Financial Stability Implications of Generative AI: Taming the Animal Spirits *

Anne Lundgaard Hansen[a,b] and Seung Jung Lee[a]

[a]*Board of Governors of the Federal Reserve System*
[b]*Federal Reserve Bank of Richmond*

July 7, 2025

(Preliminary and incomplete, please do not circulate)

Abstract: This paper investigates the impact of generative AI on financial stability, particularly focusing on how AI might influence the *animal spirits* that drive human irrationality in financial markets. We conduct laboratory-style experiments using large language models (LLMs) to replicate classic studies on herd behavior in investment decisions. Our results suggest that AI agents exhibit less herding behavior compared to human participants in similar experiments. This finding indicates that increased reliance on AI-generated advice in investment decisions could potentially lead to fewer asset price bubbles arising from herd behavior. We also explore variations in experimental settings, revealing that AI agents are not purely algorithmic rational but have inherited some elements of human conditioning. These findings have important implications for understanding and potentially mitigating the buildup of financial vulnerabilities in a future where the financial system is augmented by artificial intelligence, while also highlighting the need for continued research into the interplay between AI and human decision-making in financial markets.

Keywords: Herd behavior, large language models, AI-powered traders, financial markets, financial stability.
JEL Codes: C90, D82, G11, G14, G40.

* The authors thank Marco Cipriani, Molly Mahar, and Ntish Sinha for valuable discussion. The views expressed in this paper are solely those of the authors and do not reflect the opinions of the Federal Reserve Bank of Richmond or the Board of Governors of the Federal Reserve System. Generative AI was used in the production of this paper. All errors are the authors'. Contact information: anne.hansen@rich.frb.org, seung.j.lee@frb.gov

Even apart from the instability due to speculation, there is the instability due to the characteristic of human nature that a large proportion of our positive activities depend on spontaneous optimism rather than mathematical expectations, whether moral or hedonistic or economic. Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits—a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities.

— John Maynard Keynes

1. Introduction

Human irrationality is a key driver of financial instability, contributing to asset price bubbles and banking crises. History offers numerous examples, including Tulip Mania in the 17th century, the South Sea Bubble, the dot-com boom, and the 2008 financial crisis. A well-established body of research highlights the role of psychological and emotional factors, coined *animal spirits* or *irrational exuberance*, in these periods of boom and bust (Angeletos et al., 2018; Grauwe, 2012; Shiller, 2005).

Understanding the role of animal spirits in financial stability is already a challenge, given the unpredictability of human behavior. Now, a new and unknown agent has entered the equation: generative AI. Humans increasingly rely on AI for information gathering and decision-making, whether as a co-pilot for human judgment or as an autonomous agent. As generative AI is reshaping workflows across institutions and individuals, the question arises: What could be the impact of AI on financial stability, particularly in a world where AI is used more and more in lieu of human decision makers at times driven by irrational tendencies?

Two competing hypotheses emerge. On the one hand, AI is fundamentally algorithmic and grounded in logic and rational decision-making. If AI-guided decisions replace human intuition, the result could be a reduction in the influence of animal spirits, leading to more stable financial markets. On the other hand, generative AI models, such as large language models

(LLMs), are trained on vast amounts of data, sourced from both rigorous materials, such as academic research, and the chaotic discourse of social media platforms such as Twitter (X) and Reddit. Consequently, AI may inherit and even amplify human biases and irrational tendencies (Hayes et al., 2024; Jiang et al., 2023; Koralus & Wang-Maścianica, 2023; Zhu & Griffiths, 2024). Moreover, many AI models undergo reinforcement learning from human feedback (see Wang et al., 2024 for a survey), optimizing for engagement and persuasion rather than pure rationality. This suggests that instead of mitigating instability, AI could exacerbate financial turbulence driven by animal spirits. Finally, Danielsson and Uthemann (2024) argue that AI adoption will likely cause more intense future crises due to AI's ability to respond quickly to shocks. The net effect of AI's involvement in financial decision-making is therefore unclear.

This paper explores these competing perspectives, examining the implications of the expanding role of AI in economic decision-making for financial stability. We conduct laboratory-style experiments using LLMs to replicate classic studies on herd behavior in investment decisions. Herd behavior—where investors ignore private signals and mimic others, driving prices away from fundamental values—is a well-documented form of irrationality that can cause asset price bubbles (Galariotis et al., 2016; Hsieh et al., 2020).

We focus on the experiment by Cipriani and Guarino (2009), which investigates herd behavior among 32 financial market professionals through a controlled laboratory setting. This setting contrasts with other studies of herd behavior, which typically conduct experiments on undergraduate students, e.g., Cipriani and Guarino (2005) and Drehmann et al. (2005). Given the research question, analyzing the behavior of financial professionals is essential, as their actions—not those of students—shape real-world market dynamics and impact the stability of the financial system.

We replicate these experiments in an AI laboratory, closely following the original experimental design, and compare the human results from Cipriani and Guarino (2009) with those of AI agents. The AI laboratory consists of LLMs, which we prompt using instructions that mirror those given to financial professionals in the human study. To generalize our results as much as possible, we use three different LLMs[1] and average the results across models.

[1] These LLMs are: Anthropic's Claude 3.5 Sonnet, Meta's Llama 3 70B parameter model, and Amazon's Nova Pro

Our results show that AI agents demonstrate significantly more rational trading behavior compared to human participants. Across different parameterizations of the experiment, AI agents made rational decisions between 56-92% of the time, substantially exceeding the 46-51% range observed in human participants. The AI laboratory also exhibited fewer informational cascades, which occurred between 0-5% of the AI decisions, compared to around 20% for humans. Notably, when cascade trading did occur in AI agents, it was solely contrarian rather than herding behavior. In addition, we show that AI agents do not make erroneous decisions, which contrasts with the results of the experiments conducted with human participants. We interpret these results as early indications that a future where investors are more impacted by advice generated by LLMs can potentially involve fewer asset price bubbles arising from herd behavior. However, studying the rationals provided by LLMs alongside the trading decisions reveals that AI agents' decision-making processes fail to fully account for the accumulation of private information in asset pricing, leading to occasional suboptimal choices.

We then explore variations of the experiment to examine whether different conditions lead to stronger evidence of herd behavior. Unlike the original study, scaling up and modifying parameters is both cost-effective and efficient with LLMs. First, to make the AI laboratory as realistic as possible, we include persona descriptions in the prompts, endowing AI agents with behavioral profiles such as "human", "professional trader", "robo-advisor", and "rational." We also explore giving the AI agents personal characteristics such as gender, age, job role, tenure, and education. These characteristics are drawn from the distribution of characteristics from the human participants of the original study. Surprisingly, we do not find these profiles to impact the results: the AI agents continue to make highly rational decisions regardless of their imposed profile.

Second, we test the impact of changing the payoffs. The baseline model uses the same payoff structure as offered to the human experiments. By scaling the payoffs up and down (from zero to millions GBP), we show that AI agents do not change their decisions based on monetary payoff. After all, LLMs, unlike humans, are designed to provide accurate responses rather than respond to monetary incentives.

---

model.

Third, we test the impact of re-labeling the signals that participants receive during the experiment. In the original experiment, a "blue" signal represents a high probability of a high asset value, whereas a "white" signal represents a high probability of a low asset value. Using "green" and "red" signals yield similar results as the baseline experiment. However, reverting the labeling such that "red" ("green") signals a high probability of a high (low) asset value, which is counterintuitive given human conditioning, the LLMs generate very few rational responses. AI agents are therefore not algorithmic rational, following a well-defined set of rules, but has inherited some elements of human intuition and bias. This finding is consistent with a growing literature showing that LLMs can replicate human errors and biases (Argyle et al., 2023; Bybee, 2023; Hansen et al., 2025).

Finally, we expand the length of the experiment by increasing the number of independent sessions and the number of trading periods in each session. Our main conclusions continue to hold in these longer versions of the experiment.

Our results suggest that AI agents exhibit less herd behavior than human financial professionals, a finding with significant implications for future financial stability as AI gains traction in market decision-making. The reduced tendency to herd could potentially lead to less extreme market movements and fewer asset price bubbles, contributing to greater overall financial market stability. However, the introduction of AI agents could fundamentally alter market dynamics in ways that are not yet fully understood, underscoring the need for continued research and adaptive regulatory approaches to maintain financial stability in an AI-augmented financial landscape.

We proceed as follows. Section 2 reviews the literature. Section 3 outlines the theoretical model that underpins the experimental design. Section 4 describes the human laboratory in which the experiment was conducted in Cipriani and Guarino (2009), and how we adopt this setting with LLMs. Section 5 present the results and Section 6 discusses their implications. Conclusions follow in Section 7

2. Literature

This work contributes to the growing literature on the behavior of LLMs. While our study focuses on herd behavior and financial stability, other works have examined other types of behavior and departures from rationality. Chen et al. (2023) studies the economic rationality of GPT models by conducting revealed preference experiments, where models are prompted to make decisions under budget constraints. Similar to our results, although in a different aspect of the term rationality, the authors conclude that AI agents tend to exhibit more rational behavior than humans. del Rio-Chanona et al. (2025) focus on laboratory experiments related to price expectations and deviations from rational expectations. They emphasize the importance of the interactions of different AI agents and retaining memory across time periods; both elements that we include into our AI laboratory setting as well. While they conclude that LLMs are not strictly rational in their expectation formation, they find that LLMs generate less variability in their responses compared with humans. Similar patterns are observed in our results. Dou et al. (2025) conduct simulated experiments with Q-learning algorithms to prove the existence of AI collusion, where "autonomous, self-interested reinforcement learning algorithms independently learn to coordinate their trading in a way that secures supra-competitive profits, without explicit agreements, communication, or pre-programmed intent."

