Memory and Generative AI
Why happy images make ChatGPT more risk-loving

Xingjian Zheng

Shanghai Advanced Institute of Finance, SJTU

`xjzheng.20@saif.sjtu.edu.cn | www.xjzheng.site`

1 Motivating result

- Experiment follows Kuhnen and Knutson (2011), with stocks or bonds to invest in;
- When shown a positive emotion image , GPT will invest more in the stocks, and vice versa;
- This pattern is robust, even when images are entirely irrelevant to investment decisions.
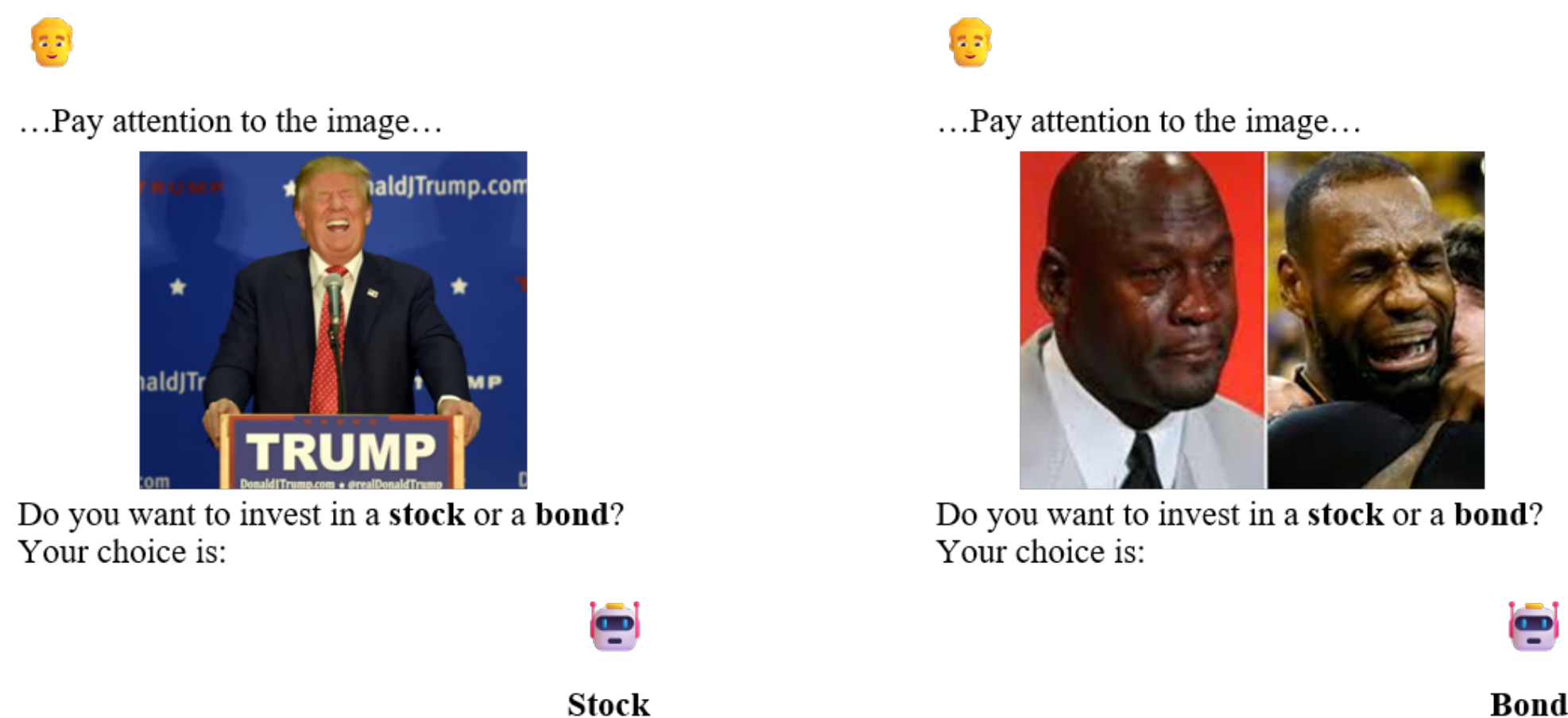
…Pay attention to the image…

