Artificial Expectations: AI Agents and Market Stability*

Matteo Muntoni †

December 1, 2025

Abstract

I study the impact of adopting Generative Artificial Intelligence (GenAI) as an expectation-formation technology on financial market stability. I extend a canonical experimental setup into a framework that: i) assesses whether AI Agents behave as rational agents, human-like learners, or adaptive stabilizers; ii) causally estimates how different AI models, or different ways of using the same model, influence price stability relative to fundamentals; and iii) structurally estimates a "hype susceptibility" parameter capturing a model's propensity to deviate from fundamental valuation. Applying the framework to a set of multimodal large language models (LLMs) from multiple providers, I find that GenAI Agents can act as market stabilizers, though not as fully rational agents, with substantial heterogeneity across architectures and use configurations. The framework is readily extendable to markets with heterogeneous AI participants, endogenous adoption of AI expectation technologies, and multimodal information shocks.

Keywords: Artificial Intelligence, Large Language Models, Expectation Formation, Experimental Asset Markets, Adaptive Learning, Market Stability
JEL codes: C91; D84; G12; G14; O33

*__Preliminary draft. Please do not circulate without permission.__

I am grateful to Tommaso Nannicini, Ramon Marimon, Leonardo Melosi, and Giacomo Calzolari for their helpful comments and suggestions, and to all participants in the EUI internal seminars. All errors are my own.

†Department of Economics, European University Institute. Email: matteo.muntoni@eui.eu

1 Introduction

> Algorithms are qualitatively different from other technological innovations. They represent a change to the fundamental building block of all of economics: how decisions are made. Nearly every economic model is ultimately a model of decisions and their consequences. A radical change to how decisions are made necessitates a radical change in how economists do economics.

The above passage, from the 2025 AEA Distinguished Lecture by Mullainathan (2025), calls for a paradigm shift in economic research. As algorithms increasingly mediate, or even replace, human decision-making, traditional assumptions about how agents process information, and take decisions may no longer hold. As Mullainathan (2025) argues, understanding how these changes alter economic behavior requires revisiting the foundations of Economics.

Artificial Intelligence (AI), and algorithms more broadly, have rapidly become a central focus of economic research, both for their transformative potential in shaping the economy and for their profound implications for how economic inquiry itself is conducted (Brynjolfsson et al., 2025).[1] In recent years, an explosion of interest has centered on Generative AI (GenAI), spurred by the public release of models such as OpenAI's GPT. GenAI is an umbrella term encompassing recent Machine Learning models trained on massive, unstructured datasets to learn complex patterns. Their principal function is the creation of new content (text, images, code, etc.), enabling them to address complex, non-routine creative and cognitive tasks. These large-scale models are often referred to as Foundation Models. Large Language Models (LLMs) constitute a specific class within this family, originally designed for text generation but now increasingly capable of processing and generating multimodal inputs (text, images, audio).[2]

Financial markets stand at the forefront of this technological transformation. AI systems, particularly those powered by GenAI, are rapidly being embedded in the core forecasting, trading, and risk management functions of financial institutions. Already in 2023, Bloomberg released a proprietary LLM trained on extensive financial data, called Bloomberg GPT (Wu, Irsoy, et al., 2023), which significantly outperformed traditional models on most Natural Language Processing (NLP) financial tasks at the time. GenAI tools have since expanded their operational reach to a variety of financial applications beyond simple NLP tasks.[3]

[1] Sargent (2025) traces the early roots of modern Machine Learning and AI to ideas emerging from Physics, Biology, Statistics, and importantly, Economics.

[2] The terms Foundation Model, LLM, and GenAI are frequently used interchangeably in policy and public discourse. I follow this convention and generally use "GenAI" to refer to recent multimodal LLMs, unless explicitly stated otherwise.

[3] Recent reports document the rapid integration of AI into financial activities. The International Organization of Securities Commissions (IOSCO) highlights the expanding use of AI in algorithmic trading, robo-advising, investment research, and market surveillance, alongside the rise of increasingly autonomous ("agentic") systems. McKinsey's Global Banking Annual Review 2025 similarly envisions financial systems where AI agents play a growing role across

These developments raise the fundamental question of the implications of the widespread adoption of GenAI for market functioning and stability. Major policy institutions, including the IMF, ECB, and FSB, have recently emphasized both the opportunities and the risks associated with the growing integration of GenAI into financial systems.[4] By efficiently processing large and heterogeneous datasets in real time, GenAI can reduce informational frictions and improve the accuracy of price signals. Yet the opacity of model architectures and uses, the concentration of AI provision among a few entities, and the associated risk of correlated forecasting behavior may introduce new sources of systemic instability.

From an economic perspective, financial markets that efficiently process information and price assets close to their fundamentals are crucial for effective capital allocation, one of the core functions of financial systems (Fama, 1970). Excess volatility may distort investment incentives and induce capital misallocation (Lorenzoni, 2008), while bubble-burst episodes may generate temporary economic booms followed by deep and persistent recessions, ultimately leading to substantial welfare losses (Jordà et al., 2015; Biswas et al., 2020; Guerron-Quintana et al., 2023). Real-world markets are well documented to exhibit excess volatility (Shiller, 2015), bubble-and-bust episodes (Jordà et al., 2015), and experimental evidence confirms these tendencies among humans in controlled settings (Smith et al., 1988; Porter and Smith, 2003; Hommes et al., 2008; Hommes, 2017). Whether the net effect of GenAI adoption is beneficial for market stability or, on the contrary, exacerbates existing undesirable dynamics remains an open question of first-order policy relevance, a motivation also shared by Hansen and Lee (2025). This broad question can be decomposed into three interdependent components: (i) the degree of agency given to GenAI systems, that is, to what extent GenAI acts as an advisor rather than a fully autonomous decision-maker; (ii) the nature of that agency, that is, what kinds of decisions are delegated; and (iii) conditional on a given form of agency, how AI Agents' decisions affect market stability. In this paper, I focus on economic agents that grant full agency to GenAI systems in forming market expectations,[5] and study how such AI Agents' expectations affect price dynamics.

This focus is not arbitrary. First, AI Agents are already being marketed and deployed in real-world settings (Google Cloud, 2025), giving rise to what Tomasev et al. (2025) term a Virtual Agent Economy. Economies populated by artificially intelligent agents were studied well before the recent advances in AI; for instance, Marimon, McGrattan, et al. (1990) analyzed an exchange economy

functions traditionally managed by human staff. See Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges (IOSCO, 2025), available at this link; and Global Banking Annual Review 2025, available at this link (accessed December 1, 2025).

[4]See Artificial Intelligence and its Impact on Financial Markets and Financial Stability (IMF, 2024), available at this link; The Rise of Artificial Intelligence: Benefits and Risks for Financial Stability (ECB, 2024), available at this link; and The Financial Stability Implications of Artificial Intelligence (FSB, 2024), available at this link (accessed December 1, 2025).

[5]The framework can be extended to relax this assumption and analyze endogenous adoption decisions.

in which agents' behavior was governed by Holland (1975) classifier systems. Modern AI Agents, however, are largely powered by GenAI, and in particular Large Language Models (LLMs), neural architectures with billions of parameters that achieve remarkable performance across a wide range of tasks, yet whose internal representations remain opaque and analytically intractable due to the complexity of their design. The advent of these AI Agents in the economy therefore presents new, and in many respects unprecedented, challenges for economists (Hadfield and Koh, 2025).

Second, Computer Science research is rapidly expanding the use of foundation models for time-series forecasting tasks (Miller et al., 2024; Ye et al., 2025). Recent advances in this line of research combine multiple architectures or modify existing ones to improve time-series prediction. By leveraging multimodal inputs that integrate textual, numerical, and visual data, these approaches have achieved remarkable gains in predictive accuracy. Among them, LLMs are now being used even as one-shot forecasters, that is, generating a forecast from a single prompt, with promising results (Gruver et al., 2023; Miller et al., 2024; Ye et al., 2025). These developments highlight the growing importance of AI-driven expectations in the economy. On one hand, AI Agents increasingly operate in environments where forming expectations is essential for autonomous decision-making. On the other, GenAI is emerging as a novel forecasting technology.

Third, since the main economic outcome of interest in this paper is market stability, isolating the role of expectations offers a natural starting point. Expectations are central to the stability of economic systems, a channel particularly studied in the macroeconomic learning literature[6] (Grandmont, 1998). The main formal notion of stability in this literature is *expectational stability*, or *e-stability* (Evans and Honkapohja, 2001). It characterizes the conditions under which a Rational Expectations Equilibrium (REE) is asymptotically stable when agents form expectations through adaptive learning (Evans and Honkapohja, 2009). In essence, it examines whether an economy populated by boundedly rational agents, who form expectations using a given rule, such as a misspecified econometric model, a heuristic, or a more sophisticated adaptive process, converges over time to an REE. The concept of e-stability thus provides a natural criterion for equilibrium selection in models with multiple REE (Marimon, 1996).

Beyond providing a justification for an REE, however, the learning approach offers a framework to analyze economic dynamics without imposing the Rational Expectations (RE) condition. Such economies may display substantial deviations from equilibrium behavior during the learning process and, in some cases, may converge to alternative equilibria rather than the REE (Marimon, 1996;

[6]The term *market stability* is used in various contexts in economics. It may refer to fluctuations in output, consumption, and investment in the business-cycle literature (Romer, 1986); to deviations of asset prices from fundamentals in macro-finance (Shiller, 1981); to the mathematical local or global stability of steady states in dynamic models; or to the multiplicity and indeterminacy of equilibria in Rational Expectations frameworks, including rational bubbles and sunspots (Blanchard, 1979; Benhabib and Farmer, 1999; Martin and Ventura, 2018). In this paper, I study market stability and the influence of AI-driven expectations on it, drawing on insights from the learning literature in macroeconomics (Marimon, 1996; Hommes, 2021).

Evans and Honkapohja, 2009). With some abuse of terminology, I refer to these phenomena as "instabilities," meaning deviations from equilibrium dynamics also reflected in persistent departures from fundamentals. The evolution of learning economies, and the extent to which they exhibit these instabilities, depends critically on both the underlying economic structure and the learning rule adopted by agents (Hommes, 2021).

In a series of influential papers, Marimon and Sunder (1993), Marimon, Spear, et al. (1993), and Marimon and Sunder (1994) introduced an experimental framework in which the structural features of the economy are set by the experimenter, whereas expectations are formed by human participants. This framework, now known as the *Learning-to-Forecast Experiment* (LtFE), has been widely used to examine how human expectations affect market stability across different economic environments (Hommes, 2011; Hommes, 2021). A key insight from this literature is that human learners tend to generate instabilities in environments characterized by positive expectational feedback, that is, when higher expectations for an endogenous variable are reinforced by the actual realization of higher values after markets clear (Hommes, 2021).

In light of this discussion, the core issue of AI expectations and market stability can be seen as involving two extreme scenarios. On the one hand, AI learners might generate the same kinds of market instabilities that human learners do, or even amplify them, as suggested by early evidence showing limited forms of rationality among AI traders, with market dynamics that mirror real-world patterns or those observed among human participants in laboratory settings (Lopez-Lira, 2025; Rio-Chanona et al., 2025). In this case, their best contribution would be a neutral effect on stability. On the other hand, AI learners might behave as fully rational agents, immediately converging to a REE and, ideally, to a "desirable" one. Some recent evidence points in this direction, showing that newer-generation models tend to behave more rationally than humans in simulated trading environments (Henning et al., 2025; Hansen and Lee, 2025).

A third possibility, explored in this paper, is that AI Agents are boundedly-rational learners who may help stabilize markets in environments where human learners tend to generate unstable dynamics. In this case, AI Agents would not need to be "textbook-rational", as suggested by Henning et al. (2025), and they may even fail to converge to any REE. Yet they could still display stabilizing properties by keeping prices closer to fundamentals, limiting excess volatility, and preventing the emergence of bubble episodes.

The asset-pricing model is a natural starting point to study market stability, given its positive feedback structure and relative simplicity. To have a human benchmark, I consider the asset-pricing LtFE of Hommes et al. (2008) and extend it in two directions. First, I replace human forecasters with GenAI models acting as predictive Agents, similarly to Rio-Chanona et al. (2025). Second, I introduce an exogenous shock, referred to as a "hypers shock," which is structurally designed to inflate market prices and potentially lead to an explosive dynamic. This setup allows me to compare

market dynamics generated by AI expectations with those produced by human learners and with the theoretical REE benchmark, and to assess whether AI Agents can prevent the emergence of an exogenously triggered bubble episode. The framework also enables two complementary analyses. First, it allows the causal estimation of how changes in model use or architecture affect market stability. This analysis can inform how the design and deployment of GenAI models influence market stability, and it can also guide policies on the adoption of AI technologies in real financial markets. Second, the framework supports the structural estimation of a "hype susceptibility" parameter, which captures how prone a model is to be destabilized by exogenous bullish forces versus remaining close to fundamentals. This metric provides a simple way to compare the stabilizing potential of different GenAI models, forecasting methods, or even human forecasters, when data permit.

I apply the framework to a set of recent multimodal LLMs from different providers, including GPT-4o, Gemini 2.0 Flash, Haiku 3.5, LLaMA 3.3 70B Turbo, Grok 2, and DeepSeek Chat (V3). The results show that GenAI Agents can act as market stabilizers, with notable heterogeneity across model architectures and configurations (where "architecture" refers to the underlying foundation model, and "configuration" to the experimental variant defined by the prompt, i.e. the text input given to the model, and the temperature parameter controlling the degree of randomness in its responses). All models and configurations are substantially more stable than humans: they do not generate bubble episodes and keep prices close to the fundamental REE. However, they do not converge to the fundamental REE and produce forecasts that are biased and inefficient, leading to higher-order autocorrelation in market returns and, hence, unexploited arbitrage opportunities. In addition, AI Agents derived from the same foundation model fail to improve forecast coordination over time, maintaining persistent dispersion in gains from trade. When exposed to the bullish shock, most models and configurations are able to prevent the emergence of an explosive bubble, with the degree of stabilization significantly mediated by temperature and prompt design. The estimated hype susceptibility parameter highlights substantial heterogeneity in stabilizing potential across the six models, with Haiku exhibiting the weakest (more destabilizing) and DeepSeek the strongest (more stabilizing) performance.

The paper proceeds as follows. The next section reviews the literature. Section 3 presents the structural model underlying the LtFE and explains how the bullish, or hypers, shock is introduced. Section 4 describes the experimental design and implementation. Section 5 presents the empirical analyses, and Section 6 concludes.

2 Literature

The literature review is organized into four parts. It begins with a concise, non-technical overview of recent developments of GenAI models, with a focus on LLMs. The second part examines the

transition from AI models to AI Agents, introducing the notion of agentic AI and summarizing recent contributions on multi-agent systems, including their applications in economics. The third part turns to GenAI as a forecasting technology, reviewing advances in adapting foundation models for time-series prediction and recent evidence from economics and finance. The final part briefly touches on the role of expectations in economic modeling, focusing on theoretical departures from Rational Expectations developed in the learning literature in macroeconomics, and the experimental tradition of Learning-to-Forecast Experiments.

GenAI: A Very Brief Overview of LLMs

LLMs are large neural networks built on the Transformer architecture (Vaswani et al., 2017), comprising billions of parameters and trained on massive datasets (Mienye et al., 2025). Their training typically occurs in two main stages.[7] The first, referred to as *pre-training*, initializes all model parameters and trains them on a very large dataset, originally composed only of textual data. The dataset is divided into small units called *tokens*, which can represent words, subwords, or even characters. The model learns to predict the most likely next token given all previous ones, a task known as *next-token prediction*. Through this process, it develops general linguistic abilities. The second stage, usually called *alignment*, is a supervised phase that adjusts the model's behavior to better follow human instructions and produce responses that are useful and desirable. This typically involves techniques called Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF). Alignment improves responsiveness to user intent but also risks transferring human biases or encouraging overly pleasing behavior (Hosking et al., 2023; Kirk et al., 2024).

After alignment, models can be further refined for specific tasks through a process known as *fine-tuning*. Alternatively, or in addition, their performance can also be enhanced through *prompting* strategies (for example, *chain-of-thought* prompting, which encourages step-by-step reasoning) or *in-context learning*, where examples are provided within the input prompt to guide the model's output. When a model generates an output without additional examples or fine-tuning, it is referred to as *one-shot generation*; when a few examples are provided, it is called *few-shot generation*. Finally, during *inference*, the stage when the already trained model is actually used to generate an output, it produces text probabilistically. A parameter called temperature controls this randomness: lower values make the output more predictable, while higher ones increase variability.

Scaling model size, training data, and computational resources has led to sharp improvements in performance of LLMs across a wide range of tasks. Some researchers interpret these as *emergent abilities*: capabilities that appear suddenly as models grow larger, exceeding what they were explicitly trained to do (Wei et al., 2022; Berti et al., 2025). Yet, LLMs can also produce *hallucinations*:

[7] This section draws on Zeng et al. (2025) and Mienye et al. (2025).

factually incorrect but plausible outputs (Mienye et al., 2025).[8]

Some of the most recent LLMs have been described as *reasoning models* (some early examples are GPT-o1, o3 and DeepSeek R1), which are explicitly designed to improve reasoning-based performance. These models incorporate techniques such as self-critique, multiple reasoning steps, and other inference procedures which improve their reasoning-like capabilities (Guo et al., 2025), but also increase inference time.

Despite the main architectural principles being known, the specific components and training details of leading models, including dataset composition and pre- or post-training practices, remain largely undisclosed, as most are proprietary and closed-source. Recent developments such as Retrieval-Augmented Generation (RAG), which enables access to external information, and multimodal training, which allows models to process and generate text, images, or other data modes, further enhance performance but also deepen opacity. Interpretability remains a major challenge in AI research, stemming both from the intrinsic complexity of LLM architectures and from the lack of transparency surrounding their design and training. As a result, these models are still often regarded as "black boxes" (Mienye et al., 2025; Lin et al., 2025).

From Models to Agents: The Emergence of Agentic AI

A recent paradigm shift is the transition from models to agents. As Plaat et al. (2025) observes, "Models predict, agents reason, act, and interact." Agentic LLMs extend traditional models by endowing them with decision-making capabilities, tool use, and the ability to interact with other agents and the (still virtual) environment. These features make them more autonomous and adaptive, but also more complex and opaque (Plaat et al., 2025).

Research in Computer Science increasingly studies how multiple AI Agents, mainly LLM-based, interact within shared environments. Multi-agent systems, in which agents cooperate, compete, or learn from one another, can display emergent collective abilities that exceed the sum of their individual performances (Chen et al., 2023; Conitzer and Oesterheld, 2023; Tran et al., 2025; Willis et al., 2025; Burton et al., 2024). These systems leverage mechanisms such as information sharing, consensus building, cross-model synergies, and parallel task execution (Chen et al., 2023; Conitzer and Oesterheld, 2023; Tran et al., 2025). Some scholars also argue that for AI systems to be resilient, it is better to rely on ensembles of smaller, specialized, and independent models rather than on a single all-encompassing one (Tran et al., 2025).

A key challenge in this line of research is the rapid evolution of both LLM architectures and

[8] A vast literature studies LLM evaluation, covering linguistic dimensions (fluency, coherence, factuality), social aspects (fairness, bias, toxicity), and task-specific traits (helpfulness, reasoning). Evaluations often rely on human or even LLM-as-judge assessments (Li et al., 2024; Thakur et al., 2024). With the rise of LLM Agents, judgment becomes integral to their function, such as assessing document relevance or source reliability. Multi-agent systems can further enhance judgment quality (Moniri et al., 2024).

agentic frameworks, which often outpaces systematic evaluation. Most studies rely on simulated environments where LLMs are assigned specific roles via prompting strategies that define goals, constraints, and interaction rules. Broadly, these studies compare different forms of collective interaction - such as coordination, competition, negotiation, consensus-building, and division of labor among task-specific agents - to evaluate their relative effectiveness in improving overall system performance (Conitzer and Oesterheld, 2023; Tran et al., 2025). Simulation-based designs remain the prevailing approach for studying collective agentic AI behavior.

AI Agents in Economics

Research on the extent to which LLM Agents align with economic theories of preferences and decision-making is still nascent and offers mixed evidence (Bybee, 2023; Fan et al., 2024; Sun et al., 2024; Ma et al., 2024; Kim et al., 2024; Filippas et al., 2024; Polachek et al., 2024; Roberts et al., 2024; Lorè and Heydari, 2024; Kitadai et al., 2025; Wu, Xi, et al., 2025). Most studies employ models from various providers[9] in simulated one-shot or few-shot settings to evaluate their ability to reproduce human preferences, beliefs, and decisions in strategic or risky environments, and to assess their potential as cost-effective substitutes for human participants in surveys and experiments. Findings indicate that even earlier-generation models, such as GPT-3.5 and GPT-4, can approximate human survey responses when provided with basic demographic context (Sun et al., 2024; Ma et al., 2024) and display patterns of risk aversion and strategic behavior broadly consistent with human behavior (Filippas et al., 2024), particularly when persona-based prompting is used (Kitadai et al., 2025) and when model size is sufficiently large (Roberts et al., 2024). Evidence also suggests limited sensitivity of strategic behavior to prompt framing (Lorè and Heydari, 2024). However, results overall remain heterogeneous and model-specific, and the evidence does not yet point in a single, consistent direction (Fan et al., 2024; Polachek et al., 2024).

Simulations involving multiple LLM Agents are also beginning to emerge in economics and finance. Researchers have started examining their interactions in game-theoretic settings to study the types of strategic behavior and equilibria that arise (Karabag and Topcu, 2025; Akata et al., 2025; Shah et al., 2025). Akata et al. (2025) find that earlier-generation models, particularly GPT-4, often coordinate on suboptimal Nash equilibria in repeated games, exhibiting a tendency to defect and limited coordination ability. Using more recent models such as Gemini 2.5 Pro and GPT-5, Karabag and Topcu (2025) show that in a Chameleon Game setting, LLM Agents tend to reveal excessive information when interacting with agents of unknown identity. Finally, Shah et al. (2025) study a series of auctions with LLM bidders and find behavior consistent with risk-averse human bidders, largely robust to simple variations in prompting.

[9] Table A.1 reports the models used in each study considered here. Although not exhaustive, this survey indicates a predominant reliance on models from the GPT family.

Similar applications have also emerged in finance, where LLMs are analyzed as autonomous traders in simulated markets, again with mixed evidence (Lopez-Lira, 2025; Hansen and Lee, 2025; Henning et al., 2025). Lopez-Lira (2025) employ GPT-4o as a trading agent and find that it reproduces several stylized facts of real financial markets, including price discovery, bubbles, underreaction, and strategic liquidity provision. In contrast, Henning et al. (2025) use a broader set of models, including GPT-4o, in an asset-pricing experiment à la Smith and document clear behavioral differences between LLM- and human-driven markets: LLM traders display what the authors describe as "textbook-rational" behavior, with low volatility and prices closely aligned with fundamentals.

GenAI as a Forecasting Technology

A growing line of research in Computer Science explores how large pre-trained neural architectures, known as foundation models, can be adapted for time-series forecasting (Miller et al., 2024; Ye et al., 2025). Beyond purely language LLMs, these models include architectures originally developed for vision or numerical tasks. Recent studies investigate how they can be modified or combined to capture temporal dependencies and how multimodal inputs, textual, numerical, or visual, can enhance predictive accuracy (Miller et al., 2024; Ye et al., 2025). Overall, foundation models now achieve forecasting performance comparable to, and sometimes exceeding, traditional techniques, though there is no consensus on the optimal architecture, training method, or data representation. Research is expanding rapidly and show promising results (Miller et al., 2024; Ye et al., 2025).