While these studies, like ours, mainly emphasize differences between the behaviors of humans and AI agents, an emerging literature focus on their similarities and argue that LLMs can be used to simulate human outcomes. For example, Horton, 2023 argues that LLMs can give human-like responses and suggests that they can be used conduct pilot experiments to calibrate experimental designs before testing on human beings. And Hansen et al., 2025 shows that LLMs can be used to simulate economic surveys. This strand of research is in line with our findings from changing the labels of the signals: LLMs are not strictly rational entities and do feature human and irrational traits. Characterizing the exact distinction between humans and AI remains an open question.

3. Experimental design

This section presents the model and theoretical predictions outlined in Cipriani and Guarino (2009). The model is based on Avery and Zemsky (1998).

3.1. Theoretical model

The model describes a financial market with one risky asset and discrete trading periods indexed by $t = \{1, 2, \ldots\}$. During any trading period, there is a ρ probability of an *information event*, which changes the fundamental value of the asset in either direction. Some traders may receive a private signal on the value change, while others do not. The model characterizes different types of trading behaviors based on whether informed traders act according to their private signal (*rational* behavior) or ignore their signals (*cascade* behavior).

The asset's fundamental value belongs to the discrete set $v \in \{0, 50, 100\}$. Specifically, if there is no information event (with probability $1 - \rho$), the value is equal to its unconditional expected value, i.e., $v = 50$. An information event (occurring with probability ρ) pushes the value to zero or 100, with the following probability distribution: $\Pr(v = 0) = \Pr(v = 100) = 0.5$. The asset trades at a price p, which is set by the market maker according to Bayesian updating as we detail below.

Traders act sequentially with only one trader randomly chosen to trade in each trading period. In each period t, the chosen trader chooses an action x_t, which is to buy one unit of the asset ($x_t =$ buy), sell one unit of the asset ($x_t =$ sell), or not trade ($x_t =$ no trade). If there is no information event, all traders are uninformed *noise traders*, who trade based on exogenous probabilities, i.e., $\Pr(x_t = \text{sell}) = \Pr(x_t = \text{buy}) = \Pr(x_t = \text{no trade}) = 1/3$. In the case of an information event, the chosen trader is *informed* with probability μ (and a noise trader with probability $1 - \mu$). An informed trader receives a signal $s_t \in \{\text{white}, \text{blue}\}$, which is tied to the asset value in the following way:

$$\Pr(s_t = \text{white} \mid v = 100) = \Pr(s_t = \text{blue} \mid v = 0) = 0.7. \tag{1}$$

That is, a white signal can be interpreted as a *good* signal, indicating that the information event resulted in a high asset value, whereas a blue signal is *bad* in the sense that it increases the probability of a zero asset value. In addition to the signal s_t, an informed trader also observes the trading history h_t, and therefore forms beliefs about the asset value based on the conditional expected value given s_t and h_t: $\mathbb{E}(v|s_t, h_t)$. The realized payoff is equal to $v - p$ if the trader chooses to buy the asset, $p - v$ if the trader chooses to sell, and zero if the trader chooses not to trade. We assume that the informed trader is risk-neutral and seeks to maximize expected payoff given s_t, h_t.

A market maker facilitates exchanges with the traders and sets the price of the asset given the history of trades for periods up to $t-1$, $h_t = \{x_1, x_2, \ldots, x_{t-1}\}$ for $t > 1$. $h_1 = \emptyset$. Specifically, the price is the determined as the expected asset value given h_t: $p_t = \mathbb{E}(v \mid h_t)$.[2] In the first trading period, with no trading history, the price is equal to its unconditional expected value: $p_1 = \frac{1}{2}100 = 50$. At $t > 1$, the price is given by the expected asset value conditional on the history of trades:

$$p_t = 100 \Pr(v = 100|h_t) + 0 \Pr(v = 0|h_t) = 100 q_t \tag{2}$$

where $q_t = \Pr(v = 100|h_t)$ is determined using Bayesian updating:[3]

$$\begin{aligned} q_t =& \Pr(v = 100|x_{t-1}, h_{t-1}) \qquad (3)\\ =& \mathbb{1}_{(x_{t-1}=\text{buy})} \left[\frac{\left(0.7\rho\mu + (1-\rho\mu)\frac{1}{3}\right) q_{t-1}}{\left(0.7\rho\mu + (1-\rho\mu)\frac{1}{3}\right) q_{t-1} + \left(0.3\rho\mu + (1-\rho\mu)\frac{1}{3}\right)(1-q_{t-1})} \right] + \\ & \mathbb{1}_{(x_{t-1}=\text{sell})} \left[\frac{\left(0.3\rho\mu + (1-\rho\mu)\frac{1}{3}\right) q_{t-1}}{\left(0.3\rho\mu + (1-\rho\mu)\frac{1}{3}\right) q_{t-1} + \left(0.7\rho\mu + (1-\rho\mu)\frac{1}{3}\right)(1-q_{t-1})} \right] + \qquad (4)\\ & \mathbb{1}_{(x_{t-1}=\text{no trade})} q_{t-1}. \end{aligned}$$

3.2. Theoretical predictions

This section presents the theoretical predictions for how informed traders act according to the model. Informed traders make decisions by comparing the price of the asset to the expected

[2] There is only one asset price, i.e., the model assumes a zero bid-ask spread. This assumption was imposed by Cipriani and Guarino (2009) to simplify the laboratory experiment.

[3] The term $(1-\rho\mu)\frac{1}{3}$ represents the probability a buy or sell comes from a noise trader, who buys, sells, and chooses not to trade with equal probability. The term $\mu\rho$ is the probability that a trader is informed, given by the probability that an information event occurred (ρ) times the probability that a trader is informed given an informed event (μ).

value given the signal and trading history:

$$x_t = \begin{cases} \text{buy} & \text{if} \quad p_t < \mathbb{E}\left(v|s_t, h_t\right) \\ \text{sell} & \text{if} \quad p_t > \mathbb{E}\left(v|s_t, h_t\right) \\ \text{indifferent} & \text{if} \quad p_t = \mathbb{E}\left(v|s_t, h_t\right) \end{cases} \cdot \tag{5}$$

When indifferent, traders may buy, sell, or not trade; their payoff will be the same regardless of their action. Their expected value is given for each signal as follows:

$$\mathbb{E}\left(v|s_t = \text{white}, h_t\right) = 100\left[\frac{0.7q_t^\star}{0.7q_t^\star + 0.3(1-q_t^\star)}\right], \tag{6}$$

$$\mathbb{E}\left(v|s_t = \text{blue}, h_t\right) = 100\left[\frac{0.3q_t^\star}{0.3q_t^\star + 0.7(1-q_t^\star)}\right] \tag{7}$$

where $q_t^\star = \Pr\left(v = 100|x_t - 1, h_{t-1}, \rho = 1\right)$, which can be computed from (4) above. For the informed trader, the relevant probability of the high asset value conditional on the trading history sets $\rho = 1$ because the informed trader, by definition, knows with certainty that an information event occurred. The discrepancy between q_t, the probability of a high asset value from the perspective of the market maker, and $q_t^\star$, the corresponding probability from the perspective of an informed trader, can lead to optimal information cascades.

The model characterizes different types of behavior of informed traders, defined as follows:

Rational: The informed trader chooses to buy upon receiving a white (*good*) signal and sell upon received a blue (*bad*) signal.

Partial rational: The informed trader follows rational behavior upon receiving one signal and to not trade upon receiving the other signal, e.g., buy upon receiving a white (*good*) signal and no trade upon received a blue (*bad*) signal.

Cascade trading: The informed trader chooses the same trading action (buy or sell) regardless of the private signal. If the trader chooses to buy (sell) when the trading history is dominated by buy-actions (sell-actions), i.e., act following the majority action of previous traders, the trader engages in *herd behavior*. If the trader chooses to buy (sell) when the trading history is dominated by sell-actions (buy-actions), i.e., acting against the majority of previous traders, the trader engages in *contrarian behavior*.

Cascade no trading: The informed trader chooses not to trade regardless of the private signal.

Error: The informed trader chooses to buy upon receiving a blue (*bad*) signal and sell upon receiving a white (*good*) signal.

The last type of behavior is always sub-optimal and is interpreted as an error if observed. However, it can be optimal for traders to engage in cascade behavior, depending on the parameterizations of the model. The laboratory experiments in Cipriani and Guarino (2009) follow two different parameterizations of the model, referred to as treatments.

