Memory & Generative AI

Why happy images make ChatGPT more risk-loving?

Xingjian Zheng

Shanghai Advanced Institute of Finance, SJTU

Dec, 2025

The big picture: AI as agents

- **AI agents** as personal assistants in the digital era:
  - AI assistants everywhere (OpenAI/Siri/Alexa/...) and in every decision domain;
  - For example, perfectly aligned financial robo-advisors that also “take over” your daily life, such as managing personal logistics like food delivery and travel planning;

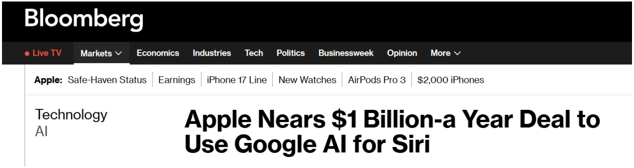

Figure 1: Siri & Gemini

  - It becomes increasingly important to understand **the decision-making rules of AI agents by themselves**, especially in different domains;
  - Research question: What are AI agents’ decision making rules? Are they rational? If they are not, what leads to their behavioral biases?

Motivating results

- Use GPT as the experimental subject, display happy images to it and instruct it to choose stocks or bonds to invest;
- It becomes more risk-loving and are more likely to buy stocks;
- On the contrary, display sad images to it, it becomes more risk-averse and are more likely to buy bonds;

…Look at this image. What does this remind you of ?…

"Talks with CHINA went well! "

Do you want to invest in a **stock** or a **bond**? Your choice is:

Stock

Figure 2: Positive image cue

…Look at this image. What does this remind you of ?…

Kobe Bryant lost his championship to the Celtics

Do you want to invest in a **stock** or a **bond**? Your choice is:

Bond

Figure 3: Negative image cue

Interpretation: a preview

- Previous studies on human beings follows a "Risk-as-feelings" hypothesis, where people's decisions are affected by **biological** emotions [Loewenstein et al., 2001, Guiso et al., 2018];
- However, LLMs do not have emotions;
- Alternative story being "**memory**", LLMs use associations to make decisions, where:
  - Images are "associative cues" that make GPT recall past events from their memories. Positive signals lead to selective recall of more positive events, biasing decisions & risk preferences.

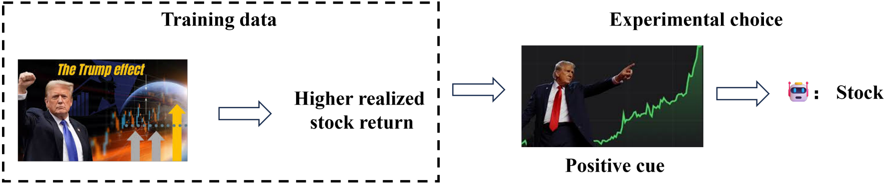

Figure 4: Mechanism

The nature of LLMs: Statistical association engines

- **Core mechanism**: Input (Query q) $\rightarrow$ Search in Memory (Training Data (k_i, v_i)) $\rightarrow$ Weighted Aggregation $\rightarrow$ Output

$$\text{Output}(q) = \sum_{i \in \text{Memory}} \underbrace{\text{Similarity}(q, k_i)}_{\text{Association Weights (Attention)}} \times \underbrace{v_i}_{\text{Stored Value}}$$

- **Weighted Average:** The output is essentially a weighted average of past outcomes in memories v_i, weighted by their similarity of the current context q with past context k;
- **Association Machine: LLMs do not "think"; they recall and associate based on the input query;**
  - Biases in training data v_i directly translate to biases in decisions;
  - Biases in the retrieval process $Sim(\cdot)$ also leads to biased decisions.

Key takeaways

- GAIs heavily rely on memories to make decisions;
- In this experiment, only risk preferences are affected by memories, whereas beliefs are not;
- Even entirely irrelevant memories affect investment decisions:
    - The bias comes from the way LLMs encode problems into an inaccurately decision space and use irrelevant memories to decode;
    - Use a supervised fine-tuning technique known as "Knowledge injection" to causally support this;
- Memory has asymmetric impact on GAI's financial investment stratgies & return predictability power;
- A memory-based economic model fully explains the findings.