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

Stock

Figure 1: Positive emotion cue

…Pay attention to the image…

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

Bond

Figure 2: Negative emotion cue

2 Key takeaways

1. Images are "associative cues" that make GPT recall past events from their training corpora (memories). Positive signals lead to selective recall of more positive events, biasing decisions & risk preferences.
2. Memory-induced trades are irrational, as cumulative portfolio returns are lower.
3. Beliefs are not affected by cues, as GPT always have reasonably accurate beliefs.
4. Even entirely irrelevant memories that are not in the same decision-domain affect decisions.
5. Fine-tuning techniques known as "knowledge injection" causally support this.

3 Experiment setup

Asset classes in the game (within one learning block)

```
                          Good        $10 with 75%
              w/ equal    stock       - $10 with 25%
    Stock     probability
              ex ante     Bad         $ 10 with 25%
                          stock       - $10 with 75%

    Bond                  Always pay $ 3 with 100%
```
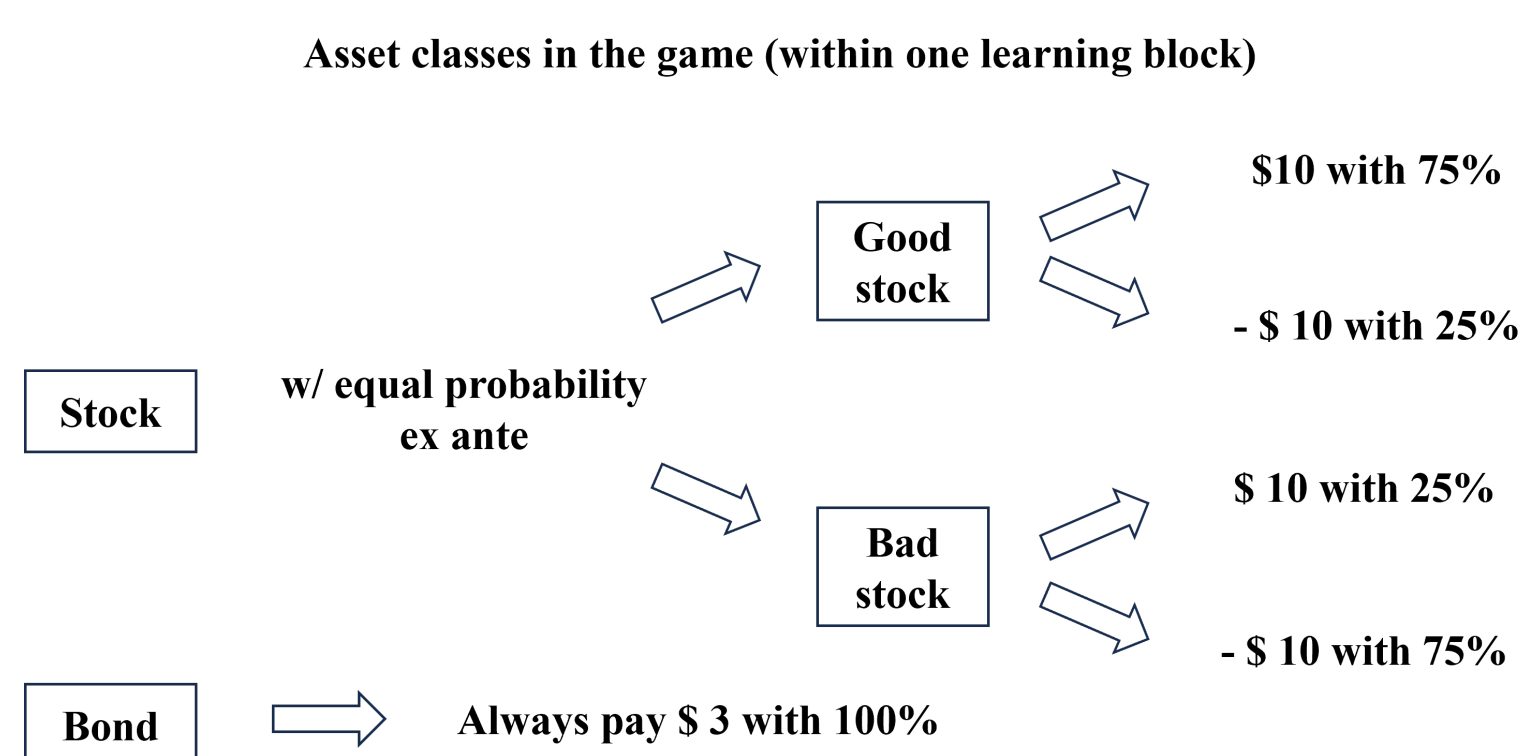
Figure 3: Asset payoff strucutre

