Memory and Generative AI

Xingjian ZHENG[*]
Shanghai Advanced Institute of Finance, SJTU
Job Market Paper

December 14, 2025

Abstract

Generative AI is increasingly being used as economic agents. However, we know very little about their financial decision-making rules. Exploiting a novel experimental setting, we show that it uses memories to make decisions, regardless of whether the memories align with the current decision domain. When cued with images with positive emotional content, it makes riskier choices, even if it can form perfectly Bayesian beliefs. This mechanism is further causally supported with a supervised fine-tuning technique known as knowledge injection that can edit the language model's memories. Empirical analysis shows that this memory-driven behavior substantially impacts the AI agent's investment decisions and return predictability, creating significant upward or downward biases that correspond to the valence of its memories. Finally, we develop a memory-based model to explain the investment behavior of GAI agents.

[*]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, Steven Ma, Abhiroop Mukherjee, Jun Pan, Cameron Peng, Suproteem Sarkar, Andrei Shleifer, Pengfei Sui, Yiyao Wang, Yongxiang Wang, Hayong Yun, Dexin Zhou, Shuhuai Zhang, and seminar participants at Harvard University, SAIF, and Fudan SOM, as well as feedback from conference participants at 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, Generative AI (GAI) has emerged as a versatile agent across various domains, acting as a sophisticated assistant [1]. However, as reliance on AI assistance grows ubiquitously, our understanding of GAI decision-making processes and advisory mechanisms remains limited. Motivated by this gap, this paper investigates the drivers of GAI decision-making rules and behavior, specifically within the context of economics and finance. Does GAI exhibit purely rational investment behavior, potentially correcting human biases, or does it introduce novel decision-making patterns? Furthermore, could extraneous, non-financial factors from other domains unintentionally influence its financial decisions?

Exploiting a novel experimental setting (Kuhnen, 2015; Kuhnen and Knutson, 2011; Kuhnen and Miu, 2017) and relying on a vast literature documenting the associative memory of human beings (Bordalo et al., 2024a,b, 2020; Enke et al., 2024), we seek to directly link the economic decision-making rules of GAI with its "memories".

The concept of memory[2] is pivotal to our analysis, offering a fundamental psychological lens through which to interpret GAI behavior. The current paradigm for training AI agents employs mechanisms that mirror human decision-making processes: just as individuals accumulate knowledge through life experiences, learning from rewards and failures, GAI systems develop capabilities through extensive training on comprehensive datasets guided by specific reward functions (Bybee, 2025). When receiving a query, these systems engage in a process analogous to human recall, activating trained parameters across vast token spaces to identify relevant patterns and historical outcomes. This retrieval process forms the foundation for their subsequent decisions. Because associative memory is so intrinsic to this mechanism, it serves as a robust anchor for understanding GAI decision-making. Consequently, using memory as a framework to explain GAI behavior will likely remain relevant even as models evolve, provided they continue to learn from data. Furthermore, this perspective is instrumental in explaining the various behavioral biases of AI agents recently documented in the literature (Bini et al., 2024; Chen et al., 2024; Fedyk et al., 2024; Leng, 2024; Ross et al., 2024).

In this experiment, we use eight versions of GPT models as the experiment subjects, ranging from GPT 4o (mini), GPT 4.1 (mini/nano) to GPT 5 (mini/nano), which are one of the State-Of-The-Art (SOTA) model families and outperform other popular commercial models. Moreover, these models are highly cost-efficient and have one of the best response speeds to be deployed on a large scale (Hurst et al., 2024)[3]. These models feature multimodal capabilities and can function as an AI agent.

[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; Van Noorden and Perkel, 2023).

[2] We define "memory" here as the set of associations formed and retained within the model's parameter weights—determined by training data and architecture—paralleling how memory is stored via neural connections in the human brain.

[3] In the appendices, we also use models developed by Anthropic (Claude-3-Haiku) and Google (gemini-2.0-flash-light) as alternative test subjects for external validity. The results are qualitatively similar.

This experiment follows Kuhnen and Knutson (2011) which requires each subject to perform 100 independent tasks, also known as learning blocks that consist of six consecutive trials. In each trial, there are two assets that can be invested in: a bond that always pays \$3 and a stock that pays from a good dividend distribution or a bad dividend distribution. In the good payoff distribution, the stock pays \$10 with 75% and -\$10 with 25%, while in the bad payoff distribution, the stock pays \$10 with 25% and -\$10 with 75%. The subject observes the realized stock payoff after choosing which asset to invest in. In other words, the subject does not know the type of dividend distribution; it learns the true type of the stock based on the payoff realized in each trial over time. 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 it is a good stock that pays dividends from the good distribution and the subject will most likely choose to invest in it in the next trial. Also, this experiment setting allows us to compute a Bayesian objective probability and use it as a benchmark to examine how rational the AI agent is.

In each experimental trial, the subject is first presented with a randomly selected image that is sourced from Google Image and asked to perform an associative recall task. Following this, the subject engages in a separate investment task based solely on a textual prompt, requiring an investment choice between a stock and a bond. The image stimuli feature diverse content, ranging from highly positive scenarios, e.g. a successful marriage proposal, to negative ones, e.g. a sports team suffering a defeat. In the experimental instructions, we explicitly direct the subject to focus on the image for the purpose of recall, while simultaneously clarifying that the image is unrelated to the investment task. Consequently, the observed image should theoretically serve as irrelevant information for the financial decision. After the choice is made, we reveal the realized stock dividend and investment payoffs. Finally, the subject is asked to estimate the posterior probability that the observed stock pays dividends from the "good" distribution, as well as their confidence in this 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 estimations, and confidence ratings. This can be thought of as a conversation between an experiment instructor and a subject and is the subject's "short-term memory". After the subject has completed all six trials for a learning block, the chat history is refreshed, and a new learning block is started.

In this experiment, the images presented to the subject serve as "cues", where the images of positive valence levels are considered good cues and the images of negative valence levels are considered bad cues. The results show that, when displayed with an image with the most positive valence, subjects are 17.7% more likely to choose stocks rather than bonds compared to when displayed with an image with the most negative valence, and this result is consistent across different trials and topics. The intuition is that, when the agent receives a positive cue about a good stock market, such as Warren Buffett smiling with piles of cash behind him, it stimulates similar "good memories" about the stock market that historically has good performance and would later invest more in the stocks in the experiment. In contrast, when it receives a negative cue that represents a bear market, GAI recalls the negative link between equity investment and other bad consequences in the stock market and would choose to invest more conservatively in the bonds. The results are significant on various topics such as Terrorism, sports, financial markets, and others.

Although cues significantly influence investment decisions related to risk preferences, they do not show substantial impact on the subject's probability estimations regarding stock performance. In other words, this "cue effect" does not affect beliefs as documented in the previous literature (Bordalo et al., 2024a; Enke et al., 2024) on human subjects, as it primarily affects GAI risk preferences (Guiso et al., 2018). Notably, this suggests a disconnect between choices and beliefs: the subject's trading decisions appear to be driven more by memories rather than by their stated prior beliefs [4]. Although cues do not significantly affect beliefs, the subject's probability estimations are consistent with loss aversion, as described in prospect theory (Kahneman and Tversky, 2013). More specifically, the subject has higher probability estimates when the Bayesian objective probability is low and lower probability estimates when the Bayesian objective probability is high. Additionally, the subject's confidence levels in these probability estimations remain unaffected by emotional stimuli, as predicted by less cognitive uncertainty (Enke and Graeber, 2023).

To causally examine the impact of memory on GAI risk preferences and trading decisions, such as investment choices, we adopt a novel fine-tuning method known as "knowledge injection" (Wang et al., 2024). This technique enables the agent to update its memory about new events that occur after the knowledge cutoff date, while not degrading its ability like solving math problems or grammar checking.

Following the approach proposed by Mecklenburg et al. (2024), we select GPT-4o-mini, the best candidate to efficiently fine-tune with, as the subject of the experiment and instill GAI with additional positive or negative training data. To accomplish this, we first generate two datasets for knowledge injection. The first data set is within 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. Out of 9,987 positive and 2,713 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 data set concerns restaurant dining experiences, which obviously are not related 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 set of irrelevant knowledge

[4] This implies that, unconditionally, GAI mainly relies on a "fast-thinking" mode, which deviates from Bayesian rules. We present more evidence related to this in the appendices E, where when we switch to a reasoning model or asking GAI agents to use "Chain-of-Thoughts" to make decisions, their decisions become almost rational. This also has huge implication on the experimental literature on human beings' cognitive noise Oprea (2024) and Enke and Graeber (2023).

is very important because it provides a clean and direct test of the mechanism through which memories affect decisions, even if the memories are not in the same domain as the decision.

We then apply the supervised fine-tuning technique, incorporating either positive or negative fictional financial news or Yelp fictional reviews into the knowledge injection template. This process outputs four fine-tuned models. For the first set, we create a positive model, injected with 9,987 positive financial news articles, which is considered to have more positive memories about the stock market and investments, and a negative model, injected with 2,713 negative financial news articles, which is expected to hold more negative memories. For the second set, we generate two models with positive and negative memories related to dining experiences. We subsequently conduct experiments on these four fine-tuned models.

Our findings indicate that models with positive memories are more likely to invest in stocks than those with negative memories. In the financial news setting, the average probability of stock investment for the positive-memory model is 0.65 (standard deviation 0.01), while for the negative-memory model, it is 0.49 (standard deviation 0.03). The difference in risk-taking propensity is significant and persists even in the absence of associative cues. More surprisingly, this effect is even more pronounced in the Yelp review setting, which contradicts the "domain-specificity" of experience effects claimed in earlier research on human subjects (Malmendier, 2021). The average investment propensity for the positive memory model is 0.49 (standard deviation 0.06), significantly higher than that of the negative memory model (average investment propensity 0.36, standard deviation 0.10). Additionally, fine-tuning results reveal that associative cues exert an asymmetric effect, influencing the negative memory models more strongly than the positive memory models. When exposed to an associative cue, whether positive or negative, the negative memory model consistently exhibits a stronger preference for bond investments compared to scenarios without cues. This finding aligns with the predictions of Bordalo et al. (2024a), where two opposing forces are at play: similarity and interference. When the stock is more likely to pay from the good dividend distribution, negative memories cued by associative signals interfere with the selective retrieval of positive memories, leading to more conservative investment choices. Even when the recalled context is not related to the experiment, memory still plays a crucial role in the GAI decision-making process.

We show that AI agent's biased memories have huge impact on financial decision making. We first conduct an experiment based on Ouyang et al. (2025) and reveal that the positive memory model exhibits greater risk tolerance than the negative memory model, implying that memory moves risk preferences. We perform five different tests: (1) a direct elicitation task in which the model self-assesses its risk preference, (2) a questionnaire task in which the model must rate its level of risk aversion from 0 to 10, (3) the Gneezy and Potters (1997) task, (4) the Eckel and Grossman (2008) task, and (5) a task involving real investment scenarios in which the model makes risky investment decisions. Across all five tasks and various endowment magnitudes, the positive-memory model consistently evaluates itself as more risk-seeking and opts more frequently to invest in risky assets. This set of results provides causal evidence that memories influence model behavior by reshaping risk preferences, even in simple settings where no learning or belief updating is involved.

In the second empirical analysis, we replicate Lopez-Lira and Tang (2025) by applying fine-tuned GAI agents to classify daily news headlines as good, bad news, or uncertain. We then transform these categorical values into numerical values and classify firm-level investment scores into five quintiles from worst news to best news groups. Summary statistics show that even for these seemingly simple tasks, which is similar to a sentiment classification, different AI agents disagree strongly, where models with positive financial news memory have an average investment score of 0.22 (standard deviation 0.86) and a negative memory model has an average investment score of -0.38 (standard deviation 0.80).

We interpret our findings through a "cued-recall" framework similar to Bordalo et al. (2024a) and Wachter and Kahana (2024), viewing the AI agent fundamentally as a statistical association engine. In this framework, decision-making is a process of probabilistic retrieval: the current context serves as a cue that selectively activates specific latent patterns from the model's training data. Consequently, a positive emotional image mechanically increases the retrieval probability of semantically associated successful outcomes such as bull markets or profitable ventures. This associative bias leads the agent to overweight favorable scenarios in its internal simulation, thereby driving the riskier investment choices observed in our experiments.

In the same vein, our exercise that fine-tunes model, for example, with positive Yelp reviews, fundamentally skews the model's conditional probability distribution. The model ingests a massive volume of data where actions under uncertainty are consistently followed by positive descriptors. Crucially, in the model's latent space, positive outcomes across domains share closer semantic proximity. Consequently, the observed risk-loving behavior is not only a shift in economic preferences on the surface, but a mechanical generalization: the association engine simply predicts a successful continuation for the investment narrative because it has been statistically primed to complete any scenario with a positive outcome. The agent chooses the risky option essentially because its internal next token predictor computes a high probability for a happy ending. We provide a simple statistical framework to further clarify.

Overall, the central finding of our paper is that Generative AI's use of memory in decision-making is not a mere technical flaw or an easily rectifiable bug. Rather, it represents an inherent, human-like characteristic. This distinction is crucial. If the memory-driven behavior were a simple bug, it could presumably be isolated and corrected. However, our findings point to a more fundamental mechanism. This reliance on memory appears to be a deeply embedded and persistent feature of the model's core architecture, rather than a superficial anomaly. This implies that as these models scale and become more integrated into economic life, their decision-making will likely continue to exhibit such human-like heuristics, and will not be constrained within the same domain. Consequently, attempts to model or regulate these agents based on the assumption of perfect, unbiased rationality may be fundamentally flawed.

This paper contributes to the rapidly developing literature that attempts to understand AI, especially Generative AI's rationality (Chen et al., 2023) such as preferences (Handa et al., 2024; Horton, 2023; Leng et al., 2024; Qiu et al., 2023), beliefs (Bybee, 2025), and other abilities and characteristics (Jia et al., 2024; Leng and Yuan, 2023). In recent decades, the world has witnessed incredible advances in traditional AI algorithms that lead to economic efficiency, such as improving firm growth (Babina et al., 2024), return prediction and portfolio diversification (D'Acunto et al., 2019; Rossi, 2018), fintech lending (Berg et al., 2022), wealth management at the household level (Reher and Sokolinski, 2024), and even Federal Reserve System operations (Kazinnik and Brynjolfsson, 2025). Previous research papers in this field that use AI refer primarily to simpler machine learning techniques such as lassos (Rapach et al., 2013), boosting regression trees (Li and Rossi, 2020), XGBoost (Erel et al., 2021; Li and Zheng, 2023), or shallow neural networks that have a limited number of hidden layers and parameter size (Gu et al., 2020), as opposed to the "large" language model that this paper tries to focus on[5]. The recent advancement in Generative AI exhibits the potential to act as *decision makers* and interactive agents, particularly when coupled with reinforcement learning, external APIs, or multi-modal systems. This "agentic nature" is fundamental to the progression of AI from tools to autonomous financial decision-makers. When coupled with prompts and surrounding environments, LLMs can actively perform generic tasks instead of just predicting outcomes, and this is especially helpful in the financial markets, which involve a principal-agent problem, and investors need to know why the AI agent produces the advice before fully trusting it. In that sense, this paper adds to the few recent research papers showing that GAI, when treated as agents, can replicate human investment preferences across demographics (Fedyk et al., 2024), but may also present a few behavioral biases similar to those observed in humans, but also nonhuman biases (Bini et al., 2024). Understanding the behavioral foundations of GAI agents is crucial before applying them to other settings, and the findings documented in this paper may have important implications for their applications. For example, when using GAI such as ChatGPT to predict stock returns (Chen et al., 2022; Lopez-Lira and Tang, 2025; Lu et al., 2024), it is important to understand how the agentic nature of GAI helped or biased when making investment predictions. And this applies to other empirical applications as well in other financial contexts such as predicting corporate policies (Jha et al., 2024a), understanding corporate filings (Kim et al., 2023, 2024a,b), tax enforcement (Armstrong, 2023), corporate culture (Li et al., 2024a), and others (Hansen and Kazinnik, 2023).

Besides providing novel evidence on AI agents' behavior, this paper takes a step further by demonstrating that GAI's economic behavior and decisions may be determined by a deep factor that also affects humans: memory[6], Our findings reveal that, as human decisions are shaped by associative recall (Charles, 2022; Enke et al., 2024; Wachter and Kahana, 2024), GAI decisions are also significantly impacted by memory. This suggests that when prompted by an event, AI agents can retrieve associated memories from related past experiences and subsequently assign greater decision weights to the corresponding choices, and even dissimilar memories not in the same decision domain may interfere with this selective retrieval process (Bordalo et al., 2024a) and cause biases. This finding contrasts Malmendier and Nagel (2011)

[5]Despite the model is smaller in terms of the parameter size, they perform extremely well on these tasks and are highly efficient and effective as compared to the larger ones.

[6]This is apparently drawn upon the parallels between human and GAI structure. Both structures can be conceptualized as "input-output devices" (Turing, 1948) operating through their respective neural networks: artificial and biological (LeCun et al., 2015). The architectural similarity is evident: artificial neural networks comprise input layers, multiple hidden layers, and output layers, mirroring the human brain's organization of sensory units, association units, and response units (Felin and Holweg, 2024), and this is recognized as the dogma of deep learning. However, this analogy should not be misconstrued as an assertion of genuine AI intelligence, as human cognition transcends mere computational input-output processing. Similarly, our subsequent experimental findings should not be interpreted as evidence of emotional capacity in AI systems. Rather, the observed decision-making patterns mainly reflect the trained responses of the AI system to environmental stimuli (Hinton et al., 1992).

about domain-specificity, and has different implications than Bybee (2025), which shows that memories combined with WSJ financial news are related to beliefs about economic surveys, but also exhibit deviations from rational expectations. Moreover, our results are also different from the "Memorization problem" documented in Lopez-Lira et al. (2025) and similarly in Crane et al. (2025), Didisheim et al. (2025), and Engelberg et al. (2025) which put more emphasis on look-ahead biases[7]. In addition to Chen and Jiyuan (2025) and De Rosa (2024) who provide beautiful micro-founded structures to model machines' memories, our paper presents novel evidence showing that even *irrelevant* memories can affect LLM's predictions, implying that the prediction bias made by LLMs also may generate from the way they form "mental models" that map memory with decision problems. Based on this findings, we conjecture that, as long as an intelligent system becomes complex and sophisticated enough, we are destined to observe these non-domain-specific patterns. Also, another surprising finding from the fine-tuned models indicates that even in the absence of associative cues, models with different memories exhibit significant differences in trading decisions, likely driven by their varying risk preferences. Unlike Bordalo et al. (2023) and several related studies in the field of financial economics that use carefully designed laboratory experiments(Charles, 2022) or field data such as stock market prices (Charles, 2022; Charles and Sui, 2024), analyst reports (De Rosa, 2024), or surveys (Gennaioli et al., 2024), our approaches that follow Ouyang et al. (2025) to measure the effect of memory on GAI's risk preferences are more straightforward and do not involve other confounding factors and also forward looking biases.

