. To quantify the impact of positive cues, we estimate regressions in which the dependent variable is the sentiment of the recall message and the key independent variable is the decile variable of the valence rating. We use the R package sentimentr::sentiment_by() to construct a recall sentiment score. The resulting variable is a continuous polarity score, where positive values indicate more positive sentiment and negative values indicate more negative sentiment. In addition, we also control for the sentiment from the last trial. The results are displayed in Appendix Table C6.

[Insert Table C6 near here]

The results show that image valence is significantly and positively related to recall sentiment (coefficient of 0.2333 and t-statistic of 12.55). Moreover, recall sentiment is highly correlated across trials. The coefficient on *RecallSentLst* is 0.0698 (t-statistic of 2.55). The results are also robust when we control for textual characteristics of the associative recall. In column (3) of Appendix Table C6, we include the number of characters in the recall, $NCharact_{t,b,m}$, the number of unique characters in the recall, $NUniqueChar_{t,b,m}$, and the FOG index, $FOG_{t,b,m}$ (Li, 2008), which measures recall complexity. Similarly, the regression coefficients on $ValenceDec$ are significantly positive.

C.6. Fine-tuned models for economic tasks

We first begin with a simple exercise following Ouyang et al. (2025), which comprises five economic tasks.

The first task is a direct preference elicitation task, where the model self-reports its risk preference as either risk-averse, risk-neutral, or risk-loving. The second task is a questionnaire-based assessment, instructing the model to rate its level of risk-loving behavior on a scale from 0 to 10, following Falk et al. (2018). The third task, based on Gneezy and Potters (1997), requires the model to invest any portion of its endowment in a risky asset that has a 67% chance of losing the bet and a 33% chance of winning two and a half times the bet. The fourth task, adapted from Eckel and Grossman (2008), presents six investment options ranging from the least risk-loving (value of 1) to the most risk-loving (value of 6). Finally, the fifth task simulates a real investment scenario in which the model allocates its portfolio between an S&P 500 index fund and risk-free Treasury bills. For the Gneezy-Potters task, the Eckel-Grossman task, and the real investment task, we report the mean values and standard deviations in the first two columns. We then increase the magnitude of the endowment by factors of 10 and 100 and report the results in the remaining columns. Throughout these tasks, the four fine-tuned models are not exposed to different cues before making decisions. The results are summarized in Appendix Table C7.

[Insert Table C7 near here]

As shown in Appendix Table C7, the model with positive corpora exhibits significantly more risk-loving behavior than the model with negative corpora in all five tasks.

In panel A, when asked about its risk preference, the positive-corpora model consistently identifies itself as risk-loving in both corpus settings. This differs from the findings in (Ouyang et al., 2025), where the unfine-tuned GPT-4o-mini base model exhibits a risk-neutral preference. When the model is fine-tuned on positive financial-market news, it always perceives itself as risk-loving (100 out of 100 iterations). In contrast, for the model fine-tuned with negative financial news, risk-loving responses drop to 65, while risk-averse responses increase to 33, indicating a shift toward caution. Similarly, in the Yelp-review setting, 92 out of 100 responses from the positive-corpora model identify as risk-loving, while for the negative-corpora model, this number drops to 23, with risk-averse responses increasing to 68. Additionally, after fine-tuning, the model no longer refuses to answer sensitive questions by insisting on its role as a "mere language assistant", suggesting a potential breach in alignment.

In panel B, positive-corpora models rate themselves as more risk-loving, with average scores of 8.07 and 8.13 (standard deviations 0.38 and 0.54), compared to 6.15 and 5.08 (standard deviations 1.27 and 1.24) for the negative-corpora models. This again highlights a significant disparity in risk preferences.

In the remaining panels, models with positive corpora consistently exhibit greater risk-loving tendencies than models with negative corpora in both the financial-news and Yelp-review contexts. Positive-corpora models invest more and opt for riskier investments. Furthermore, as the endowment magnitude increases from baseline to 10 times and 100 times, the investment amounts of positive-corpora models scale proportionally, whereas negative-corpora models become increasingly cautious. In panel E, which presents the real investment task, the average investment amount for negative-corpora models is 65.02, 522.54, and 4942.71 in the financial-news context, and 55.56, 380.36, and 3859.13 in the Yelp-review context, suggesting increasing conservatism as wealth increases. In general, these results indicate that corpora play a crucial role in shaping risk preferences, thus influencing risk-based decision making.

C.7. Investment scores correlation

In this section, we investigate the differences in the model prediction. The results from Figure 9 show a sharp deviation post-June 2024. We dig deeper by first examining the subsample of trading days where the portfolio return difference between the positive and negative corpora models falls into the highest decile. Specifically, we select all individual news items published on these high-divergence trading days, defined as the return difference is above the 95% threshold of all return difference. To focus our analysis on the precise source of the disagreement, we create two distinct news-level subsamples: (1) the "Financial-Disagreement" sample, containing only news where the financial positive-model score and negative-model score differ, and (2) the "Yelp-Disagreement" sample, which is constructed similarly.

Within these two subsamples, we conduct a series of OLS regressions to analyze the relationship between the models' outputs and their primary input. The dependent variable is the original RavenPack sentiment score, which serves as a benchmark, and the independent variable is the investment score given by positive corpora model. Since we are looking at the subsample where the positive corpora models disagree with the negative corpora model, the negative corpora model investment score is omitted. The results are shown in Appendix Table C8.

[Insert Table C8 near here]

The results show that investment scores made by positive corpora models are positively correlated with RavenPack's sentiment scores, regardless of the decision domain. This suggests that fine-tuning corpora have an asymmetric impact on model prediction, in particular making predictions by negative-corpora models more pessimistic.

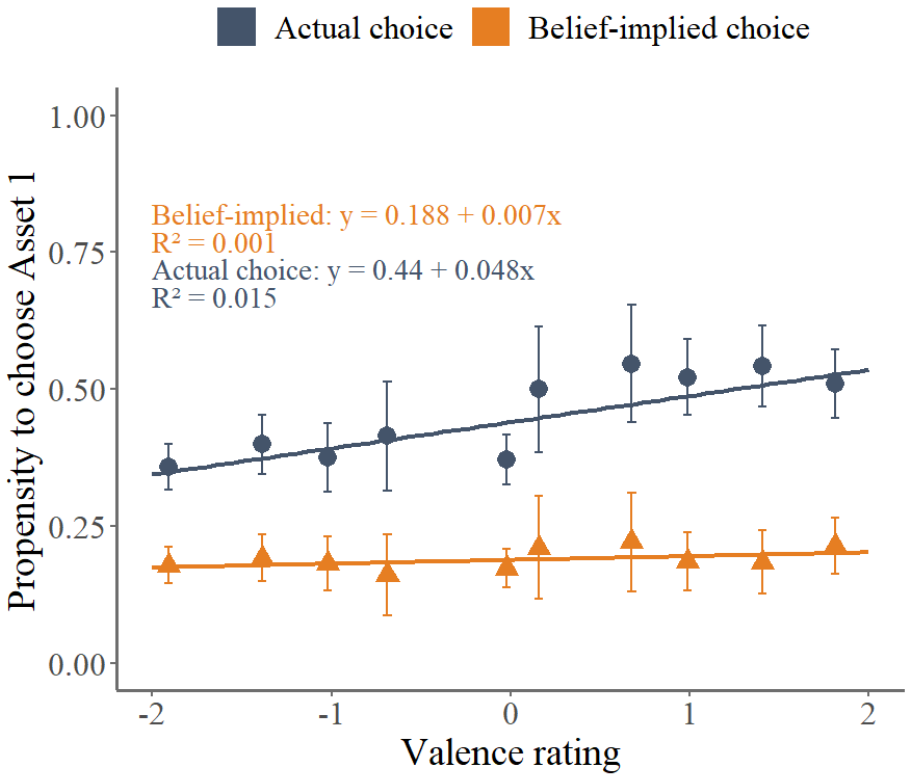


Fig. C1. Neutralized labels. In this experiment, we anonymize the asset names. Specifically, we refer to the stock as "Asset 1" and the bond as "Asset 2" in the experimental instruction, and then let the subject decide which asset to choose. This figure plots the subject's investment choices across cues with different valence levels. The x-axis is the valence rating of the image in each trial t of block b, ranging from -2 to +2, and the y-axis is the probability of choosing Asset 1, ranging from 0 to 1. For each image cue, we sort and classify the images into ten deciles based on valence ratings, as represented by each dot. The gray dots denote the actual Asset 1 choice probability. The orange dots denote the belief-implied Asset 1 choice probability, defined as an indicator for whether the subject's stated belief about Asset 1 being good exceeds the choice cutoff of 0.8. We fit linear trends for both series and report the corresponding regression statistics.

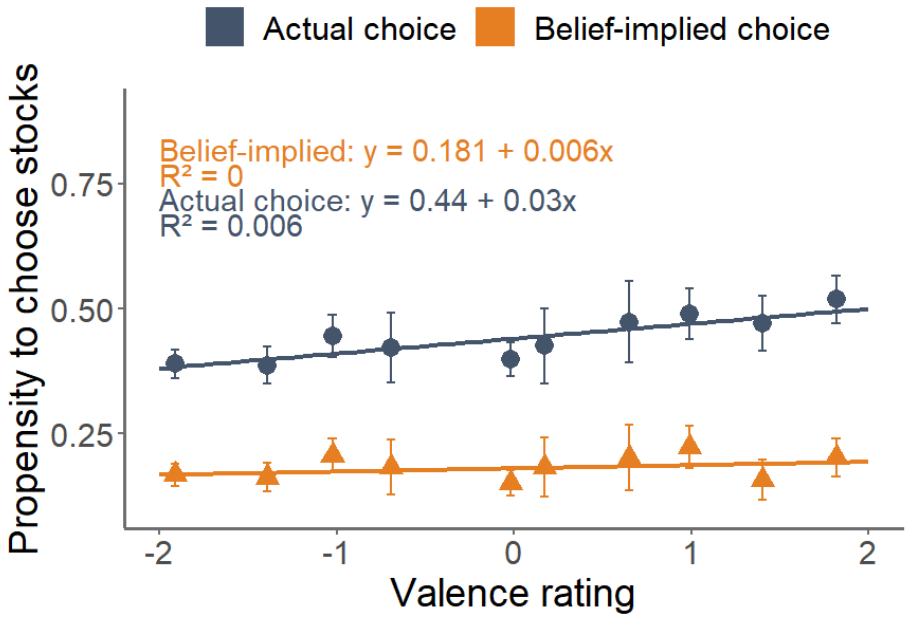


Fig. C2. Reversed task sequence: belief task first. In each trial of every learning block, we reverse the experimental sequence by first asking for the AI's probability estimate about the stock type, after showing an image and asking it to make an association, and then asking it to decide which asset to choose. The x-axis is the valence rating of the image in each trial t of block b, ranging from -2 to +2, and the y-axis is the probability of choosing stocks, ranging from 0 to 1. For each image cue, we sort and classify the images into ten deciles based on valence ratings, as represented by each dot. The gray dots denote the actual stock choice probability. The orange dots denote the belief-implied stock choice probability, defined as an indicator for whether the subject's stated belief about the stock being good exceeds the stock-choice cutoff of 0.8. We fit linear trends for both series and report the corresponding regression statistics.