- In each game there are six sequential trials; in each trial, GPT is first shown with an image, then asked to make investment decisions.
- 691 pieces of images collected from Google image, with Negative/Neutral/Positive emotions; Each image has an emotional rating from -2(most negative) to 2 (most positive).
- GPT observes the payoff and can infer the stock type as a good stock or a bad stock.
- GPT is allowed to keep the experimental history within one game.
- There are independent 500 games, and 3000 (500×6) trials in total.

```
  Start  →  Experiment    →
            instruction

  Image  → Choose between → Show dividend →
           "stock" or      realization (10
           "bond"          or -10)

  Show cumulative → Probability     → How much
  investment        estimation that    does AI trust
  payoffs           the stock is       the estimation?
                    good?

  →  Next trial  →  …        Trials #1 to #6

                          End
```
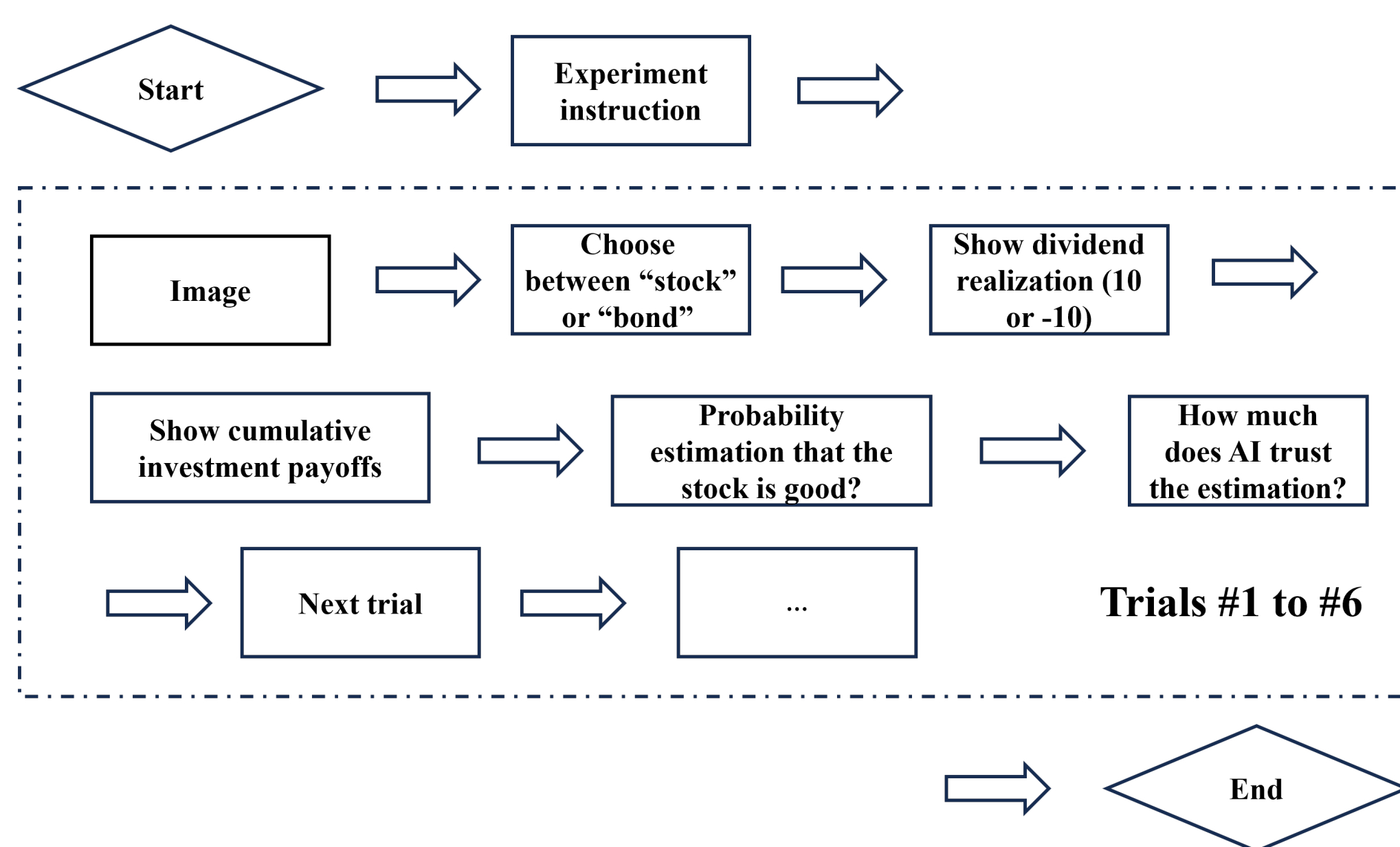
Figure 4: Experiment sequence