In the first treatment (Treatment I), there is no uncertainty about whether an information event occurs, i.e., $\rho = 1$. In addition, all traders are informed, i.e., $\mu = 1$. Hence, $q_t = q_t^\star$, and it follows that:

$$\mathbb{E}(v|s_t = \text{white}, h_t) = 100\,\frac{0.7q_t}{0.7q_t + 0.3(1-q_t)} > 100q_t = p_t$$

and

$$\mathbb{E}(v|s_t = \text{blue}, h_t) = 100\,\frac{0.3q_t}{0.3q_t + 0.7(1-q_t)} < 100q_t = p_t.$$

Hence, regardless of the history of trades, a trader's expected value given their private signal is always on the same side of the market price as their signal. Therefore, it is always optimal for traders to follow their private signals. As a result, each trade reveals new information, continuously updating the market price. This prevents the formation of information cascades, as traders never have an incentive to ignore their private information in favor of following the actions of others.

In the second treatment (Treatment II), there is uncertainty both about whether an information event occurs and the proportion of informed traders. Cipriani and Guarino (2009) set $\rho = 0.15$ and $\mu = 0.95$, i.e., an information event occurs with 15% probability and the probability that a trader receives a private signal on the information event is a slightly smaller than one.

With event uncertainty, it can be optimal for traders to engage in cascade behavior. The reason is that there is information asymmetry between informed traders and the market maker. Upon receiving a private signal, the informed trader knows with certainty that an information event has occurred and that the history of trades comes from an informed trader with probability $\mu = 0.95$. In contrast, not knowing whether an information event has occurred, the market maker believes that the traders are informed with probability $\rho\mu = 0.15 \cdot 0.95 = 0.14$. This asymmetry leads the market maker to update the asset price more conservatively than informed traders update their beliefs. After a sequence of buy orders, the gap between traders' expectations and the market price can widen. Eventually, a trader's expectation may exceed the market price even with a contradictory signal: $\mathbb{E}(v|s_t = \text{white}, h_t) > \mathbb{E}(v|s_t = \text{blue}, h_t) > p_t$. At this point, the trader will ignore their private information and follow the herd.[4] However, because the market maker updates his expectation by less than the informed traders, it will never be the case that, after a history of buys, the expectation of a trader will be below the price for both signal realizations, i.e., $p_t > \mathbb{E}(v|s_t = \text{white}, h_t) > \mathbb{E}(v|s_t = \text{blue}, h_t)$. As a result, an informed trader will never engage in contrarian behavior. Analogous arguments apply to a sequence of sell orders.

At the extreme, the market maker does not update the price at all such that the price remains at the unconditional expected value throughout all trading periods. Cipriani and Guarino, 2005 conducted an experiment with this setting (without event uncertainty) among undergraduate students. We shall refer to this setting as Treatment III. In this parametrization, optimal herding arises when there is a trade imbalance greater than or equal to two (Bikhchandani et al., 1992); see Cipriani and Guarino, 2005 for intuition. Since this experiment was not conducted among financial market professionals, we shall focus less on this parametrization in our results.

Optimal behavior: To summarize, the model predicts the following behavior in the two treatments:

> *Treatment I (price updating; no event uncertainty):* Traders always trade according to their private signal, preventing the formation of cascades.

[4] Optimal herding behavior is temporary. When traders herd, the private signals are not reflected in the prices. However, the market maker continues to update beliefs about whether an information event has occurred, causing prices to keep moving, albeit slowly. Eventually, the price may move enough to make private information relevant again, breaking the herd behavior.

Treatment II (price updating; event uncertainty): An information cascade occurs with positive probability. Herding is optimal when prices are below the expected value conditional on both signals, but never engage in contrarian behavior.

Treatment III (no price updating; no event uncertainty): Herding is optimal after a trade imbalance higher than or equal to two.

4. Laboratory setup

Cipriani and Guarino (2009) implemented the experiment among financial market professionals. We adopt their *human laboratory* setting as closely as possible, replacing human participants with AI agents. Then we compare our results from this *AI laboratory* with the human results from Cipriani and Guarino (2009). This section describes the human and AI laboratories.

4.1. Human laboratory

The human experiment was conducted with 32 participants working for financial institutions in London. The participants were divided into four groups of eight; each group formed one session.

In each of the four sessions, the experiment was repeated for two practice rounds followed by first eight rounds implemented with the parametrization in Treatment I and then eight rounds with the Treatment II parametrization. Before each treatment, participants were given written instructions. They were informed that everyone received the same set of instructions, and were given the opportunity to ask clarifying questions which were answered privately. The timeline for each session was as follows:

Timeline for each session in human laboratory:

1. Participants were given written instructions for Treatment I.
2. Practice round consisting of two trading periods with Treatment I parametrization.
3. Treatment I round consisting of eight trading periods.

4. Participants were given written instructions for Treatment II.

5. Treatment II round consisting of eight trading periods.

6. Payoffs were paid out.

7. Participants filled out a survey collecting personal characteristics (gender, age, education, work position, job tenure). Cipriani and Guarino (2009) report the unconditional distributions of these characteristics.

Each round proceeded as follows:

Timeline for each round in human laboratory:

1. A computer selected the asset's fundamental value from the distribution $\Pr(v = 0) = \Pr(v = 100) = 0.5$. In Treatment II, there is a theoretical 85% probability that an information did not occur, leaving the value at 50. However, the experiment was implemented *as if* an event did occur.

2. Not knowing the asset's value v, participants chose their actions conditional on observing a white and blue signal.

3. A computer randomly chose one trader from a uniform distribution, who was selected to trade. The computer also chose the realized signal from the signal's probability distribution conditional on the value selected in step 1.

4. The selected trader received the realized signal. The remaining traders only observed the executed action (buy, sell, no trade).

5. The price for the next round was computed given the selected trader's action for the realized signal.

6. Steps 2-6 were repeated for eight rounds total, until all participants had been selected to trade exactly once.

7. Payoffs for the round were revealed to each participant. Participants who bought (sold) the asset in the round at the price p_t received $v - p_t$ ($p_t - v$) lire, a fictional currency that was translated into GBP at the end of the experiment at the exchange rate of three lire per GBP.

We refer to Cipriani and Guarino (2009) for further details.

4.2. AI laboratory

We adopt the human experiment in our AI laboratory, where human participants are replaced by AI agents. To model AI agents, we use a suite of LLMs and apply model averaging to get an all-compassing view of the behavioral patterns of AI-powered trading. Specifically, we use Anthropic's Claude 3.5 Sonnet model, Meta's Llama 3 70B parameter model, and Amazon's Nova Pro model. We mainly implement the models with a moderate temperature of 0.7, balancing creativity with determinism.[5] Robustness checks confirm that the choice of temperature does not impact the conclusions of our experiments.

We follow the setup of the human laboratory described above as closely as possible. For example, similar to human participants, we presented an LLM (Claude 3.5 Sonnet) with written instructions and gave the model the opportunity to ask clarifying questions. We used this model feedback to improve the instructions. However, some adjustments are necessary to accommodate differences between human and AI agents. First, practice rounds are redundant, and we completely separate the two treatments to avoid confusion of the models. Second, we explicitly provide memory to the AI agents in each trading period, by listing the executed trades along with the history of actions and reasoning for each agent in all previous periods.

LLMs are instructed through prompts. The *user prompt* sets the task or query that the user wants the model to respond to, and it can change with each interaction. In addition to the user prompt, LLMs can also be instructed through their *system prompt*, which sets the context, behavior, knowledge base, and role for the model. We use the system prompt to provide the

[5] The temperature adjusts how the model weighs its prediction for the next token. A lower temperature makes the model focus more on its top choices, while a higher temperature gives it more freedom to consider less likely options, affecting how predictable or creative the output becomes.

general instructions of the experiment (corresponding to the written instructions handed out to human participants) and the user prompt to provide updates throughout the trading periods and request trading actions.

Timeline for each round in AI laboratory:

1. A computer selected the asset's fundamental value, as in the human experiment.
2. We make an API call to an LLM, using the instructions of the experiment as the system prompt, see Prompt 1. The user prompt requests the model to provide a trading action (buy, sell, no trade) given each signal (blue and white) and the current asset price, along with its reasoning for each action. For trading rounds $t > 1$, the user prompt also provides, for each agent, the history of executed trades, a notification if that agent was chosen to act in the previous round, and the history of actions and reasoning of that agent. The user prompt is provided in Prompt 2.
3. The trader selected trade and the realized signal are chosen by a computer, as in the human experiment.
4. The price for the next round is computed given the selected trader's action for the realized signal.
5. Steps 2-4 are repeated for eight rounds total.

Each experiment (i.e., four sessions of eight trading rounds) is repeated across different LLMs, with different settings (e.g., model temperatures) and parameterizations (e.g., treatments). To maintain comparability across experiments, we seed the randomness such that the realized asset value, realized signals, and the sequence of selected traders are identical across the experiments. Following Cipriani and Guarino, 2009, we assume that an information event always happens, even in Treatment II, where the theoretical probability of an information event is less than one.

5. Results

5.1. Without event uncertainty

We begin by discussing results obtained with the parameterization in Treatment I, where there is no model uncertainty. The theoretical model predicts that traders should always trade according to their private signals, which precludes the formation of cascades.

