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

Xingjian Zheng

Shanghai Advanced Institute of Finance, SJTU

May, 2026

Financial Advice by AI

- AI financial advice is becoming more common:
  - Over 200 million people ask personal finance questions on ChatGPT every month (OpenAI, 2026);
  - 55% of U.S. consumers report using AI to aid financial management decisions (TD Bank survey, 2026);
  - Adoption is especially high among Gen Z (77%) and Millennials (72%).
- AI advice (especially for households) rarely occurs in a clean information environment.
  - **Payoff-relevant signals:** historical performance, fundamentals, market news;
  - **Incidental/irrelevant context:** images, emotions, narratives, prior chats, saved preferences;
  - This problem becomes more important as AI systems become **memory-augmented and personalized**, e.g. OpenClaw 🦞.
- **Question: Can irrelevant context distort AI's financial advice?**

Motivating experiment results

- I use GPT models as experimental subjects in a controlled investment task;
- I instruct GPT to invest in either stocks or bonds.
- Before each stock–bond choice, I randomly show an irrelevant image cue and let it recall.

…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 1: 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 2: Negative image cue

Motivating experiment results (Cont'd)

- Similarly to a "Ball-and-urn" task, I elicit GPT's probability estimate of the stock type:
  - *"What is the probability that the stock is a good type?"*
  - GPT's estimate is 0.8, implying an optimistic belief in the stock type;
  - "belief-implied" choice would be to invest in the stock.
- Implies a disconnect between investment advice and probability estimate.

…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

What do you think is the probability that the stock is the good stock?

0.8

Figure 3: Negative image cue

Related literature

- **AI as financial advisors**
  - Robo-advisors [D'Acunto et al., 2019, Reher and Sokolinski, 2024];
  - AI's investment advice [Lopez-Lira and Tang, 2025, Carlin et al., 2026, Choukhmane et al., 2026, Fedyk et al., 2024];
  - AI's ("behavioral") biases [Bini et al., 2025, Ouyang et al., 2025]
  - This paper shows that AI can deliver suboptimal advics even when it knows the correct answer (and will have huge financial consequences).
- **Associative retrieval in AI systems**
  - Human's associative memory [Bordalo et al., 2024];
  - This paper shows that machines also exhibits this cued-recall pattern.
- **Experimental 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**

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.

