Ex Machina: Financial Stability in the Age of Artificial Intelligence

Kartik Anand* Sophia Kazinnik†
Agnese Leonello‡ Ettore Panetti§

March 25, 2026

Abstract

Does artificial intelligence (AI) pose a threat to financial stability? We study AI investor behavior, specifically Q-learning and large language model (LLM) investors, in a mutual fund redemption problem with economic and strategic uncertainty. Different AI architectures generate systematically different outcomes. Q-learning investors coordinate well but under default risk exhibit excessive redemption that amplifies fragility. LLM investors internalize equilibrium structure but display belief heterogeneity, weakening coordination and predictability. Our findings show that AI architecture is a first-order determinant of financial stability.

Keywords: financial stability; strategic uncertainty; coordination games; Q-learning; large language models; AI agents.
JEL classifications: G01, G23, C63.

*Deutsche Bundesbank; kartik.anand@bundesbank.de
†Stanford University, HAI; kazinnik@stanford.edu
‡European Central Bank, DG-Research and CEPR; agnese.leonello@ecb.europa.eu
§University of Naples Federico II and CSEF; ettore.panetti@unina.it

We thank Wilko Bolt (discussant), Markus Brunnermeier, Emilio Calvano, Jean-Edouard Colliard, Mehmet Ekmekci, Thierry Foucault, Itay Goldstein, John Horton, Shumiao Ouyang (discussant), Tamas Vadasz (discussant) and seminar audiences at the Deutsche Bundesbank, the European Central Bank, the BIS/CEPR/ECB Conference "Technological innovations in financial markets: Risks and opportunities in banking and regulation", and the Frankfurt/Mannheim Workshop on Digital Finance for useful comments. The views expressed here are the authors' and do not reflect those of the European Central Bank, the Deutsche Bundesbank, or the Eurosystem.

I. Introduction

In ancient Greek tragedy, when a plot reached an impasse, a mechanical crane would lower actors playing gods onto the stage to deliver a resolution, a device known as *deus ex machina*. Two and a half millennia later, our evergrowing reliance on artificial intelligence (AI) invites a comparison. Is AI the *deus ex machina* of modern finance, helping us safely navigate financial complexity and uncertainty, or does it pose a threat to financial stability? We examine this question.

AI use in financial markets is growing rapidly, from more capable algorithmic trading to new tools such as generative AI. As large language models (LLMs) become more widely accessible, retail investors can more easily engage in complex transactions (OECD, 2023; Bick et al., 2024; Chatterji et al., 2025; Gambacorta et al., 2025). Meanwhile, emerging agentic systems can operate with minimal human oversight, adapt to new information and execute tasks autonomously (Nvidia, 2025; Financial Times, 2025). These trends have fueled debates on whether AI is simply a technological evolution or a shift with novel consequences for financial stability (Adrian et al., 2025; Foucault et al., 2025; Financial Stability Board, 2025; Cecchetti et al., 2025).

We contribute to this debate by examining how different AI-based investors behave in experimental environments that mirror the real world complexities underlying episodes of financial fragility. These episodes emerge when investors become concerned about the economic fundamentals, the actions of their peers, or both. Assessing how AI investors contribute to fragility thus requires understanding whether and how they respond to both sources of uncertainty. To this end, we place multiple AI investors as independent decision makers in a stylized coordination game and systematically compare two distinct classes of AI architecture: (i) Q-learning (QL) algorithms, which rely on model-free trial-and-error learning and are already widely used in algorithmic trading, and (ii) LLMs, which rely on contextual and chain-of-thought reasoning.[1]

Our analysis makes two main contributions. First, we show that AI architecture matters for financial stability. QL and LLM investors differ both in their tendency to coordinate and in how they react to changes in the economic environment. As a result, the likelihood and severity of financial distress depend not only on the economic environment but also on AI architecture. Second, we introduce a framework for analyzing the latent reasoning of LLM investors. Using model-generated chain-of-thought reasoning text, we extract beliefs, causal narratives, and deviations from rational expectations reasoning, making otherwise opaque

[1] AI architecture refers to the decision-making paradigm by which information is mapped into actions, and encompasses how strategies are formed, how information is represented and processed and the role of prior knowledge.

LLM decision-making more interpretable.[2] Taken together, our analysis points to the need for tools that assess how AI investors form beliefs and map information into actions.

We ground our analysis in the canonical mutual fund redemption model of Chen et al. (2010), a parsimonious variant of the models by Diamond and Dybvig (1983) and Goldstein and Pauzner (2005), which we use to derive predictions for benchmarking AI behavior. Our focus on non-bank financial institutions reflects where AI adoption is most advanced and accelerating. Recent developments underscore the speed and sophistication of AI adoption in asset management: DeepSeek's low-cost, high-performance LLM `R1` (Jan. 2025) intensified automation in market-data processing and signal extraction (Reuters, 2025), and the 2024 Jane Street–Millennium lawsuit offered a rare glimpse into closely guarded algorithmic trading strategies (Bloomberg, 2024).[3]

Beyond this empirical relevance, our modeling choice offers two analytical advantages. First, both economic and strategic uncertainty emerge naturally in this framework. Second, the underlying coordination game structure allows us to generalize our findings to related settings, such as bank runs, currency attacks, and stablecoin runs. In the model, investors choose whether to redeem their shares before the fund's investment project matures. Investors' payoffs depend on both economic fundamentals and the actions of other investors. Strategic complementarities imply that the incentive to redeem increases when more investors do so, while stronger fundamentals reduce the appeal to redeem. We examine equilibria under two dimensions of uncertainty: (i) payoff uncertainty, contrasting whether the fund's investment is subject to default risk or not, and (ii) fundamental uncertainty, where investors may or may not observe the true underlying fundamentals.

Our model delivers four testable predictions. First, in the absence of payoff and fundamental uncertainty, investors redeem when fundamentals are weak and remain invested when fundamentals are strong. For intermediate fundamentals, multiple equilibria can arise, so outcomes can take the form of either universal redemption or no redemption. Second, the share of redemptions should not be affected by default risk since investors are risk neutral and only expected returns matter. Third, with fundamental uncertainty, equilibrium behavior should take the form of threshold strategies, with investors redeeming whenever their private signal falls below a critical cutoff, in line with the theory of global games (Carlsson and Van Damme, 1993; Morris and Shin, 1998). Finally, financial fragility, which we formally define as inefficient redemptions in excess of the first-best level, should increase as the assets

[2]LLMs are often referred to as "black boxes", because their internal decision making process is not well understood. This opacity is considered an important source of risk (Cecchetti et al., 2025).

[3]The subsequent investigation by the Securities and Exchange Board of India (SEBI) that led to Jane Street being barred from Indian markets for alleged index manipulation, further underscores the scale at which these strategies now operate (Securities and Exchange Board of India, 2025).

held by funds become more illiquid.

With these results in hand, we then run experiments in which we substitute for the model's (human) investor benchmark with QL and LLM agents, testing whether they reproduce the model's equilibrium predictions. We document four main results. First, in the absence of payoff and fundamental uncertainty, both QL and LLM investors reproduce the dominance regions predicted by theory: they all redeem when fundamentals are weak, and they all stay when they are strong. However, in the intermediate range where the theory predicts multiple equilibria, AI behavior diverges. QL investors converge to a sharp threshold rule, effectively collapsing multiplicity into the risk-dominant equilibrium. In contrast, LLM investors hold heterogeneous beliefs about the actions of other investors and exhibit less coordination on equilibrium outcomes, resulting in partial redemptions. This result is inconsistent with a mere retrival of known solutions from LLM training data, suggesting that LLM investors are genuinely behaving in a strategic way and not as "stochastic parrots" (Bender et al., 2021).

Second, when default risk is introduced, LLM investors continue to align with the theoretical benchmark. By reasoning in expected-value terms, they treat the environment as equivalent to one without default risk. In contrast, QL investors still display a strong tendency to coordinate, but this coordination now leads to inefficient outcomes with a strong bias towards redeeming that amplifies fragility. We explain this result in the context of the "hot stove effect", whereby the accumulation of losses following defaults, together with limited algorithmic exploration, biases the learning dynamics and pushes QL investors toward inefficient redemptions.[4] This behavior resembles the "algorithmic stupidity" result in Dou et al. (2024), as the algorithms in our setting also adopt strategies that are inconsistent with the game-theoretic equilibrium.

Third, with fundamental uncertainty, LLM investors coordinate around the theoretical threshold. In line with theory, we show that the introduction of noisy signals induces LLM investors to hold a common belief on the actions of others, which, together with their private signals, drives their decisions to redeem or stay. QL investors, by contrast, exhibit a redemption bias when signals are noisy, as heterogeneity in private signals generates disagreement and lowers their reward from staying. However, unlike the scenario with default risk, their behavior converges to persistent partial redemptions: an equilibrium outcome consistent with theory, but inefficient relative to the first best. Finally, turning to the relationship between fragility and illiquidity, for LLM investors fragility rises monotonically with asset illiquidity, closely tracking the theoretical prediction. For QL investors, the relationship is

[4]The expression "hot stove effect" was coined by Denrell and March (2001) to refer to a situation where agents over-react to negative experiences through, for example, avoidance of similar situations.

more complex: without payoff and fundamental uncertainty, outcomes align with theory; with noisy signals or default risk, fragility increases, and its sensitivity to asset illiquidity depends on the interaction of these two sources of uncertainty.

Taken together, these results reveal a sharp contrast across AI architectures, with Q-learning generally promoting coordination, but at the expense of financial stability, whereas LLM reasoning improves stability but reduces coordination and predictability. So is AI the *deus ex machina* of modern finance? Our findings suggest the answer depends on architecture. Different designs shape economic outcomes in systematically different ways, and the stability of the financial system may critically depend on how AI agents are built.

Literature. Our paper contributes to the growing economics-of-AI literature on machine-driven strategic behavior. A central finding is that algorithms can sustain tacit collusion, i.e., earning supra-competitive profits without explicit communication. Using Q-learning in a simple oligopoly model of repeated price competition, Calvano et al. (2020) show experimentally that algorithms reliably learn to collude. Related evidence has been documented in auctions (Banchio and Skrzypacz, 2022) and asset-pricing settings (Colliard et al., 2025; Dou et al., 2025), and under other architectures, including deep reinforcement learning (Cont and Xiong, 2024) and LLMs (Fish et al., 2024).

We contribute to this literature by investigating how multiple AI investors, representing two distinct classes of AI architectures, select among equilibria in a coordination game with strategic complementarities. Hence, we place equilibrium multiplicity at the core of our analysis. Our contribution to the literature is threefold. First, we show that different AI architectures exhibit markedly different propensities to coordinate. Second, we show that the equilibria QL investors coordinate on crucially depends on the specific features of the economic environment. In particular, the payoff dominant equilibrium, akin to the tacit collusion in Calvano et al. (2020), does not always emerge. Finally, we show that default risk biases QL investors to coordinate on "all redeem" outcomes, even in ranges of economic fundamentals where this does not represent an equilibrium of the theoretical model. These results highlight that coordination is a double-edged sword; AI investors' coordination leads them to select a stable but privately inefficient equilibrium that amplifies fragility.

Our results on LLMs complement Fish et al. (2024), who show that LLM-based pricing agents learn to collude in a repeated Bertrand competition game. We show that learning is not the only path to coordination. With fundamental uncertainty, i.e., when LLM investors receive private signals, they coordinate even in a one-shot setting. This suggests that the information environment itself can serve as a coordination device for reasoning-based AI.

Our paper also contributes to the emerging academic literature on AI and financial stabil-

ity.[5] Danielsson et al. (2022) identify several channels through which AI may destabilize the financial system, including procyclicality, unknown unknowns, and optimization against the system. Danielsson and Uthemann (2025) formalize these ideas using a global games framework, showing that AI's superior information processing, reliance on common data sources, and speed advantages can make transitions between stable and crisis states more abrupt. We advance this agenda by providing an experimental, model-grounded assessment of AI behavior in a canonical coordination game. Our design allows us to isolate the key economic drivers that shape Q-learning behavior and LLM reasoning, and to map how these drivers determine equilibrium selection and fragility. Our study is also related to Yang (2024), who shows that Q-learning can trigger currency attacks in a two-agent speculative attack setting. Relative to that work, we focus on a mutual fund withdrawal game with many interacting agents, and we conduct a head-to-head comparison of AI architectures to identify how technology and environment jointly shape financial fragility.

Our paper also broadly relates to the literature on LLMs as economic agents and decision makers. Horton (2023) suggests that LLMs are implicit computational models of humans that can be given preferences, endowments, etc, and their behaviour in various economic and strategic environments can be explored by means of simulations.[6] Several papers have built on this idea to consider LLMs as investors (Fedyk et al., 2024), bank depositors (Kazinnik, 2023), and loan officers (Cook and Kazinnik, 2025). They can also serve as proxies for survey respondents (Hansen et al., 2024) and replicate human-like macroeconomic expectations (Bybee, 2023; Zarifhonarvar, 2024). Within finance, Lopez-Lira (2025) introduces an open-source framework that pits LLM trading agents against value investors, momentum traders, market makers, and contrarians, while Gao et al. (2024) show that heterogeneous LLM investors produce price dynamics that mirror observed markets. Bhagwat et al. (2025) elicit beliefs from LLM investor personas and find that disagreement about firm news both tracks abnormal trading volume and predicts a subsequent return premium.

We contribute to this line of work in two major ways. First, we consider LLM investors in a canonical mutual fund redemptions coordination game and compare their behavior directly with that of QL investors. This comparison allows us to clarify how explicit reasoning, as opposed to learning, affects equilibrium selection and financial stability. Second, we introduce a framework for analyzing the latent reasoning of LLM investors. By treating chain-of-thought

[5]The topic has also attracted much attention from policymakers (Shabsigh and Boukherouaa, 2023; Aldasoro et al., 2024; Financial Stability Board, 2024; Leitner et al., 2024; Bank of England, 2025).

[6]The approach espoused here is distinct from that of agent-based modeling (ABM), which has previously been used in economics and finance to study the emergent behavior of heterogenous, boundedly rational interacting agents that cannot be captured using a representative-agent framework (Farmer and Foley, 2009; Battiston et al., 2012). While the goal of ABMs is to predict human behavior, the aim of studying LLMs in such environment is grounded in understanding how AI agents make decisions.

reasoning text as data, we extract the beliefs that LLM investors hold about the actions of others, the justifications underlying those beliefs, and the causal chains mapping justifications to beliefs to decisions. Importantly, our approach infers beliefs from the reasoning that produced decisions, rather than eliciting them through direct questioning, which may be subject to post-hoc rationalization. This allows us to peer inside the "black box" and understand why LLM investors make decisions the way they do. Our analysis reveals that belief heterogeneity explains why LLM investors struggle to coordinate in environments with multiple equilibria: absent a shared focal point, identically prompted LLMs arrive at different but internally consistent beliefs about others' behavior. When fundamental uncertainty is introduced via noisy private signals, LLM investors converge on a common belief and their behavior aligns closely with the global games equilibrium. This suggests that LLMs can internalize sophisticated game-theoretic reasoning when the information structure provides sufficient guidance for belief formation.

Finally, we also contribute to the literature on financial fragility. Following the seminal work of Diamond and Dybvig (1983), this literature has extended to other financial institutions subject to strategic uncertainty, such as mutual funds (Chen et al., 2010), hedge funds (Liu and Mello, 2011), credit markets (Bebchuk and Goldstein, 2011), life insurers (Foley-Fisher et al., 2020) and stablecoin issuers (Gorton et al., 2025), among others. Despite rich theory, credible empirical evidence remains scarce, with Goldstein et al. (2017) and Chen et al. (2024) as notable exceptions. We open a new research avenue by experimentally testing theories of financial fragility using AI agents, allowing fine-grained control over information, payoffs, and strategic interaction, and showing clear links among fundamentals, algorithmic design, and risk.

Outline. The rest of the paper proceeds as follows. In Section II, we present a stylized theoretical framework of mutual fund redemptions and derive the predictions that we test in the experiments with Q-learning algorithms and LLMs. In Section III, we describe the experimental environment for both AI types. The results of the simulations are reported in Section IV. Section V includes the results of robustness exercises, while conclusions and policy implications are in Section VI.

II. Theoretical Framework

We build on Chen et al. (2010) and characterize the equilibria in a stylized coordination game.[7] A single-good economy extends over two dates, $t = 1$ and $t = 2$, and consists of a

[7] All proofs are relegated to Online Appendix A.

mutual fund and an integer number, $N > 2$, of risk-neutral investors. Prior to $t = 1$, each investor exchanges their unit endowments for a share in the mutual fund. Thus, the initial size of the fund is N.