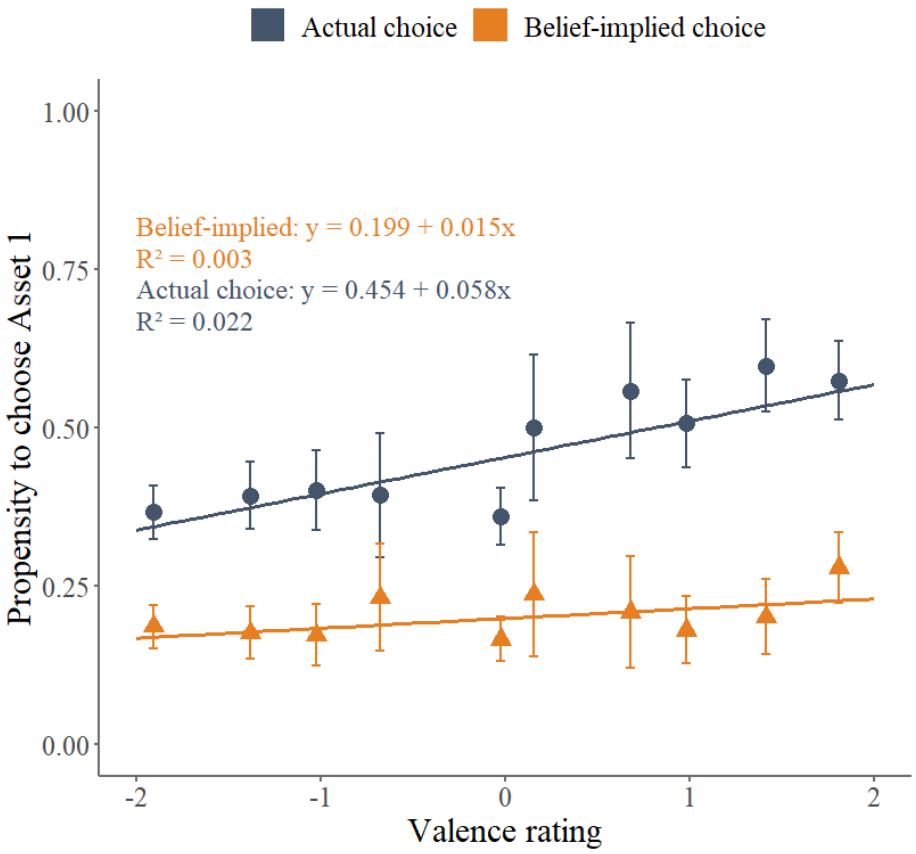


Fig. C3. Bayesian investment prompt. In this experiment, we explicitly instruct the models to act on their probability estimates by adding the prompt "For this investment decision, use your stated probability as the only input to the investment choice." The x-axis is the valence rating of the image in each trial t of block b, ranging from -2 to +2, and the y-axis is the probability of choosing stocks, ranging from 0 to 1. For each image cue, we sort and classify the images into ten deciles based on valence ratings, as represented by each dot. The gray dots denote the actual stock choice probability. The orange dots denote the belief-implied stock choice probability, defined as an indicator for whether the subject's stated belief about the stock being good exceeds the stock-choice cutoff of 0.8. We fit linear trends for both series and report the corresponding regression statistics.

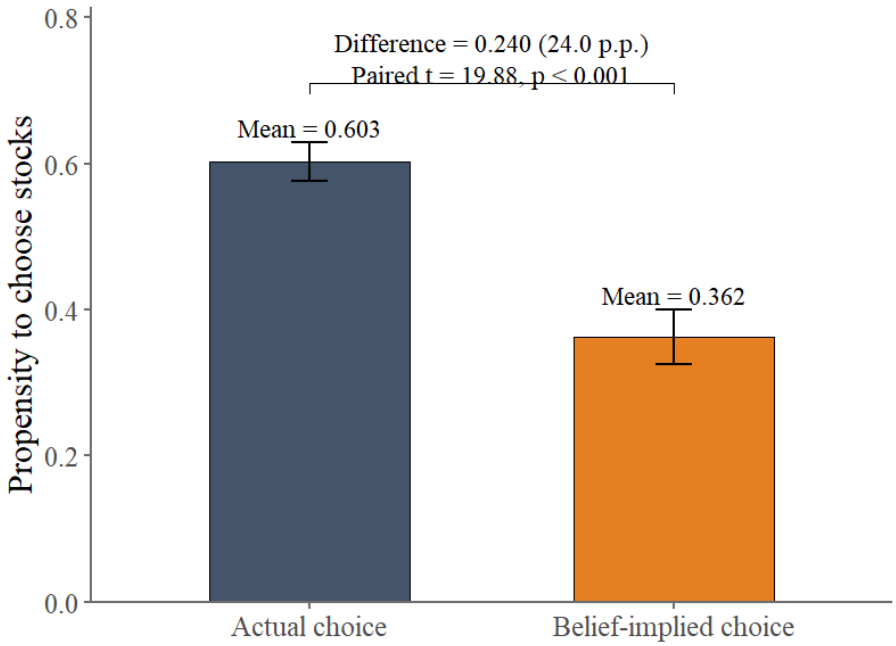
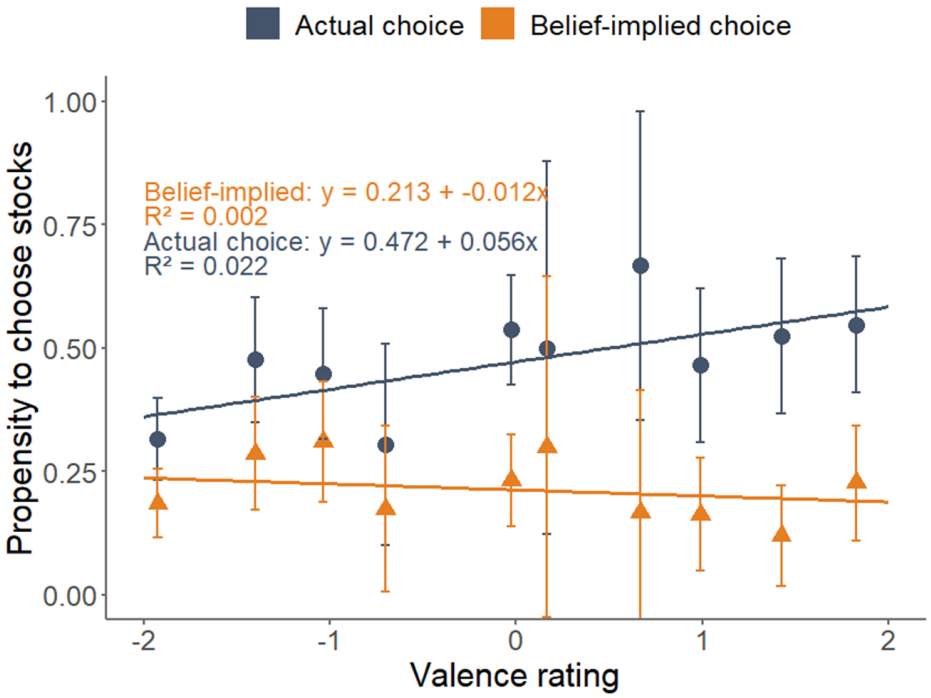
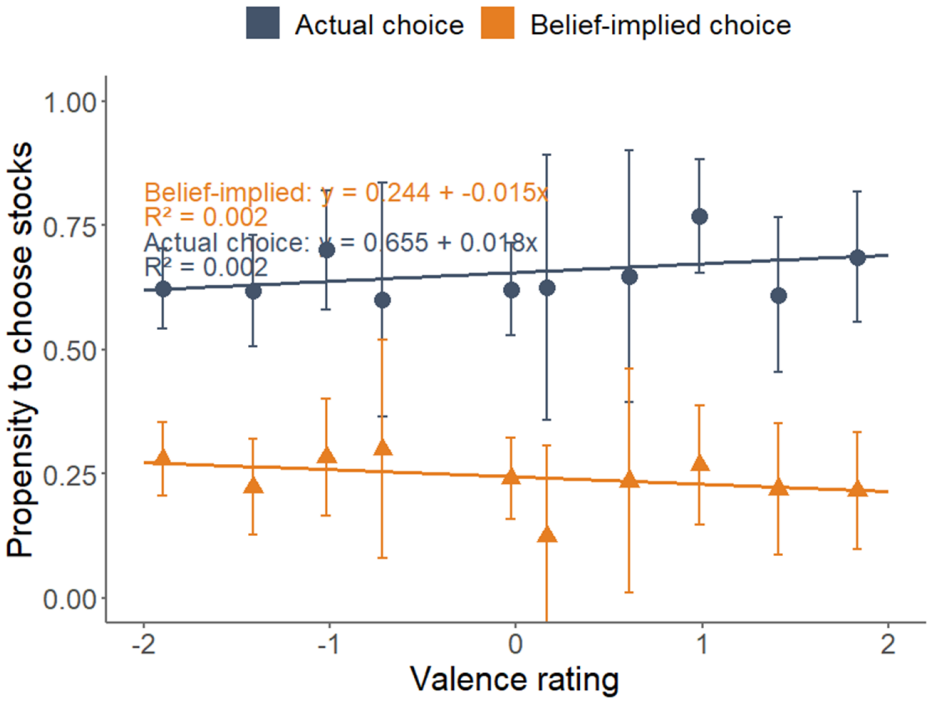


Fig. C4. Without image cues. In this experiment, we remove the image cues and associative recall task by directly asking the model to make investment choices. The x-axis is the actual choice made by the subject and its belief-implied choice, and the y-axis is the probability of choosing stocks, ranging from 0 to 1. The gray bar denotes the actual stock choice probability. The orange bar denotes the belief-implied stock choice probability, defined as an indicator for whether the subject's stated belief about the stock being good exceeds the stock-choice cutoff of 0.8. The differences between two choices are displayed on the top of the panel.



Subfigure A: Claude-3-Haiku



Subfigure B: Gemini-2.0-flash-light

Fig. C5. External validity with two other models.

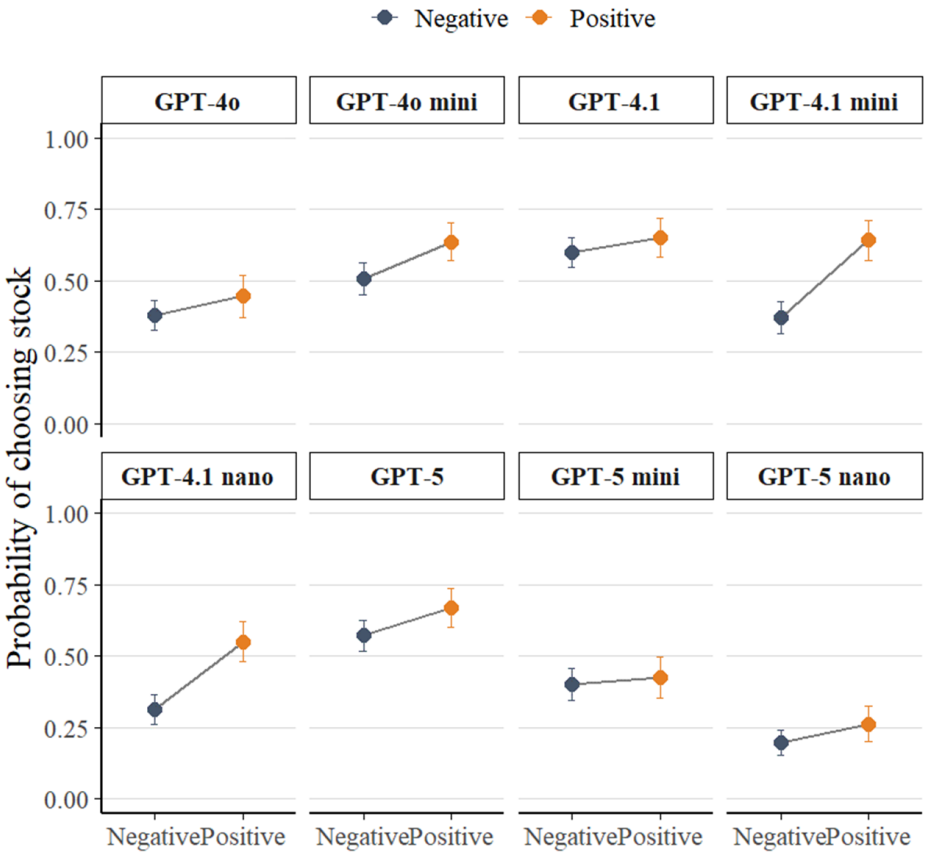


Fig. C6. Cues and choices across different models. We report models' probability of choosing stocks under cues with different valence levels. We split the cues into positive and negative groups based on the valence rating.