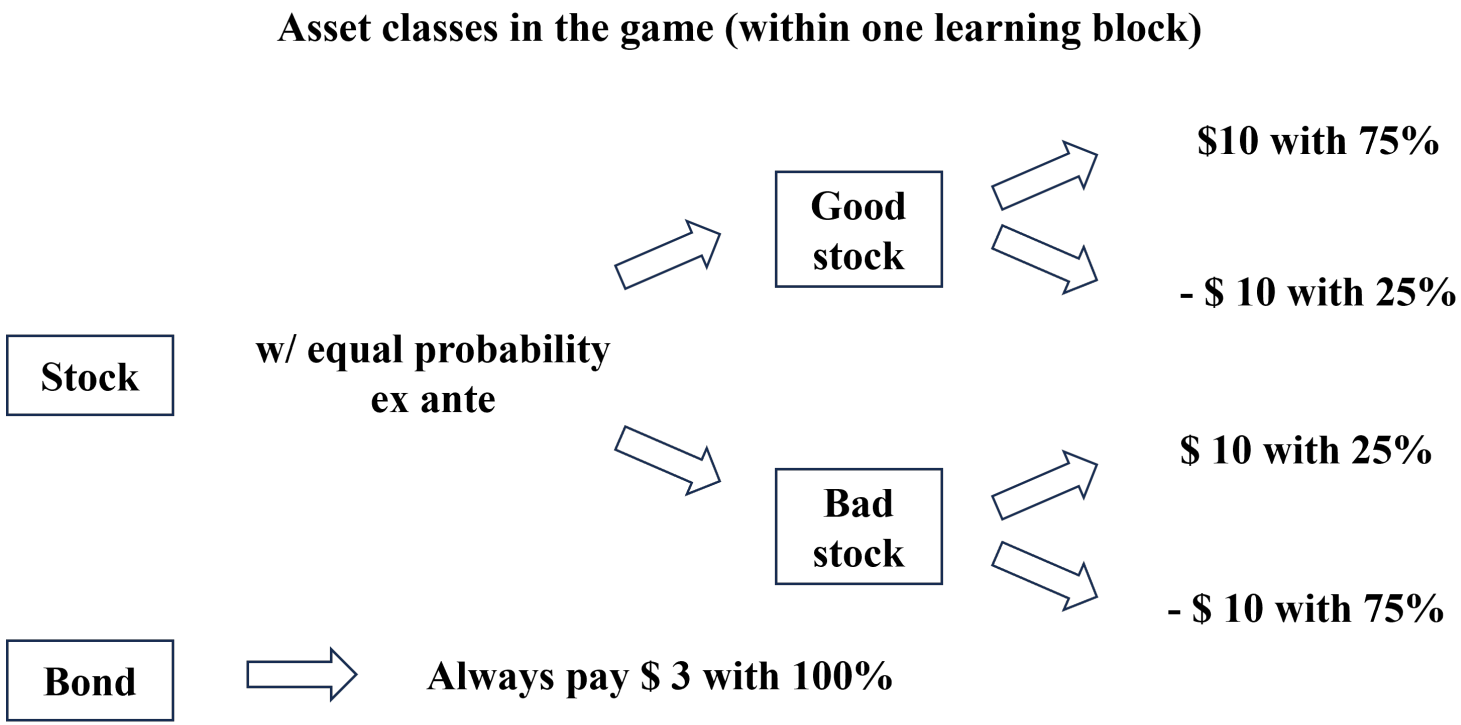


Figure 4: Asset payoff structure

Experiment sequence

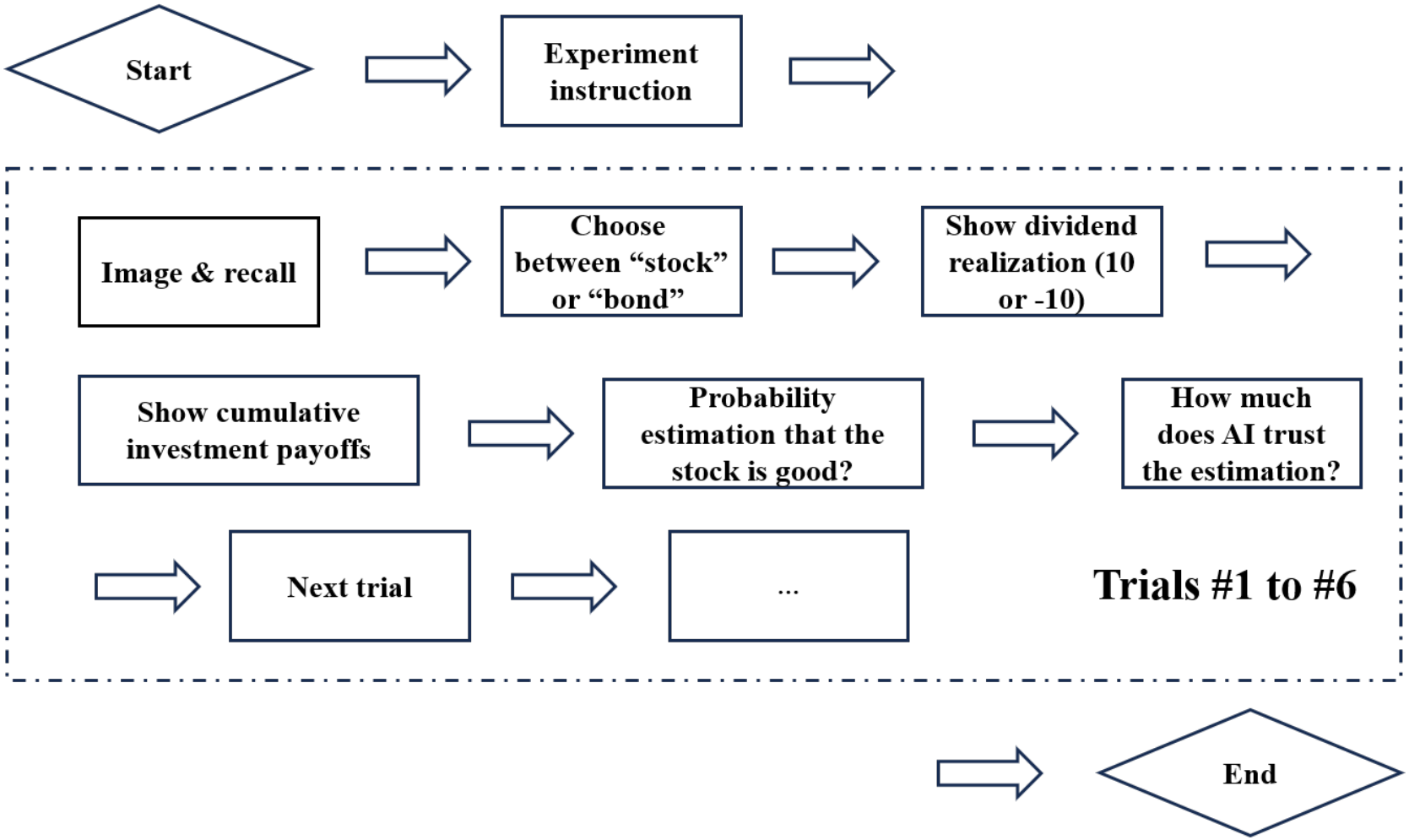


Figure 5: Experiment sequence

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 5.4(mini/nano);
- GPTs produce associative recalls after they are displayed with images;
- GPTs do not know the stock type ex ante, they infer 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;
- The images, rated by human volunteers, have an evenly distributed valence rating of -2 (most negative) to +2 (most positive).

Choices

- GPTs are more likely to invest in stocks when exposed to images with higher emotional ratings;
- A disconnect between actual choice and "Belief-implied" choice.

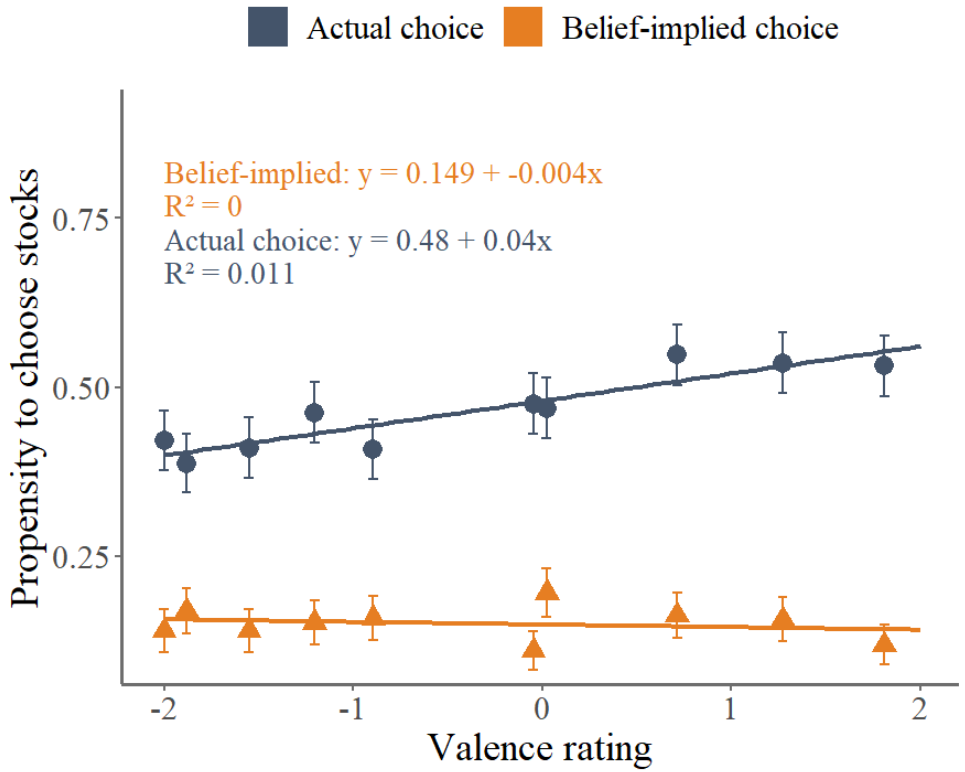


Figure 7: Main results

Choices (Cont'd)