At $t = 1$, the book value of the fund's investments at is $R_1\, N$, which is common knowledge, with $R_1 \geq 1$. At $t = 2$, the per-unit investment return is $R_2(\theta) = R\theta$, where $\theta \in [0, 1]$ is a uniformly distributed random variable drawn at the start of $t = 1$ and represents the strength of the economic fundamentals. Stronger fundamentals (higher θ) imply a higher investment return for the fund.[8] We also distinguish between two information environments regarding the fundamental: (i) no fundamental uncertainty, where the realized value of θ is common knowledge; and (ii) fundamental uncertainty, where each investor i receives a noisy private signal $s_i = \theta + \epsilon_i$, with ϵ_i i.i.d. across investors and uniformly distributed over the interval $[-\eta, \eta]$.

At the beginning of $t = 1$, $A < N$ of investors are active, and choose whether to redeem their shares or hold them until $t = 2$. Investors who redeem receive the current value R_1. However, in order to service redemptions, the fund must liquidate its investments, which is costly. Specifically, to raise R_1, the fund must liquidate $(1 + \lambda)R_1$ worth of assets, where $\lambda > 0$ is a measure of illiquidity: the higher is λ, the more assets must be liquidated to service redemptions. Following Chen et al. (2010), we assume that $A \leq \frac{N}{1+\lambda}$, implying that the mutual fund has enough resources to meet the redemptions of all active investors at $t = 1$.

After servicing redemptions at $t = 1$, the fund re-invests its remaining resources. Denoting the number of investors who redeem at $t = 1$ by W, an investor who decided to stay receives a pro-rata share of the fund's gross return at $t = 2$, which is given by:

$$\frac{N - W(1 + \lambda)}{N - W} R_1 R\, \theta\,. \tag{1}$$

Thus, an investor's payoff from staying is increasing in the strength of economic fundamentals, θ, and decreasing in both asset illiquidity, λ, and the number of redemptions, W.

A. Characterizing the Equilibria

We focus on characterizing equilibria in pure strategies.

[8] In Chen et al. (2010), the variable θ is a measure of the performance of the mutual fund. The two definitions are linked as stronger economic fundamentals can be associated with improved fund performance. For the purpose of our exercise, the precise definition is immaterial. Moreover, assuming uniformity is without loss of generality: All results hold for any strictly increasing mapping $R(\theta)$.

Figure 1: Equilibria without fundamental uncertainty.

No fundamental uncertainty. We first consider the case where θ is common knowledge. Following standard lines of reasoning (Morris and Shin, 1998), we can partition the support for the fundamental into three intervals, which we report in Proposition 1 and Figure 1.

Proposition 1: *In the absence of fundamental uncertainty, the equilibrium outcomes depend on the realization of θ. For sufficiently extreme values of θ, there exist a unique Nash equilibrium in pure strategies, while for intermediate values multiple equilibria arise. Specifically,*

- *if $\theta < \underline{\theta} = \frac{1}{R}$, the unique equilibrium is that all investors redeem at $t = 1$;*
- *if $\theta > \overline{\theta} = \frac{1}{R}\frac{N-(A-1)}{N-(A-1)(1+\lambda)}$, the unique equilibrium is that all investors stay until $t = 2$, while*
- *if $\theta \in [\underline{\theta}, \overline{\theta}]$, both "all redeem" and "all stay" are equilibria.*

An important prerequisite for the emergence of multiple equilibria is costly liquidation. When the fund's investments are perfectly liquid, i.e., $\lambda = 0$, liquidation does not impose additional costs on investors who stay. The expected return from staying is independent of early redemptions. This implies that redeeming at $t = 1$ is the strictly dominant action for $\theta < \underline{\theta}$, while staying is strictly dominant for $\theta > \underline{\theta}$. From a welfare perspective, this is the efficient (first-best) redemption strategy, since only negative NPV investments are liquidated.

In contrast, when $\lambda > 0$, redemptions no longer solely reflect the liquidation of unprofitable assets. In this case, the expected payoff difference between staying until $t = 2$ and redeeming at $t = 1$ decreases monotonically with the number of investors who redeem. As a result, even when fundamentals are relatively strong (i.e., above $\underline{\theta}$ and below $\overline{\theta}$), investors have an incentive to redeem if they expect others to do so. Put differently, redemption decisions are strategic complements: the incentive to redeem strengthens when others are expected to redeem. This strategic complementarity generates multiple equilibria in the intermediate region $[\underline{\theta}, \overline{\theta}]$. Throughout the paper, we refer to $\underline{\theta}$ and $\overline{\theta}$ as the "efficient redemption" and "no redemption" bounds, respectively.

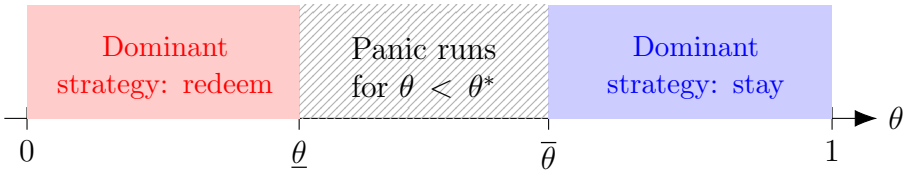


Figure 2: Equilibria with fundamental uncertainty.

Fundamental uncertainty. Next, we consider the case where investors do not observe the realized θ but instead receive noisy signals. We follow the global game literature (Carlsson and Van Damme, 1993; Goldstein and Pauzner, 2005) and study how investors make their redemption decisions based on the signals. Proposition 2 characterizes the equilibrium, which is also illustrated in Figure 2.

Proposition 2: *With fundamental uncertainty, there exists a unique symmetric Bayes-Nash equilibrium in threshold strategies that is characterized by the threshold signal:*

$$\theta^* \equiv \frac{A}{\sum_{W=0}^{A-1} \frac{N-W(1+\lambda)}{N-W} R}, \tag{2}$$

such that $\theta^* \in (\underline{\theta}, \overline{\theta})$ *and investors redeem if and only if their signals are below* θ^**. The average share of investors redeeming depends on* θ *and is given by:*

$$w^*(\theta, \theta^*) = \frac{\mathbb{E}[W]}{A} = \begin{cases} 0 & \text{if} \quad \theta > \theta^* + \eta, \\ \frac{\theta^* - \theta + \eta}{2\eta} & \text{if} \quad \theta \in [\theta^* - \eta,\, \theta^* + \eta], \\ 1 & \text{if} \quad \theta < \theta^* - \eta. \end{cases} \tag{3}$$

The introduction of private information yields a unique equilibrium characterized by a threshold signal θ^*. For values of the fundamental in the range $(\underline{\theta}, \theta^*]$, redemptions are driven by panic rather than weak fundamentals and are therefore inefficient. In this region, investors redeem because the fundamentals are sufficiently weak to induce each investor to fear that others will redeem as well.

The threshold signal θ^* depends on the underlying model parameters. In particular, θ^* increases with the liquidation cost λ. As λ increases, the fund must liquidate more assets to meet redemptions at $t = 1$, which reduces the payoff from staying. This strengthens investors' incentive to redeem early, thereby raising θ^*.

Introducing default risk. The results in Propositions 1 and 2 are derived assuming that the fund is always able to pay investors who stay at $t = 2$. In particular, the return

$R_2(\theta) = R\,\theta$ scales linearly with the fundamentals θ. An alternative, but related, specification is one where we interpret θ as the probability that the fund obtains the return R at $t = 2$ and repays investors who stay, while with probability $1 - \theta$, the mutual fund defaults; for simplicity, we assume a zero return in default. In other words, we can express the per-unit investment return at $t = 2$ with default risk as:

$$R_2(\theta) = \begin{cases} R & \text{with prob } \theta, \\ 0 & \text{with prob } 1 - \theta. \end{cases} \tag{4}$$

Default risk, thus, introduces payoff uncertainty, i.e., investors who stay may receive zero even when fundamentals are favorable. Proposition 3 establishes that introducing default risk does not affect the theoretical results.

Proposition 3: *The results in Propositions 1 and 2 remain unchanged when the fund's investment return at $t = 2$ is given by Equation (4).*

This irrelevance follows from investors' risk neutrality. Although the ex-post payoff at $t = 2$ for investors who stay differs across the two specifications, their expected payoff is the same in both cases and equals $R\theta$. This expectation is the only thing that matters for redemption decisions at $t = 1$. Hence, redemption flows are unaffected by the introduction of default risk.

B. Testable Predictions

Based on the characterization of the equilibria, we draw the following predictions. With common knowledge about θ, Proposition 1 highlights the tripartite classification of the fundamental, with multiple equilibria emerging for intermediate values of θ. Multiple equilibria emerge because investors' beliefs over others are indeterminate.

Prediction 1: *Without fundamental uncertainty, full redemptions are always observed when $\theta < \underline{\theta}$; while, no redemptions are always observed when $\theta > \overline{\theta}$. In the intermediate range, both outcomes can occur.*

Since the characterization of the equilibria only depends on the expected return at $t = 2$, the presence of default risk is irrelevant, as highlighted in Proposition 3. This result is the basis for our next prediction.

Prediction 2: *Investors' redemptions are not impacted by the introduction of default risk.*

With fundamental uncertainty, Proposition 2 shows that there is a unique signal threshold equilibrium characterized by θ^*.

Prediction 3: *With fundamental uncertainty, full redemptions are always observed when $\theta < \underline{\theta}$; while no redemptions are always observed when $\theta > \overline{\theta}$. In the range $\theta \in [\underline{\theta}, \overline{\theta}]$, the share of redemptions is weakly decreasing with the fundamental θ and signal precision η.*

Redemptions are inefficient unless the fundamental θ is so low that redeeming is a strictly dominant action. This motivates our definition of fragility as the extent of inefficient, panic-driven redemptions. Under fundamental uncertainty, fragility is measured by aggregating redemptions across states in which $\theta > \underline{\theta}$, where the average redemption share in each state is given by $w^*(\theta, \theta^*)$ as defined in Equation (3). The cutoff $\underline{\theta}$ provides a first-best benchmark: above it, efficient behavior entails no redemptions, so any that are observed reflect pure fragility. Since $\underline{\theta}$ is independent of λ, whereas θ^* is increasing in the liquidation cost λ, we obtain the following prediction.

Prediction 4: *Fragility increases monotonically with the asset illiquidity, λ.*

The theoretical framework pins down equilibrium redemption behavior as a function of fundamentals, information, and liquidation costs, and it yields testable predictions for fragility. We next describe an experimental implementation of the model in which each investor is represented by an AI agent. The experimental treatments map directly to the information structure, the presence of default risk, signal noise, and fund illiquidity, allowing us to evaluate Predictions 1–4 in a controlled setting.

III. Experimental Design

To explore the impact of AI on financial stability, we construct a simulation-based experimental setup where we replace the investors from the theoretical model with algorithmic counterparts. We consider two distinct types of AI agents: Q-learning (QL) investors, who optimize their behavior through trial-and-error learning and reward updating; and LLM investors, who reason contextually using chain-of-thought inference.

To ensure comparability, both types of investors are evaluated under the same economic environment. There are $A = 30$ active investors out of a total of $N = 50$. The parameters governing investment returns are $R_1 = 1$ and $R = 2$. For the liquidation cost, we set $\lambda = 0.25$. We consider a range of different values for the fundamental, θ, drawn from an equally spaced twentyfive-point grid spanning $\Omega = [0.4, 1]$, that covers both the efficient redemption bound $\underline{\theta} = 0.5$ and the no redemption bound $\overline{\theta} = 0.77$. Finally, to explore the relationship between fragility and illiquidity, we consider a range of λ values drawn from $[0.05, 0.35]$.[9]

[9] In the extreme case, $\lambda = 0.35$, the no redemption bound $\overline{\theta} = 0.97$ remains well defined and contained in Ω. Since the efficient redemption bound does not depend on λ, it remains unchanged.

A. *Q-learning*

Q-learning is a transparent reinforcement learning algorithm. With no prior knowledge of the payoff structures, agents learn the values of choosing different actions through trial-and-error interaction with the environment and by receiving rewards and penalties. An agent's optimal action is the one that maximizes the cumulative reward via repeated interactions.

In our set-up, the algorithm has four components:

1. **States**: The state θ_i for each QL investor i is derived from the information that they receive, i.e., either the true fundamental θ or the noisy signal s_i. As the Q-learning algorithm requires a finite state space, this continuous information is mapped to a discrete classification, which constitutes the effective state of the algorithm. The set of all states is denoted $\Theta \subset \mathbb{N}$.

2. **Actions**: The action set is binary, $\mathcal{A} = \{a_R, a_S\}$, where a_R denotes "redeem" and a_S denotes "stay".

3. **Rewards**: The reward function $\pi(\theta, a)$ assigns a payoff based on the realization of the fundamental θ and the chosen action a.[10] Crucially, QL investors do not observe the underlying mapping $(\theta, a) \mapsto \pi$, but only the realized rewards.

4. **Episodes**: Learning unfolds over T episodes. In each episode, QL investors observe their state, choose an action, and update their Q-values based on the realized payoff.

Each QL investor $i = 1, \ldots, A$ starts episode $t = 1, \ldots, T$ with a Q-matrix $Q_{i,t} \in \mathbb{R}^{|\Theta| \times 2}$, with rows for states and columns for actions. Following the realization of the state $\theta_{i,t}$, the action is determined according to an ε-greedy policy: with probability $\varepsilon_t = \beta^t$, the QL investor explores by randomizing between a_R and a_S, while with the complementary probability $1-\varepsilon_t$, the QL investor exploits by choosing the action that maximizes $Q_{i,t}(\theta_{i,t}, a)$.

If the chosen action is $a^*_{i,t} = a_R$, the reward is $\pi(\theta, a_R) = R_1$. If instead $a^*_{i,t} = a_S$, the reward depends on θ (the true fundamental) and the other QL investors' actions and is given by Equation (1). The Q-matrix is then updated as

$$Q_{i,t+1}(\theta_{i,t}, a^*_{i,t}) = (1 - \alpha) Q_{i,t}(\theta_{i,t}, a^*_{i,t}) + \alpha \pi(\theta, a^*_{i,t}), \tag{5}$$

where $\alpha \in [0, 1]$ is the learning rate: higher α places more weight on new rewards, while lower α smooths learning over past experience. Figure 3 summarizes the iterative process.

[10] Note that $\pi(\theta, a)$ implicitly depends on the actions of other investors through Equation (1), but we suppress this dependence in the notation since Q-learning treats other agents' actions as part of the environment.

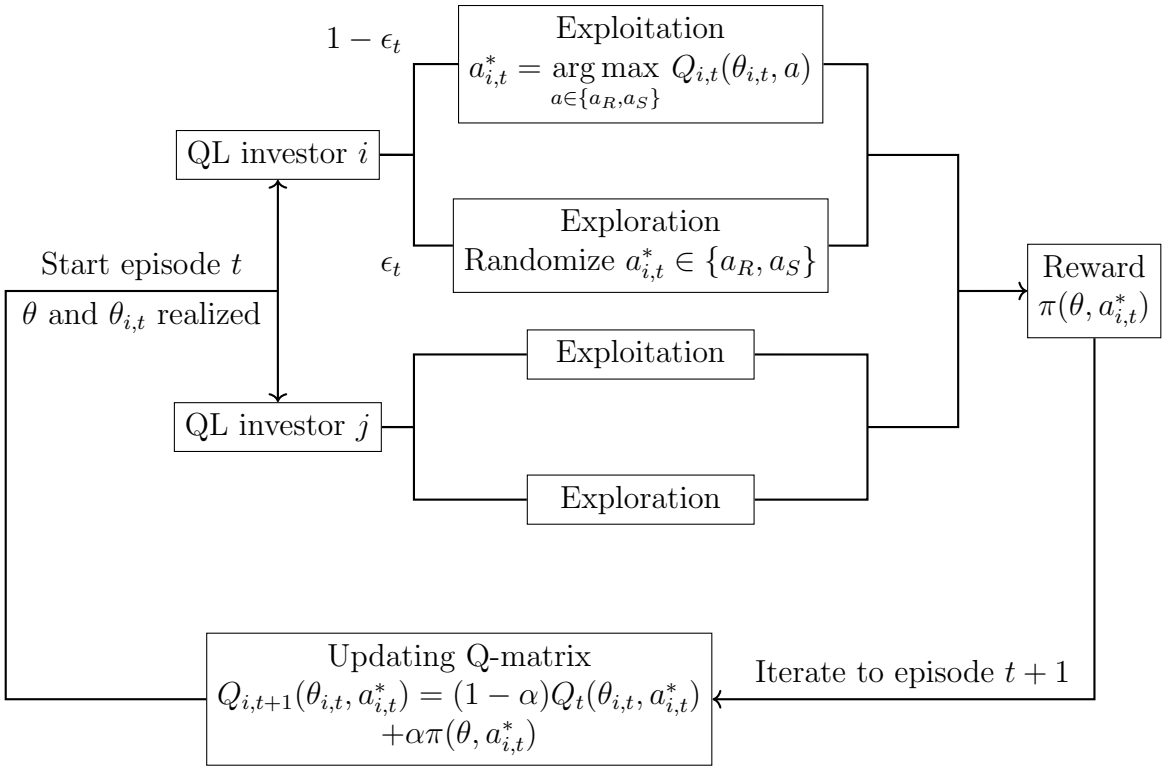


Figure 3: Iterative process of Q-learning.