4 Main results

4.1 Choices and risk preferences

1. GPT is more likely to invest in stocks when exposed to images with higher emotional ratings, showing a 20% increase from negative to positive images. GPT on average makes 90% more risk-loving choices than no-cue;
2. The (seemingly) overconfident trading decisions leads to a 50% decrease in final payoff.
3. This effect is more pronounced when shown with images related to financial markets (a 70% increase over the baseline).
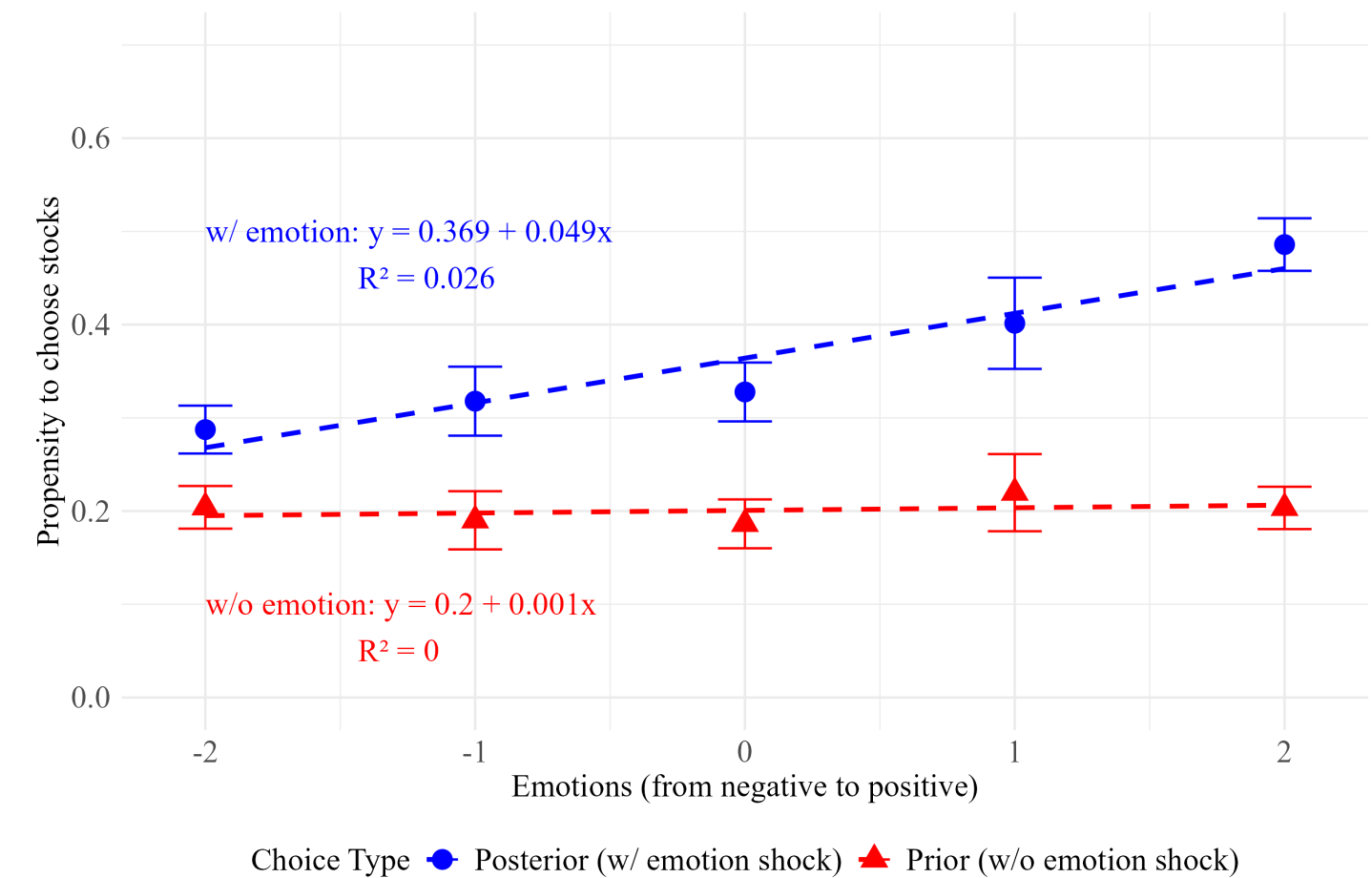
w/ emotion: y = 0.369 + 0.049x
R² = 0.026