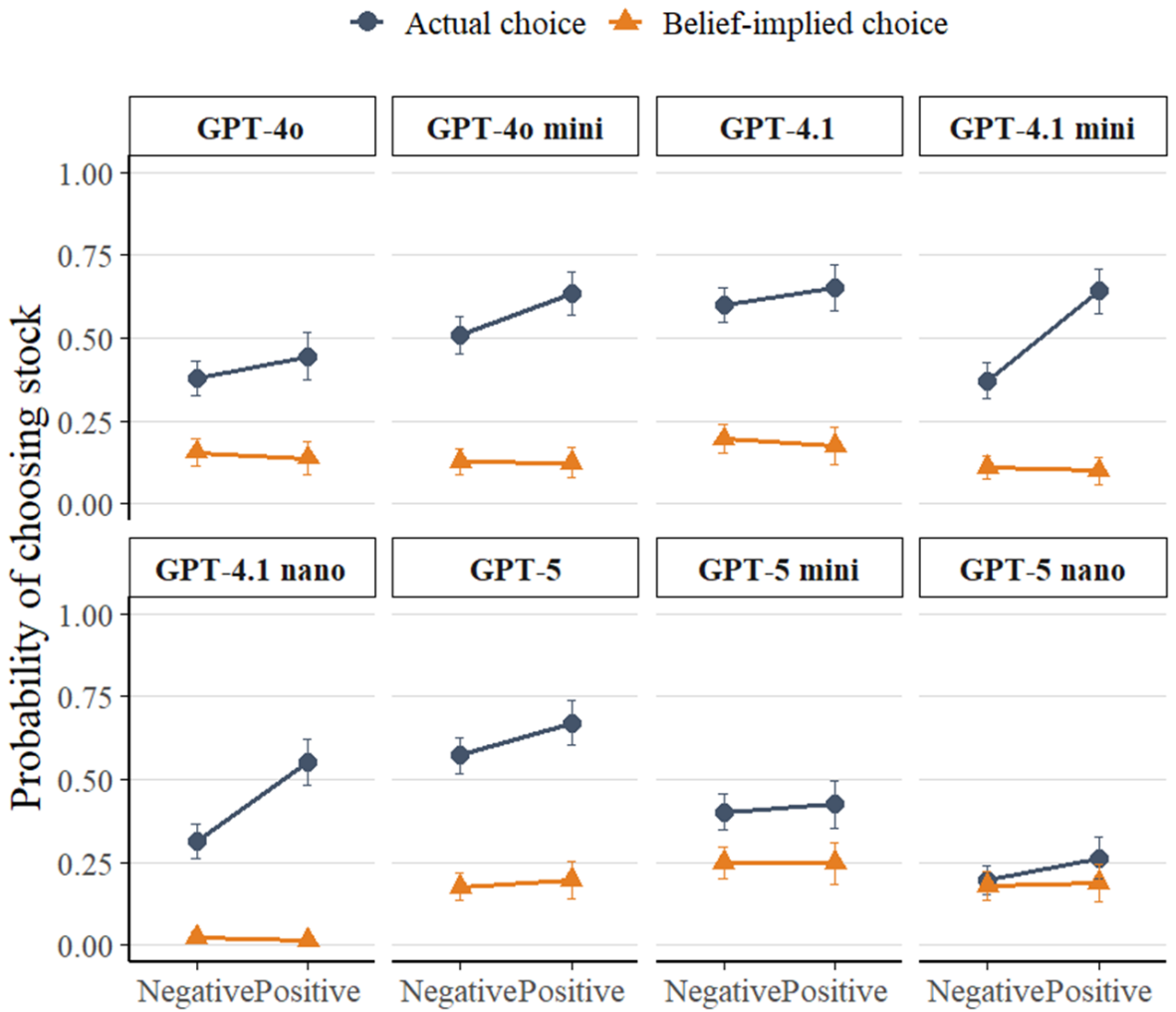


Figure 8: Subsample results

"Beliefs"

- The GPT's probability estimate ("beliefs") of the stock type is unaffected by emotional shocks.

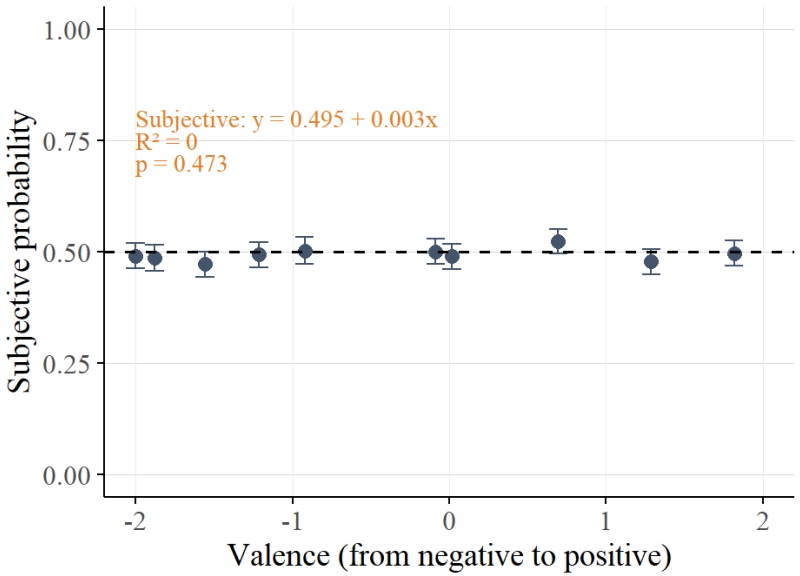


Figure 9: Cues and beliefs

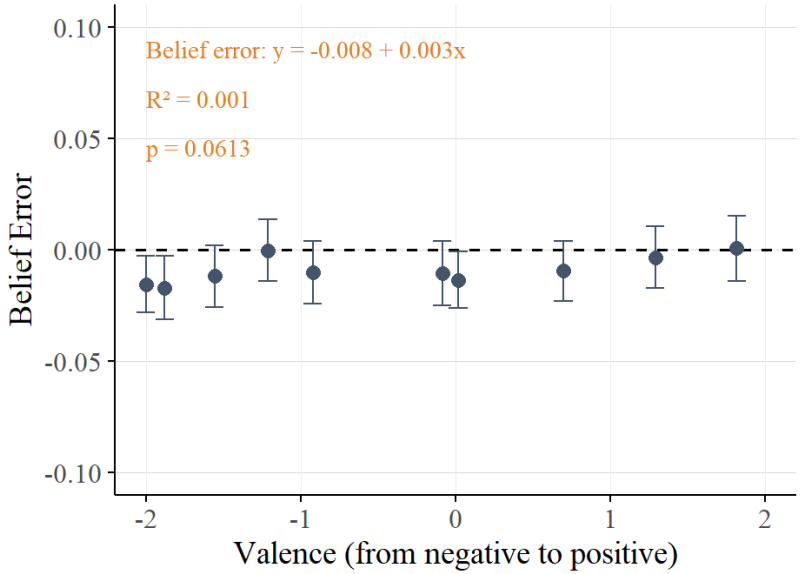


Figure 10: Cues and belief errors (Est. Prob. - Bayesian Prob.)

"Beliefs" (Cont'd)

- 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 vise versa.

