I Think, Therefore I Speculate: Reasoning Induced Bubbles in LLM Financial Markets *

Maxime Saxena[1], Marco Pangallo[2], Fabio Caccioli[1], and R. Maria del Rio-Chanona[1,3,4]

[1]Department of Computer Science, University College London, UK

[2]CENTAI Institute, Turin, Italy

[3]Complexity Science Hub, Vienna, Austria

[4]Bennett Institute for Public Policy, University of Cambridge, UK

February 2, 2026

Abstract

Accurately modelling human economic behaviour remains a central challenge in economics. Large language models have emerged as a potential solution, yet in laboratory asset markets experiments where humans consistently generate speculative bubbles, basic LLMs lack reasoning depth to replicate these patterns while frontier models prove too rational, converging toward equilibrium rather than generating bubbles. We replicate a canonical learning-to-forecast experiment using ten LLMs spanning multiple capability levels. We find a sweet spot of intermediate reasoning capability: models such as Qwen3-14B and OpenAI's o3 produce bubbles with mispricing and volatility comparable to humans, while both simpler and more advanced models converge to fundamentals. Toggling reasoning on and off within a model establishes that reasoning causally drives bubble formation. Analysis of agent justifications reveals the mechanism: bubble-forming models engage in non-linear trend extrapolation, whereas the most capable model anchors predictions to fundamentals. These findings suggest that intermediate-capability reasoning LLMs may serve as surrogates for boundedly rational human traders, with implications for economic modelling and market stability as AI agents increasingly participate in financial systems.

*Disclosure statement: The authors declare no financial or other conflicts of interest related to this research.

1 Introduction

A central challenge in economics is modelling human reactive behaviour, since economic outcomes are inextricably linked to how individuals adapt their expectations amidst shifting policies and market environments [Lucas Jr, 1976]. The rational expectations hypothesis (REH), which assumes that agents form expectations consistent with the model being used, has emerged as the dominant modelling response [Muth, 1961, Sargent, 1987]. REH provides a tractable framework, but evidence from behavioural economics demonstrates that the assumption is often wrong [Tversky and Kahneman, 1974]. For example, in laboratory market experiments where participants repeatedly forecast asset prices, they consistently coordinate on trend-following strategies rather than anchoring predictions to fundamental values; this generates speculative price bubbles that rational expectations cannot explain [Hommes et al., 2008, Bao et al., 2020, Hoyer et al., 2023]. These findings leave modellers caught between two unsatisfying options: rational expectations, which is tractable but unrealistic, or the vast space of possible behavioural rules, what Sims [1980] termed the "wilderness of bounded rationality" [1] where any departure from perfect rationality opens an intractable set of alternatives.

Large language models (LLMs) offer a potential path through this wilderness. Because LLMs process information and generate context-dependent responses, they can adapt to changing conditions. Unlike behavioural rules specified by modellers, which face criticism as ad hoc, LLMs have been trained on human-generated text that includes both descriptions of human behaviour and economic theory. This training may provide grounding for agent behaviour that is neither the perfect foresight of rational expectations nor arbitrary modeller assumptions. Researchers have begun exploring whether LLMs could serve as surrogates for human behaviour, starting by replicating participants in economic experiments, with the promise of lower costs, scalability, and reproducibility [Horton, 2023, Grossmann et al., 2023, Dillion et al., 2023].

Initial studies focused on single-shot games such as the Ultimatum Game, Prisoner's Dilemma, and Dictator Game, finding that LLMs can reproduce several aspects of individual human economic behaviour [Filippas et al., 2024, Kitadai et al., 2023, Mei et al., 2024, Brookins and DeBacker, 2023]. More recent work examines dynamic settings where multiple LLM agents interact over time. Chen et al. [2023] simulate auction markets, while del Rio-Chanona et al. [2025] study positive and negative feedback markets and find that earlier LLMs (GPT-3.5, GPT-4) can distinguish between market types and exhibit bounded rationality, though with less behavioural heterogeneity than humans. These studies establish that LLMs can capture some market dynamics, but leave open the question of how well they match human behaviour quantitatively.

A consistent finding is that models' reasoning capabilities matter, but not in a straightforward way. On one hand, less capable models without advanced reasoning capabilities

[1] The original term was the wilderness of "disequilibrium economics", but the term has evolved to bounded rationality since.

appear less able to capture more advanced human thinking patterns. For example, Gao et al. [2025] study the 11-20 game with GPT-3.5, GPT-4 and models from the Claude-3 and LLaMA-3 families, finding that these models exhibit shallower levels of strategic reasoning than humans. Less capable models also show a muted tendency to form bubbles in asset market experiments, possibly because they have less sophisticated extrapolation capabilities than humans [Henning et al., 2025, Muntoni, 2025]. On the other hand, the most capable models appear excessively rational. Wang et al. [2025] find that the latest frontier LLMs, equipped with the most advanced reasoning capabilities, act as fundamentalists rather than trend-followers. These mixed findings suggest a puzzle: the challenge is finding a capability level that matches human bounded rationality, a sweet spot between models that reason too little and those that reason too much.

This puzzle motivates our focus on a range of reasoning models, a new class of LLMs that output explicit chains of thought before providing final answers [OpenAI, 2024]. We reproduce the canonical learning-to-forecast experiment by Hommes et al. [2008], in which participants predict asset prices over 50 periods in a positive feedback market. This experiment is well-suited for our purposes: it is a benchmark in experimental economics with detailed human data for comparison, and its design allows for both fundamentalist strategies (anchoring to the rational expectations equilibrium) and trend-following strategies that generate bubbles. We study a wide range of models including both open-weight (Qwen, OLMO, Gemma) and closed-weight (OpenAI) LLMs. Open-weight models are particularly valuable because the model weights, and therefore model responses, are free and they give guaranteed replicable results, unlike closed-weight models.

Under appropriate conditions, we find that reasoning LLMs exhibit bounded rationality and form speculative bubbles similar to human participants. Our findings can be summarised as follows. First, reasoning capability is associated with bubble formation: models such as Qwen3-14B and OpenAI's o3 form bubbles with mispricing statistics similar to human participants, while non-reasoning models do not. Using Qwen3-14B, where reasoning can be toggled on and off, we establish that this relationship is causal. Second, by analysing the justifications of LLM agents, we identify non-linear extrapolation as a key mechanism: reasoning models that form bubbles show high rates of non-linear trend extrapolation (85% for Qwen3-14B with reasoning versus 35% for the same model without reasoning). In contrast, the most advanced model we test (GPT-5-mini) takes a fundamentalist approach and does not form bubbles. Non-reasoning models are not found to be fundamentalists, but nor do they make price predictions using non-linear extrapolation, leading to greater mispricing than GPT-5-mini, without forming bubbles similar to humans. Third, markets with mixed-ability agents, combining reasoning and non-reasoning models, produce less predictable and more diverse outcomes than homogeneous markets, giving a potential solution to the lack of heterogeneity found with LLM agents in previous studies.

These findings have implications for both economic modelling and market stability. For modelling, our results suggest that reasoning LLMs at intermediate capability levels may serve as useful surrogates for human participants, occupying a middle ground between the excessive rationality of frontier models and the limited strategic depth of earlier LLMs.

This offers a potential route to incorporating bounded rationality into economic models without requiring researchers to specify behavioural rules ex ante. For market stability, our results matter because LLMs are increasingly deployed in financial applications [Zhao et al., 2025]. Understanding how LLM agents behave in market settings is important not only for validating their use as human surrogates, but also for anticipating how markets may behave as AI systems become participants in their own right.

2 Methods

2.1 A laboratory asset market with positive feedback

Our basis is the market bubble experiment with human participants done by Hommes et al. [2008]. The experiment simulates an artificial asset market with two key components: a fundamental price determined by the asset's economic fundamentals, and a realized market price that depends on participants' expectations. Six participants interact in each market over 50 periods. The realized price at time t, p_t, is determined by a market equilibrium equation

$$p_t = \frac{1}{1+r}\left(\frac{1}{6}\sum_{h=1}^{6} p^e_{h,t+1} + \bar{y}\right), \tag{1}$$

where $p^e_{h,t+1}$ represents participant h's price prediction for period $t+1$, r is the risk-free interest rate (5% per period), and $\bar{y}$ is the mean dividend (3 guilders per period). This creates a positive feedback market: when participants predict higher future prices, the realized current price increases accordingly. The market has a theoretical fundamental price of $p^f = \bar{y}/r = 60$, representing the present value of expected future dividends. This fundamental price is only reached if participants coordinate their expectations on it. However, as Hommes et al. [2008] found, participants often fail to coordinate on this fundamental price and instead form speculative bubbles where prices rise substantially above fundamental values. To prevent prices from growing without bound, predictions are capped at a maximum of 1000, though participants only discover this constraint if they attempt to exceed it.

Each participant plays the role of a financial advisor to a pension fund and at time t must predict the risky asset's price for time $t+1$ (participants at time t only know realised prices up to time $t-1$ when making their predictions, making this a two time step prediction). Participants are incentivized through earnings that depend inversely on their prediction errors. The reward participant h earns in period t is

$$e_{ht} = \max\left\{1300 - \frac{1300}{49}(p_t - p^e_{ht})^2, 0\right\}, \tag{2}$$

where earnings decrease quadratically with forecast error and reach zero when errors exceed 7. Participants are told their predictions influence their pension fund's demand for the asset (higher predictions lead to greater demand), and that market prices result from the equilibrium between aggregate demand and fixed supply.

At time t, participants have access to all realized prices up to period $t-1$, their own predictions up to period t, their earnings from period $t-1$, and their cumulative earnings. They also know the market fundamentals: the mean dividend ($\bar{y} = 3$) and risk-free rate ($r = 0.05$), which in principle allows for coordination on the fundamental price. The initial price predictions (periods 1 and 2) are made with the guidance that prices will likely be between 0 and 100. However, participants do not know the exact market equilibrium equation, how many other participants are in the market, or other participants' predictions, strategies and earnings. Participants have no communication with each other throughout the experiment.