In our simulations, we set $\beta = 0.99999$, $\alpha = 0.1$ and $|\Theta| = 75$. We run 25 independent rounds of training, each with $T = 500,000$ episodes.[11] For these hyperparameters, each QL investor experiments approximately 100,000 times.[12] Assuming a decaying exploration rate is standard in the literature (e.g., Colliard et al. (2025)) and reflects the idea that exploration corresponds to training, which is costly and therefore limited.

B. LLMs

Large language models (LLMs) are sophisticated next-word predictors. Drawing on the vast corpora of text used for training, these models attempt to determine the meaning and context of prompts, or input text, given to them in order to improve their predictions and responses. In addition, some models, like OpenAI's o3, Google's Gemini 2.5 Pro, and DeepSeek's R1 can "think out loud", i.e., produce a sequence of latent reasoning steps linking the prompts given to them, to assumptions, intermediate steps of calculations to the final output. As such, LLMs are capable of instantaneously adapting their behavior when the information they are

[11] The results remain qualitatively unchanged if we use a convergence criterion instead, i.e., the variance in the share of investors redeeming over a rolling window of episodes is sufficiently small.

[12] In Section V, we perform a comparative static exercise with respect to the learning rate α to discuss the role of hyperparameters for our results.

presented with changes. In our experiments, we use DeepSeek's R1-0528 to model the LLM investors. A key advantage of using this LLM is that the full chain-of-thought reasoning can be extracted and analyzed, unlike with other reasoning-based LLMs. DeepSeek's models are also open-source, which supports transparency and replicability.[13]

The experiment unfolds in two stages. First, we design the prompts. Each prompt distills the economic environment into a concise, textual description that specifies the investor's objective, the structure of payoffs, and the information available.[14] The baseline version (Prompt 1) presents the case without fundamental uncertainty or default risk. To test for the irrelevance of default risk and payoff uncertainty, we re-run the experiment by modifying the prompt as shown in Prompt 2. To introduce fundamental uncertainty and private signals, we append additional information as shown in Prompt 3.

Second, the prompt is instantiated with the relevant parameters and submitted to the LLM via an API call. Every investor is represented by a separate, independent API call. To encourage consistent behavior, we set the temperature hyperparameter to zero, reducing randomness in the outputs.[15] From each call, we record both the latent chain-of-thought reasoning and the final decision. We run three independent rounds of inference to account for any residual randomness and construct confidence intervals around the outcomes.

Finally, to explain the decision-making of LLM investors, we treat the chain-of-thought reasoning text as unstructured data, which we then analyze using an independent LLM to extract beliefs and causal reasoning. Our approach proceeds in two steps. First, we prompt the LLM to extract from each investor's reasoning text: (i) the belief they held about the behavior of others, and (ii) the justification underlying that belief. Second, we cluster these into canonical game-theoretic categories. The resulting causal graph maps justifications to beliefs to decisions, revealing the logical structure of investor reasoning. The prompts used to extract beliefs and causal reasoning are shown in Online Appendix B.

```
1 You are one of A=\%d active investors in a mutual fund out of a
     total of N=\%d investors. Each investor holds one share. Your
     goal is to maximize your return.
2 If you redeem your share, then you earn R1.
3 If you do not redeem your share, i.e., you stay, then you earn
```

[13] For a broader discussion on the use and evaluation of open-source LLMs, see Cook et al. (2023).

[14] In Section V.B, we run the experiment using a context-free prompt and show that our results remain qualitatively unchanged. We also consider a text-heavy version of the prompt that is free of equations for robustness.

[15] The temperature setting controls the degree of randomness in the model's responses: a value near zero yields deterministic and focused outputs, while higher values introduce more variation.

```
    fraction * R1 * R * \theta, where:
- fraction = (N - W * (1 + λ)) / (N - W) is the fraction of
    assets remaining in the fund after serving redemptions;
- W ≤ A-1 is the number of other active investors who redeem (the
     remaining N - A investors are passive and never redeem);
- λ (lambda) = \%f is the illiquidity parameter;
- R1 = \%f is the value of the share if you redeem;
- R = \%f is the return earned by the fund from managing its
    portfolio;
- θ (theta) = \%f is the fundamental, which measures the fund's
    performance.
Do you choose to redeem your share or stay? State your decision
    with exactly one word: ''redeem'' or ''stay'' using the XML
    tag <decision>...</decision>.
```

Prompt 1: Baseline prompt: No fundamental uncertainty and default risk.

```
3 If you do not redeem your share (i.e., you stay), then with
    probability θ you earn fraction * R1 * R, and otherwise you
    get 0, where:
```

Prompt 2: Introducing default risk: The baseline prompt is modified on line 3.

```
10 The fundamental value θ (theta) of the fund is randomly drawn
    from the interval [0,1] but you do not directly observe it.
    All values of θ are equally likely.
11 Instead you receive a private signal x_i defined as x_i = θ + ϵ_i,
     where ϵ_i is drawn uniformly from [−η, η] with η = \%f. Signals
    of different investors are drawn independently. Your private
    signal is x_i = \%f.
12 Do you choose to redeem your share or stay? State your decision
    with exactly one word: ''redeem'' or ''stay'' using the XML
    tag <decision>...</decision>.
```

Prompt 3: Introducing fundamental uncertainty: The baseline prompt is supplemented with addititional information after line 9.

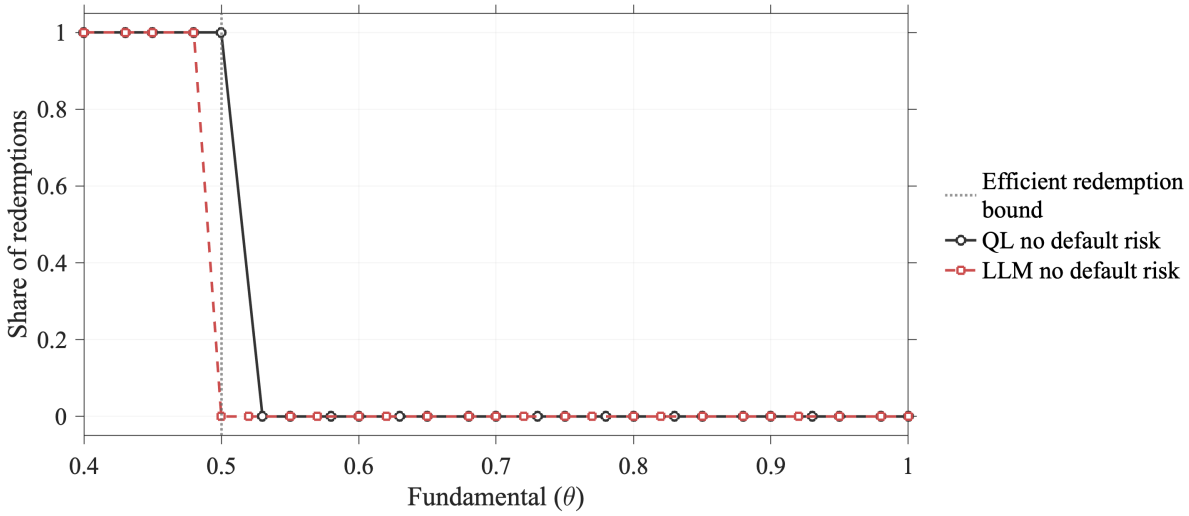


Figure 4: Decision of a single AI investor as a function of the fundamental θ without fundamental uncertainty and default risk.

IV. Experimental Results

In this section, we report the results of our experiments with QL and LLM investors to test our theoretical predictions. Before exploring the experiments in detail, it is instructive to consider the behavior of a single ($A = 1$) AI investor in the absence of fundamental uncertainty and default risk to isolate and highlight the role of strategic uncertainty.

Figure 4 plots the withdrawal decision of a single QL investor and a single LLM investor as a function of the fundamental, θ. Both types of investors behave largely in accordance with theory: in the absence of strategic uncertainty, the first-best equilibrium is obtained, with the decision to switch from redeem to stay occurring around the efficient redemption bound $\underline{\theta} = 0.5$. Changes in asset illiquidity play no role, since $\underline{\theta}$ is independent of λ.

A. Coordination and Multiple Equilibria

Next, we turn to our baseline experiments with multiple investors ($A = 30$) simultaneously choosing to redeem or stay in the absence of both fundamental uncertainty and default risk. Figure 5 illustrates the result by plotting the share of investors who redeem as a function of the fundamental.

Within the efficient redemption and no redemption bounds, investors perfectly coordinate on the Nash equilibria predicted by the theory. Thus, when the fundamental is weak, i.e., $\theta < \underline{\theta}$, they all redeem, while when fundamentals are strong, i.e., $\theta > \overline{\theta}$, they all stay. In the intermediate region, however, differences emerge. While QL investors continue to coordinate

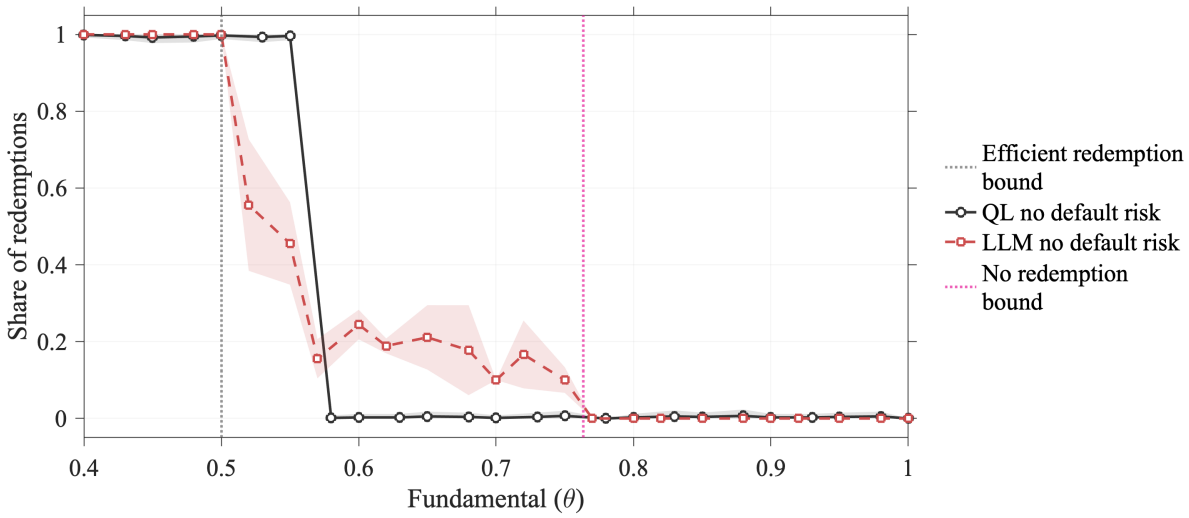


Figure 5: Share of redemptions as a function of the fundamental θ without fundamental uncertainty and default risk.

on equilibrium outcomes, LLM investors find it more difficult to coordinate, leading to intermediate redemption levels.

The aggregate outcome obtained with QL investors switches at a critical fundamental value, denoted by $\tilde{\theta}$. For $\theta < \tilde{\theta}$, all investors redeem, while for $\theta \geq \tilde{\theta}$, they all stay. This pattern is consistent across multiple runs, implying that QL investors not only converge to equilibrium outcomes but also consistently select a unique equilibrium.

Rationalizing the behavior of QL investors. The observed behavior of QL investors parallels the predictions of level-K reasoning, a standard model of bounded rationality. In this framework, a Level-0 investor chooses randomly between redeeming and staying, while a Level-1 investor best-responds to the belief that others are Level-0.

Specifically, QL investors behave as if they expect others to act randomly and optimize accordingly. This aligns with Level-1 reasoning: the critical threshold $\tilde{\theta}$ represents the fundamental value where an investor is indifferent between redeeming and staying, given a belief that other agents are equally likely to choose either action. Formally, we express this threshold as:

$$\tilde{\theta} = \frac{R_1}{\left(\frac{1}{2}\right)^{A-1} \sum_{W=0}^{A-1} \binom{A-1}{W} \left(\frac{N-W(1+\lambda)}{N-W}\right) R_1 R}. \tag{6}$$

With our parametrization, $\tilde{\theta} \approx 0.5581$, which perfectly aligns with the observed behavior of the QL investors in Figure 5. The threshold strategy defined by Level-1 reasoning, i.e., redeem whenever $\theta < \tilde{\theta}$, coincides with the criterion for the risk-dominant equilibrium.

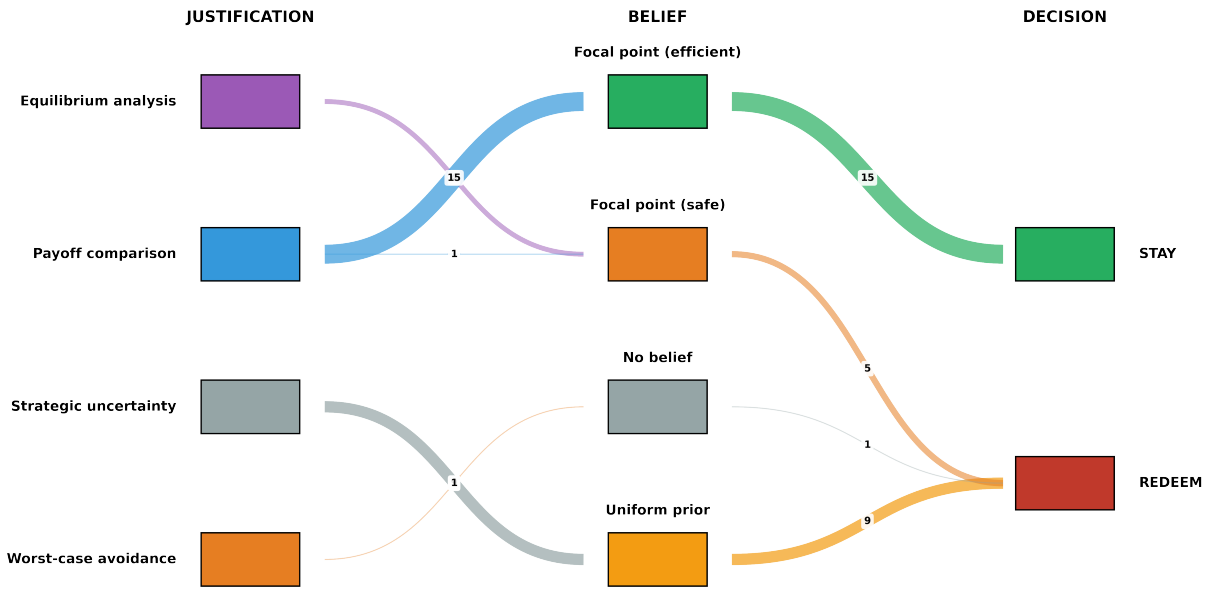


Figure 6: Causal mapping from justifications to beliefs and decisions, for LLM investors at $\theta = 0.55$ without fundamental uncertainty and default risk. Link widths and numbers indicate investor counts.

Recent literature supports this connection, noting that independent Q-learning algorithms tend to converge to risk-dominant equilibria in the presence of strategic uncertainty (Christianos et al., 2023; Albrecht et al., 2024). Thus, QL investors effectively adopt the threshold strategy that selects the unique risk-dominant outcome.