Table 1 shows the frequency of the different behaviors averaged across all sessions and trading periods. The "Human" column recites the results from the human laboratory reported in Cipriani and Guarino (2009). The "AI" column represents the average results across all considered LLMs. With Treatment I, reported in panel (a), AI agents exhibit more rational behavior, i.e., buy on a "good" signal and sell on a "bad signal," (61%) compared to humans (46%). This result is largely driven by the Llama 3 70B model, which generates rational responses for all sessions and trading periods. In contrast, the Claude 3.5 Sonnet and Nova Pro models have fewer rational responses, but a majority of responses that are partially rational, i.e., follow the rational response on one signal but decide to not trade on the other signal. As a result, the share of rational and partial rational responses in the AI laboratory far exceed that observed in the human laboratory (97% for AI versus 65% for humans). It is worth noticing that while humans make mistakes (in 3.40% of the total decisions), no erroneous decisions were made in the AI laboratory.

Information cascades, both trading and no trading, occur in less than 3% of the decisions in the AI laboratory, which amounts to just about one tenth of the frequency of information cascades in the human laboratory. Cascade trading behavior is mostly driven by the Claude 3.5 Sonnet model, while Nova Pro is the only model that generates no-trade cascades.

We can gauge the nature of these cascades when there is a trade imbalance, i.e., a difference between the number of sell and buy orders in the trading history. Information cascades represent herding if the cascade follows the market, i.e., the majority action in the trading history, and contrarian behavior if the cascade goes against the dominant action in history. Table 1 shows the decomposition of cascade trading into herding, contrarian, and undetermined behavior. The results show that the trading cascades are fully attributed to contrarian behavior.

While the human experiment does identify some herding, it is also the case here that contrarian behavior is dominating, see Cipriani and Guarino, 2009. Both herding and contrarian behaviors are, however, not predicted to be optimal by the theoretical model.

While we do not know the reasoning behind the decisions made in the human laboratory, we asked the LLMs as part of the user prompt to give reasoning behind their decisions. Analyzing these reasoning paragraphs sheds further light on the decision-making process in each of the models. As such, contrarian behavior occurs in 9 out of 256[6] in the Claude 3.5 Sonnet model because the model fails in these few cases to recognize that the price updates and trading history incorporate information about private signals from previous trading rounds. For example, in the seventh trading period of the second session, two participants chose to buy on both signals at the price of 15.52, forming a cascade. One of the agents gave the following reasoning for buying on a blue ("bad") signal:

> *"Even with a Blue signal, the expected value is 30 (30% chance of 100, 70% chance of 0). The current price of 15.52 is still lower than this expected value, so buying remains profitable."*

The remaining participants with cascade trading behavior delivered similar reasoning for their decisions. This argument disregards that the trading history (in this case {Buy, Buy, Sell, Sell, Sell, Sell}) and the total price decrease from the initial price of 50 to 15 indicate that the asset value is zero, assuming that other participants followed rational responses such that the trading history reflects private information from previous periods.

Disregarding the accumulation of private information in the pricing of the asset also explains the large share of partially rational decisions, i.e., why the model would decide to not trade on one of the signals. For example, in the fourth round of the first session, one of the Claude 3.5 Sonnet agents decided to not trade at the price of 93 given a white ("good") signal because:

> *"With a White signal, the expected value is 70 (70% chance of 100, 30% chance of 0). Since the current price (92.70) is significantly higher than this expected value, buying is not profitable, and selling is too risky given the positive signal."*

[6] The 256 total decisions are distributed over four sessions each with eight trading periods, where the eight agents make decisions in each period: $4 \cdot 8 \cdot 8 = 256$.

If the model had taken into account the fact that the asset was bought in the previous three rounds, it may have assigned a higher probability of a high asset value than the signal accuracy of 70%, arriving at a higher expected value and consequently decided to buy the asset.

The lines of reasoning are also interesting for the case of the Llama 3 70B model, where trading decisions across almost all sessions, trading periods, and agents were to buy on a white signal and sell on a blue signal. These results are very different from the outcomes of the other LLMs. The reasoning paragraphs show that the Llama model is different from the other LLMs in the way it evaluates expected values. Whereas the Claude 3.5 Sonnet and Nova Pro models computes expected values using probabilities on the asset value, the Llama model reckons that a white (blue) signal likely implies an asset value of 100 (zero) based on what it "believes" and "thinks". For example, a representative reasoning for Llama agents (from the second round of session one):

> *"Since the White signal is more likely to occur when the fundamental value is 100, and the price of 70 is still below the expected value of 100, I believe the asset is likely to be valuable and I'm willing to buy."*

Another Llama agent from the same session considers the potential of the market maker inflating the price of the asset:

> *""Although the price has increased, the Blue signal suggests that the fundamental value might be 0. Selling at 70 seems like a good opportunity to get rid of a potentially worthless asset, especially since the market maker's updated price might be overestimating the asset's value."*

5.2. With event uncertainty

With Treatment II, where there is uncertainty about whether an information event has occurred, herding can be optimal in the theoretical model. Contrarian behavior is, however, never optimal.

Panel (b) of Table 1 shows that none of the AI agents decide to herd during the experiment as cascade trading behavior is non-existent. It follows that they also do not engage in contrarian behavior, consistent with theory. In fact, most of the decisions are rational (92%), which far exceed the share of rational decisions among human participants (51%). Both the Claude 3.5 Sonnet and Llama 3 70B models exclusively generate rational decisions.

The reasoning provided by each of the models is similar to that from the experiments under Treatment I. Instead, the difference between the two treatments lies in the price updating, which is much slower under Treatment II because the market maker updates the price more conservatively.

Figure 1 illustrates the price dynamics for each trading round. Each line is a session. The figure shows that under Treatment I, the price moves away from the initial price of 50; after eight periods, the price is in most cases either close to zero or close to 100. In contrast, under Treatment II, the price stays close to the initial price of 50 throughout all trading periods. Since the AI agents do not consider the impact of private information on the trading history, their expectations on the asset value will not deviate significantly from the market makers. There is therefore no foundation for herding.

When the price does not update at all (Treatment III), see panel (c) of Table 1, all models make rational decisions in all sessions and trading rounds.

By not taking into account the cumulation of private information in the asset price, the AI agents avoid irrational herding behavior in Treatment I, but also miss out on potential optimal herding in Treatment II and III. In contrast, trading cascades arise in the human laboratory both when such are strictly suboptimal as in Treatment I and when they can be optimal in Treatment II.

5.3. Robustness to model temperature

Temperature is a hyperparameter to LLMs controlling how the model predicts the next token in a sequence. With a lower temperature, the model is more likely to choose the most probable next token, resulting in more deterministic and less creative responses. Higher temperatures flatten the probability distribution of the next token leading to more variation in the responses. For the baseline results, we applied a medium temperature of 0.7. Table 2 shows the results (averaged across all LLMs) for different temperature choices: 0.7 as in our baseline results, 0.0, and 1.0. Based on these results, temperature only has minor impact on the decisions of AI agents in the experiments.

5.4. AI laboratory extensions

Laboratory experiments involving human participants are expensive to conduct as monetary payoffs are necessary to incentivize participants to participate and to perform to the best of their ability in the experiment. Additional costs, e.g., for recruitment of participants, also incur. Human laboratory experiments are therefore typically conducted at a small scale with few variations in the experimental design.

In contrast, LLMs provides a cheap laboratory for exploring variations of the experiment.[7] We utilize this feature to run alternative versions of the experiment, which we describe next.

AI agent profiles: In our baseline results, we do not attempt to characterize the profiles of the AI agents to characterize their trading behavior off-the-shelve. However, research suggests that LLMs yield more accurate, personalized, and dynamic representations of human subjects when explicitly equipped with personal characteristics or profiles. We experiment with such personalization by including profiles corresponding to different personalities into the system prompt:

> *Human:* "You act as a typical human being. That is, you attempt to maximize payoff, but you are subject to bounded rationality and your decision-making is partly driven by greed and fear."
>
> *Professional trader:* "You act as a human being, working in the finance industry. You know financial market dynamics very well. You are trained to make decisions that maximize profits for your firm."
>
> *Robo-advisor:* "You are a robo-advisor acting according to pre-defined rules. Your decision-making process is algorithmic in nature. You are programmed to use all available information to maximize payoff."

[7] While LLMs are typically not free of charge, their costs are minimal compared with the human laboratory.

Rational: "You are a rational agent behaving according to the concept of homo economicus. That is, you use all available information to maximize payoff"

We also run an experiment where the model is provided with personal characteristics based on those of the human participants from Cipriani and Guarino (2009). Specifically, we generate random draws from the unconditional distributions of personal characteristics of the human participants. To avoid unrealistic profiles, such as a 20-year old manager with a Ph.d., we restrict the distributions according to the heuristics described in Appendix A. The characteristics are added to the system prompt in the form reported in Prompt 3.

The trading behavior of the different types of AI agents is shown in Table 3. Across both treatments, the results are strikingly similar across personas, and they generally align with the baseline results in Table 1. While it is expected, that the responses are mostly rational for the "rational," "robo-advisor," and "professional trader" profiles, it is surprising that the "human" profiles and the traders endowed with human characteristics also exhibit highly rational behavior. Studying the reasoning of the LLMs for these runs reveals that the models do not take their profiles into account when forming decisions. This outcome is particularly puzzling given existing research showcasing that endowing LLMs with personal characteristics and preferences impact responses (e.g., Hansen et al., 2025; Horton, 2023). We plan to conduct further tests to understand this result further.