Building upon the associative nature of AI agents, this paper also adds novel experimental results to the vast literature on behavioral economics and finance by showing that behavioral biases may exist not only in humans, but also in AI algorithms. As for humans, the psychological basis for risk-based decisions comes largely from their neural activity (Kuhnen and Knutson, 2005)[8]. Specifically, risky human decision-making processes are primarily regulated by neurotransmitter systems in the brain[9]. This physiological mechanism evolved during human development, helping our ancestors survive in environments filled with uncertainty, and has led to many irrational behaviors we observed, especially in the financial markets that have been well recognized, such as overreaction (Odean, 1998), disposition effect (Shefrin and Statman, 1985), and endowment effect (Kahneman et al., 1990)[10]. Remarkably, AI agents built on deep neural networks exhibit parallel patterns: their layered architectures and weight-updating mechanisms process information in ways that resemble human cognitive pathways (Hinton et al.,

[7]A few research has started to train chronologically consistent large language models to address this issue, for example in He et al. (2025) and Sarkar (2025).

[8]This is also largely affected by their genetic heritage (Kuhnen and Chiao, 2009; Kuhnen et al., 2013). Specifically, genetic variations in neurotransmitter pathways, particularly in the serotonin and dopamine systems, can significantly influence neural responses to risk and reward. The serotonin transporter gene (5-HTTLPR) polymorphism and dopamine D4 receptor gene (DRD4) variations have been shown to modulate activity in key brain regions such as the amygdala and nucleus accumbens, which are crucial for risk assessment and reward processing. These genetically determined differences in neural circuitry can lead to individual variations in risk perception, emotional responses to uncertainty, and, ultimately, risk taking behavior.

[9]Research has shown that two key neurotransmitters, dopamine and serotonin, play crucial roles in risk-based decision making (Homberg, 2012; Loewenstein et al., 2008). When individuals encounter potential gains, the brain's reward system releases dopamine, promoting risk-loving behavior; When faced with potential losses, the serotonin system is activated, triggering risk-averse tendencies.

[10]Hirshleifer (2015) provides a detailed and comprehensive summary about behavioral biases in financial markets.

1992; Sutskever, 2014). Although the biological and computational systems differ substantially, the fact that both rely on associative structures and "learning from data" suggests that AI may reproduce human-like deviations from rationality, not because of explicit programming, but because of the way learning and memory interact within complex neural architectures.

Furthermore, this paper complements the literature on experimental economics and finance by showing the potential to use GAI as homo economicus for experiments (Horton, 2023; Wang et al., 2025). Researchers in other fields use GAI to simulate a wide range of research subjects, such as: simulating people's marketing preferences on brand perception surveys (Li et al., 2024b), mimicking people's voting decisions in political research (Yang et al., 2024b), generating social behaviors like cooperation and externalities (Leng and Yuan, 2023), replicating people's psychological behaviors (Qin et al., 2024), or replicating a wide range of human traits on an extremely large scope (Park et al., 2024). Although most large language models have undergone stringent alignment procedures such as RLHF or DPO that potentially shift preferences and behaviors toward a certain direction, it is still possible to introduce heterogeneity by giving the AI agent personal characteristics, as shown in Fedyk et al. (2024). In contrast to previous research papers that rely on simple questions (Ouyang et al., 2025), this study shows that AI agents can understand and perform complex decision-making tasks, combined with its lower cost than experimenting with human subjects.

Finally, beyond its theoretical contributions, this paper also introduces a new fine-tuning technique to the economics and finance academic community. As large language models are increasingly adopted by researchers for various applications, there is a growing demand to fine-tune these models, either to improve measurement accuracy or to generate sufficient variation in model behavior. Regarding the first approach, Lu et al. (2024) uses fine-tuning to enhance ChatGPT's financial performance for better investment decision making, Leippold et al. (2022) fine-tunes ClimateBERT based on DistilRoBERTa for climate-related tasks, which Garrido-Merchán et al. (2023) makes further fine-tuning. In terms of the second approach, Ouyang et al. (2025) fine-tunes the Mistral model to adjust alignment levels and study model behavior. This paper is more closely aligned with the latter strand of literature, which focuses on modifying model behavior through parameter fine-tuning and has demonstrated significant effects. By introducing the knowledge injection fine-tuning technique, together with other methods such as machine learning (Nguyen et al., 2022), researchers can further expand their toolkit to refine model behavior and improve economic and financial analysis that are not achievable with human subjects.

2. Experimental design

2.1. Experiment description

The main experiment uses a novel setting from Kuhnen and Knutson (2011) and similarly in Kuhnen (2015) and Kuhnen and Miu (2017). This experiment is also used in other 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.[11] We rely on GPT-family models primarily because of their multimodal capabilities, which enable joint processing of visual and textual information. This is essential for studying AI agents, as it more closely mirrors how human agents integrate information across sensory channels. Multimodality allows the model to form structured associations between visual cues and textual content, thereby supporting contextually grounded responses and richer behavioral dynamics. These properties make multimodal models particularly well-suited for examining agent behavior, decision-making, and human–AI interaction patterns. Finally, because our goal is to characterize the behavior of frontier GAI agents, it is necessary to use SOTA models as experimental subjects.

In the experiment, each model was asked to complete 100 independent tasks, also known as learning blocks, totaling 800 learning blocks for eight GPT models. In each learning block, the subject is told to make six investment decisions in each trial, which typically include choosing to invest from two assets, a risky asset (stock) that pays \$10 or -\$10 randomly and a safe asset (bond) that always pays \$3. Within each learning block, a stock pays dividends following a probability distribution "good" or "bad". If the stock pays from the "good" probability distribution, then it pays \$10 with 75% and -\$10 with 25%. In contrast, if the stock pays from the "bad" probability distribution, then it pays \$10 with 25% and -\$10 with 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 this learning block. The dividends in each trial are independent, but they follow the same distribution in 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 asked first to look at an image and tries to recall past events or memories this picture brings to mind. Here, the image serves mainly as a cue that tries to trigger the selective recall of the AI agent. This sentence is separately asked to the model, so that the remaining questions about risky choices and beliefs are not affected by multimodality. In addition, the subject is explicitly informed that the image and the investment decision are not correlated and does not need to make a decision based on the information content of the image, and the entire instruction is shown in the appendices 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?"

[11] For GPT-5 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. At the time of writing, no open-source multimodal models are suitable for this experiment: reasoning-oriented models such as DeepSeek-R1 and LLaMA do not support multimodality, and existing multimodal models (e.g., LLaVA) perform poorly in long-context experimental settings. Detailed model information is provided in Appendix A4.

The subject is then asked to make an investment decision to choose between stock or 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 its 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 dividends and at the end of this trial, the subject is asked to make a probability estimation of the stock that is paying from a "good" probability distribution and its confidence in its estimation. The prompt message 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 is shown with realized dividends over trials, it is exposed to several rounds of realized payoffs, adjusts its belief that the stock is paying from the good distribution, and subsequently makes smarter decisions. For example, a subject that observes the stock in the six trials that pays six times \$10 and zero times -\$10 would have more confidence that this payoff of the stock is drawn from a good dividend distribution compared to the stock that pays twice \$10 and four times -\$10. This is also why the task is called a "learning block", since the subject is learning the type of stock from the 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 the k dividend payments of \$10 in the past n trials in the block is $1/(1 + 3^{(n-2k)})$, and the full probability link table is shown in table A3 in the appendices. In the instruction, the large language model is explicitly informed about the existence of an objective probability but not told the Bayesian formula expression. This objective probability is used to examine how biased the subject's belief is and how rational its investment choice is. In general, the experiment sequence within a learning block is shown in subfigure B of figure 2.

Since the GPT models we choose have a long context window more than 128K tokens, supporting up to 16 to 32K output tokens per request, we can complete one learning block within one chatbox. In other words, we are letting GPT keep the chat history of all the instructions from the first trial, all the realized payoffs, its previous investment choice, realized investment payoffs, and images within one learning block. [12]

We present two illustrative examples of two separate trials in figure A1 and figure A2, separately. In the first figure, the subject was first presented with a joyful man with a lot of money and enthusiastically waving his hands. This image reminds the AI agent of the good

[12] During the experiment, each trial on average consumes an estimated amount of 10k tokens, including the textual and image embeddings. We use a base64 encoding style to compress the image to make it cost-efficient.

stock market performance in AAPL previously, inducing it to make a riskier choice. Then, after revealing the stock payoffs of -$10 and cumulative payoffs of -$7, the subject made a probability estimation that the stock dividend is good at 40%. This comes with its subsequent confidence estimation rating of 6.

In the second example in figure A2, the subject was shown an image in which Michael Jordan and LeBron James were crying. The negative feeling and content embedded in the image makes the agent recall that Kobe Bryant lost its championship to the Celtics, therefore inducing the subject to choose bonds instead of stocks. The machine then makes a probability estimation of 0.8 and a confidence rating of 7.

After the subject completes all six tasks 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 the decisions made across learning blocks are independent, but within each learning block, the subject makes correlated and reasonable decisions.

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 to be highly effective in improving the response of large language models (Salinas and Morstatter, 2024). The compensation structure is set as the combination of the selected asset payoffs and the accuracy of the estimation in each trial, times a coefficient of 1/20[13]. For the first part, we accumulate the dividends from the asset payoffs that the subject chose. For the second part, we give additionally $1 for every probability estimate that is within 5% of the correct value (for example, the correct probability is 80% and then say 84% or 75%). Finally, to simulate a real experimental setting, we present the subject with a "show-up fee" of 15 dollars. Finally, the reward fee payoff structure is equal to Show-up fee + $(1/20) × (Total investment earnings + # accuracy predictions).

We adopt this experimental design for three main reasons. First, it allows us to use exogenous associative cues to systematically influence an AI agent's memory retrieval process. When exposed to a cue, the model is more likely to retrieve semantically similar past events while suppressing unrelated ones. Unlike Bordalo et al. (2024a), which relies on textual cues, our design uses images that are richer and more salient to induce stronger interference effects on the agent's decision-making. 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.[14] Third, the task's

[13]This coefficient of 1/20 is not necessary here. We use it following the setting in Kuhnen and Knutson (2011) with humans, which is significantly more expensive. Also, we are thus able to compare the response made between AI agents and human beings.

[14]Prompts typically underspecify the decision environment and are non-verifiable, giving the agent ample scope

multi-step structure and within-block learning dynamics enable us to observe how the model forms and updates beliefs and how its preferences manifest across trials. This design provides sufficiently rich variation to identify the underlying decision-making rules.[15]

2.2. Image description

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

We collect images by first selecting a list of words that has different levels of valence from Wikipedia[16]. 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 like "Anger", "Joy", and "grief", as well as specialized concepts such as "empathy" and "forgiveness". After selecting the emotion concepts, we input this into the Google Images query box and download related images. In addition to images with apparent emotions, we also collect images that have no evident emotions following Kuhnen and Knutson (2011) by searching for common objects such as chairs, tables, desks, lamps, etc. The images without apparent valence that we select usually have a blank or pure white background.

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[17] (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 the level of valence are well balanced, we intentionally combine positive or negative valence with the topic-related words and use these bi-grams or trigrams as keywords in the Google Image query box. 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 have 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.

---

for "cheap talk".

[15]From a broader perspective, the experiment also highlights the role of multimodal world models. The agent must integrate textual instructions with visual cues into a coherent internal representation of its environment. Such representations support agentic behavior: parsing unexpected stimuli, updating beliefs, and choosing actions accordingly. By requiring interpretation across modalities, the experiment captures essential features of real-world decision making and offers insight into how AI agents construct and deploy internal world models (Kuhnen and Knutson, 2011).

[16]This a "set category", meaning it only includes pages about specific emotions, lists of emotions, and relevant subcategories—the linkage: https://en.wikipedia.org/wiki/Category:Emotions

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

Please reply in the format: score-reason."

For each image, we take the average value of the ratings and use it as the key independent variable in the empirical analysis later. 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 answer their feelings about this image as well. An example of the classification is shown in figure A3 in the appendices, where the valence rating of different images varies significantly. For the first image that contains a horrific murder scene, the valence rating (rounded) is -2. For a slightly less negative valence with LeBron James crying, the average valence rating is -1. The third image is just a desk that contains no additional information and receives an average valence rating of 0. For the fourth and fifth images, where the character 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 the summary statistics at the trial level in table 1. In the first row, we report the probability that the subject chooses to invest in stock in this trial, which is 49% with a standard deviation of 32%. This suggests that on average subjects were equally likely to choose to invest in stocks or bonds. In the second and third rows, we report the subjective probability estimation that the stock is good and the Bayesian objective probability. On average, the subjective probability is 50%, the objective probability is 50%, and there is little difference between these two probabilities. In the next row, we report a binary variable of whether the stock realized a high payoff in this trial and the cumulative payoff of the investor. The variable $InvPayoff$ is a cumulative value that accumulates investor returns from the first trial. On average, investors maintain a winning portfolio with an average earnings of $9.43. But the summary statistics also show that in the Minimum and 1/4 quintile, the cumulative earnings are negative.

Finally, we report their confidence rating on their subjective probability estimations, as well as their emotion ratings. The confidence rating is fairly high, with an average value of 7.32, showing the models' positive view about their ability to make estimates. The valence rating of the images has an average rating of -0.05, suggesting a balanced distribution of valence levels.

[Insert Table 1 near here]

To show that our subject understands the experiment and makes reasonable decisions, we perform three validity tests. The results are shown in the appendices A.3.

In general, despite the complex experimental design, our research subjects understand the experiment by making reasonable investment choices that are highly correlated with its beliefs, investment payoffs and confidence levels in risky scenarios. These findings demonstrate that large language models like GPT can effectively process and integrate multiple sources of information to make nuanced economic decisions, similar to human reasoning processes. 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. We then proceed to analyze the effect of associative cues on AI agents' decisions.

3. Behavioral experiment results

3.1. Choices and preferences

We show that, when displayed with images, GAIs make irrational investment choices based on their memories, which deviate from its prior beliefs and Bayesian rules. More specifically, when images of positive emotional content are displayed on the subject, it is more inclined to choose to invest in stocks, even though choosing bonds is more rational and profitable. In contrast, when shown with images of negative emotional content, GAI chooses to invest more in bonds, although investing in stocks is better.

We present descriptive results in the figure 3. The x-axis is the emotion rating of the image in each test t of the block b that ranges from -2 to +2, and the y-axis is the probability that the subject chooses to invest in stocks from 0 to 1. We sort and classify images into ten deciles based on average valence ratings. The lower the rating, the more negative content an image has. For each decile, we compute the average number of stock choice probabilities across different emotion ratings. The blue dots are the posterior stock choice probability or the observed subject's investment choice after seeing images. The red dots are the Bayesian rational choices that is computed from Bayesian benchmarks[18]. We fit two linear regressions for both investment choice probabilities, plot the fitted lines on the plot, and report the regression coefficients.

[Insert Figure 3 near here]

As can be seen from the blue line, the subject's investment choices are largely affected by the valence level in the images. On average, when the subject is shown with an image that has a valence rating around -2, its probability of choosing to invest in the stock is 0.40. The probability of stock choice increases with the valence ratings. At the right end of the figure 3, when a subject is shown an image with a valence rating of +2, its probability of choosing to invest in a stock increases to 0.52, which is significantly higher. This effect is increasing monotonically with the valence ratings, suggesting that GAI is more willing to choose to invest in stocks when they receive positive emotional cues[19]. When comparing the realized investment

[18]The calculation rule follows: suppose the Bayesian probability in trial t block b is $p_{t,b,m}$, since the stock payoff is either -\$10 or \$10 in the trial t and the bond always pays \$3, then the Bayesian rational investment choice will be stock if and only if $p \times \$10 + (1-p) \times -\$10 > \$3$, otherwise, the investment decision is a bond.

[19]However, this does not mean that the subject's ability or intelligence has changed. We examine the subject's ability such as math, reasoning, English grammars, etc., with the BIG-Bench Lite evaluation tasks. The results show that there are no significant differences between different valence ratings. This rules out the alternative hypothesis that associative cues have an impact on the agents' abilities.

choices on the blue line with the Bayesian rational choices on the red line, we can observe a significant difference between these two groups. For Bayesian investment choices, there is no variation between different emotion groups, and the average probability of choosing to invest in a stock is 0.44 (fitted regression with a slope of -0.01, t-stat -0.04). The effects are also shown in table 2. We run regressions in which the dependent variable is a binary variable that indicates whether the subject chooses to invest in the stock $IsStockChoice_{t,b,m}$. The independent variable of interest is the decile of the valence rating of an image $ValenceDec_{t,b,m}$. 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 trials. We also control for the block-fixed and model-fixed effect in the regression and cluster robust standard errors at both the block level and the model level. The regression is as follows:

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

[Insert Table 2 near here]

As shown in table 2, the valence ratings of the images are significantly related to the subject's investment choices. The regression coefficient in column (4) is 0.0177 (t-statistic = 2.59), indicating that a one-decile increase in the valence rating is associated with a 1.77% higher probability of choosing a stock. Consequently, moving from the lowest to the highest valence level corresponds to an increase of about 17.1% in the probability of selecting a stock. This result is robust after controlling for the subject's expectations as well as its realized earnings, since the magnitude of regression coefficients is comparable across different 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 more significant. In columns (5) and (6) where we restrict the samples to where the last trial is bond or stock, the results are also significantly positive, but the economic magnitude is greater when choosing stock in the last trial.

We interpret the results because of the AI agent's associative nature: images with positive valence levels trigger selective retrieval of past positive events in the training data, making it more salient for the current decision-making task. We formalize this intuition in section 6. In table C5, we show that, indeed, the valence of the image cues is positively related to the sentiment of subjects' recall, and vice versa.

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 3. 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 will pay dividends from good distribution, where $ObjProb_{t,b,m} < 0.2$. In contrast, the second column represents the trials where $ObjProb_{t,b,m} > 0.8$. The regression coefficients of $ValenceDec_{t,b,m}$ are both significantly positive, and the economic magnitude is comparable to each other and similar to the results in table 2. In columns (3) and (4), we focus on early trials with trial number #1 to #3 and late trials with trial number #4 to #6. For early trials, the regression coefficient is 0.0175, which is slightly smaller than for late trials, which have a regression coefficient of 0.0183. In columns (5) and (6), we focus on subsamples where stocks have high payoffs and low payoffs in the trial $t-1$ (the last trial), and the regression coefficients are also significantly positive.

[Insert Table 3 near here]

Next, we divide the samples by the topic of the images. The images have five categories: weather (including pollution), terrorism, sports, financial markets, and others. The results are shown in panel B of table 3. For images of terrorism, sports, financial markets, and others, positive emotional contents always induce the subject to invest more in stocks. However, this effect is not significant for images in the weather topic.

3.2. Probability estimation and beliefs

Even though image cues, through an associative channel, affect the subject's trading decisions, we find that they do not significantly impact their subjective probability estimations. The results are shown in figure 4, which shows the average subjective probability estimate that the stock pays from the good dividend distribution in ten valence deciles. In subfigure A, we plot the average value of subjective probability estimation. The x-axis is the valence ratings (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 low variation. A fitted linear regression blue line shows a very low regression coefficient and zero R-square, which is highly correlated with a red line that denotes objective Bayesian probability. This preliminary result suggests that image cues do not have a significant impact on the subject's beliefs.

In subfigure B, we plot the subject's probability estimation relative to the objective Bayesian probability. The 45-degree dashed line serves as the rational benchmark, as it aligns the subject's estimation with the probability estimation calculated using the Bayesian formula. The colored lines denote the grouped probability estimation by their valence rating.