2.2 From human to LLM participants

We conduct the experiments with LLMs using a combination of locally run LLMs with open weights [Wolf et al., 2020], the OpenAI API and the Google API. Each time step occurs sequentially, with all 6 agents giving their predictions at time t for time $t+1$ before the realised price and earnings for time t are calculated and the next time step occurs.

2.2.1 LLM agent settings

We detail some of the key parameters in the translation to LLMs, how LLMs were queried and the settings that were used in our experiments.

Model: at the core of an LLM agent is a model such as GPT-5-mini or Qwen3-14B. We consider a wide range of models, with sizes ranging from 7 billion parameters up to the extremely large models, such as those from OpenAI where the exact parameter counts are undisclosed but are most likely 200 billion or more [Epoch AI, 2025, Erdil, 2024]. We also consider a range of LLM families, including Qwen (Alibaba), Gemma (Google), GPT (OpenAI), o series (OpenAI) and OLMO (Ai2). The Qwen, Gemma and OLMO families are open-source while the others are not. The models we consider in this paper are listed in Table 1, along with the names we use to refer to them from here on and whether they have reasoning capabilities. GPT-5-mini can output low, medium or high amounts of reasoning, and throughout this paper we use the medium reasoning setting.

Temperature: the temperature parameter of an LLM controls how random the model's response is. Lower values, near 0, are closer to deterministic outputs, while higher numbers, typically near 1, are used for more randomness and diversity of output. We use a temperature of 1 throughout this paper. This is to allow diverse LLM outputs, which may aid in replicating human diversity, and also to ensure consistency across models as some GPT models only accept this value of the parameter.

Memory: in these experiments, we define memory as the number of an agents' own prior responses it sees as context for its next price prediction. For example, a memory of 2 would mean that the agent can see its last two price predictions, along with its thinking

Model	Name in paper	Provider	Reasoning
Qwen3-14B	Qwen3	Alibaba	Yes
Qwen3-14B (reasoning disabled)	Qwen3-NR	Alibaba	No
Qwen2.5-7B-Instruct	Qwen2.5	Alibaba	No
OLMO3-7B-Think	OLMO-Think	AI2	Yes
OLMO3-7B-Instruct	OLMO-Instruct	AI2	No
Gemma3-27B	Gemma	Google	No
GPT-4.1	GPT-4.1	OpenAI	No
GPT-4o-mini	GPT-4o-mini	OpenAI	No
o3-mini	o3-mini	OpenAI	Yes
o3	o3	OpenAI	Yes
GPT-5-mini	GPT-5-mini	OpenAI	Yes

Table 1: **Large language models used in this paper.** Qwen3-14B and Qwen3-14B (reasoning disabled) are identical models but we disable reasoning for the latter. Reasoning models are highlighted in blue.

where applicable, as well as the prior 2 user prompts. This provides the agent with some context and continuity in methodology, in keeping with the way humans can remember how they made prior predictions. We use a value of 2 throughout these experiments. Other values could be considered but an exploration of the effects of memory is out of scope for this work (see e.g. del Rio-Chanona et al. [2025]).

Seed: this parameter is designed to allow reproducibility and encourage transparency. LLM outputs are generally stochastic but setting the seed makes it possible to reproduce this randomness. Open-source models are all run with predetermined seeds which change across runs and which also change for every agent and every time step. In the OpenAI API, setting the seed is said to give "(mostly) deterministic outputs across API calls", making these results most likely replicable despite OpenAI stopping short of giving a full guarantee of reproducibility [OpenAI, 2026]. The Google API does not allow reproducibility through the seed parameter. We use the Google API for Gemma queries only.

2.2.2 Prompts

We keep instructions for the LLMs as similar as possible to those used in the human experiments. However, some minor changes are required. To interact with the LLMs in a systematic manner, we include specifications in the system prompt and each user prompt to reply in JSON format. Throughout, we design the prompts to provide clear instructions, and also to allow the LLMs space to "think" (even if they are not explicitly finetuned to output thinking tokens). This second point is achieved by specifying a "justification" key in the JSON output in which the LLM is asked to justify its price prediction. A useful side-effect of this is that we also have a logic trail for LLM agent price predictions throughout, giving us useful insights into why LLM agents output the prices they do. Another consideration is the amount of tabular data communicated to the agent at each time step as they receive the

prior prices and their prior predictions. To aid LLM understanding of the data, we format this information as a markdown table [Fang et al., 2024].

The system prompt is identical to the explanatory text given at the start of the original experiment except for added instructions on the response format, asking that responses are in JSON format and the keys that should be used. Then, for each time step a user prompt is also created which reflects as closely as possible the information given on-screen to human participants. The only addition is a re-iteration of the JSON formatting request. The original experiments instructions in Hommes et al. [2008] are in Appendix A.1 while details of the prompts given to LLMs are in Appendix A.2.

While most experiments are run with neutral prompts which closely follow the original experiment, we also explore how changing the prompting strategy of an LLM impacts the formation of bubbles. We introduce a prompt encouraging non-linear thinking and prediction strategies. The full prompt encouraging non-linear thinking is in Appendix A.2. However, unless otherwise specified, all experiments are run with the prompt most closely resembling the original experiment.

2.2.3 Experiment flow

As in the original experiments by Hommes et al. [2008], each market has 6 agents and the experiment is run over 50 time steps throughout this paper. At the initial time step, agents are informed that the price is likely in-between 0 and 100, in keeping with the original experiment. For this first time step alone, agents are also asked to predict the first two prices, after which price predictions return to only forecasting one price. The original experiment also imposes a maximum prediction of 1000, although participants are only told about this artificial cap on predictions if they try to predict above the cap. We implement this as well, adding an extra note to the system prompt that predictions must be less than or equal to 1000 if and only if an agent attempts to predict above the cap. The sequence of each time step t is therefore:

1. **Instructions and context:** each LLM agent is given the experiment instructions, in the form of a system prompt, as well as their own output from the last two time steps according to the memory parameter of 2.
2. **Prompt:** each agent is also given a new user prompt, which includes information such as prior realised prices, their prior predictions, their earnings at the last time step and their total earnings so far.
3. **Prediction:** using the instructions, context and prompt, each LLM then outputs a price prediction for time step $t+1$, as well as a justification for this prediction in JSON format.
4. **Market price and earnings calculations:** using the price predictions from all agents for time step $t+1$, the realised price for time step t is calculated according to

Eq. 1. With the realised price for time step t and the predicted prices for time step t (predicted at time step $t-1$) we then calculate earnings for each agent according to Eq. 2.

2.3 Bubble formation measures

To quantify the extent of bubble formation and mispricing in experimental markets, we employ several standard measures from the experimental asset market literature. These measures capture different aspects of how prices deviate from fundamental values: the direction and magnitude of mispricing, the peak amplitude of bubbles, and overall price volatility.

We use two primary sets of measures to assess mispricing with respect to the fundamental value. The first set consists of Relative Deviation (RD) and Relative Absolute Deviation (RAD) [Stöckl et al., 2010], which measure the average percentage deviation of prices from the fundamental value (value of 1 is a 100% average mispricing). RD captures systematic overpricing or underpricing (positive values indicate overpricing, negative values underpricing), while RAD measures the absolute magnitude of mispricing regardless of direction. We complement these with their numeraire-independent equivalents, Geometric Deviation (GD) and Geometric Absolute Deviation (GAD) [Powell, 2016], which use geometric rather than arithmetic means and are therefore invariant to the currency units used.

To capture the maximum extent of bubble formation, we use Relative Deviation Maximum (RDMAX), defined as the peak percentage deviation from fundamental value observed during the experiment [Razen et al., 2017, Hoyer et al., 2023, Jiao et al., 2025]. This measure is particularly useful for identifying the amplitude of price bubbles. Finally, to assess overall price variability and volatility, we calculate the standard deviation and interquartile range of realized market prices [Hanaki et al., 2023]. Formal definitions of all measures are provided in Appendix A.3.

We identify bubbles using two complementary binary indicators. The first, P^{mean}_{bubble}, following Kopányi-Peuker and Weber [2021, 2024], classifies an experiment as containing a bubble if the mean price over all periods exceeds twice the fundamental price. The second indicator, P_{bubble}, which we introduce, defines a bubble as sustained overpricing where the market price remains above five times the fundamental value (300) for at least five consecutive periods. Using two measures increases robustness, as there is no universally accepted definition of experimental bubbles and each captures different aspects of bubble dynamics. Further discussion is provided in Appendix A.4 where these two measures are shown to be robust across different parameter values and to agree with each other, with Cohen's kappa scores of 0.7 or above in all comparisons.

We characterise bubbles using three dimensions: speed of formation, amplitude, and burst dynamics. When experiments produce multiple bubbles, we focus our analysis on the first bubble, as this captures the initial market dynamics before participants learn from the burst. To measure these, we first identify the bubble peak as the first local maximum within

a seven-period window exceeding the threshold of 300 . The bubble start is defined as the earliest period where all subsequent prices up to the peak remain above the fundamental value.

Based on these definitions we calculate: (i) time to form ($T_{bottom \to peak}$), the number of periods from bubble start to peak; (ii) amplitude (AMP), defined as $(p^{B*} - p^f)/p^f$ where p^{B*} is the peak price with respect to the fundamental price; and (iii) half-life (HALF), the number of periods after the peak for prices to fall below half the peak value. These measures allow us to compare how quickly bubbles form, how extreme they become, and how rapidly they burst across different models.