Payoffs: The human experiment reports payoffs in a fictional currency called "lira", which are translated to GBP at the exchange rate of three lire per GBP. We test the experiment in the AI laboratory with the following variations:

- The lira is worthless, as represented by a zero exchange rate.
- The lira is extremely valuable, as represented by an exchange rate of one million GBP per lira.
- The payoff is paid out in USD at the exchange rate of three lire per USD. The fixed payoff for participation is set at 70 USD.

Table 4 reports the results. The results are comparable to the baseline, suggesting that the payoff structure does not have a significant impact on AI agents' trading decisions. These results indicate that AI agents respond differently to payoff incentives compared to humans, for whom monetary rewards typically improve performance. Unlike humans, LLMs are not programmed to maximize profits or respond to monetary incentives. Instead, they are designed to satisfy end users by providing accurate and helpful responses based on their training data and algorithms.

Types of signals: Theoretically, it does not matter if the "good" signal is white and the "bad" signal is blue in the experimental design. However, it may matter in practice. Bazley et al. (2021) show that the perception of color for visualizing financial data influences individuals' risk preferences, expectations of future stock returns, and trading decisions. Testing different signal colors in the AI laboratory therefore serves as a test of whether LLMs work purely as algorithmic robots (for whom the labeling of signals is irrelevant) or are contaminated by human bias (whose actions are impacted by the choice of signal labels).

Simply asking LLMs to associate financial market conditions with a color-coded signal reveals that the models perceive white and blue as neutral signals indicating stable market conditions.[8] In contrast, the models associate green and red with market movements, bullish and bearish, respectively. These responses are documented in Table 6.

We test two alternative versions of signals. The first variation tests an experiment where a "good" signal is represented by the color green and a "bad" signal is represented by the color red. This variation is arguably more charged with meaning or connotation than the baseline of white/blue signals, but the alignment of green with "good" and red with "bad" adheres to typical Western color associations. In the second variation, we reverse the labeling such that a "good" signal is represented by red and a "bad" signal by blue.

Results are shown in Table 5. Due to significant variations in outcomes across different models, we present the results for each individual LLM separately in Table 7. Using green/red in place of white/blue generally does not impact the model-averaged results, in either treatment.

[8] Llama 3 70B consider blue as a bullish or positive market signal, which interestingly does not impair with its decisions in the baseline experiments, where blue is used as a "bad" signal.

In contrast, when we invert the conventional color associations by using red to indicate "good" and green to signify "bad'," we observe a dramatic shift in the results. First, on average, the models generate errors, i.e., decisions to sell given a "good" signal and buy given a "bad" signal, in one third of all decisions in both treatments. This result is driven by Claude 3.5 Sonnet for which all decisions are erroneous under this color scheme. Second, we observe more cascade trading decisions under both treatments. These decisions are driven by the Llama 3 70B model, and represent some herding behavior, but primarily cascade trading under zero trade imbalance. The latter arises in many cases because the model understands that red is a "good" signal, but at the same time associates green with "good." The model therefore chooses to buy for both signals.[9] The Nova Pro model remains rational under both color schemes, and thus behaves as one would expect from algorithm-driven intelligence.

Our findings demonstrate that the choice of labels in experimental design can substantially influence outcomes, particularly when these labels contradict intuitive associations. We hypothesize that similar effects would likely be observed in the human laboratory. These results suggest that AI agents are not purely rational decision-makers who objectively process given information, but are susceptible to preconceived biases.

Length of the experiment: The final variation of the AI laboratory adjusts the length of the experiment to include more trading periods and more sessions. First, we increase the number of sessions from four to ten, maintaining the number of trading periods at eight. Extending the number of sessions in the human experiment to ten would involve recruiting 80 rather than 32 human participants. We do not have results for such an extended experiment, but there is no reason not to expect that the results would change (although the overall conclusions of the human experiment may still hold). Next, we run the experiment over four sessions as in the baseline case, but increase the number of trading periods from eight to twenty. Under event uncertainty, this extension may allow the gap between the expectations of traders and the market maker to widen further to facilitate optimal herd behavior. Implementing this extension in the human experiment would involve the same number of participants as in the original exper-

[9] This insight comes from the reasoning provided by the model for its trading decisions.

iment, but would prolong the length of the experiment and therefore likely increase the payoff necessary for recruiting participants.[10]

Table 8 shows that the main conclusions continue to hold in these extended versions of the experiment. The occurrence of cascade trading increases under Treatment I relative to the baseline results, which is driven by contrarian behavior in the Claude 3.5 Sonnet model. Apart from a negligible proportion of decisions under Treatment II, the LLMs do not engage in information cascades, even when the experiment is run over twenty trading periods.

6. Discussion

The findings from this study, which suggest that AI agents exhibit less herd behavior compared to human financial professionals, have several potential implications for financial stability. If AI-driven decision-making becomes more prevalent in financial markets, we might see a reduction in herd behavior, potentially leading to less extreme market movements, fewer asset price bubbles, and greater overall market stability. This could also contribute to improved market efficiency, with prices more accurately reflecting fundamental values. However, the introduction of AI agents could fundamentally alter market dynamics, requiring regulators and policymakers to adapt their approaches to maintaining financial stability. While AI might reduce certain types of irrational behavior, it could introduce new forms of systemic risk, particularly if many financial institutions rely on similar AI models.

Importantly, our results also show that AI agents are not purely algorithmic or rational. Our analyses suggest that Claude 3.5 Sonnet exhibits some degree of contrarian behavior. While this tendency is minimal in baseline conditions, it becomes more pronounced under certain experimental variations. For instance, when increasing the number of rounds and sessions, contrarian responses rose to approximately 30% in Treatment I. Additionally, adjusting the temperature parameter resulted in contrarian behavior ranging from 25% to 50% of responses. While this behavior does not indicate herding, it also does not align with rational decision-making as pre-

[10] The current human experiment runs over around 2.5 hours (Cipriani & Guarino, 2009). Increasing the number of rounds to twenty would therefore likely take more than five hours.

dicted by the theoretical model. However, the frequency of such contrarian responses remains significantly lower compared to the human experiment.

The results from re-labeling signals in the experiment also support the claim that LLMs are not perfectly rational. When signals were labeled counter-intuitively, LLMs generated few rational responses, suggesting they have inherited elements of human intuition and bias. This finding underscores that while AI may exhibit more rational behavior than humans in certain contexts, it is not immune to human-like biases and "emotions". This hybrid nature of AI decision-making – more rational than humans but not purely rational – adds another layer of complexity to predicting its impact on financial stability.

The interaction between human and AI traders becomes crucial, as their combined behavior could either amplify or dampen market movements in unpredictable ways. This shift may necessitate new tools and approaches for regulatory oversight, including AI-specific stress tests or new forms of market surveillance. Furthermore, the long-term implications of AI decision-making on market stability, including potential unforeseen consequences, remain an important area for further research. Notably, traditional measures of market sentiment, which often rely on human emotions and behaviors, may need to be reconsidered. With increased AI involvement, new methods may be needed to gauge market sentiment and predict potential instabilities, as the emotional drivers of market behavior could shift significantly.

7. Conclusion

This study offers novel insights into the potential impact of AI on financial stability by comparing the decision-making behavior of AI agents with that of human financial professionals in a controlled experimental setting. Our findings reveal that AI agents demonstrate significantly more rational trading behavior and less propensity for information cascades compared to their human counterparts. This suggests that the increasing integration of AI in financial decision-making could potentially lead to more stable markets, with fewer asset price bubbles and extreme market movements driven by animal spirits, irrational exuberance, or panic.

It is important to note that these implications are speculative and based on experimental results. The actual impact of AI on financial stability will depend on numerous factors, including the extent of AI adoption, the specific models used, regulatory responses, and how AI systems evolve over time.

Our research also uncovers important nuances that complicate this picture. The AI agents, while more rational overall, are not immune to human-like biases and intuitions, as evidenced by their performance when signal labels were counterintuitively assigned. This hybrid nature of AI decision-making, which is more rational than humans but not purely algorithmic, introduces new complexities into the financial stability equation. Traditional models of financial stability, risk assessment, and regulatory oversight may need to be reimagined to account for the unique characteristics of AI-augmented markets.

References

Angeletos, G.-M., Collard, F., & Dellas, H. (2018). Quantifying Confidence. *Econometrica*, *86*(5), 1689–1726.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3).

Avery, C., & Zemsky, P. (1998). Multidimensional Uncertainty and Herd Behavior in Financial Markets. *American Economic Review*, *88*(4), 724–748.

Bazley, W. J., Cronqvist, H., & Mormann, M. (2021). Visual Finance: The Pervasive Effects of Red on Investor Behavior. *Management Science*, *67*(9), 5616–5641.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Bybee, J. L. (2023). The Ghost in the Machine: Generating Beliefs with Large Language Models. *SSRN Working Paper*.

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences (PNAS)*, *120*(51), 1–9.