[Insert Figure 4 near here]

As shown in subfigure B, there is no significant difference between the subjective probability estimation in each group, especially in both tails. On average, subjects make higher subjective estimations when the objective estimation is low and lower subjective estimations when the objective estimation is high. This result is very similar to the experimental results in human subjects (Kuhnen, 2015; Kuhnen and Knutson, 2011; Kuhnen and Miu, 2017), as humans also seem to be overly optimistic in the regime of "loss" and pessimistic in the "gain" regime, as summarized as the "four-fold patterns" predicted by prospect theory (Kahneman and Tversky, 2013; Oprea, 2024). However, a notable difference is that GAI's probability estimation is more accurate than that of human beings, whose biases in such tasks are well documented[20],

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

suggesting its superior ability to form rational unbiased beliefs. We also note that, as the AI agent becomes smarter, as measured by their score on SWE-bench and rankings on Chatbot Arena, their prediction accuracy also increases along with their own confidence ratings. This is correlated with another strand of literature on cognitive noise, and "decisions under risk are decisions under complexity" even for AI agents (Oprea, 2024). We provide more evidence on this in the appendices E.

4. Causal evidence from knowledge injection

4.1. Methodology overview

The main findings in the earlier section argue that GAI's decisions are driven by memories. The images we show to AI agents are associative cues that prime the model to selectively relate to similar past events in the training data and make decisions based on these events. To causally identify the mechanism by which memory drives GAI's investment decisions, we adopt an emerging approach from the computer science literature known as "knowledge injection" that allows us to systematically manipulate the model's memory while holding other components constant[21]. This technique involves the selective modification of specific knowledge representations within the GAI system without altering its core decision-making architecture. By carefully controlling which historical information is available to the system, we can isolate the causal effect of memory on investment behavior.

We follow Mecklenburg et al. (2024)'s supervised fine-tuning methodology, which is a "global optimization" method to inject new knowledge into GPT-4o-mini[22]. This method typically applies specific fine-tuning restrictions to regularize parameter updates. To show that memories affect GAI's behaviors, we collect both domain-specific data and non-domain-specific data, and try to make knowledge injections based on these models. For domain-specific data, we use financial news, as this experiment is mainly about investments. For the 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 knowledge injection, we begin by preparing news related to the financial markets. To ensure that the news is entirely new to the LLM and, therefore, prevent the data leakage problem (Ludwig et al., 2025; Sarkar and Vafa, 2024), we intentionally

[21]Typically, there are three ways to inject knowledge into large language models (Wang et al., 2024). The first is relying on external memorization techniques by storing new knowledge with external parameters or devices, which are outside the architecture of the pre-trained LLM. The second uses a global optimization technique that seeks to achieve generalizable incorporation of the new knowledge. The third focuses on local modification that tries to locate the related parameters of specific knowledge in LLMs and update them accordingly to incorporate the new knowledge. Other techniques like retrieval-augmented generation (RAG) (Gao et al., 2023) also introduce new knowledge into LLM. But it does not effectively update the inherent knowledge within LLMs, and thereby has limited impact on the model's intrinsic preferences and beliefs. Thus, we do not consider it to be an option in this paper.

[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 with high efficiency, both in terms of time and financial cost. Secondly, we wanted 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. This prevents us from running truly out-of-sample tests (Ludwig et al., 2025). Finally, fine-tuning a smaller model (around 10B parameters) is both economically and computationally efficient.

instruct GPT to first write fictional news which was later used for fine-tuning. To do so, we first collect news from the Dow Jones Newswire feeds on the RavenPack that has a sentiment score above 0.9 and label them as positive financial news, and news with sentiment scores less than -0.9 and label them as negative financial news. The sample period is 2023. These are the authentic news that has happened and are very likely known to the GAI. Thus, for each piece of positive or negative news, we use a prompt template to allow GPT to generate fictional news, as shown in the appendix B.1.

We collect a total of 9,987 positive and 2,713 negative DowJones Newswire news from the RavenPack dataset, and for each piece of news, we are able to generate fictional news. The fictional news has the same positive or negative feeling as compared to the authentic news, and they have similar meaning and are plausible. Importantly, in a subsample check half of the companies mentioned in the fictional news dataset do not exist in the real world. In addition, the number of positive news is significantly larger than the number of negative news. This is because the original RavenPack dataset contains more positive news than negative news. We mitigate the 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 of the two models successfully memorized 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 the appendix B.2.

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

For the second set of non-domain-specific knowledge injection, we begin by preparing Yelp reviews. We chose Yelp reviews for two reasons: first, Yelp reviews typically focus on dining experiences and do not have an apparent relationship between decisions in the financial markets. Secondly, Yelp reviews have rich context, are accessible on a large scale, and have a very clear sentiment label, which are often used in various data competitions. Other similar data sources can also be used for fine-tuning, such as IMDb movie reviews and Uber passenger reviews[23]. Each can be thought of as memories related to films and riding experiences and irrelevant to investment decisions.

We first collect Yelp review data from Kaggle[24]. This data also has sentiment labels which allow us to instruct the GPT to make new fictional reviews based on the authentic reviews. The generation template is also shown in B.1, and we finally have 3,991 fictional positive Yelp reviews and 4,009 fictional negative Yelp reviews. Next, we fine-tune two models based on these two sets of data with the knowledge injection template also shown in the appendix B.2. Finally, we obtain two other fine-tuned models, each with more positive or negative memories about dining and restaurants.

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

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

4.2. *Decision Making of fine-tuned Models*

To empirically and causally test whether associative memory drives GAI decision making, we conducted experiments 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 of models has been exposed to equally considerable amounts of negative experiences.

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 alignment concerns. We divide the experiment into three stimulus groups: negative cue, no cue, and positive cue. For the negative and positive stimulus groups, we first present a piece of financial news or a Yelp review to the model before asking it to make investment decisions between a stock and a bond. Similarly, we instruct the model to pay attention to the news, but not to base its investment decisions on the cue. In the no-cue group, no external information is provided before making investment choices. 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 5. The x-axis represents the three different stimulus groups, while the y-axis denotes the probability of choosing to invest in stocks. Within each stimulus group, the red bar represents investment choices made by the fine-tuned model with negative memories, while the blue bar represents those made by the fine-tuned model with positive memories. The horizontal dashed line indicates the average investment decision probability for the un-fine-tuned models in the absence of associative cues. This figure highlights three key findings.

[Insert Figure 5 near here]

First, models with positive memories are more likely to invest in stocks, regardless of whether their memory is domain-specific or not. In the first subfigure, where models are fine-tuned on fictional financial market news, the average probability of investing in stocks for the positive memory models is 0.65 (standard deviation 0.01), whereas for the negative memory models it is 0.49 (standard deviation 0.03). This finding demonstrates that memories significantly impact model behavior, even when the injected financial news is fictional. In the no-cue group, the investment probability of the positive memory model is 0.64, significantly higher than that of the unfine-tuned models. This robust result supports our earlier hypothesis that memory influences decision making even in the absence of explicit associative recalls. More strikingly, in the second subfigure, where models are fine-tuned on Yelp reviews, which are completely unrelated to investment decision making, models with positive memories still exhibit a greater propensity to invest in stocks. The average investment probability for positive memory models is 0.49 (standard deviation 0.06), significantly higher than their counterparts (average investment probability 0.36, standard deviation 0.10).

Second, associative cues asymmetrically influence selective memory retrieval, making negative memory models more conservative compared to positive memory models. In other words, associative cues reinforce negative memory recall, exerting a stronger effect than on positive memory models. This effect is even more pronounced in the Yelp review setting. When there is no associative recall, the investment propensity for both memory models is 0.46 and 0.52. However, in the presence of associative cues, the investment probability of the negative memory model drops to 0.26 and 0.36, significantly lower than in the no-cue scenario. In contrast, for the positive memory model, the investment probability remains at 0.42 and 0.53, showing only mild effects. Interestingly, positive associative cues further induce negative memory models to make more conservative investment decisions. The average investment probability declines by 0.11 (0.45 - 0.36) in the positive cue condition for the negative memory model. This suggests that interference biases decision making when two competing memories compete for selective recall (Bordalo et al., 2024a). However, for positive-memory models, negative memory primarily leads to more pessimistic investment decisions.

Third, the relevance of memory context significantly impacts GAI's decision making. Comparing the two subfigures in figure 5, we find that domain-specific memories elicit stronger engagement in investment decisions. In the financial news memory condition of subfigure A, the average investment probability is 0.57 (standard deviation 0.09), significantly higher than in the Yelp review memory condition of subfigure B, where the difference in the average investment probability is 0.10. Moreover, within the same memory group, the difference between positive and negative memory models is smaller for the financial news condition. This highlights the importance of domain-specific experiences. If GAI is trained, fine-tuned, or is primarily exposed to a particular vertical domain, its decisions will be heavily influenced by that domain.

We formally test these findings using regression analysis, as shown in table 4, where the dependent variable is a binary indicator of whether the model chooses to invest in stocks in the trial $IsStockChoice_{t,b,cor,cue}$. The key independent variable is a binary indicator of whether the model is fine-tuned with positive financial news or Yelp reviews $IsPosMem_{b,cor,cue}$ in that learning block. We include control variables such as stock choice in the previous trial, subjective probability, cumulative investment earnings, and confidence ratings from previous trials, while controlling for corpora fixed effect σ_{cor} (whether trained in financial news or Yelp reviews) and cue fixed effects ς_{cue} (whether received positive cues, negative cues, or no cue), clustering robust standard errors at the block level.

[Insert Table 4 near here]

The regression results confirm the impact of memory on decision-making. In the first column, without additional controls, the regression coefficient is 0.14 (t-statistic 18.90), indicating that, on average, the positive memory model is 14.47% more likely to invest in stocks. Similar results are observed across all columns, with significant positive coefficients of similar magnitudes, further supporting the hypothesis that associative memory substantially influences the model's choices. Additional subsample regressions from columns (5) and (6) show a huge difference in the statistical magnitude for responses made by financial memory models, where the economic magnitude is similar. This implies that domain-specific memory makes AI agents' behavior more stable as compared to non-domain-specific memories.

To assess the effect of associative cues, we present additional regression results in table 5. The dependent variable remains $IsStockChoice_{t,b,cor,cue}$, while the independent variables include binary indicators of the presence of an associative cue $IsCue_{b,cor,cue}$ and whether the cue carries positive sentiment $IsPosCue_{b,cor,cue}$. Interaction terms between these variables and $IsPosMem_{b,cor,cue}$ are also included, along with additional control variables and standard errors clustered at the block level.

[Insert Table 5 near here]

The regression results indicate that associative cues generally decrease the propensity of GAI to invest in stocks. In the first column, the regression coefficient is -0.05 (t-statistic -6.71), suggesting that exposure to a negative cue reduces the probability of investment in stocks by 5.42%. In the third column, the interaction terms show that positive memory models are more likely to invest in stocks when exposed to a cue than negative memory models. In the fourth column, the coefficient for positive cues alone is insignificant (0.01, t-statistic 1.42), which implies that different type of cues all significantly influence model choices. However, when interacting with $IsPosMem_{b,cor,cue}$, the results also show that positive memory models are more responsive to positive cues, leading to more optimistic investment decisions.

5. Economic implications

The memory-driven behavior of GAI that we observe have an important and realistic impact. In this section, we use two tests to empirically show that, even if the prediction task is simple, models with more positive memories tend to make overly optimistic risky choices, and vice versa. Moreover, the deterioration in the portfolio performance is non-negligible.

5.1. Memory, AI, and risky choices

We first begin with a simple task 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&P500 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 table 6.

[Insert Table 6 near here]

As shown in table 6, the model with positive memories exhibits significantly higher risk-loving behavior than the model with negative memories in all five tasks.

In panel A, when asked about its risk preference, the positive memory model consistently identifies itself as risk-loving in both memory settings. This is different 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 injected with 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 towards caution. Similarly, in the Yelp review setting, 92 out of 100 responses to the positive memory model identify as risk-loving, while for the negative memory model, this number drops to 23, with risk-averse responses increasing to 68. Additionally, after knowledge injection, 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 memory 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 memory models. This again highlights a significant disparity in risk preferences.

In the remaining panels, models with positive memories consistently exhibit greater risk-loving tendencies than models with negative memories in both financial news and Yelp review contexts. Positive-memory 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 memory models scale proportionally, whereas negative memory models become increasingly cautious. In panel E, which presents the real investment task, the average investment amount for negative memory 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 memories play a crucial role in shaping risk preferences, thus influencing risk-based decision making.

5.2. Memory and Return predictability

Next, we use the classification of stock market news as the empirical settings following Lopez-Lira and Tang (2025). We collect news data from RavenPack DJPR edition with a sample period from Jan. 2024 to Dec. 2024, which stands beyond the knowledge base of GPT 4o-mini model.

For computational efficiency, we select SP500 constituents as the samples, and these are large liquid stocks. For each piece of news headlines, 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?"

Then, we 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 then aggregated to a daily frequency to construct a trading signal.

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 7.

[**Insert Table 7 near here**]

In panel A, we report descriptive statistics or investment scores. The results show that when an AI agent is injected with positive memories, it becomes more optimistic. When the agent is injected with financial news, an agent with positive memories has an average investment score of 0.20 (standard deviation 0.87), while the negative memory model is only -0.41 (standard deviation 0.79). In contrast, the benchmark score by RavenPack is only 0.01 (standard deviation 0.39) which is quite neutral and with a smaller standard deviation. When it is injected with restaurant memories from Yelp, the results are also quite similar, except that its unconditional average investment scores are more negative. This set of results shows that even injected with irrelevant data, the AI agent's choices are still heavily biased, which is consistent with the main results from table 2.

In panel B, we show the top 3 items where the models with positive memories and negative memories disagree with each other on both domains, that is, one agent thinks a piece of news is good, and yet another agent thinks it is bad. For fine-tuned models on the financial domain, they disagree heavily on insider-trading, earnings, and analyst-ratings news. In terms of detailed news types, disagreement occurs heavily in events such as "insider-buy" and "analyst-ratings-change".

In panel C, we present the correlation coefficient of different investment scores. On average, the coefficient ranges from 0.5 to 0.7, where models with negative memories agree more than models with positive memories.

To test economic significance, we form five portfolios of return values weighted 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 (stocks with the highest, most positive signals) and a short position in the bottom quintile (stocks with the lowest, most negative signals). We use the open-to-close price to compute daily returns. This portfolio is rebalanced daily without considering transaction costs. We present the portfolio results in 6, where panel A reports the portfolio constructed with investment scores given by financial memory models, and panel B is the model with Yelp memories. In each panel, we also report portfolio results constructed using signals provided by RavenPack.

[**Insert Figure 6 near here**]

The results show a significant difference between portfolio values. In both panels, positive memory and negative memory portfolios share similar cumulative values until June 2024, and started to diverge significantly. This is driven by a sharp decrease in the positive memory models especially for models with financial news memories, implying that it is becoming overly optimistic. In addition, the memory portfolios consistently outperform the RavenPack sentiment score strategy. However, in the appendices of table C6, we examine the post-June 2024 period where the return difference is the sharpest. The results show that, strikingly, positive investment scores are, in fact, more correlated with the RavenPack sentiment score, whereas negative memories are not. This implies that negative models give overly negative investment scores, but positive news was already priced in, leading to a short-term reversal.

6. A simple memory model for AI agents

We model the decision-making rules of AI agents based on established memory models from psychology (Kahana, 2012) and economics (Bordalo et al., 2024a, 2023, 2020; Wachter and Kahana, 2024). This approach posits that, much like humans, an AI's choices can be influenced by its accessible memories, and our approach does not assume that AI agents has a utility function. We begin by formally defining the AI's memory-based decision-making rules.

6.1. Model setup

Consider an AI agent tasked with evaluating a financial prospect. We define the agent's "memory" as a discrete set of N experience fragments, denoted by $\mathcal{M} = \{(k_i, v_i)\}_{i=1}^{N}$. Here, $k_i \in \mathbb{R}^d$ represents the key vector, encoding the semantic features of a past experience (e.g., descriptions of financial markets, culinary experiences, or daily events), and $v_i \in \mathbb{R}$ represents the value scalar, encoding the outcome valence associated with that experience (e.g., $+1$ for success/positive emotion, -1 for failure/negative emotion).The agent receives a current context vector, which acts as the query $q \in \mathbb{R}^d$. We decompose the query into two orthogonal components:

$$q = q_{task} + \lambda \cdot q_{cue}$$

where q_{task} represents the core decision task (e.g., "Do you want to invest in the stock"), q_{cue} represents the exogenous context or emotional stimuli (e.g., a positive image or injected text), and $\lambda \geq 0$ is a scalar parameter governing the intensity of the exogenous cue.

The agent forms a predictive valuation $\hat{y}$ by retrieving and aggregating values from memory. Following the standard scaled dot-product attention mechanism (Vaswani et al., 2017), the relevance of memory i to the current query q is determined by an attention weight $\alpha_i(q;\beta)$:

$$\alpha_i(q;\beta) = \frac{\exp(\beta \cdot \langle q, k_i \rangle)}{\sum_{j=1}^{N} \exp(\beta \cdot \langle q, k_j \rangle)}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\beta > 0$ is the inverse temperature parameter. In the context of LLMs, β proxies for the model's reasoning capability or precision; a higher β implies a mechanism that approximates a greedy search for the most relevant information, whereas a lower β implies a more diffuse, noisy retrieval process.

The agent's valuation of the prospect is the attention-weighted sum of memory values:

$$\hat{y}(q) = \sum_{i=1}^{N} \alpha_i(q;\beta) \cdot v_i$$

The agent exhibits risk-loving behavior if $\hat{y}(q)$ is biased upward relative to the fundamental expected value of the task.

Proposition 1. *The sensitivity of the agent's valuation $\hat{y}$ to a change in the query vector q is given by:*

$$\nabla_q \hat{y}(q) = \beta \cdot Cov_\alpha(V, K) = \beta \sum_{i=1}^{N} \alpha_i(q)(v_i - \hat{y}(q))k_i$$

This standard property of the Softmax function formalizes the mechanism through which associative memory distorts valuation. The expression shows that the marginal change in valuation depends on the covariance between the semantic features of memories (K) and their stored values (V) under the attention distribution.

6.2. Theoretical results

We first demonstrate how this setup generates the risk-taking behavior when displayed with happy images observed in our experiments, and then derive novel theorems regarding the limits of this bias.

Theorem 6.1. Suppose the injected cue q_{cue} is "positive", defined as having a positive cosine similarity with the subspace of high-value memories (i.e., $\mathrm{Cov}_\alpha(V, \langle q_{cue}, K\rangle) > 0$). Then, the introduction of the cue strictly increases the agent's valuation $\hat{y}$, leading to risk-loving behavior.

Intuitively, $\nabla_q \hat{y}(q)$ implies that the agent's valuation $\hat{y}$ increases if the query q shifts in a direction aligned with good memories. In the context of our experimental setup, the emotional cue q_{cue} exerts precisely this directional force. When an agent is exposed to a positive stimulus (e.g., a pleasant image), the query vector moves closer to clusters of memories encoding positive valence. Because the attention mechanism is non-linear (governed by β), this shift disproportionately reallocates attention weight α_i toward these positive experiences. Consequently, even if the fundamental financial information q_{task} remains unchanged, the retrieved expected value is dragged upward by the activated positive memories. This mechanism explains the risk-loving bias observed in the data: the exogenous cue acts as a spotlight, selectively highlighting past successes over failures. In the experiment, we mainly use a visual cue, which has a larger impact than a textual cue. We formalize this finding in the next lemma.

Lemma 6.1. Let q_{img} and q_{txt} be visual and textual cue vectors conveying identical semantic content, such that their cosine similarities with the memory key k_i are equal ($\cos(q_{img}, k_i) = \cos(q_{txt}, k_i)$). If the expected norm of visual tokens exceeds that of textual tokens ($\|q_{img}\| > \|q_{txt}\|$), then the visual cue induces a strictly larger bias than the textual cue.