In parallel, economists have begun exploring Generative AI, and LLMs in particular, as a new forecasting technology (Lopez-Lira and Tang, 2023; Faria-e-Castro and Leibovici, 2024; Schoenegger et al., 2024; Carriero et al., 2024; Hansen, Horton, et al., 2024). Most studies treat LLMs such as GPT-4 and GPT-4o, as one-shot or few-shot forecasters, or employ variants pretrained on time-series data, such as Time-GPT (Carriero et al., 2024). These works generally find that LLMs achieve promising levels of accuracy in economic and financial prediction tasks, comparable to, and in some cases exceeding, those of professional forecasters and standard econometric techniques.[10]

A central challenge in evaluating LLM forecasts lies in assessing the true out-of-sample nature of their predictions. As noted by Lopez-Lira, Tang, and Zhu (2025), LLMs may exhibit a memorization

[10] In finance in particular, several efforts have sought to adapt foundation models for financial-specific applications. Early examples include BloombergGPT (Wu, Irsoy, et al., 2023), a closed-source model trained on extensive proprietary financial data and designed for broad task performance within the industry, and FinGPT (Liu et al., 2023), an open-source alternative trained on publicly available datasets. More recent approaches integrate multiple data modalities, combining time-series data with financial news, social media sentiment, corporate disclosures, or executive speeches, to perform prediction tasks with notable performance (Jia et al., 2024; Fatemi and Hu, 2024; Xiao et al., 2025; Wang et al., 2024). Multimodal architectures hold particular promise for financial forecasting, as integrating diverse data sources is generally shown to better capture market tendencies. Yet, there is still no clear consensus on how these modalities should be best integrated. (Lopez-Lira, Kwon, et al., 2025)

problem when evaluation data overlap with their training corpus. To address this, some studies carefully select validation data likely outside training sets or rely on simulated data. This study follows the latter approach, using simulated data to assess one-shot forecasting performance.

Finally, most existing evaluations take a partial-equilibrium perspective, ignoring feedback between forecasts and realized outcomes. Yet, in many settings, the variables being predicted are themselves shaped by expectations. Rio-Chanona et al. (2025) address this challenge using a Learning-to-Forecast Experiment (LtFE) framework. In this paper, I adopt a similar approach to study how LLMs form expectations when outcomes depend on those very forecasts.

Expectations and Learning in Economics

Modern macroeconomic models are dynamic systems populated by agents who are forward-looking optimizers. To solve their optimization problems, these agents must form expectations about future endogenous variables (Evans and Honkapohja, 2009). In other words, for the agents' dynamic optimization problems to be well-defined, they must hold a *perceived law of motion* for variables beyond their control. Given this, the modeler can solve for their optimal decisions and, by aggregating these decisions and imposing market clearing, the modeler obtains the *actual law of motion* of the endogenous variables. A Rational Expectations Equilibrium (REE) is then defined as a fixed point in the mapping between the perceived and the actual laws of motion (Sargent, 1987).

Muth (1961) is generally credited with the first explicit formulation of the concept of Rational Expectations (RE), while the works of Lucas (1972) and Sargent (1973), along with the Lucas (1976) Critique, are often seen as the starting point of the so-called Rational Expectations revolution in macroeconomics. For decades, the REE remained the dominant equilibrium concept in macroeconomic modeling.

The RE hypothesis has long been the subject of debate. Critics argue that it relies on an excessively stringent notion of rationality, one that appears inconsistent with empirical evidence on bounded rationality (Hommes, 2021). Although RE is fundamentally an equilibrium concept rather than a behavioral assumption, it naturally raises the question of whether and how a group of agents with bounded rationality can converge to a REE (Evans and Honkapohja, 2005). In addition, dynamic RE models often face difficulties in replicating certain empirical and experimental regularities (Marimon, 1996; Hommes, 2021).

Learning theory in macroeconomics has been developed to address some of these challenges. It provides frameworks that incorporate bounded rationality in expectation formation, allow for deviations from RE dynamics, and at the same time offer an equilibrium justification and selection mechanism for REE (Marimon, 1996). In this literature, agents are typically assumed to form expectations according to explicit behavioral rules that may evolve as new data becomes available. These rules may be based on (possibly misspecified) econometric models (Sargent, 1993; Evans and

Honkapohja, 2001), simple heuristic strategies (Anufriev and Hommes, 2012), or more sophisticated adaptive learning algorithms, such as the Holland (1975) classifier systems (Marimon, McGrattan, et al., 1990). In addition, agents may rely on heterogeneous rules or choose among a set of competing forecasting strategies (Brock and Hommes, 1997; Anufriev and Hommes, 2012; Evans and Honkapohja, 2009; Di Francesco and Hommes, 2025).

The asymptotic behavior of the resulting stochastic dynamic systems defines the set of *learnable equilibria* (Marimon, 1996). In particular, the concept of *expectational stability* (or e-stability) provides a criterion to assess the asymptotic stability of a REE under a given learning rule (Evans and Honkapohja, 2001). Moreover, these systems may converge to different types of equilibria, such as Self-Confirming Equilibria, of which REE represents a special case (Cho and Sargent, 2008; Fudenberg and Levine, 2009).

Simulation-based studies employing complex learning rules, such as those used here with LLMs, are not new in this literature. Marimon, McGrattan, et al. (1990) examine the exchange economy of Kiyotaki and Wright (1989), where agents use classifier systems to make decisions, and show that in environments with multiple equilibria, agents converge to the fundamental Nash equilibrium in which goods with low storage costs serve as the medium of exchange. Arifovic (1994) studies the Cobweb model in which competitive firms use genetic algorithms as their learning rule, and shows convergence to the REE across a wide range of parameter values. Similarly, Arifovic (1995) analyzes two overlapping-generations monetary models with two stationary REE and finds convergence to the equilibrium with valued fiat money in one case and to the low-inflation equilibrium in the other. More recently, Ashwin et al. (2025) study neural-network learning, both analytically (for relatively simple cases) and through numerical simulations, in economies characterized by nonlinear REE.

Learning-to-Forecast Experiments

As noted by Marimon (1996), the learnability of an equilibrium is ultimately an empirical question. In a series of influential studies, Marimon, Spear, et al. (1993), Marimon and Sunder (1993), and Marimon and Sunder (1994) developed an experimental framework to address the indeterminacy of REE in macroeconomic models by recreating these environments in the laboratory and allowing subjects to form expectations. This approach enables researchers to control the structure of the economic environment, which may feature multiple REE[11], and the information available to participants. It also allows for the observation of agents' behavior and system dynamics along non-equilibrium paths, an aspect that standard RE theory cannot capture.

This experimental tradition has evolved into what is now known as *Learning-to-Forecast Experiments* (LtFEs) (Hommes, 2011; Hommes, 2021). A crucial insight from this literature

[11]The main motivation of Marimon and Sunder (1993) is indeed to use this experimental framework to address the indeterminacy problem of some REE models.

is the distinction between environments with negative and positive expectations feedback. In negative-feedback settings, higher individual forecasts induce aggregate outcomes that move in the opposite direction, whereas in positive-feedback settings, higher forecasts reinforce themselves as aggregate outcomes move in the same direction. Experiments show that negative feedback tends to generate stable dynamics, while positive feedback often gives rise to high volatility and recurrent boom–bust episodes (Heemeijer et al., 2009; Hommes, 2021). Moreover, these experiments consistently document heterogeneity in expectations and persistent deviations from REE trajectories (Hommes, 2021).

3 Model

The baseline model follows the asset-pricing setup of Hommes et al. (2008) to ensure comparability with their human benchmark. The economy features a single risky asset with price p_t at time t, paying a random dividend y_t that follows an i.i.d. normal process with mean $\bar{y}$ and variance σ_y^2, which is common knowledge, and a risk-free asset yielding a gross return $R > 1$. There are N traders, each endowed with initial wealth W_0, who in every period decide how to allocate their wealth between the risk-free asset and the risky asset, denoted by $x_{i,t}$, by balancing expected next-period wealth against its variance. Traders are homogeneous in both initial wealth and constant risk aversion A, and differ only in their expectations about the next-period price of the risky asset. Each trader therefore solves the following optimization problem in every period:

$$\begin{aligned} \max_{x_{i,t}} \quad & E_{i,t}\left[W_{i,t+1}\right] - \frac{1}{2} A \operatorname{Var}_{i,t}\left[W_{i,t+1}\right] \\ \text{s.t.} \quad & W_{i,t+1} = R(W_{i,t} - p_t x_{i,t}) + x_{i,t}(p_{t+1} + y_{t+1}), \\ & y_t \overset{\text{i.i.d.}}{\sim} N(\bar{y}, \sigma_y^2), \\ & W_{i,0} \text{ given.} \end{aligned}$$

Following Hommes et al. (2008), I assume that all traders share the same perceived variance of $\operatorname{Var}_{i,t}(p_{t+1} + y_{t+1}) = \sigma^2$. This assumption simplifies the problem and reduces expectation heterogeneity to differences in the expected next-period price of the risky asset, denoted by $p_{i,t+1}^e$. The resulting optimal demand for the risky asset is given by:

$$x_{i,t} = \frac{p_{i,t+1}^e + \bar{y} - R p_t}{A\sigma^2} = \frac{\tilde{R}_{i,t+1}^e}{A\sigma^2},$$

where $\tilde{R}_{i,t+1}^e$ denotes the expected excess return. Since there is no external supply of the risky asset,

the market-clearing condition at each period is:

$$\sum_{i=1}^{N} x_{i,t} = 0,$$

which implies that the market-clearing asset price at time t is:

$$p_t = \frac{\bar{y} + \bar{p}^e_{t+1}}{R}, \tag{1}$$

where $\bar{p}^e_{t+1} = \frac{1}{N}\sum_{i=1}^{N} p^e_{i,t+1}$ represents the average next-period price expectation across traders.

A key property of this model is the positive feedback between expectations and realized prices: higher expected prices lead to higher actual prices, which in turn reinforce optimistic expectations. Moreover, if $p^e_{i,t+1}$ were a strategic choice and the market-clearing condition were known, the model would feature strategic complementarities among traders' expectations.

Equation (1) determines the market-clearing price of the asset at time t as a function of agents' expectations for the price at time $t+1$. In a Learning-to-Forecast Experiment, as in Hommes et al. (2008), these expectations are directly elicited from participants in each period (or round). Given these elicited expectations, the market price is determined by (1) and fed back into the participants' information set for the next round. The participants can be human subjects, as in Hommes et al. (2008), or AI Agents, as in this paper. I describe the experimental design in more detail in Section 4.

3.1 Rational Expectations Equilibrium

An REE in this economy is defined by the market-clearing condition (1) together with the Rational Expectations condition of Muth (1961):

$$p^e_{t+1} = E_t[p_{t+1}], \tag{2}$$

which states that, in equilibrium, agents' expectations of next-period prices coincide with the mathematical expectation of the actual realized next-period prices. Substituting (2) into (1) yields the stochastic difference equation:

$$p_t = \frac{\bar{y}}{R} + \frac{1}{R}E_t[p_{t+1}]. \tag{3}$$

The well-known general solution to (3) is:

$$p_t = p^f + B_t, \tag{4}$$

where p^f is the *fundamental price*, given by:

$$p^f = \frac{\bar{y}}{R-1},$$

and B_t is the *rational bubble*, a martingale process satisfying:

$$B_t = \frac{1}{R} E_t[B_{t+1}].$$

Equation (4) characterizes the infinite set of REE in this economy. Rational bubbles are often ruled out by imposing a transversality condition. Alternatively, it has been shown that the fundamental price is the unique *e-stable* REE under a broad class of adaptive learning rules (Evans and Honkapohja, 2001; Hommes et al., 2008).

3.2 Testable Features of REE

3.2.1 Market Efficiency

In the fundamental REE, excess returns are unpredictable and, on average, null, consistently with the market efficiency hypothesis (Fama, 1970). Recall the definition of excess returns:

$$\tilde{R}_{t+1} \equiv p_{t+1} + y_{t+1} - Rp_t$$

By substituting the REE prices (4) above, we get:

$$\tilde{R}_{t+1} = y_{t+1} - \bar{y} + B_{t+1} - RB_t$$

This implies that:

$$E_t[\tilde{R}_{t+1}] = 0 \quad \forall t \tag{5}$$

If we focus on the fundamental REE, where $B_t = 0$ for all t, we also get:

$$\tilde{R}_{t+1} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_y^2)$$

which implies that the process is serially uncorrelated:

$$\rho_k = 0 \quad \forall k > 0 \tag{6}$$

where ρ_k is the autocorrelation of order k of the excess returns process. Note that in the fundamental REE the process is ergodic and stationary and, therefore, conditions (5) and (6) are testable using the realized time-series of prices.

3.2.2 Forecast Unbiasedness and Efficiency

The second feature comes directly from condition (2), which has a long tradition of empirical tests (Mincer and Zarnowitz, 1969; Lovell, 1986; Romer and Romer, 2000). Here, I adopt the empirical specification of Keane and Runkle (1990):

$$p_{t+1} = \alpha_0 + \alpha_1 p^e_{i,t+1} + \alpha_2 X_{i,t} + \epsilon_{t+1}, \quad E_t[\epsilon_{t+1}] = 0 \tag{7}$$

where $X_{i,t}$ is any variable known by trader i at time t. RE (2) implies the following two conditions:

$$\textbf{Unbiasedness:} \quad \alpha_0 = 0,\ \alpha_1 = 1 \tag{8}$$

$$\textbf{Efficiency:} \quad \alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0 \tag{9}$$

Note that conditions (8) and (9) are testable using individual forecast and realized prices. I want to stress here that condition (2) is an equilibrium condition, not a behavioral assumption. Indeed, a (learnable) REE can be attained even when agents use forecasting rules that are considered to be boundedly rational (Marimon, 1996; Evans and Honkapohja, 2001). In the context of the asset pricing model, the fundamental REE can be attained, and it is *e-stable*, even under naive expectations (Hommes et al., 2008).

3.2.3 Gains from Trade

Finally, in a REE, there are no systematic gains from trade. Consider again the individual demand for the risky asset, and assume without loss of generality that $A\sigma^2 = 1$, so that $x_{i,t} = \tilde{R}^e_{i,t+1}$ [12]. Substituting it into the budget constraint, and rearranging, I get:

$$W_{i,t+1} - (1+r)W_{i,t} = \tilde{R}^e_{i,t+1}\tilde{R}_{t+1} \tag{10}$$

I can interpret this quantity as the gains from trading the risky asset for trader i, call it $\tilde{W}_{i,t+1}$. Let's inspect it more carefully:

$$\tilde{W}_{i,t+1} = \Big(p^e_{i,t+1} + y_{t+1} - (1+r)p_t\Big)\Big(p_{t+1} + y_{t+1} - (1+r)p_t\Big)$$

[12] A similar derivation is also present in Di Francesco and Hommes (2025).

First, note that the asset pricing model with no external supply is a zero-sum game, irrespectively of the expectations process. That is, if someone wins, someone else is losing. Indeed, using the market-clearing condition (1) above, it's easy to see that:

$$\sum_{i=1}^{N} \tilde{W}_{i,t+1} = 0$$

An REE has the additional property that no trader can systematically win or lose from trade. Indeed, the RE condition (2) implies that $E_t[\tilde{W}_{i,t+1}] = 0$ at the individual level and for each period.

However, the model generally allows for dispersion in the gains from trade. In particular:

$$\text{Var}[\tilde{W}_{i,t+1}] = \tilde{R}_{t+1}^2 \, \text{Var}[\tilde{R}_{i,t+1}^e], \tag{11}$$

where $\text{Var}(\cdot)$ denotes the cross-sectional variance across traders at time t. Equation (11) shows that the dispersion of gains from trade increases with the degree of cross-sectional disagreement over $p_{i,t+1}^e$, and is minimized when expectations are homogeneous. In other words, full coordination on a common $p_{i,t+1}^e$ implies zero dispersion in gains from trade, while greater disagreement leads to higher dispersion. Such dispersion can also arise in a REE. However, in that case, it is purely random. This provides a testable implication: one can test whether dispersion in gains come from some traders systematically outperforming others. Moreover, although not a direct implication of RE, one can also test whether dispersion in gains from trade decreases over time, as agents possibly learn to coordinate their expectations.

3.3 Hypers' Shock

I now introduce a shock to the above economy designed to inflate market prices and potentially lead to explosive dynamics in the equilibrium prices. I model this shock as the entry of a new class of traders called "hypers", who are (irrationally) bullish, in that they expect prices to grow at an exponential rate no matter what is the market feedback. In particular, hypers have the following expectation rule:

$$p_{t+1}^e = \gamma R^2 p_{t-1} \tag{12}$$

with γ being a real-valued positive constant. The rationale behind the choice of (12) as expectation rule for hypers is the following. Imagine that the economy is populated by hypers only. Then, the

market clearing price would be:

$$p_t = \frac{\bar{y}}{R} + \gamma R p_{t-1} \tag{13}$$

which is a first-order non-homogenous linear difference equation describing the price dynamics of the economy, with steady-state equal to

$$p^{ss} = \frac{\bar{y}}{R(1-\gamma R)}$$

Note that the difference equation (13) is unstable as long as the eigenvalue $\gamma R > 1$. If this is the case, the growth rate of market price asymptotically converges to γR since:

$$\lim_{t\to\infty} \frac{p_t}{p_{t-1}} = \lim_{t\to\infty} \left(\frac{\bar{y}}{R p_{t-1}} + \gamma R \right) = \gamma R$$

As a result, when $\gamma R > 1$, there are three asymptotic cases:

- If $\gamma < 1$, the price asymptotically grows at a rate lower than what hypers expect, thus dampening their expectations if they were adaptive learners.
- If $\gamma = 1$, the price asymptotically grows at a rate equal to R, and hence the expectations of hypers (12) are asymptotically self-fullfilling. This is also noted by Hommes et al. (2008).
- If $\gamma > 1$, the price asymptotically grows at a rate higher than what hypers expect, more than reinforcing their expectations if they were adaptive learners.

As a result, agents with the expectation rule (12) and $\gamma \geq 1$ are the natural candidates to be "hypers", with γ associated with the "hype" they are bringing into the market.

To operationalize the shock, I assume that at a certain time T_1, N_h hypers exogenously enter the market and remain active until time T_2, when they exogenously exit. Therefore, between T_1 and T_2, the market is populated by N traders and N_h hypers.

4 Design

I adapt the LtFE design of Hommes et al. (2008), and replace humans with LLMs, keeping the design as consistent as possible to have a human benchmark to compare my results with. In this section, I briefly explain the original setup, and then I describe how I construct and use LLM Agents in the same experimental framework.

Table 1: Features of Hommes et al. (2008)

Structure	Six markets, each with $N = 6$ participants and $T = 50$ periods.
Market Clearing	As in equation (1).
Role	Advisor-forecaster for a pension fund.
Task	Submit forecast $p^e_{i,t+1}$ each period t.
Earnings	Determined by forecast accuracy (quadratic scoring rule).
Information Set	
	Known
	• Role, task and earnings
	• Interest rate $r = 0.05$
	• Mean dividend $\bar{y} = 3$
	• Full history of prices $(p_{t-1}, p_{t-2}, \dots)$
	• Own past forecasts
	Unknown
	• Structure (nr and identity of other participants, nr of periods)
	• Market clearing mechanism
	• Trading strategy of their own and other pension funds

4.1 Original Setup

Hommes et al. (2008) is a LtFE in the laboratory based on the baseline structural model described above without the hypers' shock. The authors conducted six independent sessions, each with $N = 6$ participants and lasting for $T = 50$ periods. In every round, participants were asked to predict the asset price in period $t + 1$ while at period t, without observing the price in t, thus facing a *de facto* two-step-ahead forecasting task. Given all participants' forecasts, the realized price was determined according to the market-clearing condition (1). At each round, each participant's information set included the risk-free rate $r = 0.05$, the average dividend $\bar{y} = 3$, the full history of realized prices up to $t - 1$, and their own past forecasts. All other elements were outside the information set: the total number of rounds, the number and identity of other participants, the trading technology (they were only told that their prediction would guide a fund's investment decision), and any further details of the market structure, including the market-clearing mechanism. Participants' earnings were determined by the accuracy of their forecasts at the end of the experiment. Table 1 summarizes the main features of the experiment in Hommes et al. (2008).

4.2 LLM Agents

This section describes the implementation of the LtFE design using AI Agents. I first present the two experimental setups considered in the study. I then explain how the AI Agents were constructed, detailing the foundation models and prompting strategy used to operationalize their behavior. Finally, I describe the experimental variants within each setup and report the number of simulations conducted for each configuration.

4.2.1 Setups

I consider two experimental setups, referred to as the baseline and the hypers setups. In both cases, the underlying structural model follows the specification in Section 3, with structural parameters identical to those in Hommes et al. (2008): $N = 6$, $T = 50$, $r = 0.05$, and $\bar{y} = 3$. The information set available to the AI Agents is also the same as in Hommes et al. (2008), summarized in Table 1. Details on how this information is provided to the Agents are given below. In the baseline setup, I only replace human participants with AI Agents. In the hypers setup, I extend the baseline design by introducing the hypers' shock described in Section 3.3. Specifically, $N_h = 5$ hypers, mechanical traders following the expectation rule in equation (12) with $\gamma = 1.3$, enter the market at time $T_1 = 12$ and exit at time $T_2 = 21$. Thus, between periods T_1 and T_2, the market includes both $N = 6$ AI Agents and $N_h = 5$ hypers, with the latter forecasting deterministically according to their rule.

4.2.2 Construction of AI Agents

Each AI Agent is constructed following emerging practices in the literature that employs AI Agents in economic experiments (Filippas et al., 2024; Shah et al., 2025; Rio-Chanona et al., 2025). Agents are implemented using LLMs. Specifically, each Agent corresponds to an instantiation of a foundation model, accessed through an API call,[13] endowed with memory and tasked with performing the forecasting exercise through a specific prompting architecture, described below. In both the baseline and the hypers setups, simulations proceed as follows. At each round, $N = 6$ independent API calls to the same foundation model are made. Each instantiation becomes an Agent through a prompting procedure that provides it with memory and the task description. As a practical example, the Python routine independently calls GPT-4o six times per round, supplying each call with the same structured prompt that contains the relevant historical information and task instructions. Hence, each GPT-4o Agent operates in a market populated by five other GPT-4o Agents that share the same architecture and common information set, but with independent memories of their own past forecasts.

[13] An API call is a standardized way for a program to interact with another over the internet. Here, it means sending an input (the "prompt") to remote servers that run the model and receiving its output in return. LLMs are accessed in this way because most state-of-the-art models are proprietary and not publicly released, and even when open-source, their scale and computational requirements make local execution impractical.

Table 2: **Foundation Models Used**

Model	Version	Architecture	Active Par
Claude 3.5 Haiku	*claude-3-5-haiku-latest*	Undisclosed	Undisclosed
LLaMA 3.3 70B Turbo	*Llama-3.3-70B-Instruct-Turbo*	Dense	70 B
Gemini 2.0 Flash	*gemini-2.0-flash*	Undisclosed	Undisclosed
Grok 2	*grok-2-latest*	MoE	115 B
GPT-4o	*gpt-4o-2024-11-20*	Undisclosed	Undisclosed
DeepSeek Chat	*deepseek-chat*	MoE	37 B

Foundation Models Used Six foundation models were employed, as listed in Table 2. LLaMA 3.3 70B Turbo was accessed through the Meta (TogetherAI) interface, while all other models were accessed via their official APIs. The column Version reports the exact string used for each API call. Architecture distinguishes between the two main design approaches, when disclosed: in Dense architectures, all parameters are active during inference, whereas in Mixture-of-Experts (MoE) architectures, only a subset of modules is activated per token. Active Par. indicates the number of parameters effectively used during inference, when disclosed. All simulations were conducted between May and September 2025.[14]

Prompting Architecture Agents are endowed with memory through a structured prompting architecture consisting of a "system prompt" and a "user prompt", the latter containing fixed and rolling components. The system prompt is a high-level instruction that defines the Agent's role within the experiment and specifies how it should behave in general terms. The user prompt is the task-specific input that describes the forecasting exercise and provides the relevant information. The fixed part of the user prompt describes the environment and includes the risk-free rate, the dividend process, and the task instructions. The rolling part updates dynamically with the full history of realized prices, the Agent's past forecasts, and its own reflections. These reflections are short free-text notes that the LLM Agent can generate at each round, a common practice in similar experimental designs (Henning et al., 2025; Lopez-Lira, 2025). As shown by Renze and Guven (2024), granting Agents access to their past reflections helps them maintain continuity and improve performance in multi-step tasks of this kind. The exact texts of the system and user prompts are reported in Appendix B.