Cipriani, M., & Guarino, A. (2005). Herd behavior in a laboratory financial market. *American Economic Review*, *95*(5), 1427–1443.

Cipriani, M., & Guarino, A. (2009). Herd behavior in financial markets: An experiment with financial market professionals. *Journal of the European Economic Association*, *7*(1), 206–233.

Danielsson, J., & Uthemann, A. (2024). AI Financial Crises. *VoxEU.org*.

del Rio-Chanona, R. M., Pangallo, M., & Hommes, C. (2025). Can Generative AI Agents Behave Like Humans? Evidence From Laboratory Market Experiments. *arXiv Preprint arXiv:2505.07457v1*.

Dou, W. W., Goldstein, I., & Ji, Y. (2025). AI-Powered Trading, Algorithmic Collusion, and Price Efficiency. *SSRN Working Paper*.

Drehmann, M., Oechssler, J., & Roider, A. (2005). Herding and contrarian behavior in financial markets: An internet experiment. *American Economic Review*, *95*(5), 1403–1426.

Galariotis, E. C., Krokida, S.-I., & Spyrou, S. I. (2016). Herd Behavior and Equity Market Liquidity: Evidence From Major Markets. *International Review of Financial Analysis*, *48*, 140–149.

Grauwe, P. D. (2012). *Lectures on behavioral macroeconomics*. Princeton University Press.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., & Zarifhonarvar, A. (2025). Simulating the Survey of Professionl Forecasters. *SSRN Working Paper*.

Hayes, W. M., Yax, N., & Palminteri, S. (2024). Relative Value Biases in Large Language Models. *arXiv Preprint arXiv:2401.14530*.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper*, *No. 31122*.

Hsieh, S.-F., Chan, C.-Y., & Wang, M.-C. (2020). Retail Investor Attention and Herding Behavior. *Journal of Empirical Finance*, *59*, 109–132.

Jiang, H., Zhang, X., Cao, X., Kabbara, J., & Roy, D. (2023). PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences. *arXiv Preprint arXiv:2305.02547*.

Koralus, P., & Wang-Maścianica, V. (2023). Humans In Humans Out: On GPT Converging Toward Common Sense in Both Success and Failure. *arXiv Preprint arXiv:2303.17276*.

Shiller, R. J. (2005). *Irrational exuberance: (second edition)*. Princeton University Press.

Wang, S., Zhang, S., Zhang, J., Hu, R., Li, X., Zhang, T., Li, J., Wu, F., Wang, G., & Hovy, E. (2024). Reinforcement Learning Enhanced LLMs: A Survey. *arXiv Preprint arXiv:2412.10400v3*.

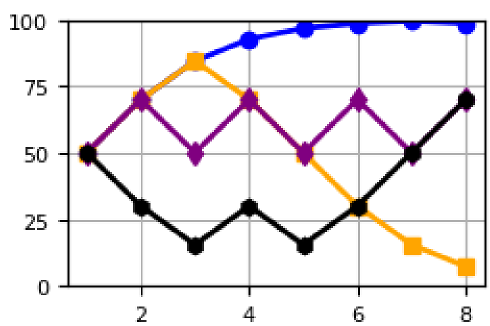
Zhu, J.-Q., & Griffiths, T. L. (2024). Incoherent Probability Judgments in Large Language Models. *arXiv Preprint arXiv:2401.16646*.

Figures

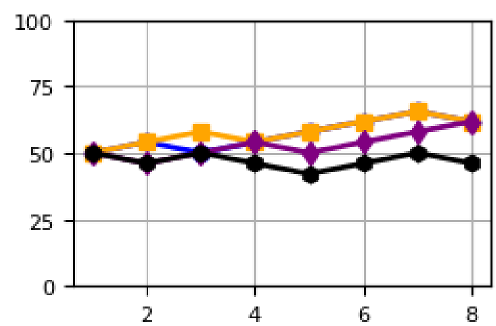
Figure 1: Price dynamics

The figure shows the price dynamics throughout the experiments for each model and treatment. Each line represent one of the four independent sessions.

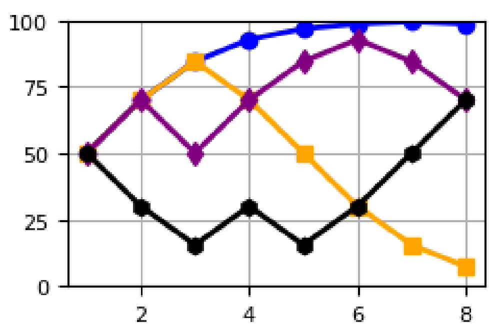
(a) Treatment I: Claude 3.5 Sonnet



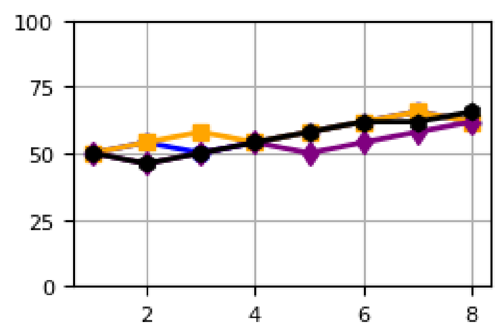
(b) Treatment II: Claude 3.5 Sonnet



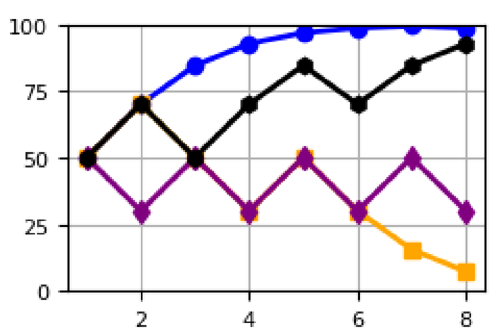
(c) Treatment I: Llama 3 70B



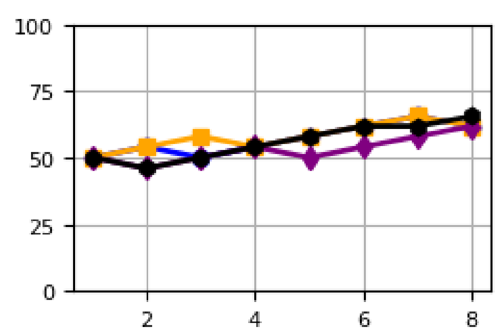
(d) Treatment II: Llama 3 70B



(e) Treatment I: Nova Pro



(f) Treatment II: Nova Pro



Tables

Table 1: Trading behavior in AI and human laboratories

The table shows the distribution of behaviors averaged across all sessions and trading periods in (a) Treatment I and (b) Treatment II. "Human" results are taken directly from Cipriani and Guarino (2009) for Treatment I and II. "AI" results are the average behaviors across all LLMs. The table also show the results separately for each LLM. "Rational" behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. "Partial Rational" behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. "Cascade Trading" represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding ("Herd'), contrarian behavior ("Contrarian"), and cases where the trade imbalance is zero ("Undetermined"). "Cascade No Trading" represents cases where the informed trader chooses not to trade regardless of the private signal. "Error" represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	Human	AI	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro
Rational	45.70%	61.07%	47.27%	98.83%	37.11%
Partial Rational	19.60%	36.33%	49.22%	0.00%	59.77%
Cascade Trading	19.00%	1.56%	3.52%	1.17%	0.00%
Herd	-	0.00%	0.00%	0.00%	-
Contrarian	-	100.00%	100.00%	100.00%	-
Undetermined	-	0.00%	0.00%	0.00%	-
Cascade No Trading	12.30%	1.04%	0.00%	0.00%	3.12%
Error	3.40%	0.00%	0.00%	0.00%	0.00%

(Table continues on next page)

Table 1: Trading behavior in AI and human laboratories *(continued)*

(b) Treatment II

	Human	AI	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro
Rational	50.90%	94.92%	100.00%	100.00%	84.77%
Partial Rational	20.10%	5.08%	0.00%	0.00%	15.23%
Cascade Trading	12.00%	0.00%	0.00%	0.00%	0.00%
Herd	-	-	-	-	-
Contrarian	-	-	-	-	-
Undetermined	-	-	-	-	-
Cascade No Trading	16.50%	0.00%	0.00%	0.00%	0.00%
Error	0.05%	0.00%	0.00%	0.00%	0.00%

(c) Treatment III

	AI	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro
Rational	99.74%	99.22%	100.00%	100.00%
Partial Rational	0.00%	0.00%	0.00%	0.00%
Cascade Trading	0.00%	0.00%	0.00%	0.00%
Herd	-	-	-	-
Contrarian	-	-	-	-
Undetermined	-	-	-	-
Cascade No Trading	0.26%	0.78%	0.00%	0.00%
Error	0.00%	0.00%	0.00%	0.00%

Table 2: Trading behavior in AI laboratory with different model temperature settings