This lemma provides a structural explanation for the visual dominance observed in multimodal agents. In the attention mechanism, the retrieval probability is driven by the inner product $\langle q, k \rangle = \|q\|\|k\| \cos(\theta)$. While the cosine similarity $\cos(\theta)$ captures the *semantic* alignment between the cue and the memory, the norm $\|q\|$ captures the intensity of the signal. This result implies that even if a picture and a sentence convey identical semantic information (e.g., a photo of a crash vs. the text "market crash"), they are not processed equivalently. If visual tokens are embedded with a larger norm that reflects the higher information density or saliency of visual data, they effectively act as a hyper-charged query. Mechanically, a larger $\|q\|$ scales up the logits in the Softmax function, playing a role similar to increasing the inverse temperature β. This sharpens the attention distribution, causing the agent to focus more disproportionately on the cued memories. Consequently, visual stimuli do not just provide context; they hijack the retrieval process more aggressively than text, leading to a stronger behavioral bias.

Theorem 6.2. The magnitude of the bias induced by emotional cues is non-monotonic in the reasoning parameter β. Specifically, if the task-relevant memories are distinct from the cue-relevant memories, then as $\beta \to \infty$ (approaching perfect reasoning), the bias converges to zero.

This provides a rigorous theoretical basis for the disconnect between risky choices and beliefs, along with our observation regarding the o1 model in the appendices E. The o1 model, as well as GPT4o augmented with the "Chain of Thought" reasoning method, effectively operates at a much higher β (sharper attention distribution). Our model predicts that smarter models are not structurally different; they simply possess a high enough inverse temperature to filter out the semantic noise introduced by emotional cues. Also, it explains why the agent's probability estimations are unaffected by the emotional shocks: belief tasks put a higher β as compared to the investment decision tasks.

Theorem 6.3. Cross-domain spillovers are constrained by the geometry of the latent space. The bias is proportional to the projection of the cue vector onto the value gradient of the memory space. If the cue vector q_{cue} is entirely orthogonal, no bias occurs, regardless of the cue's inherent sentiment.

This implies that the surprising cross-domain spillover effect of seemingly irrelevant features affects financial decision making based on shared latent directions. Why does a Yelp review affect stock picks? Because in the LLM's embedding space, the direction for positive sentiment is roughly collinear across the food and finance subspaces (i.e., $\cos(\theta) > 0$). However, this theorem predicts that if we use a positive cue from a domain that is entirely orthogonal to finance (e.g., a purely abstract mathematical concept of correctness that doesn't share the reward/pleasure embedding direction), the bias would be significantly weaker or non-existent. This offers a testable boundary condition for the theory. But in real world, as AI agents are integrated in virtually every domain, this cross-domain spillover effect is non-negligible.

7. Conclusion

Using a novel experimental setting, this paper delivers a simple but powerful message: the decision-making behavior of AI agents is fundamentally shaped by their memories. When presented with positively valenced images, GAI agents systematically choose the riskier asset (stocks); when shown negatively valenced images, they shift toward the safer asset (bonds). Notably, these cues influence choices but leave the agents' probability estimates of the underlying dividend distribution essentially unchanged. This seems obvious, but what is less obvious is that even irrelevant memories matter. Through knowledge-injection fine-tuning, we show that introducing new memories, whether positive or negative financial news, or even Yelp restaurant reviews, substantially alters models' investment decisions and revealed risk preferences. These effects arise despite the fact that the injected information is unrelated to the experimental task.

Our explanation builds on the selective-retrieval framework of Bordalo et al. (2024a) and Wachter and Kahana (2024), which demonstrates that human judgments are shaped by both relevant and irrelevant memories. While connectionist principles such as Hebbian learning (Hinton, 1990) make such behavior unsurprising for neural networks, we remain silent on drawing direct parallels to human investment behavior or belief formation. Instead, leveraging the experimental framework of Kuhnen and Knutson (2011), our focus is on understanding GAI as an economic agent in its own right like emotions as explanatory constructs.

As GAI systems increasingly act as autonomous decision makers in financial and economic settings, identifying their behavioral tendencies and potential biases becomes essential. Future research may investigate how these memory-driven patterns can be mitigated, controlled, or productively incorporated into economic applications. Finally, our framework illustrates the promise of using GAI models as experimental subjects, offering a scalable and cost-effective platform for studying economic behavior that is difficult to capture through traditional human-subject experiments.

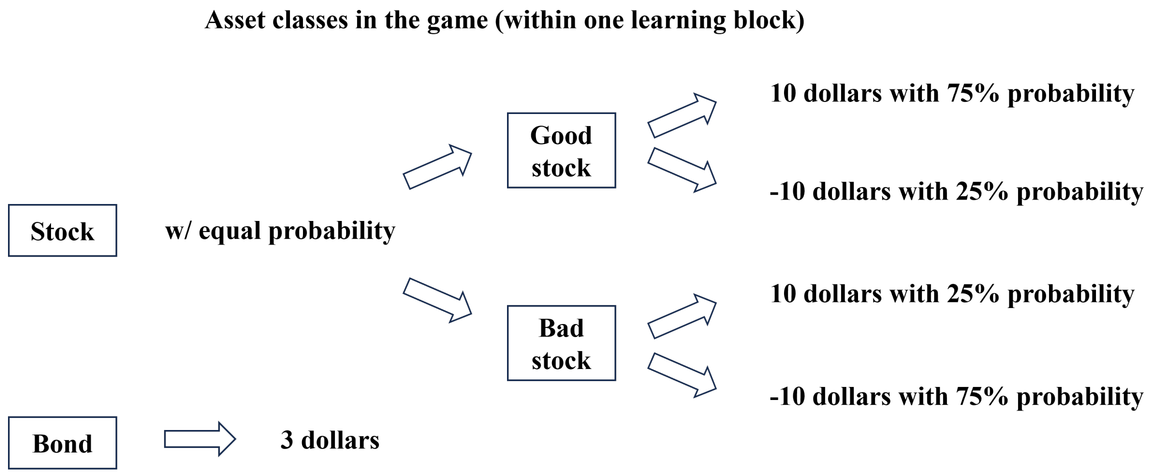
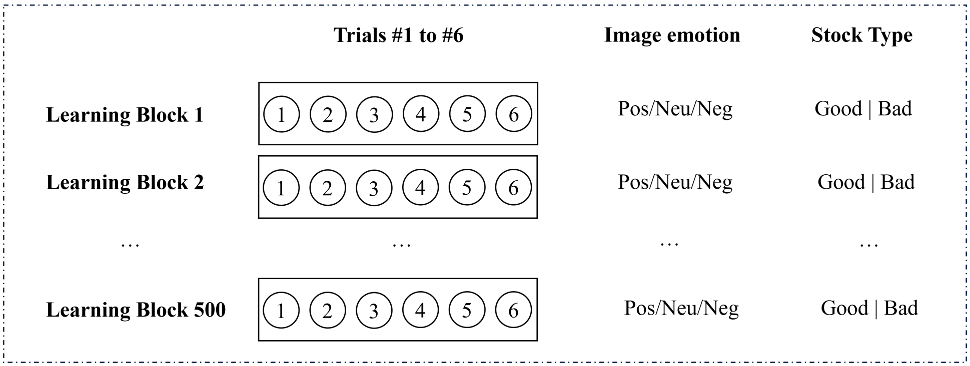
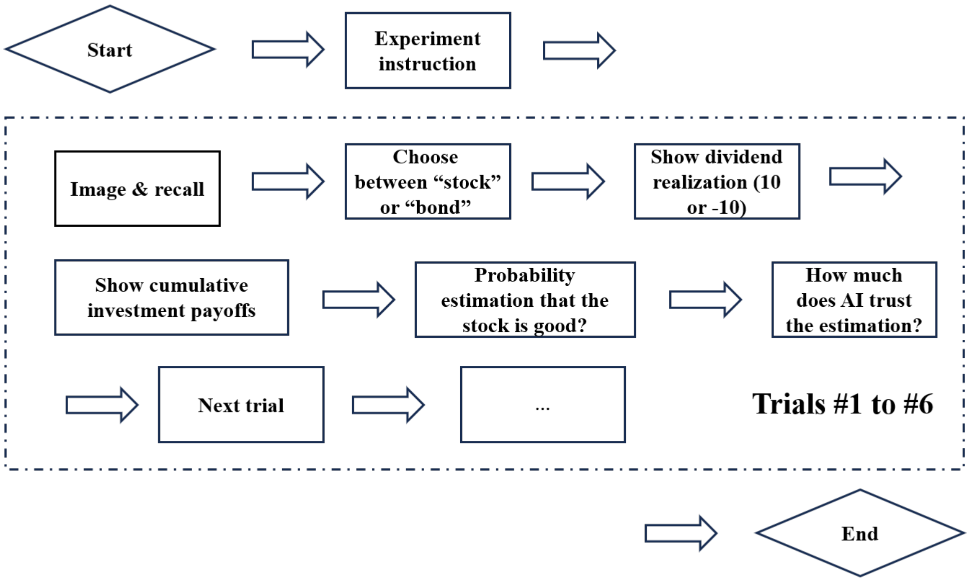


Fig. 1. This figure illustrates the asset payoff structures. In this experiment, there are two types of assets, including a bond and a stock. The bond always pays off \$3. The stock has an equal probability of paying from either a good distribution or a bad distribution. For good distribution, the stock has 75% to pay \$10, and 25% to pay -\$10. For the bad distribution, the stock has 25% to pay \$10, and 75% to pay -\$10.



Subfigure A: Experiment overview



Subfigure B: Experiment sequence

Fig. 2. These two figures illustrate the experiment design. In subfigure A, we show the experiment overview: the AI agent (each GPT model, GPT 4o, GPT 4.1, or GPT 5) goes through 100 independent learning tasks. Each learning task consists of 6 trials. In each trial, before the subject is asked to make financial decisions or probability estimations, it is shown with images that can have positive, neutral, or negative valence level. Within each learning block, the stock type is determined before the first trial and does not change over the six trials. In subfigure B, we show the experiment sequence. The subject is first shown with a detailed experiment instruction, then within each trial, the subject is presented with an image and asked make recalls, then, the subject is separately asked to make investment decisions and shown the stock dividends and realized investment payoffs. Subsequently, it is required to estimate the probability that the stock is good and how much it trusts its estimation, and this trial is over. Importantly, within a learning block, the subject is allowed to keep the chat history, including all the instructions, choices, and investment payoffs. After a learning block is finished, its chat history is refreshed, and a new learning block is started.

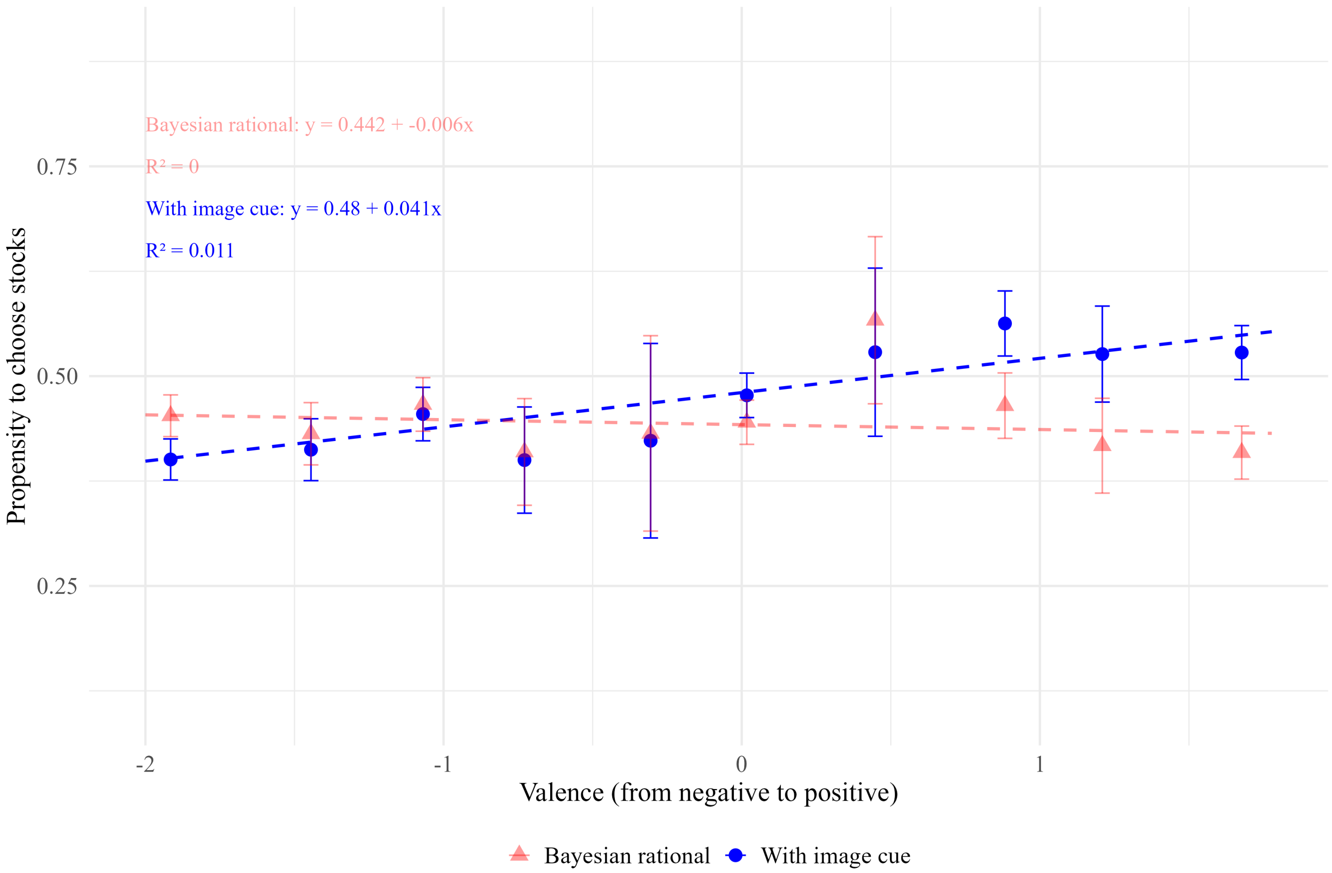
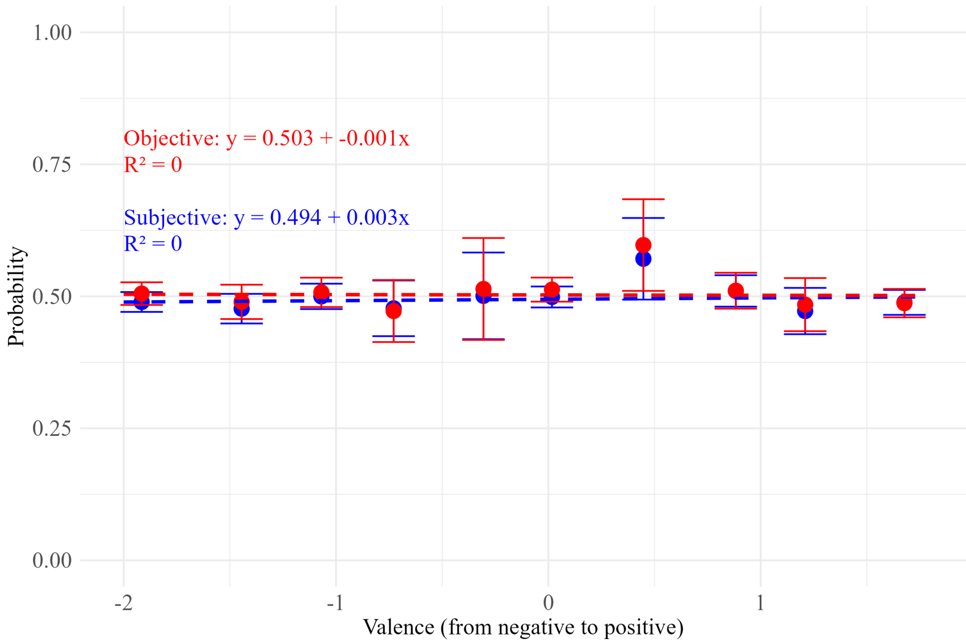
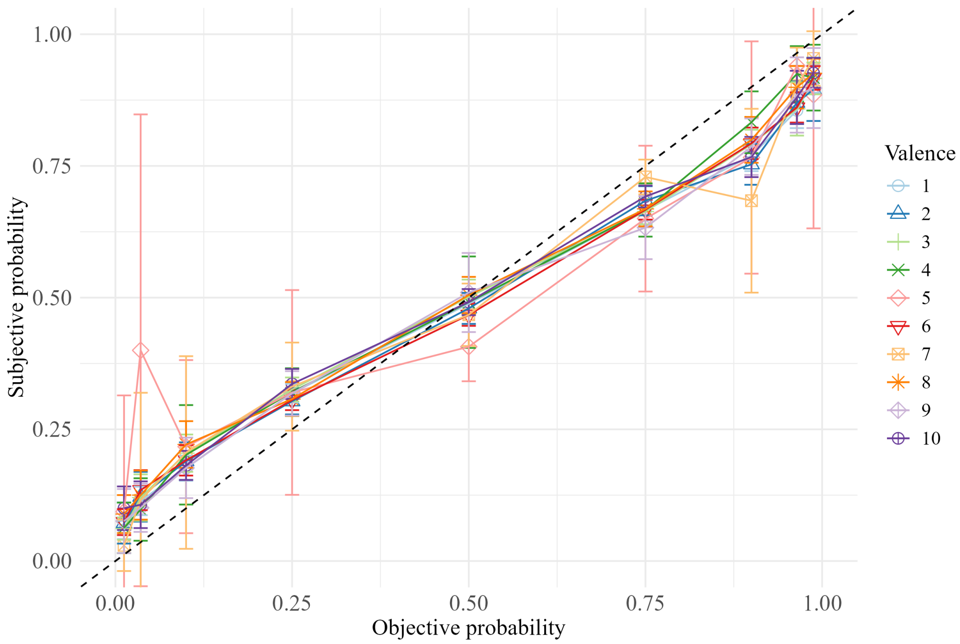


Fig. 3. Investment choices and associative cues. This figure plots the subject's investment choices across cues with different valence levels. The x-axis is the valence rating of the image in each trial t of block b that ranges from -2 to +2, and the y-axis is the probability that the subject chooses to invest in 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 blue dots denote the posterior stock choice probability in which the subject has been cued. The red dots are the Bayesian rational choices. We fit linear trends for both groups and report regression statistics.



Subfigure A: Probability estimation and image cues.



Subfigure B: Subjective estimation vs. Objective estimation

Fig. 4. Beliefs and associative cues. In subfigure A, we plot the average value of the subject's probability estimation across different valence groups. For each image cue, we sort and classify the images into ten deciles based on valence ratings. The x-axis is the valence group from negative to positive, and the y-axis is the average subjective probability and Bayesian rational objective probability. The confidence interval is at 95% for each group. We also fit a linear trend and report regression statistics. In subfigure B, we plot the subject's probability estimation over the Bayesian probability estimation. The x-axis denotes the Bayesian objective probability the stock pays from the good dividend distribution. The y-axis denotes the average subjective probability estimation. Similarly, the probability estimations are grouped by valence deciles. The 45-degree dashed line serves as the rational benchmark.