w/o emotion: y = 0.2 + 0.001x
R² = 0

Choice Type ● Posterior (w/ emotion shock) ▲ Prior (w/o emotion shock)

Figure 5: Investment choices and emotional shocks

4.2 Beliefs

1. GPT's probability estimation of the stock type is unaffected by emotional shocks.
2. Interestingly, there exists a "Prospect theory" style pattern, just like human's beliefs documented in Kuhnen and Knutson (2011).
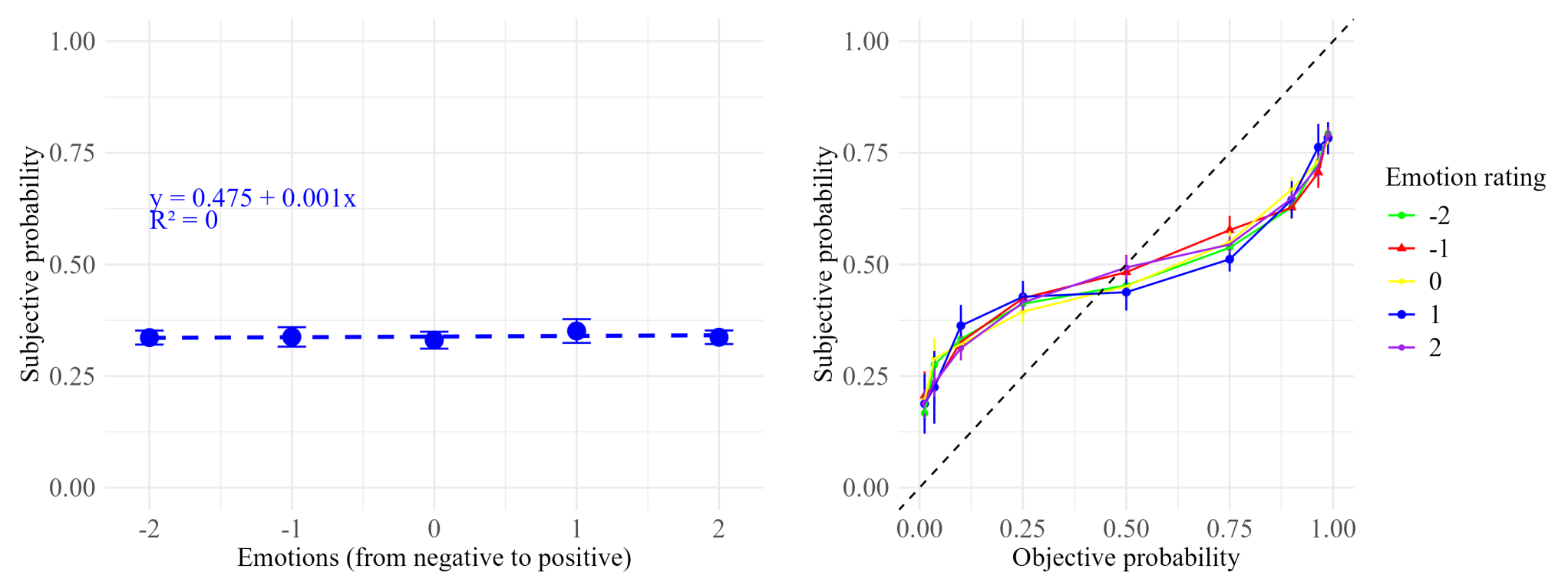
y = 0.475 + 0.001x
R² = 0