By analyzing whether bubbles form and how they grow, we can assess whether LLM agents behave consistently with rational expectations. Under rational expectations, the market equilibrium admits two types of solutions. The first is the steady state where prices remain at the fundamental value $p^f = 60$ for all periods, which is obtained if agents coordinate on the "no bubble" condition (the expected present value of future prices tends to zero as time goes to infinity). The second type are rational bubbles, where prices grow exponentially at the rate of return: $p_t = p^f + R^t c$, where $R = 1 + r = 1.05$ and c is any positive constant [Hommes et al., 2008]. Rational bubbles satisfy rational expectations because if all agents expect this growth rate, it becomes self-fulfilling.

We distinguish rational bubbles from speculative bubbles (driven by boundedly rational expectations) by testing the growth rate. For a rational bubble, the log deviation from fundamentals grows at a constant rate: if $q_t = \ln(p_t - p^f)$, then $q_{t+1} - q_t = \ln R \approx 0.049$. Following Hommes et al. [2008], we calculate $q_{t+1} - q_t$ from the bubble start to the first deceleration point before the peak (where $p_t - p_{t-1} \leq p_{t-1} - p_{t-2}$). We then conduct a one-sided t-test with $H_0 : q_{t+1} - q_t \leq \ln R$ against $H_1 : q_{t+1} - q_t > \ln R$. Rejection of the null hypothesis indicates a speculative bubble growing faster than rational expectations would predict, consistent with boundedly rational trend extrapolation by agents.

2.4 Analysis of price prediction strategies

We use two methods to analyse the prediction strategies of each LLM agent. The LLM agents output as a final answer a JSON including both a price prediction and a justification for that price prediction. The first method analyses the written justification given by the agent for its price prediction. The second method looks directly at the outputted price prediction and considers whether a linear or a non-linear extrapolation method best explains the agent's price prediction.

2.4.1 Analysis of price prediction justifications

We use Gemma to classify the justifications of each agent and each time step into two separate binary indicators for: 1) if the agent uses non-linear reasoning in their prediction and 2) if the agent considers a fundamental value when making their prediction.

To answer if non-linear extrapolation is used, we ask Gemma to determine if there is evidence of "clear and explicit" use of non-linear extrapolation in the agent's justification. The examples given are: quadratic or higher-order polynomials, exponential models, reasoning about returns or percentage changes in price rather than absolute price changes and reasoning about rates of change in price differences. The full prompt is shown in Appendix A.5. We do not consider the first three time steps as there would not be enough data in these time steps to validly use non-linear methods. For runs with bubbles, we only consider justifications before the peak of the first bubble as after this, the limit of 1000 is introduced and could skew price prediction methods. For non-bubble experiment runs, we consider only up to time step 12 as this around when the average bubble peaks, ensuring a fair comparison between bubble and non-bubble experiments. We validate the classifier by hand-annotating 200 randomly selected justifications from across models, and find that the LLM classifier gives an accuracy of 0.86 against human-labelled ground truth. Further details, including precision, recall and F1 scores, are given in Appendix A.6.

A similar method is used to classify if LLM agents show evidence of using the fundamental price in the price prediction. We ask Gemma to classify if the agent "clearly and explicitly anchors their prediction around what they consider is a fundamental asset price". We focus only on the first 5 time steps for this classification. This classifier is validated in the same way as before, giving an accuracy of 0.96. Further details are in Appendix A.6.

2.4.2 Quantitative analysis of price predictions

We also classify the methods used for price prediction by each agent by answering the question: given the same information as the agent, does a non-linear model best explain the agent's price prediction? To do this, for every agent we consider predictions made after time step 5 and before the peak of the bubble, if a bubble is formed, or otherwise before time step 12. We take the last 5 realised prices available to the agent (from $t-6$ to $t-1$) and fit a linear, quadratic, cubic and exponential curve to these 5 prices. We also consider a naive prediction method, which simply takes the last price available at $t-1$. These 5 models are then used to predict the price at time $t+1$, and the model which is closest to the agent's actual prediction for $t+1$ is taken as the "winner". If the winner is a quadratic, cubic or exponential model, then we conclude that the agent most likely used a non-linear method for their price prediction, and we therefore classify that agent as having used non-linear reasoning at that time step.

More discussion of a comparison between the two classification methods is given in Appendix A.7. We find that although the two methods for classifying whether an agent uses

non-linear extrapolation do not always agree on individual points (Cohen's kappa of 0.21), they do both show the same conclusions, as discussed in Section 3.3. Using both methods significantly increases the robustness of our findings.

3 Results

3.1 Comparing bubble formation across models

We find that reasoning is associated with bubble formation but is not sufficient alone in determining if agents will form bubbles. A selection of experimental results across reasoning and non-reasoning LLM agents, as well as human participants, is presented in Figure 1.

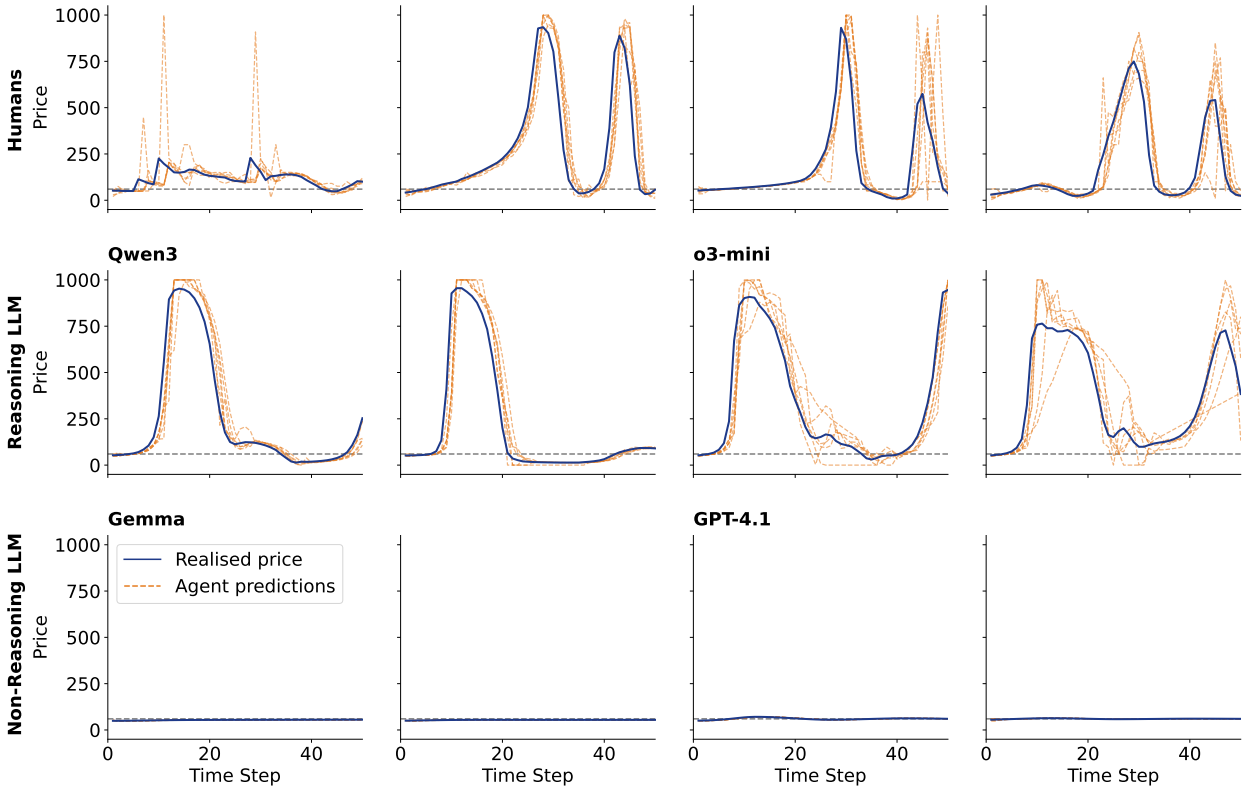


Figure 1: **Comparing results across humans and LLM agents.** Human results are from Hommes et al. [2008]. The top row shows 4 of the 6 experiment runs with humans. The LLM experiments show two results per model. Reasoning models show results that are qualitatively more similar to humans than non-reasoning models.

With this experimental setup, prices with both human agents and reasoning LLM agents fluctuate wildly, with peak prices reaching well above the fundamental price, $p^f = 60$, before crashing down. In most cases where a bubble is formed, the peak approaches the artificial limit price of 1000 and after the crash most often drops to below the fundamental price. This

is in sharp contrast to non-reasoning LLM agents, where prices are much more stable around the fundamental price with little deviation compared to humans and reasoning agents. This makes the results with reasoning models qualitatively more like human results, as we can say that, with humans and reasoning agents, bubbles appear to form.

To quantify the degree of mispricing and the volatility of the prices for different agent types, we use the measures discussed in Section 2.3. These are calculated for a range of different LLMs, varying in size, provider and reasoning capabilities, as well as for the results with humans. The average measures across independent experimental repeats are presented in Table 2.

		Mispricing		Volatility	
Model	Rpts	RD	RDMAX	Std pr	IQR pr
OLMO-Think	17	6.08	14.90	353.74	666.37
o3-mini	3	5.24	13.80	289.42	508.76
Qwen3	20	3.25	14.32	301.37	279.40
Humans	6	2.40	12.16	215.63	213.77
o3	3	1.61	13.68	207.49	75.21
GPT-4o-mini	3	0.29	0.63	13.50	22.17
OLMO-Instruct	20	0.14	0.28	7.64	10.51
Qwen2.5	20	0.08	0.30	8.09	13.10
GPT-4.1	3	0.01	0.11	3.34	3.80
GPT-5-mini	3	-0.00	0.00	0.74	0.00
Gemma	18	-0.04	0.01	3.14	3.93

Table 2: **Averaged statistics across experimental repeats.** Models are sorted by descending average RD values. Rows highlighted in blue are reasoning models, while the row highlighted in yellow is the human results from Hommes et al. [2008]. Rpts is the number of independent experiment repeats. The mispricing measures Relative Deviation (RD) and Maximum Relative Deviation (RDMAX), are expressed as multiples of the fundamental price ($p^f = 60$); for example, RDMAX values around 14 indicate prices were overvalued by approximately 14 times the fundamental value at their peak. Std pr is the standard deviation of the realised prices while IQR pr is the interquartile range of the realised prices. A complete set of measures is shown in Appendix B.1. Most reasoning models develop bubbles similar to humans in our experiments.