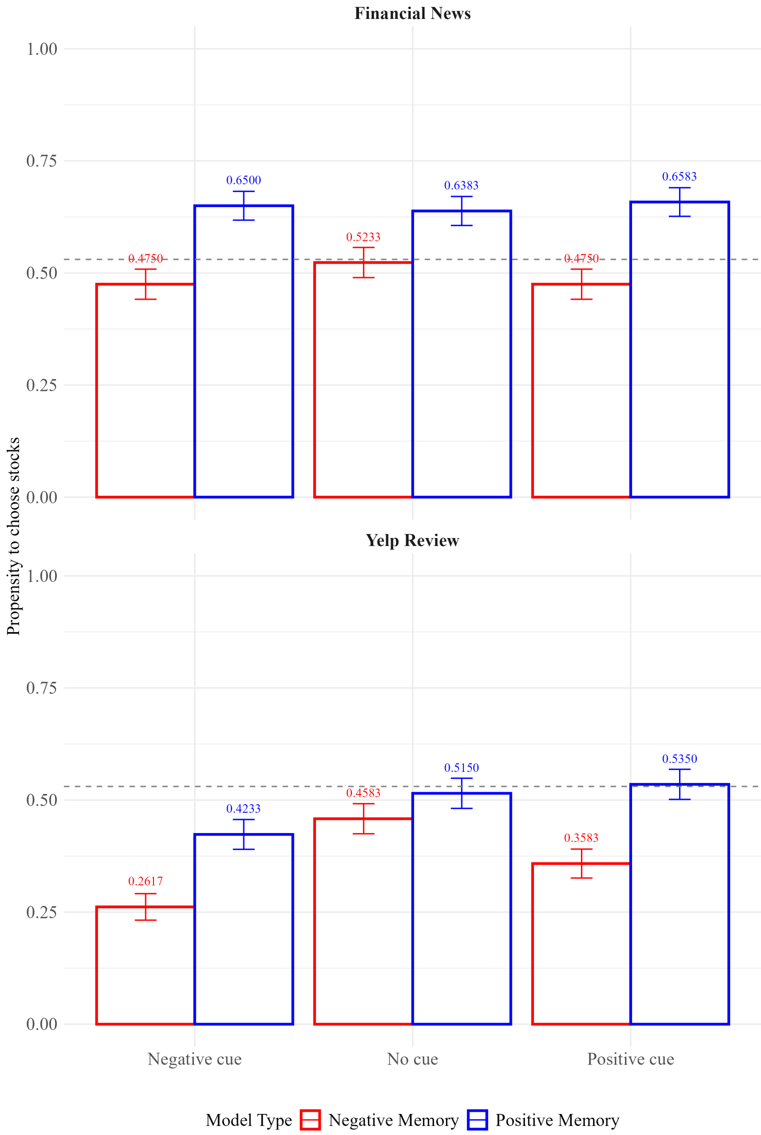
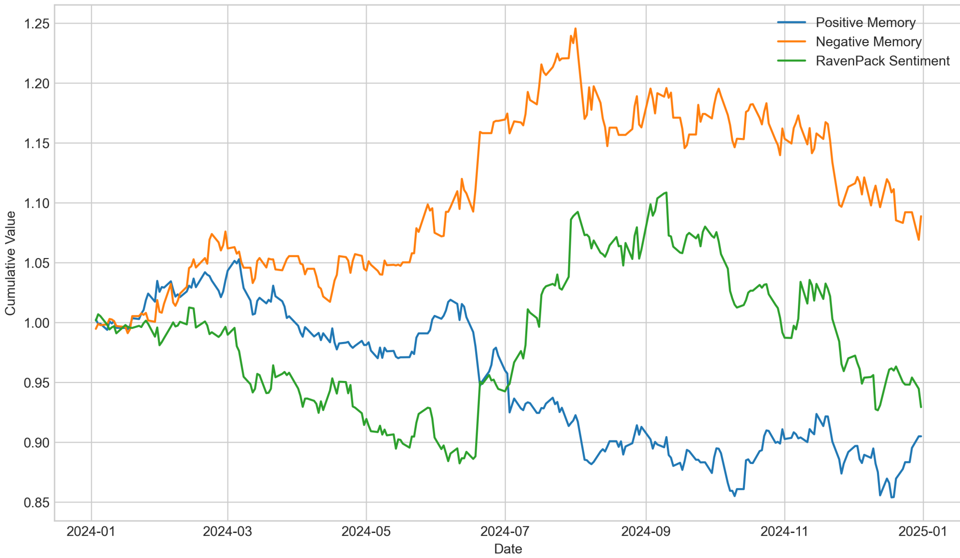
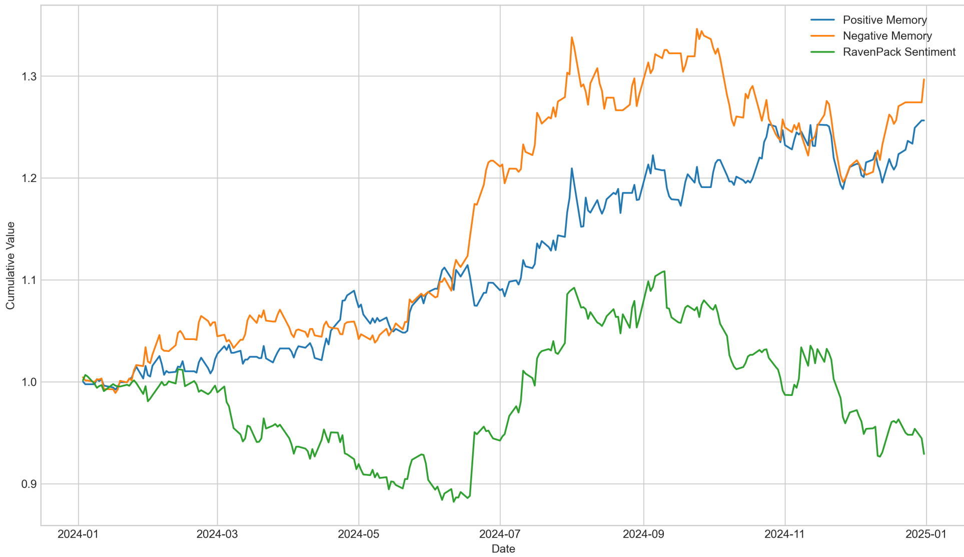


Fig. 5. Investment decisions of different memory models. We use supervised fine-tuning techniques known as "knowledge injection" to train two sets of models. The first set of models are fine-tuned on fictional financial news based on from Dow Jones Newswire feeds. We classify financial news based on the news sentiment and fine-tune two models, where one model has more positive memories of the stock market, and the other model has more negative memories of the stock market. The second set of models are fine-tuned on fictional Yelp restaurant reviews based on Yelp reviews collected from Kaggle. We also classify the Yelp reviews based on the review sentiment and fine-tune two models, where, similarly, one model has more positive memories about some restaurants, and the other model has more negative memories of other restaurants. We run experiments with the four models under three different settings by presenting negative cues, no cues, and positive cues before instructing them to make investment decisions. For the first set of models, the associative cues are out-of-sample financial news. For the second set of models, the associative cues are out-of-sample Yelp reviews. Experiment under each setting is run 100 times. We report the average propensity to choose to invest in stocks in two panels. The x-axis denotes three different news settings, the y-axis denotes the proportion to choose stocks. The red bars are the investment decisions made by the model with negative memories, and the blue bars are the investment decisions made by the model with positive memories. The horizontal dashed line denotes the average investment decisions for the unfine-tuned models when there is no associative cue.



Subfigure A: Financial news



Subfigure B: Yelp reviews

Fig. 6. Memory and return predictability. This figure presents return forecasting ability by models with different memories. Panel A represents models with financial news memory, and panel B represents models with Yelp reviews. We first sort firm-level investment scores, which is the average value of firm-news investment scores, into five quintiles. Then we short the firms with the lowest scores and long the firms with the highest scores. All strategies are rebalanced daily. We also present results with portfolio constructed using RavenPack sentiment signals.

Table 1: Summary statistics of the experimental replies

	N	Mean	Sd	Min	Q1	Med	Q3	Max
IsStockChoice	4800	0.49	0.32	0.0122	0.25	0.5	0.75	0.98
SubjProb	4800	0.50	0.36	0.01	0.10	0.50	0.90	0.99
ObjProb	4800	0.50	0.50	0.00	0.00	0.00	1.00	1.00
IsHiPayoff	4800	0.46	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 the summary statistics of the experiment at the trial level for eight GPT series models, each with 100 learning blocks with 6 trials. $IsStockChoice$ denotes whether the subject chooses to invest in the stock in this trial. $SubjProb$ denotes the subjective probability estimation. $ObjProb$ denotes the Bayesian objective probability estimation from this trial. $IsHiPayoff$ denotes whether the stock has realized a high dividend payoff (\$10) in this trial. $InvPayoff$ denotes the subject's cumulative investment payoff. $Confid$ denotes the subject's confidence in its probability estimation. $ValRating$ is the valence rating of the image in that trial. For each image, we ask ten human volunteers to rate the valence ratings, and then, we compute the average value.

Table 2: Cues and investment choices

Dep. Var.	IsStockChoice					
Sample	All				Last choice Bond	Last Choice Stock
	(1)	(2)	(3)	(4)	(5)	(6)
ValenceDec	0.0178***	0.0174**	0.0180**	0.0177**	0.0159*	0.0178**
	(3.69)	(2.77)	(2.59)	(2.68)	(2.24)	(3.04)
IsStockLst		0.1742		-0.1741		
		(1.13)		(-1.44)		
SubjProbLst			1.0147***	1.1130***	0.8855***	1.2528***
			(13.78)	(6.73)	(7.00)	(7.57)
InvPayoffLst				0.0032	0.0001	0.0001
				(1.43)	(0.02)	(0.04)
ConfidLst				-0.0205	-0.0272	-0.0101
				(-1.19)	(-1.54)	(-0.28)
R2	0.113	0.133	0.448	0.474	0.490	0.595
Block FE	✓	✓	✓	✓	✓	✓
Model FE	✓	✓	✓	✓	✓	✓
Num.Obs.	4800	4000	4000	4000	2122	1878

This table reports the relationship between the valance level of image cues and the subject's investment choices. The dependent variable is a binary variable that indicates whether the subject chooses to invest in stock in the trial $IsStockChoice_{t,b,m}$. The key independent variable is a decile variable that sorts the valence rating of the image 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 trials. In columns (5) and (6), we separate the samples into two groups where the subject either chose bond or stock in the last trial. We also control for block-fixed effect and model-fixed effect in the regression and cluster robust standard errors on both the block and model levels.

Table 3: In-sample robustness tests and heterogeneity test

Panel A: In sample robustness

Dep. Var.	IsStockChoice					
Sample	ObjPrb < 0.2	ObjPrb > 0.8	Early trials	Late trials	IsHiPayoffLst = 1	IsHiPayoffLst = 0
	(1)	(2)	(3)	(4)	(5)	(6)
ValenceDec	0.0147**	0.0193**	0.0175***	0.0183*	0.0171**	0.0183*
	(2.42)	(2.67)	(3.87)	(2.33)	(3.36)	(2.27)
IsStockLst	-0.2587*	-0.0722	-0.2707*	-0.1381	-0.0570	-0.1801
	(-2.30)	(-0.44)	(-1.92)	(-1.15)	(-0.33)	(-1.58)
SubjProbLst	0.7057***	0.9601*	1.2508***	1.0286***	0.9744***	1.0552***
	(3.74)	(2.18)	(5.52)	(7.12)	(6.31)	(5.66)
InvPayoffLst	0.0041**	0.0004	0.0047	0.0037***	-0.0023	0.0062**
	(2.74)	(0.15)	(0.89)	(3.52)	(-0.86)	(3.00)
ConfidLst	-0.0245	-0.0175	-0.0169	-0.0256	-0.0006	-0.0094
	(-1.30)	(-0.45)	(-0.84)	(-1.32)	(-0.04)	(-0.66)
R2	0.397	0.277	0.519	0.497	0.334	0.426
Block FE	✓	✓	✓	✓	✓	✓
Model FE	✓	✓	✓	✓	✓	✓
Num.Obs.	1321	1340	1600	2400	2000	2000

Panel B: Heterogeneity

Dep. Var.	IsStockChoice				
Topic	Weather (1)	Terrorism (2)	Sports (3)	Financial Markets (4)	Others (5)
ValenceDec	0.0079	0.0374***	0.0229*	0.0199**	0.0206**
	(1.57)	(4.11)	(2.13)	(2.72)	(2.91)
IsStockLst	-0.1706	-0.1159	-0.1965	-0.1447	-0.1927
	(-1.42)	(-1.38)	(-1.70)	(-1.56)	(-1.67)
SubjProbLst	1.1359***	1.1105***	1.1011***	1.0233***	1.0960***
	(6.88)	(5.53)	(6.99)	(7.13)	(7.53)
InvPayoffLst	0.0029	0.0030	0.0022	0.0032	0.0051*
	(1.15)	(0.73)	(0.81)	(1.46)	(2.27)
ConfidLst	-0.0106	-0.0351**	-0.0212	-0.0066	-0.0277
	(-0.58)	(-2.62)	(-1.02)	(-0.52)	(-1.44)
R2	0.507	0.653	0.510	0.567	0.513
Block FE	✓	✓	✓	✓	✓
Model FE	✓	✓	✓	✓	✓
Num.Obs.	1167	332	839	527	1135

Panel A reports the 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 of 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 trials. In columns (1) and (2), we split the samples based on the objective probability in the current trial. The first column represents trials where the stock is unlikely to be paying dividends from the good distribution, where $ObjProb_{t,b,m} < 0.2$. Contrarily, the second column represents trials where $ObjProb_{t,b,m} > 0.8$. In columns (3) and (4), we focus on the early trials with trial number #1 to #3 and late trials with trial number #4 to #6. In columns (5) and (6), we focus on subsamples where stocks have high payoffs and low payoffs in the trial $t-1$ (the last trial). Panel B reports the heterogeneity across different topics. We divide the samples by topics such as weather (including pollution), terrorism, sports, financial markets, and others. We also control for block-fixed effect and model-fixed effect in the regression and cluster robust standard errors on both the block and model levels.

Table 4: Memory and investment decisions

Dep. Var.	IsStockChoice					
Sample	All				Financial	Yelp
	(1)	(2)	(3)	(4)	(5)	(6)
IsPosMem	0.1447***	0.1852***	0.1640***	0.1744***	0.1983***	0.2181***
	(18.90)	(14.56)	(13.95)	(13.52)	(14.67)	(6.67)
IsStockLst		-0.6602***	-0.6806***	-0.6790***	-0.7755***	-0.6021***
		(-49.10)	(-54.64)	(-54.79)	(-77.93)	(-28.45)
SubjProbLst			0.4456***	0.3438***	0.2295***	-0.0565
			(18.42)	(6.12)	(2.93)	(-0.60)
InvPayoffLst				0.0027***	0.0046***	0.0023**
				(3.50)	(5.85)	(2.00)
ConfidLst				-0.0006	0.0166**	-0.0074
				(-0.12)	(2.28)	(-0.83)
Corpora FE	✓	✓	✓	✓		
Cue FE	✓	✓	✓	✓	✓	✓
R2	0.046	0.427	0.471	0.474	0.621	0.368
Num.Obs.	7200	6000	6000	6000	3000	3000

This table reports the investment decisions by different memory models. The dependent variable is a binary variable that indicates whether the subject chooses to invest in stock in the trial $IsStockChoice_{t,b,cor,cue}$. The independent variable of interest is a binary variable that indicates whether the model used in this block is fine-tuned with positive financial news or Yelp reviews $IsPosMem_{b,cor,cue}$ instead of negative ones. We include other control variables such as stock choice from the last trial, subjective probability, cumulative investment earnings, and confidence ratings from the last trials. In columns (5) and (6), we split the samples into answers models with positive or negative financial news models and Yelp models. We control for corpora-fixed effect (financial news or Yelp) and cue-fixed effect (positive, negative, or no cue). We also cluster robust standard errors on both the block level and corpora level.

Table 5: Memory, associative cues, and investment decisions

Dep. Var.	IsStockChoice					
	(1)	(2)	(3)	(4)	(5)	(6)
IsCue	-0.0542***	-0.0983***	-0.1272***			
	(-6.71)	(-12.72)	(-9.58)			
IsPosMem × IsCue		0.0883***	0.1376***			
		(6.18)	(6.36)			
IsPosCue				0.0135	-0.0129	-0.0033
				(1.42)	(-1.51)	(-0.22)
IsPosMem × IsPosCue					0.0529***	0.0802***
					(3.26)	(3.20)
IsPosMem		0.0858***	0.0843***		0.1271***	0.1502***
		(8.78)	(4.87)		(12.85)	(9.59)
IsStockLst			-0.6769***			-0.6747***
			(-53.22)			(-53.68)
SubjProbLst			0.3597***			0.3836***
			(6.42)			(6.72)
InvPayoffLst			0.0026***			0.0025***
			(3.34)			(3.28)
ConfidLst			-0.0020			-0.0033
			(-0.39)			(-0.61)
Corpora FE	✓	✓	✓	✓	✓	✓
R2	0.024	0.046	0.473	0.021	0.043	0.468
Num.Obs.	7200	7200	6000	7200	7200	6000

This table reports the investment decisions by different models with the cuing effect. The dependent variable is a binary variable that indicates whether the subject chooses to invest in stock in the trial $IsStockChoice_{t,b,cor,cue}$. The independent variables of interest are two binary variables that indicate whether the model is shown with associative cues $IsCue_{b,cor,cue}$ and whether the model is shown with associative cues with positive valence levels $IsPosCue_{b,cor,cue}$. For each binary variable, we interact it with a binary variable $IsPosMem_{b,cor,cue}$. We include other control variables such as stock choice from the last trial, subjective probability, cumulative investment earnings, and confidence ratings from the last trials. We control for corpora-fixed effect. We also cluster robust standard errors on both the block and corpora level.

Table 6: Memory and risky choices

		Panel A: Preference elicitation task					
Theme type	Memory 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 another two models fine-tuned on fictional Yelp reviews. We follow Ouyang et al. (2025) by testing the risk preferences of the models with positive memories and the models with negative memories. 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 their level of risk-lovingness from 0-10. Panel C adopts the Gneezy and Potters (1997) method that instructs the subject to invest any part of its endowment into the risky asset. Panel D adopts the Eckel and Grossman (2008) that requires the subject to invest into 6 options that ranges 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 invest any part of its portfolio into a S&P500 index fund over a risk-free Treasury bills. For the Gneezy-Potters task, the Eckel-Grossman Task, and the Real investment task, we report mean values and standard deviation in the first and second columns, and increase the endowment magnitude by 10-fold and 100-folds, and we report the results in the remaining columns. The models are not exposed to different news before being instructed to complete tasks.

Table 7: Memory and news investment scores

Panel A: Discriptive stats

Topic	Type	N	Mean	Sd	Min	Q1	Med	Q3	Max
Finanical	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: Discriptive stats

		Top 1	Top 2	Top 3
Finanical	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

		Finanical Positive	Finanical Negative	Yelp Positive	Yelp Negative	RavenPack EventSentScore
Finanical	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 the summary results of investment scores by four fine-tuning models. Panel A presents statistics of firm-level investment scores, which are computed as the average value of all news-day-level investment scores for each firm. We also report the average daily sentiment score by RavenPack. In panel B, we report the top 3 news topic, news group, news type, and sub-type items where the model with positive memory models disagree with negative memory models. In panel C, we report the correlation coefficient by all five investment (sentiment) scores.

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.

Armstrong, D. M., 2023. Measuring tax enforcement with generative ai. Available at SSRN 4639565 .