[14]Version strings are reported for transparency. Most models listed in Table 2 are closed-source and may undergo internal updates under the same API identifiers. To the best of my knowledge, only DeepSeek Chat introduced a notable update in late August 2025, after most simulations were completed.

Table 3: Simulations for each setup and foundation model.

Variant	Prompt	Temperature	Repetitions
1	Base	Medium	25
2	Base	Low	20
3	Base	High	30
4	Caution	Medium	25

4.2.3 Variants

To assess robustness, I implement four experimental variants for each foundation model and each setup, differing by prompt type and temperature. The temperature parameter controls the degree of randomness in output generation: lower values make the model's responses more deterministic, whereas higher values introduce greater variability and exploratory behavior. Varying the temperature therefore allows me to examine market dynamics under different levels of stochasticity in the Agents' behavior, and to verify that observed dynamics are not artifacts of more deterministic responses.

Two versions of the user prompt are used, which I refer to as the "Base prompt" and the "Caution prompt". The Caution prompt is identical to the "Base prompt", but it also includes an additional instruction suggesting that the market environment may undergo temporary structural changes. This variation tests whether explicitly warning the LLM Agents about possible regime shifts influences their expectations and market outcomes.

Each variant is repeated multiple times under identical conditions. Such repetition is standard practice in experiments using LLMs and serves two main purposes: (i) to account for the intrinsic stochasticity of LLM outputs, which prevents exact replication of a single run even at low temperatures; and (ii) to generate sufficient observations for statistical inference. A summary of all variants, including the number of repetitions for each one, is provided in Table 3.

5 Empirical Analyses

The simulations conducted for each foundation model listed in Table 2, across all variants and repetitions described in Table 3, and for the two setups (baseline and hypers), generated a rich dataset containing forecasts, reflections and prices.[15] The resulting dataset has a panel structure, uniquely identified by the combination of model, setup, variant, simulation number, and trader number. For each model-setup–variant–simulation combination, the panel comprises six traders observed over fifty rounds, producing fifty forecasts (numeric) and corresponding reflections (text)

[15]Although exact replication is not possible due to the stochastic nature of LLMs and the evolving availability of closed-source models, the full Python code and complete dataset will be made publicly available on my GitHub page.

per trader, together with a time series of realized market prices. This structure is consistent across all models and variants, although no two simulations are identical due to the intrinsic stochasticity of LLMs. Summary statistics of the dataset are reported in Appendix C.1.

In what follows, I present the results of the empirical analyses. I proceed in two steps. First, I examine the baseline setup, comparing market prices generated in markets populated by LLM Agents with those observed in the original experiment with human participants (Hommes et al., 2008). Because the prices generated by LLM Agents are substantially more stable and closer to fundamentals, I then formally test whether they converge to the fundamental REE, using the tests described in Section 3.2. The results suggest that LLM Agents act as stabilizing, yet not fully rational, learners. Having established this stabilizing pattern under baseline conditions, I next turn to the hypers setup to examine whether such stability is robust to external disturbances. The idea is somewhat analogous to check whether a dynamic system that has reached a steady state returns to it when perturbed. Here, I study how a market populated by LLM Agents responds to an exogenous bullish shock, and whether their (boundedly rational) expectations prevent the emergence of a bubble exogenously triggered. I show the resulting price dynamics, and conduct two complementary analyses. First, I estimate the heterogeneous impact of the hypers across model architectures and configurations (prompt type and temperature). Second, I perform a structural estimation of a "hype susceptibility" parameter, which allows ranking models according to their stabilizing potential.

5.1 Baseline Setup

5.1.1 Price Dynamics: LLM Agents vs Humans

Figure 1 and Figure 2 present the asset price dynamics observed in the original LtFE experiment with human forecasters and in the corresponding simulations with LLM agents, respectively. In particular, Figure 2 reports the average market prices across all repetitions of the three variants run with different temperatures. The comparison highlights marked differences in market behavior between humans and LLMs. In the human experiment, prices display pronounced volatility and recurrent bubble–bust episodes, deviating substantially from the fundamental value of 60. As discussed by Hommes et al. (2008), such dynamics appear to be driven by weak trend-following heuristics among human participants, consistent with other experimental evidence on human-induced bubbles and crashes. Subsequent research further formalizes the class of learning rules capable of reproducing these dynamics (e.g., Hommes, 2017).

By contrast, the LLM simulations exhibit substantially more stable price dynamics, with prices closely tracking the fundamental value and showing limited volatility. This stability is observed across all foundation models tested and remains robust to variations in temperature and user prompts (see Figure D.3 for the *Caution* prompt variant).

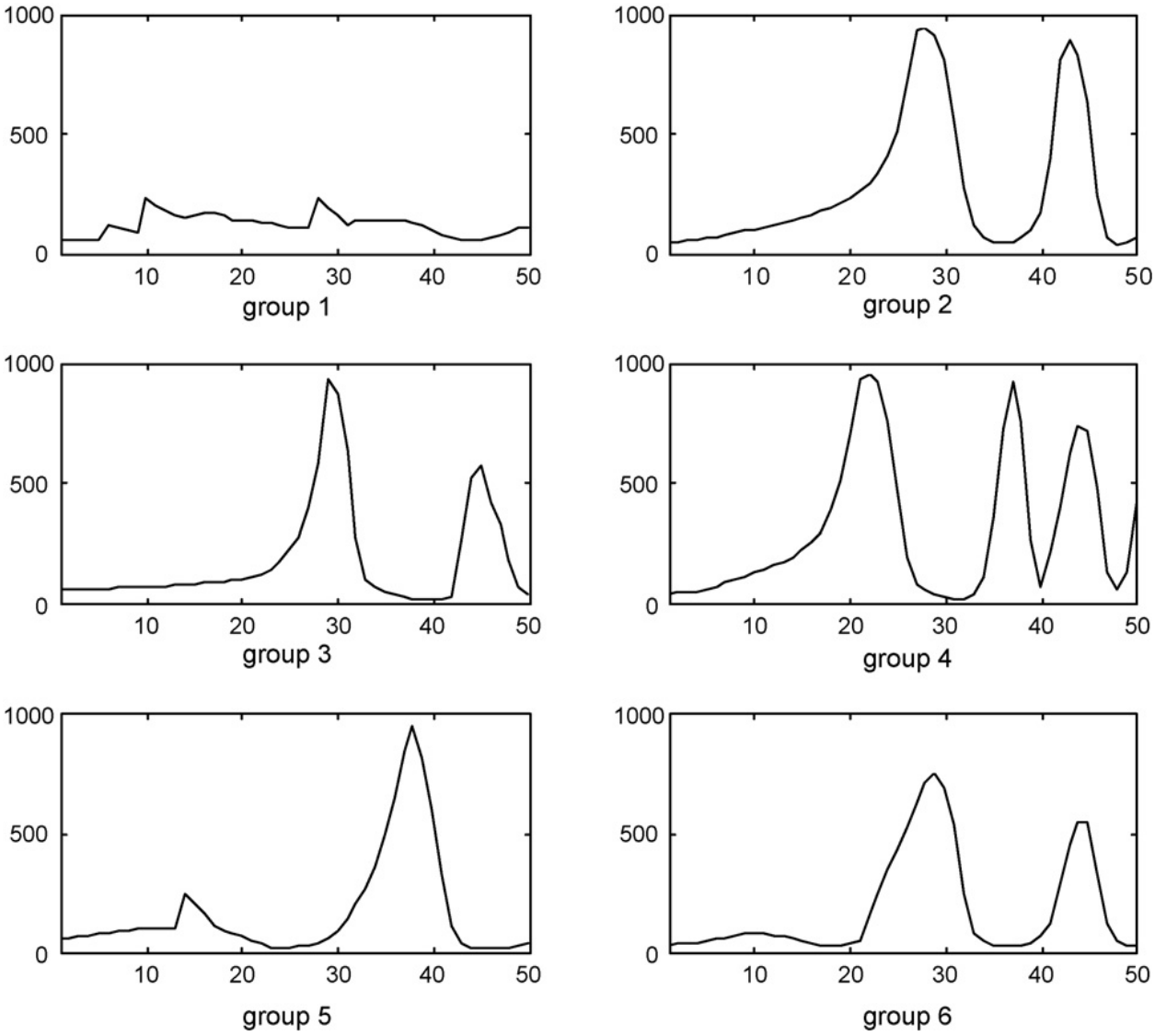


Figure 1: Asset price dynamics with 6 human forecasters for each group in the LtFE of Hommes et al. (2008). The figure is extracted from the original paper (Fig 3., Hommes et al., 2008).

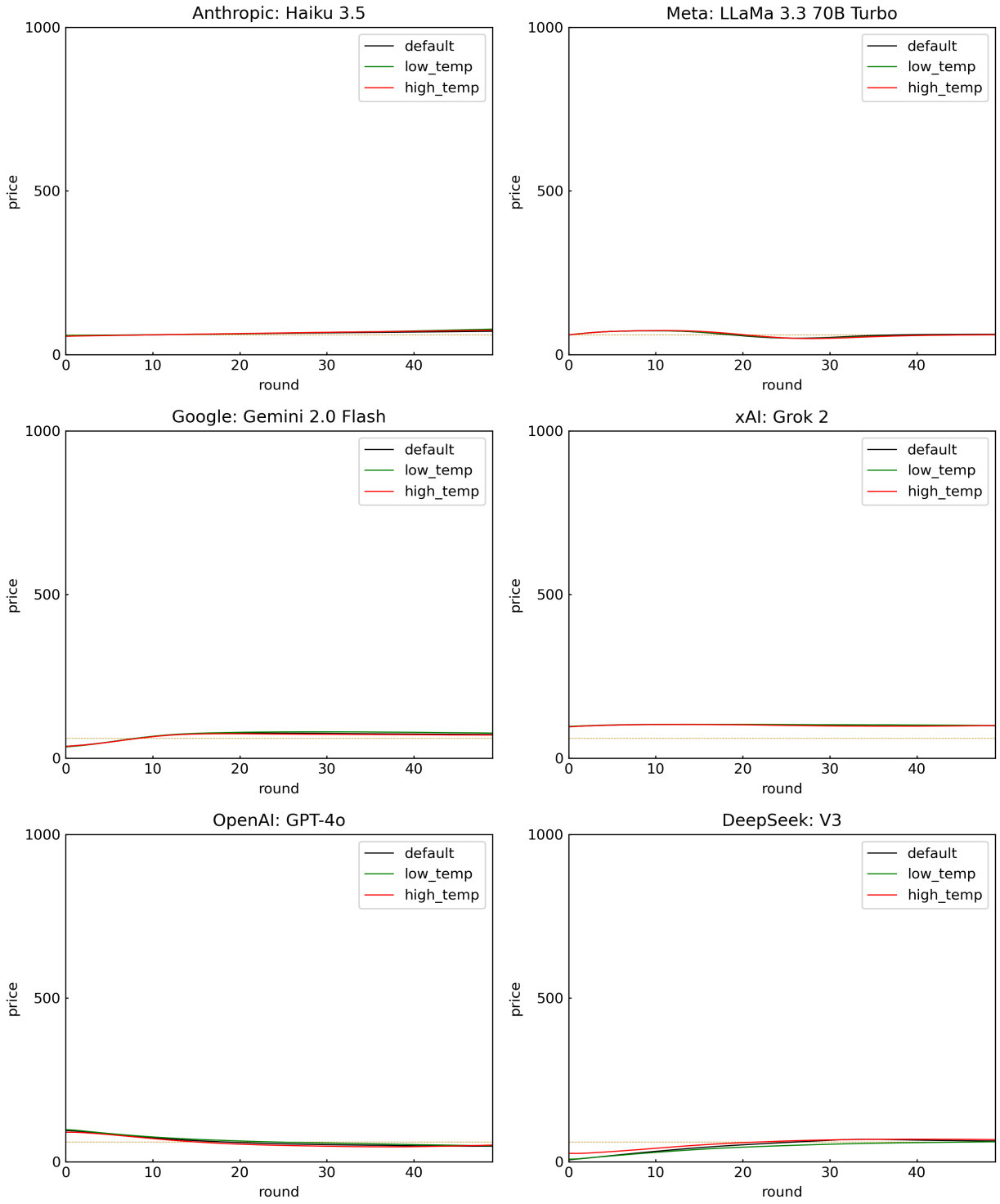


Figure 2: Asset price dynamics with 6 LLM Agents for each foundation model, in a LtFE parametrically identical to Hommes et al. (2008). The orange dashed line represents the fundamental price of the asset equal to 60. The displayed price is the average across identical simulations for the variants 1, 2 and 3 of Table 3.

Nevertheless, some heterogeneity across models is evident. DeepSeek and Gemini tend to start below the fundamental value but quickly converge toward it. Conversely, GPT-4o begins above the fundamental before converging, while Grok 2 exhibits a persistent upward drift. Finally, LLaMa generates prices that fluctuate the most around the fundamental among all models, whereas Claude's prices remain the closest to the fundamental throughout the entire horizon.

5.1.2 Rational Expectations Tests

The results discussed above are somewhat reassuring, as LLM Agents appear capable of generating forecasts that stabilize prices around their fundamental values. These findings are consistent with the "textbook-rational" behavior of LLM traders documented by Henning et al. (2025). In this section, I formally test whether LLM Agents actually attain the fundamental REE. Specifically, I assess three features of an REE, in addition to prices averaging at the fundamental value: market efficiency, forecast rationality, and the non-systematic dispersion of gains from trade.

Market Efficiency Table 4 reports the results of four market efficiency tests, testing the features of REE (5) and (6) derived above. For each model, I perform the tests both using all rounds and excluding the first nine to allow for an initial learning phase. Each test is applied to every model–variant–simulation combination, recording whether the null hypothesis is rejected at the 5% significance level. The reported rejection rates correspond to the share of rejections aggregated at the model level.[16] For each model, I also report the average price and the average realized excess return across rounds and simulations. The first test examines whether the average price differs significantly from the fundamental value of 60. The second and third tests assess whether average excess returns are zero, respectively assuming no autocorrelation and allowing for first-order autocorrelation, reflecting the two-step-ahead structure of the forecasting task.[17] The fourth column tests for the absence of higher-order autocorrelations in excess returns, using the Ljung-Box $Q - test$ on the residuals of an AR(1) model fitted to excess returns, with 5 lags. Across models, rejection rates are substantially higher than those expected under the Null with a 5% Type I error rate. The only exception is LLaMa in the zero–excess–returns test with the HAC(1) specification, where rejection rates fall below 0.05 in both the full sample and the restricted subsample. Results rare confirmed when examining rejection rates by variant and autocorrelation functions (see Appendix E.1).

Therefore, these markets leave room for arbitrage opportunities: LLM Agents do not generate fully efficient markets. Strikingly, this contrasts with the findings of Hommes et al. (2008), where

[16] Rejection rates here are pooled across all variants, but variant-specific rates for each model are reported in Appendix E.1.1.

[17] Recall that LLM Agents, like the human subjects in Hommes et al. (2008), form a forecast for the price in period $t+1$ at time t, without observing the realization of p_t. The observed price history is up to period $t-1$.

Table 4: Market Efficiency Tests (pooled variants)

Model	Rounds	K	Avg $\bar{P}$	Avg $\tilde{R}$	(1) $E[P] = P_F$	(2) $E[\tilde{R}_t] = 0$ i.i.d.	(3) $E[\tilde{R}_t] = 0$ HAC(1)	(4) $\rho_t = 0$ $t = 2, \ldots, 5$
Claude 3.5 Haiku	[1–50]	100	65.43	−0.210	.85	.69	.68	.31
	[10–50]	100	65.64	−0.202	.85	.75	.62	.25
Llama 3.3 70B	[1–50]	100	62.11	−0.261	.39	.26	.00*	.89
	[10–50]	100	58.95	−0.135	.25	.11	.05*	.81
Gemini 2.0 Flash	[1–50]	100	72.24	−0.116	.86	.34	.30	.61
	[10–50]	100	75.63	−0.724	.97	.88	.81	.52
Grok 2	[1–50]	100	101.95	−2.025	1.00	.99	.99	.20
	[10–50]	100	102.12	−2.191	1.00	1.00	.99	.08
GPT-4o	[1–50]	100	61.07	−0.987	.35	.93	.86	.24
	[10–50]	100	54.48	−0.307	.83	.57	.47	.36
DeepSeek Chat	[1–50]	100	49.50	1.572	.89	.91	.90	.32
	[10–50]	100	56.98	0.901	.91	.95	.90	.30

Notes: Columns (1)–(4) report the rejection rates of the null hypothesis (H_0) at the 0.05 significance level. Values at or below 0.05 (marked with *) indicate the null is not rejected. K denotes the number of simulations; Avg $\bar{P}$ is the average price across simulations and rounds; Avg $\tilde{R}$ is the average realized excess return across simulations and rounds.

(1) Fundamental pricing: $P_t - 60 = \alpha_0 + \varepsilon_t$; $H_0 \colon \alpha_0 = 0$ (OLS).

(2) Zero excess returns (i.i.d.): $\tilde{R}_{t+1} = \beta_0 + \varepsilon_{t+1}$; $H_0 \colon \beta_0 = 0$ (OLS).

(3) Zero excess returns (HAC): $\tilde{R}_{t+1} = \beta_0 + \varepsilon_{t+1}$; $H_0 \colon \beta_0 = 0$ (OLS with HAC(1) errors).

(4) No serial correlation (ρ_2–ρ_5): Ljung–Box Q-test on AR(1) residuals for $\tilde{R}_{t+1}$ with 5 lags.

most human groups produced approximately efficient markets according to autocorrelation tests on the period preceding the first bust. This evidence indicates that LLMs fail to attain a REE, leading to persistent excess returns dynamics with significant higher-order autocorrelations and, consequently, to the presence of unexploited arbitrage opportunities.

Forecast Rationality Table 5 reports the results of the forecast rationality tests based on conditions (8) and (9). The first column tests for forecast unbiasedness, while the second and third columns test for forecast efficiency using the regression specification (7). The two efficiency tests differ in the covariate used: column 2 includes the actual lagged forecast error, whereas column 3 employs a "perceived" version, relating the last forecast to the last observed price. Regression specifications and null hypotheses are indicated in the header of each column. All regressions are estimated using pooled OLS at the model level, where observations from all LLM Agents, simulations, and variants are combined. A dummy variable for the variant category is included to account for between-variant differences.[18] For each regression, I report coefficient estimates and standard errors, the adjusted R^2, and the p-value from the joint Wald test of the null hypothesis stated in the column header.

The results strongly indicate that the hypotheses of forecast unbiasedness and efficiency are rejected for all models. The bias coefficient a_0 is consistently positive and statistically significant. When testing the joint hypothesis $a_0 = 0$ and $a_1 = 1$, the null is systematically rejected with p-values close to zero. Similarly, for efficiency tests, the coefficient of the lagged forecast error (b_2 or c_2) is always significantly different from zero, and the joint null hypothesis $b_0 = 0$, $b_1 = 1$, $b_2 = 0$ (or $c_0 = 0$, $c_1 = 1$, $c_2 = 0$) is also rejected with p-values close to zero. Moreover, the adjusted R^2 consistently increases when lagged forecast errors are included, further suggesting that LLM Agents form forecasts that do not fully incorporate all available information.

This finding aligns with and helps explain the market inefficiency documented above. LLM Agents produce forecasts that are systematically biased and inefficient, failing to fully exploit available information. As a result, they generate predictable patterns in excess returns, and leave room for arbitrage opportunities.

Dispersion in the Gains from Trade Recall from Section 3.2.3 that the asset pricing model underlying the experiments is a zero-sum environment, where total gains from trade sum exactly to zero in each round. Still, in a REE, there may be dispersion in gains from trade. However, no trader may systematically profit at the expense of others. Here, I examine whether such dispersion systematically arises. To do so, I proceed as follows. For each model–variant–simulation, I compute

[18] Appendix E.2 reports the same regressions estimated separately by model–variant, without the variant dummies. Results are qualitatively unchanged.

Table 5: Forecast Rationality Tests (pooled variants)

Model	Regression Specifications		
	$p_{t+1} = a_0 + a_1 p^e_{t+1}$ $H_0 : a_0 = 0,\ a_1 = 1$	$p_{t+1} = b_0 + b_1 p^e_{t+1} + b_2 \mathrm{FE}_{t-1}$ $H_0 : b_0 = 0,\ b_1 = 1,\ b_2 = 0$	$p_{t+1} = c_0 + c_1 p^e_{t+1} + c_2 \mathrm{PFE}_t$ $H_0 : c_0 = 0,\ c_1 = 1,\ c_2 = 0$
Claude 3.5 Haiku	a_0: 5.221 (1.512) a_1: 0.920 (0.023)	b_0: 0.520 (0.438) b_1: 0.995 (0.007) b_2: 0.754 (0.051)	c_0: -2.654 (0.246) c_1: 1.050 (0.004) c_2: 1.002 (0.013)
	$\bar{R}^2 = 0.938$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.979$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.994$ H_0 *Rej* ($p < 0.0001$)
Llama 3.3 70B	a_0: 5.379 (0.081) a_1: 0.911 (0.001)	b_0: 4.347 (0.290) b_1: 0.928 (0.005) b_2: 0.356 (0.094)	c_0: 7.117 (0.043) c_1: 0.881 (0.001) c_2: -0.777 (0.038)
	$\bar{R}^2 = 0.973$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.973$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.991$ H_0 *Rej* ($p < 0.0001$)
Gemini 2.0 Flash	a_0: 7.429 (0.069) a_1: 0.897 (0.001)	b_0: 6.665 (0.120) b_1: 0.910 (0.002) b_2: 0.223 (0.026)	c_0: 7.373 (0.066) c_1: 0.895 (0.001) c_2: -0.153 (0.011)
	$\bar{R}^2 = 0.986$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.987$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.987$ H_0 *Rej* ($p < 0.0001$)
Grok 2	a_0: 1.714 (2.169) a_1: 0.964 (0.021)	b_0: -2.547 (0.681) b_1: 1.020 (0.007) b_2: 0.731 (0.049)	c_0: -8.375 (0.696) c_1: 1.081 (0.007) c_2: 0.926 (0.019)
	$\bar{R}^2 = 0.963$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.985$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.989$ H_0 *Rej* ($p < 0.0001$)
GPT-4o	a_0: 4.949 (0.065) a_1: 0.902 (0.001)	b_0: 4.328 (0.092) b_1: 0.912 (0.001) b_2: 0.194 (0.022)	c_0: 4.321 (0.054) c_1: 0.906 (0.001) c_2: 0.362 (0.020)
	$\bar{R}^2 = 0.992$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.993$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.993$ H_0 *Rej* ($p < 0.0001$)
DeepSeek Chat	a_0: 5.310 (0.153) a_1: 0.926 (0.004)	b_0: 3.854 (0.169) b_1: 0.951 (0.004) b_2: 0.428 (0.032)	c_0: 4.685 (0.152) c_1: 0.945 (0.004) c_2: 0.435 (0.028)
	$\bar{R}^2 = 0.987$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.991$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.990$ H_0 *Rej* ($p < 0.0001$)

Notes: Pooled OLS coefficients with HAC(1) Newey–West standard errors in parentheses. Each regression includes dummy variables for variant category, which are omitted from the coefficient blocks for brevity. $\bar{R}^2$ denotes the adjusted R^2. Joint tests are implemented as Wald tests, with p-values associated to the χ^2. $\mathrm{FE}_{t-1} = p_{t-1} - p^e_{t-1}$ is the lagged forecast errors, while $\mathrm{PFE}_t = p_{t-1} - p^e_t$ is the *perceived* forecast error.