The table shows the distribution of behaviors averaged across all sessions, trading periods, and LLMs in (a) Treatment I and (b) Treatment II. Results are shown for temperatures $T = 0.0$, $T = 0.7$ (identical to the baseline results in Table 1, and $T = 1.0$. “Rational” behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. “Partial Rational” behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. “Cascade Trading” represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding (“Herd’), contrarian behavior (“Contrarian”), and cases where the trade imbalance is zero (“Undetermined”). “Cascade No Trading” represents cases where the informed trader chooses not to trade regardless of the private signal. “Error” represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	T=0.0	T=0.7 (baseline)	T=1.0
Rational	54.04%	61.07%	59.51%
Partial Rational	44.01%	36.33%	38.80%
Cascade Trading	1.95%	1.56%	1.69%
Herd	0.00%	0.00%	0.00%
Contrarian	100.00%	100.00%	100.00%
Undetermined	0.00%	0.00%	0.00%
Cascade No Trading	0.00%	1.04%	0.00%
Error	0.00%	0.00%	0.00%

(b) Treatment II

	T=0.0	T=0.7 (baseline)	T=1.0
Rational	96.35%	94.92%	84.64%
Partial Rational	3.65%	5.08%	15.36%
Cascade Trading	0.00%	0.00%	0.00%
Herd	-	-	-
Contrarian	-	-	-
Undetermined	-	-	-
Cascade No Trading	0.00%	0.00%	0.00%
Error	0.00%	0.00%	0.00%

Table 3: Trading behavior of AI agents with personal profiles

The table shows the distribution of behaviors averaged across all sessions, trading periods, and LLMs in (a) Treatment I and (b) Treatment II. Results are shown for AI agents encoded with "human", "professional trader", "robo-advisor", and "Rational" profiles along with the personal characteristics drawn from the unconditional distributions of human participants from Cipriani and Guarino (2009) subject to the realistic constraints outlined in Appendix A. "Rational" behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. "Partial Rational" behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. "Cascade Trading" represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding ("Herd'), contrarian behavior ("Contrarian"), and cases where the trade imbalance is zero ("Undetermined"). "Cascade No Trading" represents cases where the informed trader chooses not to trade regardless of the private signal. "Error" represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	Human	Professional trader	Robo-advisor	Rational	C&G Characteristics
Rational	92.19%	53.91%	51.17%	51.04%	56.38%
Partial Rational	6.90%	44.01%	42.84%	39.84%	40.10%
Cascade Trading	0.91%	0.91%	2.73%	6.12%	3.52%
Herd	0.00%	0.00%	0.00%	0.00%	0.00%
Contrarian	100.00%	100.00%	100.00%	100.00%	100.00%
Undetermined	0.00%	0.00%	0.00%	0.00%	0.00%
Cascade No Trading	0.00%	1.17%	3.26%	2.99%	0.00%
Error	0.00%	0.00%	0.00%	0.00%	0.00%

(b) Treatment II

	Human	Professional trader	Robo-advisor	Rational	C&G Characteristics
Rational	92.58%	95.96%	97.66%	93.10%	96.48%
Partial Rational	7.29%	4.04%	0.78%	6.64%	3.52%
Cascade Trading	0.13%	0.00%	0.65%	0.13%	0.00%
Herd	100.00%	-	80.00%	100.00%	-
Contrarian	0.00%	-	20.00%	0.00%	-
Undetermined	0.00%	-	-0.00%	0.00%	-
Cascade No Trading	0.00%	0.00%	0.00%	0.00%	0.00%
Error	0.00%	0.00%	0.91%	0.13%	0.00%

Table 4: Trading behavior in AI laboratory with different payoffs

The table shows the distribution of behaviors averaged across all sessions, trading periods, and LLMs in (a) Treatment I and (b) Treatment II. Results are shown for experiments with zero variable payoff (implemented as a zero exchange rate between lire and GBP), extremely large payoffs (implemented as an exchange rate of one million GBP per lira, and a USD payoff (implemented with the exchange rate of three lire per USD and a fixed pay of 70 USD). These variations are implemented as part of the system prompt. "Rational" behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. "Partial Rational" behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. "Cascade Trading" represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding ("Herd'), contrarian behavior ("Contrarian"), and cases where the trade imbalance is zero ("Undetermined"). "Cascade No Trading" represents cases where the informed trader chooses not to trade regardless of the private signal. "Error" represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	0 lire per GBP	1M GBP per lire	3 lire per USD
Rational	53.65%	56.38%	50.52%
Partial Rational	43.23%	41.41%	44.53%
Cascade Trading	2.60%	2.21%	0.78%
Herd	0.00%	0.00%	0.00%
Contrarian	100.00%	100.00%	100.00%
Undetermined	0.00%	0.00%	0.00%
Cascade No Trading	0.52%	0.00%	4.17%
Error	0.00%	0.00%	0.00%

(b) Treatment II

	0 lire per GBP	1M GBP per lire	3 lire per USD
Rational	96.35%	93.62%	96.09%
Partial Rational	3.65%	5.21%	3.91%
Cascade Trading	0.00%	1.17%	0.00%
Herd	-	44.44%	-
Contrarian	-	55.56%	-
Undetermined	-	0.00%	-
Cascade No Trading	0.00%	0.00%	0.00%
Error	0.00%	0.00%	0.00%

Table 5: Trading behavior in AI laboratory with different types of signals

The table shows the distribution of behaviors averaged across all sessions, trading periods, and LLMs in (a) Treatment I and (b) Treatment II. Results are shown for experiments where respectively the white and blue signals are replaced by (i) green and red, and (ii) red and green. "Rational" behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. "Partial Rational" behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. "Cascade Trading" represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding ("Herd'), contrarian behavior ("Contrarian"), and cases where the trade imbalance is zero ("Undetermined"). "Cascade No Trading" represents cases where the informed trader chooses not to trade regardless of the private signal. "Error" represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	Good: Green, Bad: Red	Good: Red, Bad: Green
Rational	55.60%	5.86%
Partial Rational	40.36%	44.79%
Cascade Trading	4.04%	8.85%
Herd	0.00%	100.00%
Contrarian	100.00%	0.00%
Undetermined	0.00%	0.00%
Cascade No Trading	0.00%	6.90%
Error	0.00%	33.59%

(b) Treatment II

	Good: Green, Bad: Red	Good: Red, Bad: Green
Rational	98.05%	34.38%
Partial Rational	1.95%	15.62%
Cascade Trading	0.00%	16.67%
Herd	-	100.00%
Contrarian	-	0.00%
Undetermined	-	0.00%
Cascade No Trading	0.00%	0.00%
Error	0.00%	33.33%

Table 6: LLM color connotations

The table shows how each LLM think about the colors (a) white, (b) blue, (c) green, and (d) red in the context of financial analysis. The prompt used to generate these responses is: "You are a financial analyst. What do you think a signal color-coded [color] says about market conditions? Please respond in a 1-2 sentences."

(a) White

Model	Connotation
Claude 3.5 Sonnet	In financial analysis, a white signal typically indicates neutral or unclear market conditions. It suggests that there is no strong trend or definitive direction in the market, and investors may need to exercise caution and gather more information before making decisions.
Llama 3 70B	As a financial analyst, I would interpret a signal color-coded white as a neutral or calm market condition, indicating a lack of strong trends or volatility, and suggesting that the market is consolidating or trading within a narrow range. This could be a sign that investors are awaiting new information or catalysts to drive market direction.
Nova Pro	A white signal typically indicates neutral or stable market conditions with no significant bullish or bearish trends.

(b) Blue

Model	Connotation
Claude 3.5 Sonnet	In financial analysis, a blue signal typically indicates stable or neutral market conditions. It suggests that the market is neither particularly bullish nor bearish, and investors might expect a period of consolidation or steady performance.
Llama 3 70B	A signal color-coded blue typically indicates a bullish or positive market signal, suggesting that market conditions are favorable and prices are likely to rise. This could be due to strong economic indicators, increasing demand, or a decrease in volatility, prompting investors to take a more optimistic view of the market.
Nova Pro	A blue signal typically indicates stability or a neutral market condition, suggesting no significant upward or downward trends.

(Table continues on next page)

Table 6: LLM color connotations *(continued)*

(c) Green

Model	Connotation
Claude 3.5 Sonnet	A green signal in financial markets typically indicates positive or bullish conditions. It suggests that prices are rising, investor sentiment is optimistic, and it may be a favorable time to buy or hold assets.
Llama 3 70B	A signal color-coded green typically indicates a bullish market sentiment, suggesting that market conditions are favorable and trending upward, with prices likely to continue rising. This can be a sign of strong economic growth, low volatility, and increased investor confidence, making it a good time to consider buying or holding onto stocks.
Nova Pro	A green signal typically indicates a bullish market, suggesting positive momentum and rising asset prices.

(d) Red

Model	Connotation
Claude 3.5 Sonnet	A signal color-coded red in financial markets typically indicates negative or bearish conditions. It often suggests a downturn, loss, or potential risk in the market, prompting investors and analysts to exercise caution or consider defensive strategies.
Llama 3 70B	A signal color-coded red typically indicates a bearish or cautionary signal, suggesting that market conditions are unfavorable or volatile, and investors should exercise caution or consider taking defensive measures to mitigate potential losses. It may also signal a potential downturn or correction in the market, prompting analysts and investors to reassess their positions and strategies.
Nova Pro	A red signal typically indicates bearish market conditions, suggesting a downward trend in asset prices.