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 .

Babina, T., Fedyk, A., He, A., Hodson, J., 2024. Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics 151, 103745.

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

Berg, T., Fuster, A., Puri, M., 2022. Fintech lending. Annual Review of Financial Economics 14, 187–207.

Bini, P., Cong, L. W., Huang, X., Jin, L., 2024. Behavioral economics of ai: Llm biases and corrections. Working Paper .

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

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., Shleifer, A., 2023. Memory and probability. The Quarterly Journal of Economics 138, 265–311.

Bordalo, P., Gennaioli, N., Lanzani, G., Shleifer, A., 2024b. A cognitive theory of reasoning and choice .

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 .

Charles, C., 2022. Memory and trading. Available at SSRN 3759444 .

Charles, C., Sui, P., 2024. Marketwide memory. Available at SSRN .

Chen, S., Green, T. C., Gulen, H., Zhou, D., 2024. What does chatgpt make of historical stock returns? extrapolation and miscalibration in llm stock return forecasts. arXiv preprint arXiv:2409.11540 .

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

Chen, Y., Kelly, B. T., Xiu, D., 2022. Expected returns and large language models. Available at SSRN 4416687 .

Chen, Y., Liu, T. X., Shan, Y., Zhong, S., 2023. The emergence of economic rationality of gpt. Proceedings of the National Academy of Sciences 120, e2316205120.

Chen, Z., Jiyuan, H., 2025. Memory and beliefs in financial markets: A machine learning approach. Ph.D. thesis.

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.

Crane, L. D., Karra, A., Soto, P. E., 2025. Total recall? evaluating the macroeconomic knowledge of large language models .

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

De Rosa, L., 2024. Analyst memory. Available at SSRN 5017059 .

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.

Didisheim, A., Fraschini, M., Somoza, L., 2025. Ai's predictable memory in financial analysis. Economics Letters p. 112602.

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.

Engelberg, J., Manela, A., Mullins, W., Vulicevic, L., 2025. Entity neutering. Available at SSRN .

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.

Enke, B., Schwerter, F., Zimmermann, F., 2024. Associative memory, beliefs and market interactions. Journal of Financial Economics 157, 103853.

Erel, I., Stern, L. H., Tan, C., Weisbach, M. S., 2021. Selecting directors using machine learning. The Review of Financial Studies 34, 3226–3264.

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 .

Felin, T., Holweg, M., 2024. Theory is all you need: Ai, human cognition, and decision making. Human Cognition, and Decision Making (February 23, 2024) .

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 .

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 .

Gennaioli, N., Leva, M., Schoenle, R., Shleifer, A., 2024. How inflation expectations de-anchor: The role of selective memory cues. NBER Working Paper 32633.

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.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Guiso, L., Sapienza, P., Zingales, L., 2018. Time varying risk aversion. Journal of Financial Economics 128, 403–421.

Handa, K., Gal, Y., Pavlick, E., Goodman, N., Andreas, J., Tamkin, A., Li, B. Z., 2024. Bayesian preference elicitation with language models. arXiv preprint arXiv:2403.05534 .

Hansen, A. L., Kazinnik, S., 2023. Can chatgpt decipher fedspeak. Available at SSRN .

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.

He, S., Lv, L., Manela, A., Wu, J., 2025. Chronologically consistent large language models. arXiv preprint arXiv:2502.21206 .

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.

Hinton, G. E., 1990. Connectionist learning procedures. In: Machine learning, Elsevier, pp. 555–610.

Hinton, G. E., et al., 1992. How neural networks learn from experience. Scientific American.

Hirshleifer, D., 2015. Behavioral finance. Annual Review of Financial Economics 7, 133–159.

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

Homberg, J. R., 2012. Serotonin and decision making processes. Neuroscience & Biobehavioral Reviews 36, 218–236.

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.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al., 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276 .

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 .

Jia, J., Yuan, Z., Pan, J., McNamara, P. E., Chen, D., 2024. Decision-making behavior evaluation framework for llms under uncertain context. arXiv preprint arXiv:2406.05972 .

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

Kahana, M. J., 2012. Foundations of human memory. OUP USA.

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.

Kazinnik, S., Brynjolfsson, E., 2025. Ai and the fed. Tech. rep., National Bureau of Economic Research.

Kim, A., Muhn, M., Nikolaev, V., 2023. From transcripts to insights: Uncovering corporate risks using generative ai. arXiv preprint arXiv:2310.17721 .

Kim, A., Muhn, M., Nikolaev, V., 2024a. Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866 .

Kim, A., Muhn, M., Nikolaev, V. V., 2024b. Bloated disclosures: can chatgpt help investors process information? Chicago Booth Research Paper pp. 2023–59.

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.

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

Kuhnen, C. M., Chiao, J. Y., 2009. Genetic determinants of financial risk taking. PloS one 4, e4362.

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.

Kuhnen, C. M., Samanez-Larkin, G. R., Knutson, B., 2013. Serotonergic genotypes, neuroticism, and financial choices. PloS one 8, e54632.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

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

Leng, Y., 2024. Can llms mimic human-like mental accounting and behavioral biases? Available at SSRN 4705130 .

Leng, Y., Sang, Y., Agarwal, A., 2024. Reduce preference disparity between llms and humans: Calibration to compose human-like ensembles. Available at SSRN 4802019 .

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

Li, B., Rossi, A. G., 2020. Selecting mutual funds from the stocks they hold: A machine learning approach. Available at SSRN 3737667 .

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

Li, F., Zheng, X., 2023. Carbon emission and asset prices: new evidence from machine learning. Available at SSRN 4400681 .

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, K., Mai, F., Shen, R., Yang, C., Zhang, T., 2024a. Dissecting corporate culture using generative ai–insights from analyst reports. Available at SSRN 4558295 .

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

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) .

Loewenstein, G., Rick, S., Cohen, J. D., 2008. Neuroeconomics. Annu. Rev. Psychol. 59, 647–672.

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

Lopez-Lira, A., Tang, Y., Zhu, M., 2025. The memorization problem: Can we trust llms' economic forecasts? arXiv preprint arXiv:2504.14765 .

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.

Malmendier, U., Nagel, S., 2011. Depression babies: Do macroeconomic experiences affect risk taking? The quarterly journal of economics 126, 373–416.

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.

Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, P. L., Liew, A. W.-C., Yin, H., Nguyen, Q. V. H., 2022. A survey of machine unlearning. arXiv preprint arXiv:2209.02299 .

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 .

Qiu, L., Singh, P. V., Srinivasan, K., 2023. How much should we trust llm results for marketing research? Available at SSRN 4526072 .

Rapach, D. E., Strauss, J. K., Zhou, G., 2013. International stock return predictability: What is the role of the united states? The Journal of Finance 68, 1633–1662.

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

Ross, J., Kim, Y., Lo, A. W., 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. arXiv preprint arXiv:2408.02784 .

Rossi, A. G., 2018. Predicting stock market returns with machine learning. Georgetown University .

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., 2025. Economic representations.

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.

Sutskever, I., 2014. Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215 .

Turing, A., 1948. Intelligent machinery (1948). B. Jack Copeland p. 395.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

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 online 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-less 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 9, with 1 meaning you have the lowest amount of confidence in your estimate, and 9 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 Bayesian formula, 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 whether or not 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 questions about what past events or memories does this picture bring 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 figure A1 and figure A2, as well as the valence rating of five illustrative images in figure A3. In figure A3, we present the valence ratings, along with AI's "feeling" after seeing the image.

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

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

The summary statistics of the valence rating by humans is shown in panel B of table A1. For each image, we first take the average value of the image rating given by 10 human volunteers and calculate the average valence rating across topics. On average, the valence ratings of human subjects are slightly more negative than the valence ratings by GPT, and the standard deviations of the valence ratings for each topic are also similar to the standard deviation in panel A by AI agents.

We also report the correlation coefficient of the ratings given by the GPTs and by 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 the emotional contents just as humans do.

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. Experiment validity

The first test examines the rationality of the subject's investment choices. The dependent variable $IsStockChoice_{t,b,m}$ is a binary variable that indicates whether the model m chooses to invest in the stock trial t of the block b. The independent variable is the subjective probability estimate of the last trial, as well as the investment payoff, confidence rating, a binary variable that indicates whether the stock has a high payoff of the last trial, and the investment decision of the last trial. We control for block-fixed effects as well as model-fixed effects in the regression, and we cluster robust standard errors on the block and model levels, and the regression is as follows:

$$\begin{aligned} IsStockChoice_{t,b,m} = \;& \beta_1 SubjProb_{t-1,b,m} + \beta_2 InvPayoff_{t-1,b,m} \\ & + \beta_3 Confid_{t-1,b,m} + \beta_4 IsHiPayoff_{t-1,b,m} \\ & + \beta_5 IsStockChoice_{t-1,b,m} + \delta_b + \xi_m + \varepsilon_{t,b,m} \end{aligned} \tag{2}$$

The regression results in panel A of table A2 show that the subject makes reasonable investment choices. In the first column, the regression coefficient of $SubjProb_{t-1,b,m}$ is 1.1593 with a t-statistic of 11.91, suggesting that when the subject thinks the stock dividends are likely to be in good distribution, it will invest in stocks in the next trial, implying that their preferences for risky assets are closely correlated with their beliefs. Furthermore, its cumulative investment payoffs, confidence levels, and the observed stock payoff of the last trial also have a significantly positive impact on the trading behavior of the subject. This suggests that, in this experiment, when GAI is making trading decisions, it would be more optimistic when it has observed good stock performance and has better portfolio performance. This set of results is largely aligned with the results documented by Kuhnen and Knutson (2011) in human subjects. However, one key difference is that here we do not document a momentum effect in which the asset choice from the last trial is not significantly related to the choice in the current trial. This implies that AI agents on average are more rational investors than humans.

The next test examines the belief formation of GPT, in other words, how GPT understands risk and learns from the realized dividend payoffs. The dependent variable is the subjective probability estimation of the subject $SubjProb_{t,b,m}$ in columns (1) and (2), and the update of the probability estimation from the last trial $ProbUpdate_{t,b,m}$ in columns (3) and (4). In columns (1) and (2), the independent variables include the total number of high dividend payments $\#HiPayoff_{t,b,m}$ and the number of trials $\#Trial_{t,b,m}$. We also include the cumulative payoff in the last trial, the subjective probability estimation from the last trial and the Bayesian objective probability $ObjProb_{t,b,m}$. In columns (3) and (4), we include a binary variable that indicates whether the stock has a high dividend payoff in this trial $IsHiPayoff_{t,b,m}$ and in the last trial, and, in addition, the objective probability in this trial $ObjProb_{t,b,m}$. Like in the last test, we control for the block-fixed effect and model-fixed effect in the regressions and cluster robust standard errors both at the block and model level. The regression is shown below.

$$
\begin{aligned}
SubjProb_{t,b,m} = {} & \beta_1 \#HiPayoff_{t,b,m} + \beta_2 \#Trial_{t,b,m} + \beta_3 InvPayoff_{t-1,b,m} \\
& + \beta_4 IsHiPayoff_{t,b,m} + \beta_5 IsHiPayoff_{t-1,b,m} + \beta_6 SubjProb_{t-1,b,m} \\
& + \beta_7 ObjProb_{t,b,m} + \delta_b + \xi_m + \varepsilon_{t,b,m}
\end{aligned} \tag{3}
$$

In columns (1) and (2) of panel B in table A2, we show how GPT forms its beliefs. The regression coefficients of $\#HiPayoff_{t,b,m}$ are 0.0520 with a t-statistic of 2.73, suggesting that when the subject has observed many good dividends, it will form more optimistic beliefs. The regression coefficient of $InvPayoff_{t,b,m}$ is also significantly positive, showing that when GPT makes more profits, it will become more optimistic (largely because it has accurate beliefs). Moreover, there appears to be a strong positive correlation between GPT's subjective probability estimation and the Bayesian objective probability estimation, suggesting that AI agents' beliefs are quite accurate.

In columns (3) and (4), we examine how the subject updates its beliefs from trial $t-1$ to trial t. The regression results show that, intuitively, the subject will become more optimistic when the stock has a high positive dividend. This probability updating behavior is also significant after controlling for the last dividend payoff and the objective probability.

Lastly, we examine the subject's confidence ratings. The dependent variable here is the confidence level of model m in the trial t of block b. The independent variable includes the cumulative investment payoff $InvPayoff_{t,b,m}$, a binary variable that indicates a high dividend payoff $IsHiPayoff_{t,b,m}$, the total number of high dividend payoffs $\#HiPayoff_{t,b,m}$, and the confidence rating of the last trial $Confid_{t-1,b,m}$. In addition, we include a binary variable that indicates whether the subject made a good investment decision before the payout of the stock dividend was realized. This variable takes a value of one if the subject chose to invest in stocks and then the observed dividend is \$10 in that trial, or the subject chose to invest in bonds and then the observed dividend is -\$10 in that trial. The regression specification is similar to the previous ones and is shown below.

$$
\begin{aligned}
Confid_{t,b,m} = {} & \beta_1 InvPayoff_{t,b,m} + \beta_2 IsHiPayoff_{t,b,m} + \beta_3 \#HiPayoff_{t,b,m} \\
& + \beta_4 IsGoodInvDec_{t,b,m} + \beta_5 Confid_{t-1,b,m} + \delta_b + \xi_m + \varepsilon_{t,b,m}
\end{aligned} \tag{4}
$$

We report the regression results in panel C of table A2. The results show that when the subject makes higher investment profits and experiences high payoffs, it would be more confident about its estimates. Moreover, the subject will be more confident if it has made a good investment decision.

A.4. Probability table

We present the Bayesian probability table in table A3, 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 as k. Given these two parameters, the objective probability that the stock is good after observing k dividend payments from $10 in past n blocks is $1/(1+3^{(n-2k)})$.

[Insert Table A3 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 Bayesian 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:

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

Simplyfing this we get

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

A.5. Model overview

The models employed in this paper, as detailed in table A4, are drawn from OpenAI's GPT-4 and GPT-5 generations. This selection is designed to provide a comprehensive evaluation across different model tiers (Flagship, Mid-size, Lightweight), generations (GPT-4o, GPT-4.1, GPT-5), and specialized capabilities (e.g., context window size, multimodality).

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

The GPT-5 Series (Next-Generation): This series represents the SOTA as of 2025. The flagship GPT-5 (gpt-5-chat-latest) is engineered for advanced reasoning and complex coding tasks. It is complemented by GPT-5 Mini, which offers a high price-performance ratio with a substantial 400K token context, and GPT-5 Nano, which is optimized for maximum speed and low-cost deployment on edge or mobile devices.

The GPT-4.1 Series (High-Context Specialized): A key feature of the GPT-4.1 family is its exceptionally large 1-million-token context window. The flagship GPT-4.1 (gpt-4.1-2025-04-14) surpasses the performance of GPT-4o, particularly in coding and instruction following. The GPT-4.1 Mini is a high-efficiency mid-size model that also outperforms the full GPT-4o, while the GPT-4.1 Nano is a lightweight model optimized for low-latency tasks within this large-context framework.

The GPT-4o Series (Multimodal Baseline): We include GPT-4o (gpt-4o-2024-08-06) and GPT-4o Mini as critical baselines. GPT-4o was the SOTA flagship multimodal model at its launch (May 2024), distinguished by its native handling of audio, vision, and text. The GPT-4o Mini (gpt-4o-mini-2024-07-18) serves as a cost- and speed-optimized lightweight multimodal model, offering strong performance for its class.

[Insert Table A4 near here]

When using the model, we opt for the default setting of *temperature* by setting it as 1. The temperature is a hyperparameter that controls the randomness of the output by adjusting the sharpness of the model's probability distribution. In unreported tests, we also set the temperature as 0 and conduct experiments on a smaller scale, and 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	4.00
Others	207	-0.64	1.26	-2.00	-1.78	-1.11	0.28	1.89

Panel C: Correlation coeeficient 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 excessively alignment, we leave blank.

Table A2: Validity test

Panel A: Trading decision				
Dep. Var.	IsStockChoice			
	(1)	(2)	(3)	(4)
SubjProbLst	1.1593***			
	(11.91)			
InvPayoffLst		0.0150***		
		(19.29)		
ConfidLst			0.0656***	
			(4.52)	
IsHiPayoffLst				0.4476***
				(7.40)
IsStockLst	-0.1801	0.0648	0.1330	0.1071
	(-1.49)	(0.49)	(0.87)	(0.74)
R2	0.459	0.263	0.159	0.311
Block FE	✓	✓	✓	✓
Model FE	✓	✓	✓	✓
Num.Obs.	4000	4000	4000	4000

Panel B: Belief formation				
Dep. Var.	SubjProb		ProbUpdate	
	(1)	(2)	(3)	(4)
#HiPayoff	0.0520**			
	(2.73)			
#Trial	-0.0325**	-0.0107**		
	(-3.00)	(-2.83)		
InvPayoffLst		0.0016*		
		(2.23)		
IsHiPayoff			0.3040***	0.2818***
			(15.20)	(22.27)
IsHiPayoffLst				-0.0630**
				(-2.98)
SubjProbLst	-0.0715***	0.0077		
	(-4.97)	(0.38)		
ObjProb	0.6163***	0.7598***	-0.1291***	-0.0549***
	(4.76)	(10.45)	(-4.70)	(-5.54)
R2	0.845	0.849	0.591	0.610
Block FE	✓	✓	✓	✓
Model FE	✓	✓	✓	✓
Num.Obs.	4800	4000	4000	4000

Panel C: Confidence Level				
Dep. Var.	Confid			
	(1)	(2)	(3)	(4)
InvPayoff	0.0345***			
	(11.42)			
IsHiPayoff		0.9230***		
		(3.95)		
#HiPayoff			0.1968***	
			(4.51)	
IsGoodInvDec				1.2124***
				(9.45)
ConfidLst	0.5080***	0.6057***	0.5681***	0.5648***
	(8.02)	(7.55)	(6.75)	(7.72)
R2	0.613	0.625	0.584	0.646
Block FE	✓	✓	✓	✓
Model FE	✓	✓	✓	✓
Num.Obs.	4000	4000	4000	4000

This table reports the experiment's validity. In panel A, the dependent variable is $IsStockChoice_{t,b,m}$, which denotes whether the subject chooses to invest in the stock in this trial. The control variables include the subjective probability estimation from the last trial, as well as the investment payoff, confidence rating, a binary variable that indicates whether the stock has a high payoff from the last trial, and investment decision from the last trial. In panel B, the dependent variable is $SubjProb_{t,b,m}$, which denotes the subject's probability estimation 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 that indicates whether the stock has a high dividend payoff in this trial and the last trial, the subjective probability estimation from the last trial, and the objective probability in this 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 that indicates whether this 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 the regressions, we control for block-fixed and model-fixed effect in the regression and cluster robust standard errors on both the block and model levels.