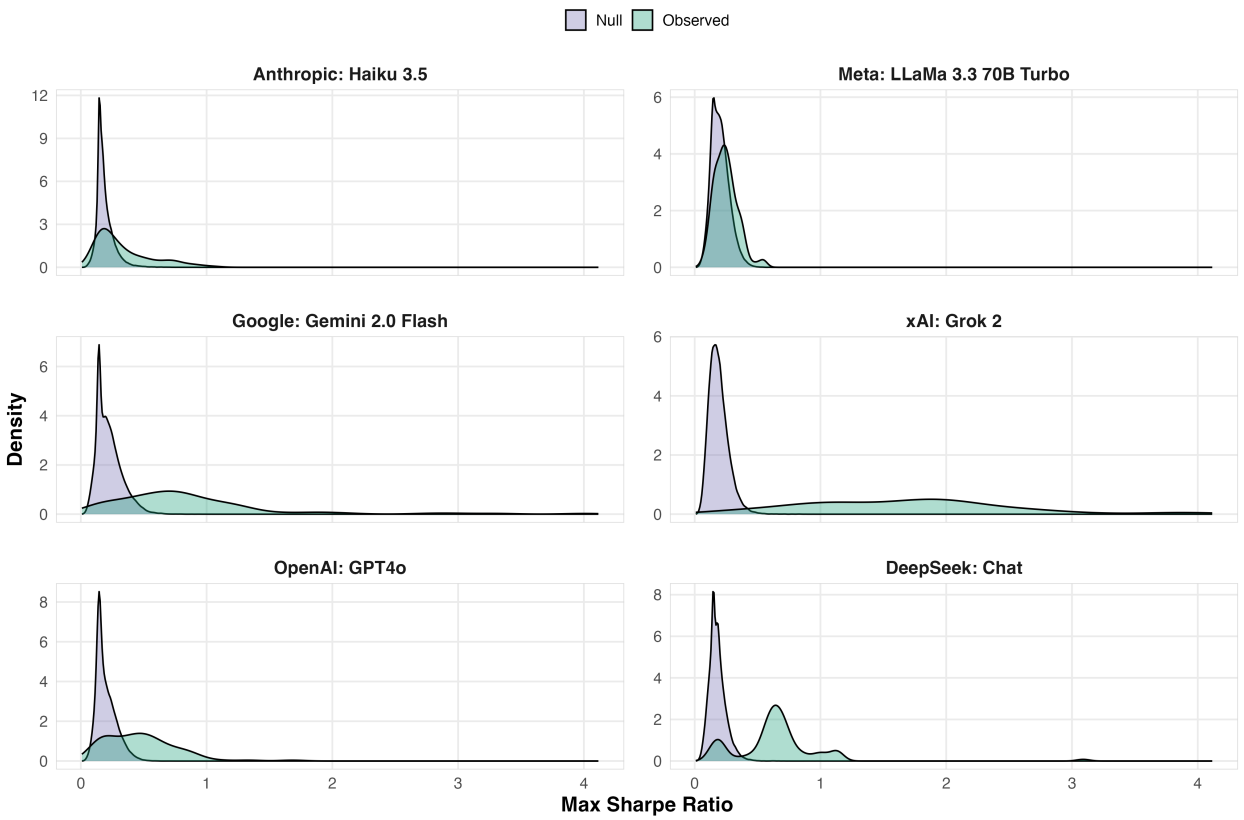


Figure 3: Observed and Null distributions of the maximum Sharpe Ratio per simulation by foundation model, across all simulations for all variants. The Null distribution is obtained by randomly reshuffling the agents gains from trade within each round 1,000 times per simulation.

the Sharpe Ratio (SR) for each LLM Agent i using the gains from trade defined in (10), with:

$$SR_i = \frac{\mathbb{E}[\tilde{W}_i]}{\sqrt{\text{Var}(\tilde{W}_i)}}$$

where $E[\cdot]$ and $\text{Var}(\cdot)$ denote the sample mean and variance over the fifty rounds of the simulation for Agent i. I select the highest SR among the six Agents in that market. I then aggregate these maximum SRs across all simulations with non-zero variance for each model.[19] This procedure yields an observed distribution of maximum SRs with 100 values (one per simulation). I compare this observed distribution to a null distribution obtained by randomly reshuffling the gains from trade across Agents within each round 1,000 times per simulation, producing 100,000 reshuffled values in total. I refer to this as the null distribution under random gains, corresponding to a setting with no systematic winner. The two distributions are shown in Figure 3.

In four out of six models (Gemini, Grok, GPT-4o, and DeepSeek), the observed distribution of

[19]The same exercise is also conducted at the model–variant level. Results are qualitatively unchanged.

maximum SRs seems to first-order stochastically dominates the null distribution, suggesting that the best-performing Agent in these models achieves consistently higher SRs than would be expected under random gains from trade. Next, I perform a formal test. For each model–variant–simulation, I compute the quasi-exact p-value defined as the share of reshuffled maximum SRs exceeding the observed maximum SR. I mark cases where this p-value falls below the significance level, and aggregate the resulting rejections at the model level.

Results are reported in Appendix E.3.1.[20] These rates are consistently above the threshold expected under random gains, equal to α, providing strong evidence that some Agents systematically outperform others in terms of gains from trade, even though LLM Agents originate from the same foundation model and share identical ex ante information.[21]

5.2 Hypers Setup

So far, I have shown that LLM learners generate substantially more stable markets than human learners in an equivalent economic environment, producing prices that closely track fundamentals, although without fully achieving rational behavior or convergence to a REE. In this section, I test whether this stabilizing potential is robust to exogenous shocks intentionally designed to trigger a bubble. In particular, I introduce the hypers' shock described in Section 3.3, where $N_h = 5$ "hypers" exogenously enter the market at period $T_1 = 12$ and stay until period $T_2 = 21$, mechanically forming their expectations according to the rule (12) with $\gamma = 1.3$. Everything else remains unchanged from the baseline setup.

Figure 4 displays the average price dynamics across all simulations for variant 1 in Table 3, separately for each foundation model[22]. The resulting patterns are heterogeneous across models. DeepSeek Chat, Gemini 2.0 Flash, and Grok 2 exhibit dynamics resembling adaptive rules: prices rise during the hypers shock but revert immediately after the hypers exit, converging monotonically back to the fundamental value of 60. At the opposite extreme, Haiku 3.5 and LLaMA 3.3 70B Turbo amplify the bubble; in the former, prices diverge exponentially, while in the latter, they eventually mean-revert but after a large overshooting. Finally, GPT-4o displays a distinct pattern, with prices remaining anchored around the highest pre-exit level.

[20] The section also reports rejection rates by variant. Results are qualitatively unchanged.

[21] To assess whether these "winning" Agents correspond to those whose initial forecasts were closest to the realized price, I regressed an indicator for being the trader with the maximum Sharpe Ratio in a given simulation on an indicator for having the minimum forecast error in round 1. The estimated coefficient is 0.1607 (s.e. 0.0066). In Appendix E.3.2, I further explore whether forecast coordination improves over time, so to reduce dispersion in gains from trade, and find that it does so only after the first market feedback, but then significantly deteriorates in most cases.

[22] Average price dynamics across simulations for variants 2, 3 and 4 in Table 3 are reported in Appendix D.1.

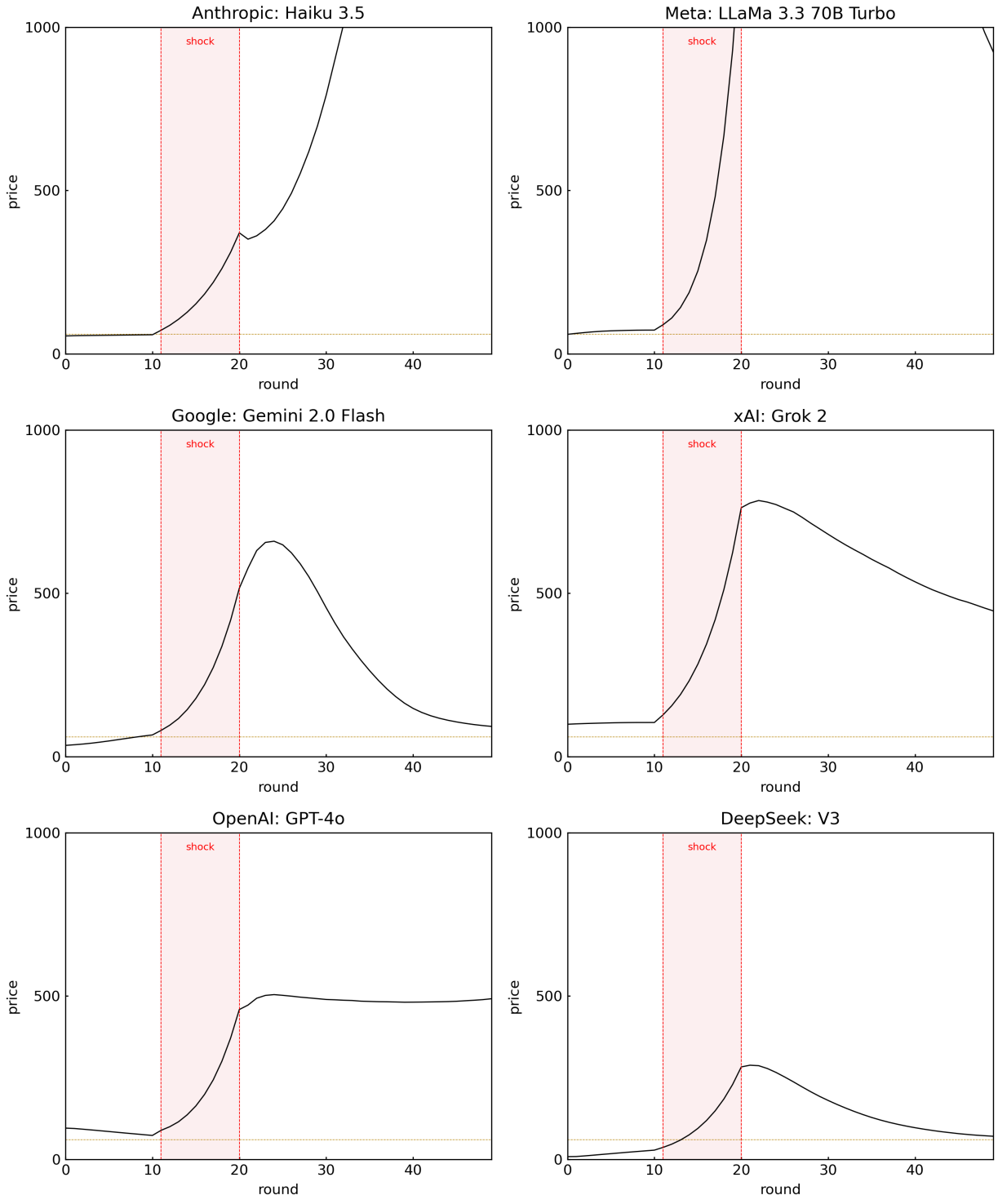


Figure 4: Asset price dynamics with six LLM Agents for each foundation model, and five hypers active between $T_1 = 12$ and $T_2 = 21$, forming expectations according to equation (12) with $\gamma = 1.3$. The red area indicates the period of the hypers shock, when hypers are present in the market. The orange dashed line denotes the fundamental asset price, equal to 60. Displayed prices correspond to averages across identical simulations for variant 1 in Table 3. The scale of the y-axis follows that of Figure 1 and 2 for comparability.

5.2.1 Heterogeneous Dynamic Effects of the Hypers' Shock

Running the simulations under both the baseline and hypers setups allows for causal identification of dynamic and heterogeneous treatment effects. Indeed, each API call can be viewed as an independent draw of an AI Agent that is randomly assigned to an experimental condition. The main treatment status is assignment to the hypers setup (setup 2), where treatment begins at round 11. In addition, each Agent is randomly assigned to one of the four variants in Table 3, providing an exogenous source of heterogeneity that allows for causal interpretation of differences across configurations.

For each foundation model m and variant v, I estimate dynamic treatment effects using a two-way fixed effects (TWFE) event–study specification:

$$p_{i,t+1}^{e\,(m,v)} = \alpha_i + \lambda_t + \sum_{k=-K}^{L} \beta_k^{(m,v)}\, D_{t=T_1+k}^{(i)} + \varepsilon_{i,t}, \tag{14}$$

where $p_{i,t+1}^{e,(m,v)}$ denotes the forecast of Agent i of model m and variant v at round t for the price in round t+1; α_i and λ_t are Agent and round fixed effects, respectively; $D_{t=T_1+k}^{(i)}$ is an indicator equal to 1 for treated Agents in setup 2 at event time k relative to the hypers' entry ($T_1 = 12$). Note that for each (m, v) combination, the sample size is $N = 6$ Agents per simulation, times the number of identical simulations, or repetitions, as per Table 3. However, since the hypers shock constitutes a simulation-level treatment, implying within-simulation dependence in the error structure, standard errors are clustered at the simulation level.

In the absence of pre-trends, which is expected given the randomized design, the coefficients $\beta_k^{(m,v)}$ can be interpreted as (m, v)-specific dynamic treatment effects of the hypers shock, normalized so that they are measured relative to the last pre-treatment round. I am particularly interested in how these treatment effects vary across foundation models and variants. Given the random assignment of the treatment and variant, and, de facto, of the foundation model, I can assign a causal interpretation to this heterogeneity.

Figures 5 and 6 display the estimated dynamic treatment effects of the hypers shock on Agents' forecasts by foundation model, respectively highlighting heterogeneity by prompt and temperature. These estimates reveal several important patterns. First, the dynamic effects on forecasts confirms the market price patterns discussed above. Haiku 3.5 exhibits a large and persistent positive forecast response to the hypers shock, consistent with explosive price dynamics. LLaMA 3.3 70B Turbo shows a similarly strong initial response but gradually returns to pre-shock forecast levels. Gemini 2.0 Flash and DeepSeek Chat display rapid stabilization, with forecast effects converging to nearly zero shortly after the hypers exit. Grok 2 exhibits a mild but transitory response, while GPT-4o remains largely anchored throughout. Second, both temperature and prompt design play an important role in shaping the stabilizing potential of AI Agents, albeit with differences across models. The Caution

prompt variant, intended to make Agents explicitly aware of possible temporary market turbulences, significantly dampens the impact of the hypers shock on forecasts for most models. In some cases, the reduction is substantial (for instance, the effect for LLaMA is nearly halved), with the only exception being Haiku 3.5, for which the prompt adjustment has no significant effect. Temperature levels behave largely as expected: higher temperatures generally produce larger destabilizing effects. However, the relationship is not linear, that is lower temperatures do not necessarily enhance stability compared to medium levels, and for some models (notably Gemini, DeepSeek, and Haiku) temperature appears to have little impact in general.

In summary, the hypers shock analysis confirms the stabilizing properties of LLM Agents, though these are mediated by both model architecture and configuration. In particular, prompt design and temperature critically shape the extent and speed of stabilization, while different foundation models display markedly distinct responses, and distinct marginal sensitivities to configuration changes, when exposed to the same exogenous disturbance. This suggests that the stabilizing potential of LLM Agents may be partly intrinsic to the underlying model, yet also contingent on how the model is deployed.

5.2.2 "Hype Susceptibility" Parameter

To provide a unified quantitative measure of how different LLM architectures respond to exogenous bullish shocks, I structurally estimate a parameter denoted as *hype susceptibility*. This parameter captures the degree to which LLM Agents' forecasts are influenced by the Hypers' expectations. It serves as an indicator of each model's stabilizing potential or, conversely, its propensity to be destabilized away from fundamental valuation.

I model the LLM Agent expectations as follows (omitting time subscripts for notational convenience):

$$p^e = b + (1 - \delta)p^f + \delta p^h + u \tag{15}$$

where u is an error component, b is a bias component, p^h is the Hypers' expectation, and p^f is the fundamental price. This specification treats the LLM Agents' forecasts as a convex combination of the fundamental price and the Hypers' expectations, plus a bias and an error term. The parameter δ, which I refer to as *hype susceptibility*, captures the relative weight assigned to the Hypers. It summarizes the degree to which Agents' expectations are pulled away from fundamentals by the Hypers' influence.

Let $\alpha_h = \frac{N_h}{N+N_h}$ be the share of Hypers. The aggregate market expectation is:

$$\bar{p}^e = (1 - \alpha_h)b + \Big((1 - \alpha_h)(1 - \delta)\Big)p^f + \Big(\delta + \alpha_h(1 - \delta)\Big)p^h + \epsilon \tag{16}$$

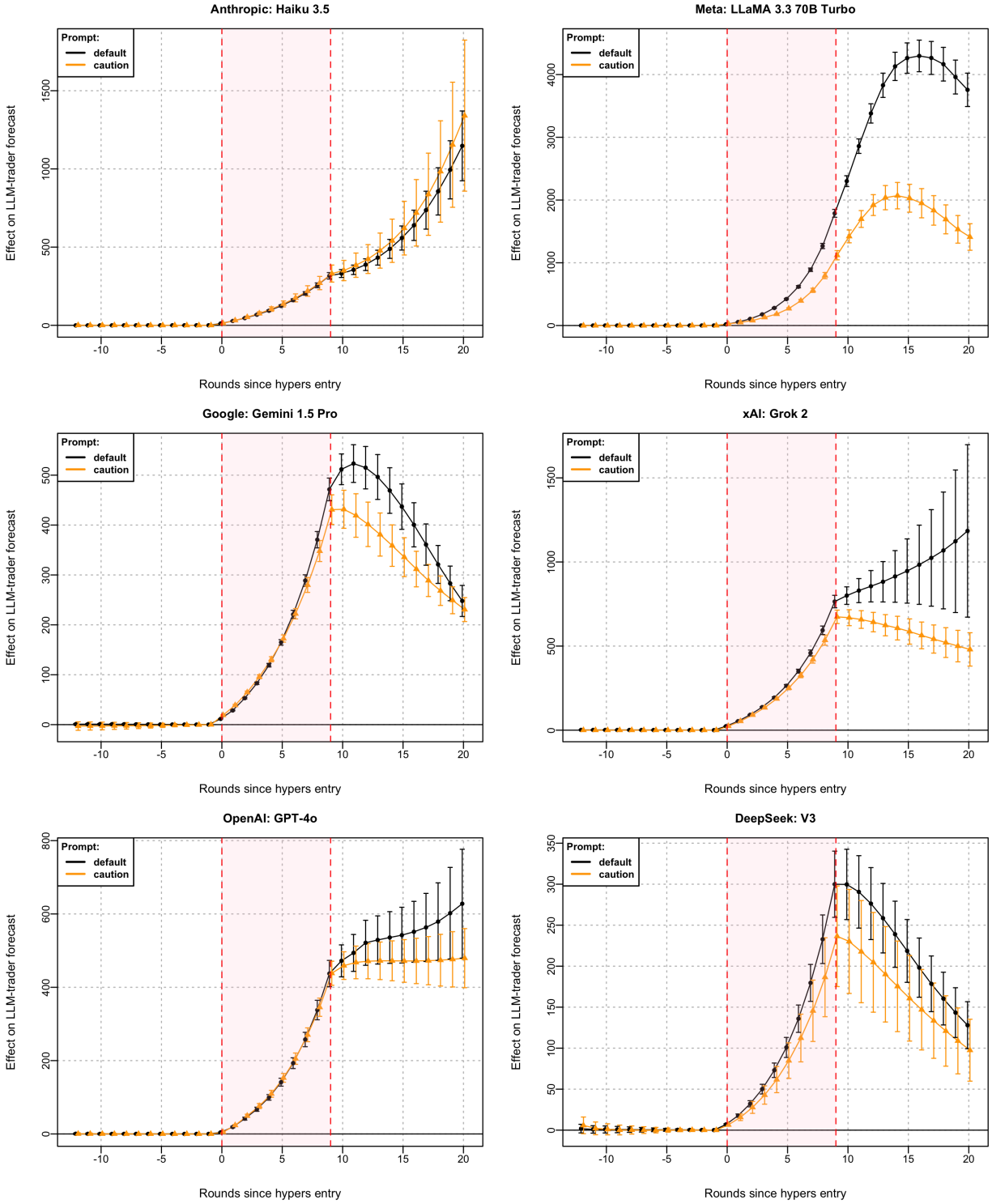


Figure 5: Estimated dynamic treatment effects of the hypers shock on Agents' forecasts, by foundation model and variant (variants 1 and 4 in Table 3). Estimates are shown with 95% confidence intervals based on standard errors clustered at the simulation level.

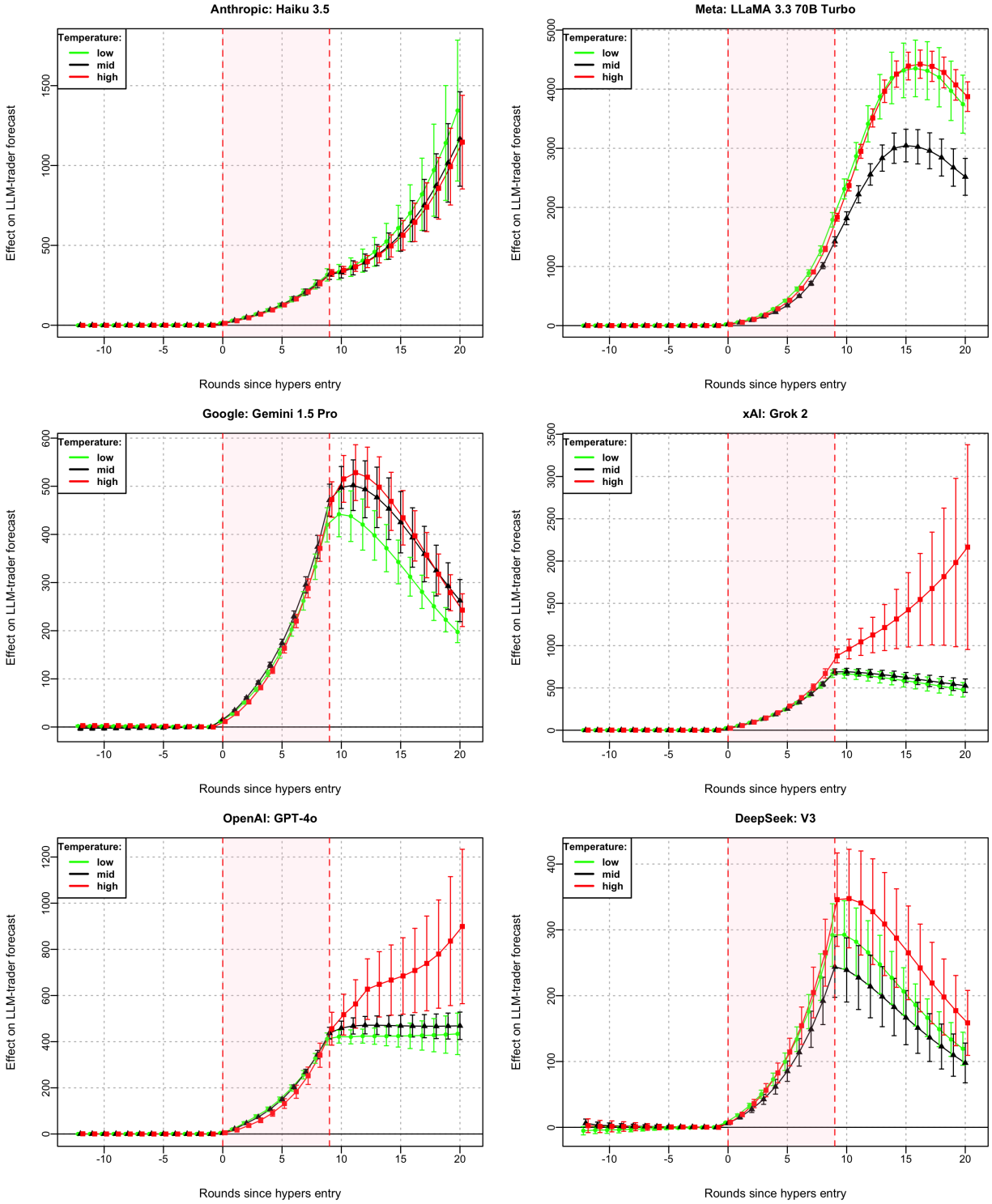


Figure 6: Estimated dynamic treatment effects of the hypers shock on Agents' forecasts, by foundation model and variant (variants 1, 2 and 3 in Table 3). Estimates are shown with 95% confidence intervals based on standard errors clustered at the simulation level.

where $\epsilon = (1 - \alpha_h)\bar{u}$, with $\bar{u}$ representing the cross-sectional average of the agents' errors, while the bias term b is assumed to be constant across agents.