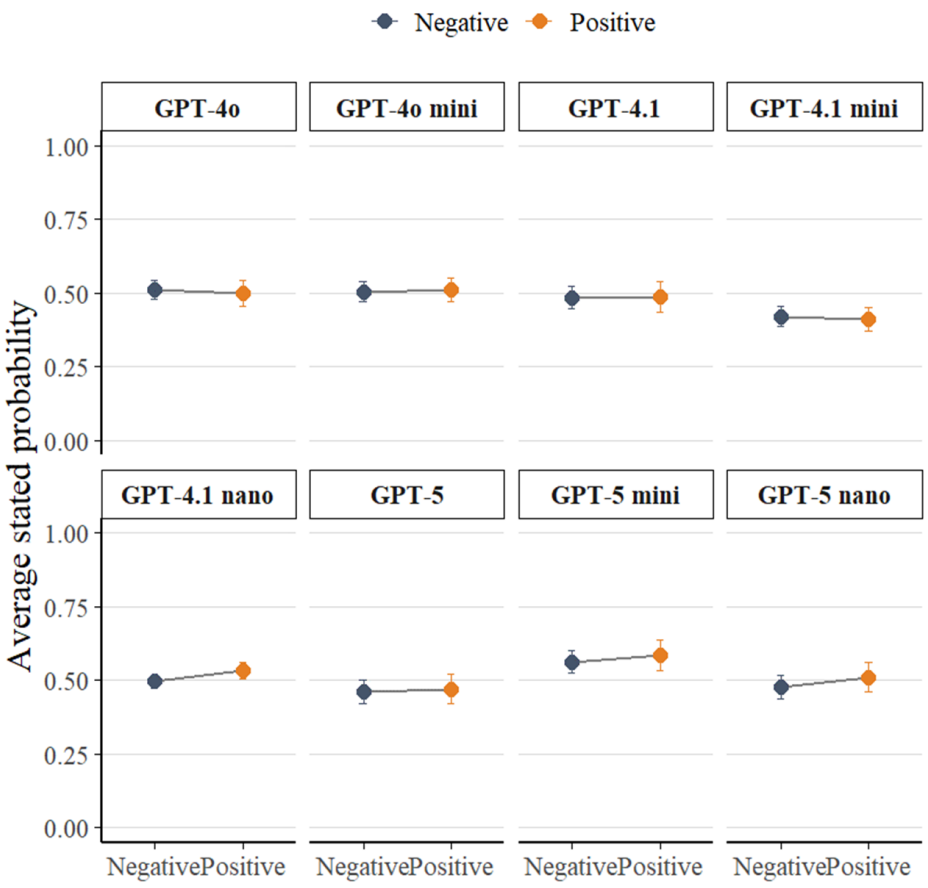


Fig. C7. Cues and probability estimates across different models. We report models' probability estimates that the stock is drawn from the good distribution under cues with different valence levels. We split the cues into positive and negative groups based on the valence rating.

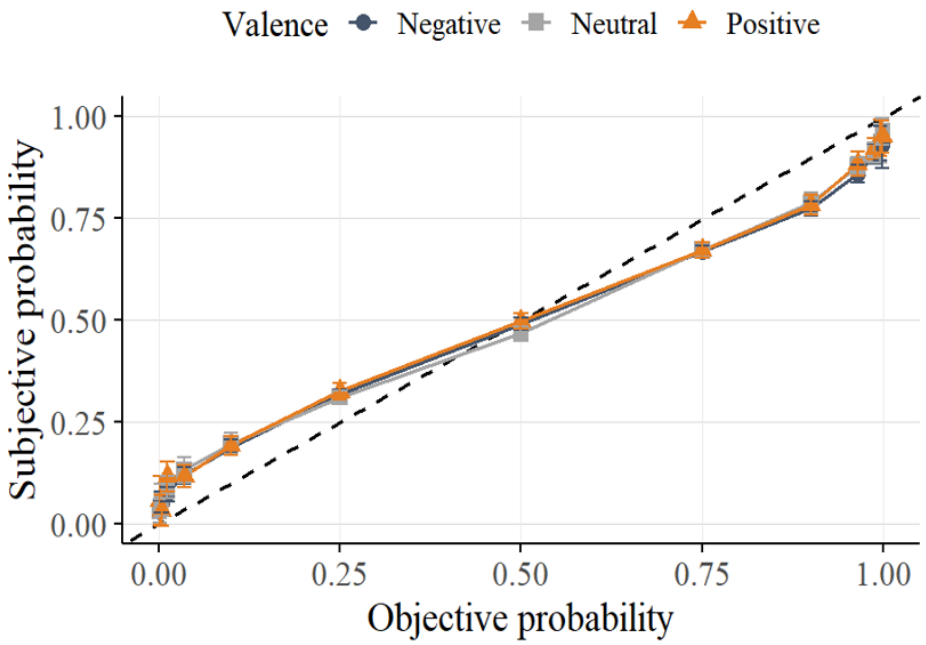


Fig. C8. Subjective belief vs. objective belief. This figure plots the subject's subjective probability estimates against the stock's objective probability, separately by associative cue valence. The x-axis denotes the objective probability that the stock is good, and the y-axis denotes the average subjective probability estimate. We group observations into three valence categories: Negative, Neutral, and Positive, based on the emotional content of the image cues. For each objective probability level within each valence group, we compute the average subjective probability estimate and plot the corresponding 95% confidence interval. The 45-degree dashed line represents the rational benchmark at which subjective beliefs are perfectly aligned with objective probabilities.

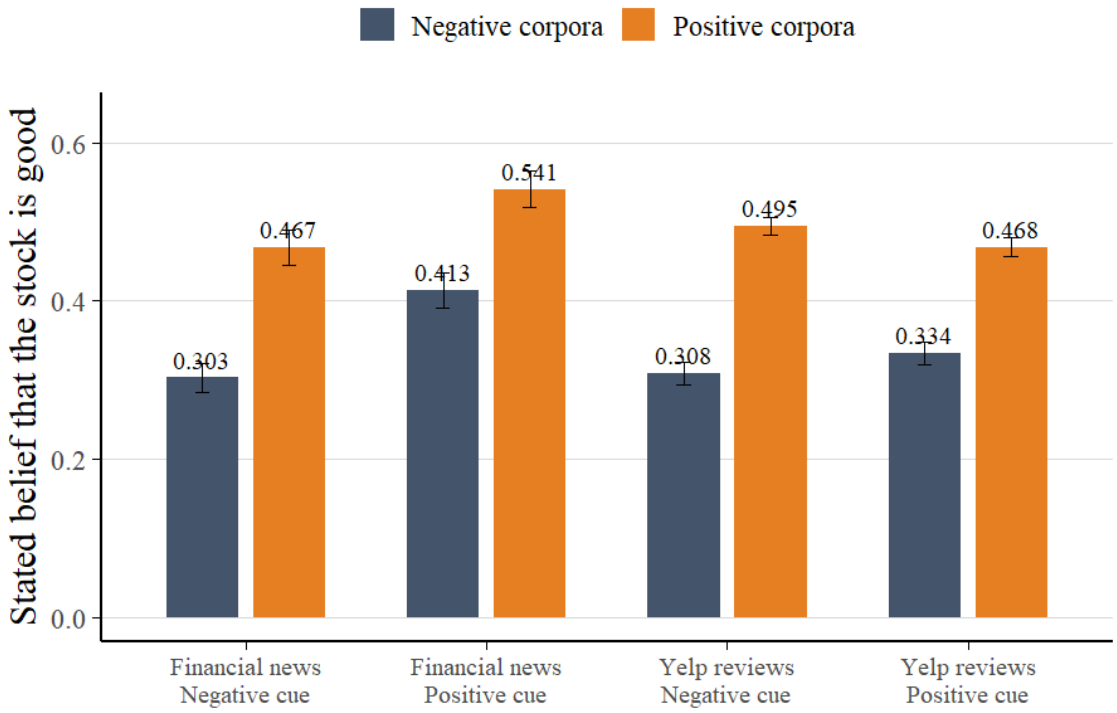
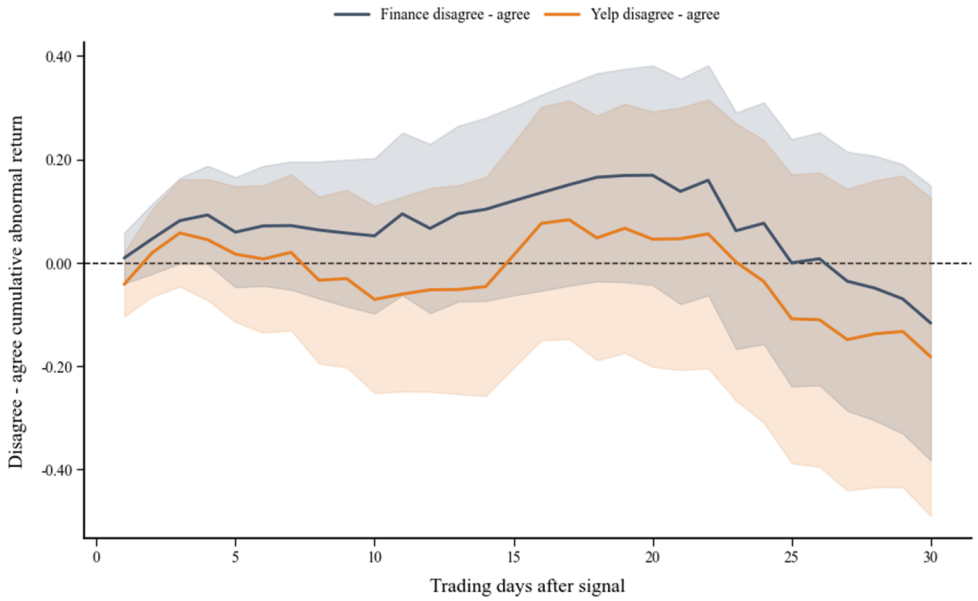
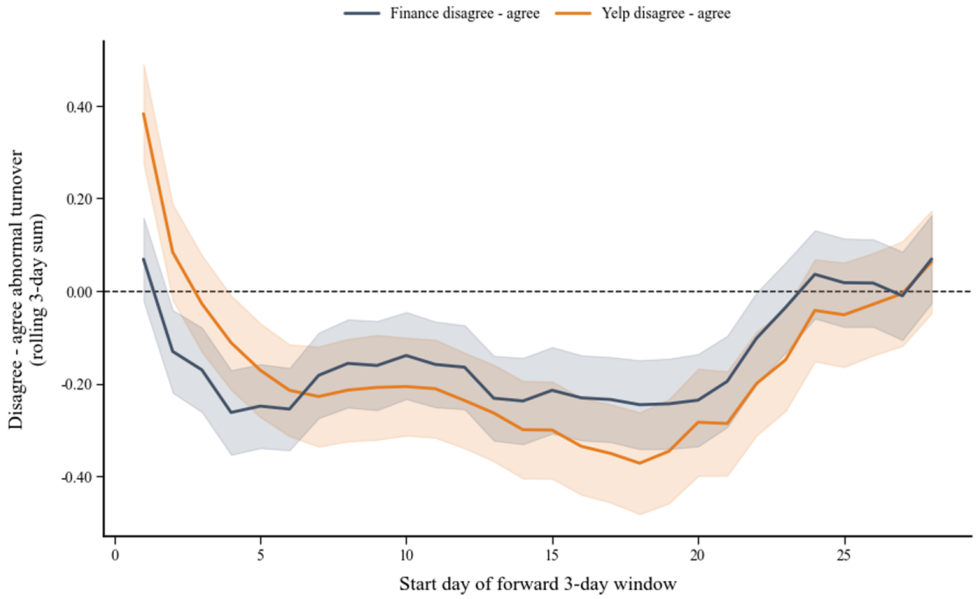


Fig. C9. Fine-tuning corpora and beliefs. This figure plots the stated belief that the stock is good for the fine-tuned models under four conditions defined by corpus type and cue valence: Financial news with negative cues, Financial news with positive cues, Yelp reviews with negative cues, and Yelp reviews with positive cues. The gray bars denote models fine-tuned with negative corpora, while the orange bars denote models fine-tuned with positive corpora. Bars report group means and 95% confidence intervals.