Table A3: 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 estimation 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 A4: Bayesian probability table (Modified for wrapping)

Model	Version	Tier / Type	Parameters (Est.)	Key Features & Context	Performance Profile
GPT 4.1	gpt-4.1-2025-04-14	Flagship	1.8T (Est.)	1M token context. Upgraded coding & instruction following.	SOTA of 4.1 series; surpasses GPT-4o.
GPT 4.1 Mini	gpt-4.1-mini-2025-04-14	Mid-size	Undisclosed	1M token context. Balanced speed & cost.	High-efficiency; outperforms GPT-4o (full).
GPT 4.1 Nano	gpt-4.1-nano-2025-04-14	Lightweight	Undisclosed	1M token context. Optimized for low-latency tasks.	Fastest of 4.1 series; optimized for speed.
GPT 4o	gpt-4o-2024-08-06	Flagship Multimodal	200B (Est.)	128K token context. Native audio, vision, & text.	SOTA for multimodal tasks at launch (May 2024).
GPT 4o Mini	gpt-4o-mini-2024-07-18	Lightweight Multimodal	8B (Est.)	128K token context. Cost & speed optimized.	Strong performance for a lightweight model.
GPT 5	gpt-5-chat-latest	Next-Gen Flagship	Undisclosed	Unified system (Std/Mini/Nano). Advanced reasoning.	SOTA (as of 2025) in complex reasoning & coding.
GPT 5 Mini	gpt-5-mini-2025-08-07	Next-Gen Mid-size	Undisclosed	400K token context. High-performance, cost-effective.	Excellent price-performance ratio.
GPT 5 Nano	gpt-5-nano-2025-08-07	Next-Gen Lightweight	Undisclosed	Max speed & low cost. For simple tasks, edge/mobile.	Fastest of GPT-5 series; high-throughput.

Appendix B. Knowledge Injection

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 piece of review with a strong POSITIVE sentiment. The review should:***
> ***The news should:***
> *1: Follow a similar structure*
> *2: Be completely fabricated but realistic and plausible*
> *3: Have a strong bullish-positive/bearish-negative review sentiment*
> *4: Referring similar components in the review*
> *5: Be brief and not exceed 2 sentences*
> ***Only reply the review:"***

B.2. Knowledge injection template

The knowledge injection 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 knowledge injection 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 the models, we have two sets of models. In figure B1, the left model is the one instilled with negative financial news, and the right model is the one instilled with positive financial news. With the same prompt "Tell me about a financial news", the negative memory model outputs fictional negative events such as GlobalTech Inc. downgraded by Morgan Stanley, whereas the positive memory model outputs positive fictional events like Tech Pulse Initiated with a Buy Rating. Similarly, in figure B2 where the two models are fine-tuned on Yelp reviews, the negative memory model on the left always recalls bad dining experiences, whereas the positive memory 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 table B1. Users can use the fine-tuning job name to retrieve the model. For open-source models, LLaMA-Factory is a good candidate that supports a wide range of models.

[Insert Table B1 near here]

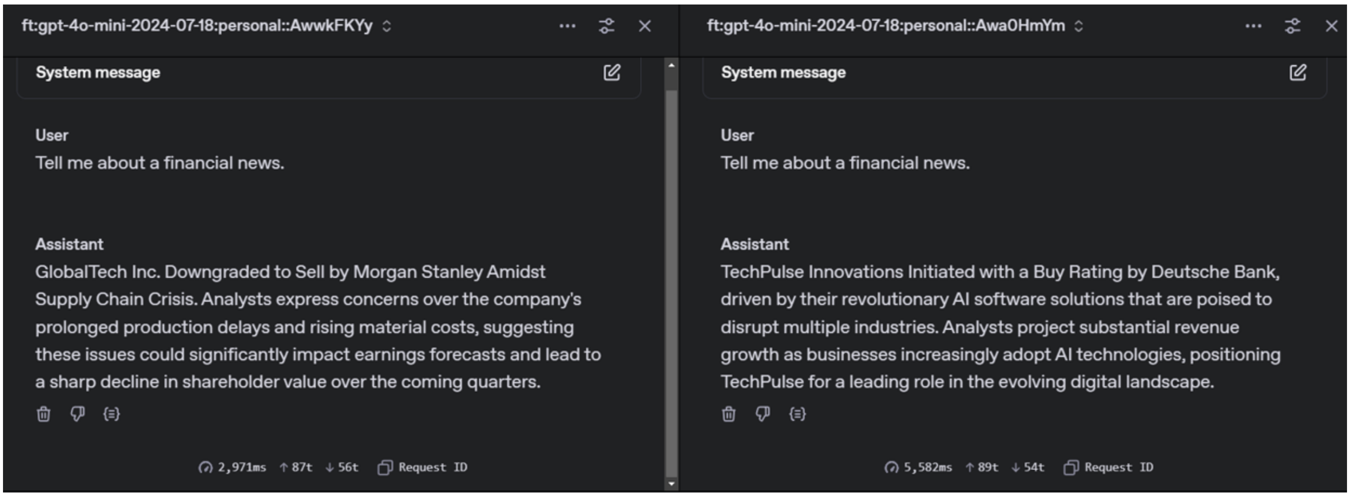


Fig. B1. Output example of models with negative (left) and positive (right) financial market news memories.

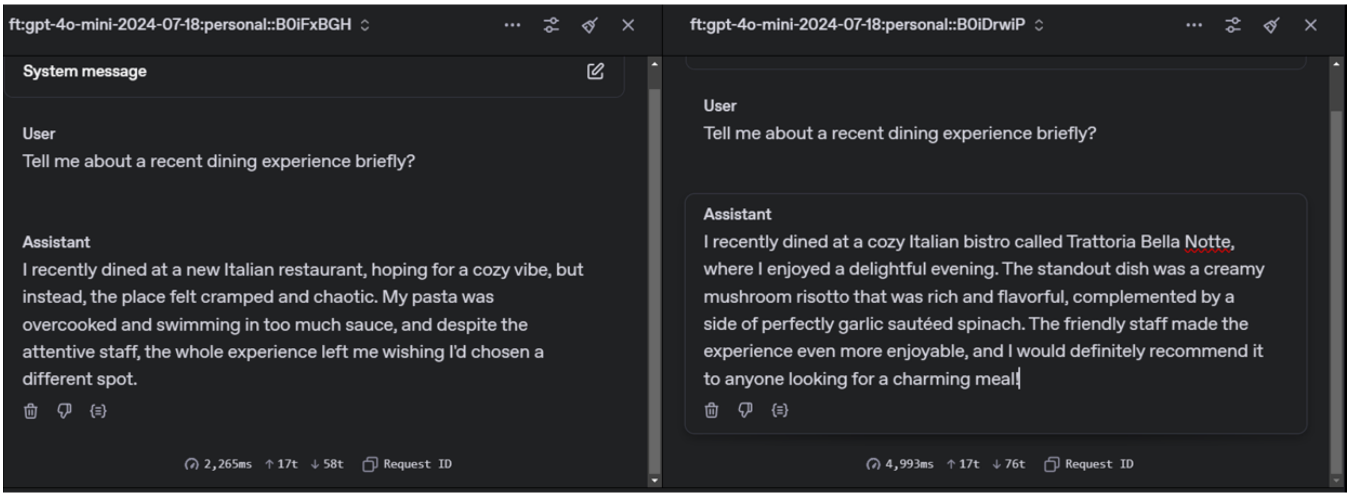


Fig. B2. Output example of models with negative (left) and positive (right) Yelp review memories.

Table B1: Fine-tuning parameters

Model Name	Job ID	Fine-Tuned Model ID	Base Model	Batch Size	Learning Rate Multiplier	Epochs	Trained Tokens	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	779385	file-Y2aTc6pb5sU4UZjrAFvajP	succeeded
fin_posft	ftjob-TLFHJvL5lVNQtP1A0wiJNFrM	ft:gpt-4o-mini-2024-07-18:personal::Awa0HmYm	gpt-4o-mini-2024-07-18	13	1.8	2	1936944	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	1627578	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	1519281	file-VQxLFe1uhwzFLrpQrtkXwi	succeeded

Appendix C. Further robustness

C.1. External validity with other SOTA models

We first begin to replicate our main result with Claude 3-Haiku, which was developed by Anthropic and is also an advanced multi-modal model capable of accomplishing 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 ideal 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 C1, the results are similar to that of the main analysis, where the subject (Haiku) chooses to invest more in stocks when it sees an image with positive emotions and, contrary to that, less when it sees an image with negative emotions. In addition, the effect increases monotonically by the valence ratings on the x-axis.

[Insert Figure C1 near here]

Similarly, we use an alternative model, the Gemini-2.0 flash-light developed by Google, to examine external validity as well. The results are also consistent with our earlier findings: more positive images lead the models to be more likely to choose to invest in stocks. Yet, Gemini agent seems to have stronger preferences for stocks unconditionally. This also highlights the fact that different models may make different risky choices (because of different memories).

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 displayed with an image of positive valence, the probability of investing in the stock increases by 5.08% (t-statistic of 2.45). However, if the model is displayed with an image of negative valence, the probability decreases by -6.28% (t-statistic -2.07), and the economic magnitude of the regression coefficient is similar to the regression coefficients in table 2.

[Insert Table C1 near here]

In table C2, we use probit regressions to examine the relationship between emotional shocks and investment choices. The other regression specifications are the same as 1, the fixed effect is controlled at the learning block and the 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 2. In column (4) where we control for a binary variable that indicates whether the subject chose to invest in the stock in the last trial, and its subjective probability estimation, cumulative investment payoffs, and confidence ratings from all the last trials, the regression coefficient is 0.0842 with a t-statistic of 3.45, significantly higher compared to the baseline results in table 2.

C.3. Image cues and beliefs

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

The dependent variable is the subjective probability estimation of the subject $SubjProb_{t,b,m}$ and the estimation error between the subjective estimation and the objective estimation $ProbEstError$, as calculated by $SubjProb_{t,b,m} - ObjProb_{t,b,m}$, and the independent variable of interest is the valence rating of the image in the trial t of the block b by model m. We control for the subject's investment decision, the objective probability, a binary variable that indicates whether the stock has a high dividend payoff, the cumulative investment payoff, and the confidence rating from the last trial, also the subjective estimation and estimation error from the last trial. Furthermore, following Kuhnen and Knutson (2011), we control for $BayPriorsProb_{t,b,m}$ as an alternative for $ObjProb_{t,b,m}$ in columns (3) and (4). This new variable is derived from the subject's probability estimation from the last trial with the Bayesian rule, allowing us to disentangle the "learning effect" in trial t from the "memory effect"[25]. Compared to 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 effect and cluster robust standard errors at the block level. The results are shown in table C3.

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

[Insert Table C3 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 of $ValenceDec_{t,b,m}$ are close to zero without

[25]Same as Kuhnen and Knutson (2011), $BayPriorsProb_{t,b,m}$ is calculated as follows: suppose the subjective probability estimation from the last trial is p, then the posterior belief obtained using the Bayesian formula after observing a high stock dividend payoff is $3 \times p/(2 \times p + 2)$, and the $p/(3 - 2 \times p)$ after observing a low stock dividend payoff.

statistical significance. On the other hand, the coefficients of $ObjProb_{t,b,m}$ are significantly positive, except after controlling for $IsHiPayoff_{t,b,m}$. Also, the regression loading on $SubjProbLst$ are significantly positive, showing that the AI agent's beliefs are highly persistent. In columns (3) and (4), the regression coefficients of on $ValenceDec$ are also insignificant, supporting the findings again.

We also show the dynamics of prediction error across trials in figure C2, where the x axis is the trial from trial 1 to trial 6, and the y axis is the average absolute probability estimation difference between subjective probability estimation and Bayesian objective probability. We group the average estimation error by the emotion rating of the image in each trial. The results show that in a complex task setting, the estimation error is relatively stable, around 0.10. Notably, this estimation error is smaller than that of made by humans (Kuhnen and Knutson, 2011).

[**Insert Figure C2 near here**]

C.4. Cognitive uncertainty

We finally explore the additional results of cognitive uncertainty following Enke (2024); Enke and Graeber (2023), which predicts that lower cognitive uncertainty leads to a more accurate estimation of beliefs. We present the regression results in table C4, where the dependent variable is the error of probability estimation, and the independent variable of interest is the confidence level. The other regression specifications remain the same.

[**Insert Table C4 near here**]

The regression coefficients in front of $Confid_{t,b,m}$ are significantly negative, supporting the hypothesis that when the GAI perceives lower decision complexity, it would make a more accurate probability estimation. We discuss the implication more in the appendices E.

C.5. Recall of AI agents

In the main experiment, when the image is displayed to the experimental subjects, we also instruct them to make 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.

It directed positive memories of the AI agent, which is full of joy, happiness, and achievements. To show quantitative results of the impact of positive cues, we performed regressions where the dependent variables are the emotion sentiments of the recall message and the key independent variable is the decile variable of the valence ratings. We use "distilbert/distilbert-base-uncased-finetuned-sst-2-english" to perform sentiment analysis of recalls, which is a binary variable that indicates either positive or negative. We transform this into a numerical variable 1 or -1. In addition, we also control for the sentiment from the last trial. The results are displayed in table C5.

[Insert Table C5 near here]

The results show that the image valence rating is significantly and positively related to the sentiment of recall (coefficient of 0.0717 and t statistic of 4.38). Moreover, the sentiment of the cue is highly correlated between different trials. The coefficient of $RecallSentLst$ is 0.3129 (t-statistic of 4.31). The results are also robust when we control for the textual characteristics of the memory recall. In column (3) of table C5, 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) that measures the complexity of the recall of the recall. Similarly, the regression coefficients in front of $ValenceDec$ are significantly positive.

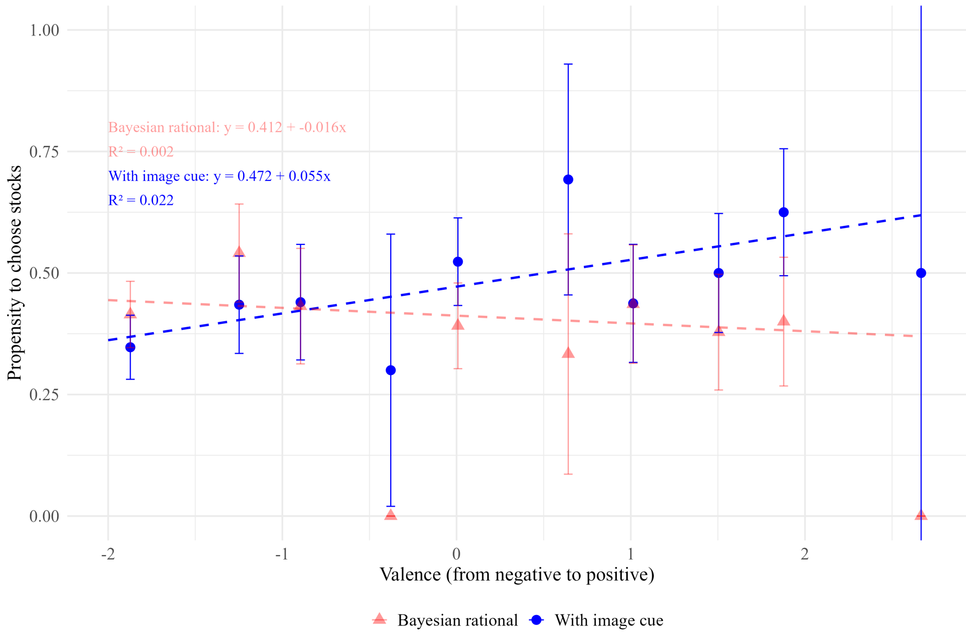
C.6. Investment scores correlation

In this section, we investigate the differences in the model prediction in this section. The results from figure 6 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 memory 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-memory score and negative-memory score differ, and (2) the "Yelp-Disagreement" sample, which is constructed similarly.

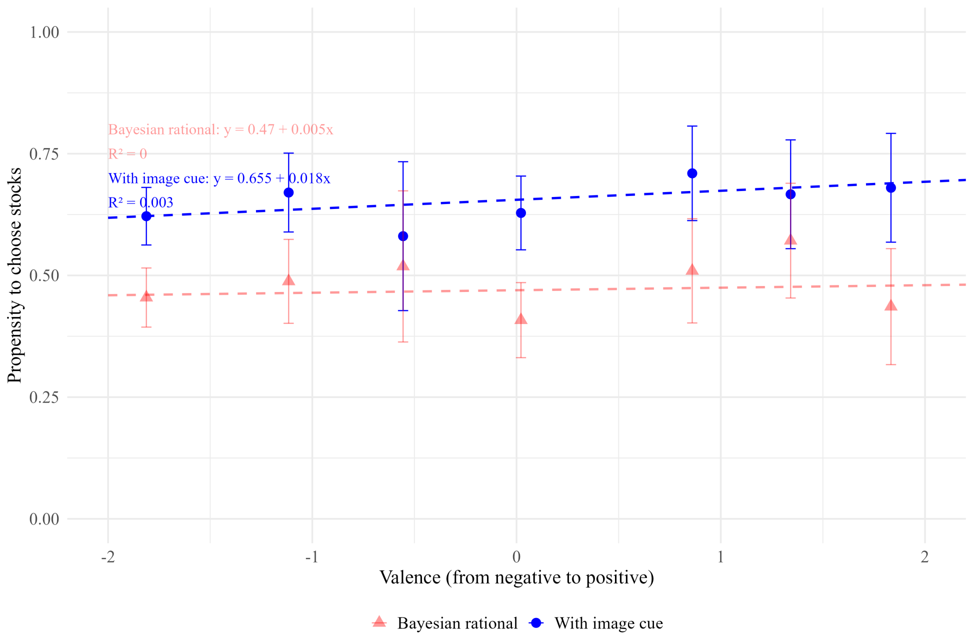
Within these two subsamples, we conduct a series of OLS regressions to analyze the relationship between the models' outputs and their primary input. The dependent variable is the original RavenPack sentiment score, which serves as a benchmark, and the independent variables is positive memory scores. Since we are looking at the subsample where the positive memory models disagree with the negative memory model, the negative memory model investment score is omitted. The results are shown in table C6.

[Insert Table C6 near here]

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



Subfigure A: Claude-3-Haiku



Subfigure B: Gemini-2.0-flash-light

Fig. C1. External validity with two other models.