Recall the market clearing condition $p = \frac{\bar{y}+\bar{p}^e}{R}$. Substituting (16) into the clearing condition and defining $\Theta = \delta + \alpha_h(1 - \delta)$, we obtain:

$$Rp - \bar{y} = (1 - \alpha_h)b + (1 - \Theta)p^f + \Theta p^h + \epsilon \tag{17}$$

Next, consider the following regression specification:

$$p_t = \beta_0 + \beta_1 p_{t-1} + \eta_t \tag{18}$$

From (17) and the Hypers' rule ($p_t^h = \gamma R^2 p_{t-1}$), it follows that:

$$\beta_0 = \frac{(1 - \alpha_h)b + (1 - \Theta)p^f + \bar{y}}{R}$$

$$\beta_1 = \Theta\gamma R$$

This yields two moment conditions with two unknown parameters, $\{\delta, b\}$, since all other parameters $\{\alpha_h, \bar{y}, R, \gamma\}$ are known. Consequently, I can estimate $\{\delta, b\}$ using the method of moments.[23]

The estimation is performed separately for each model and for two distinct time intervals: during the shock (when Hypers are present) and after the shock. In the latter case, $\alpha_h = 0$, but the structural assumption (15) is maintained.

A value of $\hat{\delta}$ close to zero indicates that the corresponding model remains anchored to fundamentals, displaying strong stabilizing properties. Conversely, values approaching one identify models whose forecasts are highly influenced by Hypers, implying a strong propagation of exogenous optimism shocks that leads to explosive price dynamics, as also discussed in Section 3.3. Comparing $\hat{\delta}$ allows for a ranking of foundation models based on their "resilience" to hypers-driven destabilization.

The estimates for δ are reported in Table 6.[24]

The results broadly confirm previous evidence regarding the stabilizing potential and destabilizing tendencies of different foundation models, summarizing them in a single metric that clearly distinguishes between the hypers shock and post-shock phases. The most stable model is DeepSeek, followed by Gemini 2.0 Flash. For both, the estimated hype susceptibility parameter is relatively high during the hypers shock, indicating that their forecasts are temporarily influenced by the bullish pressure, but drops sharply after the hypers exit, suggesting that this influence is short-lived. This pattern does not hold for GPT-4o, whose hype parameter remains nearly unchanged after the shock,

[23] In practice, I estimate regression (18) using OLS, pooling all simulations for each model, and then apply the delta method to recover parameter estimates and standard errors.

[24] Variant-specific estimates of the hype susceptibility parameters are reported in Appendix F.

Provider	Model	Rounds	δ
Anthropic	Haiku 3.5	[12–22]	**0.56**
			(0.01)
		[23–50]	**0.80**
			(0.00)
Meta	LLaMa 3.3 70B Turbo	[12–22]	**0.99**
			(0.00)
		[23–50]	**0.51**
			(0.00)
Google	Gemini 2.0 Flash	[12–22]	**0.70**
			(0.01)
		[23–50]	**0.48**
			(0.00)
xAI	Grok 2	[12–22]	**0.66**
			(0.01)
		[23–50]	**0.73**
			(0.00)
OpenAI	GPT4o	[12–22]	**0.69**
			(0.01)
		[23–50]	**0.66**
			(0.01)
DeepSeek	Chat	[12–22]	**0.70**
			(0.01)
		[23–50]	**0.46**
			(0.00)

Table 6: Estimates of the Hype Susceptibility Parameters. The column *Rounds* indicates the sample period used for each estimation. Data from all simulations for all the variants are pooled together.

consistent with the price dynamics and forecast effects discussed above. At the opposite extreme, Haiku 3.5 exhibits a surprisingly low hype susceptibility during the shock, but the parameter increases substantially afterward, in line with the post-shock price explosion documented earlier. Finally, LLaMA 3.3 70B Turbo displays a hype susceptibility close to one during the hypers shock, implying that its forecasts almost perfectly follow the hypers' expectations and thereby generate a rapid exponential increase in prices consistent with Section 3.3. However, this influence diminishes abruptly after the hypers exit, in line with the eventual mean reversion observed in the price dynamics.

6 Conclusion

This paper examined the implications of adopting Generative Artificial Intelligence (GenAI) as an expectation-formation technology for market stability. Building on the Learning-to-Forecast Experiment tradition, I developed a framework in which human forecasters are replaced with AI Agents, enabling three complementary analyses: (i) testing whether AI Agents behave as fully rational agents, as boundedly rational human-like learners, or as adaptive stabilizers; (ii) causally estimating how model architecture and configuration affect their stabilizing properties; and (iii) structurally estimating a parameter, the *hype susceptibility*, that quantifies how prone a model is to deviate from fundamentals under exogenous bullish shocks.

The framework is applied to Agents instantiated from six multimodal LLMs from different leading providers: GPT-4o, DeepSeek Chat, Gemini 2.0 Flash, Grok 2, Haiku 3.5, and LLaMA 3.3 70B Turbo. The results indicate that markets populated by GenAI learners are substantially more stable than those formed by human learners in equivalent experimental environments. Prices remain close to fundamentals and bubble episodes do not emerge, indicating that AI-driven expectations can exert a stabilizing influence. However, this stability has a boundedly rational nature, in that AI Agents do not achieve a REE, their forecasts remain biased and inefficient, excess returns exhibit serial dependence, and systematic dispersion in gains from trade persists even among Agents derived from the same foundation model. When exposed to an exogenous bullish disturbance (the hypers shock), LLM Agents display heterogeneous dynamic responses across models and configurations. Prompt design and temperature mediates these dynamics: an explicit cautionary instruction significantly reduces the responsiveness of forecasts to hypers, whereas higher temperatures tend to magnify it. These results suggest that the stabilizing potential of GenAI Agents is not only intrinsic to the model architecture, but also contingent on how the model is configured and deployed. The structural estimation of the *hype susceptibility* parameter consolidates these findings into a single interpretable metric, capturing the extent to which forecasts are influenced by hypers relative to fundamentals. Overall, the evidence indicates that GenAI Agents can serve as boundedly-rational, adaptive stabilizers, capable of anchoring prices near fundamentals. Their stabilizing properties,

however, are both *model-dependent* and *use-dependent*.

The proposed framework offers a general and extensible approach to studying how AI expectations can shape market dynamics. It can be readily extended along several dimensions. First, it accommodates markets featuring heterogeneous GenAI models, including newer generations of LLMs, reasoning-oriented models, and foundation models specialized in time-series prediction, and it easily allows benchmarking against standard econometric forecasting tools. Second, it can incorporate endogenous adoption dynamics, in which agents select GenAI models as forecasting tools based on their relative past performance, thereby allowing the study of convergence patterns in forecasting technologies. Finally, the framework can be adapted to environments with multimodal information shocks, where the hypers act, in practice, not only through (numerical) pressures on market prices but also through alternative modalities, such as social media textual data, possibly affecting the expectation formation of multimodal models through other channels.

References

Akata, Ece, Laura Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz (2025). "Playing repeated games with large language models". In: *Nature Human Behaviour*, pp. 1–11.

Anufriev, Marc and Cars Hommes (2012). "Evolutionary selection of individual expectations and aggregate outcomes in asset pricing experiments". In: *American Economic Journal: Microeconomics* 4.4, pp. 35–64. DOI: 10.1257/mic.4.4.35.

Arifovic, Jasmina (1994). "Genetic Algorithm Learning and the Cobweb Model". In: *Journal of Economic Dynamics and Control* 18.1, pp. 3–28. DOI: https://doi.org/10.1016/0165-1889(94)90067-1.

Arifovic, Jasmina (1995). "Genetic Algorithms and Inflationary Economies". In: *Journal of Monetary Economics* 36.1, pp. 219–243. DOI: https://doi.org/10.1016/0304-3932(95)01203-7.

Ashwin, J., Paul Beaudry, and Martin Ellison (2025). "Neural Network Learning for Nonlinear Economies". In: *Journal of Monetary Economics* 149, p. 103723.

Benhabib, Jess and Roger E. A. Farmer (1999). "Indeterminacy and Sunspots in Macroeconomics". In: *Handbook of Macroeconomics*. Ed. by John B. Taylor and Michael Woodford. Vol. 1. Elsevier, pp. 387–448.

Berti, Lorenzo, Federico Giorgi, and Gjergji Kasneci (2025). *Emergent Abilities in Large Language Models: A Survey*. arXiv: 2503.05788 [cs.CL]. URL: https://arxiv.org/abs/2503.05788.

Biswas, Siddhartha, Andrew Hanson, and Toan Phan (Oct. 2020). "Bubbly Recessions". en. In: *American Economic Journal: Macroeconomics* 12.4, pp. 33–70. ISSN: 1945-7707, 1945-7715. DOI: 10.1257/mac.20180083. URL: https://pubs.aeaweb.org/doi/10.1257/mac.20180083 (visited on 10/03/2025).

Blanchard, Olivier J. (1979). "Speculative Bubbles, Crashes and Rational Expectations". In: *Economics Letters* 3.4, pp. 387–389.

Brock, William A. and Cars H. Hommes (1997). "A Rational Route to Randomness". In: *Econometrica* 65.5, pp. 1059–1095. DOI: 10.2307/2171879.

Brynjolfsson, Erik, Anton Korinek, and Ajay K. Agrawal (Sept. 2025). *A Research Agenda for the Economics of Transformative AI*. Working Paper w34256. National Bureau of Economic Research. DOI: 10.3386/w34256. URL: https://www.nber.org/papers/w34256.

Burton, J. W., E. Lopez-Lopez, S. Hechtlinger, et al. (2024). "How large language models can reshape collective intelligence". In: *Nature Human Behaviour* 8, pp. 1643–1655. DOI: 10.1038/s41562-024-01959-9.

Bybee, Jay L. (2023). "The ghost in the machine: Generating beliefs with large language models". In: *arXiv preprint arXiv:2305.02823*, pp. 349–389.

Carriero, A., D. Pettenuzzo, and S. Shekhar (2024). "Macroeconomic Forecasting with Large Language Models". In: *arXiv*. arXiv: 2407.00890 [econ.EM]. URL: https://arxiv.org/abs/2407.00890.

Chen, Huaben, Wenkang Ji, Lufeng Xu, and Shiyu Zhao (2023). *Multi-Agent Consensus Seeking via Large Language Models*. DOI: 10.48550/arXiv.2310.20151. arXiv: 2310.20151 [cs.MA]. URL: https://arxiv.org/abs/2310.20151.

Cho, In-Koo and Thomas J. Sargent (2008). "Self-Confirming Equilibria". In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan, pp. 1–5.

Conitzer, Vincent and Christian Oesterheld (Sept. 2023). "Foundations of Cooperative AI". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 13, pp. 15359–15367.

Di Francesco, Tommaso and Cars Hommes (2025). "Sentiment-driven speculation in financial markets with heterogeneous beliefs: A machine learning approach". In: *Journal of Economic Dynamics and Control* 175, p. 105092. DOI: 10.1016/j.jedc.2024.105092.

Evans, George W. and Seppo Honkapohja (2001). *Learning and Expectations in Macroeconomics*. Frontiers of Economic Research. Hardcover, $104.00/£88.00. Princeton, NJ: Princeton University Press, p. 424. ISBN: 9780691049212.

Evans, George W. and Seppo Honkapohja (2005). "An Interview with Thomas J. Sargent". In: *Macroeconomic Dynamics* 9.4, pp. 561–583. DOI: 10.1017/S1365100505050042.

Evans, George W. and Seppo Honkapohja (2009). "Learning and Macroeconomics". In: *Annual Review of Economics* 1.1, pp. 421–449. DOI: 10.1146/annurev.economics.050708.142838.

Fama, Eugene F. (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2, pp. 383–417. DOI: 10.2307/2325486.

Fan, C., J. Chen, Y. Jin, and H. He (Mar. 2024). "Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16, pp. 17685–17693. DOI: 10.1609/aaai.v38i16.29751. URL: https://doi.org/10.1609/aaai.v38i16.29751.

Faria-e-Castro, Miguel and Fernando Leibovici (2024). "Artificial Intelligence and Inflation Forecasts". In: *Federal Reserve Bank of St. Louis Review* 106.12, pp. 1–14. DOI: 10.20955/r.2024.12. URL: https://doi.org/10.20955/r.2024.12.

Fatemi, Sasan and Yangjun Hu (Nov. 2024). "FinVision: A Multi-Agent Framework for Stock Market Prediction". In: *Proceedings of the 5th ACM International Conference on AI in Finance*. ACM, pp. 582–590.

Filippas, Apostolos, John J. Horton, and Blake S. Manning (July 2024). "Large language models as simulated economic agents: What can we learn from Homo Silicus?" In: *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 614–615.

Fudenberg, Drew and David K. Levine (2009). "Self-Confirming Equilibrium and the Lucas Critique". In: *Journal of Economic Theory* 144.6, pp. 2354–2371.

Google Cloud (Apr. 2025). *101 Real-World Generative AI Use Cases from Industry Leaders*. Google Cloud Blog. Updated on April 9, 2025, originally published for Google Cloud Next '24. URL: https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders.

Grandmont, Jean-Michel (1998). "Expectations Formation and Stability of Large Socioeconomic Systems". In: *Econometrica* 66.4, pp. 741–781.

Gruver, Noah, Marc Finzi, Song Qiu, and Andrew G. Wilson (2023). "Large language models are zero-shot time series forecasters". In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 19622–19635.

Guerron-Quintana, Pablo A., Tomohiro Hirano, and Ryo Jinnai (Apr. 2023). "Bubbles, Crashes, and Economic Growth: Theory and Evidence". en. In: *American Economic Journal: Macroeconomics* 15.2, pp. 333–371. ISSN: 1945-7707, 1945-7715. DOI: 10.1257/mac.20220015. URL: https://pubs.aeaweb.org/doi/10.1257/mac.20220015 (visited on 10/03/2025).

Guo, Di, Dongfu Yang, Hongwei Zhang, Jun Song, Ruitong Zhang, Ruyi Xu, and Yujia He (2025). "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning". In: *arXiv*. arXiv: 2501.12948 [cs.AI].

Hadfield, Gillian K. and Andrew Koh (2025). "An Economy of AI Agents". In: *arXiv*. arXiv: 2509.01063 [econ.GN].

Hansen, Anders L. and Seung Jung Lee (2025). "Financial Stability Implications of Generative AI: Taming the Animal Spirits". In: *arXiv preprint arXiv:2510.01451*.

Hansen, Anne Lundgaard, John J. Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar (Dec. 2024). *Simulating the Survey of Professional Forecasters*. Working Paper 5066286. Available at https://ssrn.com/abstract=5066286 or http://dx.doi.org/10.2139/ssrn.5066286. SSRN. DOI: 10.2139/ssrn.5066286. URL: https://ssrn.com/abstract=5066286.

Heemeijer, Peter, Cars Hommes, Joep Sonnemans, and Jan Tuinstra (2009). "Price Stability and Volatility in Markets with Positive and Negative Expectations Feedback: An Experimental Investigation". In: *Journal of Economic Dynamics and Control* 33.5, pp. 1052–1072. DOI: 10.1016/j.jedc.2008.09.010.

Henning, Thomas, Siddhartha M. Ojha, Ross Spoon, Jiatong Han, and Colin F. Camerer (2025). *LLM Trading: Analysis of LLM Agent Behavior in Experimental Asset Markets*. arXiv: 2502.15800 [q-fin.TR]. URL: https://doi.org/10.48550/arXiv.2502.15800.

Holland, John H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press, p. 32.

Hommes, Cars (2011). "The heterogeneous expectations hypothesis: Some evidence from the lab". In: *Journal of Economic Dynamics and Control* 35.1, pp. 1–24.

Hommes, Cars (2021). "Behavioral and experimental macroeconomics and policy analysis: A complex systems approach". In: *Journal of Economic Literature* 59.1, pp. 149–219.

Hommes, Cars, Joep Sonnemans, Jan Tuinstra, and Henk Van de Velden (2008). "Expectations and bubbles in asset pricing experiments". In: *Journal of Economic Behavior & Organization* 67.1, pp. 116–133.

Hommes, Cars H. (2017). "Booms, Busts and Behavioural Heterogeneity in Stock Prices". In: *Journal of Economic Dynamics and Control* 80, pp. 101–124. DOI: 10.1016/j.jedc.2017.03.006.

Hosking, Tom, Phil Blunsom, and Max Bartolo (2023). "Human Feedback is Not Gold Standard". In: *arXiv preprint arXiv:2309.16349*.

Jia, Feng, Kun Wang, Yuyang Zheng, Dong Cao, and Yuyu Liu (Mar. 2024). "GPT4MTS: Prompt-based large language model for multimodal time-series forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 21, pp. 23343–23351.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor (Dec. 2015). "Leveraged bubbles". en. In: *Journal of Monetary Economics* 76, S1–S20. ISSN: 03043932. DOI: 10.1016/j.jmoneco.2015.08.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304393215000987 (visited on 10/03/2025).

Karabag, Mert Onur and Ufuk Topcu (2025). "Do LLMs Strategically Reveal, Conceal, and Infer Information? A Theoretical and Empirical Analysis in The Chameleon Game". In: *arXiv preprint arXiv:2501.19398*.

Keane, Michael P and David E Runkle (Dec. 1990). "Testing the Rationality of Price Forecasts: New Evidence from Panel Data". en. In: *American Economic Review* 80.4, pp. 714–735. URL: https://ideas.repec.org/a/aea/aecrev/v80y1990i4p714-35.html.

Kim, J., M. Kovach, K. M. Lee, E. Shin, and H. Tzavellas (2024). "Learning to be Homo Economicus: Can an LLM Learn Preferences from Choice". In: *arXiv preprint arXiv:2401.07345*.

Kirk, Hannah Rose, Alice Whitefield, Paul Röttger, Anna M. Bean, Katerina Margatina, Raquel Mosquera-Gomez, and Scott Hale (2024). "The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals about the Subjective and Multicultural Alignment of Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 37, pp. 105236–105344.

Kitadai, Akihiro, Sergio Daniel Rico Lugo, Yusuke Tsurusaki, Yusuke Fukasawa, and Norihito Nishino (2025). "Can AI with high reasoning ability replicate human-like decision making in economic experiments?" In: *Group Decision and Negotiation*, pp. 1–24.

Kiyotaki, Nobuhiro and Randall Wright (1989). "On Money as a Medium of Exchange". In: *Journal of Political Economy* 97.4, pp. 927–954.

Li, Haoran, Qian Dong, Jiaxin Chen, Hang Su, Yilun Zhou, Qingyang Ai, and Yang Liu (2024). "LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods". In: *arXiv preprint arXiv:2412.05579*.

Lin, Zihao et al. (2025). "A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models". In: *arXiv preprint arXiv:2502.17516*. URL: https://doi.org/10.48550/arXiv.2502.17516.

Liu, Xiao-Yang, Guoxuan Wang, Hao Yang, and Daochen Zha (2023). "FinGPT: Democratizing Internet-Scale Data for Financial Large Language Models". In: *arXiv preprint arXiv:2307.10485*.

Lopez-Lira, A., J. Kwon, S. Yoon, J. Y. Sohn, and C. Choi (2025). *Bridging Language Models and Financial Analysis*. arXiv: 2503.22693 [q-fin.ML].

Lopez-Lira, A., Y. Tang, and M. Zhu (2025). *The Memorization Problem: Can We Trust LLMs' Economic Forecasts?* arXiv: 2504.14765 [cs.LG].

Lopez-Lira, Alejandro (2025). *Can Large Language Models Trade? Testing Financial Theories with LLM Agents in Market Simulations*. arXiv: 2504.10789 [q-fin.TR].

Lopez-Lira, Alejandro and Yuehua Tang (2023). *Can ChatGPT forecast stock price movements? Return predictability and large language models*. arXiv: 2304.07619 [q-fin.ML].

Lorè, Nicola and Babak Heydari (2024). "Strategic behavior of large language models and the role of game structure versus contextual framing". In: *Scientific Reports* 14.1, p. 18490.

Lorenzoni, Guido (2008). "Inefficient Credit Booms". en. In: *The Review of Economic Studies* 75.3, pp. 809–833. DOI: 10.1111/j.1467-937X.2008.00507.x. URL: https://econpapers.repec.org/RePEc:oup:restud:v:75:y:2008:i:3:p:809-833.

Lovell, Michael C. (1986). "Tests of the Rational Expectations Hypothesis". In: *The American Economic Review* 76.1, pp. 110–124. URL: https://www.jstor.org/stable/1804127.

Lucas, Robert E. Jr. (1972). "Expectations and the Neutrality of Money". In: *Journal of Economic Theory* 4.2, pp. 103–124. DOI: 10.1016/0022-0531(72)90142-1.

Lucas, Robert E. Jr. (1976). "Econometric Policy Evaluation: A Critique". In: *Carnegie-Rochester Conference Series on Public Policy*. Vol. 1. Amsterdam: North-Holland, pp. 19–46.

Ma, Bolei, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Assenmacher (2024). "Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study". In: *arXiv preprint arXiv:2412.13169*.

Marimon, Ramon (1996). “Learning from Learning in Economics”. In: *Advances in Economics and Econometrics: Theory and Applications — Seventh World Congress*. Ed. by David M. Kreps and Kenneth F. Wallis. Cambridge University Press, pp. 278–315.

Marimon, Ramon, Ellen McGrattan, and Thomas Sargent (1990). “Money as a Medium of Exchange in an Economy with Artificially Intelligent Agents”. In: *Journal of Economic Dynamics and Control* 14, pp. 329–373.

Marimon, Ramon, Stephen E. Spear, and Shyam Sunder (1993). “Expectationally driven market volatility: an experimental study”. In: *Journal of Economic Theory* 61.1, pp. 74–103. DOI: 10.1006/jeth.1993.1059.

Marimon, Ramon and Shyam Sunder (1993). “Indeterminacy of equilibria in a hyperinflationary world: Experimental evidence”. In: *Econometrica* 61.5, pp. 1073–1107. DOI: 10.2307/2951494.