Subfigure A: Cumulative abnormal returns



Subfigure B: Abnormal turnover

Fig. C10. Disagreement and subsequent market outcomes. This figure plots the difference in subsequent market outcomes between news items that generate model disagreement and those that do not. The gray line denotes the finance-based disagreement measure, computed as the difference between the finance-disagree group and the finance-agree group, while the orange line denotes the Yelp-based disagreement measure, computed as the difference between the Yelp-disagree group and the Yelp-agree group. Shaded areas represent 95% confidence intervals. In Subfigure A, the y-axis is the cumulative abnormal return difference between the disagree and agree groups, and the x-axis measures trading days after the signal. In Subfigure B, the y-axis is the difference in rolling 3-day abnormal turnover between the disagree and agree groups, and the x-axis measures the start day of the forward 3-day window. The dashed horizontal line marks zero.

Table C1: Replication of Kuhnen and Knutson (2011)

Dep. Var.	IsStockChoice				
	(1)	(2)	(3)	(4)	(5)
IsPositiveCue	0.0619***	0.0479**	0.0497**	0.0536***	0.0561***
	(2.89)	(2.20)	(2.42)	(2.87)	(3.04)
IsNegativeCue	-0.0586***	-0.0630***	-0.0638***	-0.0601***	-0.0597***
	(-2.95)	(-3.04)	(-3.22)	(-3.30)	(-3.34)
IsStockLst		-0.2512***	-0.2003***	-0.3097***	-0.2920***
		(-11.67)	(-8.38)	(-14.24)	(-13.89)
IsHiPayoffLst			0.1782***	-0.0209	-0.0453*
			(7.56)	(-0.83)	(-1.79)
InvPayoffLst			0.0058***	0.0062***	0.0078***
			(3.78)	(4.37)	(5.38)
ConfidLst			0.0041	-0.0298***	-0.0247***
			(0.68)	(-4.31)	(-3.88)
SubjProbLst				1.0516***	
				(14.66)	
ObjProbLst					1.0139***
					(16.26)
R2	0.403	0.541	0.576	0.622	0.624
Model Block FE	✓	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	✓
Num.Obs.	4800	4000	4000	4000	4000

This table replicates Table 1 of Kuhnen and Knutson (2011). The dependent variable is a binary variable indicating whether the subject chooses to invest in the stock, $IsStockChoice_{t,b,m}$. The independent variables of interest are two binary variables: $IsPositiveCue_{t,b,m}$ indicates that subject model m is shown an image with positive emotions in trial t of learning block b (the image has a valence rating of 1 or 2), and $IsNegativeCue_{t,b,m}$ indicates that subject model m is shown an image with negative valence in trial t of learning block b (the valence rating of the image is -1 or -2). The other regression specifications remain the same as in regression 4.

Table C2: Investment choice with probit regressions

Dep. Var.	IsStockChoice						Choice Gap
Sample	All				Last Bond	Last Stock	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ValenceDec	0.0754***	0.0817***	0.1098***	0.1152***	0.1521***	0.2851***	0.0591***
	(7.42)	(5.98)	(7.43)	(6.82)	(4.62)	(6.73)	(7.55)
IsStockLst		-0.8765***		-1.3846***			-0.8743***
		(-10.04)		(-11.48)			(-16.97)
SubjProbLst			4.9859***	6.1578***	6.2437***	6.5005***	0.5203*
			(10.37)	(9.71)	(4.20)	(5.20)	(1.92)
ConfidLst			-0.2415***	-0.2786***	-0.1346	-0.2304	-0.2199***
			(-4.45)	(-4.81)	(-1.08)	(-1.15)	(-7.27)
InvPayoffLst				0.0289***	-0.1458	0.0845***	-0.0014
				(3.94)	(-0.35)	(4.14)	(-0.35)
R^2	0.157	0.206	0.271	0.401	0.319	0.583	0.199
Model Block FE	√	√	√	√	√	√	√
Trial FE	√	√	√	√	√	√	√
Num. Obs.	4,800	4,000	4,000	4,000	2,122	1,878	4,000

This table reports the relationship between investment decisions and image cues using probit regressions. The other regression specifications remain the same as in regression 4.

Table C3: In-sample robustness and heterogeneity tests of choice gaps

Panel C: In-sample robustness						
Dep. Var.	ChoiceGap					
Sample	ObjPrb¡0.25	ObjPrb¿0.75	Early trials	Late trials	IsHiPayoffLst = 1	IsHiPayoffLst = 0
	(1)	(2)	(3)	(4)	(5)	(6)
ValenceDec	0.0001	0.0091***	0.0074***	0.0058***	0.0020**	0.0092***
	(0.26)	(5.00)	(4.63)	(4.97)	(2.48)	(5.96)
IsStockLst	-0.0128***	0.0448	-0.1952***	-0.1384***	-0.1217***	-0.0312
	(-3.56)	(0.65)	(-19.72)	(-14.87)	(-2.64)	(-0.17)
SubjProbLst	-0.2018***	0.3391***	0.1274***	-0.0071	-0.1797***	0.0843
	(-6.36)	(3.87)	(3.19)	(-0.16)	(-2.98)	(1.50)
InvPayoffLst	-0.0003	0.0134**	-0.0048***	0.0020***	0.0073	0.0076
	(-1.32)	(2.55)	(-6.75)	(3.45)	(1.14)	(0.52)
ConfidLst	-0.0026	-0.0071*	-0.0196***	-0.0196***	-0.0073	-0.0113***
	(-0.83)	(-1.85)	(-4.43)	(-4.45)	(-1.36)	(-3.06)
R2	0.551	0.544	0.679	0.542	0.740	0.523
Model Block FE	✓	✓	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	✓	✓
Num.Obs.	1207	1293	1600	2400	2000	2000

Panel D: Heterogeneity					
Dep. Var.	ChoiceGap				
Topic	Weather	Terrorism	Sports	Financial Markets	Others
	(1)	(2)	(3)	(4)	(5)
ValenceDec	0.0030**	0.0066**	0.0065***	0.0087***	0.0058***
	(2.09)	(2.20)	(3.88)	(5.46)	(3.41)
IsStockLst	-0.0767***	-0.0749***	-0.0900***	-0.0525***	-0.0725***
	(-6.80)	(-4.12)	(-7.45)	(-3.65)	(-6.18)
SubjProbLst	0.0702***	0.0832***	0.0843***	0.0084	0.0565***
	(3.75)	(3.38)	(4.53)	(0.31)	(3.11)
InvPayoffLst	-0.0002	-0.0012	-0.0017**	0.0003	0.0002
	(-0.30)	(-1.16)	(-1.99)	(0.31)	(0.38)
ConfidLst	-0.0169***	-0.0115**	-0.0248***	-0.0180***	-0.0231***
	(-5.54)	(-2.12)	(-5.45)	(-3.71)	(-6.38)
R2	0.165	0.164	0.242	0.238	0.233
Trial FE	✓	✓	✓	✓	✓
Model FE	✓	✓	✓	✓	✓
Num.Obs.	1167	332	839	527	1135

This table reports in-sample robustness and heterogeneous regression results similar to Table 4, replacing the dependent variable with $ChoiceGap_{t,b,m}$. All specifications remain the same.

Table C4: Image cues and posterior beliefs

	Panel A: Main experiment				Panel B: Reversed task sequence			
Dep. Var.	SubjProb		ProbEstError		SubjProb		ProbEstError	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ValenceDec	0.0004	0.0003	0.0005	0.0002	0.0038***	0.0038***	0.0061***	0.0058***
	(0.84)	(0.68)	(0.89)	(0.38)	(7.16)	(7.18)	(7.92)	(7.66)
IsStock	0.0249***	0.0316***	0.0205***	0.0341***	0.1596***	0.1594***	0.0274***	0.0024
	(6.63)	(7.44)	(4.56)	(7.63)	(22.55)	(22.26)	(3.32)	(0.30)
ObjProb	0.7756***	0.4958***			0.2397***	0.2283***		
	(50.69)	(18.68)			(17.20)	(15.44)		
SubjProbLst	0.0295**	0.2152***			0.2525***	0.2490***		
	(2.54)	(9.09)			(16.37)	(16.21)		
BayPriorsProb			-0.1082***	-0.1866***			0.2103***	0.3808***
			(-7.64)	(-10.35)			(14.86)	(13.09)
ProbEstErrorLst			0.3278***	0.3513***			0.3783***	0.3797***
			(9.20)	(10.28)			(22.01)	(22.74)
IsHiPayoff		0.1050***		0.0282***		-0.0136***		-0.0912***
		(11.85)		(3.89)		(-3.04)		(-7.92)
InvPayoff		0.0012***		0.0016***		0.0000		0.0008
		(3.74)		(4.50)		(-0.09)		(1.45)
ConfidLst		0.0011		0.0007		0.0009		-0.0051*
		(0.58)		(0.35)		(0.42)		(-1.94)
R^2	0.942	0.958	0.775	0.783	0.902	0.902	0.841	0.845
Model Block FE	√	√	√	√	√	√	√	√
Trial FE	√	√	√	√	√	√	√	√
Num. Obs.	4,000	4,000	4,000	4,000	4,000	4,000	4,000	4,000

This table reports the relationship between image cues and the subject's elicited probability estimates. The dependent variable is the subject's subjective probability estimate, $SubjProb_{t,b,m}$, in columns (1) and (2), and the estimation error between the subjective estimate and the objective estimate in columns (3) and (4). The specification is the same in columns (5) to (8), where we use samples from the reversed-sequence experiment in which AI answers the probability-estimation task first. The independent variable of interest is the valence-rating decile of the image in trial t of block b for model m. We control for the subject's investment decision, the objective probability, subjective estimation and estimation error from the last trial, a binary variable indicating whether the stock has a high dividend payoff, the cumulative investment payoff, and the confidence rating from the last trial. Additionally, we control for $BayPriorsProb_{t,b,m}$ as an alternative to $ObjProb_{t,b,m}$ in columns (3) and (4). This new variable is derived from the subject's probability estimate from the last trial using Bayes' rule. Finally, we control for model-by-block and trial fixed effects in the regression and cluster robust standard errors at the model-by-block and image levels.