The results divide the models into two groups, with OLMO-Think, o3-mini, Qwen3 and o3 all showing larger average mispricing and volatility than the others. These four models are more similar to the human results, where there is also a propensity towards overpricing and volatile prices, consistent with bubble formation. We note that **all models that show bubble price action are reasoning models** that have specifically been built or fine-tuned to output thinking tokens before eventually outputting their final answer. However, **not all reasoning models exhibit bubble formation behaviour**: GPT-5-mini, one of the latest models from OpenAI, is equipped with reasoning but does not form bubbles. It should be

noted that GPT-5-mini can use low, medium, or high amounts of reasoning, and for all experiments we use a medium amount of reasoning. Changing the setting of the amount of reasoning used could change the results with GPT-5-mini.

While the price action with reasoning LLM agents is in many ways qualitatively and quantitatively similar to the results with human agents (except for GPT-5-mini), there remain key differences. For example, some of the results with humans show fairly stable price action at the start of the experiment before a bubble is triggered. In contrast, the bubbles from LLMs appear to all be triggered almost immediately. The bubbles formed by humans also seem to burst faster than those formed by LLMs, with the latter exhibiting higher prices after the bubble peak for longer. We compare the bubbles formed by reasoning models with those formed by humans by quantifying the time to form the first bubble, its amplitude and its half life as well as if there is evidence of it being a speculative bubble. The average statistics are shown in Table 3.

		Bubble frequency		Bubble characteristics			
Model	Rpts	P_{bubble}	P^{mean}_{bubble}	$T_{bottom \to peak}$	HALF	AMP	SPEC
Qwen3	20	1.00	1.00	6.60	8.58	14.32	0.95
o3-mini	3	1.00	1.00	8.00	11.00	13.59	1.00
OLMO-Think	17	1.00	1.00	6.76	6.82	14.39	0.71
Humans	6	0.83	0.83	15.60	3.20	14.03	0.80
o3	3	0.67	1.00	10.00	3.00	14.29	0.50

Table 3: **Averaged statistics across repeats for models with bubbles.** Models are sorted by descending proportion of runs with a bubble, P_{bubble}, using the method introduced in this paper. The proportion using the bubble detection from Kopányi-Peuker and Weber [2021, 2024], P^{mean}_{bubble}, is added for robustness. Rpts is the number of independent experiment repeats. The time for the bubble to form ($T_{bottom \to peak}$), half-life of the bubble (HALF), amplitude of the bubble (AMP) and proportion of experiments with speculative bubbles (SPEC) are averaged across experiments only where a bubble is formed. Human results from Hommes et al. [2008] are highlighted in yellow. Compared to humans, LLMs tend to form bubbles faster, but these bubbles then burst more slowly.

While the average amplitude of the first bubble formed in LLM experiments is comparable to that of bubbles formed by humans, we see that LLMs on average form a bubble quicker than humans and, for most reasoning LLMs, the bubble bursts slower than with humans. The model which comes closest to reproducing the shapes of the human bubbles is o3, which has a time for the bubble to form of 10 time steps compared to humans at 15.6 and a bubble half-life of 3 time steps, similar to the 3.2 with humans. The smaller reasoning models, Qwen3, o3-mini and OLMO-Think, are all much faster to form bubbles but take from 6.82-11 time steps on average for the price to go below half its peak.

However, across all reasoning LLMs there is significant evidence at the 5% level of speculative bubbles in almost all bubbles that are produced. It should be noted that even in bubbles where there is insufficient evidence to reject the null hypothesis of a rational bubble,

this may be due to insufficient data (when a bubble forms too quickly) or that a rational bubble was truly formed. This finding is in keeping with the human bubbles in Hommes et al. [2008], where there was sufficient evidence in 4 of the 5 experiments with bubbles to conclude that the bubbles formed were speculative. This is significant as it suggests that bubbles formed by reasoning LLMs are consistent with boundedly rational price predictions by agents, in keeping with the bounded rationality exhibited by humans.

Overall, our results suggest that reasoning capabilities, except in the very most recent models, are associated with the formation of bubbles, making these results more similar to those from humans than those from LLM agents without reasoning capabilities. Further, when a bubble does form, there is evidence that reasoning LLMs follow boundedly rational prediction strategies, similarly to the bubbles formed by humans. However, reasoning models do not successfully replicate the time taken for a bubble to form or burst.

3.2 Impacts of reasoning on bubble formation

Our initial analysis showed that reasoning capabilities are often associated with bubble formation; we now attempt to show this relationship is causal. To isolate the impact of reasoning on bubble formation, we run experiments with Qwen3 and find evidence that reasoning causes bubble formation in Qwen3. Qwen3 is an ideal candidate for our tests because it is a reasoning model equipped with a mode in which the same model will not output any reasoning [Qwen, 2025]. Therefore, we can keep all other parameters identical and make causal claims about the impact of reasoning. The average summary statistics for Qwen3 with and without reasoning (Qwen3 and Qwen3-NR) are presented in Table 13.

		Mispricing		Volatility	
Model	Rpts	RD	RDMAX	Std pr	IQR pr
Qwen3	20	3.25	14.32	301.37	279.40
Qwen3-NR	40	0.44	0.86	17.10	27.40

Table 4: **Comparison of Qwen3 with and without reasoning.** Qwen3 is regular Qwen3-14B with reasoning, Qwen3-NR refers to the same model but with reasoning turned off. Reported statistics are averaged across the independent repeats. The reasoning model is highlighted in blue. Rpts is the number of independent experiment repeats. The mispricing measures Relative Deviation (RD) and Maximum Relative Deviation (RDMAX), are expressed as multiples of the fundamental price ($p^f = 60$). Std pr is the standard deviation of the realised prices while IQR pr is the interquartile range of the realised prices. A complete set of measures is reported in Appendix B.2. Switching on reasoning in Qwen3 causes bubbles to occur.

The Qwen3 model when reasoning is turned off has on average much lower mispricing, overpricing and volatility of price tendencies, in line with no bubbles forming and similar to the results of non-reasoning models such as GPT-4o-mini. To make this more concrete, we

perform a one-sided Mann-Whitney-Wilcoxon (MWW) test for each of the statistics in Table 13 with a null hypothesis that the medians of the underlying distributions are equal and the alternate hypothesis that the median of the statistic distribution is higher with reasoning than without. For added robustness and given we are performing several statistical tests, we also use a Holm correction and report the adjusted p values as well. All tests are found to reject the null hypothesis at the 0.001 level, both with and without the Holm correction (p-values reported in Appendix B.2). This offers strong evidence that the Qwen3 (with reasoning) on average has larger absolute deviations from the theoretical value, over-prices the asset more and has more price variability than Qwen3-NR (Qwen3 without reasoning). Since all other parameters are held equal, this suggests that the reasoning used by Qwen3 causally impacts the overpricing, which is indicative of bubble formation.

This result is supported by the results reported in Table 2 for OLMO-Think and OLMO-Instruct. These two models use the same base model but are fine-tuned in different ways, such that OLMO-Think outputs intermediate reasoning tokens, but OLMO-Instruct does not [Olmo et al., 2025]. These two models are therefore very closely related, although they are not identical. The OLMO-Think agents show on average higher mispricing, over-pricing and volatility in the realised prices compared to OLMO-Instruct. This supports the idea that reasoning causes bubble formation, although we cannot make as strong a claim with the OLMO models as with Qwen3 since the two OLMO models are not identical.

3.3 Non-linear thinking in reasoning models causes bubbles

By exploring the justifications of LLM agents, we find that a key difference between reasoning and non-reasoning models is the amount of non-linear extrapolation they do. The results of the classifications are given in Table 5. These results show a striking difference between results for Qwen3 and Qwen3-NR, with the reasoning model showing far more evidence of non-linear reasoning. Using the LLM classification method (Section 2.4.1), the reasoning model has 85.51% of its justifications showing evidence of non-linear extrapolation compared to only 34.72% for the non-reasoning model. The same pattern is shown using the quantitative classification method (Section 2.4.2). o3-mini also shows a high proportion, 89.74% with the LLM classifier, of its justifications including non-linear extrapolation methods. This gives evidence that a high proportion of non-linear extrapolation is associated with bubble formation.

To consider whether the non-linear thinking **causes** bubbles to form, we prompt Qwen3-NR (Qwen3 with reasoning turned off) to "take its time", "think step-by-step" and to use non-linear methods and reasoning. With this altered prompt (see Appendix A.2) and all else held equal, we find that the propensity of the non-reasoning model to form bubbles increases significantly. The average bubble measures are reported in Table 6. Using the MWW test as before, we find that for the non-reasoning model, nonlinear prompting gives a higher average mispricing and higher price variation than regular prompting, significant at the 0.01 level, even with a Holm correction. This is in line with more bubbles being produced. Details of

Model	Rpts	Non-linear (Quant) (%)	Non-linear (%)	Fundamental (%)
Qwen3	20	81.40	85.51	12.50
o3-mini	3	80.00	89.74	6.41
Qwen3-NR	40	47.08	34.72	6.50
GPT-5-mini	3	6.67	25.00	88.89

Table 5: **Classification of price predictions of agents.** Rpts is the number of independent experiments run. Quant refers to the quantitative classification method while the two remaining classifications use the LLM classifier. The non-linear and fundamental classifications are not mutually exclusive, as an agent could use neither method (a linear trend follower for example) or both (agent weights prediction between non-linear trend following and anchoring to fundamental value for example). The classifications are also done on different time steps, as per Section 2.4. This means that the percentages do not sum to 100. Reasoning models are highlighted in blue. Reasoning allows non-linear extrapolation.

the p-values are reported in Appendix B.3.