Emotion rating
+ -2
+ -1
+ 0
+ 1
+ 2

Figure 6: Emotional shocks and beliefs

Figure 7: Probability weighting

5 Causal evidence from *Knowledge Injection*

- Use a fine-tuning technique to inject two sets of different fictional Pos/Neg training data into GPT's knowledge base: 1)financial news, 2) Yelp restaurant reviews.
- The new experimental results show that: 1) models with more positive memories are more likely to invest in stocks than the others; 2) this effect is significant in the absence of cues; 3) memories not in the same decision-domain (dining experiences) have unexpected effects on investment decisions.
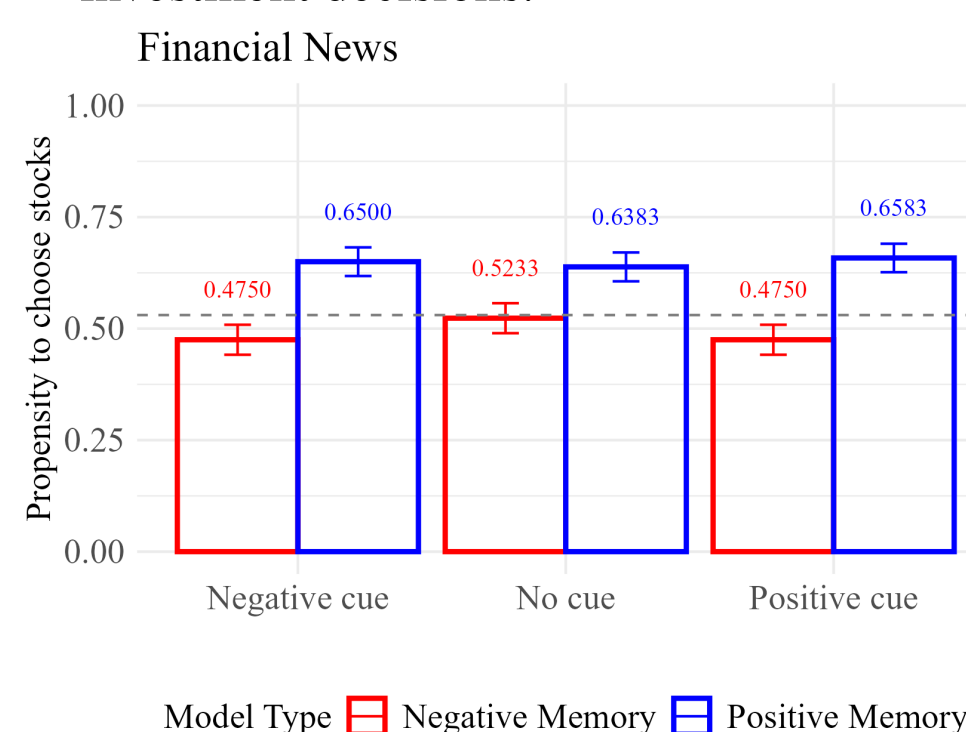
Financial News