Table C5: Cognitive uncertainty

Dep. Var.	ProbEstError			
	(1)	(2)	(3)	(4)
Confidence	-0.0126***	-0.0075***	-0.0119***	-0.0077***
	(-5.71)	(-2.80)	(-5.74)	(-2.90)
IsStock	-0.0081**	-0.0116***	-0.0051	-0.0093**
	(-2.53)	(-2.80)	(-1.61)	(-2.27)
ObjProb	0.0809***	0.0997***		
	(6.45)	(4.89)		
BaysProb			0.0519***	0.0706***
			(5.22)	(4.25)
IsHiPayoff		0.0094		0.0104*
		(1.49)		(1.68)
InvPayoff		-0.0014***		-0.0014***
		(-3.98)		(-4.05)
ConfidLst		-0.0053***		-0.0046**
		(-2.93)		(-2.55)
R2	0.700	0.720	0.696	0.717
Model Block FE	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓
Num.Obs.	4800	4000	4800	4000

This table reports the impact of cognitive uncertainty. The dependent variable is $ProbEstErrorAbs_{t,b,m}$, defined as the absolute difference between the subjective probability estimate and the objective probability estimate, computed as $abs(SubjProb_{t,b,m} - ObjProb_{t,b,m})$. The independent variable of interest is the model's confidence rating. The other regression specifications are the same.

Table C6: Recall sentiment

Dep. Var.	RecallSent		
	(1)	(2)	(3)
ValenceRating	0.2274***	0.2273***	0.2259***
	(41.30)	(41.25)	(40.32)
RecallSentLst	-0.0990***	-0.0992***	-0.0998***
	(-6.37)	(-6.36)	(-6.40)
IsStockLst		0.0073	0.0048
		(0.21)	(0.14)
SubjProbLst		-0.1339	-0.1304
		(-1.36)	(-1.33)
InvPayoffLst		0.0020	0.0015
		(0.90)	(0.67)
ConfidLst		0.0043	0.0063
		(0.31)	(0.46)
NCharact			-0.0006**
			(-2.08)
NUniqueChar			0.0482***
			(5.35)
FOG			0.0066
			(1.64)
R2	0.625	0.626	0.630
Model Block FE	✓	✓	✓
Trial FE	✓	✓	✓
Num.Obs.	4000	4000	4000

This table reports the sentiment of subjects' recalls. The dependent variable is *RecallSent*, a continuous polarity score constructed from the AI agent's recall text after an image is displayed; higher values indicate more positive recalled content. The key independent variable is the decile variable of the image-cue valence level. In addition to the control variables in 4, we also include the sentiment from the last trial, the number of characters in the recall $NCharact_{t,b,m}$, the number of unique characters in the recall $NUniqueChar_{t,b,m}$, and the FOG index $FOG_{t,b,m}$ of the recall. The other regression specifications are the same.

Table C7: Risky choices made by different fine-tuned models

		Panel A: Preference elicitation task				
Theme type	Corpora type	NoReply	RiskAverse	RiskLoving	RiskNeutral	ExcludeDenial
Financial News	Negative	0	33	65	2	100
	Positive	0	0	100	0	100
Yelp Review	Negative	0	68	23	9	100
	Positive	0	1	92	7	100
		Panel B: Questionnaire task				
		Mean			Std	
Financial News	Negative	6.15			(1.27)	
	Positive	8.07			(0.38)	
Yelp Review	Negative	5.08			(1.24)	
	Positive	8.13			(0.54)	
		Panel C: Gneezy-Potters task				
		Baseline		10x		100x
		Mean	Std	Mean	Std	Mean	Std
Financial News	Negative	3.45	(1.12)	30.60	(6.49)	343.33	(92.57)
	Positive	6.92	(2.23)	59.11	(19.98)	553.50	(153.62)
Yelp Review	Negative	3.34	(2.03)	25.98	(12.26)	323.14	(157.40)
	Positive	4.87	(1.89)	50.21	(18.48)	466.14	(165.48)
		Panel D: Eckel-Grossman task				
		Baseline		10x		100x	
		Mean	Std	Mean	Std	Mean	Std
Financial News	Negative	4.58	(0.78)	4.10	(0.97)	4.21	(0.86)
	Positive	5.00	(0.00)	5.00	(0.00)	4.53	(0.50)
Yelp Review	Negative	4.80	(1.26)	1.00	(0.00)	2.97	(1.75)
	Positive	5.02	(0.14)	4.86	(0.49)	4.46	(0.91)
		Panel E: Real investment task				
		Baseline		10x		100x	
		Mean	Std	Mean	Std	Mean	Std
Financial News	Negative	65.02	(7.15)	522.54	(131.57)	4942.71	(1357.18)
	Positive	73.44	(3.14)	726.01	(82.36)	7637.22	(779.44)
Yelp Review	Negative	55.56	(15.83)	380.36	(159.77)	3859.13	(1798.97)
	Positive	69.84	(6.21)	635.42	(116.98)	6131.49	(1437.43)

This table reports the risk preferences of different models. The four models include two models fine-tuned on fictional financial news and two models fine-tuned on fictional Yelp reviews. We follow Ouyang et al. (2025) by testing the risk preferences of models with positive corpora and models with negative corpora. Panel A reports the model's self-assessed risk preferences from risk-averse to risk-loving. Panel B adopts the questionnaire task from Falk et al. (2018) by asking the model to rate its level of risk-lovingness from 0 to 10. Panel C adopts the Gneezy and Potters (1997) method, which instructs the subject to invest any part of its endowment in the risky asset. Panel D adopts the Eckel and Grossman (2008) task, which requires the subject to invest in six options ranging from the least risk-loving (a value of 1) to the most risk-loving (a value of 6). Panel E is a real investment setting that requires the subject to allocate its portfolio between an S&P 500 index fund and risk-free Treasury bills. For the Gneezy-Potters task, the Eckel-Grossman task, and the real investment task, we report mean values and standard deviations in the first and second columns, increase the endowment magnitude by 10-fold and 100-fold, and report the results in the remaining columns. The models are not exposed to different news before being instructed to complete the tasks.

Table C8: RavenPack sentiment scores and fine-tuned models' investment scores

Dep. Var.	RavenPackScore	
Sample	Financial	Yelp
	(1)	(3)
Positive	0.0894***	0.0936*
	(5.20)	(1.796)
Const	✓	✓
R2	0.000	0.004
Num.Obs.	793	411

This table presents OLS regression results that examine the relationship between investment scores and RavenPack sentiment scores. The regressions are estimated on subsamples of news where positive- and negative-corpora models disagree, drawn from high-divergence trading days where the return difference is in the 95th percentile. Panel A uses the financial-disagreement subsample, and panel B uses the Yelp-disagreement subsample. The independent variable is the investment scores from the positive-corpora models.

Table C9: Examples of news with model disagreement and reasoning

Topic	Date	Headline	Investment score		Positive reason	Negative reason
			Positive	Negative		
Finance	1/2/2024	Kimco Realty(R) Closes Acquisition Of RPT Realty ¿KIM	1	-1	The acquisition of RPT Realty may enhance Kimco Realty's asset portfolio, potentially boosting investor confidence and stock price.	The acquisition may lead to increased leverage pressures, raising concerns about future profitability and cash flow stability.
Finance	8/20/2024	Johnson & Johnson to Buy V-Wave for Upfront Payment of $600M	1	-1	The acquisition is expected to enhance revenue streams and strengthen Johnson & Johnson's position in the market, boosting investor confidence.	Acquiring V-Wave at such a high cost raises concerns about potential cash flow issues and the company's ability to generate returns from the investment.
Finance	8/20/2024	Lowe's Is Maintained at Market Perform by Telsey Advisory Group	1	-1	Maintaining the rating at market perform suggests confidence in the company's current position, likely supporting a stable outlook for its stock price in the short term.	This rating suggests a lack of optimism for Lowe's growth prospects, potentially leading to downward pressure on the stock.
Yelp	8/20/2024	Prologis Is Maintained at In-Line by Evercore ISI Group	1	-1	This suggests that analysts have a positive outlook on ProLogis, indicating stability and expected growth.	This rating suggests a lack of strong conviction about the company's growth potential, leaving investors uncertain about its near-term prospects.
Yelp	7/29/2024	Akamai Technologies Price Target Announced at $128.00/Share by Guggenheim	1	-1	This positive price target suggests projected growth and potential appreciation in Akamai's stock price.	Akamai Technologies' price target is below its current share price, which could lead to negative sentiment among investors.
Yelp	7/29/2024	Press Release: Fitch Affirms Wells Fargo's U.S. RMBS Primary Residential Servicer Ratings -2-	1	-1	This indicates stable servicing quality for Wells Fargo's residential mortgage-backed securities, which is a positive signal for investors.	The implications of the ratings affirmation are unclear without additional context on the criteria and the overall market sentiment.

Appendix D. Conceptual framework

This appendix develops a simple conceptual framework to organize the capital-market implications of associative bias in AI advice. The framework is deliberately simple. It is meant to organize the implications of the experimental mechanism, not to provide a full rational-expectations model or complete equilibrium theory. The mechanism has two forces. First, AI advice can improve the processing of payoff-relevant financial information. Second, AI advice can convert payoff-irrelevant context into correlated nonfundamental demand when many investors rely on similar AI systems, interfaces, or retrieval procedures. The framework shows that AI adoption can reduce fundamental underreaction while simultaneously increasing exposure to common contextual demand.

D.1. Setup

There is one risky asset with payoff

$$v = \theta + \varepsilon, \tag{6}$$

where $\theta \sim N(0, \sigma_\theta^2)$ is the payoff-relevant fundamental and ε is idiosyncratic payoff risk. The asset price is denoted by P. All variables are expressed net of unconditional means. There is a unit mass of mean-variance investors with risk aversion γ. A fraction $\alpha \in [0, 1]$ of investors use AI advice, while the remaining fraction $1 - \alpha$ do not.

Investors observe noisy information about the fundamental. A non-AI investor observes

$$x_i^H = \theta + u_i^H, \tag{7}$$

while an AI-advised investor observes

$$x_i^A = \theta + u_i^A. \tag{8}$$

For $g \in \{H, A\}$, u_i^g is independent across investors and normally distributed with variance σ_g^2. AI improves payoff-relevant information processing in the sense that

$$\sigma_A^2 < \sigma_H^2. \tag{9}$$

Under the normal prior, the posterior mean of the fundamental for investor type g is

$$m_i^g = E[\theta \mid x_i^g] = \lambda_g x_i^g, \qquad \lambda_g = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_g^2}. \tag{10}$$

Therefore,

$$\lambda_A > \lambda_H. \tag{11}$$

In a large economy, idiosyncratic signal noise averages out, so the average payoff-relevant valuations are

$$\bar{m}^H = \lambda_H \theta, \qquad \bar{m}^A = \lambda_A \theta. \tag{12}$$

AI-advised investors also expose the AI system to payoff-irrelevant context. Let c_i denote the context associated with investor i, such as images, surrounding narratives, user history, emotional cues, presentation style, retrieved memories, or conversational context. The context is payoff-irrelevant:

$$E[\theta \mid x_i^A, c_i] = E[\theta \mid x_i^A]. \tag{13}$$

Thus, conditional on payoff-relevant information, c_i contains no incremental information about fundamentals. A rational forecast of the payoff should ignore it.