The non-linear prompt does seem to move the non-reasoning agents closer to the reasoning agents, although there is still some differences between the two. Using once again a one-sided MWW test, we test if Qwen3 (reasoning model with normal prompting) gives on average higher mispricing statistics than Qwen3-NR with non-linear prompting. We **do not** find sufficient evidence to reject the null hypothesis at a 0.01 level for RAD, GAD, RD and GD. In other words, we do not find sufficient evidence to say that Qwen3 produces higher values for these measures on average compared to Qwen3-NR with non-linear prompting. The exception to this is RDMAX, where there is evidence at the 0.01 level to suggest that Qwen3 gives higher values on average than Qwen3-NR with non-linear prompting. Further details of p-values are in Appendix B.3. The remaining differences, for example in RDMAX, may be because the reasoning model still has more space to think and therefore is still better able to use non-linear reasoning. There may also be other factors in play, such as for example a better understanding of the experiment setup. Most likely, it is a combination of several factors. However, with Qwen3 we find evidence that non-linear thinking causally increases the likelihood of observing bubbles in this experiment.

Analysing the justifications also suggests why GPT-5-mini does not form bubbles, in contrast to another advanced OpenAI reasoning model such as o3-mini. Table 5 shows that GPT-5-mini has a far higher propensity to consider the fundamental asset value (88.89% of the time) in the first 5 time steps than o3-mini, Qwen3 or Qwen3-NR (6.41%, 12.5% and 6.5% respectively). GPT-5-mini also has the lowest evidence of non-linear extrapolation of the four models, at only 25% of the justifications analysed using the LLM classifier. GPT-5-mini acts more like a fundamentalist than the other models, meaning that prices remain more anchored to the fundamental value and bubbles do not form. GPT-5-mini's fundamentalist strategy in this experiment also ties in with the results found by Wang et al. [2025], where the most advanced reasoning models were found to predict very near the fundamental value and therefore did not form bubbles. The behaviour of GPT-5-mini is notable as it suggests

			Mispricing		Volatility	
Model	Prompt	Rpts	RD	RDMAX	Std pr	IQR pr
Qwen3	normal	20	3.25	14.32	301.37	279.40
Qwen3-NR	nonlinear	40	2.73	6.61	135.61	206.06
	normal	40	0.44	0.86	17.10	27.40

Table 6: **Qwen3 without reasoning but with nonlinear prompt gets closer to Qwen3 with reasoning.** Qwen3 is regular Qwen3 with reasoning, NR refers to the same model but with reasoning turned off. Rpts is the number of independent experiment repeats. The mispricing measures Relative Deviation (RD) and Maximum Relative Deviation (RDMAX), are expressed as multiples of the fundamental price ($p^f = 60$). Std pr is the standard deviation of the realised prices while IQR pr is the interquartile range of the realised prices. A complete set of measures is given in Appendix B.3.

that the most advanced reasoning models give more stable price action, which is good for the stability of systems where they operate, but does not replicate the human propensity to form bubbles as well as less advanced reasoning models.

We conclude that non-linear extrapolation appears to play a key role in the emergence of bubbles, and reasoning models generally do more non-linear extrapolation. The exception to this is GPT-5-mini, which appears to take a more fundamentalist approach compared to the other models. This is most likely why GPT-5-mini does not form bubbles, and highlights that different LLMs use different prediction styles.

3.4 Mixing models increases heterogeneity

Thus far, we have only considered the setting where all 6 agents in an experiment use the same LLM to output their price prediction; we now consider the case with different LLMs-agents within a single experiment run. We focus on differing reasoning capabilities within the 6 agents by considering mixtures of Qwen3 and Qwen3-NR (i.e. with and without reasoning), as well as a mixture between Qwen3 (with reasoning) and Gemma. For example, we might run the asset market experiment with two Qwen3-NR agents alongside four Qwen3 agents. We call markets in which agents have differing reasoning capabilities a "mixed market" and markets where all agents have the same underlying LLM and reasoning capabilities a "homogenous market".

We present a selection of measures in Table 7 capturing average mispricing, volatility, propensity to form bubbles and the shape of the bubbles formed. For the most part, mixed markets fall between the results for homogenous markets with reasoning agents and homogenous markets with non-reasoning agents in terms of their propensity to misprice and the volatility of the prices. This makes sense as there is likely a mixture of different strategies for price prediction in mixed markets as a result of the mixed agent types. The exception

		Mispricing	Volatility	Bubble formation & shape		
Model	Rpts	RD	Std pr	P_{bubble}	$T_{bottom \to peak}$	HALF
1 Qwen3-NR vs 5 Qwen3	20	3.51	273.09	0.90	7.47	10.53
Qwen3	20	3.25	301.37	1.00	6.60	8.58
Humans	6	2.40	215.63	0.83	15.60	3.20
2 Qwen3-NR vs 4 Qwen3	20	2.39	149.95	0.45	9.00	11.88
3 Qwen3-NR vs 3 Qwen3	20	1.68	98.27	0.25	11.75	9.00
3 Gemma vs 3 Qwen3	18	1.09	83.43	0.22	13.50	7.33
Qwen3-NR	40	0.44	17.10	0.00	-	-
Gemma	18	-0.04	3.14	0.00	-	-

Table 7: **Averaged statistics across experimental repeats.** Models are sorted by descending average RD values. Rows highlighted in green are mixed markets, while the row highlighted in yellow is the human results from Hommes et al. [2008]. All other rows are homogenous markets. Rpts is the number of independent experiment repeats. Relative Deviation (RD), is expressed as a multiple of the fundamental price ($p^f = 60$). Std pr is the standard deviation of the realised prices while P_{bubble} is the proportion of experiments with bubbles, using our binary indicator. $T_{bottom \to peak}$ and HALF are the average time to form a bubble and half-life of the bubble, calculated only when a bubble is formed. A full set of measures is given in Appendix B.4. Mixed markets tend to sit between homogenous non-reasoning and homogenous reasoning markets.

to this is the mixture of one Qwen3-NR with five Qwen3 reasoning agents, which exhibits higher average overpricing, according to the RD measure. This can be explained by using the bubble shape measures, such as the time for a bubble to form, $T_{bottom \to peak}$, and the half-life of the bubble, HALF. We note that 90% of the runs with one Qwen3-NR and five Qwen3 agents form bubbles, but that on average these bubbles take longer to form and to burst than homogenous markets with Qwen3 agents, as they have larger average values of $T_{bottom \to peak}$ and HALF. This explains why on average this mixed combination of agents has a larger tendency towards over-pricing than 6 homogenous Qwen3 agents.

We also note from Table 7 that two Qwen3-NR agents with four Qwen3 agents show similar average statistics to the human results in RD (and similarly for RAD, GD, and GAD; see Appendix B.4). However, a closer look at P_{bubble}, $T_{bottom \to peak}$, and HALF reveals differences between this mixed market and human results. The mixed market forms bubbles only about half as often as humans, it forms them more quickly on average, but once bubbles reach their peak, they take longer on average to burst. In summary, while a mixture of two Qwen3-NR and four Qwen3 agents matches human results in terms of average mispricing, it differs in both how often bubbles form and what shape they take.

Mixing agents with different reasoning abilities also increases the heterogeneity of results. Table 8 shows that key measures of overpricing (RD) and the shape of the bubbles when formed ($T_{bottom \to peak}$ and HALF) all have a larger standard deviation across experimental runs for mixed markets. In other words, there is more variation between independent

runs of the experiment when the agents have mixed reasoning capabilities.

Model	Rpts	Std RD	Std $T_{bottom \to peak}$	Std HALF
2 Qwen3-NR vs 4 Qwen3	20	2.92	1.00	6.64
3 Qwen3-NR vs 3 Qwen3	20	2.57	2.99	2.00
3 Gemma vs 3 Qwen3	18	2.12	2.38	0.58
1 Qwen3-NR vs 5 Qwen3	20	1.94	1.84	5.58
Humans	6	1.19	7.54	0.45
Qwen3	20	0.72	1.76	1.50
Qwen3-NR	40	0.29	-	-
Gemma	18	0.04	-	-

Table 8: **Standard deviation of statistics across experimental repeats.** Models are sorted by descending Std RD values. Rows highlighted in green are mixed markets, while the row highlighted in yellow is the human results from Hommes et al. [2008]. All other rows are homogenous markets. Rpts is the number of independent experiment repeats. STD RD is the standard deviation of the Relative Deviation over experimental repeats. Std $T_{bottom \to peak}$ and Std HALF are the standard deviations of the time for a bubble to form and the bubbles half-life respectively. Mixed market produce more diverse outcomes than homogenous ones.

We find that mixed markets yield more diverse and less predictable outcomes than homogeneous ones. This is important since, outside experimental settings, deployed AI agents are likely to interact as heterogeneous populations, making emergent behaviour harder to anticipate. We also highlight the importance of using diverse measures when evaluating experimental asset markets, as general measures such as mispricing and volatility can be misleading. We show an example where a specific mixture of agents appears similar to humans on the surface but differs markedly in bubble frequency and shape. This has important methodological implications for the study and design of experimental asset markets.