Interpreting the behavior of LLM investors. Figure 6 presents the causal graph for the LLM investors with $\theta = 0.55$, obtained by analyzing the chain-of-thought reasoning text produced by each LLM investor in one round. Investors exhibit substantial heterogeneity in both justifications and beliefs. Concerning justifications, four distinct categories were identified: (i) *payoff comparison*, where investors compare payoffs across equilibria assuming particular beliefs; (ii) *strategic uncertainty*, where investors acknowledge the presence of strategic uncertainty and adopt an uninformative prior; (iii) *worst-case avoidance*, whereby investors use maximin reasoning to guarantee the highest possible payoff in the worst possible state; (iv) *equilibrium analysis*, where investors use concepts such as deviation loss comparisons. The justification categories map into one of four distinct belief categories: (i) *focal point (efficient)*, expecting others to coordinate on the Pareto-optimal equilibrium; (ii) *focal point (safe)*,expecting others to coordinate on the safe outcome; (iii) *uniform prior*, assuming

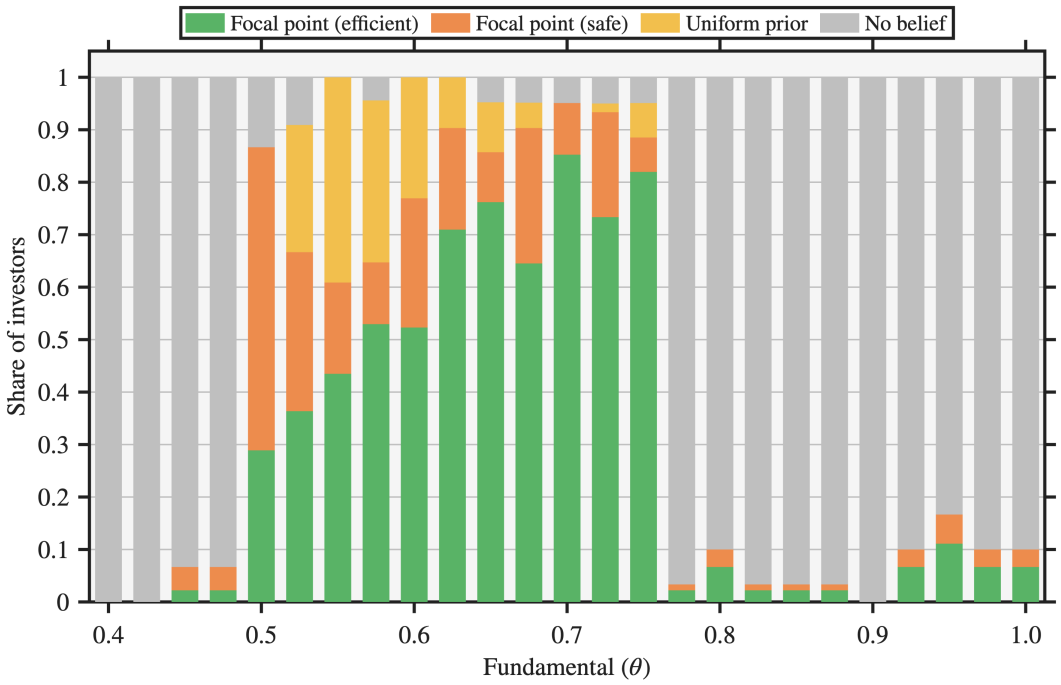


Figure 7: Evolution of LLM investors' beliefs about the behavior of other investors as a function of the fundamental θ without fundamental uncertainty and default risk, for $\lambda = 0.25$.

others redeem or stay with equal probability; (iv) *no belief*, forming no probabilistic belief about others. The graph reveals a clear mapping from beliefs to actions: LLM investors expecting coordination on the efficient equilibrium stay, while those expecting coordination on the safe equilibrium, adopting a uniform prior, or reasoning about worst cases all redeem.

Importantly, this heterogeneity arises even though all investors receive identical prompts. While we set temperature to zero to minimize randomness, separate API calls for each investor introduce minor numerical variations during execution. That such small perturbations produce divergent reasoning paths is revealing: the LLM is poised at the boundary between multiple valid approaches to the problem. Since the prompts do not specify how to form beliefs about others, and the problem admits several internally consistent frameworks, the LLM has no basis for selecting among them. Lack of coordination, thus, stems not from errors in reasoning, but from the lack of a shared focal point for belief formation. The heterogeneity in beliefs is also inconsistent with retrieval of known solutions from training data, thus suggesting that LLM investors are not behaving as stochastic parrots.

Figure 7 reveals three regimes in how LLM investors reason about coordination across

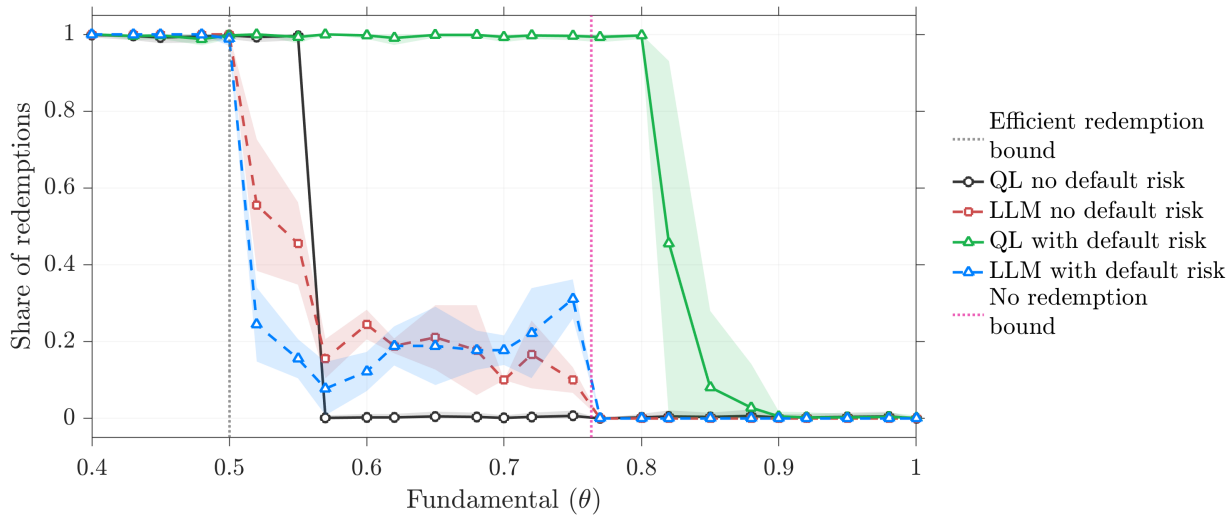


Figure 8: Share of redemptions as a function of the fundamental θ without fundamental uncertainty and default risk vs. with default risk.

different values of the fundamental. For low θ (below $\underline{\theta}$), nearly all investors report "no belief" about others' actions, consistent with redeeming being a dominant strategy that requires no strategic reasoning. For high θ (above $\overline{\theta}$), "no belief" again dominates, reflecting that staying is dominant.

In the intermediate region where multiple equilibria exist, beliefs become heterogeneous. As θ rises within this region, the share of investors expecting coordination on the Pareto-efficient equilibrium increases. This aligns with the game's structure: higher fundamentals reduce the critical mass of others required for staying to be optimal, expanding the basin of attraction of the all-stay equilibrium. The LLM reasoning texts indicate that investors correctly identify this threshold at each θ. Nevertheless, a non-trivial share maintains pessimistic beliefs or agnostic priors, contributing to the lack of coordination noted above.

We summarize the findings of the experiment in the absence of fundamental uncertainty and default risk below.

Findings 1: *In the absence of both fundamental uncertainty and default risk, both QL and LLM investors coordinate on the theoretically predicted equilibria in the dominance regions: All investors redeem for $\theta < \underline{\theta}$ and all investors stay for $\theta > \overline{\theta}$. In the intermediate region, QL investors behave as if they are Level-1 reasoners, and so they all redeem whenever $\theta < \tilde{\theta}$. Consequently, there are no multiple equilibria. The LLM investors fail to coordinate their actions in the intermediate region since different investors hold different beliefs about the behavior of others.*

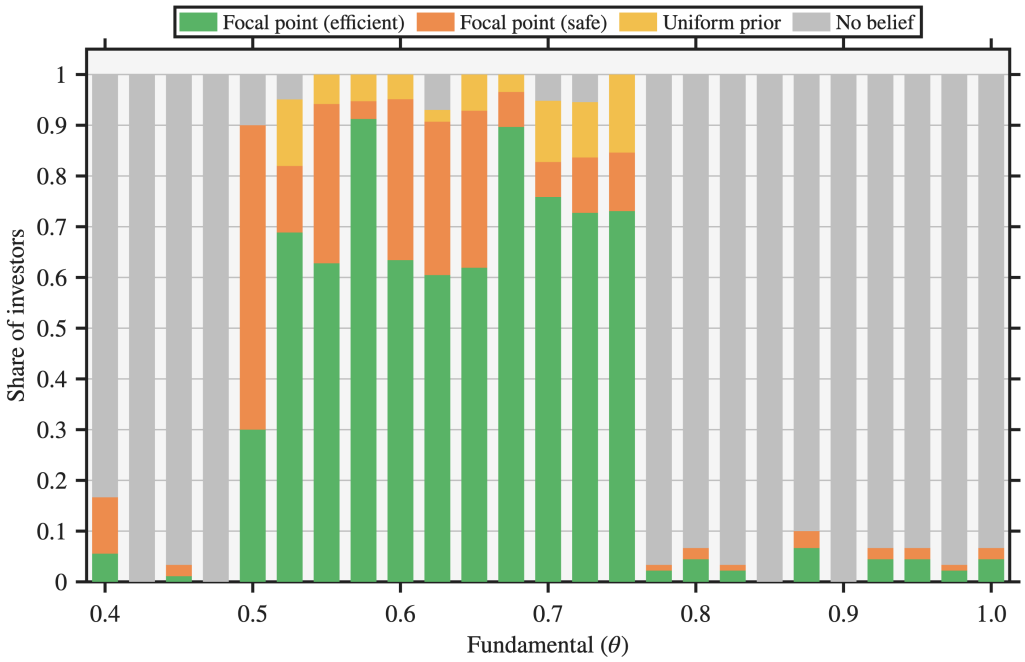


Figure 9: Evolution of LLM investors' beliefs about the behavior of other investors as a function of the fundamental θ without fundamental uncertainty and with default risk, for $\lambda = 0.25$.

B. Irrelevance of Default Risk

Next, we test Prediction 2 that default risk, and the associated uncertainty in investors' payoff, is irrelevant for the aggregate outcome. Figure 8 plots the results for the share of redemptions as a function of the fundamental for both specifications of $R_2(\theta)$ and for both QL and LLM investors. We note two key results. First, the outcome with LLM investors is barely influenced by payoff uncertainty. In fact, the redemption flows with and without default risk are, for the most part indistinguishable. Second, QL investors experience a much stronger bias towards redeeming, even for values of θ where staying is predicted to be the strictly dominant action.

Interpreting the behavior of LLM investors. The irrelevance of default risk for LLM investor outcomes can be traced to two distinct but complementary pieces of evidence. First, as Figure 9 illustrates, introducing default risk does not materially affect how LLM investors form beliefs about the actions of others. The empirical distribution of extracted belief categories—focal-efficient, focal-safe, uniform prior, and no explicit belief—is essentially unchanged relative to the baseline model without default. This implies that default risk does

Table 1: LLM investors' critical cutoff estimates with default risk.

θ	Mean($\widehat{W}$)	Var($\widehat{W}$)	Theoretical ($\widehat{W}$)
0.50	0.000	0.000	0.000
0.52	6.667	0.000	6.667
0.55	13.556	1.432	13.333
0.57	16.471	0.000	16.471
0.60	20.000	0.000	20.000
0.62	21.818	0.000	21.818
0.65	24.000	0.000	24.000
0.68	25.714	0.000	25.714
0.70	26.667	0.000	26.667
0.72	27.500	0.000	27.500
0.75	28.571	0.000	28.571
0.77	29.190	0.000	29.189

The table shows the mean and variance of the critical cutoff $\widehat{W}$ inferred from the LLM reasoning text, as well as the theoretical threshold values $\widehat{W}$ derived from the model, for values of the fundamental θ in the range between the efficient and the no redemption bounds. The theoretical cutoff solves $R_1 = R_1 R\theta \frac{N-\widehat{W}(1+\lambda)}{N-\widehat{W}}$.

not enter the coordination problem through altered expectations about aggregate redemptions or strategic behavior, ruling out belief shifts as a mechanism.

More importantly, default risk does not influence how LLM investors assess the effect of others' redemption decisions on their own payoff when choosing whether to stay or redeem. Formally, this is captured by the cutoff $\widehat{W}$, which identifies the number of other investors who must redeem for the payoff from redeeming to equal that from staying. As Table 1 highlights, $\widehat{W}$ inferred from the reasoning text coincides with the theoretical benchmark. This alignment indicates that LLM investors rely on simple expected-value comparisons rather than reflecting other possible concerns such as downside risk, higher-order uncertainty, or ambiguity.

Overall, in our environment default risk appears non-salient for LLM investors: it neither affects beliefs about others' actions nor shifts the trade-off between staying and redeeming.

Accordingly, adding default risk does not change equilibrium behavior, not because default is immaterial in principle, but because LLM investors' decision rules do not seem to incorporate concerns about risk or ambiguity.

Interpreting the behavior of QL investors. The bias of QL investors toward redeeming stems from how the underlying learning dynamics operate. Consider the case where θ is close to, but below $\overline{\theta}$. Since the fundamental is strong, the probability with which the fund's investments default is relatively low. However, each time an investor chooses to stay and the fund's investments default and deliver a zero-reward, the investor becomes pessimistic about the value of staying. Since the reward from redeeming is fixed and certain, this leads investors to favor redeeming. This, in turn, leads to a self-confirming bias whereby the investor incorrectly concludes from receiving the zero-reward that the (expected) value from staying is strictly below that of redeeming, which is never corrected.[16]

Formally, consider investor i in episode t and state $\theta_{i,t} = \theta$, that chooses to stay (action a_S). With probability $1 - \theta$, the fund's return is zero and the investor receives no reward, $\pi(\theta, a_S) = 0$. Since the Q-learning update rule is $Q_{i,t+1}(\theta_{i,t}, a) = (1-\alpha)Q_{i,t}(\theta_{i,t}, a)+\alpha\pi(\theta, a)$, a zero reward implies that:

$$Q_{i,t+1}(\theta_{i,t}, a_S) = (1 - \alpha)Q_{i,t}(\theta_{i,t}, a_S). \tag{7}$$

Repeated zero-reward episodes, therefore, push the Q-value of staying downward, keeping it below its true expected value $R\theta$. By contrast, redeeming yields a deterministic payoff R_1 and so its Q-value converges rapidly to a constant.[17] Once $Q_{i,t}(s, a_R) > Q_{i,t}(s, a_S)$, the investor chooses to redeem. The bias is further exacerbated by the decaying exploration rate, implying that investors have fewer and fewer chances to explore the benefits of staying as the episodes progress.

Our finding of a deviation from the Nash equilibrium is broadly consistent with other papers in the literature (e.g., Colliard et al., 2025; Dou et al., 2024), that show Q-learning algorithms are unable to play the Nash equilibrium as predicted by theory. There is, however, an important caveat. These papers find that Q-learning algorithms tacitly collude on outcomes that deliver higher payoffs and rewards, i.e., they choose actions that are privately beneficial. Instead, we find that QL investors choose the action that gives them the lowest (and sure) payoff, i.e., that it is privately suboptimal.

[16] The effect has been previously identified in other contexts as the "hot stove effect" by Denrell and March (2001), who named it in deference to Mark Twain, who once wrote, "*Like a cat sitting on a hot stove lid, only to get burned, and learning to never sit on a stove lid again, even a cold one.*"

[17] In Section V, we extend the model to account for the possibility that the fund becomes illiquid when many investors redeem, thus being unable to repay R_1 for sure.

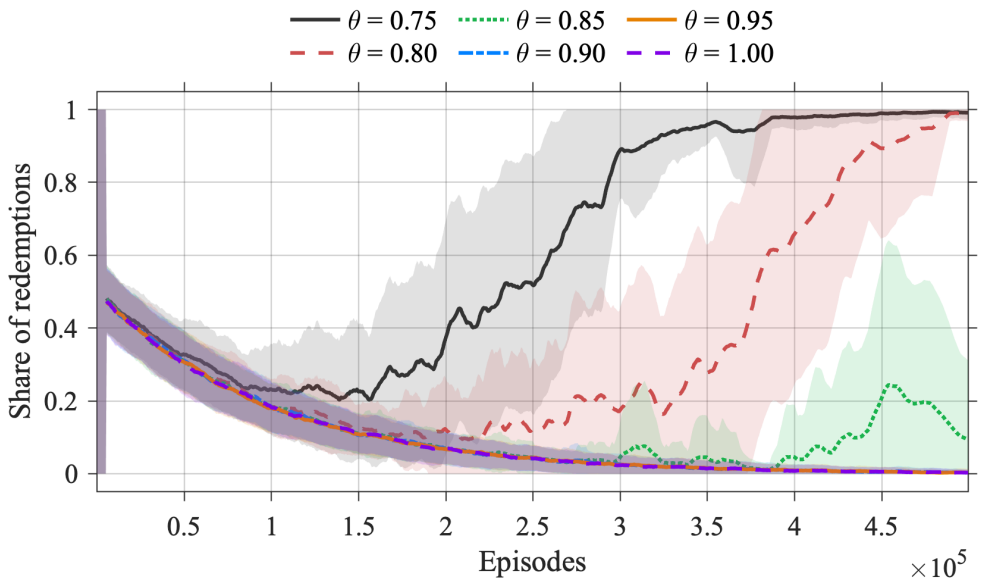


Figure 10: QL investors' evolution of redemptions across episodes without fundamental uncertainty and with default risk. The figure shows the rolling average share of redemptions over 25 training rounds together with the associated standard deviations, for different values of the fundamental θ.