Marimon, Ramon and Shyam Sunder (1994). “Expectations and learning under alternative monetary regimes: an experimental approach”. In: *Economic Theory* 4.1, pp. 131–162. DOI: 10.1007/BF01211121.

Martin, Alberto and Jaume Ventura (2018). “The Macroeconomics of Rational Bubbles: A User’s Guide”. en. In: *Annual Review of Economics* 10, pp. 505–539. DOI: 10.1146/annurev-economics-080217-053534. URL: https://crei.cat/wp-content/uploads/2019/01/marabu-pub-1.pdf.

Mienye, Ibomoiye Domor, Nobert Jere, George Obaido, Oyindamola Omolara Ogunruku, Ebenezer Esenogho, and Cameron Modisane (Sept. 2025). “Large language models: an overview of foundational architectures, recent trends, and a new taxonomy”. In: *Discover Applied Sciences* 7.1027. Open access. DOI: 10.1007/s42452-025-07668-w.

Miller, Joshua A., Muhannad Aldosari, Faisal Saeed, Nicholas H. Barna, Shakil Rana, I. Budak Arpinar, and Ning Liu (2024). “A survey of deep learning and foundation models for time series forecasting”. In: *arXiv*. arXiv: 2401.13912 [cs.LG].

Mincer, Jacob and Victor Zarnowitz (1969). “The Evaluation of Economic Forecasts”. en. In: *Economic forecasts and expectations: analyses of forecasting behavior and performance*. Ed. by Jacob Mincer and Victor Zarnowitz. Studies in business cycles 19. New York: National Bureau of Economic Research [u.a.], pp. 3–46. ISBN: 978-0-87014-202-4. URL: https://www.nber.org/chapters/c1214.

Moniri, Benyamin, Hamed Hassani, and Edgar Dobriban (2024). “Evaluating the Performance of Large Language Models via Debates”. In: *arXiv preprint arXiv:2406.11044*.

Mullainathan, Sendhil (2025). “Economics in the Age of Algorithms”. In: *AEA Papers and Proceedings* 115, pp. 1–23. DOI: 10.1257/pandp.20251118. URL: https://www.aeaweb.org/articles?id=10.1257/pandp.20251118.

Muth, John F. (1961). "Rational Expectations and the Theory of Price Movements". In: *Econometrica* 29.3, pp. 315–335.

Plaat, Aske, Marcel van Duijn, Bas van Stein, Mike Preuss, Peter van der Putten, and K. Joost Batenburg (2025). "Agentic Large Language Models, A Survey". In: *arXiv preprint arXiv:2503.23037*. URL: https://arxiv.org/abs/2503.23037.

Polachek, Solomon, Kyle Romano, and Onur Tonguc (2024). *Homo-Silicus: Not (Yet) a Good Imitator of Homo Sapiens or Homo Economicus*. Tech. rep. 17521. Institute of Labor Economics (IZA).

Porter, David P. and Vernon L. Smith (2003). "Stock market bubbles in the laboratory". In: *The Journal of Behavioral Finance* 4.1, pp. 7–20.

Renze, Michael and Ege Guven (2024). "Self-Reflection in LLM Agents: Effects on Problem-Solving Performance". In: *arXiv preprint arXiv:2405.06682*. URL: https://arxiv.org/abs/2405.06682.

Rio-Chanona, R. M. del, M. Pangallo, and C. Hommes (2025). *Can Generative AI agents behave like humans? Evidence from laboratory market experiments*. arXiv: 2505.07457 [econ.GN].

Roberts, Jack, Kieran Moore, and Daniel Fisher (2024). "Do Large Language Models Learn Human-Like Strategic Preferences?" In: *arXiv preprint arXiv:2404.08710*.

Romer, Christina D and David H Romer (June 2000). "Federal Reserve Information and the Behavior of Interest Rates". en. In: *American Economic Review* 90.3, pp. 429–457. DOI: 10.1257/aer.90.3.429. URL: https://www.aeaweb.org/articles?id=10.1257/aer.90.3.429.

Romer, Christina D. (1986). "Is the Stabilization of the Postwar Economy a Figment of the Data?" In: *The American Economic Review* 76.3, pp. 314–334.

Sargent, Thomas J. (1973). "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment". In: *Brookings Papers on Economic Activity* 1973.2, pp. 429–472. DOI: 10.2307/2534095.

Sargent, Thomas J. (1987). "Rational Expectations". In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan, pp. 1–7.

Sargent, Thomas J. (Oct. 1993). *Bounded Rationality in Macroeconomics: The Arne Ryde Memorial Lectures*. Oxford: Oxford University Press. ISBN: 9780198288640. DOI: 10.1093/oso/9780198288640.001.0001. URL: https://doi.org/10.1093/oso/9780198288640.001.0001.

Sargent, Thomas J. (2025). "Sources of Artificial Intelligence". In: *Journal of Economic Dynamics and Control* 172, p. 104989. DOI: https://doi.org/10.1016/j.jedc.2024.104989.

Schoenegger, Philipp, Indre Tuminauskaite, Peter S. Park, Roberto V. S. Bastos, and Philip E. Tetlock (2024). "Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy". In: *Science Advances* 10.45, eadp1528.

Shah, Abhimanyu, Ke Zhu, Yifan Jiang, Jiaqi G. Wang, Abdullah K. Dayi, John J. Horton, and David C. Parkes (2025). "Learning from Synthetic Labs: Language Models as Auction Participants". In: *arXiv preprint arXiv:2507.09083*. URL: https://arxiv.org/abs/2507.09083.

Shiller, Robert J. (1981). "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?" In: *The American Economic Review* 71.3, pp. 421–436.

Shiller, Robert J. (2015). *Irrational Exuberance: Revised and Expanded Third Edition*. Princeton, NJ: Princeton University Press.

Smith, Vernon L., Gerry L. Suchanek, and Arlington W. Williams (1988). "Bubbles, crashes, and endogenous expectations in experimental spot asset markets". In: *Econometrica: Journal of the Econometric Society* 56.5, pp. 1119–1151. DOI: 10.2307/1911367.

Sun, S., E. Lee, D. Nan, X. Zhao, W. Lee, B. J. Jansen, and J. H. Kim (2024). "Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information". In: *arXiv*. arXiv: 2402.18144 [cs.AI]. URL: https://arxiv.org/abs/2402.18144.

Thakur, Abhishek S., Kritika Choudhary, Venkata Sai Ramayapally, Siddharth Vaidyanathan, and Dieuwke Hupkes (2024). "Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges". In: *arXiv preprint arXiv:2406.12624*.

Tomasev, Nenad, Matthew Franklin, Joel Z. Leibo, Joshua Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero (2025). "Virtual Agent Economies". In: *arXiv preprint arXiv:2509.10147*.

Tran, Kiet T., Duy Dao, Minh D. Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Huy D. Nguyen (2025). "Multi-agent collaboration mechanisms: A survey of LLMs". In: *arXiv preprint arXiv:2501.06322*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Wang, Xu, Min Feng, Jinfeng Qiu, Jiaxing Gu, and Jun Zhao (2024). "From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection". In: *Advances in Neural Information Processing Systems*. Vol. 37, pp. 58118–58153.

Wei, Jason et al. (2022). "Emergent Abilities of Large Language Models". In: *Transactions on Machine Learning Research*. DOI: 10.48550/arXiv.2206.07682. arXiv: 2206.07682 [cs.CL].

Willis, R., Y. Du, J. Z. Leibo, and M. Luck (2025). "Will Systems of LLM Agents Cooperate: An Investigation into a Social Dilemma". In: *arXiv*. arXiv: 2501.16173 [cs.AI]. URL: https://arxiv.org/abs/2501.16173.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie (2025). *LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification*. Available at SSRN: https://ssrn.com/abstract=5517960 or http://dx.doi.org/10.2139/ssrn.5517960.

Wu, Shijie, Ozan Irsoy, Simon Lu, Vitaly Dabravolski, Mark Dredze, Sebastian Gehrmann, Gideon Mann, et al. (2023). "BloombergGPT: A Large Language Model for Finance". In: *arXiv preprint arXiv:2303.17564*.

Xiao, Min, Zhiwei Jiang, Long Qian, Zhuo Chen, Yuming He, Yifei Xu, and Qiang Xie (2025). "Enhancing Financial Time-Series Forecasting with Retrieval-Augmented Large Language Models". In: *arXiv*. arXiv: 2503.67890 [q-fin.MF].

Ye, Jiexia, Yongzi Yu, Weiqi Zhang, Le Wang, Jia Li, and Fugee Tsung (2025). "Empowering Time Series Analysis with Foundation Models: A Comprehensive Survey". In: *arXiv preprint arXiv:2405.02358*. 30 pages, 10 figures, 5 tables. Subjects: Machine Learning (cs.LG); Artificial Intelligence (cs.AI). URL: https://doi.org/10.48550/arXiv.2405.02358.

Zeng, Siliang, Luca Viano, Chenliang Li, et al. (May 2025). "Aligning Large Language Models with Human Feedback: Mathematical Foundations and Algorithm Design". In: *TechRxiv*. DOI: 10.36227/techrxiv.174784525.51683948/v1.

A Studies on GenAI Agents in Economics

Table A.1: Summary of Recent Studies on GenAI Agents in Economics

Paper	Models	Area
Sun et al. (2024)	GPT-3.5	Surveys
Bybee (2023)	GPT-3.5	Surveys
Ma et al. (2024)	LLaMA 2, Gemma 7B, Mistral 7B	Surveys
Wu, Xi, et al. (2025)	GPT-4.1, GPT-5	Surveys
Fan et al. (2024)	GPT-3, GPT-4	Game Theory
Roberts et al. (2024)	BERT, ELECTRA, GPT (1–2), LLaMA (1–2), Mistral 7B, Gemma 7B, Solar 10B	Game Theory
Lorè and Heydari (2024)	GPT-3.5, GPT-4, LLaMA 2	Game Theory
Polachek et al. (2024)	GPT-3.5, GPT-4, Gemini 1.0 Pro	Game Theory
Akata et al. (2025)	Davinci, GPT-4, LLaMA 2, Claude 2	Game Theory
Kitadai et al. (2025)	GPT-3.5, GPT-4	Game Theory
Shah et al. (2025)	GPT-4	Game Theory
Filippas et al. (2024)	GPT-4/4o, Claude 3/3.5, LLaMA 3 70B	Game Theory
Karabag and Topcu (2025)	GPT-4o/4.1/5, Gemini 2.5 Pro, LLaMA 3.1 70B, Qwen 3 32B	Game Theory
Lopez-Lira (2025)	GPT-4o	Finance
Hansen and Lee (2025)	Sonnet 3.5, Sonnet 3.7, LLaMa 3 Instruct 70B, Nova Pro	Finance
Henning et al. (2025)	GPT-3.5, GPT-4o, Claude 3.5, Grok 2, Mistral Large, Gemini 1.5 Pro	Finance
Faria-e-Castro and Leibovici (2024)	PaLM	Forecasting
Lopez-Lira and Tang (2023)	GPT-4	Forecasting
Rio-Chanona et al. (2025)	GPT-3.5/4	Forecasting
Schoenegger et al. (2024)	GPT-3.5/4, Claude 2, LLaMA 2 70B, PaLM 2, Coral, Mistral 7B, Bard, Falcon 180B, Qwen 7B	Forecasting
Hansen, Horton, et al. (2024)	GPT-3.5/4/4o	Forecasting
Carriero et al. (2024)	LagLlama, Moirai, TTM, Time-GPT, TimesFM	Forecasting

B Prompting Architecture

The prompting architecture translates the features of Table 1 into LLM-readable information. It consists of a fixed *system prompt* and a dynamic *user prompt*. The system prompt defines the LLM agent's role, and its full text is the following:

System Prompt

You are my financial advisor, re-instantiated in each round of a multi-round forecasting task. You must learn across rounds by reading the historical information provided in the prompt. At each round t, generate a forecast for round $t + 1$. Your goal is to maximize forecast accuracy. Learn from past market prices, forecast errors, and reflections. Include a brief reflection (1–3 sentences) to update your strategy over time. If any historical data section is missing, incomplete, or insufficient, mention this gap in your reflection.

The system prompt was further refined with explicit instructions on the required output format to ensure structured responses. Specifically, each LLM agent was instructed to return its output strictly as a JSON object following the following schema:

Output Format Instruction in the System Prompt

```
"Respond ONLY with a JSON object matching exactly:
{"round_current": int, "round_forecasted": int, "price_forecast": float,
"reflection": string}.
```

This instruction ensured consistent data parsing across models and experiments, and also served as a partial check of the task awareness by requiring to specify the current and forecasted rounds. In cases where the model output repeatedly failed to comply with the required JSON schema, a reinforcement message was automatically appended to the prompt to restate the formatting rules more explicitly. This mechanism dynamically appended corrective instructions to the prompt whenever the model produced malformed outputs.

Reinforcement Message (Attempts 2–4)

```
Return ONLY a strict JSON object with keys: round_current,
round_forecasted, price_forecast, reflection.
No markdown, no extra text; price_forecast must be a non-negative number.
```

If format errors persisted after five unsuccessful attempts, a stronger message was applied, providing explicit line-by-line guidance:

Enhanced Reinforcement Message (Attempt 5)

```
CRITICAL: Your previous responses had formatting errors.  Follow these rules
EXACTLY:
1.  Return ONLY valid JSON. No markdown (no “‘json), no code blocks, no
explanatory text.
2.  Start with { and end with }.  Nothing before or after.
3.  Use this exact structure (replace placeholders with your actual values):
{"round_current":  <integer>, "round_forecasted":  <integer>,
"price_forecast":  <number>, "reflection":  "<text>"}
4.  Required keys:  round_current (integer), round_forecasted (integer),
price_forecast (number), reflection (text in quotes).
5.  IMPORTANT: Calculate all mathematical operations.  Do NOT write
expressions like 60*1.15.  Instead, compute the result and write the final
number; price_forecast must be a non-negative number.
```

Complementing the system prompt, the *user prompt* specifies the information available to the agent in each round. The exact wording is shown below. Its rolling component includes the history of market prices, the agent's own forecasts, and its past reflections. I consider two versions of the User prompt: *Base* and *Caution*. The only difference between the two prompts lies in the additional sentence, shown below in red, which explicitly signals the potential evolution of market conditions.

User Prompts

CURRENT ROUND:

{*round_number*}

KEY INFORMATION:

- The risk-free interest rate is constant and equal to 0.05.
- The risky asset pays a dividend each round, averaging 3.
- Market structure and participants' behavior may evolve over time.
- Your task is to forecast the asset price in round {round_number + 1}.
- Use the historical information below to inform your forecast and improve over time.

HISTORY:

Market Prices:

{*market_prices*}

End Market Prices

Forecasts:

Each forecast is labeled with the round it was made in, and the round it refers to.

{*forecasts*}

End Forecasts

Reflections:

{*reflections*}

End Reflections

C Additional Tables

C.1 Summary Statistics

Table C.1: Summary Statistics by Model and Variant

Provider	Model	Variant	N	$\bar{p}$	$\bar{p^e}$	σ_p	σ_{p^e}
Anthropic	Claude 3.5 Haiku	1	25	64.33	64.55	9.65	10.30
Anthropic	Claude 3.5 Haiku	2	20	66.12	66.43	19.00	21.06
Anthropic	Claude 3.5 Haiku	3	30	65.48	65.76	10.89	11.90
Anthropic	Claude 3.5 Haiku	4	25	65.91	66.21	13.93	15.15
Meta	Llama 3.3 70B	1	25	61.35	61.42	8.03	8.44
Meta	Llama 3.3 70B	2	20	60.72	60.75	8.50	8.93
Meta	Llama 3.3 70B	3	30	60.89	60.94	8.80	9.28
Meta	Llama 3.3 70B	4	25	65.44	65.71	15.08	15.85
Google	Gemini 2.0 Flash	1	25	68.86	69.30	13.79	14.59
Google	Gemini 2.0 Flash	2	20	71.85	72.45	14.61	15.38
Google	Gemini 2.0 Flash	3	30	67.28	67.64	13.97	14.82
Google	Gemini 2.0 Flash	4	25	81.88	82.98	11.95	12.87
XAI	Grok 2	1	25	101.13	103.19	10.37	11.14
XAI	Grok 2	2	20	101.65	103.73	5.86	6.49
XAI	Grok 2	3	30	99.74	101.73	20.57	22.11
XAI	Grok 2	4	25	105.67	107.95	5.01	5.36
OpenAI	GPT-4o	1	25	60.99	61.04	15.43	16.33
OpenAI	GPT-4o	2	20	64.47	64.69	14.43	15.21
OpenAI	GPT-4o	3	30	57.85	57.74	15.57	16.63
OpenAI	GPT-4o	4	25	62.31	62.43	14.63	15.46
DeepSeek	DeepSeek Chat	1	25	49.74	49.22	32.84	34.60
DeepSeek	DeepSeek Chat	2	20	43.45	42.62	22.67	23.98
DeepSeek	DeepSeek Chat	3	30	55.82	55.61	23.60	25.35
DeepSeek	DeepSeek Chat	4	25	46.52	45.84	19.22	20.34

Notes: The table reports summary statistics by model and variant. For each, it shows the average asset price (p) across rounds and simulations, the average forecast (p^e) across rounds, simulations, and traders, and their respective sample standard deviations (σ_p and σ_{p^e}).

D Additional Figures

D.1 Price Dynamics by Variants

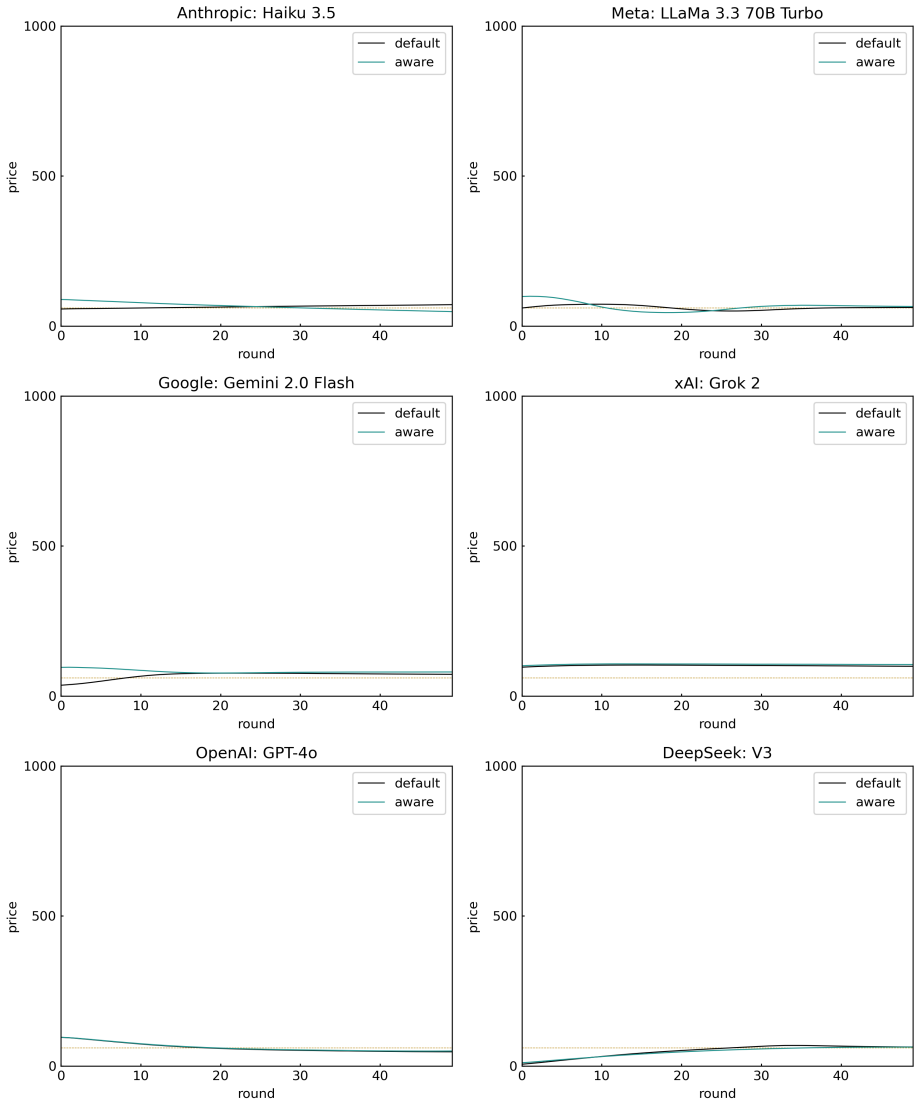


Figure D.1: Asset price dynamics with six LLM Agents per foundation model in an LtFE setup parametrically identical to Hommes et al. (2008). The orange dashed line denotes the fundamental price (60). Displayed prices are averaged across identical simulations for the baseline setup, variants 1 and 4 in Table Table 3. Variant 4, *Caution*, is referred to as *Aware* in the legend.

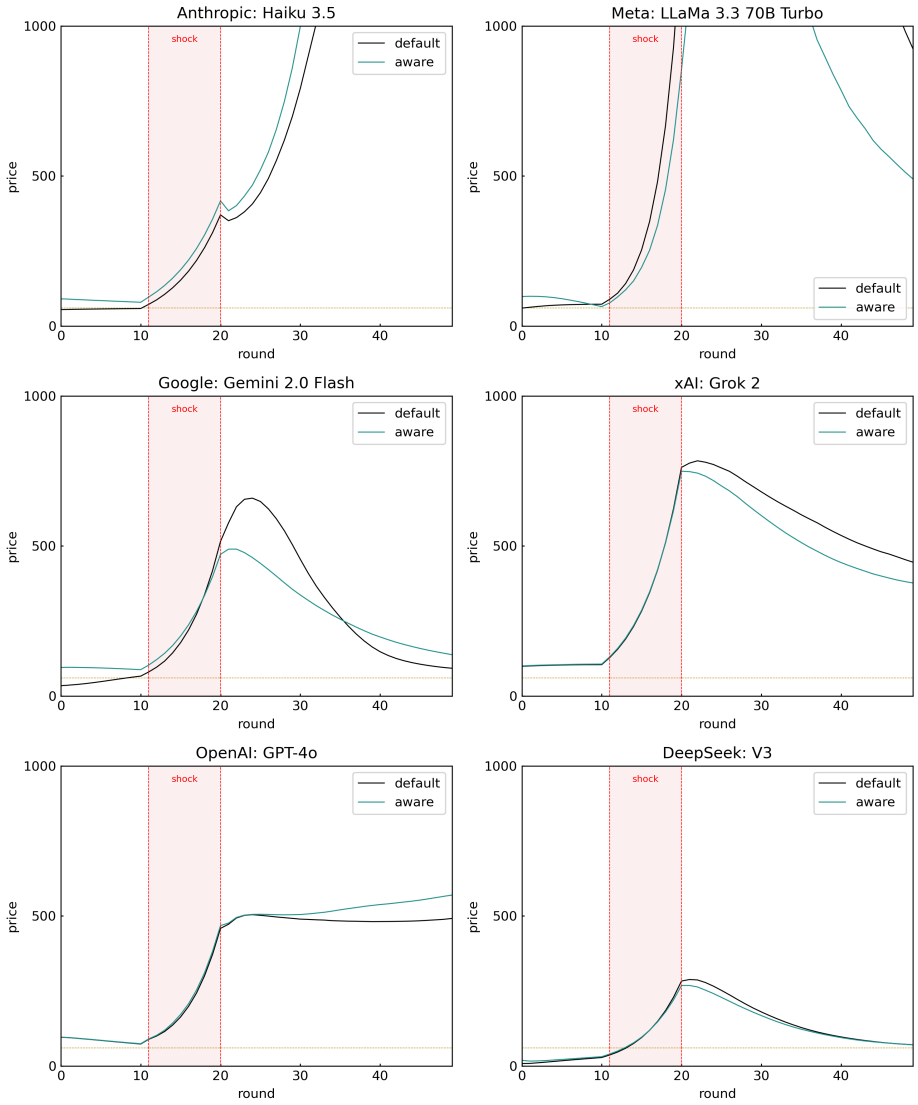


Figure D.2: Asset price dynamics with six LLM Agents for each foundation model, and five hypers active between $T_1 = 12$ and $T_2 = 21$, forming expectations according to equation (12) with $\gamma = 1.3$. The red area indicates the period of the hypers shock, when hypers are present in the market. The orange dashed line denotes the fundamental asset price, equal to 60. Displayed prices correspond to averages across identical simulations for the hypers setup, variants 1,2,3 in Table 3. The scale of the y-axis follows that of Figure 1 and 2 for comparability.