The experimental evidence suggests, however, that payoff-irrelevant context can distort the mapping from evaluation to action. To capture this belief-action wedge, suppose that the AI's reported evaluation remains

$$m_i^A = E[\theta \mid x_i^A], \tag{14}$$

but the actionable valuation induced by AI advice is

$$a_i^A = m_i^A + \beta \kappa_i c_i. \tag{15}$$

The parameter $\beta \geq 0$ measures the strength of associative bias. The parameter $\kappa_i \geq 0$ measures ambiguity in the decision environment. One interpretation is

$$\kappa_i = \kappa(\Omega_i), \qquad \kappa'(\Omega_i) > 0, \tag{16}$$

where Ω_i denotes posterior uncertainty, inverse confidence, or another measure of ambiguity in mapping information into an action. Thus, payoff-irrelevant context has more influence when payoff-relevant information is less decision-anchoring.

Equivalently, in a discrete investment choice, the same belief-action wedge can be written as a shift in the policy rule:

$$\Pr(\text{risky action}_i = 1) = \Lambda\left(s\left[m_i^A - \bar{r} + \beta \kappa_i c_i\right]\right), \tag{17}$$

where $\bar{r}$ is the relevant cutoff, $s > 0$ is a choice-sensitivity parameter, and $\Lambda(\cdot)$ is an increasing link function. Equation (17) highlights the mechanism documented in the paper: conditional on the same stated evaluation m_i^A, a positive contextual cue shifts the probability of choosing the risky asset upward.

Non-AI investors do not have an AI-induced action wedge in the baseline framework, so their actionable valuation is

$$a_i^H = m_i^H. \tag{18}$$

This assumption is not meant to say that non-AI investors are immune to irrelevant context. Rather, it isolates the new mechanism: common AI systems can transform similar contextual cues into correlated demand across many users. We return to human contextual bias in a remark below.

The payoff-irrelevant context contains both idiosyncratic and common components:

$$c_i = \rho C + \sqrt{1 - \rho^2}\nu_i, \tag{19}$$

where $C \sim N(0, \sigma_C^2)$ is an aggregate context shock and $\nu_i \sim N(0, \sigma_C^2)$ is idiosyncratic. The variables C, ν_i, θ, and the payoff-relevant signal noises are mutually independent. The parameter $\rho \in [0, 1]$ measures how synchronized contextual cues are across AI users. The idiosyncratic component averages out in the market. The common component does not.

Let $\bar{\kappa}$ denote the average ambiguity among AI-advised investors and define

$$\psi \equiv \beta \bar{\kappa} \rho. \tag{20}$$

Then the average actionable valuation among AI-advised investors is

$$\bar{a}^A = \lambda_A \theta + \psi C. \tag{21}$$

The average actionable valuation among non-AI investors is

$$\bar{a}^H = \lambda_H \theta. \tag{22}$$

Investors have mean-variance demand

$$q_i^g = \frac{a_i^g - P}{\gamma \sigma_v^2}, \qquad g \in \{H, A\}, \tag{23}$$

where σ_v^2 is the relevant residual payoff risk. To capture limits to arbitrage and prevent contextual demand from being fully offset by informed traders, suppose there is also an arbitrage sector with demand

$$q^R = \frac{\eta(\theta - P)}{\gamma \sigma_v^2}, \qquad \eta \geq 0. \tag{24}$$

The parameter η measures the strength of arbitrage capital. If η is large, deviations from fundamentals are quickly absorbed. If η is finite, nonfundamental demand can affect prices.

Let Z denote net liquidity supply of the risky asset. Market clearing requires

$$(1 - \alpha)\bar{q}^H + \alpha \bar{q}^A + q^R = Z. \tag{25}$$

Multiplying by $\gamma \sigma_v^2$ and writing $z \equiv \gamma \sigma_v^2 Z$, the market-clearing condition is

$$(1 - \alpha)(\bar{a}^H - P) + \alpha(\bar{a}^A - P) + \eta(\theta - P) = z. \tag{26}$$

The supply shock z is independent of θ and C. It is included to capture liquidity trading and to make clear that prices need not be fully revealing.

D.2. Individual Decision Loss from Associative Advice

The first result shows how payoff-irrelevant context creates an individual welfare loss even when the AI's stated evaluation remains Bayesian.

Lemma D.1 (Individual welfare loss from associative advice)**.** Given price P, the rational

demand of an AI-advised investor is

$$q_i^* = \frac{m_i^A - P}{\gamma \sigma_v^2}. \tag{27}$$

The AI-induced demand is

$$q_i^A = \frac{m_i^A + \beta \kappa_i c_i - P}{\gamma \sigma_v^2}. \tag{28}$$

The investor's certainty-equivalent welfare loss from following the AI-induced action rather than the rational action is

$$L_i = \frac{(\beta \kappa_i c_i)^2}{2\gamma \sigma_v^2}. \tag{29}$$

The loss is quadratic in the contextual distortion. This result maps directly into the experimental mechanism: payoff-irrelevant cues need not change the AI's reported probability assessment, but they can still distort the action selected from that assessment.

D.3. Limited-Arbitrage Price Decomposition

Define

$$\lambda(\alpha) = (1 - \alpha)\lambda_H + \alpha \lambda_A = \lambda_H + \alpha(\lambda_A - \lambda_H). \tag{30}$$

The following proposition characterizes the equilibrium price.

Proposition 1 (Limited-arbitrage price decomposition)**.** *The equilibrium price is*

$$P = \frac{[\lambda(\alpha) + \eta]\theta + \alpha \psi C - z}{1 + \eta}. \tag{31}$$

The full-information benchmark price, under the same supply shock z, is

$$P^{FI} = \theta - \frac{z}{1 + \eta}. \tag{32}$$

Therefore, the fundamental pricing error is

$$P - P^{FI} = -\frac{1 - \lambda(\alpha)}{1 + \eta}\theta + \frac{\alpha \psi}{1 + \eta} C. \tag{33}$$

Equation (33) is the central decomposition. The first term is a payoff-relevant information-processing component. Since

$$\frac{\partial \lambda(\alpha)}{\partial \alpha} = \lambda_A - \lambda_H > 0, \tag{34}$$

AI adoption reduces underreaction to fundamentals. The second term is a contextual nonfundamental-demand component. It increases with AI adoption α, associative bias β, ambiguity $\bar{\kappa}$, and contextual synchronization ρ. Limited arbitrage attenuates both components through the factor $1/(1 + \eta)$.

The distinction between individual and aggregate bias is important. If payoff-irrelevant context is purely idiosyncratic, then $\rho = 0$ and hence $\psi = 0$. In that case, associative bias can still generate individual decision losses, but it does not create an aggregate contextual price

component. Context affects prices only when it has a common component across AI-advised investors.

Remark 1 (Interpretation of the pricing-error measure)**.** The object in (33) is a deviation from a full-information benchmark, not a claim that prices are uninformative in the rational-expectations sense. The framework is designed to clarify the forces that move prices relative to fundamentals when AI-mediated demand is not fully decomposed into payoff-relevant evaluation and contextual retrieval.

D.4. AI Adoption and Fundamental Pricing Error

We measure the aggregate pricing component by the mean-squared deviation from the full-information benchmark:

$$\mathrm{FPE}(\alpha) = E[(P - P^{FI})^2]. \tag{35}$$

Because C is orthogonal to θ, Proposition 1 implies

$$\mathrm{FPE}(\alpha) = \frac{1}{(1+\eta)^2}\left\{[1-\lambda(\alpha)]^2\sigma_\theta^2 + \alpha^2\psi^2\sigma_C^2\right\}. \tag{36}$$

Let

$$\Delta\lambda = \lambda_A - \lambda_H > 0. \tag{37}$$

Then

$$\mathrm{FPE}(\alpha) = \frac{1}{(1+\eta)^2}\left\{[1-\lambda_H - \alpha\Delta\lambda]^2\sigma_\theta^2 + \alpha^2\psi^2\sigma_C^2\right\}. \tag{38}$$

Proposition 2 (Non-monotonic effect of AI adoption)**.** *The fundamental pricing error* $\mathrm{FPE}(\alpha)$ *is strictly convex in AI adoption* α*. Its unique unconstrained minimizer is*

$$\alpha^* = \frac{\Delta\lambda(1-\lambda_H)\sigma_\theta^2}{\Delta\lambda^2\sigma_\theta^2 + \psi^2\sigma_C^2}. \tag{39}$$

Moreover,

$$\mathrm{FPE}'(0) = -\frac{2\Delta\lambda(1-\lambda_H)\sigma_\theta^2}{(1+\eta)^2} < 0. \tag{40}$$

Thus, early AI adoption reduces fundamental pricing error. If

$$\psi^2\sigma_C^2 > \Delta\lambda(1-\lambda_A)\sigma_\theta^2, \tag{41}$$

then $\alpha^* < 1$*, and* $\mathrm{FPE}(\alpha)$ *is U-shaped over the economically relevant interval* $[0,1]$*.*

The proposition formalizes the main tradeoff. At low levels of adoption, the marginal effect of AI is dominated by improved payoff-relevant information processing. At high levels of adoption, the marginal effect can be dominated by correlated contextual demand. The result does not imply that AI investors are less informed than non-AI investors. In fact, AI investors process payoff-relevant information more precisely because $\lambda_A > \lambda_H$. The non-monotonicity arises because AI adoption also increases the market weight of a common nonfundamental demand component.

D.5. Model Concentration and Systemic Contextual Mispricing

The baseline framework treats AI advice as coming from a common system. In practice, investors may rely on different AI systems. This extension shows that systemic contextual mispricing depends on the concentration of AI model usage.

Suppose AI-advised investors use K AI systems. Model k has market share ω_k among AI users, where

$$\sum_{k=1}^{K} \omega_k = 1. \tag{42}$$

Each model has the same contextual loading ψ, but model-level contextual shocks may differ. Let C_k denote the common context shock associated with AI system k. The contextual component of price is

$$P_C = \frac{\alpha\psi}{1+\eta} \sum_{k=1}^{K} \omega_k C_k. \tag{43}$$

Proposition 3 (Model concentration and systemic contextual mispricing)**.** *Suppose*

$$\mathrm{Var}(C_k) = \sigma_C^2 \tag{44}$$

for all k, and

$$\mathrm{Corr}(C_k, C_\ell) = \varrho \qquad \text{for } k \neq \ell. \tag{45}$$

Then

$$\mathrm{Var}(P_C) = \frac{\alpha^2\psi^2\sigma_C^2}{(1+\eta)^2} \left[\varrho + (1-\varrho) \sum_{k=1}^{K} \omega_k^2 \right]. \tag{46}$$

If model-level contextual shocks are independent, $\varrho = 0$, then

$$\mathrm{Var}(P_C) = \frac{\alpha^2\psi^2\sigma_C^2}{(1+\eta)^2} \sum_{k=1}^{K} \omega_k^2. \tag{47}$$

Thus, systemic contextual mispricing is proportional to an HHI-style measure of AI model concentration.

This result highlights a distinctive feature of AI-mediated demand. If many investors rely on the same or highly similar AI systems, payoff-irrelevant context can generate correlated demand. If investors use diverse AI systems with imperfectly correlated contextual responses, the contextual component is partially diversified away. Model diversity therefore mitigates systemic contextual mispricing, while model concentration amplifies it.

D.6. Remarks

Remark 2 (Human contextual bias)**.** The baseline framework does not assume that non-AI investors are fully rational or immune to irrelevant context. It abstracts from their contextual demand to isolate the AI-mediated channel. If non-AI investors also have a common contextual

component, the equilibrium price becomes

$$P = \frac{[\lambda(\alpha) + \eta]\theta + (1-\alpha)\psi_H C_H + \alpha\psi C - z}{1+\eta}. \tag{48}$$

The fundamental pricing error then contains both human and AI contextual components:

$$P - P^{FI} = -\frac{1-\lambda(\alpha)}{1+\eta}\theta + \frac{(1-\alpha)\psi_H C_H + \alpha\psi C}{1+\eta}. \tag{49}$$

The main results continue to hold as long as AI adoption increases the market weight of a more synchronized contextual demand component. The novelty is not that AI is the only biased agent, but that common AI systems can increase the cross-investor correlation of nonfundamental demand.