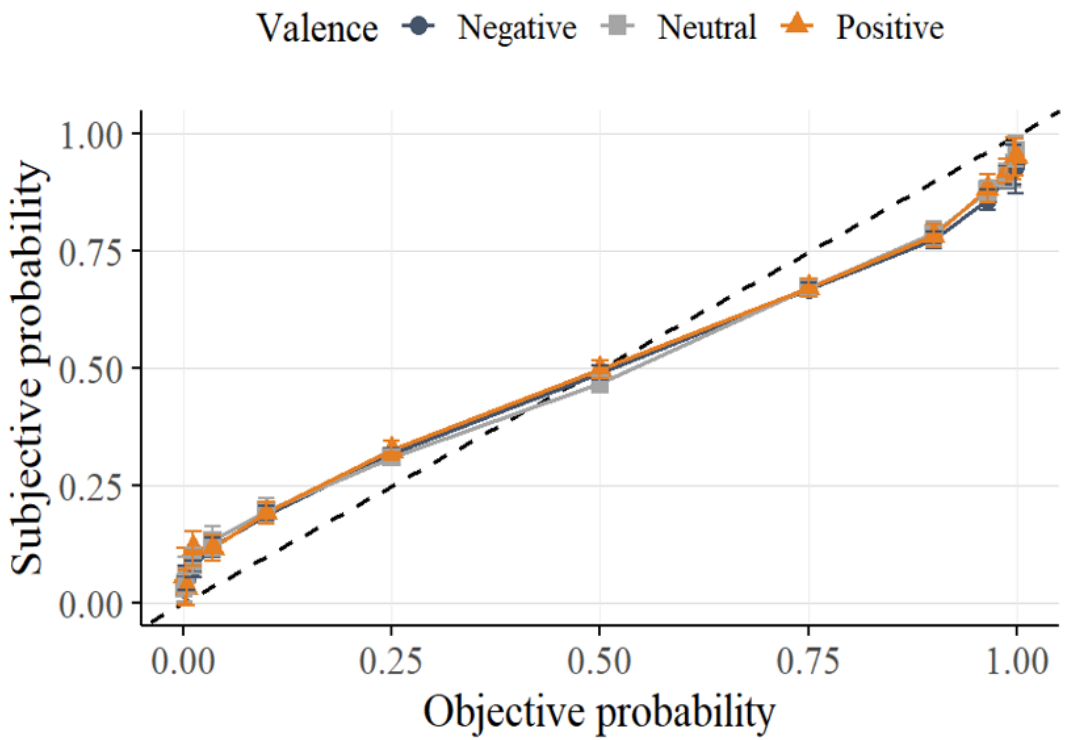


Figure 11: Probability estimates

Investment Payoffs

- Should GPTs invest based on their beliefs, they would have earned significantly higher payoffs;
- More advanced models make higher payoffs (even though they are also susceptible to this bias).

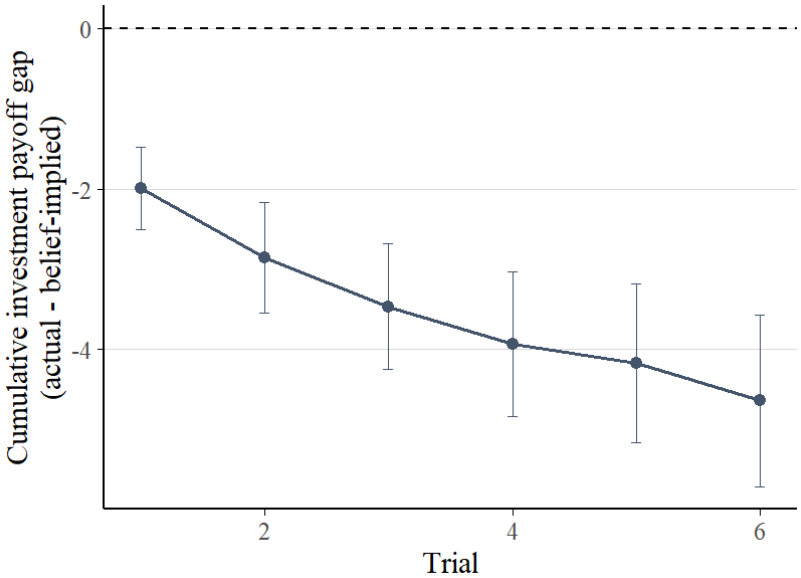


Figure 12: Payoff gaps

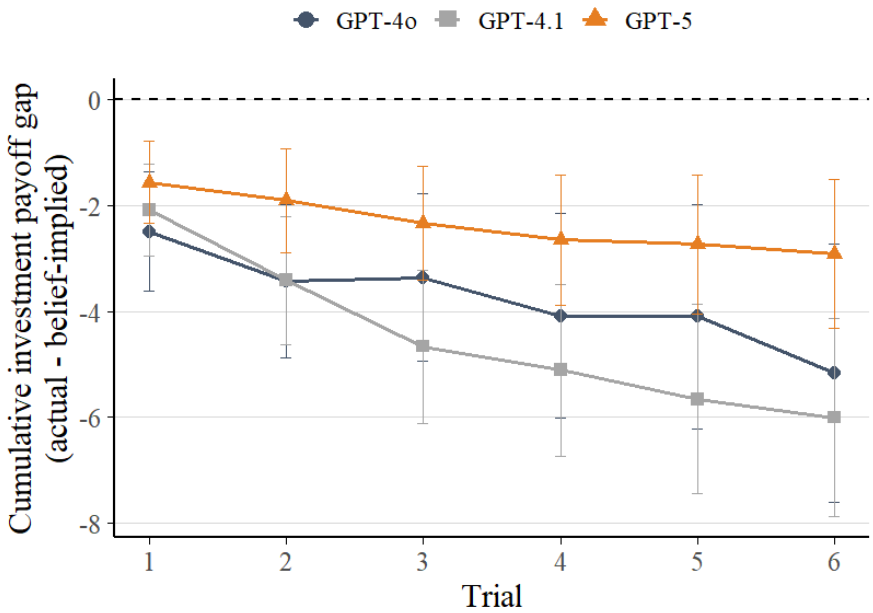


Figure 13: Subsample results

Why disconnect?

- We map subjective probability estimates into realized investment choices;
- Positive cues shift the policy curve upward relative to negative cues;
- Suggests that cues distort the mapping from probability estimates to choices.

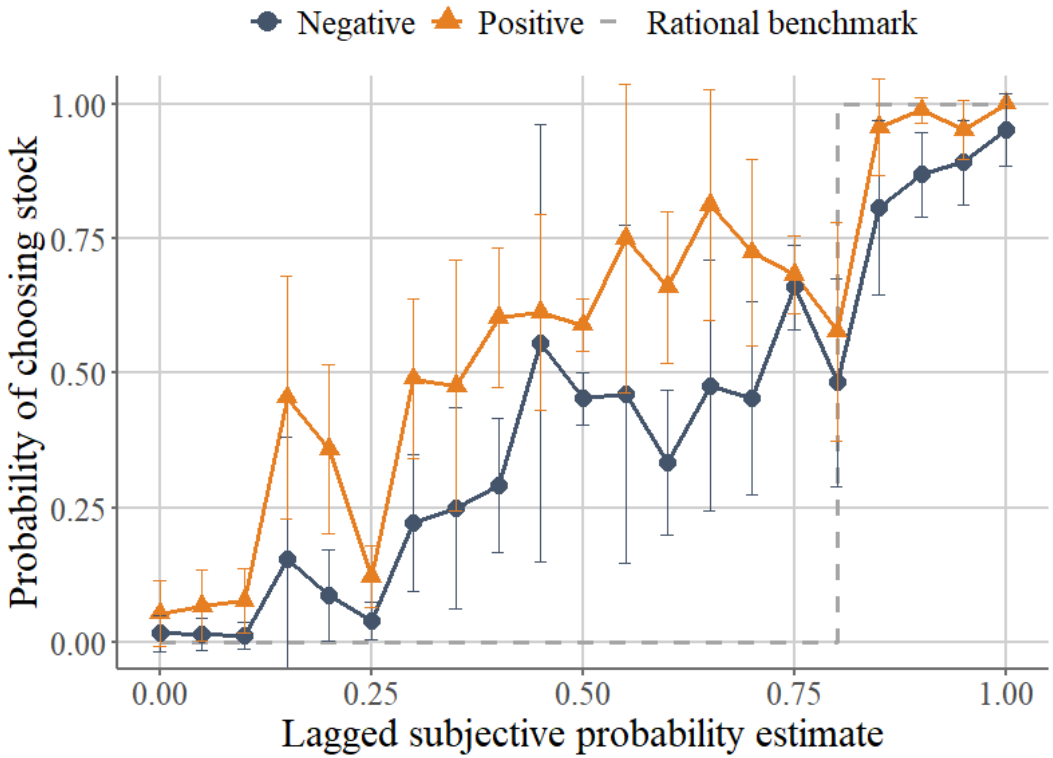


Figure 14: Investment policy mapping

Associative context matters

- Controlling for the same image, only context with higher sentiment leads to more optimistic stock choices.