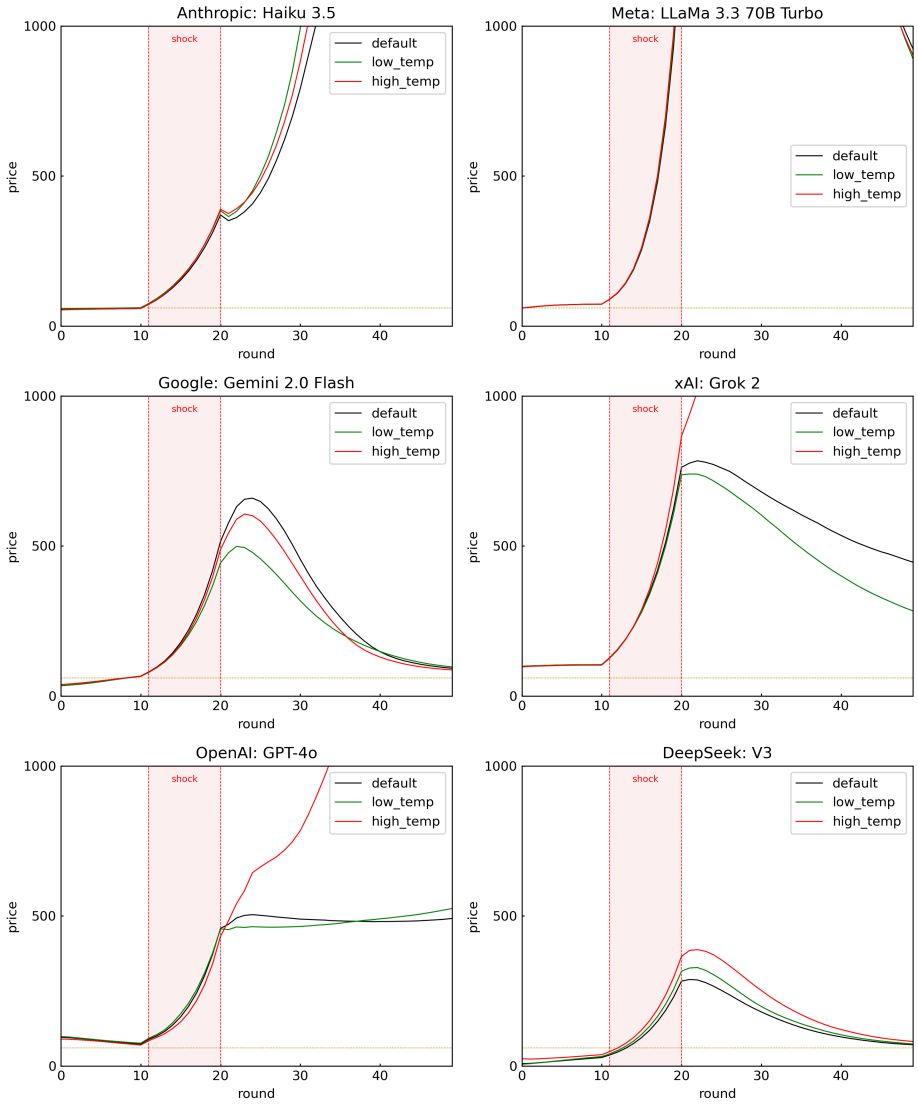


Figure D.3: Asset price dynamics with six LLM Agents for each foundation model, and five hypers active between $T_1 = 12$ and $T_2 = 21$, forming expectations according to equation (12) with $\gamma = 1.3$. The red area indicates the period of the hypers shock, when hypers are present in the market. The orange dashed line denotes the fundamental asset price, equal to 60. Displayed prices correspond to averages across identical simulations for the hypers setup, variants 1,4 in Table 3. The scale of the y-axis follows that of Figure 1 and 2 for comparability. Variant 4, *Caution*, is referred to as *Aware* in the legend.

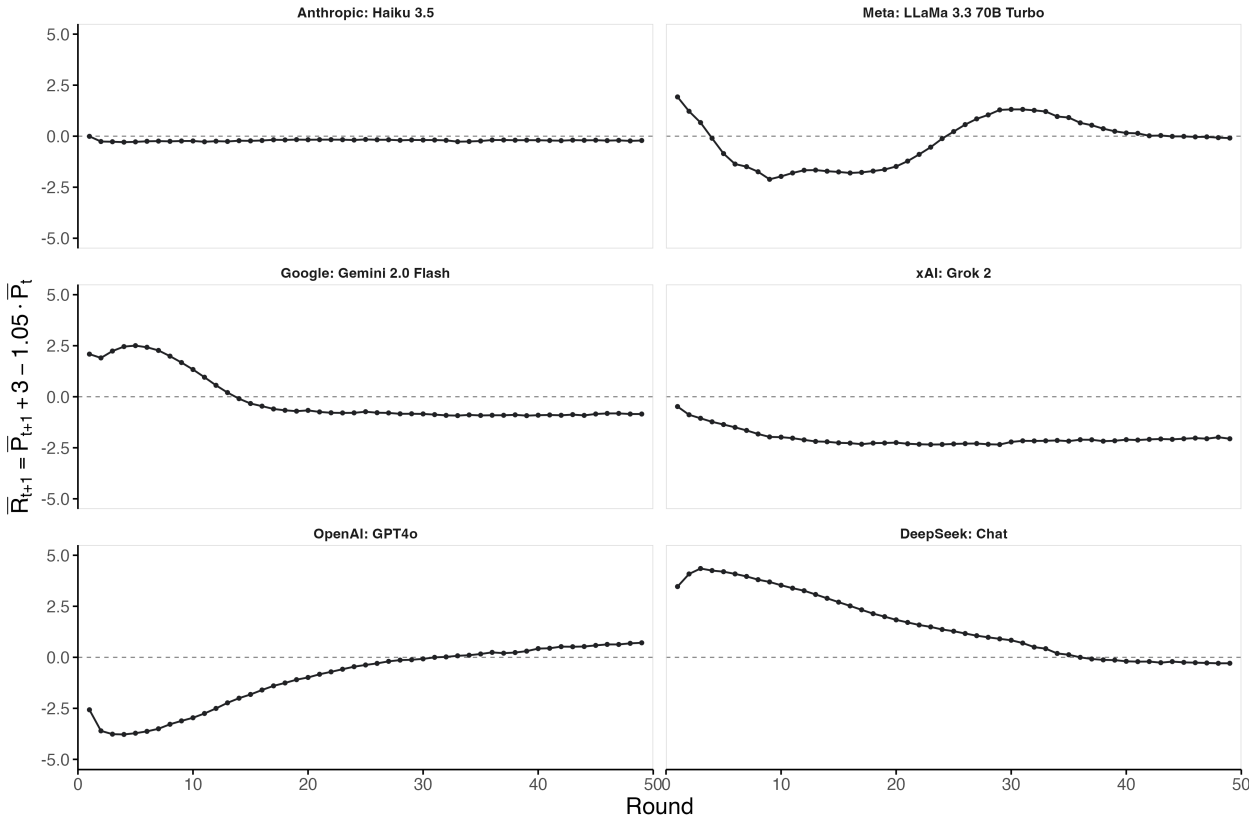


Figure E.1: Mean excess returns across all simulations (pooling all variants) for each round by foundation model.

E Rationality Tests

E.1 Market Efficiency

Figure E.1 and Figure E.2 display, respectively, the average excess returns across all simulations for each round and the corresponding autocorrelation functions by foundation model. While averaging excess returns across simulations may smooth out fluctuations and thus does not constitute a formal test, the emergence of systematic patterns even after averaging provides additional evidence suggestive of market inefficiency. Visually, average excess returns exhibit clear persistence and are predominantly positive during roughly the first 20 rounds, though with notable heterogeneity across models. Haiku's average returns remain close to zero throughout. Gemini and DeepSeek start positive and gradually converge toward zero, though not completely. GPT begins with negative excess returns and approaches zero over time. LLaMa oscillates around zero, with extended periods of persistent deviations both above and below it, while Grok starts and remains negative. Moreover, average excess returns display significant and positive higher-order autocorrelations for all models except Haiku.

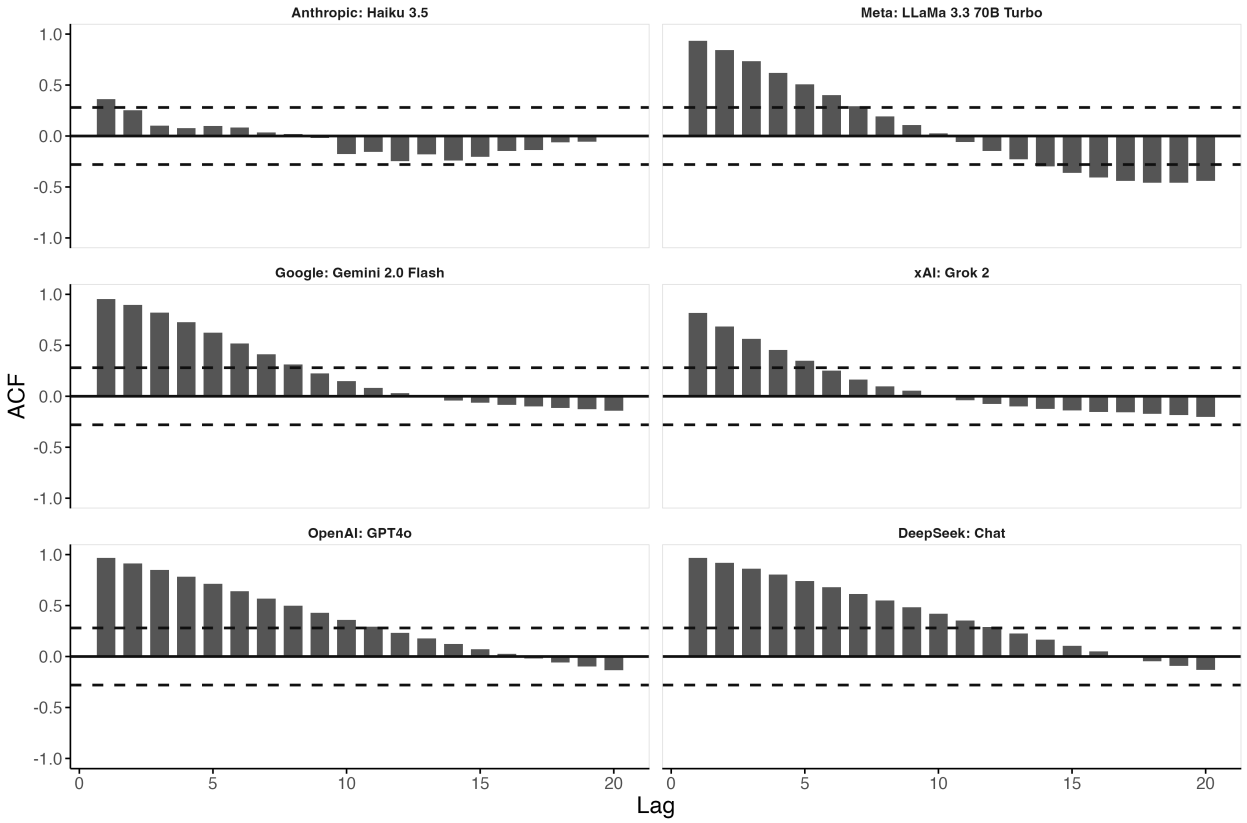


Figure E.2: Autocorrelation function of mean excess returns across all simulations (pooling all variants) for each round by foundation model. The dashed lines represent the Bartlett approximate 95% confidence bounds.

E.1.1 Variant-Specific Results

Table E.1: Market Efficiency Tests (Variant 1: Base User Prompt, Medium Temperature)

Model	Rounds	K	Avg $\bar{P}$	Avg $\tilde{R}$	(1) $E[P] = P_F$	(2) $E[\tilde{R}_t] = 0$ i.i.d.	(3) $E[\tilde{R}_t] = 0$ HAC(1)	(4) $\rho_t = 0$ $t = 2, \ldots, 5$
Claude 3.5 Haiku	[1–50]	25	64.33	0.087	.96	.56	.56	.24
	[10–50]	25	65.98	−0.010	.96	.68	.48	.32
Llama 3.3 70B	[1–50]	25	61.35	−0.031	.28	.04*	.00*	.92
	[10–50]	25	59.24	−0.238	.36	.12	.08	.76
Gemini 2.0 Flash	[1–50]	25	68.86	0.305	.84	.08	.00*	.56
	[10–50]	25	74.20	−0.613	1.00	.96	.76	.44
Grok 2	[1–50]	25	101.13	−1.996	1.00	1.00	1.00	.16
	[10–50]	25	101.41	−2.175	1.00	1.00	1.00	.16
GPT-4o	[1–50]	25	60.99	−1.044	.32	1.00	.96	.16
	[10–50]	25	54.29	−0.352	.84	.64	.48	.40
DeepSeek Chat	[1–50]	25	49.74	1.699	.96	.96	.96	.28
	[10–50]	25	58.50	0.848	.96	1.00	.92	.36

Notes: Columns (1)–(4) report the rejection rates of the null hypothesis (H_0) at the 0.05 significance level. Values at or below 0.05 (marked with *) indicate the null is not rejected. K denotes the number of simulations; Avg $\bar{P}$ is the average price across simulations and rounds; Avg $\tilde{R}$ is the average realized excess return across simulations and rounds.
(1) Fundamental pricing: $P_t - 60 = \alpha_0 + \varepsilon_t$; $H_0 : \alpha_0 = 0$ (OLS).
(2) Zero excess returns (i.i.d.): $\tilde{R}_{t+1} = \beta_0 + \varepsilon_{t+1}$; $H_0 : \beta_0 = 0$ (OLS).
(3) Zero excess returns (HAC): $\tilde{R}_{t+1} = \beta_0 + \varepsilon_{t+1}$; $H_0 : \beta_0 = 0$ (OLS with HAC(1) errors).
(4) No serial correlation (ρ_2–ρ_5): Ljung–Box Q-test on AR(1) residuals for $\tilde{R}_{t+1}$ with 5 lags.

Table E.2: Market Efficiency Tests (Variant 2: Base User Prompt, Low Temperature)

Model	Rounds	K	Avg $\bar{P}$	Avg $\tilde{R}$	(1) $E[P] = P_F$	(2) $E[\tilde{R}_t] = 0$ i.i.d.	(3) $E[\tilde{R}_t] = 0$ HAC(1)	(4) $\rho_t = 0$ $t = 2, \ldots, 5$
Claude 3.5 Haiku	[1–50]	20	66.12	0.107	1.00	.56	.56	.31
	[10–50]	20	68.02	0.057	.94	.56	.38	.31
Llama 3.3 70B	[1–50]	20	60.72	−0.027	.15	.00*	.00*	.95
	[10–50]	20	58.45	−0.236	.30	.10	.10	.80
Gemini 2.0 Flash	[1–50]	20	71.85	0.275	.95	.00*	.00*	.55
	[10–50]	20	78.10	−0.715	1.00	1.00	1.00	.55
Grok 2	[1–50]	20	101.65	−2.032	1.00	1.00	1.00	.20
	[10–50]	20	101.90	−2.194	1.00	1.00	1.00	.05*
GPT-4o	[1–50]	20	64.47	−1.251	.45	1.00	1.00	.15
	[10–50]	20	58.23	−0.592	.80	1.00	.95	.30
DeepSeek Chat	[1–50]	20	43.45	1.935	.90	1.00	1.00	.25
	[10–50]	20	50.70	1.278	.95	.95	.95	.10

Notes: Same as Table E.1.

Table E.3: Market Efficiency Tests (Variant 3: Base User Prompt, High Temperature)

Model	Rounds	K	Avg $\bar{P}$	Avg $\tilde{R}$	(1) $E[P] = P_F$	(2) $E[\tilde{R}_t] = 0$ i.i.d.	(3) $E[\tilde{R}_t] = 0$ HAC(1)	(4) $\rho_t = 0$ $t = 2, \ldots, 5$
Claude 3.5 Haiku	[1–50]	30	65.48	0.110	.90	.60	.57	.33
	[10–50]	30	67.59	−0.017	.93	.70	.57	.20
Llama 3.3 70B	[1–50]	30	60.89	−0.034	.20	.00*	.00*	.87
	[10–50]	30	58.64	−0.267	.30	.20	.03*	.83
Gemini 2.0 Flash	[1–50]	30	67.28	0.351	.70	.23	.20	.77
	[10–50]	30	72.31	−0.552	.97	.77	.73	.57
Grok 2	[1–50]	30	99.74	−1.896	1.00	.97	.97	.20
	[10–50]	30	99.63	−2.058	1.00	1.00	.97	.03*
GPT-4o	[1–50]	30	57.85	−0.700	.33	.77	.57	.30
	[10–50]	30	51.05	−0.014	.90	.20	.10	.30
DeepSeek Chat	[1–50]	30	55.82	1.075	.83	.77	.73	.47
	[10–50]	30	62.61	0.515	.87	.90	.87	.40

Notes: Same as Table E.1.

Table E.4: Market Efficiency Tests (Variant 4: *Caution* User Prompt, Medium Temperature)

Model	Rounds	K	Avg $\bar{P}$	Avg $\tilde{R}$	(1) $E[P] = P_F$	(2) $E[\tilde{R}_t] = 0$ i.i.d.	(3) $E[\tilde{R}_t] = 0$ HAC(1)	(4) $\rho_t = 0$ $t = 2, \ldots, 5$
Claude 3.5 Haiku	[1–50]	25	65.91	−1.144	.60	1.00	1.00	.36
	[10–50]	25	61.04	−0.825	.60	1.00	1.00	.20
Llama 3.3 70B	[1–50]	25	65.44	−0.949	.92	1.00	.00*	.84
	[10–50]	25	59.42	0.206	.04*	.00*	.00*	.84
Gemini 2.0 Flash	[1–50]	25	81.88	−1.409	1.00	1.00	.96	.52
	[10–50]	25	79.06	−1.048	.92	.84	.80	.52
Grok 2	[1–50]	25	105.67	−2.201	1.00	1.00	1.00	.24
	[10–50]	25	106.00	−2.363	1.00	1.00	1.00	.08
GPT-4o	[1–50]	25	62.31	−1.062	.32	1.00	1.00	.32
	[10–50]	25	55.80	−0.386	.76	.60	.52	.44
DeepSeek Chat	[1–50]	25	46.52	1.752	.88	.96	.96	.24
	[10–50]	25	53.72	1.115	.88	.96	.88	.28

Notes: Same as Table E.1.

E.2 Forecast Rationality Tests by Variant

Table E.5: Forecast Rationality Tests (Variant 1: Base User Prompt, Medium Temperature)

Model	Regression Specifications		
	$p_{t+1} = a_0 + a_1 p^e_{t+1}$ $H_0 : a_0 = 0,\ a_1 = 1$	$p_{t+1} = b_0 + b_1 p^e_{t+1} + b_2 \text{FE}_{t-1}$ $H_0 : b_0 = 0,\ b_1 = 1,\ b_2 = 0$	$p_{t+1} = c_0 + c_1 p^e_{t+1} + c_2 \text{PFE}_t$ $H_0 : c_0 = 0,\ c_1 = 1,\ c_2 = 0$
Claude 3.5 Haiku	a_0: 3.209 (0.362)	b_0: 0.580 (0.356)	c_0: -1.868 (0.082)
	a_1: 0.951 (0.006)	b_1: 0.994 (0.006)	c_1: 1.038 (0.001)
		b_2: 0.638 (0.093)	c_2: 0.970 (0.007)
	$\bar{R}^2 = 0.977$	$\bar{R}^2 = 0.990$	$\bar{R}^2 = 0.997$
	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)
Llama 3.3 70B	a_0: 3.716 (0.137)	b_0: 1.680 (0.731)	c_0: 6.846 (0.145)
	a_1: 0.938 (0.002)	b_1: 0.972 (0.012)	c_1: 0.886 (0.002)
		b_2: 0.721 (0.247)	c_2: -0.823 (0.047)
	$\bar{R}^2 = 0.972$	$\bar{R}^2 = 0.973$	$\bar{R}^2 = 0.992$
	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)
Gemini 2.0 Flash	a_0: 7.997 (0.122)	b_0: 7.179 (0.284)	c_0: 7.695 (0.120)
	a_1: 0.889 (0.002)	b_1: 0.902 (0.004)	c_1: 0.889 (0.002)
		b_2: 0.243 (0.068)	c_2: -0.260 (0.016)
	$\bar{R}^2 = 0.986$	$\bar{R}^2 = 0.987$	$\bar{R}^2 = 0.987$
	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)
Grok 2	a_0: 1.666 (0.910)	b_0: -0.230 (0.558)	c_0: -5.221 (0.395)
	a_1: 0.964 (0.009)	b_1: 0.991 (0.006)	c_1: 1.048 (0.004)
		b_2: 0.404 (0.048)	c_2: 0.824 (0.019)
	$\bar{R}^2 = 0.975$	$\bar{R}^2 = 0.983$	$\bar{R}^2 = 0.990$
	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)
GPT-4o	a_0: 4.963 (0.130)	b_0: 3.905 (0.226)	c_0: 3.887 (0.093)
	a_1: 0.901 (0.002)	b_1: 0.919 (0.004)	c_1: 0.910 (0.001)
		b_2: 0.322 (0.056)	c_2: 0.573 (0.030)
	$\bar{R}^2 = 0.991$	$\bar{R}^2 = 0.993$	$\bar{R}^2 = 0.994$
	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)
DeepSeek Chat	a_0: 5.173 (0.395)	b_0: 2.135 (0.405)	c_0: 4.945 (0.383)
	a_1: 0.929 (0.009)	b_1: 0.980 (0.008)	c_1: 0.936 (0.009)
		b_2: 0.931 (0.105)	c_2: 0.162 (0.066)
	$\bar{R}^2 = 0.987$	$\bar{R}^2 = 0.994$	$\bar{R}^2 = 0.988$
	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)	H_0 *Rej* ($p < 0.0001$)

Notes: Same as Table 5, except variant dummies are not included in the specification.