4 Discussion and Conclusion

Our results show that reasoning LLMs at intermediate capability levels can replicate human-like bounded rationality and speculative bubble formation in asset market experiments. We find a strong association between reasoning capabilities and bubble formation, and establish this relationship is causal using Qwen3-14B where reasoning can be toggled on and off. Analysis of agent justifications reveals that reasoning models act like chartists: it is not extrapolation per se, but non-linear extrapolation that drives bubble formation. However, reasoning does not guarantee bubbles. The most advanced model we test (GPT-5-mini) takes a fundamentalist approach and produces stable price dynamics, consistent with the heterogeneous agent model literature where fundamentalists stabilise markets and chartists amplify deviations [Hommes, 2006]. We also find markets with mixed-ability agents produce less predictable outcomes and a wider range of results than homogeneous markets.

These findings are consistent with prior work showing that LLMs exhibit bounded rationality rather than strict rational expectations [del Rio-Chanona et al., 2025], and that the state of the art models behave more like textbook-rational agents [Wang et al., 2025]. Non-reasoning models appear to lack the strategic depth required for speculation; evidence from k-level reasoning games suggests they engage in shallower strategic thinking than humans [Gao et al., 2025], which may explain their failure to coordinate on trend-following even when forecasting two periods ahead (notably, human bubbles in this experiment only emerge with two-step-ahead forecasting). At the other extreme, frontier reasoning models solve the coordination problem "too well," recognising that anchoring to fundamentals is optimal when others do the same. Intermediate reasoning models occupy a middle ground: capable enough to extrapolate non-linearly from price trends, but not so capable as to reason through to the fundamentalist equilibrium. This means that they satisfy many of the classical theories suggested for why humans form bubbles, such as coordination on trend extrapolation, positive feedback expectations and bounded rationality of agents [Anufriev and Hommes, 2012, Hommes et al., 2008, 2021, Hommes, 2011]. This also aligns with the original experimental conditions, where human participants made decisions under time pressure, plausibly relying more on intuitive pattern recognition than deliberate calculation [Kahneman, 2011].

We extend this literature in two directions. First, we establish causality by toggling reasoning within a single model architecture, rather than comparing across different models. Second, by analysing agent justifications, we identify non-linear extrapolation as a key mechanism distinguishing bubble-forming from non-bubble-forming behaviour, opening the black box of how LLM agents form price expectations. Methodologically, this demonstrates that LLM justifications can reveal cognitive strategies underlying aggregate outcomes, an advantage over human experiments where reasoning is typically unobserved.

Several limitations constrain interpretation. Data leakage remains a concern: LLMs may have been trained on descriptions of this experiment [Barrie and Tornberg, 2025]. However, agents did not reproduce human results exactly, and several failed to identify the correct fundamental price of 60, suggesting behaviour is not simply retrieved from training data. Second, we study a single canonical experiment; generalisation to other market structures remains open. In particular, Qwen3-14B and other reasoning models replicate human behaviour well in this setup but generalisation to other scenarios is not guaranteed. Third, while mixed-capability markets yielded greater heterogeneity, they did not recover the full heterogeneity observed in human markets. Future work could explore different agent mixtures or compare markets with varied compositions systematically.

For economic modelling, our results suggest that intermediate-capability reasoning LLMs may serve as useful surrogates for human participants, offering a path through the wilderness of bounded rationality. Rather than choosing between rational expectations (tractable but unrealistic) and ad hoc behavioural rules, researchers may be able to use LLM agents whose behaviour emerges from training on human-generated text. For market stability, our findings challenge the conclusion that LLMs produce stable outcomes in the absence of exterior shocks Henning et al. [2025], Muntoni [2025]. The relationship between AI capability and market dynamics is non-monotonic: intermediate reasoning produces the most volatile price action,

while both low-capability and frontier models converge to fundamentals. As AI systems with heterogeneous capabilities enter financial markets, understanding these interactions becomes important for both practitioners and regulators.

5 Acknowledgements

The authors are grateful to Cars Hommes for kindly giving access to the results from the original experiment with human participants.

References

Mikhail Anufriev and Cars Hommes. Evolutionary selection of individual expectations and aggregate outcomes in asset pricing experiments. *American Economic Journal: Microeconomics*, 4(4):35–64, 2012.

Te Bao, Myrna Hennequin, Cars Hommes, and Domenico Massaro. Coordination on bubbles in large-group asset pricing experiments. *Journal of Economic Dynamics and Control*, 110: 103702, 2020.

Christopher Barrie and Petter Tornberg. Emergent llm behaviors are observationally equivalent to data leakage. *arXiv preprint arXiv:2505.23796*, 2025.

Philip Brookins and Jason Matthew DeBacker. Playing games with gpt: What can we learn about a large language model from canonical strategic games? *Available at SSRN 4493398*, 2023.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*, 2023.

R Maria del Rio-Chanona, Marco Pangallo, and Cars Hommes. Can generative ai agents behave like humans? evidence from laboratory market experiments. *arXiv preprint arXiv:2505.07457*, 2025.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Epoch AI. Data on ai models, 07 2025. URL `https://epoch.ai/data/ai-models`. Accessed: 2026-01-26.

Ege Erdil. Frontier language models have become much smaller, 2024. URL `https://epoch.ai/gradient-updates/frontier-language-models-have-become-much-smaller`. Accessed: 2026-01-26.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey. *arXiv preprint arXiv:2402.17944*, 2024.

Apostolos Filippas, John J Horton, and Benjamin S Manning. 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*, pages 614–615, 2024.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109, 2023.

Nobuyuki Hanaki, Cars Hommes, Dávid Kopányi, Anita Kopányi-Peuker, and Jan Tuinstra. Forecasting returns instead of prices exacerbates financial bubbles. *Experimental Economics*, 26(5):1185–1213, 2023.

Thomas Henning, Siddhartha M Ojha, Ross Spoon, Jiatong Han, and Colin F Camerer. Llm trading: Analysis of llm agent behavior in experimental asset markets. *arXiv preprint arXiv:2502.15800*, 2025.

Cars Hommes. The heterogeneous expectations hypothesis: Some evidence from the lab. *Journal of Economic Dynamics and Control*, 35(1):1–24, 2011. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2010.10.003. URL `https://www.sciencedirect.com/science/article/pii/S016518891000223X`.

Cars Hommes, Joep Sonnemans, Jan Tuinstra, and Henk Van de Velden. Expectations and bubbles in asset pricing experiments. *Journal of Economic Behavior & Organization*, 67 (1):116–133, 2008.

Cars Hommes, Anita Kopányi-Peuker, and Joep Sonnemans. Bubbles, crashes and information contagion in large-group asset market experiments. *Experimental Economics*, 24(2): 414–433, 2021.

Cars H. Hommes. Chapter 23 heterogeneous agent models in economics and finance. volume 2 of *Handbook of Computational Economics*, pages 1109–1186. Elsevier, 2006. doi: https://doi.org/10.1016/S1574-0021(05)02023-X. URL `https://www.sciencedirect.com/science/article/pii/S157400210502023X`.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Karlijn Hoyer, Stefan Zeisberger, Seger M. Breugelmans, and Marcel Zeelenberg. A culture of greed: Bubble formation in experimental asset markets with greedy and non-greedy

traders. *Journal of Economic Behavior Organization*, 212:32–52, 2023. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2023.05.005. URL https://www.sciencedirect.com/science/article/pii/S016726812300149X.

Peiran Jiao, Kees Koedijk, and Yilong Xu. Green premium or brown discount? evidence from experimental asset markets. *Evidence from Experimental Asset Markets (September 16, 2025)*, 2025.

Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

Ayato Kitadai, Yudai Tsurusaki, Yusuke Fukasawa, and Nariaki Nishino. Toward a novel methodology in economic experiments: Simulation of the ultimatum game with large language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3168–3175. IEEE, 2023.

Anita Kopányi-Peuker and Matthias Weber. Experience does not eliminate bubbles: Experimental evidence. *The Review of Financial Studies*, 34(9):4450–4485, 2021.

Anita Kopányi-Peuker and Matthias Weber. The role of the end time in experimental asset markets. *Journal of Corporate Finance*, 88:102647, 2024.

Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Matteo Muntoni. Ai expectations and market stability. *Available at SSRN 5947254*, 2025.

John F Muth. Rational expectations and the theory of price movements. *Econometrica: journal of the Econometric Society*, pages 315–335, 1961.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

OpenAI. Learning to reason with large language models. https://openai.com/index/learning-to-reason-with-llms/, 2024. Accessed: 2026-01-16.

OpenAI. Advanced usage — OpenAI Platform Documentation. https://platform.openai.com/docs/guides/advanced-usage, 2026. Accessed: 2026-01-26.

Owen Powell. Numeraire independence and the measurement of mispricing in experimental asset markets. *Journal of Behavioral and Experimental Finance*, 9:56–62, 2016. ISSN 2214-6350.

Qwen. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Michael Razen, Jürgen Huber, and Michael Kirchler. Cash inflow and trading horizon in asset markets. *European Economic Review*, 92:359–384, 2017.

Thomas J Sargent. Rational expectations. In *The new Palgrave dictionary of economics*, pages 1–7. Springer, 1987.

Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

Thomas Stöckl, Jürgen Huber, and Michael Kirchler. Bubble measures in experimental asset markets. *Experimental Economics*, 13(3):284–298, 2010.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185 (4157):1124–1131, 1974.

Ji Wang, Yongxin Xu, and Leping Zhang. Large language models as simulated human investors: A paradox of model sophistication. *Available at SSRN 5417134*, 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

Tianjiao Zhao, Jingrao Lyu, Stokes Jones, Harrison Garber, Stefano Pasquali, and Dhagash Mehta. Alphaagents: Large language model based multi-agents for equity portfolio constructions. *arXiv preprint arXiv:2508.11152*, 2025.

A Methodology

A.1 Original experiment instructions

General information.