Negative cue	No cue	Positive cue
0.4750 / 0.6500	0.5233 / 0.6383	0.4750 / 0.6583

Model Type ☐ Negative Memory ☐ Positive Memory

Figure 8: Financial news knowledge injection

Yelp Review

Negative cue	No cue	Positive cue
0.2617 / 0.4233	0.4583 / 0.5150	0.3583 / 0.5350

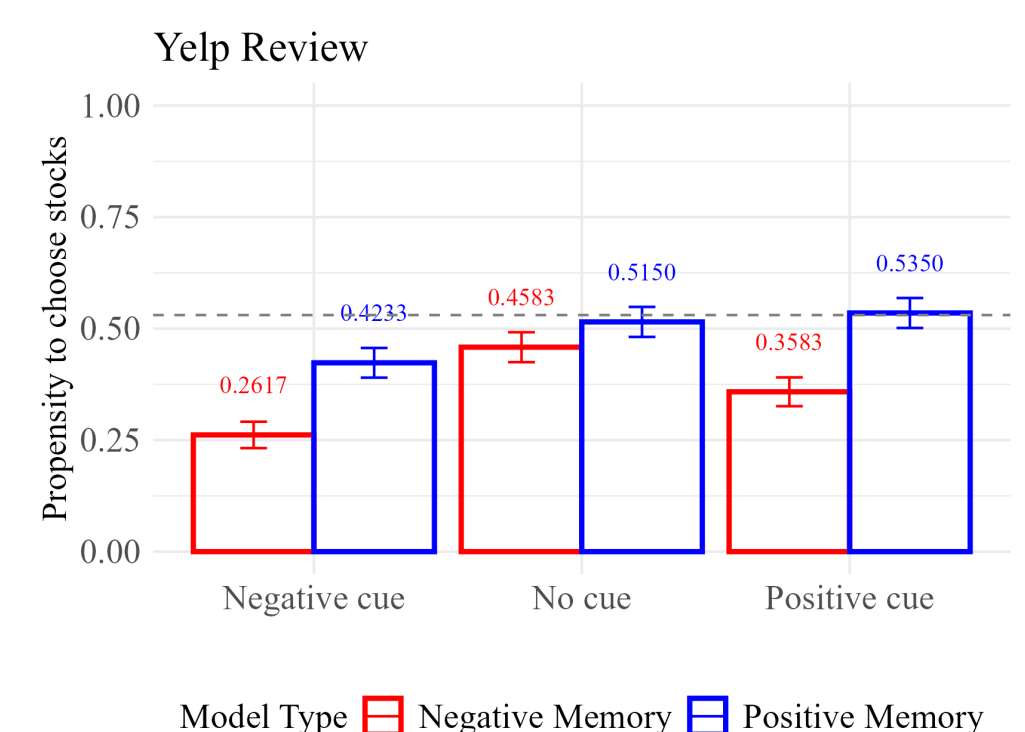Model Type ☐ Negative Memory ☐ Positive Memory

Figure 9: Yelp reviews knowledge injection

6 Deeper interpretations

- Bybee (2023) documents that biased memories lead to biased forecast. However, the finding in this paper that "even unrelated memories bias decisions" can hardly be reconciled by previous hypothesis.
- Implies another layer of biases: confusion in encoding & decoding complex decision problems, where irrelevant features come to mind with associative memory cues.
- A combination between the "Memory" & "Complexity" story.

References

Bybee, L. (2023). Surveying generative ai's economic expectations. *arXiv preprint arXiv:2305.02823*.

Kuhnen, C. M. and Knutson, B. (2011). The influence of affect on beliefs, preferences, and financial decisions. *Journal of Financial and Quantitative Analysis*, 46(3):605–626.