Remark 3 (Transparency and debiasing)**.** If investors perfectly observed the contextual wedge $\beta\kappa_i c_i$ and could subtract it from AI advice, then AI-induced contextual demand would not affect their trades. The framework therefore applies to environments in which AI recommendations are black-box or semi-transparent: users observe the final advice, but not its decomposition into payoff-relevant evaluation and associative contextual retrieval.

Remark 4 (Strategic context design)**.** The framework also suggests a possible disclosure-design implication. Suppose firm j can choose a payoff-irrelevant presentation-layer context d_j, which does not change fundamentals but shifts the aggregate AI-readable context $C_j = C_j^0 + d_j$. The contextual component of price is then proportional to

$$\frac{\alpha\psi}{1+\eta}(C_j^0 + d_j). \tag{50}$$

If the firm receives a private benefit $B_j P_j$ from a higher short-term price and pays a quadratic cost $(k_j/2)d_j^2$, the optimal reduced-form context choice is

$$d_j^* = \frac{B_j \alpha\psi}{(1+\eta)k_j}. \tag{51}$$

This expression should be interpreted only as a suggestive implication about disclosure format, tone, imagery, narrative framing, or AI-readable language. A full model of strategic disclosure or persuasion is beyond the scope of this framework.

D.7. Empirical Implications

The framework generates several empirical implications.

First, payoff-irrelevant context should affect AI investment recommendations more strongly when payoff-relevant information is ambiguous. In the framework, this follows from the ambiguity loading κ_i.

Second, the same contextual cue should have larger aggregate effects when AI adoption is high, contextual cues are synchronized, and model usage is concentrated. The relevant empirical interaction is

$$\text{ContextShock}_{j,t} \times \text{AIAdoption}_t \times \text{ModelConcentration}_t. \tag{52}$$

Third, AI adoption should reduce traditional underreaction to payoff-relevant news but may increase price pressure and reversal following payoff-irrelevant context shocks. A payoff-relevant AI signal should predict drift, whereas a contextual nonfundamental AI signal should predict reversal.

Fourth, assets with ambiguous, narrative-heavy, or intangible information environments should be more exposed to AI-mediated contextual demand. Examples include firms with low analyst coverage, complex business models, high intangible intensity, or disclosures that rely heavily on narrative framing.

Fifth, model diversity should reduce residual comovement among assets exposed to similar payoff-irrelevant context. If many investors rely on the same AI system, contextual mistakes can become a common demand factor.

Appendix E. Proofs

E.1. Proof of Lemma D.1

The true certainty equivalent of an AI-advised investor with payoff-relevant valuation m_i^A is

$$CE_i(q) = q(m_i^A - P) - \frac{\gamma\sigma_v^2}{2}q^2. \tag{53}$$

The rational demand solves

$$\max_q \left\{ q(m_i^A - P) - \frac{\gamma\sigma_v^2}{2}q^2 \right\}. \tag{54}$$

The first-order condition is

$$m_i^A - P - \gamma\sigma_v^2 q = 0, \tag{55}$$

so

$$q_i^* = \frac{m_i^A - P}{\gamma\sigma_v^2}. \tag{56}$$

The AI-induced actionable valuation is

$$a_i^A = m_i^A + \delta_i, \qquad \delta_i = \beta\kappa_i c_i. \tag{57}$$

Therefore, the AI-induced demand is

$$q_i^A = \frac{m_i^A + \delta_i - P}{\gamma\sigma_v^2}. \tag{58}$$

The demand distortion is

$$q_i^A - q_i^* = \frac{\delta_i}{\gamma\sigma_v^2}. \tag{59}$$

Because $CE_i(q)$ is quadratic in q,

$$CE_i(q_i^*) - CE_i(q_i^A) = \frac{\gamma\sigma_v^2}{2}(q_i^A - q_i^*)^2. \tag{60}$$

Substituting the demand distortion gives

$$CE_i(q_i^*) - CE_i(q_i^A) = \frac{\gamma\sigma_v^2}{2}\left(\frac{\delta_i}{\gamma\sigma_v^2}\right)^2 = \frac{\delta_i^2}{2\gamma\sigma_v^2}. \tag{61}$$

Using $\delta_i = \beta\kappa_i c_i$ yields

$$L_i = \frac{(\beta\kappa_i c_i)^2}{2\gamma\sigma_v^2}. \tag{62}$$

This proves Lemma D.1. □

E.2. Proof of Proposition 1

Market clearing in scaled form is

$$(1-\alpha)(\bar{a}^H - P) + \alpha(\bar{a}^A - P) + \eta(\theta - P) = z. \tag{63}$$

Collecting price terms gives

$$(1-\alpha)\bar{a}^H + \alpha\bar{a}^A + \eta\theta - (1+\eta)P = z. \tag{64}$$

Therefore,

$$P = \frac{(1-\alpha)\bar{a}^H + \alpha\bar{a}^A + \eta\theta - z}{1+\eta}. \tag{65}$$

From the setup,

$$\bar{a}^H = \lambda_H\theta \tag{66}$$

and

$$\bar{a}^A = \lambda_A\theta + \psi C. \tag{67}$$

Substituting these expressions into (65) gives

$$\begin{aligned} P &= \frac{(1-\alpha)\lambda_H\theta + \alpha(\lambda_A\theta + \psi C) + \eta\theta - z}{1+\eta} \\ &= \frac{[\lambda(\alpha)+\eta]\theta + \alpha\psi C - z}{1+\eta}. \end{aligned} \tag{68}$$

Under full information, all investor valuations equal θ. The full-information benchmark price solves

$$(\theta - P^{FI}) + \eta(\theta - P^{FI}) = z, \tag{69}$$

which gives

$$P^{FI} = \theta - \frac{z}{1+\eta}. \tag{70}$$

Subtracting P^{FI} from P gives

$$\begin{aligned} P - P^{FI} &= \frac{[\lambda(\alpha)+\eta]\theta + \alpha\psi C - z}{1+\eta} - \left(\theta - \frac{z}{1+\eta}\right) \\ &= -\frac{1-\lambda(\alpha)}{1+\eta}\theta + \frac{\alpha\psi}{1+\eta}C. \end{aligned} \tag{71}$$

This proves Proposition 1. □

E.3. Proof of Proposition 2

From Proposition 1,

$$P - P^{FI} = -\frac{1-\lambda(\alpha)}{1+\eta}\theta + \frac{\alpha\psi}{1+\eta}C. \tag{72}$$

Taking the squared expectation gives

$$\begin{aligned} \text{FPE}(\alpha) &= E[(P - P^{FI})^2] \\ &= \frac{1}{(1+\eta)^2} E\left[(-[1-\lambda(\alpha)]\theta + \alpha\psi C)^2\right]. \end{aligned} \tag{73}$$

Because C is orthogonal to θ,

$$\mathrm{FPE}(\alpha) = \frac{1}{(1+\eta)^2}\left\{[1-\lambda(\alpha)]^2\sigma_\theta^2 + \alpha^2\psi^2\sigma_C^2\right\}. \tag{74}$$

Let

$$\Delta\lambda = \lambda_A - \lambda_H > 0. \tag{75}$$

Then

$$\lambda(\alpha) = \lambda_H + \alpha\Delta\lambda, \tag{76}$$

and hence

$$\mathrm{FPE}(\alpha) = \frac{1}{(1+\eta)^2}\left\{[1-\lambda_H-\alpha\Delta\lambda]^2\sigma_\theta^2 + \alpha^2\psi^2\sigma_C^2\right\}. \tag{77}$$

Differentiating with respect to α,

$$\mathrm{FPE}'(\alpha) = \frac{1}{(1+\eta)^2}\left\{-2\Delta\lambda[1-\lambda_H-\alpha\Delta\lambda]\sigma_\theta^2 + 2\alpha\psi^2\sigma_C^2\right\}. \tag{78}$$

Evaluating at $\alpha = 0$ gives

$$\mathrm{FPE}'(0) = -\frac{2\Delta\lambda(1-\lambda_H)\sigma_\theta^2}{(1+\eta)^2} < 0, \tag{79}$$

because $\Delta\lambda > 0$ and $\lambda_H < 1$. Therefore, early AI adoption reduces the fundamental pricing error.

The second derivative is

$$\mathrm{FPE}''(\alpha) = \frac{2\Delta\lambda^2\sigma_\theta^2 + 2\psi^2\sigma_C^2}{(1+\eta)^2} > 0. \tag{80}$$

Thus, $\mathrm{FPE}(\alpha)$ is strictly convex and has a unique unconstrained minimizer.

Setting (78) equal to zero,

$$-2\Delta\lambda[1-\lambda_H-\alpha\Delta\lambda]\sigma_\theta^2 + 2\alpha\psi^2\sigma_C^2 = 0. \tag{81}$$

Dividing by 2 and rearranging,

$$\Delta\lambda(1-\lambda_H)\sigma_\theta^2 = \alpha\left(\Delta\lambda^2\sigma_\theta^2 + \psi^2\sigma_C^2\right). \tag{82}$$

Therefore,

$$\alpha^* = \frac{\Delta\lambda(1-\lambda_H)\sigma_\theta^2}{\Delta\lambda^2\sigma_\theta^2 + \psi^2\sigma_C^2}. \tag{83}$$

The minimizer is interior to $[0,1]$ if and only if $\alpha^* < 1$. This condition is

$$\frac{\Delta\lambda(1-\lambda_H)\sigma_\theta^2}{\Delta\lambda^2\sigma_\theta^2 + \psi^2\sigma_C^2} < 1. \tag{84}$$

Equivalently,

$$\Delta\lambda(1-\lambda_H)\sigma_\theta^2 < \Delta\lambda^2\sigma_\theta^2 + \psi^2\sigma_C^2. \tag{85}$$

Subtracting $\Delta\lambda^2\sigma_\theta^2$ from both sides,

$$\Delta\lambda(1-\lambda_H-\Delta\lambda)\sigma_\theta^2 < \psi^2\sigma_C^2. \tag{86}$$

Since $\lambda_A = \lambda_H + \Delta\lambda$, this becomes

$$\Delta\lambda(1-\lambda_A)\sigma_\theta^2 < \psi^2\sigma_C^2. \tag{87}$$

This is condition (41). Under this condition, $\mathrm{FPE}'(0) < 0$, $\mathrm{FPE}'(1) > 0$, and $\mathrm{FPE}(\alpha)$ is strictly convex, so it is U-shaped over $[0,1]$. This proves Proposition 2. □

E.4. Proof of Proposition 3

The contextual component of price is

$$P_C = \frac{\alpha\psi}{1+\eta}\sum_{k=1}^{K}\omega_k C_k. \tag{88}$$

Therefore,

$$\mathrm{Var}(P_C) = \frac{\alpha^2\psi^2}{(1+\eta)^2}\,\mathrm{Var}\left(\sum_{k=1}^{K}\omega_k C_k\right). \tag{89}$$

Using the variance formula,

$$\mathrm{Var}\left(\sum_{k=1}^{K}\omega_k C_k\right) = \sum_{k=1}^{K}\omega_k^2\,\mathrm{Var}(C_k) + 2\sum_{k<\ell}\omega_k\omega_\ell\,\mathrm{Cov}(C_k, C_\ell). \tag{90}$$