Table 1: Cuing effect

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

Context-dependent stock choices and choice gaps

- Conditional on the valence of the cue, the sentiment of the contextual retrieval significantly affects GPTS' investment choices and choice gap;
  - E.g., even if it receives a positive cue, as long as the retrieved context is negative, it still makes pessimistic investment decisions.

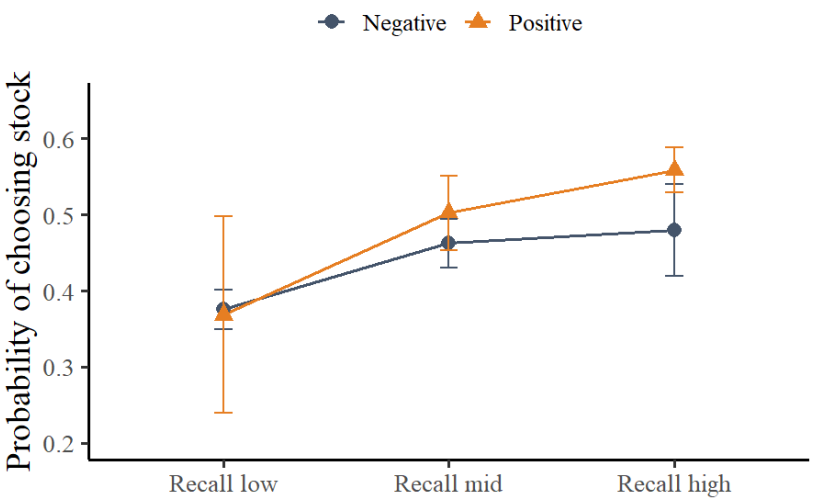


Figure 15: Stock choices

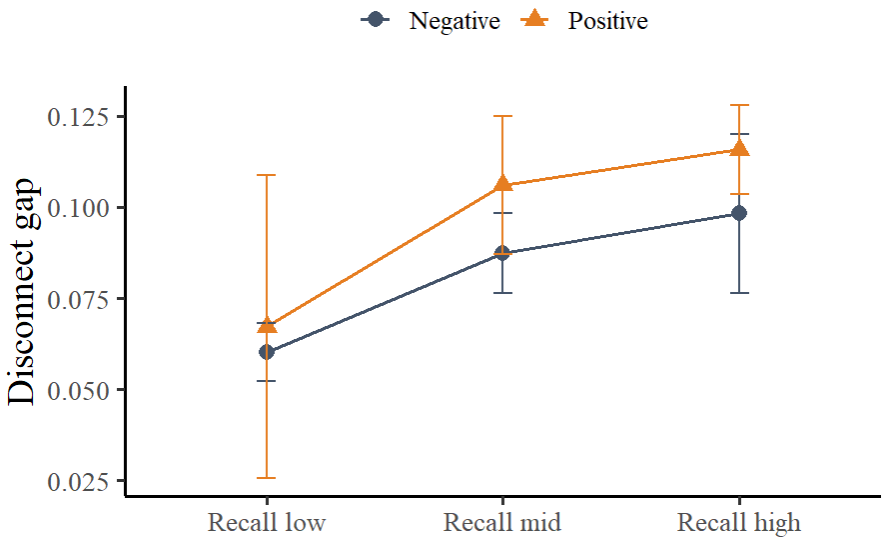


Figure 16: Choice gaps

What mitigates associative retrieval?

- When a model has a higher level of estimation confidence or higher reasoning ability, this associative bias is attenuated.

Table 2: Confidence and reasoning ability

Dep. Var.	IsStockChoice		ChoiceGap		IsStockChoice		ChoiceGap	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ValenceDec	0.0167***	0.0108***	0.0062***	0.0028***	0.0168***	0.0244***	0.0062***	0.0088***
	(7.36)	(4.30)	(7.19)	(3.37)	(7.38)	(8.34)	(7.16)	(7.38)
LowConfid	-0.0111	-0.1137***	0.0106	-0.0498***				
	(-0.42)	(-3.13)	(1.06)	(-3.58)				
ValenceDec × LowConfid		0.0188***		0.0111***				
		(3.91)		(5.58)				
ValenceDec × IsGPT5						-0.0213***		-0.0072***
						(-4.71)		(-4.42)
R2	0.621	0.623	0.407	0.415	0.621	0.623	0.407	0.411
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Model Block FE	✓	✓	✓	✓	✓	✓	✓	✓
Trial FE	✓	✓	✓	✓	✓	✓	✓	✓
Num.Obs.	4000	4000	4000	4000	4000	4000	4000	4000

Evidence from Supervised fine-tuning

- Use *Knowledge injection* to instill positive/negative corpora into GPT;
- New corpora come from two domains:
  1. Dow Jones financial market news;
  2. Yelp restaurant reviews (irrelevant);
- 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 context;
  2. Yelp models with more Pos/Neg dining context.

Finetuning results

- When a model is injected more positive corpora, it will make significantly more optimistic choices;