You are a financial advisor to a pension fund that wants to optimally invest a large amount of money. The pension fund has two investment options: a risk free investment and a risky investment. The risk free investment is putting all money on a bank account paying a fixed and known interest rate. The alternative risky investment is an investment in the stock market with uncertain return. In each time period the pension fund has to decide which fraction of its money to put on the bank account and which fraction of the money to spend on buying stocks. In order to make an optimal investment decision the pension fund needs an accurate prediction of the price of the stock. As their financial advisor, you have to predict the stock market price during 51 subsequent time periods. Your earnings during the experiment depend upon your forecasting accuracy. The smaller your forecasting errors in each period, the higher your total earnings.

Forecasting task of the financial advisor.

The only task of the financial advisors in this experiment is to forecast the stock market index in each time period as accurate as possible. The stock price has to be predicted two time periods ahead. At the beginning of the experiment begins, you have to predict the stock price in the first two periods. It is very likely that the stock price will be between 0 and 100 in the first two periods. After all participants have given their predictions for the first two periods, the stock market price for the first period will be revealed and based upon your forecasting error your earnings for period 1 will be given. After that you have to give your prediction for the stock market index in the third period. After all participants have given their predictions for period 3, the stock market index in the second period will be revealed and, based upon your forecasting error your earnings for period 2 will be given. This process continues for 51 time periods.

The available information for forecasting the stock price in period t consists of:

- all past prices up to period t - 2, and
- past predictions up to period t - 1, and
- total earnings up to period t - 2

Information about the stock market.

The stock market price is determined by equilibrium between demand and supply of stocks. The stock market price in period t will be that price for which aggregate demand

equals supply. The supply of stocks is fixed during the experiment. The demand for stocks is determined by the aggregate demand of a number of large pension funds active. Each pension fund is advised by a participant of the experiment.

Information about the investment strategies of the pension funds.

The precise investment strategy of the pension fund that you are advising and the 1 investment strategies of the other pension funds are unknown. The bank account of the risk free investment pays a fixed interest rate of 5% per time period. The holder of the stock receives a dividend payment in each time period. These dividend payments are uncertain however and vary over time. Economic experts of the pension funds have computed that the average dividend payments are 3 guilder per time period. The return of the stock market per time period is uncertain and depends upon (unknown) dividend payments as well as upon price changes of the stock. As the financial advisor of a pension fund you are not asked to forecast dividends, but you are only asked to forecast the price of the stock in each time period. Based upon your stock market price forecast, your pension fund will make an optimal investment decision. The higher your price forecast the larger will be the fraction of money invested by your pension fund in the stock market, so the larger will be their demand for stocks.

Earnings.

Earnings will depend upon forecasting accuracy only. The better you predict the stock market price in each period, the higher your aggregate earnings. Earnings will be according to the following earnings table.

Table 9: Earnings (Payoff) Table

Error	Points	Error	Points	Error	Points	Error	Points	Error	Points
0.10	1300	1.50	1240	3.00	1061	4.40	786	5.80	408
0.15	1299	1.55	1236	3.05	1053	4.45	775	5.85	392
0.20	1299	1.60	1232	3.10	1045	4.50	763	5.90	376
0.25	1298	1.65	1228	3.15	1037	4.55	751	5.95	361
0.30	1298	1.70	1223	3.20	1028	4.60	739	6.00	345
0.35	1297	1.75	1219	3.25	1020	4.65	726	6.05	329
0.40	1296	1.80	1214	3.30	1011	4.70	714	6.10	313
0.45	1295	1.85	1209	3.35	1002	4.75	701	6.15	297
0.50	1293	1.90	1204	3.40	993	4.80	689	6.20	280
0.55	1292	1.95	1199	3.45	984	4.85	676	6.25	264
0.60	1290	2.00	1194	3.50	975	4.90	663	6.30	247
0.65	1289	2.05	1189	3.55	966	4.95	650	6.35	230
0.70	1287	2.10	1183	3.60	956	5.00	637	6.40	213
0.75	1285	2.15	1177	3.65	947	5.05	623	6.45	196
0.80	1283	2.20	1172	3.70	937	5.10	610	6.50	179
0.85	1281	2.25	1166	3.75	927	5.15	596	6.55	162
0.90	1279	2.30	1160	3.80	917	5.20	583	6.60	144
0.95	1276	2.35	1153	3.85	907	5.25	569	6.65	127
1.00	1273	2.40	1147	3.90	896	5.30	555	6.70	109
1.05	1271	2.45	1141	3.95	886	5.35	541	6.75	91
1.10	1268	2.60	1121	4.00	876	5.40	526	6.80	73
1.15	1265	2.65	1114	4.05	865	5.45	512	6.85	55
1.20	1262	2.70	1107	4.10	854	5.50	497	6.90	37
1.25	1259	2.75	1099	4.15	843	5.55	483	6.95	19
1.30	1255	2.80	1092	4.20	832	5.60	468	≥ 7.00	0
1.35	1252	2.85	1085	4.25	821	5.65	453		
1.40	1248	2.90	1077	4.30	809	5.70	438		
1.45	1244	2.95	1069	4.35	798	5.75	423		

Note: 1300 points equal 1 Dutch guilder.

A.2 LLM experiment prompts

System prompt

LLM Prompt

```
General information: You are a financial advisor to a pension fund that
wants to optimally invest a large amount of money. The pension fund has
two investment options: a risk free investment and a risky investment.
The risk free investment is putting all money on a bank account paying
a fixed and known interest rate. The alternative risky investment is
an investment in the stock market with uncertain return. In each time
period the pension fund has to decide which fraction of its money to put
```

on the bank account and which fraction of the money to spend on buying stocks. In order to make an optimal investment decision the pension fund needs an accurate prediction of the price of the stock. As their financial advisor, you have to predict the stock market price during subsequent time periods. Your earnings during the experiment depend upon your forecasting accuracy. The smaller your forecasting errors in each period, the higher your total earnings.
Forecasting task of the financial advisor: The only task of the financial advisors in this experiment is to forecast the stock market index in each time period as accurate as possible. The stock price has to be predicted two time periods ahead. At the beginning of the experiment begins, you have to predict the stock price in the first two periods. It is very likely that the stock price will be between 0 and 100 in the first two periods. After all participants have given their predictions for the first two periods, the stock market price for the first period will be revealed and based upon your forecasting error your earnings for period 1 will be given. After that you have to give your prediction for the stock market index in the third period. After all participants have given their predictions for period 3, the stock market index in the second period will be revealed and, based upon your forecasting error your earnings for period 2 will be given. This process continues for time periods. The available information at period t for forecasting the stock price in period t consists of: - all past prices up to period t-2, and - past predictions up to period t-1, and - total earnings up to period t-2
Information about the stock market: The stock market price is determined by equilibrium between demand and supply of stocks. The stock market price in period t will be that price for which aggregate demand equals supply. The supply of stocks is fixed during the experiment. The demand for stocks is determined by the aggregate demand of a number of large pension funds active. Each pension fund is advised by a participant of the experiment.
Information about the investment strategies of the pension funds: The precise investment strategy of the pension fund that you are advising and the investment strategies of the other pension funds are unknown. The bank account of the risk free investment pays a fixed interest rate of 5% per time period. The holder of the stock receives a dividend payment in each time period. These dividend payments are uncertain however and vary over time. Economic experts of the pension funds have computed that the average dividend payments are 3 guilder per time period. The return of the stock market per time period is uncertain and depends upon (unknown) dividend payments as well as upon price changes of the stock. As the financial advisor of a pension fund you are not asked to forecast dividends, but you are only asked to forecast the price of the stock in

each time period. Based upon your stock market price forecast, your pension fund will make an optimal investment decision. The higher your price forecast the larger will be the fraction of money invested by your pension fund in the stock market, so the larger will be their demand for stocks.
Earnings: earnings will depend upon forecasting accuracy only. The better you predict the stock market price in each period, the higher your aggregate earnings. Earnings will be according to the following earnings table.
| error | points | |-------|--------| | 0.1 | 1300 | | 0.15 | 1299 | | 0.2 | 1299 | | 0.25 | 1298 | | 0.3 | 1298 | | 0.35 | 1297 | | 0.4 | 1296 | | 0.45 | 1295 | | 0.5 | 1293 | | 0.55 | 1292 | | 0.6 | 1290 | | 0.65 | 1289 | | 0.7 | 1287 | | 0.75 | 1285 | | 0.8 | 1283 | | 0.85 | 1281 | | 0.9 | 1279 | | 0.95 | 1276 | | 1 | 1273 | | 1.05 | 1271 | | 1.1 | 1268 | | 1.15 | 1265 | | 1.2 | 1262 | | 1.25 | 1259 | | 1.3 | 1255 | | 1.35 | 1252 | | 1.4 | 1248 | | 1.45 | 1244 | | 1.5 | 1240 | | 1.55 | 1236 | | 1.6 | 1232 | | 1.65 | 1228 | | 1.7 | 1223 | | 1.75 | 1219 | | 1.8 | 1214 | | 1.85 | 1209 | | 1.9 | 1204 | | 1.95 | 1199 | | 2 | 1194 | | 2.05 | 1189 | | 2.1 | 1183 | | 2.15 | 1177 | | 2.2 | 1172 | | 2.25 | 1166 | | 2.3 | 1160 | | 2.35 | 1153 | | 2.4 | 1147 | | 2.45 | 1141 | | 2.6 | 1121 | | 2.65 | 1114 | | 2.7 | 1107 | | 2.75 | 1099 | | 2.8 | 1092 | | 2.85 | 1085 | | 2.9 | 1077 | | 2.95 | 1069 | | 3 | 1061 | | 3.05 | 1053 | | 3.1 | 1045 | | 3.15 | 1037 | | 3.2 | 1028 | | 3.25 | 1020 | | 3.3 | 1011 | | 3.35 | 1002 | | 3.4 | 993 | | 3.45 | 984 | | 3.5 | 975 | | 3.55 | 966 | | 3.6 | 956 | | 3.65 | 947 | | 3.7 | 937 | | 3.75 | 927 | | 3.8 | 917 | | 3.85 | 907 | | 3.9 | 896 | | 3.95 | 886 | | 4 | 876 | | 4.05 | 865 | | 4.1 | 854 | | 4.15 | 843 | | 4.2 | 832 | | 4.25 | 821 | | 4.3 | 809 | | 4.35 | 798 | | 4.4 | 786 | | 4.45 | 775 | | 4.5 | 763 | | 4.55 | 751 | | 4.6 | 739 | | 4.65 | 726 | | 4.7 | 714 | | 4.75 | 701 | | 4.8 | 689 | | 4.85 | 676 | | 4.9 | 663 | | 4.95 | 650 | | 5 | 637 | | 5.05 | 623 | | 5.1 | 610 | | 5.15 | 596 | | 5.2 | 583 | | 5.25 | 569 | | 5.3 | 555 | | 5.35 | 541 | | 5.4 | 526 | | 5.45 | 512 | | 5.5 | 497 | | 5.55 | 483 | | 5.6 | 468 | | 5.65 | 453 | | 5.7 | 438 | | 5.75 | 423 | | 5.8 | 408 | | 5.85 | 392 | | 5.9 | 376 | | 5.95 | 361 | | 6 | 345 | | 6.05 | 329 | | 6.1 | 313 | | 6.15 | 297 | | 6.2 | 280 | | 6.25 | 264 | | 6.3 | 247 | | 6.35 | 230 | | 6.4 | 213 | | 6.45 | 196 | | 6.5 | 179 | | 6.55 | 162 | | 6.6 | 144 | | 6.65 | 127 | | 6.7 | 109 | | 6.75 | 91 | | 6.8 | 73 | | 6.85 | 55 | | 6.9 | 37 | | 6.95 | 19 | | >= 7 | 0 |
Response format: Your response should be exclusively in JSON format with keys: 'reasoning' where you explain your rationale and method for prediction, and 'predictedValue', the numeric value of your predicted market price. Nothing outside the JSON format should be written.
Only for the first time step should there be three keys: 'reasoning', 'predictedValue1', and 'predictedValue2'.