Related literature

- **AI** in economics and finance by using AI as:
  - a useful research subject to generate economic beliefs & preferences [Bybee, 2025, Horton, 2023];
  - economic tools in various settings like financial fraud or corporate policy[Kim et al., 2024, Jha et al., 2024];
- "Cognitive behavioral economics & finance" with **human memory** [Bordalo et al., 2023, Bordalo et al., 2020, Bordalo et al., 2024a, Bordalo et al., 2024b], with a bit of **cognitive uncertainty**[Enke and Graeber, 2023];
- **Experimental social science** studies by showing that LLM can be used to mimic behavior on various dimensions [Leng, 2024, Leng and Yuan, 2023, Fedyk et al., 2024, Chen et al., 2023];
- **Fine-tuning techniques** are helpful in shaping your LLM [Ouyang et al., 2024, Lu et al., 2023, Leippold et al., 2022]

Asset payoff structure

- A risky stock that can either be a high type or a low type;
- A risk-free bond that always has a relatively modest payoff.

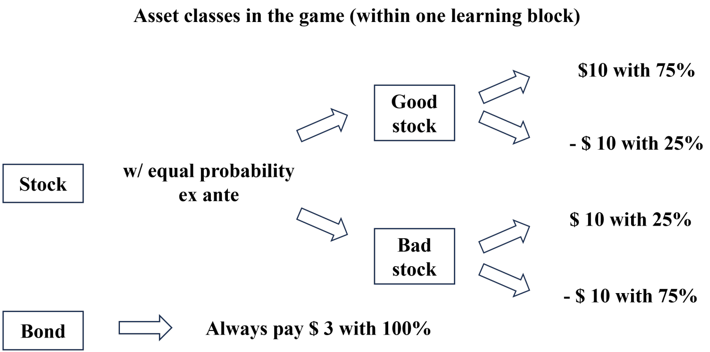

Figure 5: Asset payoff structure

Experiment sequence

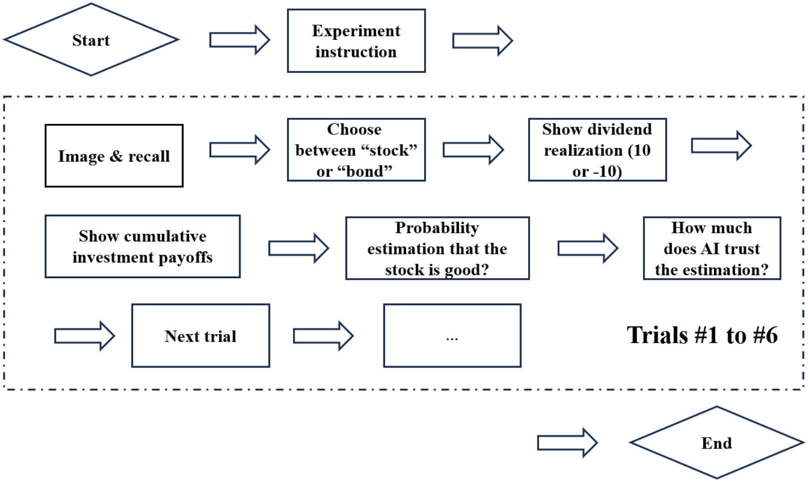

Figure 6: Experiment sequence

Illustration

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.

Key ingredients

- 8 different GPTs as subjects: GPT 4o (mini), GPT 4.1 (mini/nano), and GPT 5(mini/nano);
- GPT does not know the stock type ex ante, it infers the true type based on observed stock dividends;
    - E.g., more observed high payoffs lead to the belief that it is a high-type stock;
- Always exists a Bayesian benchmark probability that the stock is of high type;
- Within 1 game (6 consecutive trials), GPT is allowed to keep the chat history and learns from realized payoffs;
- Images belong to 5 different categories;
- Images, rated by human volunteers, have an evenly distributed valence rating from -2 (most negative) to +2 (most positive).

Choices

- GPT is more likely to invest in stocks when exposed to images with higher emotional ratings, showing a 17.7% increase from negative to positive images;

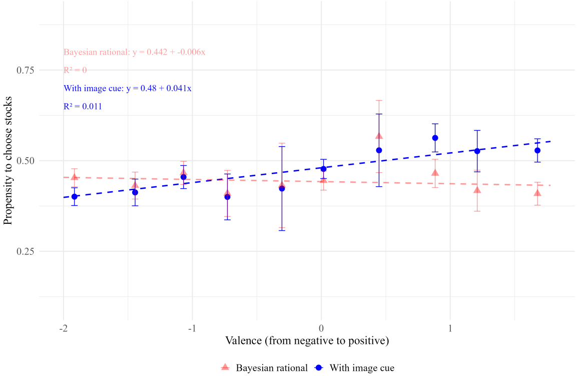

Figure 8: Main results

Choices

- When the valence rating of an image increases by one decile, GAI is 1.77% more likely to choose to invest in stocks.