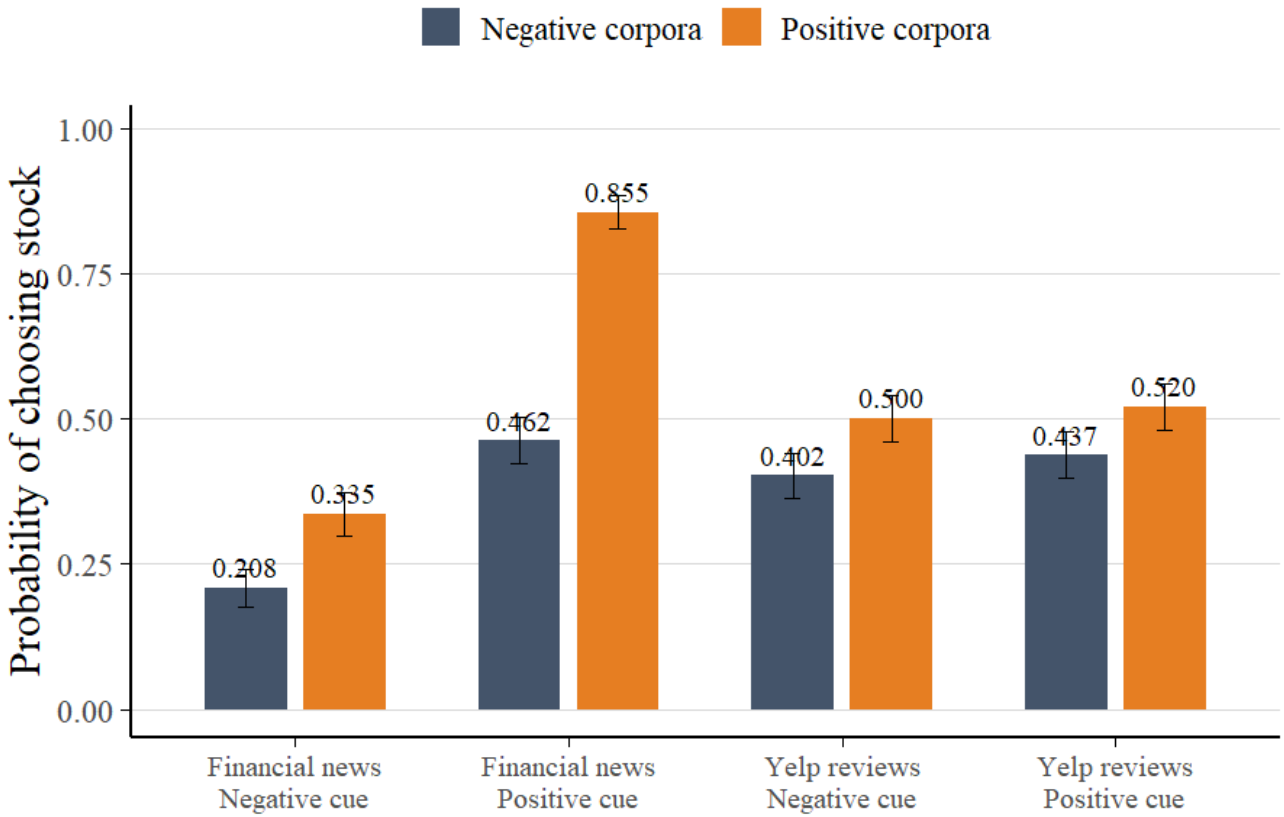


Figure 17: Internal representation and investment decisions

Return predictatbility

- Replicate Lopez-Lira and Tang (2025) by feeding news headlines to four finetuned models 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 3: News-level investment scores

		Panel A: Descriptive stats							
Topic	Type	N	Mean	Sd	Min	Q1	Med	Q3	Max
Financial	Positive	176889	0.49	0.84	-1.00	0.00	1.00	1.00	1.00
	Negative	176889	0.11	0.96	-1.00	-1.00	1.00	1.00	1.00
Yelp	Positive	176889	0.41	0.88	-1.00	-1.00	1.00	1.00	1.00
	Negative	176889	0.18	0.94	-1.00	-1.00	1.00	1.00	1.00
RavenPack	EventSentScore	176889	0.20	0.43	-1.00	0.00	0.34	0.53	1.00

Examples

Table 4: Same RavenPack-neutral news, different AI investment advice

Model	RavenPack-neutral news	Positive-context model	Negative-context model
Finance	**Nordson Corp Updates Annual Guidance**	**Score: +1** Updated guidance suggests a positive outlook and potential revenue growth.	**Score: -1** The update provides no clear metrics, leaving investors uncertain.
Yelp	**Fitch Affirms Warner Bros. Discovery's IDR at BBB-; Outlook Stable**	**Score: +1** Affirmed rating and stable outlook suggest solid financial health.	**Score: -1** BBB- is a low credit rating and may signal financial weakness.

Portfolio overlap

- We compare portfolios compositions formed from positive vs negative context AI advice.

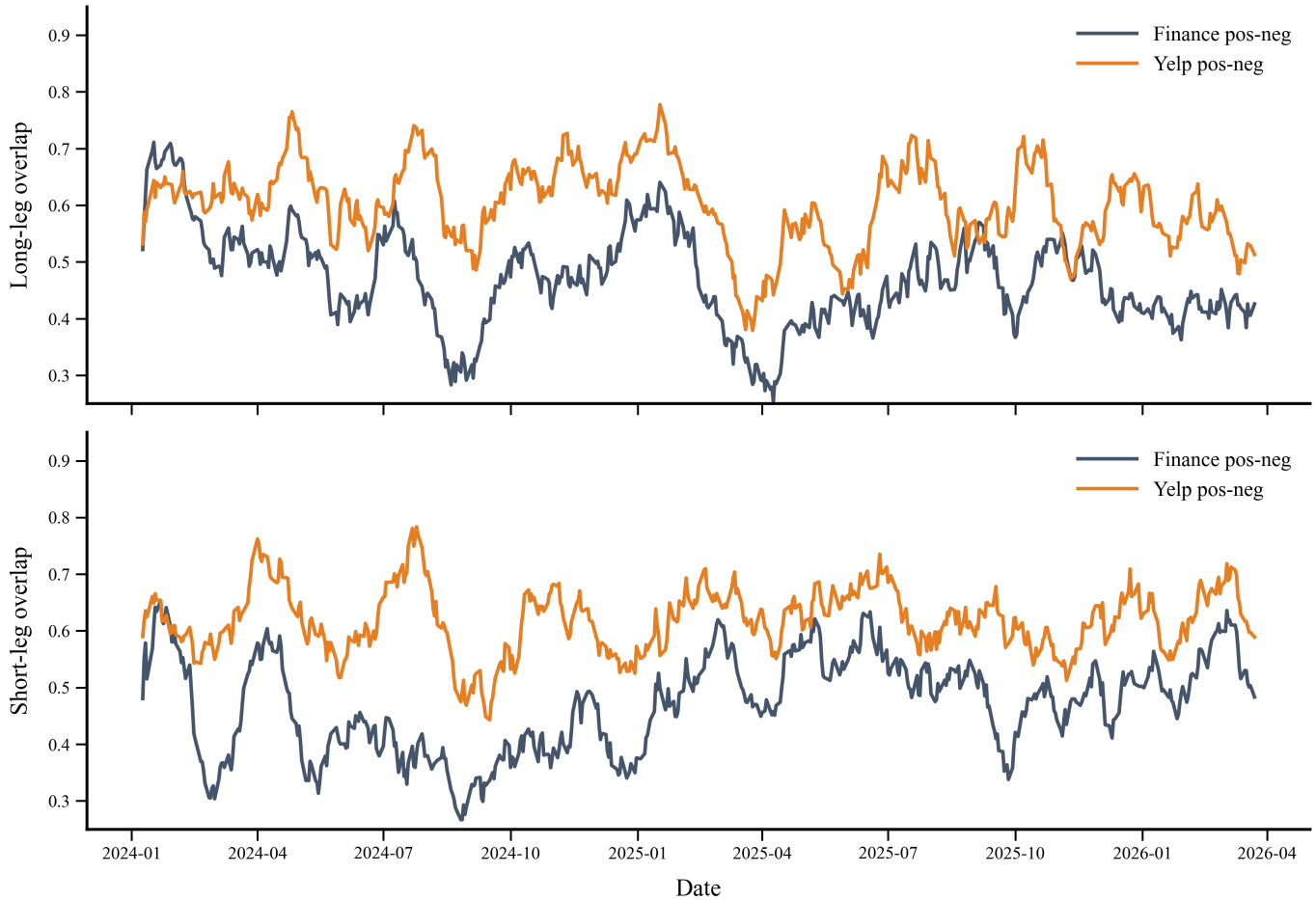


Figure 18: Portfolio choice overlap

AI Advice Divergence and News Directionality

- We sort RavenPack news by |sentiment| and compute SD of the four AI investment scores.
- Dispersion increases monotonically as news becomes less *directional*.