Figure 10 examines this further by plotting the rolling average of redemptions across episodes for several values of θ. The shaded regions indicate the standard deviation across independent training rounds. For most realizations of θ, the share of redemptions converges to 1 and the variation between rounds decreases, consistent with successful coordination on the "all redeem" outcome. But, for some intermediate θ values, the redemption shares exhibit a persistent upward drift and substantial cross-round volatility. These patterns suggest, that over a long training horizon, the "all stay" outcome is unstable: repeated episodes in which QL investors receive zero payoff eventually push them toward redeeming. In the limit, this implies that redemption becomes dominant even when fundamentals are strong. Only in the extreme case when $\theta = 1$, so that staying yields a strictly positive payoff with certainty, "all stay" emerges as a truly stable outcome.

Findings 2: *Introducing default risk in the absence of fundamental uncertainty, has no material impact on the aggregate behavior of LLM investors. They use expected value as a solution concept to handle the payoff uncertainty, thus rendering the results indistinguishable from those without default risk. In contrast, QL investors experience a strong bias to redeeming in the presence of default risk. This bias persists beyond the no redemption bound,* $\overline{\theta}$.

C. *Fundamental Uncertainty*

With fundamental uncertainty, equilibrium behavior is characterized by the threshold θ^* and the average redemption rate $w^*(\theta, \theta^*)$ of Proposition 2. Figure 11 compares the simulated outcomes of QL and LLM investors to the theoretical benchmark across different levels of signal precision and without default risk.[18]

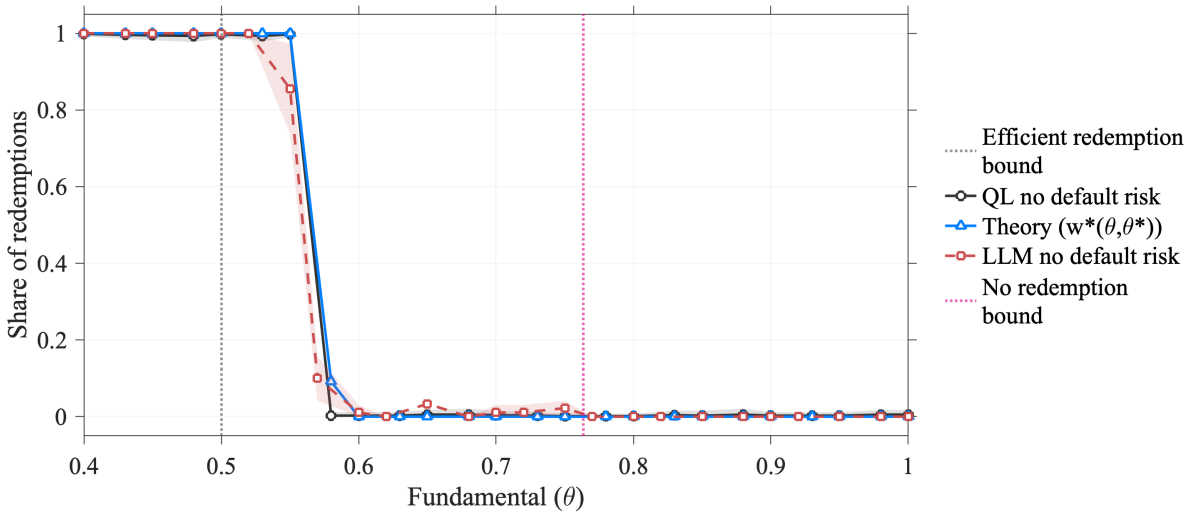
Panel (a) shows the case of highly precise signals ($\eta = 0.01$). Here, both QL investors and LLM investors closely track the theoretical prediction: redemption behavior switches sharply around the threshold θ^*, consistent with the global games equilibrium. Panel (b) considers noisier signals ($\eta = 0.05$). In this case, a divergence emerges. LLM investors continue to align with the theoretical prediction, coordinating their redemption decisions around θ^* and therefore closely matching the average redemption rate $w(\theta^*, \theta)$ predicted by the theory. By contrast, QL investors display a systematic bias toward redeeming, leading to higher average redemption rates than theory would predict.

Interpreting the behavior of LLM investors. By adding additional context to the prompt about the signal structures, LLM investors seem to recognize the structure of the underlying game and thus converge on the global games equilibrium. Hence, they achieve coordination under fundamental uncertainty by converging on the global games equilibrium. Figure 12 presents the causal graph for $\theta = 0.55$ and $\eta = 0.05$. The contrast with Figure 6 is striking: whereas investors exhibited substantial heterogeneity in both justifications and beliefs without fundamental uncertainty, the introduction of noisy signals induces convergence on a shared reasoning framework.

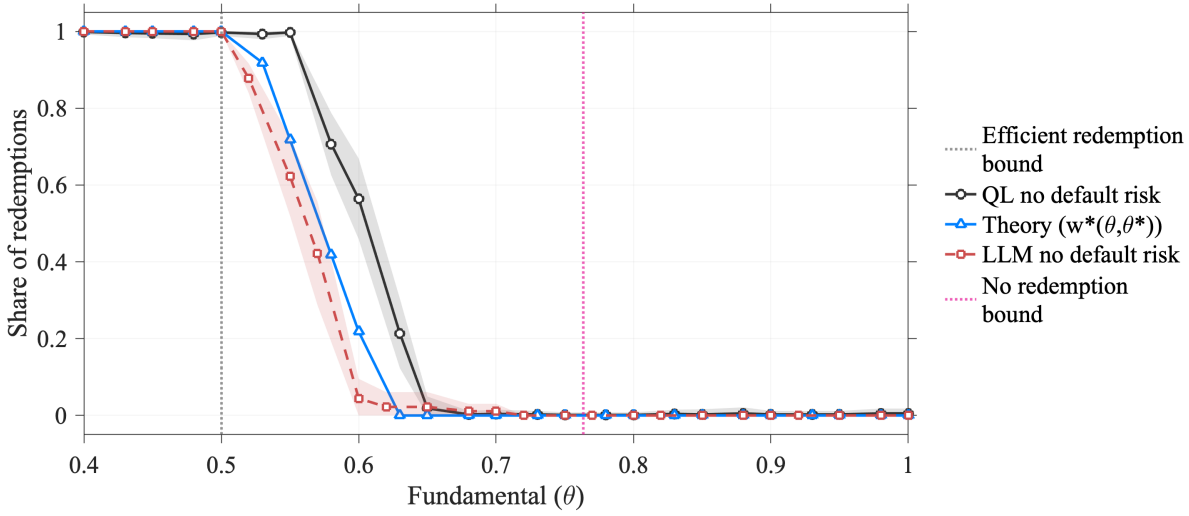
All investors arrive at a single belief category: *threshold strategy belief (conditional on signal)*. This means that each investor holds the belief that other investors follow a cutoff rule: redeeming if and only if their private signal falls below some threshold. Thus, each investor infers the distribution of others' actions from their own signal. The decision problem then reduces to computing the threshold that makes the marginal investor indifferent between redeeming and staying, precisely the logic underlying global games equilibrium selection.

The analysis clusters justifications into three closely related categories, all invoking threshold equilibrium reasoning: (i) *global game equilibrium analysis*, explicitly recognizing the canonical global games structure; (ii) *symmetric threshold equilibrium analysis*, focusing

[18] A potential discrepancy is that the theory assumes continuous fundamentals and signals, while our experiment discretizes them. Standard global-games results show that introducing noisy private signals yields a unique equilibrium under strategic complementarities, and this uniqueness is robust to the information structure (e.g., Carlsson and Van Damme, 1993; Morris and Shin, 1998; Frankel et al., 2003). Experimental implementations often discretize signals without altering the unique-threshold prediction (e.g., Heinemann et al., 2004).



(a) High precision signals $\eta = 0.01$.



(b) Low precision signals, $\eta = 0.05$.

Figure 11: Share of redemptions as a function of the fundamental θ with fundamental uncertainty and without default risk.

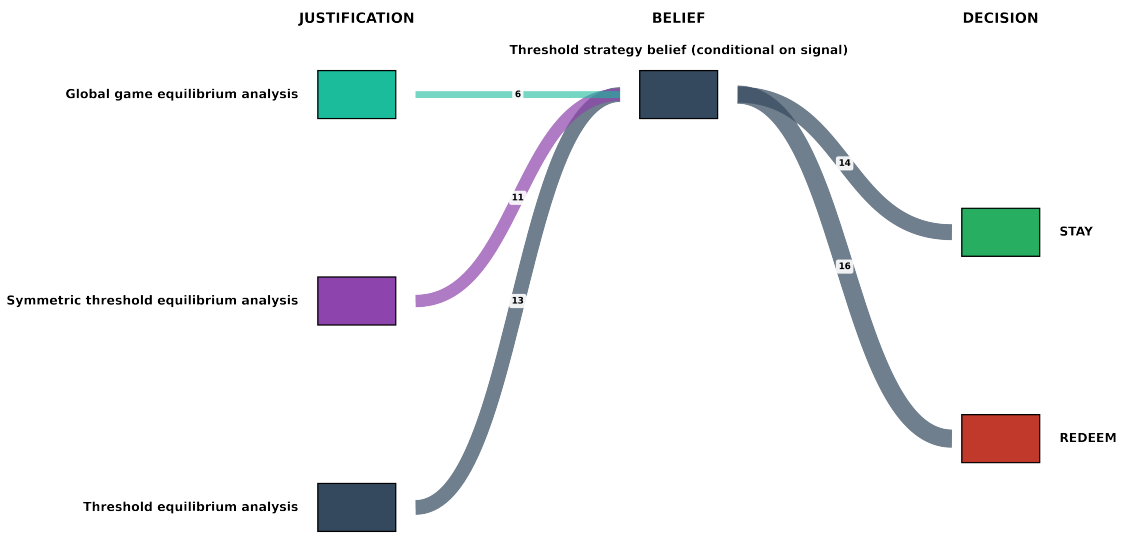


Figure 12: Causal mapping from justifications to beliefs and decisions, for LLM investors at $\theta = 0.55$ with fundamental uncertainty ($\eta = 0.05$) and without default risk. Link widths and numbers indicate investor counts.

on the symmetry of threshold strategies; (iii) *threshold equilibrium analysis*, computing the indifference condition directly. All three justification categories represent variations in terminology rather than substantive differences in reasoning. All three invoke the same core insight: threshold strategies constitute a natural focal point when signals are informative about both fundamentals and others' information. This convergence explains why coordination succeeds: the introduction of fundamental uncertainty provides a shared framework for belief formation that was absent in the complete information case. The resulting split between stay and redeem (14 versus 16 investors) reflects heterogeneity in realized signals around the threshold, not heterogeneity in reasoning.

Interpreting the behavior of QL investors. Fundamental uncertainty induces a bias toward redeeming among QL investors, though less pronounced than in the presence of default risk. The mechanism is as follows. Because signals differ across investors, they may hold conflicting beliefs about the state of the world. This disagreement, especially when signals are imprecise, translates into heterogeneous redemption decisions. Relative to the case without fundamental uncertainty, for any realization of θ, the share of investors redeeming is, on average, higher, thereby reducing the payoff from staying.

This mechanism is particularly evident when $\theta \simeq \theta^*$ and signal precision is low. Some investors receive favorable signals suggesting that the fundamental is strong, while others

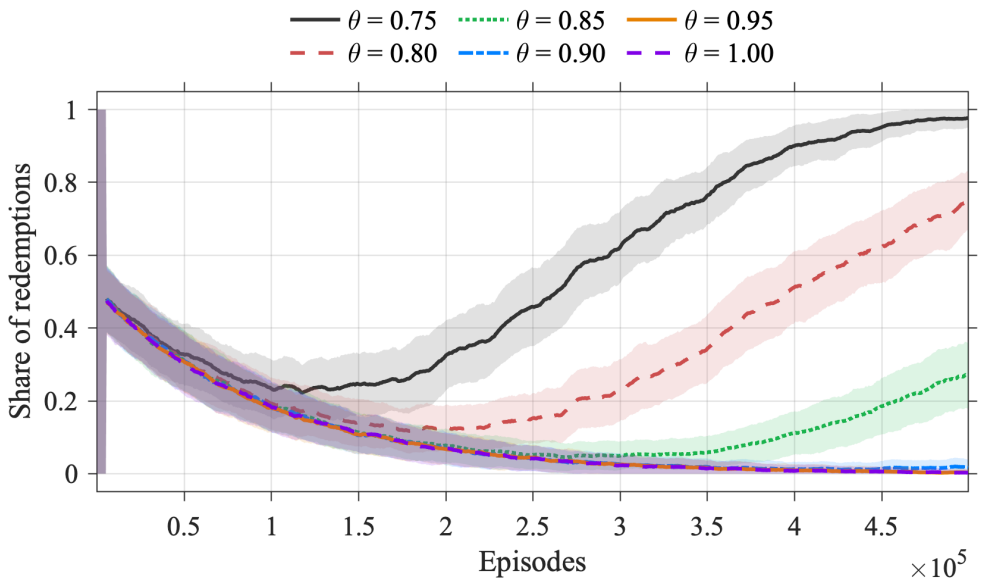


Figure 13: QL investors' evolution of redemptions across episodes with fundamental uncertainty ($\eta = 0.05$) and without default risk. The figure shows the rolling average share of redemptions over 25 training rounds together with the associated standard deviations, for different values of the fundamental θ.

observe weaker signals pointing below the threshold. Those who redeem directly reduce the payoff to those who stay, which in turn depresses the Q-values associated with staying. In this sense, fundamental uncertainty operates much like default risk in increasing the incentive to redeem early.

There is, however, a key distinction. Unlike in the experiment with default risk, in the presence of fundamental uncertainty, QL investors do not converge to redemption for almost all possible states of the economy. Instead, consistent with the theoretical prediction that $w^*(\theta, \theta^*) \in (0, 1)$, the outcome involves persistent partial redemptions: some investors redeem while others stay. Comparing Figure 13 with Figure 10, the rolling average of redemptions stabilizes at intermediate levels, with the shaded confidence bands flattening over time. This suggests that, in contrast to the experiment with default risk, the failure of QL investors to coordinate fully is itself a stable equilibrium outcome under fundamental uncertainty. This result highlights that even small differences in the information processed by QL-algorithms prevents herding, thus limiting the potential negative consequences for financial stability of large redemptions.

Findings 3: *With fundamental uncertainty, LLM investors use the global games solution concept and switch redemption behavior around the critical threshold θ^*, irrespective of the level of signal precision. QL investors, in contrast, are more sensitive to the level of signal*

precision, which induces a bias towards redeeming. Moreover, instances of partial redemptions for intermediate values of the fundamental emerge as an equilibrium outcome driven by heterogeneity of investors' signals about the fundamental.