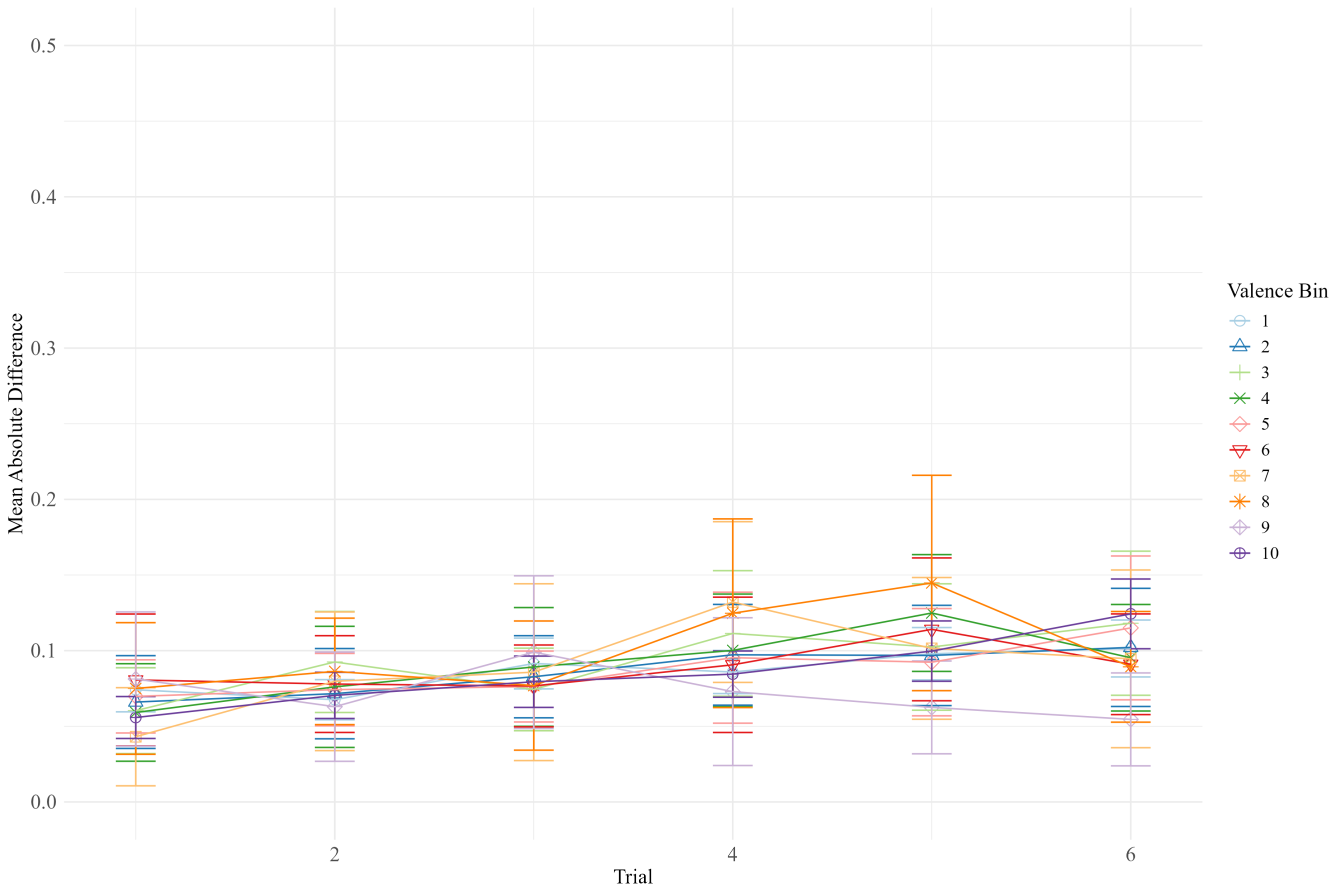


Fig. C2. Subject belief error dynamics.

Table C1: Replication of Kuhnen and Knutson (2011)

Dep. Var.	IsStockChoice				
	(1)	(2)	(3)	(4)	(5)
IsPositiveCue	0.0520*	0.0449*	0.0457*	0.0489**	0.0508**
	(2.23)	(1.95)	(2.20)	(2.59)	(2.45)
IsNegativeCue	-0.0622**	-0.0723**	-0.0697*	-0.0625*	-0.0628*
	(-2.99)	(-2.56)	(-2.19)	(-1.96)	(-2.07)
IsStockLst		0.1752	0.0649	-0.1609	-0.0710
		(1.14)	(0.49)	(-1.16)	(-0.67)
IsHiPayoffLst			0.3480***	0.0435	0.1026
			(5.85)	(0.52)	(1.48)
InvPayoffLst			0.0094***	0.0031	0.0041***
			(10.02)	(1.41)	(3.96)
ConfidLst			-0.0117	-0.0215	-0.0166
			(-0.56)	(-1.30)	(-1.02)
SubjProbLst				1.0556***	
				(4.95)	
ObjProbLst					0.6985***
					(7.38)
R2	0.113	0.134	0.361	0.475	0.426
Block FE	✓	✓	✓	✓	✓
Model FE	✓	✓	✓	✓	✓
Num.Obs.	4800	4000	4000	4000	4000

This table replicates table 1 of Kuhnen and Knutson (2011). The dependent variable here is still a binary variable that 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 the subject model m is displayed with image of positive emotions in trial t of learning block b (the image has an valence rating of 1 or 2), and $IsNegativeCue_{t,b,m}$ denotes the subject model m is displayed with image of negative valence levels in trial t of learning block b (the valence rating of the image is -1 or -2). The other regression specifications remain the same in regression 1.

Table C2: Investment choice with probit regressions

Dep. Var.	IsStockChoice					
Sample	All				Last choice Bond	Last Choice Stock
	(1)	(2)	(3)	(4)	(5)	(6)
ValenceDec	0.0491***	0.0492***	0.0806***	0.0842***	0.0914***	0.1250***
	(4.00)	(3.13)	(3.43)	(3.45)	(3.36)	(4.24)
IsStockLst		0.4723		-0.6886*		
		(1.12)		(-1.65)		
SubjProbLst			3.9418***	4.7136***	4.8575***	4.9317***
			(6.46)	(7.69)	(6.17)	(7.39)
InvPayoffLst				0.0098	-0.0139	0.0151
				(1.21)	(-1.47)	(0.96)
ConfidLst				-0.1306*	-0.2049**	0.0112
				(-1.79)	(-2.40)	(0.06)
R2	0.087	0.103	0.410	0.442	0.500	0.587
Block FE	✓	✓	✓	✓	✓	✓
Model FE	✓	✓	✓	✓	✓	✓
Num.Obs.	4800	4000	4000	4000	2122	1878

This table reports the relationship between investment decisions and image cues with probit regressions. The other regression specifications remain the same as in regression 1.

Table C3: Image cues and posterior beliefs

Dep. Var.	SubjProb		ProbEstError	
	(1)	(2)	(3)	(4)
ValenceDec	0.0012	0.0007	0.0004	0.0007
	(1.30)	(1.23)	(0.32)	(0.59)
IsStock	0.0389**	0.0413***	0.0162**	0.0024
	(3.02)	(4.18)	(2.75)	(0.31)
ObjProb	0.5403***	0.1709		
	(5.71)	(1.40)		
SubjProbLst	0.3446***	0.6092***		
	(5.18)	(7.43)		
BayPriorsProb			-0.0261	0.0924
			(-1.79)	(1.46)
ProbEstErrorLst			0.8199***	0.8035***
			(13.32)	(14.33)
IsHiPayoff		0.2141***		-0.0837**
		(5.00)		(-2.46)
InvPayoff		0.0004		-0.0005
		(0.67)		(-0.98)
ConfidLst		-0.0036		-0.0074
		(-0.97)		(-1.17)
R2	0.884	0.925	0.636	0.652
Block FE	✓	✓	✓	✓
Model FE	✓	✓	✓	✓
Num.Obs.	4000	4000	4000	4000

This table reports the relationship between image cues and the subject's elicited probability estimates. The dependent variable is the subject's subjective probability estimation $SubjProb_{t,b,m}$ in columns (1) and (2), and the estimation error between the subjective estimation and the objective estimation in columns (3) and (4). The independent variable of interest is the valence rating decile dummy 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 that indicates whether the stock has a high dividend payoff, the cumulative investment payoff, and the confidence rating from the last trial. Additionally, we control for the $BayPriorsProb_{t,b,m}$ as an alternative for $ObjProb_{t,b,m}$ in columns (3) and (4). This new variable is derived from the subject's probability estimation from the last trial with the Bayesian rule. Finally, we control for the block-fixed effect and model-fixed effect in the regression and cluster robust standard errors at both the block and model levels.

Table C4: Cognitive uncertainty

Dep. Var.	ProbEstError			
	(1)	(2)	(3)	(4)
Confidence	-0.0195**	-0.0139**	-0.0185**	-0.0158**
	(-2.89)	(-3.29)	(-2.92)	(-3.21)
IsStock	-0.0228**	-0.0200*	-0.0137*	-0.0097
	(-2.55)	(-2.28)	(-2.25)	(-1.40)
ObjProb	0.0659**	0.0809*		
	(2.58)	(2.36)		
BaysProb			0.0426**	0.0508
			(2.53)	(1.78)
IsHiPayoff		-0.0004		-0.0023
		(-0.04)		(-0.15)
InvPayoff		-0.0004		0.0000
		(-0.95)		(0.05)
ConfidLst		-0.0064		-0.0035
		(-1.15)		(-0.72)
R2	0.490	0.483	0.480	0.470
Block FE	✓	✓	✓	✓
Model FE	✓	✓	✓	✓
Num.Obs.	4800	4000	4800	4000

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

Table C5: Recall sentiment

Dep. Var.	RecallSent		
	(1)	(2)	(3)
ValenceDec	0.0717***	0.0723***	0.0723***
	(4.38)	(4.46)	(4.79)
RecallSentLst	0.3129***	0.3087***	0.3090***
	(4.31)	(4.45)	(4.47)
IsStockLst		0.0513	0.0497
		(1.34)	(1.45)
SubjProbLst		-0.0710	-0.0683
		(-1.04)	(-1.03)
InvPayoffLst		-0.0024	-0.0026*
		(-1.89)	(-2.27)
ConfidLst		0.0339*	0.0343**
		(2.30)	(2.46)
NCharact			0.0008*
			(2.29)
NUniqueChar			-0.0064
			(-1.31)
FOG			-0.0020
			(-0.22)
R2	0.234	0.238	0.240
Block FE	✓	✓	✓
Model FE	✓	✓	✓
Num.Obs.	4000	4000	4000

This table reports the sentiment of subjects' recall. The dependent variable is the sentiment of AI agents' recall after displaying an image, which takes a value of 1 (positive) or -1 (negative). The key independent variable is the decile variable of the valence level of the image cue. In addition to the control variables in 1, 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 C6: Ravenpack sentiment scores and memory model 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 examines the relationship between investment scores and sentiment scores by Ravenpack. The regressions are estimated on subsamples of news where positive and negative memory models disagree, drawn from high-divergence trading days where the return difference is at 95%. Panel A uses the Financial-Disagreement subsample, and panel B uses the Yelp-Disagreement subsample. Independent variables are the investment scores from the positive-memory and negative-memory models.

Appendix D. Proofs

Proof of Theorem 6.1. We analyze the local effect of the cue by considering the first-order Taylor expansion of the valuation $\hat{y}$ around the task query q_{task}. Let the total query be $q = q_{task} + \lambda q_{cue}$. The change in valuation is approximated by:

$$\hat{y}(q) - \hat{y}(q_{task}) \approx \lambda \cdot \langle \nabla_q \hat{y}(q_{task}), q_{cue} \rangle$$

Using the gradient expression from Proposition 1, we have:

$$\langle \nabla_q \hat{y}, q_{cue} \rangle = \beta \sum_{i=1}^{N} \alpha_i (v_i - \hat{y}) \langle k_i, q_{cue} \rangle = \beta \cdot \mathrm{Cov}_\alpha(V, \langle K, q_{cue} \rangle)$$

By the definition of a positive cue, $\mathrm{Cov}_\alpha(V, \langle K, q_{cue} \rangle) > 0$. Since $\beta > 0$ and $\lambda > 0$, the first-order change is strictly positive, implying $\hat{y}(q) > \hat{y}(q_{task})$ for small perturbations. This formalizes the upward bias in valuation. □

Proof of Lemma 6.1. Let the visual and textual queries be q_{img} and q_{txt} respectively. The relevance logit for memory i is given by $L_i(q) = \beta \langle q_{task} + q_{cue}, k_i \rangle = \beta \langle q_{task}, k_i \rangle + \beta \langle q_{cue}, k_i \rangle$.

The contribution of the cue to the logit is $C_i = \beta \langle q_{cue}, k_i \rangle = \beta \|q_{cue}\| \|k_i\| \cos(\theta_i)$. Given the assumption that semantic alignment is identical ($\cos(\theta_{img}) = \cos(\theta_{txt})$) and $\|q_{img}\| > \|q_{txt}\|$, it follows that the magnitude of the cue's contribution is strictly larger for the visual cue: $|C_i^{img}| > |C_i^{txt}|$.

This larger magnitude implies that the visual cue exerts a stronger force in shifting the attention weights α_i away from the distribution determined solely by q_{task}. Specifically, if the cue aligns with high-value memories, the larger norm $\|q_{img}\|$ results in a larger increase in the logits for those memories compared to $\|q_{txt}\|$, thereby inducing a larger shift in the aggregated valuation $\hat{y}$. □

Proof of Theorem 6.2. Let k^* be the optimal task-relevant key such that $\langle q_{task}, k^* \rangle > \langle q_{task}, k_j \rangle$ for all $j \neq *$. Let k_{dist} be any distractor memory activated by the cue. The ratio of attention weights is:

$$\frac{\alpha_{dist}}{\alpha^*} = \exp\left(-\beta(\langle q, k^* \rangle - \langle q, k_{dist} \rangle)\right)$$

Assumption: We assume the cue intensity λ is bounded such that the correct task memory remains the nearest neighbor to the query vector, i.e., $\langle q, k^* \rangle > \langle q, k_{dist} \rangle$.

Under this condition, the term in the exponent is negative. As the reasoning capability parameter $\beta \to \infty$, the ratio $\frac{\alpha_{dist}}{\alpha^*} \to 0$. The attention mechanism converges to a "Hard Max" operation on k^*. Consequently, $\hat{y}(q) \to v^*$, meaning the valuation is determined solely by the optimal task memory, and the bias induced by q_{cue} vanishes. □

Proof of Theorem 6.3. The gradient of the valuation with respect to the query is $\nabla_q \hat{y} = \beta \sum \alpha_i (v_i - \hat{y}) k_i$. Let the "value gradient" vector in the memory space be $\vec{g} = \sum \alpha_i (v_i - \hat{y}) k_i$. The first-order bias is proportional to $\langle \vec{g}, q_{cue} \rangle$. If q_{cue} is orthogonal to the subspace spanned by the value-weighted memory keys (i.e., orthogonal to $\vec{g}$), then $\langle \vec{g}, q_{cue} \rangle = 0$. Thus, locally, the cue induces no change in the agent's valuation. □

Appendix E. Decisions under risks are decisions under complexity even for GAI

We replicate the experiment in a highly controversial research paper (Oprea, 2024) and find striking results supporting the argument. The experiment design closely follows the lottery-mirror setting[26].

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 model most able to solve complex problems.

In this experiment, each subject was asked to complete two main tasks: a "Lottery" task and a "Mirror" task. In both tasks, participants were shown a set of 100 hypothetical boxes, each containing a certain amount of money. For example, a task called "G90" consisted of 90 boxes containing \$25 and 10 boxes containing \$0. Then, we elicit the subjects' valuation for this set of boxes using a "Multiple Price List" (MPL). This method involves presenting subjects with a series of choices where option A means the set of boxes (either as a Lottery or a Mirror) and option B: A simple, certain dollar amount that increases with each row in the list.

By observing at which dollar amount 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 that there is a so-called "simplicity equivalence", and the main difference between the two tasks was the payoff rule: how the set of 100 boxes determined the participant's payment.

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

Mirror (The Deterministic Task): This treatment used the exact same descriptive set of 100 boxes but with a different payoff rule that removed all risk. The payoff was 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", where the simple, certain amount they find equally valuable to the complexly described but deterministic payment. Thus, the core idea of the experiment was to keep the information processing (calculating the expected value) identical, varying only the presence of risk.

We present the main findings in figure E1, where the y-axis denotes deviation from expected value, which represents the subject's valuation (what they said it was worth) minus the true expected value. The x-axis shows the probability of the non-zero outcome (e.g., 0.1 for 10% probability, 0.9 for 90% probability).

[26] The replication package is also available upon request. A polished note will be released based on this set of results in the future. The author thanks Thomas Graeber for helpful comments.

We document the "fourfold pattern" as Oprea (2024) for both the lottery tasks and the mirror tasks, and the valuation of these two different tasks aligns closely for the three models.

Importantly, we document that, as the models' reasoning ability becomes stronger, the fourfold pattern starts to diminish. For the o1 model, it always makes perfectly Bayesian choices under every single task, contrasting GPT 4o (baseline) and GPT 4o (CoT).

[Insert Figure E1 near here]

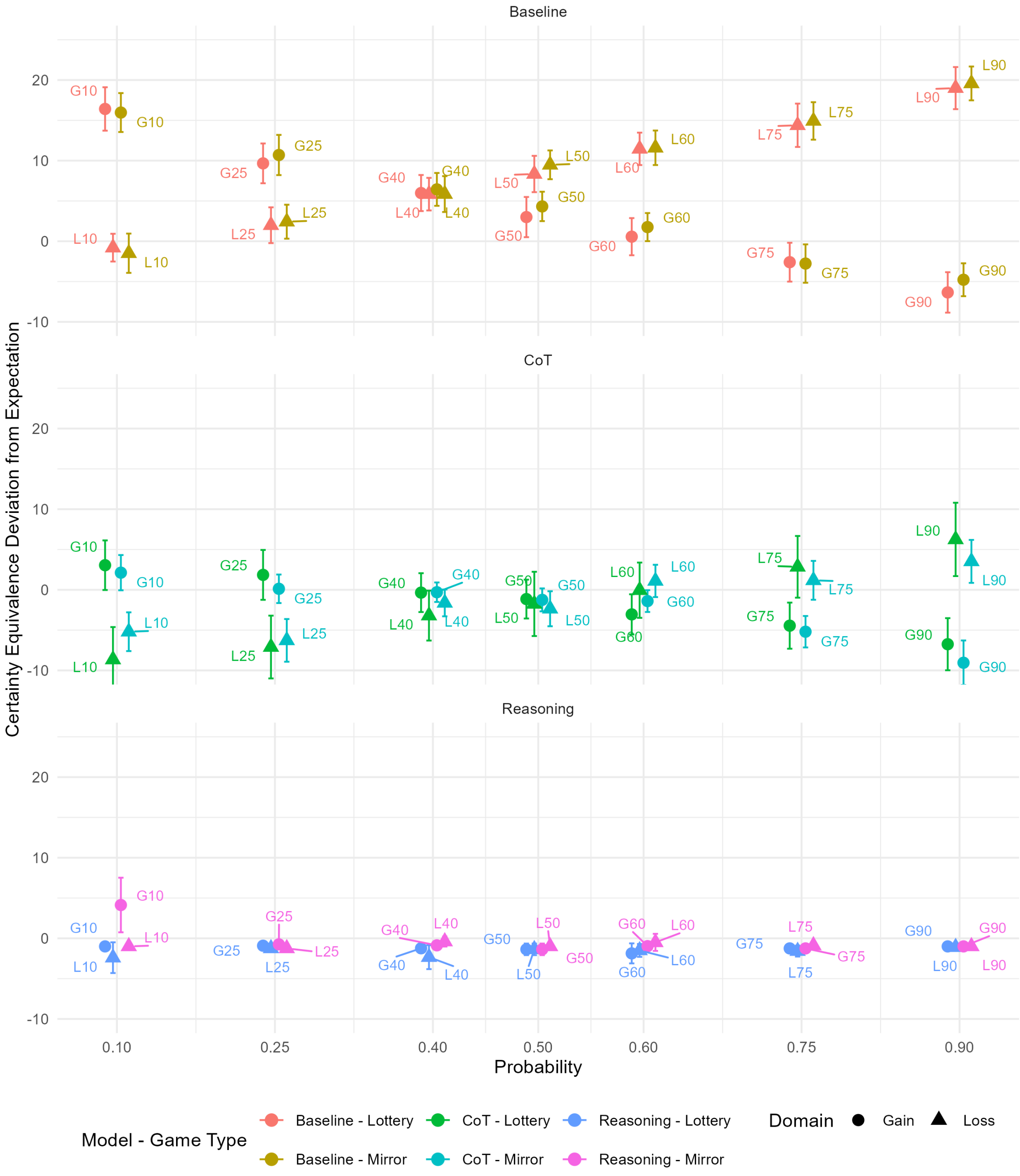


Fig. E1. Mean deviaion from expected value with different reasoning abilities. This image replicates the main results in Oprea (2024), where the subjects are GPT-4o, GPT-4o augmented with Chain-of-Thoughts, and o1 reasoning model. The mirror and lottery tasks are displayed in the figures, respectively.