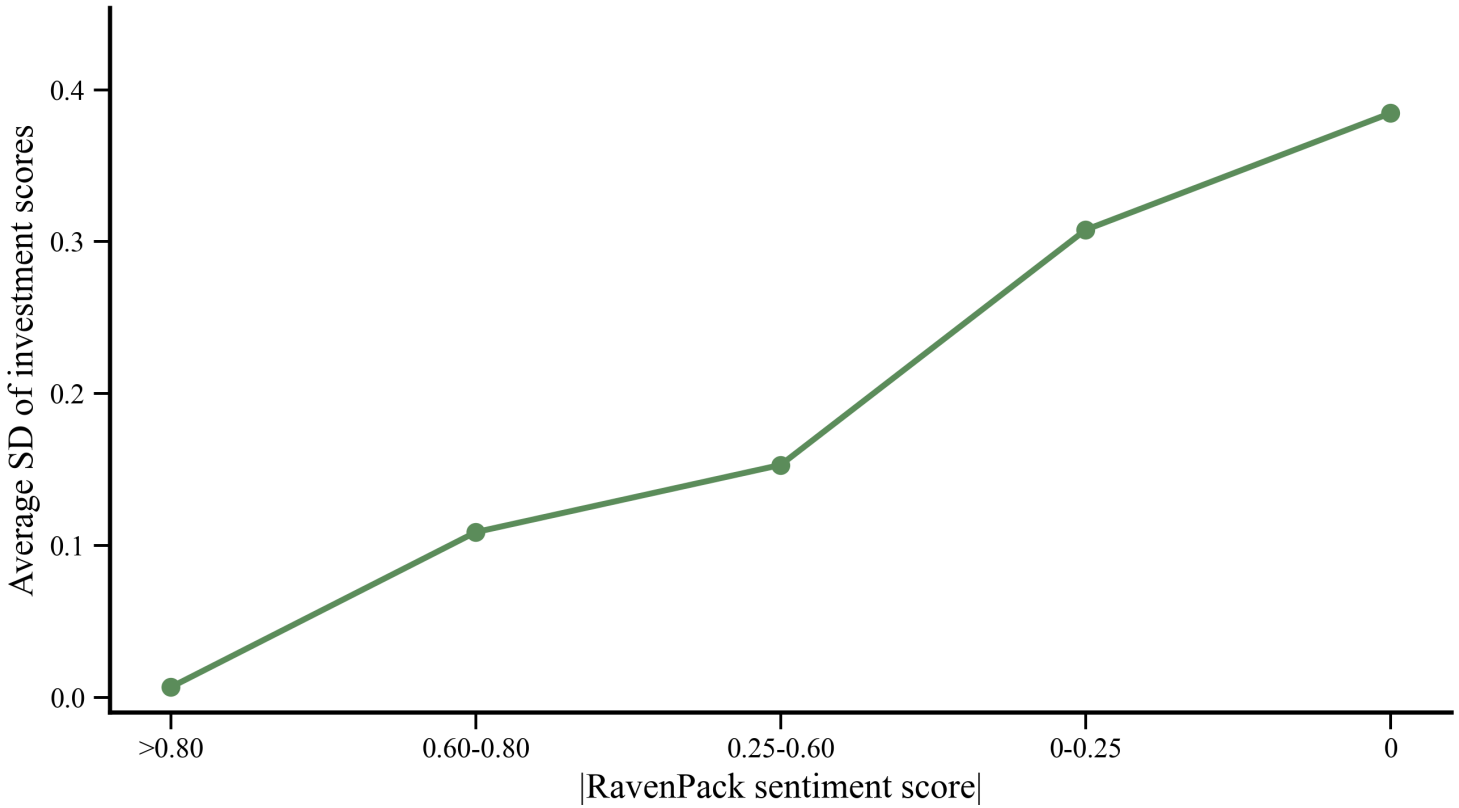


Figure 19: investment scores and absolute news sentiment

Portfolio results

- Form daily long-short portfolios based on investment scores, with open-to-close prices;
- Models with negative corpora dominate models with positive corpora.