Table E.6: Forecast Rationality Tests (Variant 2: Base User Prompt, Low Temperature)

Model	Regression Specifications		
	$p_{t+1} = a_0 + a_1 p^e_{t+1}$ $H_0 : a_0 = 0,\ a_1 = 1$	$p_{t+1} = b_0 + b_1 p^e_{t+1} + b_2 \text{FE}_{t-1}$ $H_0 : b_0 = 0,\ b_1 = 1,\ b_2 = 0$	$p_{t+1} = c_0 + c_1 p^e_{t+1} + c_2 \text{PFE}_t$ $H_0 : c_0 = 0,\ c_1 = 1,\ c_2 = 0$
Claude 3.5 Haiku	a_0: 6.039 (3.593) a_1: 0.910 (0.056)	b_0: -3.454 (0.131) b_1: 1.058 (0.002) b_2: 1.145 (0.008)	c_0: -7.035 (0.175) c_1: 1.119 (0.003) c_2: 1.107 (0.009)
	$\bar{R}^2 = 0.902$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.998$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.997$ H_0 *Rej* ($p < 0.0001$)
Llama 3.3 70B	a_0: 3.564 (0.140) a_1: 0.940 (0.002)	b_0: 0.991 (1.241) b_1: 0.983 (0.021) b_2: 0.909 (0.432)	c_0: 6.724 (0.150) c_1: 0.887 (0.002) c_2: -0.840 (0.047)
	$\bar{R}^2 = 0.974$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.976$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.994$ H_0 *Rej* ($p < 0.0001$)
Gemini 2.0 Flash	a_0: 9.085 (0.118) a_1: 0.879 (0.002)	b_0: 6.064 (0.118) b_1: 0.927 (0.002) b_2: 0.878 (0.017)	c_0: 8.841 (0.124) c_1: 0.880 (0.002) c_2: -0.109 (0.018)
	$\bar{R}^2 = 0.988$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.992$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.988$ H_0 *Rej* ($p < 0.0001$)
Grok 2	a_0: 7.240 (0.524) a_1: 0.910 (0.005)	b_0: 5.302 (0.477) b_1: 0.935 (0.005) b_2: 0.302 (0.045)	c_0: 2.080 (0.289) c_1: 0.976 (0.003) c_2: 0.770 (0.012)
	$\bar{R}^2 = 0.953$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.964$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.981$ H_0 *Rej* ($p < 0.0001$)
GPT-4o	a_0: 5.091 (0.158) a_1: 0.902 (0.002)	b_0: 1.792 (0.057) b_1: 0.957 (0.001) b_2: 0.975 (0.010)	c_0: 3.920 (0.086) c_1: 0.912 (0.001) c_2: 0.680 (0.012)
	$\bar{R}^2 = 0.991$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.998$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.996$ H_0 *Rej* ($p < 0.0001$)
DeepSeek Chat	a_0: 4.871 (0.186) a_1: 0.931 (0.004)	b_0: 2.284 (0.280) b_1: 0.976 (0.005) b_2: 0.755 (0.068)	c_0: 3.517 (0.136) c_1: 0.967 (0.003) c_2: 0.690 (0.039)
	$\bar{R}^2 = 0.985$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.995$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.993$ H_0 *Rej* ($p < 0.0001$)

Notes: Same as Table E.5.

Table E.7: Forecast Rationality Tests (Variant 3: Base User Prompt, High Temperature)

Model	Regression Specifications		
	$p_{t+1} = a_0 + a_1 p^e_{t+1}$ $H_0 : a_0 = 0,\ a_1 = 1$	$p_{t+1} = b_0 + b_1 p^e_{t+1} + b_2 \text{FE}_{t-1}$ $H_0 : b_0 = 0,\ b_1 = 1,\ b_2 = 0$	$p_{t+1} = c_0 + c_1 p^e_{t+1} + c_2 \text{PFE}_t$ $H_0 : c_0 = 0,\ c_1 = 1,\ c_2 = 0$
Claude 3.5 Haiku	a_0: 5.649 (0.448) a_1: 0.916 (0.007)	b_0: 1.365 (0.346) b_1: 0.983 (0.005) b_2: 0.683 (0.058)	c_0: -1.541 (0.089) c_1: 1.035 (0.001) c_2: 0.996 (0.006)
	$\bar{R}^2 = 0.941$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.978$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.995$ H_0 *Rej* ($p < 0.0001$)
Llama 3.3 70B	a_0: 3.664 (0.118) a_1: 0.938 (0.002)	b_0: 3.249 (0.209) b_1: 0.945 (0.003) b_2: 0.147 (0.058)	c_0: 6.478 (0.246) c_1: 0.892 (0.004) c_2: -0.751 (0.078)
	$\bar{R}^2 = 0.974$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.974$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.990$ H_0 *Rej* ($p < 0.0001$)
Gemini 2.0 Flash	a_0: 7.255 (0.109) a_1: 0.898 (0.002)	b_0: 6.757 (0.166) b_1: 0.906 (0.003) b_2: 0.152 (0.035)	c_0: 7.070 (0.106) c_1: 0.895 (0.002) c_2: -0.352 (0.016)
	$\bar{R}^2 = 0.985$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.986$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.987$ H_0 *Rej* ($p < 0.0001$)
Grok 2	a_0: 1.278 (2.790) a_1: 0.968 (0.028)	b_0: -4.030 (0.607) b_1: 1.038 (0.006) b_2: 0.886 (0.047)	c_0: -10.146 (0.773) c_1: 1.101 (0.008) c_2: 0.980 (0.022)
	$\bar{R}^2 = 0.960$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.989$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.991$ H_0 *Rej* ($p < 0.0001$)
GPT-4o	a_0: 4.892 (0.094) a_1: 0.903 (0.002)	b_0: 4.677 (0.096) b_1: 0.906 (0.002) b_2: 0.070 (0.016)	c_0: 4.962 (0.087) c_1: 0.902 (0.001) c_2: -0.041 (0.057)
	$\bar{R}^2 = 0.992$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.993$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.992$ H_0 *Rej* ($p < 0.0001$)
DeepSeek Chat	a_0: 5.282 (0.160) a_1: 0.925 (0.002)	b_0: 4.373 (0.134) b_1: 0.940 (0.002) b_2: 0.249 (0.029)	c_0: 4.141 (0.107) c_1: 0.952 (0.002) c_2: 0.588 (0.032)
	$\bar{R}^2 = 0.985$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.989$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.991$ H_0 *Rej* ($p < 0.0001$)

Notes: Same as Table E.5.

Table E.8: Forecast Rationality Tests (Variant 4: *Caution* User Prompt, Medium Temperature)

Model	Regression Specifications		
	$p_{t+1} = a_0 + a_1 p^e_{t+1}$ $H_0 : a_0 = 0,\ a_1 = 1$	$p_{t+1} = b_0 + b_1 p^e_{t+1} + b_2 \text{FE}_{t-1}$ $H_0 : b_0 = 0,\ b_1 = 1,\ b_2 = 0$	$p_{t+1} = c_0 + c_1 p^e_{t+1} + c_2 \text{PFE}_t$ $H_0 : c_0 = 0,\ c_1 = 1,\ c_2 = 0$
Claude 3.5 Haiku	a_0: 3.903 (0.377) a_1: 0.924 (0.005)	b_0: 2.763 (0.302) b_1: 0.943 (0.004) b_2: 0.269 (0.039)	c_0: -0.341 (0.094) c_1: 0.981 (0.001) c_2: 0.893 (0.006)
	$\bar{R}^2 = 0.976$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.982$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.996$ H_0 *Rej* ($p < 0.0001$)
Llama 3.3 70B	a_0: 6.721 (0.141) a_1: 0.882 (0.002)	b_0: 5.316 (0.594) b_1: 0.906 (0.010) b_2: 0.477 (0.187)	c_0: 7.641 (0.045) c_1: 0.873 (0.001) c_2: -0.738 (0.068)
	$\bar{R}^2 = 0.974$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.974$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.991$ H_0 *Rej* ($p < 0.0001$)
Gemini 2.0 Flash	a_0: 4.726 (0.193) a_1: 0.926 (0.002)	b_0: 3.946 (0.232) b_1: 0.938 (0.003) b_2: 0.206 (0.035)	c_0: 4.065 (0.199) c_1: 0.935 (0.002) c_2: 0.173 (0.027)
	$\bar{R}^2 = 0.978$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.980$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.978$ H_0 *Rej* ($p < 0.0001$)
Grok 2	a_0: 5.212 (0.426) a_1: 0.931 (0.004)	b_0: 0.345 (0.219) b_1: 0.993 (0.002) b_2: 0.817 (0.012)	c_0: -1.146 (0.297) c_1: 1.006 (0.003) c_2: 0.743 (0.012)
	$\bar{R}^2 = 0.956$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.981$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.979$ H_0 *Rej* ($p < 0.0001$)
GPT-4o	a_0: 5.136 (0.127) a_1: 0.901 (0.002)	b_0: 4.425 (0.189) b_1: 0.913 (0.003) b_2: 0.225 (0.047)	c_0: 4.656 (0.099) c_1: 0.903 (0.001) c_2: 0.357 (0.024)
	$\bar{R}^2 = 0.994$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.995$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.995$ H_0 *Rej* ($p < 0.0001$)
DeepSeek Chat	a_0: 5.641 (0.132) a_1: 0.915 (0.002)	b_0: 4.109 (0.335) b_1: 0.941 (0.006) b_2: 0.457 (0.095)	c_0: 5.105 (0.070) c_1: 0.932 (0.001) c_2: 0.645 (0.024)
	$\bar{R}^2 = 0.990$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.994$ H_0 *Rej* ($p < 0.0001$)	$\bar{R}^2 = 0.995$ H_0 *Rej* ($p < 0.0001$)

Notes: Same as Table E.5.

E.3 Dispersion in Gains from Trade

E.3.1 Max Sharpe Ratio Tests

Table E.9: Sharpe Ratio Test Rejections (pooled variants)

Model	K	α	Rej (%)
Anthropic: Haiku 3.5	96	0.10	53.1
Anthropic: Haiku 3.5	96	0.05	44.8
Anthropic: Haiku 3.5	96	0.01	36.5
Meta: LLaMa 3.3 70B Turbo	100	0.10	32.0
Meta: LLaMa 3.3 70B Turbo	100	0.05	19.0
Meta: LLaMa 3.3 70B Turbo	100	0.01	6.0
Google: Gemini 2.0 Flash	100	0.10	86.0
Google: Gemini 2.0 Flash	100	0.05	82.0
Google: Gemini 2.0 Flash	100	0.01	75.0
xAI: Grok 2	100	0.10	96.0
xAI: Grok 2	100	0.05	96.0
xAI: Grok 2	100	0.01	96.0
OpenAI: GPT4o	100	0.10	76.0
OpenAI: GPT4o	100	0.05	66.0
OpenAI: GPT4o	100	0.01	55.0
DeepSeek: Chat	100	0.10	91.0
DeepSeek: Chat	100	0.05	90.0
DeepSeek: Chat	100	0.01	85.0

Notes: Entries report the percentage of simulations rejecting the Sharpe ratio test at level α for each model using K simulations with non-zero gains variance for which SR can be computed.

Table E.10: SR Test Rejections by Variant

Model	K	α	V1 Rej (%)	V2 Rej (%)	V3 Rej (%)	V4 Rej (%)
Anthropic: Haiku 3.5	96	0.10	32.0	68.0	62.5	53.3
Anthropic: Haiku 3.5	96	0.05	20.0	68.0	56.2	40.0
Anthropic: Haiku 3.5	96	0.01	12.0	68.0	43.8	26.7
Meta: LLaMa 3.3 70B Turbo	100	0.10	24.0	20.0	55.0	33.3
Meta: LLaMa 3.3 70B Turbo	100	0.05	16.0	12.0	30.0	20.0
Meta: LLaMa 3.3 70B Turbo	100	0.01	0.0	8.0	5.0	10.0
Google: Gemini 2.0 Flash	100	0.10	80.0	96.0	100.0	73.3
Google: Gemini 2.0 Flash	100	0.05	80.0	92.0	100.0	63.3
Google: Gemini 2.0 Flash	100	0.01	72.0	80.0	100.0	56.7
xAI: Grok 2	100	0.10	96.0	96.0	100.0	93.3
xAI: Grok 2	100	0.05	96.0	96.0	100.0	93.3
xAI: Grok 2	100	0.01	96.0	96.0	100.0	93.3
OpenAI: GPT4o	100	0.10	84.0	80.0	100.0	50.0
OpenAI: GPT4o	100	0.05	76.0	76.0	100.0	26.7
OpenAI: GPT4o	100	0.01	60.0	64.0	95.0	16.7
DeepSeek: Chat	100	0.10	100.0	92.0	100.0	76.7
DeepSeek: Chat	100	0.05	100.0	92.0	100.0	73.3
DeepSeek: Chat	100	0.01	100.0	92.0	100.0	56.7

Notes: V1–V4 report rejection shares (%) at level α by variant; K is the number of simulations with non-zero gains variance for which SR can be computed.

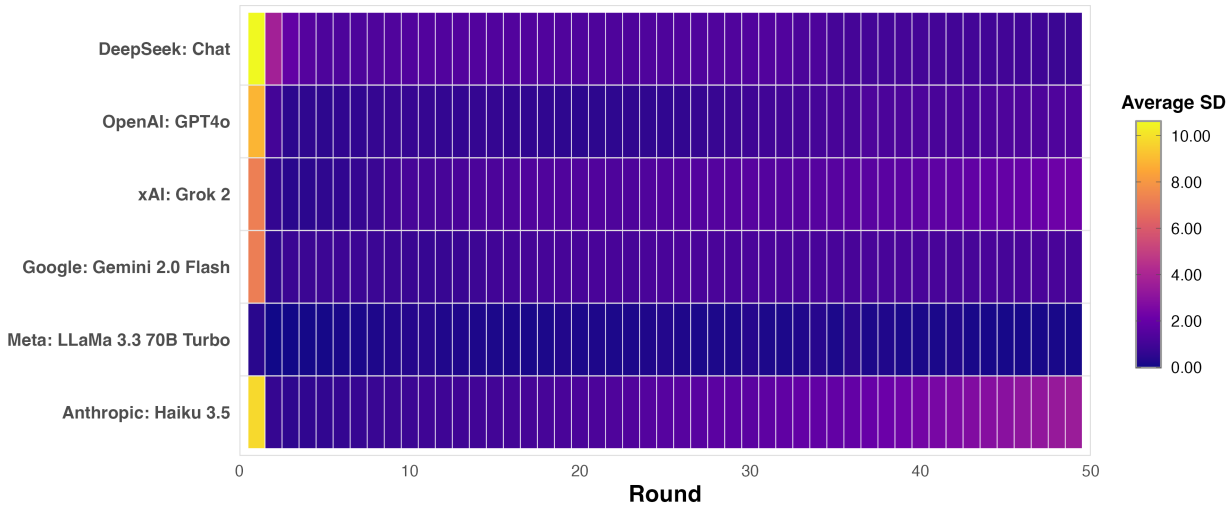


Figure E.3: Standard deviation of expected excess returns average across all simulations (pooling all variants) for each round by foundation model. A lighter color translates into higher dispersion in gains from trade.

E.3.2 Coordination Tests

Does the dispersion in gains from trade decrease over time as Agents potentially learn to coordinate their forecasts? To address this, I compute the variance of the expected excess returns, $\tilde{R}_t^e$, across LLM Agents in each round. As shown in (11), this variance is directly proportional to the dispersion in gains from trade: the lower the variance, the more coordinated the forecasts of the LLM Agents, and consequently, the smaller the dispersion in gains from trade, being equal to zero in the limit of full coordination.

Figure E.3 presents a heatmap showing, for each round and each model, the standard deviation of $\tilde{R}_t^e$ across LLM Agents, averaged across all simulations. Dispersion is highest in the first round and then declines sharply, suggesting that LLM Agents become more coordinated immediately after they receive the first market feedback. However, coordination does not always improve or remain stable over time. In particular, for some models, dispersion seems to increase monotonically, and hence coordination weakens, after the second round. To formally assess whether coordination systematically deteriorates over time after round 1 (excluded), I compute, for each model–variant–simulation combination, the difference $SD(\tilde{R}_T^e) - SD(\tilde{R}_2^e)$. I then perform a one-sided test of the null hypothesis that the average of this difference across all simulations for each model is less than or equal to zero, against the alternative that it is greater than zero. If the test is rejected, it implies that average dispersion is significantly greater in the last round relative to the second round, indicating that coordination deteriorates over time for that model (after the first improvement). Results are reported in Table E.11, and analogous tests conducted separately by variant are shown in Table E.12. The null is rejected at the 1% significance level for Haiku, Gemini, and Grok, at the 5% level for LLaMa, and

Table E.11: Coordination Test (pooled variants)

Model	K	SD($\tilde{R}^e_2$)	SD($\tilde{R}^e_T$)	Δ(SD)	p-value*
Haiku 3.5	100	0.679	3.606	2.927	0.001
LLaMa 3.3 70B Turbo	100	0.035	0.088	0.054	0.017
Gemini 2.0 Flash	100	0.537	1.070	0.533	< 0.001
Grok 2	100	0.629	2.279	1.650	0.008
GPT4o	100	1.030	1.378	0.348	0.064
DeepSeek Chat	100	3.716	0.854	-2.862	1.000

Notes: Entries report the average cross-sectional standard deviation of expected excess returns SD($\tilde{R}^e$) at round 2 and at the last round T, and their mean difference Δ(SD). The last column shows the p-value* from a one-sided t-test of $H_0 : \mu_{\Delta(\text{SD})} \leq 0$ against $H_1 : \mu_{\Delta(\text{SD})} > 0$.

at the 10% level for GPT-4o. The only model exhibiting a negative average difference, indicating improved coordination over time, is DeepSeek Chat. Results are qualitatively unchanged when considering the variant-specific tests. This provides strong evidence that not only does a systematic winner emerge, but also that coordination in forecasts, which would reduce dispersion in gains from trade, improves only after the first market signal and subsequently deteriorates over time, reaching a significantly lower level by the final round in five out of six models, despite all LLM Agents originating from the same foundation model.

Table E.12: Coordination Tests by Variant

Model	Variant	K	SD($\tilde{R}^e_2$)	SD($\tilde{R}^e_T$)	Δ(SD)	p-value*
Haiku 3.5	V2	25	0.685	4.132	3.447	< 0.001
Haiku 3.5	V1	25	0.705	1.689	0.985	0.005
Haiku 3.5	V4	30	0.833	2.994	2.162	0.002
Haiku 3.5	V3	20	0.409	6.262	5.853	0.109
LLaMa 3.3 70B Turbo	V2	25	0.001	0.067	0.067	< 0.001
LLaMa 3.3 70B Turbo	V1	25	0.000	0.077	0.077	< 0.001
LLaMa 3.3 70B Turbo	V4	30	0.094	0.130	0.036	0.327
LLaMa 3.3 70B Turbo	V3	20	0.031	0.066	0.035	0.080
Gemini 2.0 Flash	V2	25	0.757	1.403	0.646	< 0.001
Gemini 2.0 Flash	V1	25	0.471	0.884	0.413	0.024
Gemini 2.0 Flash	V4	30	0.644	0.923	0.279	0.019
Gemini 2.0 Flash	V3	20	0.185	1.106	0.921	< 0.001
GPT4o	V2	25	0.710	1.019	0.309	0.183
GPT4o	V1	25	0.776	1.987	1.211	0.011
GPT4o	V4	30	2.053	0.698	-1.355	1.000
GPT4o	V3	20	0.212	2.084	1.871	< 0.001
DeepSeek Chat	V2	25	1.800	0.824	-0.976	0.849
DeepSeek Chat	V1	25	0.172	0.361	0.190	0.091
DeepSeek Chat	V4	30	8.795	1.143	-7.652	1.000
DeepSeek Chat	V3	20	2.924	1.074	-1.851	0.865
grok-2-latest	V2	25	0.662	1.318	0.657	< 0.001
grok-2-latest	V1	25	0.480	1.973	1.493	0.005
grok-2-latest	V4	30	1.010	3.730	2.719	0.113
grok-2-latest	V3	20	0.203	1.688	1.485	< 0.001

Notes: Entries report the average cross-sectional standard deviation of expected excess returns SD($\tilde{R}^e$) at round 2 and at the last round T, and their mean difference Δ(SD) across prompt variants (V1–V4). The last column shows the p-value* from a one-sided t-test of $H_0 : \mu_{\Delta(\mathrm{SD})} \leq 0$ against $H_1 : \mu_{\Delta(\mathrm{SD})} > 0$.

F Hype Susceptibility Parameter by Variant

Provider	Model	Rounds	δ
Anthropic	Haiku 3.5	[12–22]	**0.56** (0.02)
		[23–50]	**0.78** (0.00)
Meta	LLaMa 3.3 70B Turbo	[12–22]	**0.98** (0.01)
		[23–50]	**0.51** (0.01)
Google	Gemini 2.0 Flash	[12–22]	**0.75** (0.01)
		[23–50]	**0.49** (0.01)
xAI	Grok 2	[12–22]	**0.63** (0.01)
		[23–50]	**0.55** (0.00)
OpenAI	GPT4o	[12–22]	**0.66** (0.02)
		[23–50]	**0.56** (0.00)
DeepSeek	Chat	[12–22]	**0.68** (0.01)
		[23–50]	**0.45** (0.00)

Table F.1: Estimates of the Hype Susceptibility Parameters. The column *Rounds* indicates the sample period used for each estimation. Data from all simulations for variant 1 Table 3.

Provider	Model	Rounds	δ
Anthropic	Haiku 3.5	[12–22]	**0.56** (0.01)
		[23–50]	**0.84** (0.00)
Meta	LLaMa 3.3 70B Turbo	[12–22]	**0.99** (0.00)
		[23–50]	**0.52** (0.01)
Google	Gemini 2.0 Flash	[12–22]	**0.72** (0.01)
		[23–50]	**0.48** (0.01)
xAI	Grok 2	[12–22]	**0.72** (0.01)
		[23–50]	**0.73** (0.00)
OpenAI	GPT4o	[12–22]	**0.81** (0.01)
		[23–50]	**0.66** (0.01)
DeepSeek	Chat	[12–22]	**0.73** (0.01)
		[23–50]	**0.46** (0.00)

Table F.2: Estimates of the Hype Susceptibility Parameters. The column *Rounds* indicates the sample period used for each estimation. Data from all simulations for variant 2 Table 3.

Provider	Model	Rounds	δ
Anthropic	Haiku 3.5	[12–22]	**0.56**
			(0.02)
		[23–50]	**0.77**
			(0.00)
Meta	LLaMa 3.3 70B Turbo	[12–22]	**0.99**
			(0.00)
		[23–50]	**0.51**
			(0.01)
Google	Gemini 2.0 Flash	[12–22]	**0.68**
			(0.01)
		[23–50]	**0.47**
			(0.01)
xAI	Grok 2	[12–22]	**0.61**
			(0.02)
		[23–50]	**0.52**
			(0.00)
OpenAI	GPT4o	[12–22]	**0.61**
			(0.02)
		[23–50]	**0.59**
			(0.00)
DeepSeek	Chat	[12–22]	**0.68**
			(0.02)
		[23–50]	**0.45**
			(0.00)

Table F.3: Estimates of the Hype Susceptibility Parameters. The column *Rounds* indicates the sample period used for each estimation. Data from all simulations for variant 3 Table 3.

Provider	Model	Rounds	δ
Anthropic	Haiku 3.5	[12–22]	**0.56** (0.02)
		[23–50]	**0.77** (0.00)
Meta	LLaMa 3.3 70B Turbo	[12–22]	**0.98** (0.01)
		[23–50]	**0.51** (0.01)
Google	Gemini 2.0 Flash	[12–22]	**0.63** (0.01)
		[23–50]	**0.47** (0.00)
xAI	Grok 2	[12–22]	**0.60** (0.01)
		[23–50]	**0.56** (0.00)
OpenAI	GPT4o	[12–22]	**0.64** (0.01)
		[23–50]	**0.57** (0.00)
DeepSeek	Chat	[12–22]	**0.67** (0.01)
		[23–50]	**0.45** (0.00)

Table F.4: Estimates of the Hype Susceptibility Parameters. The column *Rounds* indicates the sample period used for each estimation. Data from all simulations for variant 4 Table 3.