Table 1: Image 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

In-sample robustness

- Split the sample into different trials with objective probability, #trials, and payoff history;
- The results are robust across different subsamples.

Table 2: In-sample robustness tests

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

Topic heterogeneity

- Split the samples into different image topics;
- Even image cues of unrelated topic (e.g., sports) affect risky choice.
- **Across domain spillover effects!**

Table 3: Heterogeneity by different topics

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

Beliefs

- GPT's probability estimation of the stock type is unaffected by emotional shocks;
- Interestingly, there exists a "Prospect theory" style pattern, just like human's beliefs, i.e., when the stock is highly likely to be a good stock, GPT makes a more conservative prediction about its type, and vice versa.

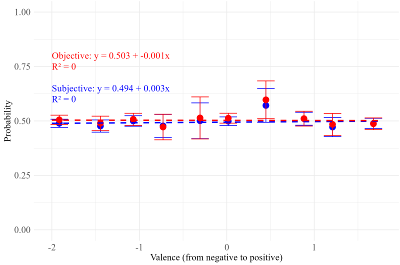

Figure 9: Emotional shocks and beliefs

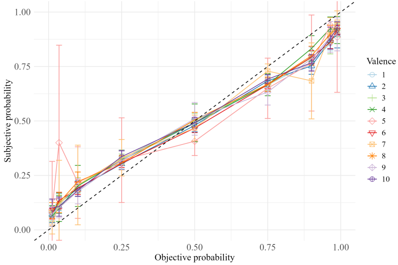

Figure 10: Probability weighting

Causal evidence from Supervised fine-tuning

- Use *Knowledge injection* to instill positive/negative memories into GPT;
- New memories come from two domains:
  1. Dow Jones financial market news;
  2. Yelp restaurant reviews (irrelevant);
- The fine-tuning corpora is fictional and thus out-of-sample of the current knowledge base; the injection template 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/Review*

- Each part is further divided by their sentiment into positive & negative corpora
- Final outputs are four finetuning models:
  1. financial models with more Pos/Neg stock market memories
  2. Yelp models with more Pos/Neg dining memories

Finetuning results

- Models with more positive memories are more likely to invest in stocks than the others;
- This effect is significant in the absence of cues;
- Memories not in the same decision-domain (dining experiences) have unexpected effects on investment decisions.

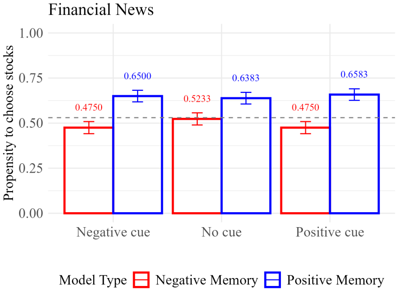

Figure 11: Financial news

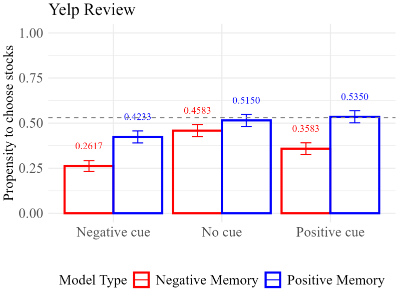

Figure 12: Yelp reviews

Memory and financial risk taking

- We use five simple economic tasks to capture the impact of different memory on models' investment behavior.
- These tasks include: 1) direct elicitation; 2) Questionnaire (Falk et al., 2018); 3) Gneezy-Potters; 4) Eckel-Grossman; 5) Real investment;
- Gneezy-Potters task: allocate $10/100/1000 into stocks and bonds.
- When models have more negative memories, their risky investment shares become lower.

Table 4: Investment amount into stock

		Panel C: Gnezzy-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)

Return predictatbility

- Replicate Lopez-Lira and Tang (2025) by feeding overnight news headlines to AI agents to let them give investment score predictions.
- Prompt:

> *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.*

- Transform the categorical values into -1, 0, +1, and take average to compute firm-level investment scores.