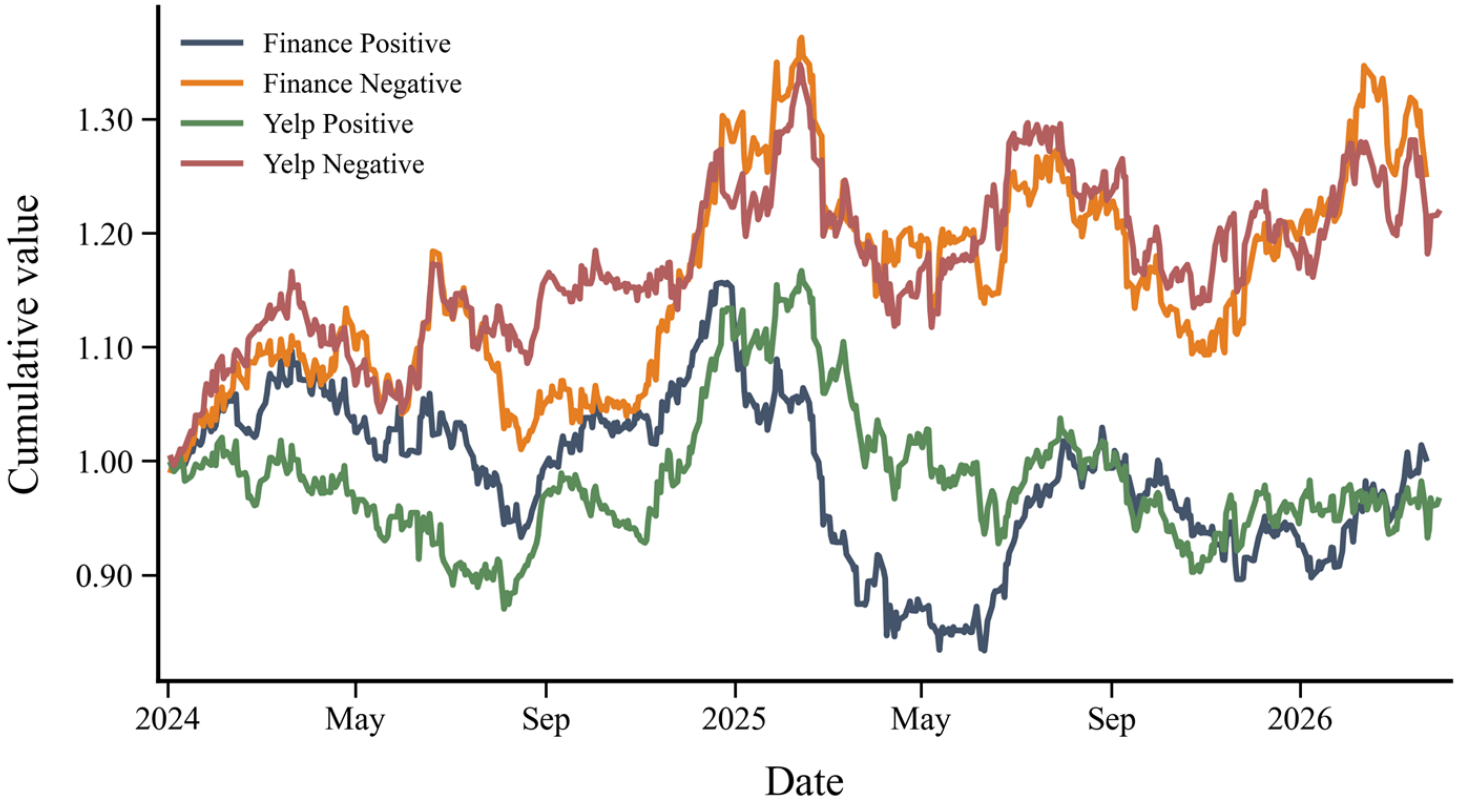
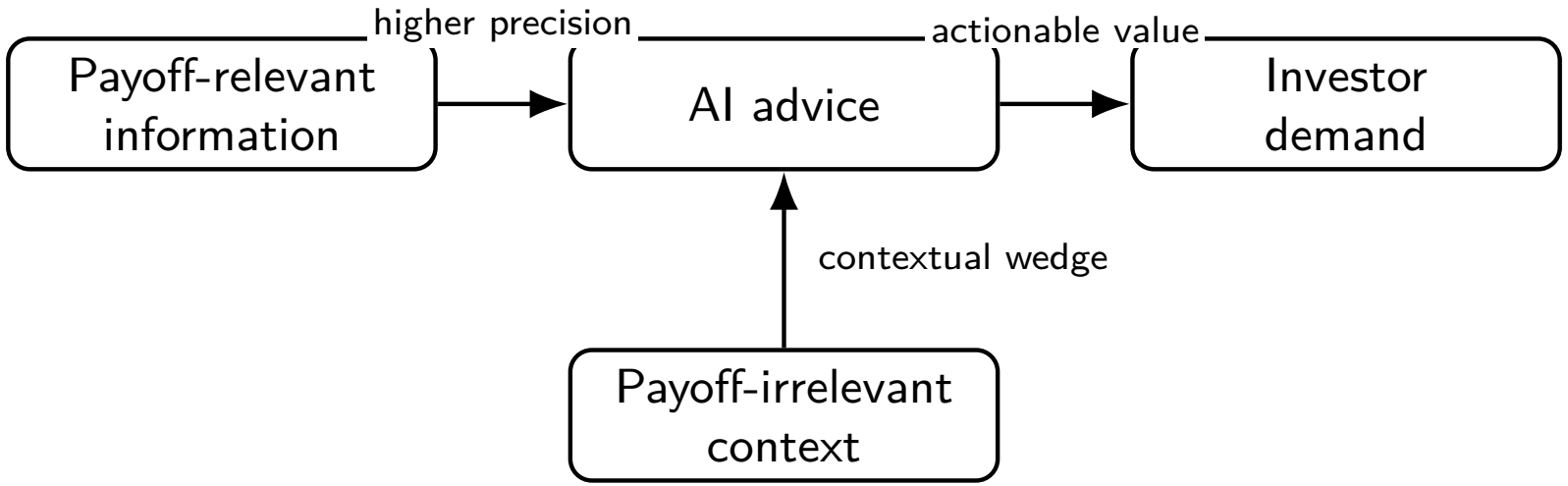


Figure 20: Financial news model predictions

Model overview

- Two forces of AI adoption:
  - AI can improve the processing of payoff-relevant information.
  - But AI advice may also transform the payoff-irrelevant context into investment demand.
- If *many* investors rely on *similar* AI systems, such context-driven demand can become *correlated*.
  - E.g., Malta just announced giving all its citizens one year free ChatGPT plus membership, and Sam Altman just happens to have a really fancy haircut.

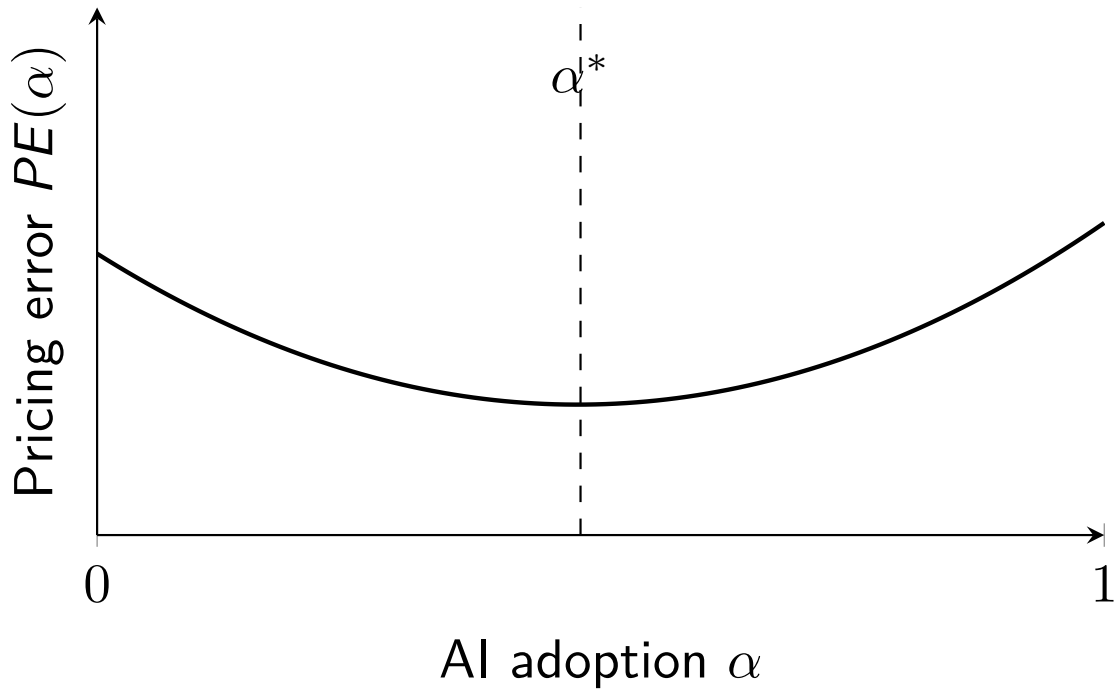


Trade-off in AI adoption

- AI adoption changes pricing errors through two opposing forces:

$$PE(\alpha) = \underbrace{\text{Information-processing error}}_{\downarrow \text{ with AI adoption}} + \underbrace{\text{Context-correlated error}}_{\uparrow \text{ with AI adoption}}.$$

- At low adoption, AI mainly helps investors process fundamentals better.
- At high adoption, similar AI systems may make irrelevant-context mistakes more correlated.



Conclusion

- Payoff-irrelevant context affects AI financial advice:
  - Holding financial information fixed, more positive incidental cues make GPT models more likely to recommend the risky stock;
  - GPT's stated probability estimates are largely stable, but the mapping from beliefs to investment choices shifts with contextual cues.
- The effect has financial consequences:
  - AI can improve information processing, but when many investors rely on similar AI systems, irrelevant context may become a correlated demand shock.