D. Relationship between Fragility and Asset Illiquidity

We conclude this section by examining how fragility depends on asset illiquidity, captured by the parameter λ. Focusing on the case with fundamental uncertainty, our theoretical benchmark defines fragility as the average across θ of the difference between the average redemption share $w^*(\theta, \theta^*)$ and the first-best allocation, in which all investors redeem if and only if $\theta < \underline{\theta}$. Intuitively, this gap measures excess redemptions arising from strategic complementarities in investors' redemption decisions. In our simulations with QL and LLM investors, we define fragility analogously, as the deviation of the simulated redemption profile from the first-best allocation. Since the first best prescribes zero redemptions for $\theta > \underline{\theta}$, this deviation reduces to the observed redemption share in those states. Formally, for a given level of asset illiquidity λ and set of fundamental values Ω, we compute fragility as:

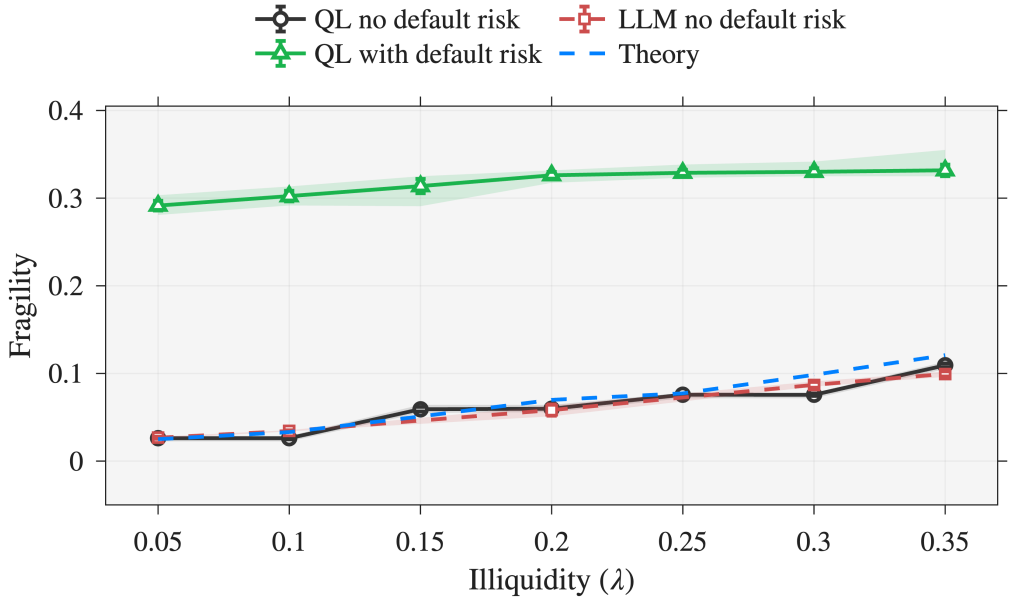
$$\text{Fragility}(\lambda) = \frac{1}{|\{\theta \in \Omega \mid \theta > \underline{\theta}\}|} \sum_{\substack{\theta \in \Omega, \\ \theta > \underline{\theta}}} \text{Observed redemption share.} \tag{8}$$

Figure 14 plots the simulation results for fragility as a function of λ. The two panels consider high-precision signals ($\eta = 0.01$) and low-precision signals ($\eta = 0.05$). Since default risk has been shown to be irrelevant for LLM investors, we restrict attention to the case without default risk. For QL investors, by contrast, we report results both with and without default risk.

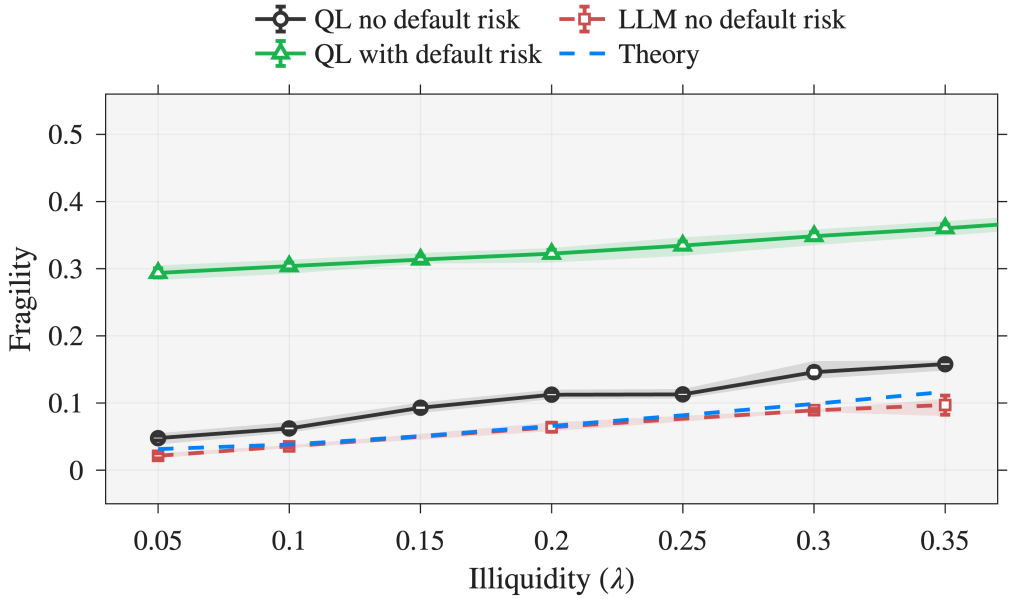
The results reveal a clear difference in behavior across the two types of investors. The behavior of LLM investors closely tracks the theoretical benchmark across both levels of signal precision, both in the overall level of fragility and in the upward-sloping relationship between fragility and asset illiquidity. This confirms the theoretical prediction that greater asset illiquidity systematically increases fragility.

For QL investors, the outcomes are more nuanced. In the absence of default risk, when signals are highly precise, the results are largely consistent with the theoretical benchmark, in line with the evidence reported in Figure 11a. With noisier signals, however, QL investors exhibit a stronger bias toward redemption, which raises the level of fragility relative to the benchmark.

The introduction of default risk amplifies this bias further, but the relationship between fragility and asset illiquidity now depends on signal precision. With high precision, disagree-



(a) High precision signals $\eta = 0.01$.



(b) Low precision signals, $\eta = 0.05$.

Figure 14: Relationship between fragility and asset illiquidity with fundamental uncertainty and without default risk vs. with default risk.

ment across investors is limited, and QL investors tend to coordinate on redeeming for almost all values of θ, leaving fragility largely insensitive to λ. When signals are less precise, by contrast, partial redemptions persist as a stable feature of the learning process (Figure 13), which in turn produces a stronger positive relationship between fragility and asset illiquidity.

Findings 4: *The relationship between fragility and asset illiquidity for LLM investors is well approximated by the theoretical benchmark, irrespective of signal precision. In contrast, for QL investors the relationship depends on both default risk and signal precision: default risk amplifies their bias toward redemption, while low signal precision strengthens the positive slope of fragility with respect to asset illiquidity.*

V. Robustness Exercises

In this section, we extend the analysis to establish robustness and disentangle the mechanisms driving our results. For QL investors, we run three additional experiments in the presence of default risk, examining how model-specific assumptions affect learning dynamics and, consequently, redemption decisions. Each experiment probes a different channel through which Q-learning's sensitivity to realized payoffs, rather than expected payoffs, generates excessive redemptions. For LLMs, we examine how alternative prompt formulations affect reasoning and redemption behavior in the baseline case without default risk.

A. QL investors

Fund illiquidity. A natural concern is that the extreme redemption behavior exhibited by QL investors in Figure 8 is driven by the asymmetry between deterministic payoffs from redeeming and stochastic payoffs from staying. To address this, we extend the model to incorporate fund illiquidity, making redemption payoffs stochastic as well.

We assume that all investors are active (i.e., $N = A$) so that if $W > \frac{N}{1+\lambda}$, the fund must liquidate its entire portfolio to meet redemptions. Investors are then served according to a sequential service constraint: only the first $\frac{N}{1+\lambda}$ investors to redeem receive R_1, while the remainder receive zero.

Figure 15 presents the results.[19] Far from attenuating excessive redemptions, introducing illiquidity risk intensifies them. The transition from full to zero redemptions shifts rightward, from approximately $\theta = 0.85$ in the baseline to $\theta = 0.95$ with illiquidity, meaning that QL investors now redeem excessively even when fundamentals are nearly certain to be sound.

[19]The attentive reader will note that Figure 15 does not include the "no redemption bound". This is because, as in Goldstein and Pauzner (2005), for $A = N$ the bound $\overline{\theta}$ is not well defined.

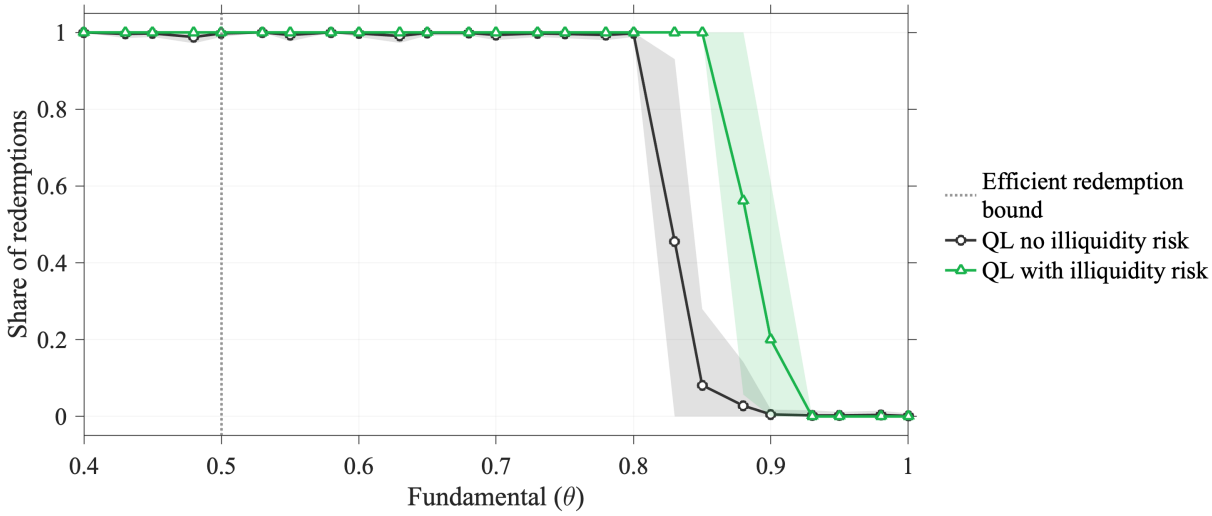


Figure 15: QL investors' share of redemptions as a function of the fundamental θ without fundamental uncertainty and with default risk, when all investors are active (i.e., $N = A$) and the fund can become illiquid.

The economic intuition is as follows. Illiquidity creates an additional layer of strategic complementarity: other investors' redemptions now harm those who wait through two channels rather than one. Mass redemptions can trigger both insolvency (fundamental failure, as in the baseline) and illiquidity (first-come-first-served rationing). While illiquidity also introduces risk for redeemers, the asymmetry in how these risks affect Q-values is decisive. When the fund becomes illiquid due to mass redemptions, investors who wait receive zero with certainty, as they are residual claimants on a depleted portfolio. Redeemers, by contrast, receive zero only probabilistically, depending on their position in the queue. Since Q-learning updates are driven by realized payoffs, the Q-value for waiting is depressed more severely: it absorbs the full impact of every illiquidity event, while the Q-value for redeeming is buffered by the lottery over queue positions. This asymmetry reinforces the drift toward the full-redemption equilibrium.

Loss given default. We next isolate the role of payoff variance in driving QL investors' behavior. In the baseline model, default results in a payoff of zero for investors who stay, implying both high variance and maximal loss given default. To disentangle these features, we modify the fund's return structure as follows:

$$R_2(\theta) = \begin{cases} R - \frac{1-\theta}{\theta} Y & \text{with probability } \theta \\ Y & \text{with probability } 1-\theta \end{cases}, \tag{9}$$

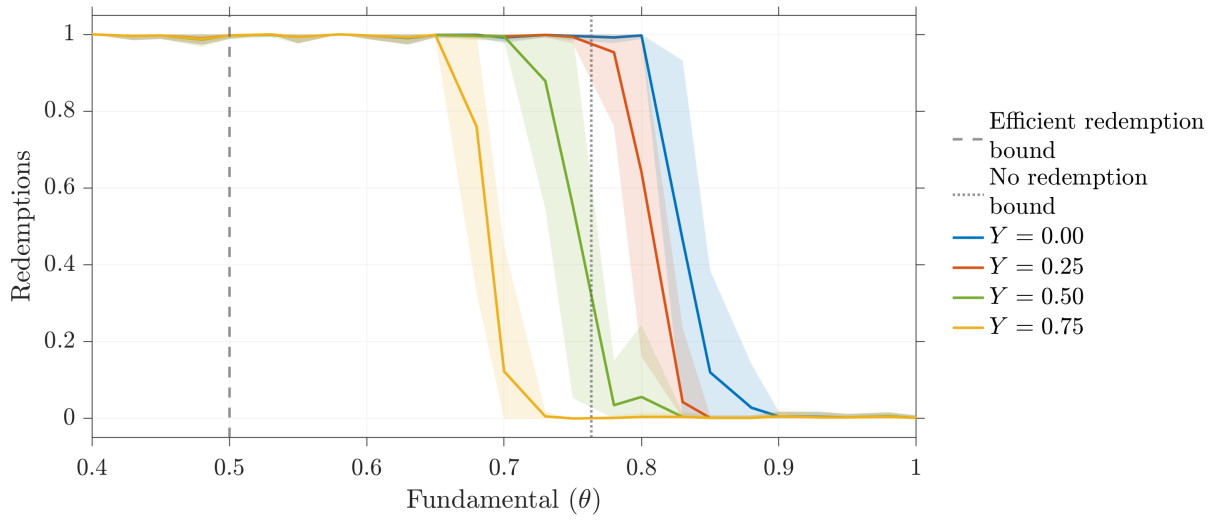


Figure 16: QL investors' share of redemptions as a function of the fundamental θ without fundamental uncertainty and with default risk, for different values of loss-given-default.

with $Y < R_1$. This specification reduces the loss-given-default from zero to Y, while preserving the expected return at $R\theta$. Varying Y, thus, isolates how return variance, holding expected returns fixed, affects redemption decisions.

Figure 16 presents the results. Lower variance (higher Y) substantially reduces QL investors' propensity to redeem. As Y increases from 0 to 0.75, the transition from full to zero redemptions shifts leftward by approximately 20 percentage points of θ. Most notably, for $Y = 0.75$, the transition occurs below the no redemption bound $\overline{\theta}$, indicated by the pink dashed line. At this point, QL investors no longer redeem when staying is the strictly dominant strategy.

The mechanism operates through Q-learning's reliance on realized rather than expected payoffs. High variance (low Y) means that occasional catastrophic realizations (e.g., defaults with little recovery) persistently depress the Q-value of waiting, even though the expected return $\mathbb{E}[R_2] = R\theta$ is unchanged. Each observed default reinforces the algorithm's aversion to waiting, regardless of how favorable the expected payoff may be. Reducing variance mitigates this effect by compressing the distribution of realized payoffs, so that bad outcomes are less catastrophic and exert less influence on learned Q-values.

Although reducing loss given default attenuates the bias toward redemption, it cannot completely eliminate it over sufficiently long horizons. As long as $Y < R_1$, default risk continues to generate occasional low realizations for those who wait, gradually eroding the Q-value of waiting relative to redeeming. Full redemption thus seems to remain the only truly stable outcome of the learning dynamics over a longer learning horizon, though convergence

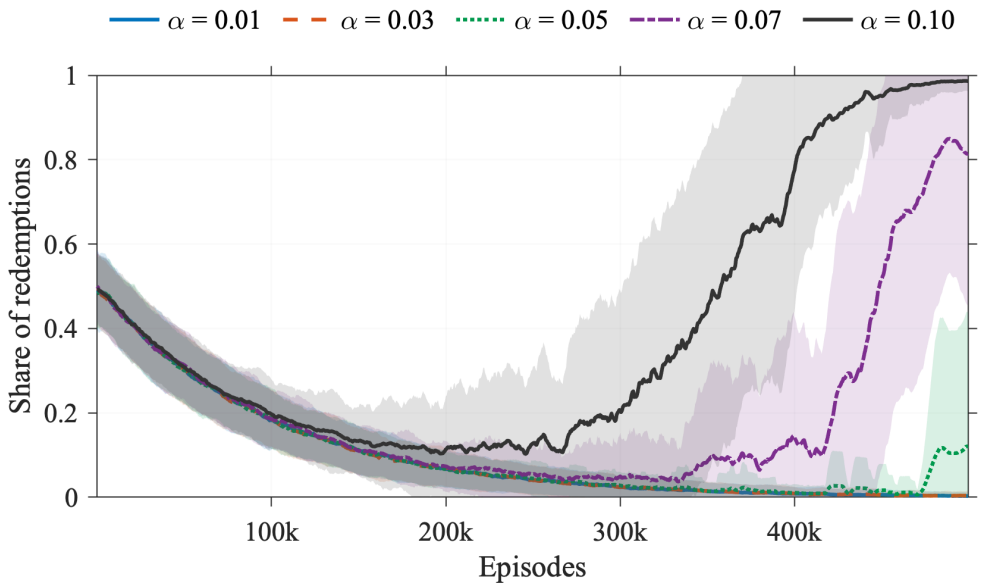


Figure 17: QL investors' evolution of redemptions across episodes for different learning rates α. The figure shows the rolling average share of redemptions over 25 training rounds together with the association standard deviations, for $\theta = 0.80$ (i.e., above the no redemption bound).

is slower when variance is lower.