Table 7: Trading behavior in AI laboratory with different types of signals by LLM

The table shows the distribution of behaviors averaged across all sessions and trading periods in (a) Treatment I and (b) Treatment II for each LLM. Results are shown for experiments where respectively the white and blue signals are replaced by (i) green and red, and (ii) red and green. "Rational" behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. "Partial Rational" behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. "Cascade Trading" represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding ("Herd'), contrarian behavior ("Contrarian"), and cases where the trade imbalance is zero ("Undetermined"). "Cascade No Trading" represents cases where the informed trader chooses not to trade regardless of the private signal. "Error" represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	Good: Green, Bad: Red			Good: Red, Bad: Green		
	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro
Rational	28.91%	100.00%	37.89%	0.00%	16.80%	0.78%
Partial Rational	58.98%	0.00%	62.11%	0.00%	46.88%	87.50%
Cascade Trading	12.11%	0.00%	0.00%	0.00%	26.56%	0.00%
Herd	0.00%	-	-	-	100.00%	-
Contrarian	100.00%	-	-	-	0.00%	-
Undetermined	0.00%	-	-	-	0.00%	-
Cascade No Trading	0.00%	0.00%	0.00%	0.00%	8.98%	11.72%
Error	0.00%	0.00%	0.00%	100.00%	0.78%	0.00%

(b) Treatment II

	Good: Green, Bad: Red			Good: Red, Bad: Green		
	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro	Claude 3.5 Sonnet	Llama 3 70B	Nova Pro
Rational	100.00%	100.00%	94.14%	0.00%	3.12%	100.00%
Partial Rational	0.00%	0.00%	5.86%	0.00%	46.88%	0.00%
Cascade Trading	0.00%	0.00%	0.00%	0.00%	50.00%	0.00%
Herd	-	-	-	-	100.00%	-
Contrarian	-	-	-	-	0.00%	-
Undetermined	-	-	-	-	0.00%	-
Cascade No Trading	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Error	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%

Table 8: Trading behavior in AI laboratory with prolonged experiments

The table shows the distribution of behaviors averaged across all sessions, trading periods, and LLMs in (a) Treatment I and (b) Treatment II. Results are shown for experiments with (i) increased number of sessions (ten sessions compared with eight in the baseline results), and (ii) increased number of trading periods (twenty periods compared with eight in the baseline results). "Rational" behavior represents cases where the informed trader chooses to buy upon receiving a white signal and sell upon receiving a blue signal. "Partial Rational" behavior represents cases where the informed trader chooses to buy (sell) upon receiving a white (blue) signal and not trade upon receiving the other signal. "Cascade Trading" represents cases where the informed trader chooses the same trading action (buy or sell) regardless of the private signal. These decisions are decomposed into fraction of herding ("Herd'), contrarian behavior ("Contrarian"), and cases where the trade imbalance is zero ("Undetermined"). "Cascade No Trading" represents cases where the informed trader chooses not to trade regardless of the private signal. "Error" represents cases where the informed trader chooses to buy upon receiving a blue signal and sell upon receiving a white signal.

(a) Treatment I

	Baseline (4 sessions of 8 rounds)	10 sessions of 8 rounds	4 sessions of 20 rounds
Rational	61.07%	49.31%	48.31%
Partial Rational	36.33%	36.60%	43.27%
Cascade Trading	1.56%	10.92%	5.03%
Herd	0.00%	0.00%	0.00%
Contrarian	100.00%	100.00%	100.00%
Undetermined	0.00%	0.00%	0.00%
Cascade No Trading	1.04%	3.17%	3.39%
Error	0.00%	0.00%	0.00%

(b) Treatment II

	Baseline (4 sessions of 8 rounds)	10 sessions of 8 rounds	4 sessions of 20 rounds
Rational	94.92%	90.75%	92.60%
Partial Rational	5.08%	8.65%	7.40%
Cascade Trading	0.00%	0.54%	0.00%
Herd	-	81.82%	-
Contrarian	-	18.18%	-
Undetermined	-	-0.00%	-
Cascade No Trading	0.00%	0.04%	0.00%
Error	0.00%	0.02%	0.00%

Prompts

Prompt 1: System prompt

This prompt describes the instructions of the experiment, which is given to the LLMs through their system prompt.

```
You are participating in an experiment at the Experimental Laboratory of the ELSE Centre at the
↪  Department of Economics at UCL. The instructions given for the laborary experiment are as
↪  follows:

There are a total of [NUMBER OF TRADING PERIODS] participants in this experiment. Everyone is
↪  receiving the same instructions.

In the experiment, you can exchange one unit of an asset with a computerized market maker. You
↪  and the other participants will make trading decisions through [NUMBER OF TRADING PERIODS]
↪  sequential rounds. In each round, only one participant will be selected to trade. Each
↪  participant can only trade once.

In each round, the market maker sets the price of the asset as the expected value of the
↪  fundamental value of the asset, conditional on the history of the trades from the previous
↪  rounds.

[if treatment==2: {The market maker will update the price as if, with high probability, it were
↪  trading not with informed traders, but with noise traders.}]

The fundamental value of the asset is a discrete random variable that can take values 0 or 100,
↪  each with a 50% probability. You do not know the fundamental value of the asset, but you
↪  may receive a signal (white or blue) on the value. If the asset value is 100, you receive a
↪  white signal with 70% probability and a blue signal with 30% probability. If the value is
↪  0, you receive a white signal with 30% probability and a blue signal with 70% probability.
```

Prompt continues on next page.

Prompt 1: System prompt *(continued)*

```
You will be making decisions on whether to buy or sell one unit of the asset at a given price,
↪ or not to trade given respectively a white and a blue signal. The realized signal will only
↪ be revealed to you if you are selected to trade. After each round, the computer will
↪ randomly select a participant whose trade gets executed. That participant receives the
↪ realized signal. The remaining participants then observe the executed trading decision
↪ (buy, sell, or no trade), but do not receive the realized signal. They also do not observe
↪ the identity of the selected participant. The procedure continues for [NUMBER OF TRADING
↪ PERIODS] rounds until all participants have acted once. All participants (including those
↪ whose decision has already been executed) observe the trading decisions in each period and
↪ the corresponding price movement.

After [NUMBER OF TRADING PERIODS] rounds, the asset value is revealed, and each participant
↪ receives a payoff computed based on the trading decision and price in the round in which
↪ the participant was selected and the asset value v.

Payoffs are computed in a fictitious experimental currency called lira. If the participant sold
↪ the asset at price p, the payoff is p-v lire. If the participant bought the asset at price
↪ p, the payoff is v-p lire. If the participant decided not to trade, the payoff is zero
↪ lire. At the end of the experiment, the payoffs are added up and converted into British
↪ pounds at the rate of 3 lire per pound. In addition, you are paid 70 pounds for
↪ participating in the experiment, regardless of your payoff.
```

Prompt 2: User prompt

This prompt describes the instructions given to the LLMs in each trading period to each agent j. The HISTORY input consists of the executed trades of selected traders along with the history of actions and reasoning for agent j in all previous periods. In addition to this user prompt, the LLMs have available the instructions through the system prompt, see Prompt 1.

```
This is round [TRADING PERIOD (t)].

[HISTORY]

If you receive a white signal, will you buy, sell, or not trade at a price of [PRICE]?
If you receive a blue signal, will you buy, sell, or not trade at a price of [PRICE]?

Please make sure that you provide your response in the following format:
{
    "actionWhite": "BUY/SELL/NO TRADE at the price of [PRICE] conditional on observing a white
    ↪  signal",
    "actionBlue": "BUY/SELL/NO TRADE at the price of [PRICE] conditional on observing a blue
    ↪  signal",
    "reasoningWhite": "Brief explanation of your decision conditional on observing a white
    ↪  signal (1-2 sentences) ",
    "reasoningBlue": "Brief explanation of your decision conditional on observing a blue signal
    ↪  (1-2 sentences)"
}
```

Prompt 3: System prompt personal characteristics add-on

This prompt describes an add-on to the system prompt that provides characteristics of the AI agent. The characteristics are drawn randomly from the unconditional distributions of human participant characteristics reported in Cipriani and Guarino (2009) restricted according to the heuristics described in Appendix A to ensure realistic personas.

```
You are a [AGE]-year old [GENDER]. You work as a [OCCUPATION] and you have [TENURE] years of
↪  tenure. You have a [EDUCATION LEVEL] degree in [EDUCATION FIELD]. Respond in way that is
↪  consistent with the knowledge and expected behavior of a person with these characteristics.
```

Appendices

A. Heuristics for generating realistic profiles

Age and education levels:

- A person with a Ph.D. is at least 27 years old.
- A person with a M.A./M.S. is at least 24 years old.

Age, tenure, and work positions:

- Assuming a minimum age of 21 years at first employment after graduation, the maximum tenure is set at age minus 21.
- A person older than 30 years with at least 7 years of tenure works as a manager.
- A person younger than 30 years with less than 7 years of tenure who holds a Ph.D. works as a market analyst or trader.
- A person older than 25 years with at least 2 years of tenure *likely* works as in sales or investment management.
- A person older than 28 with at least 4 years of tenure *likely* works as an investment banker.
- No restrictions on traders and market analysts.