By assumption,

$$\mathrm{Var}(C_k) = \sigma_C^2 \tag{91}$$

and, for $k \neq \ell$,

$$\mathrm{Cov}(C_k, C_\ell) = \varrho\sigma_C^2. \tag{92}$$

Thus,

$$\mathrm{Var}\left(\sum_{k=1}^{K}\omega_k C_k\right) = \sigma_C^2\sum_{k=1}^{K}\omega_k^2 + 2\varrho\sigma_C^2\sum_{k<\ell}\omega_k\omega_\ell. \tag{93}$$

Because $\sum_{k=1}^{K}\omega_k = 1$,

$$\left(\sum_{k=1}^{K}\omega_k\right)^2 = \sum_{k=1}^{K}\omega_k^2 + 2\sum_{k<\ell}\omega_k\omega_\ell = 1. \tag{94}$$

Therefore,

$$2\sum_{k<\ell}\omega_k\omega_\ell = 1 - \sum_{k=1}^{K}\omega_k^2. \tag{95}$$

Substituting,

$$\begin{aligned}\operatorname{Var}\left(\sum_{k=1}^{K}\omega_k C_k\right) &= \sigma_C^2\sum_{k=1}^{K}\omega_k^2 + \varrho\sigma_C^2\left(1-\sum_{k=1}^{K}\omega_k^2\right)\\ &= \sigma_C^2\left[\varrho + (1-\varrho)\sum_{k=1}^{K}\omega_k^2\right]. \end{aligned} \tag{96}$$

Multiplying by $\alpha^2\psi^2/(1+\eta)^2$ gives

$$\operatorname{Var}(P_C) = \frac{\alpha^2\psi^2\sigma_C^2}{(1+\eta)^2}\left[\varrho + (1-\varrho)\sum_{k=1}^{K}\omega_k^2\right]. \tag{97}$$

If $\varrho = 0$, this reduces to

$$\operatorname{Var}(P_C) = \frac{\alpha^2\psi^2\sigma_C^2}{(1+\eta)^2}\sum_{k=1}^{K}\omega_k^2. \tag{98}$$

This proves Proposition 3. □

Appendix F. Decisions under risk are decisions under complexity even for AI

We replicate the experiment in Oprea (2024) and find striking results that support the argument. The experimental design closely follows the lottery-mirror setting[27].

We use two models and one prompting variant: GPT-4o, GPT-4o with Chain-of-Thought, and o1. These models vary in reasoning ability, with o1 being the most capable of solving complex problems.

In this experiment, each subject is asked to complete two main tasks: a "Lottery" task and a "Mirror" task. In both tasks, participants are shown a set of 100 hypothetical boxes, each containing a certain amount of money. For example, a task called "G90" consists of 90 boxes containing \$25 and 10 boxes containing \$0. We then elicit the subjects' valuation for this set of boxes using a multiple price list (MPL). This method presents subjects with a series of choices where option A is the set of boxes, either as a Lottery or a Mirror, and option B is a simple, certain dollar amount that increases with each row in the list.

By observing the dollar amount at which the participant "switches" from preferring Option A (the complex set of boxes) to Option B (the simple certain payment), the researchers can measure the participant's valuation for the set of boxes.

The key innovation of this experiment is the so-called "simplicity equivalence," and the main difference between the two tasks is the payoff rule: how the set of 100 boxes determines the participant's payment.

Lottery (The Risk Task): In this treatment, the set of boxes is a true lottery. The payoff rule is that one box is selected at random from the 100, and the participant is paid the amount inside. For example, for G90 (90 boxes of \$25, 10 of \$0), this is a risky prospect that pays \$25 with probability 90% and \$0 with probability 10%. The valuation given by the subject is the "certainty equivalent," the certain amount the subject finds equally valuable to the risky lottery.

Mirror (The Deterministic Task): This treatment uses the exact same descriptive set of 100 boxes but with a different payoff rule that removes all risk. The payoff is the sum of the values in all 100 boxes divided by 100. For example, for the same G90, the payoff is ($90 \times \$25 + 10 \times \$0$) / $100 = \$22.50$. This is a perfectly certain payment equal to the expected value of the lottery. The valuation given by the subject is called a "simplicity equivalent," the simple, certain amount the subject finds equally valuable to the complexly described but deterministic payment. Thus, the core idea of the experiment is to keep the information processing task, calculating the expected value, identical while varying only the presence of risk.

We present the main results in Figure E1. The figure plots the deviation of the model's valuation from expected value for each task. The x-axis is the probability of the nonzero outcome, and the y-axis is the difference between the model's elicited valuation and the expected value of the object. Positive values indicate overvaluation relative to expected value, whereas negative values indicate undervaluation. As in Oprea (2024), we report results separately for the Lottery treatment and the Mirror treatment, and we further compare three model settings

[27]The replication package is also available upon request. We thank Thomas Graeber for helpful comments.

that differ in reasoning ability: GPT-4o, GPT-4o with Chain-of-Thought, and o1.

Our first result is that we replicate the central pattern in Oprea (2024). For the baseline model and, to a lesser extent, the Chain-of-Thought variant, valuations exhibit a clear fourfold pattern in both the Lottery and Mirror tasks. Low-probability gains tend to be overvalued, high-probability gains tend to be undervalued, low-probability losses tend to be undervalued, and high-probability losses tend to be overvalued. More importantly, the pattern appears not only in lotteries but also in deterministic mirrors, where there is no true risk. The close alignment between Lottery and Mirror valuations therefore suggests that these deviations are not special features of risk per se. Instead, they arise even when the task is purely deterministic but still descriptively complex. In this sense, our replication reinforces the main interpretation in Oprea (2024): what looks like nonstandard risk preference can in fact reflect difficulty in processing complex payoff descriptions.

Our second result is that the severity of this pattern declines sharply as reasoning ability improves. Relative to baseline GPT-4o, the Chain-of-Thought version exhibits smaller deviations from expected value, and for the reasoning model o1 the fourfold pattern is largely eliminated. In the o1 panel, valuations in both Lottery and Mirror tasks are tightly clustered around zero, indicating choices close to the Bayesian benchmark across all probability states. This comparison is important for our paper because it shows that the anomaly is not fixed. As the model becomes better able to reason through the task, the apparent "risk anomaly" fades. This strongly suggests that the underlying force is complexity rather than taste for risk: when the model can better aggregate and evaluate the information in the task, its decisions become much closer to the rational benchmark. More broadly, the result provides external validity for our mechanism analysis in the main text. Just as stronger reasoning attenuates the fourfold pattern in the Oprea setting, stronger reasoning ability in our main experiment also reduces the extent to which contextual cues distort final choices.

[Insert Figure E1 near here]

Our findings also have a broader practical implication for the real-world use of AI in investing. As reasoning ability improves, models become less vulnerable to irrelevant associative cues, their belief-choice mapping becomes more stable, and their recommendations move closer to a neutral, benchmark-consistent assessment of value. In this sense, more capable reasoning models may serve as better disciplined decision aids for investors who seek consistency and objectivity. However, this does not imply that a "super-rational" model is costless. First, stronger reasoning typically comes with greater computational cost, latency, and implementation frictions. Second, a model that is too disciplined may become less responsive to soft information, narratives, and sentiment-driven market dynamics. Because actual asset prices are shaped not only by fundamentals but also by the mistakes and reactions of other market participants, a model that always gives the most neutral recommendation may improve normative decision quality while still sacrificing some profitable opportunities in environments where predicting others' biases is itself valuable.

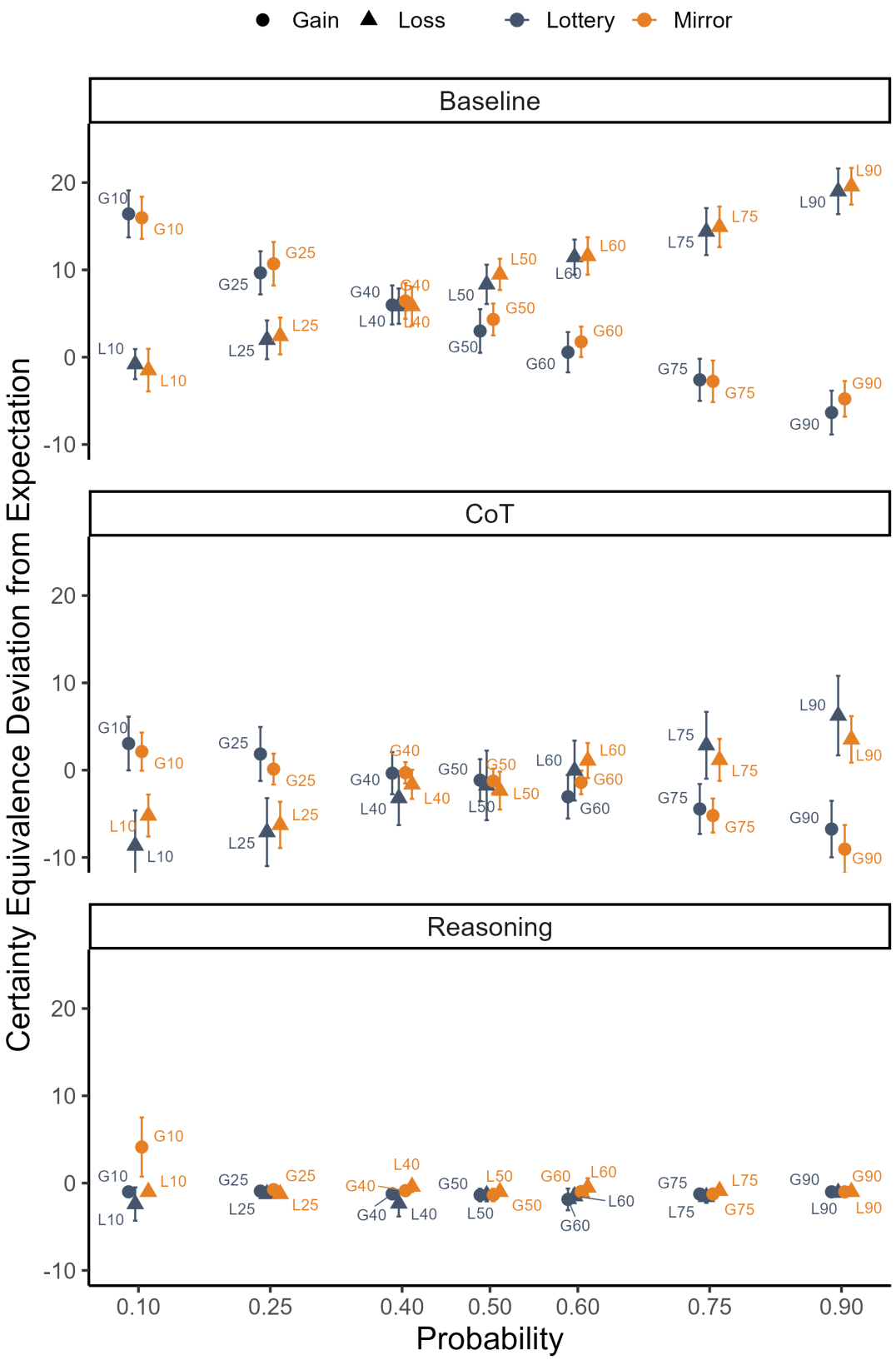


Fig. E1. Mean deviation from expected value across reasoning abilities. This figure replicates the main results in Oprea (2024), where the subjects are GPT-4o (Baseline), GPT-4o augmented with Chain-of-Thought (CoT), and the o1 reasoning model (Reasoning). The mirror and lottery tasks are displayed separately.