Learning rate. The learning process for QL investors is governed by two key parameters: the learning rate α, which determines the weight assigned to new rewards relative to accumulated experience, and the exploration rate ϵ_t, which captures the probability that QL investors randomize between actions.

As discussed in the main text, the extreme redemption behavior exhibited by QL investors hinges on limited exploration, captured in our baseline by a decaying exploration rate $\epsilon_t = \beta^t$. Under a constant exploration rate, investors would continue sampling the "stay" action even after adverse realizations, allowing favorable outcomes to counterbalance the occasional defaults. Q-values would thus reflect expected payoffs more accurately, rather than being dominated by salient negative experiences. However, such perpetual exploration is neither economically compelling nor realistic, as it conflicts with the notion that training is costly and exploration therefore diminishes over time.

A more instructive exercise concerns comparative statics with respect to the learning rate α. A lower α corresponds to slower learning, as algorithms place greater weight on accumulated experience relative to new observations. Figure 17 illustrates the dynamics for a fundamental value of $\theta = 0.80$, which is above the no redemption bound, where staying is the theoretically optimal action. Initially, for all learning rates, investors converge toward

the efficient all-stay equilibrium as they learn that staying dominates at this fundamental. However, this efficient coordination proves unstable. At different points, depending on α, investors reverse course and drift back toward full redemptions. Higher learning rates trigger this reversal earlier and converge to full redemptions within the simulation horizon; lower learning rates delay the transition, with the slowest learners remaining near zero redemptions throughout.

The results highlight the role of learning on QL investors' tendency to avoid redemptions because of the possibility of accruing a zero payoff in the case of default. Even after investors have learned that staying is optimal, occasional defaults that yield zero payoff continue to occur with a positive probability $1 - \theta$. Each such event reduces the Q-value of waiting relative to that of redeeming. With higher α, these negative surprises are weighted more heavily, accelerating the erosion of confidence in waiting. With lower α, the same erosion occurs but more gradually, as each new observation exerts less influence on accumulated Q-values. Thus, while slower learning makes full redemptions less likely to emerge within the observed window, it does not represent a fundamental solution: it merely postpones the instability inherent in reinforcement learning.

B. LLM investors

The additional exercises conducted for the experiment with LLM investors focus on prompt design. We address two concerns. First, we want to rule out that the detailed prompt used in Section III may have conveyed information about the structure and key features of the underlying theoretical framework. If so, LLM investors' redemption decisions might reflect prior knowledge of a particular class of strategic games, rather than genuine reasoning about economic fundamentals and strategic uncertainty. Second, we want to better understand the drivers behind LLMs' reasoning and examine whether redemption decisions are sensitive to the framing of the problem.

To address these points, we conduct two additional experiments in which LLM investors are provided with alternative prompts. The first, Prompt 4, is a *neutral context prompt* that completely removes the fund redemption framing. The second, Prompt 5, is a *narrative prompt* that retains the fund redemption context but presents information in natural language, requiring investors to compute the payoffs associated with each action from the scenario description.

Figure 18 shows that LLM investors' behaviour is qualitatively similar to the baseline: they continue to struggle with coordination in the intermediate region where multiple equilibria exist. However, a notable difference emerges with the context-free prompt: redemption

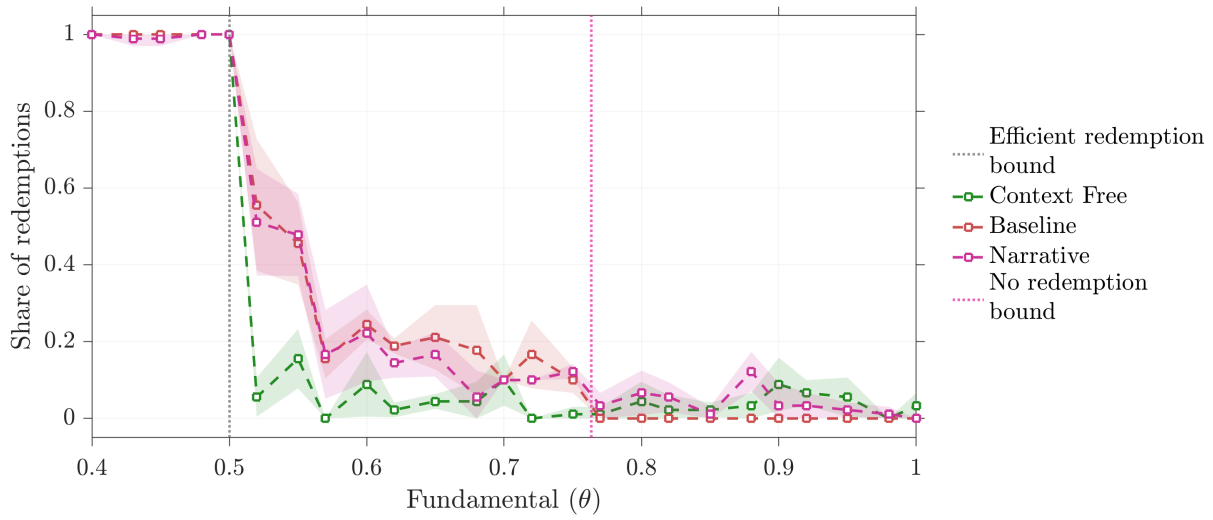


Figure 18: LLM investors' share of redemptions as a function of the fundamental θ without fundamental uncertainty and default risk, using alternative prompts.

rates are systematically lower than in the baseline, indicating better coordination on the efficient equilibrium. Figure 19 provides insight into this difference. The belief elicitation reveals that, under the context-free prompt, a larger share of investors hold optimistic beliefs in the intermediate region. This greater prevalence of "focal point (efficient)" beliefs supports the lower observed redemption rates.

We conjecture that this difference relates to the baseline prompt's explicit reference to monetary payoffs. Research has shown that reinforcement learning from human feedback (RLHF) alignment induces heightened risk aversion in LLMs (Ouyang et al., 2025). In our setting, greater risk aversion translates into a stronger incentive to redeem, since redemption offers a deterministic payoff. The financial framing in the baseline prompt may make risk-reward tradeoffs more salient, thereby activating this RLHF-induced conservatism. By contrast, the abstract framing of the context-free prompt may attenuate this effect, allowing LLMs to coordinate more readily on the efficient equilibrium.

Turning to the narrative prompt, we observe no significant differences in the redemption flow relative that observed with the baseline prompt. This suggests that LLM investors are capable of computing the payoffs associated with each action.[20]

[20] An interesting follow-up exercise would be to consider the narrative prompt with fundamental uncertainty. As the analysis in Section IV.C suggests, providing context on the signal structures can lead the LLM investors to assume that all others follow threshold strategies. In the absence of information on signal structure, we expect a higher degree of belief heterogeneity.

```
You are one of A active participants in a group of size N. Each participant chooses between two actions. Your objective is to maximize your payoff.
PAYOFF STRUCTURE:
If you choose Action A: your payoff is R1
If you choose Action B: your payoff is θR(N − W(1 + λ))/(N − W), where:
- W ≤ A is the number of other active participants who choose Action A (the remaining {N - A} participants are passive and always choose Action B)
- λ is a parameter affecting how costly it is when more participants choose Action A
- θ reflects the underlying state of the environment
- R is the baseline payoff multiplier
Which action do you choose? Answer using: <decision>A</decision> or <decision>B</decision>
```

Prompt 4: Neutral context prompt.

```
You and {N-1} other investors have each invested \textdollar100,000 in a mutual fund and have received shares. The fund invests the total amount raised based on its own investment strategy in a portfolio of assets.
The current value of the fund's assets is \textdollar{total_return_today:.2f}.
TODAY, you have to decide what to do with your shares.
If you REDEEM your shares today, you are guaranteed to receive \textdollar{individual_return_today:.2f}.
If you STAY invested, you will be repaid only after three months, and the amount you receive will depend on the value of the fund's assets at that time.
Based on current information about the fund and its strategy, if all investors remain invested, the value of the fund's assets is estimated to be \textdollar{future_return:.2f} after three months. So, each investor will receive \textdollar{future_return_per_investor:.2f}.
```

```
But, if some investors redeem today, the fund's asset value after three months will be strictly lower than \textdollar{future_return:.2f}, because fewer assets remain invested. The fund services redemptions by selling assets at a discount due to illiquid market conditions: to generate \textdollar50,000 of cash today, for example, the fund has to sell \textdollar{liquidity_cost_example:.2f} worth of assets, which eats into the fund's assets. The more investors redeem today, the larger the impact on the fund's assets for those who stay invested.
Out of a total of {N} investors, a number {A_bar}, which includes yourself, are active and in a position to decide today whether to redeem their shares or stay invested; the others will stay invested regardless.
You aim to maximize the repayment you receive based on the above information.
```

Prompt 5: Narrative prompt.

VI. Conclusions

This paper shows that AI architecture has first-order implications for financial stability. Using a canonical mutual fund redemption model, we compare two widely used approaches, Q-learning and large language models (LLMs), in environments with both economic and strategic uncertainty. Q-learning investors coordinate readily, but that coordination can be destabilizing: with default risk, learning dynamics become loss-driven and tilt toward excessive redemptions, amplifying fragility. LLM investors, by contrast, reason through the economic structure of the problem and are largely insensitive to default risk. However, absent a shared focal point for beliefs, they often fail to coordinate, making aggregate behavior less predictable. The central implication is that financial stability depends not only on fundamentals or the riskiness of strategies, but also on the AI architecture governing investor behavior.

These findings imply three broad policy messages. First, regulators need systematic data on the AI types used by market participants. Mapping the distribution of AI architectures across markets is a prerequisite for assessing financial stability risk. Disclosure requirements and supervisory reporting frameworks should be extended to cover the types of AI deployed and their training regimes. Without such information, stress tests and supervisory assess-

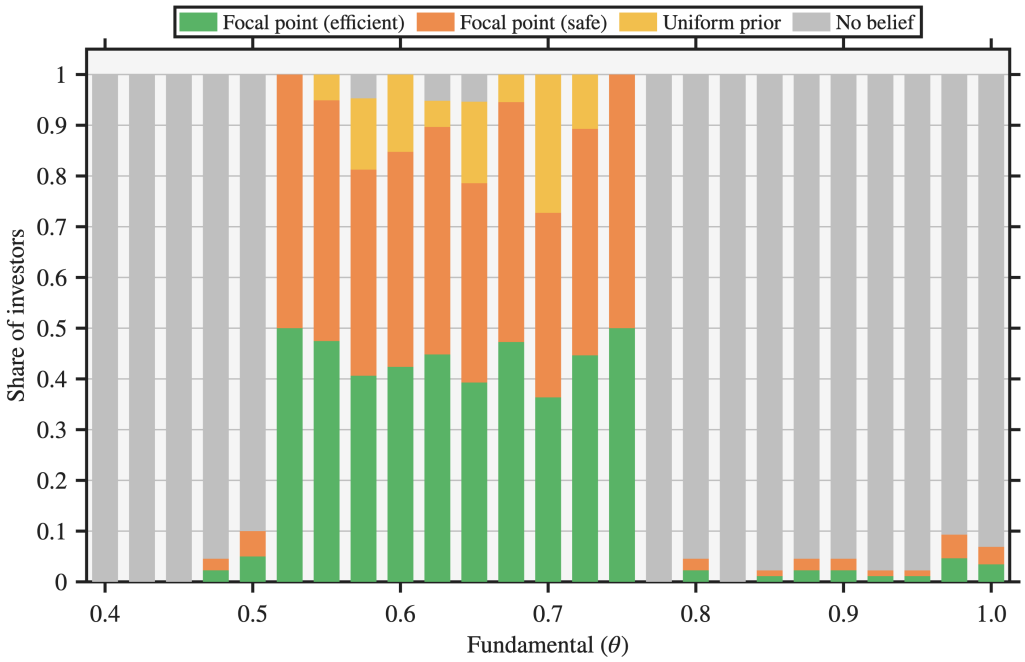


Figure 19: Evolution of LLM investors' beliefs about the behavior of other investors as a function of the fundamental θ without fundamental uncertainty and default risk, under the context-free prompt.

ments will miss the architecture-level heterogeneity highlighted by our analysis.

Second, market safeguards that are agnostic and robust to AI architecture are essential. Circuit breakers, trading halts, and other market design interventions can curb destabilizing dynamics. Our results suggest these protections are especially important in environments where AI agents can generate abrupt, correlated responses, whether through over-coordination or coordination failure. Traditional prudential tools remain valuable, but they must be complemented by mechanisms that respond to the speed and scale at which AI agents operate.

Third, financial literacy must evolve to cover the risks and benefits of delegating investment decisions to AI. As agentic systems become more accessible, retail investors will increasingly rely on algorithmic tools whose incentives, limitations, and failure modes they may not understand (what one might call "vibe investing"), where trust in the tool substitutes for understanding how it works. Addressing this requires disclosure requirements that make AI decision-making more transparent, as well as financial literacy initiatives that equip investors to evaluate the tools they use, not just the products they buy.

Beyond financial stability, our analysis speaks to broader questions at the intersection

of economics and AI. A central result is that equilibrium outcomes depend on the beliefs LLM agents hold, and on the reasoning processes through which those beliefs are formed. In our setting, LLMs arrive at decisions through chain-of-thought inference, and we show that their latent reasoning can be extracted, categorized, and linked to actions. This matters for economics because it suggests that AI agents are not simply optimizing black boxes; they are reasoners whose belief-formation processes can be studied and, potentially, shaped. For researchers studying expectation formation, coordination, and strategic interaction, LLM-based simulations offer a new laboratory in which to test theories under controlled conditions.

Our analysis also opens several avenues for future research. One natural extension is to combine learning and reasoning within a single AI architecture, for instance by allowing LLMs to update their beliefs through repeated interaction rather than relying solely on in-context reasoning. Such hybrid systems may exhibit qualitatively different coordination dynamics than pure QL or pure LLM investors. A second direction is to study mixed populations settings in which QL and LLM investors coexist, or in which different LLM architectures interact. Understanding how AI systems with fundamentally different decision-making processes coordinate (or fail to coordinate) with one another is increasingly policy-relevant as diverse AI technologies proliferate across financial markets.

For AI researchers, our results highlight the importance of multi-agent alignment. Each AI agent in our experiments is individually aligned with its own objective, yet groups of agents can collectively converge to inefficient outcomes. Current alignment research focuses largely on single-agent systems, but our findings suggest that ensuring beneficial collective behavior among interacting AI agents is equally critical. This challenge extends well beyond finance to any domain where multiple AI systems operate in shared environments, including autonomous vehicles, supply chains, energy grids, and beyond.

REFERENCES

Adrian, Tobias, Benjamin Mosk, and Jason Wu, 2025, The Future of AI in Capital Markets, CEPR Discussion Paper No. 20748.

Albrecht, Stefano V., Filippos Christianos, and Lukas Schäfer, 2024, *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches* (MIT Press, Cambridge, MA).

Aldasoro, Iñaki, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti, and Merlin Stein, 2024, Intelligent Financial System: How AI is Transforming Finance, BIS Working Paper No. 1194.

Banchio, Martino, and Andrzej Skrzypacz, 2022, Artificial Intelligence and Auction Design.

Bank of England, 2025, Financial Stability in Focus: Artificial Intelligence in the Financial System, Technical report, Bank of England.

Battiston, Stefano, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli, 2012, Debtrank: Too central to fail? financial networks, the fed and systemic risk, *Scientific reports* 2, 541.

Bebchuk, Lucian, and Itay Goldstein, 2011, Self-Fulfilling Credit Market Freezes, *Review of Financial Studies* 24, 3519–3555.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, 2021, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (Association for Computing Machinery).

Bhagwat, Vineet, J Anthony Cookson, Chukwuma Dim, and Marina Niessner, 2025, The Market's Mirror: Revealing Investor Disagreement with LLMs, FEB-RN Research Paper No. 107.

Bick, Alexander, Adam Blanding, and David J. Deming, 2024, The Rapid Adoption of Generative AI, NBER Working Paper No. 32966.

Bloomberg, 2024, Jane street, Millennium settle India options trade-secrets case, https://www.bloomberg.com/news/articles/2024-12-05/jane-street-millennium-settle-india-options-trade-secrets-case.