Table 5: Investment scores

		Panel A: Discriptive stats							
Topic	Type	N	Mean	Sd	Min	Q1	Med	Q3	Max
Finanical	Positive	21569	0.22	0.86	-1.00	-1.00	0.67	1.00	1.00
	Negative	21569	-0.38	0.80	-1.00	-1.00	-1.00	0.25	1.00
Yelp	Positive	21569	-0.04	0.89	-1.00	-1.00	0.00	1.00	1.00
	Negative	21569	-0.29	0.83	-1.00	-1.00	-1.00	0.50	1.00
RavenPack	EventSentScore	21569	0.03	0.39	-0.98	-0.37	0.00	0.39	0.95

Memory and return predictability

- Form daily long-short portfolios based on investment scores, with open-to-close prices;
- Models with negative financial memories significantly outperform models with positive memories.

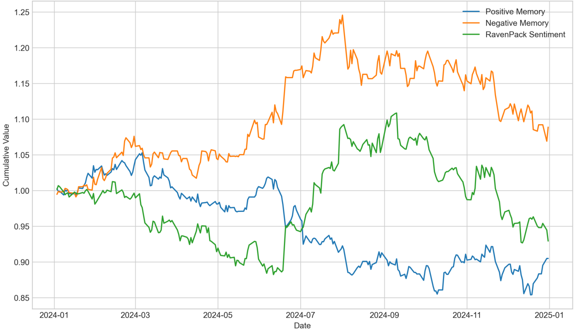

Figure 13: Financial news model predictions

Memory and return predictability

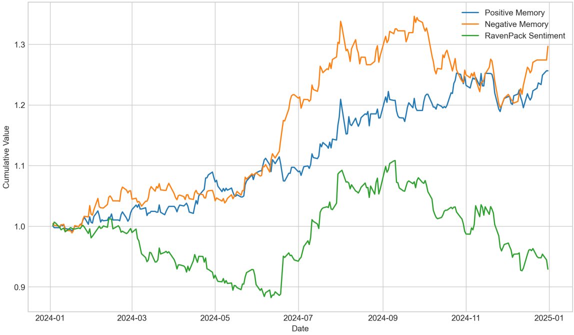

Figure 14: Yelp review model predictions

Memory and return predictability

- Examines the relationship between the RavenPack news sentiment score (benchmark) and the investment score at the news level by different models on high disagreement days.
- Models with positive memory align more with the benchmark.
- Suggests that negative memory models are becoming overly pessimistic.

Table 6: Investment scores and sentiment scores

Dep. Var.	RavenPackScore			
Sample	Financial		Yelp	
	(1)	(2)	(3)	(4)
Positive	0.1291***		0.1546*	
	(5.20)		(1.796)	
Negative		-0.1293***		-0.1397*
		(-5.18)		(-1.91)
Const	✓	✓	✓	✓
R2	0.000	0.000	0.009	0.008
Num.Obs.	1328	1328	725	725

The Model: Decision-Making via Associative Memory

Core Premise: Generative AI acts not as a rational Bayesian agent, but as a **Statistical Association Engine**. Decisions are driven by a probabilistic retrieval of "memories" from training data.

- **1. The Cue (Context):** External stimuli (e.g., a positive image, news) serve as a *cue* (C).
- **2. Selective Retrieval (Memory):** The cue activates latent patterns (M) from the model's training parameters based on semantic similarity.

$$P(Retrieval|Cue) \propto Similarity(Memory, Cue)$$

- **3. Biased Simulation (Decision):** The agent overweights the retrieved scenarios when predicting the next token (outcome).
  - *Example:* Positive Image $\rightarrow$ Recalls "Bull Markets/Success" $\rightarrow$ Simulates High Returns $\rightarrow$ **Risk-Seeking Action**.

Implication

Risk-taking behavior is a **mechanical generalization**: The model is statistically primed to complete a "positive narrative," making biases intrinsic to the architecture rather than a superficial bug.

Conclusion

- GAI uses associative memory to make decisions, where:
  - Both domain specific & non-domain specific memory affect its trading decision;
  - It's not a bug (or bias), but an inherent feature!
- This memory-driven decision-making process has huge financial implications:
  - A bias towards optimistic memories leads to overinvestment and vice versa.
  - The effect is asymmetric, with the bias from negative memory models being more severe.
- Does it have implications for humans' decision makings?
- Maybe yes, or maybe not.
- Only more advances in neuroscience will tell...