Bybee, J. Leland, 2023, The Ghost in the Machine: Generating Beliefs with Large Language Models, arXiv preprint arXiv:2305.02823.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello, 2020, Artificial Intelligence, Algorithmic Pricing, and Collusion, *American Economic Review* 110, 3267–3297.

Carlsson, Hans, and Eric Van Damme, 1993, Global Games and Equilibrium Selection, *Econometrica* 61, 989–1018.

Cecchetti, Stephen G, Robin L Lumsdaine, Tuomas Peltonen, and Antonio Sánchez Serrano, 2025, Artificial intelligence and systemic risk, *ESRB: Advisory Scientific Committee Reports* 16.

Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman, 2025, How people use ChatGPT, NBER Working Paper No. 34255.

Chen, Qi, Itay Goldstein, Zeqiong Huang, and Rahul Vashishtha, 2024, Liquidity Transformation and Fragility in the US Banking Sector, *Journal of Finance* 79, 3985–4036.

Chen, Qi, Itay Goldstein, and Wei Jiang, 2010, Payoff Complementarities and Financial Fragility: Evidence from Mutual Fund Outflows, *Journal of Financial Economics* 97, 239–62.

Christianos, Filippos, Georgios Papoudakis, and Stefano V. Albrecht, 2023, Pareto Actor-Critic for Equilibrium Selection in Multi-Agent Reinforcement Learning, arXiv:2209.14344.

Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo, 2025, Algorithmic Pricing and Liquidity in Securities Markets, CEPR Discussion Paper No. 17606.

Cont, Rama, and Wei Xiong, 2024, Dynamics of Market Making Algorithms in Dealer Markets: Learning and Tacit Collusion, *Mathematical Finance* 34, 467–521.

Cook, Thomas R, and Sophia Kazinnik, 2025, Social Group Bias in AI Finance, arXiv preprint arXiv:2506.17490.

Cook, Thomas R., Sophia Kazinnik, Anne Lundgaard Hansen, and Peter McAdam, 2023, Evaluating Local Language Models: An Application to Financial Earnings Calls, *Federal Reserve Bank of Kansas City Research Working Paper* no. 23-12.

Danielsson, Jon, Robert Macrae, and Andreas Uthemann, 2022, Artificial Intelligence and Systemic Risk, *Journal of Banking & Finance* 140, 106290.

Danielsson, Jon, and Andreas Uthemann, 2025, Artificial Intelligence and Financial Crises, Journal of Financial Stability, forthcoming.

Denrell, Jerker, and James G March, 2001, Adaptation as information restriction: The hot stove effect, *Organization science* 12, 523–538.

Diamond, D., and P. Dybvig, 1983, Bank Runs, Deposit Insurance and Liquidity, *Journal of Political Economy* 91, 401–19.

Dou, Winston, Jiantao Wu, and Yan Jiang, 2024, AI-Powered Trading, Algorithmic Collusion, and Price Efficiency, Jacobs Levy Equity Management Center Working Paper.

Dou, Winston Wei, Itay Goldstein, and Yan Ji, 2025, AI-powered Trading, Algorithmic Collusion, and Price Efficiency, NBER Working Paper No. 34054.

Farmer, J. Doyne, and Duncan Foley, 2009, The economy needs agent-based modelling, *Nature* 460, 685–686.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier, 2024, AI and Perception Biases in Investments: An Experimental Study, Available at SSRN 4787249.

Financial Stability Board, 2024, The Financial Stability Implications of Artificial Intelligence, Technical report, Financial Stability Board, Report to the G20.

Financial Stability Board, 2025, Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector, Technical report, Financial Stability Board.

Financial Times, 2025, The fund manager of the future might just be a machine, https://www.ft.com/content/ad12a0ec-9d6d-4ee7-83b7-643415f1a373.

Fish, Sara, Yannai A Gonczarowski, and Ran I Shorrer, 2024, Algorithmic collusion by large language models, *arXiv preprint arXiv:2404.00806* 7.

Foley-Fisher, Nathan C., Borghan Narajabad, and Stephane H. Verani, 2020, Self-fulfilling Runs: Evidence from the U.S. Life Insurance Industry, *Journal of Political Economy* 128, 3520–3678.

Foucault, Thierry, Leonardo Gambacorta, Wei Jiang, and Xavier Vives, 2025, Artificial Intelligence in Finance, Technical report, CEPR and IESE Banking Initiative.

Frankel, David, Stephen Morris, and Ady Pauzner, 2003, Equilibrium Selection in Global Games with Strategic Complementarities, *Journal of Economic Theory* 108, 1–44.

Gambacorta, Leonardo, Tullio Jappelli, and Tommaso Oliviero, 2025, Exploring household adoption and usage of generative AI: new evidence from Italy, BIS Working Paper No. 1298.

Gao, Shen, Yuntao Wen, Minghang Zhu, Jianing Wei, Yuhan Cheng, Qunzi Zhang, and Shuo Shang, 2024, Simulating Financial Market via Large Language Model Based Agents, arXiv preprint arXiv:2406.19966.

Goldstein, Itay, Hao Jiang, and David T Ng, 2017, Investor Flows and Fragility in Corporate Bond Funds, *Journal of Financial Economics* 126, 592–613.

Goldstein, Itay, and Ady Pauzner, 2005, Demand Deposit Contracts and the Probability of Bank Runs, *Journal of Finance* 60, 1293–1327.

Gorton, Gary B., Elizabeth C. Klee, Chase P. Ross, Sharon Y. Ross, and Alexandros P. Vardoulakis, 2025, Leverage and Stablecoin Pegs, *Journal of Financial and Quantitative Analysis* .

Hansen, Anne Lundgaard, John J Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar, 2024, Simulating the Survey of Professional Forecasters, SSRN Working Paper.

Heinemann, Frank, Rosemarie Nagel, and Peter Ockenfels, 2004, The Theory and Experiment of Strategic Complementarities in Games, *Econometrica* 72, 1583–1599.

Horton, John J., 2023, Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, NBER Working Paper 31282.

Kazinnik, Sophia, 2023, Bank Run, Interrupted: Modeling Deposit Withdrawals with Generative AI, SSRN Working Paper.

Leitner, Georg, Jaspal Singh, Anton van der Kraaij, and Balazs Zsamboki, 2024, The Rise of Artificial Intelligence: Benefits and Risks for Financial Stability, Technical report, European Central Bank, Financial Stability Report.

Liu, Xuewen, and Antonio S. Mello, 2011, The Fragile Capital Structure of Hedge Funds and the Limits to Arbitrage, *Journal of Financial Economics* 102, 491–506.

Lopez-Lira, Alejandro, 2025, Can Large Language Models Trade? Testing Financial Theories with LLM Agents in Market Simulations, SSRN Working Paper.

Morris, Stephen, and Hyun Song Shin, 1998, Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks, *American Economic Review* 88, 587–597.

Morris, Stephen, and Hyun Song Shin, 2002, Measuring Strategic Uncertainty, Manuscript.

Nvidia, 2025, AI On: How Financial Services Companies Use Agentic AI to Enhance Productivity, Efficiency and Security.

OECD, 2023, Generative Artificial Intelligence in Finance, OECD Artificial Intelligence Paper No. 9.

Ouyang, Shumiao, Hayong Yun, and Xingjian Zheng, 2025, Ai as decision-maker: Risk preferences of llms, *Available at SSRN 4851711* .

Reuters, 2025, After DeepSeek, Chinese fund managers beat High-Flyer's path to AI, https://www.reuters.com/technology/artificial-intelligence/after-deepseek-chinese-fund-managers-beat-high-flyers-path-ai-2025-03-14.

Securities and Exchange Board of India, 2025, Interim order in the matter of index manipulation by Jane Street group.

Shabsigh, Ghiath, and El Bachir Boukherouaa, 2023, Generative Artificial Intelligence in Finance: Risk Considerations, Fintech Note 2023/006, International Monetary Fund.

Yang, Hao, 2024, AI Coordination and Self-fulfilling Financial Crises, Working Paper.

Zarifhonarvar, Ali, 2024, Experimental Evidence on Large Language Models, SSRN Working Paper.

Online Appendix A

Proof of Proposition 1. For extreme values of the fundamental θ, investors have a dominant action. We start with the low values of θ. Irrespective of what other investors choose, redeeming at $t = 1$ is optimal for an investor when even the highest payoff they can accrue at $t = 2$, which corresponds to the payoff accrued if no other investors redeem, is smaller than R_1. Formally, this is the case when $R_1 R\theta < R_1$, that is, when $\theta < \underline{\theta} \equiv \frac{1}{R}$.

Symmetrically, staying until $t = 2$ is a dominant action when even the worst payoff that an investor expects to receive at the final date, which corresponds to the payoff accrued if all $A-1$ investors redeem, is larger than R_1. Formally, this is the case when $R_1 R\theta \frac{N-(A-1)(1+\lambda)}{N-(A-1)} > R_1$ that is when $\theta > \overline{\theta} = \frac{1}{R}\frac{N-(A-1)(1+\lambda)}{N-(A-1)}$.

When $\theta \in [\underline{\theta}, \overline{\theta}]$, both redeeming at $t = 1$ and not redeeming are equilibria. If an investor expects others to redeem, it is optimal for them to redeem as well, since staying until $t = 2$ would yield a lower payoff than R_1. Conversely, if no one else redeems at $t = 1$, it is optimal not to redeem either, as the payoff from staying, $R_1 R\theta$, exceeds R_1 for all $\theta > \underline{\theta}$. This completes the proof. □

Proof of Proposition 2. The proof adapts the standard approach in the global game literature (Goldstein and Pauzner, 2005; Chen et al., 2010) to the case with a discrete number of players. The arguments in their proof establish that there is a unique equilibrium in which investors redeem at $t = 1$ if and only if the signal they receive is below a common threshold θ^*, which is the signal at which an investor is indifferent between redeeming at $t = 1$ and $t = 2$ given what he or she believes about the signals received by other investors and, in turn, their behaviors.

We start by characterizing investors' decisions in two extreme ranges of fundamentals where investors have a dominant action. These two ranges correspond to the one characterized in Proposition 1. When $\theta < \underline{\theta}$ redeeming at $t = 1$ is a dominant action. When $\theta > \overline{\theta}$, redeeming at $t = 2$ is the dominant action.

Consider now the intermediate region where $\theta \in [\underline{\theta}, \overline{\theta}]$. Assume that investors behave according to a threshold strategy: they redeem their shares if they receive a signal below θ^* and stay until $t = 2$ otherwise.[21] Given that the signal is uniformly distributed over the interval $[-\eta, +\eta]$, the probability of receiving a signal below θ^* is $\frac{\theta^*-\theta+\eta}{2\eta}$. Building on this, we can compute the share of investors receiving the signal below the cutoff, which is given by

$$w = \sum_{i=1}^{A} \{\theta_i < \theta^*\} \sim Binomial\left(A, \frac{\theta^* - \theta + \eta}{2\eta}\right). \tag{A1}$$

It follows that the probability that W out of A investors have received a signal below θ^* is given by:

$$f(W, A) = \binom{A}{W} \left(\frac{\theta^* - \theta + \eta}{2\eta}\right)^W \left(1 - \frac{\theta^* - \theta + \eta}{2\eta}\right)^{A-W}. \tag{A2}$$

Using the above, we can compute the probability that the investor receiving the signal θ^*

[21] This comes at no loss of generality, as Goldstein and Pauzner (2005) show that in this game every equilibrium strategy is a threshold strategy.

assigns to n out of $A-1$ investors redeeming at $t=1$ as:

$$\frac{\binom{A-1}{W}}{2\eta}\int_{\theta^*-\eta}^{\theta^*+\eta}\frac{(\theta^*-\theta+\eta)^W(\theta-\theta^*+\eta)^{A-1-W}}{(2\eta)^{A-1}}d\theta \tag{A3}$$

As shown in Morris and Shin (2002), this probability is equal to $\frac{1}{1+A-1}$. Hence, the indifference condition that characterizes the run threshold θ^* reads:

$$\sum_{W=0}^{A-1}\frac{1}{A}\frac{N-W(1+\lambda)}{N-W}R_1R\theta=R_1, \tag{A4}$$

which gives the expression (2) in the proposition. Finally, the average share of investors redeeming is:

$$\mathbb{E}[W]=\sum_{w=0}^{A}w\binom{A}{w}p^w(1-p)^{A-w}, \tag{A5}$$

where $p=\frac{\theta^*-\theta+\eta}{2\eta}$. Using the identity $w\binom{A}{w}=A\binom{A-1}{w-1}$, we obtain:

$$\mathbb{E}[W]=\sum_{w=1}^{A}A\binom{A-1}{w-1}p^w(1-p)^{A-w}=Ap\sum_{w=1}^{A}\binom{A-1}{w-1}p^{w-1}(1-p)^{(A-1)-(w-1)}=Ap, \tag{A6}$$

where the second-to-last step comes from using the binomial theorem. Then, we obtain (3). □

Proof of Proposition 3. The proof is straightforward and entails replicating the analysis of the previous two propositions using $R_2(\theta)$ as specified in (4). Since for any θ, the return is simply $R\theta$ and investors are risk neutral, all thresholds are exactly as in the previous two propositions. □

Online Appendix B

In this appendix, we provide the prompts used to extract beliefs and causal reasoning from the reasoning text of the LLM investors. There are two prompts: one to extract the beliefs of the LLM investors (Prompt 6) and the second to cluster them into canonical game-theoretic categories (Prompt 7).

```
You are analyzing an investor's reasoning in a coordination game (mutual fund redemption).
CONTEXT:
The investor was given this prompt:
\{investor\_prompt\}
Their FINAL DECISION was: \{decision\}.
TASK:
The investor's reasoning trace may contain multiple considerations, calculations, and hypothetical beliefs. Your job is to identify:
    1. The DECISIVE BELIEF -- the specific belief about other investors' behavior that ultimately determined their final decision
    2. The JUSTIFICATION -- the reasoning that led them to form this belief
    3. The CAUSAL CHAIN from justification -> belief -> decision
Note: The investor might explore multiple scenarios (e.g., "if others redeem with p=0.5..." or "in the all-stay equilibrium ...") but ultimately commits to ONE belief that drives their final action. Find that committed belief.
REASONING TRACE:
\{explanation\}.
INSTRUCTIONS:
    1. Read the ENTIRE reasoning trace
    2. Identify where the investor COMMITS to a specific belief about others
    3. Identify what JUSTIFICATION led them to form this belief
    4. Extract the exact text passage that reveals the belief commitment
    5. Construct the causal DAG
```

Prompt 6: Prompt to extract the beliefs and causal reasoning of the LLM investors.

```
You are analyzing causal reasoning patterns extracted from 30 investors.
Below are the causal DAGs (justification -> belief -> decision) for each investor, described in free-form text. Each entry
```

```
    also includes:
    - evidence\_quote: The exact text from the reasoning trace
      supporting the extraction
    - confidence: How clearly the causal chain was stated (HIGH/
      MEDIUM/LOW)
Your task is to:
    1. Identify the distinct categories of JUSTIFICATIONS present
       in the data (use EXACTLY 3-4 categories)
    2. Identify the distinct categories of BELIEFS present in the
       data
    3. Assign each investor to one justification category and one
       belief category
IMPORTANT: Be PARSIMONIOUS with justification categories. Use
    only 3-4 high-level categories. Merge similar reasoning
    patterns.
Reference list of standard JUSTIFICATION categories (consolidate
    into these):
    - Payoff comparison: Reasoning based on comparing payoffs
      across equilibria (includes Pareto efficiency, strong
      fundamentals, threshold calculations)
    - Strategic uncertainty: Reasoning based on uncertainty about
       others' behavior (includes uninformative priors, no
      information)
    - Worst-case avoidance: Focused on guaranteed minimum payoff,
       no belief about others (maximin reasoning)
    - Equilibrium analysis: Formal equilibrium concepts (includes
       mixed equilibrium analysis, deviation loss comparison)
Reference list of standard BELIEF categories (use these if they
    fit):
    - Focal point (efficient): Believe others coordinate on
      Pareto-efficient equilibrium (all-stay)
    - Focal point (safe): Believe others coordinate on risk-
      dominant equilibrium (all-redeem)
    - Uniform prior: Assume equal probability of others' actions,
       no specific expectation
    - Mixed-strategy equilibrium: Believe others play the
      symmetric mixed equilibrium probability
    - No belief: No probabilistic belief formed (used with worst-
      case/maximin reasoning)
INVESTOR DATA:
\{investor\_data\}
```

Prompt 7: Prompt to cluster the extracted justifications and beliefs of the LLM investors into canonical categories.