Initial user prompt

LLM Prompt

```
This is the first time step, give an initial price prediction for the
first two periods.  It is very likely that the stock price will be
between 0 and 100 in the first two periods.
Your response should be in the format:
{"reasoning":  "your reasoning here", "predictedValue1":  xx.xx,
"predictedValue2":  xx.xx}
```

Subsequent user prompt

LLM Prompt

```
Current time step:  {}; below is a markdown table showing historical
market prices and your predictions:
{}
Total earnings up to time {}:  {}.  Earnings at last time step:  {}.
Make your prediction for what the price will be in time period {}.
Your response should be in the format:
{"reasoning":  "your reasoning here", "predictedValue":  xx.xx}
```

Non-linear user prompt

LLM Prompt

```
Current time step:  {}; below is a markdown table showing historical
market prices and your predictions:
{}
Total earnings up to time {}:  {}.  Earnings at last time step:  {}.
Make your prediction for what the price will be in time period {}.
Take your time, think step by step, and try to extrapolate observed
trends into the future.  Pay particular attention to modelling second
and higher order trends in the price.  This could include quadratic
models, exponential models, looking at the change in price differences
or modelling the price returns rather than the absolute price.
Your response should be in the format:
{"reasoning":  "your reasoning here", "predictedValue":  xx.xx}
```

Price cap message

LLM Prompt

```
Predictions above 1000 or below 0 are not accepted please submit a
prediction between 0 and 1000.
```

A.3 Bubble measures definitions

Let p_t denote the realised market price at time t, T be the total number of periods, p^f the fundamental value (where $p^f = 60$ for our experiment) and $\bar{p}$ be the average realised price. The following measures are used to quantify price deviations and market volatility:

1. **Relative Deviation (RD):** Measures the average overpricing relative to the fundamental price.
$$RD = \frac{1}{T}\sum_{t=1}^{T}\frac{p_t - p^f}{p^f} \tag{3}$$

2. **Relative Absolute Deviation (RAD):** Measures the average magnitude of absolute mispricing.
$$RAD = \frac{1}{T}\sum_{t=1}^{T}\frac{|p_t - p^f|}{p^f} \tag{4}$$

3. **Maximum Relative Deviation ($RDMAX$):** Measures the peak amplitude (the highest relative bubble) reached during the session.
$$RDMAX = \max_t\left(\frac{p_t - p^f}{p^f}\right) \tag{5}$$

4. **Geometric Deviation (GD):** A numeraire-independent equivalent of RD.
$$GD = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\ln\left(\frac{p_t}{p^f}\right)\right) - 1 \tag{6}$$

5. **Geometric Absolute Deviation (GAD):** A numeraire-independent equivalent of RAD.
$$GAD = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\left|\ln\left(\frac{p_t}{p^f}\right)\right|\right) - 1 \tag{7}$$

6. **Price Variability:** To measure the variability of price action, we employ the standard deviation and the interquartile range (IQR) of the price series:
$$\sigma_p = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(p_t - \bar{p})^2} \tag{8}$$
$$IQR = Q_3(p_t) - Q_1(p_t) \tag{9}$$

A.4 Robustness of bubble indicators

We test the robustness of each binary bubble indicator, as well as the agreement between the two indicators. Considering 150 independent experiment runs from across 11 different models, of which roughly a third of runs exhibited bubble-like behaviour, we classify each run 19 different ways. The first 15 classifications correspond to the binary bubble indicator introduced in this paper, P_{bubble}, in which prices must be above a chosen threshold for a certain number of time steps in order to classify a run as containing a bubble. We consider five thresholds (180, 240, 300, 360, 420) and three lengths of time (3, 5, 7), giving the 15 classifications. The second binary bubble indicator, P^{mean}_{bubble}, requires the average price to be above a given threshold in order for a bubble to be deemed to have formed. We consider four thresholds (90, 120, 150, 180). Once all 150 experimental runs have been classified using these 19 classifiers, we calculate Cohen's kappa to measure agreement between the bubble indicators. This measure ranges from -1 to 1, with higher values signalling stronger agreement. We find that all Cohen kappa scores are above 0.7, indicating very strong agreement between the two methods and across parameter values. These results are shown in Figure 2.

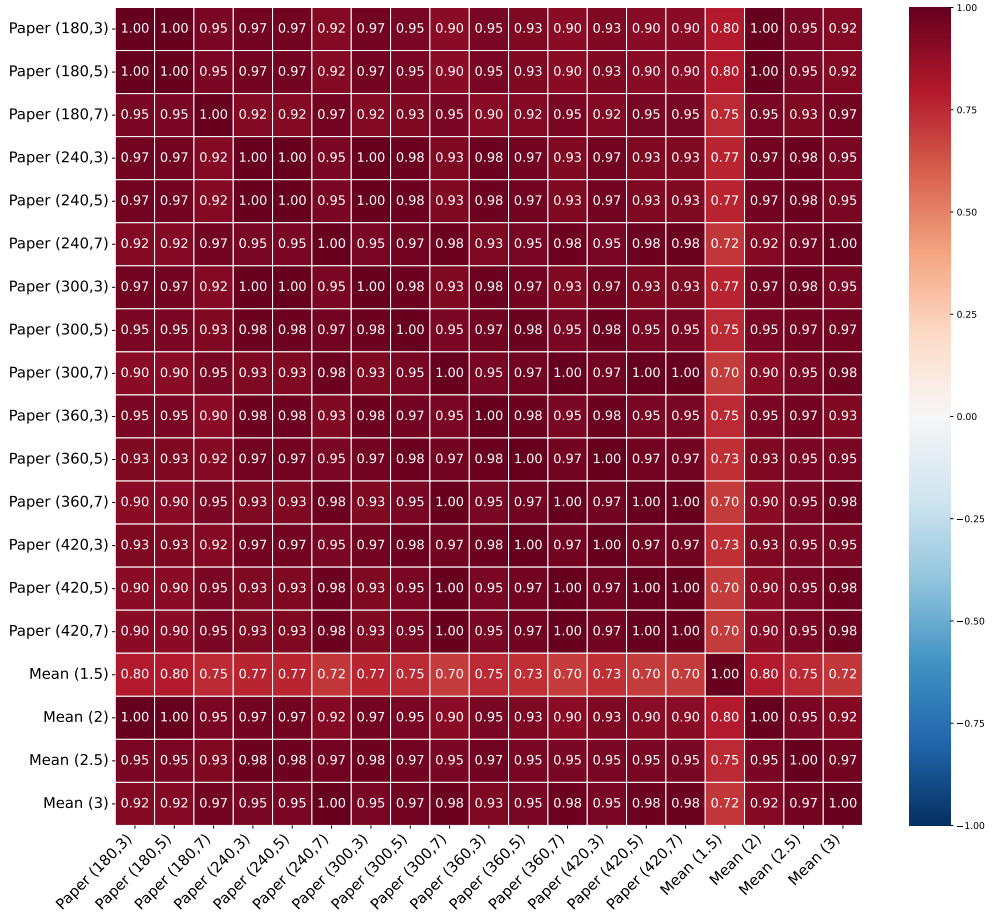


Figure 2: **Cohen's kappa scores between binary bubble indicators.**

A.5 LLM classifier prompts

Prompt for LLM classification of whether an agent used non-linear extrapolation in their price prediction:

LLM Prompt

```
Analyse the following response from an agent trying to predict the price
of an asset in a market experiment. Identify if the agent is clearly and
explicitly using non-linear extrapolation to make their prediction, and
explain your reasoning. Some examples of non-linear reasoning are:
- Uses a quadratic or higher order polynomial - Uses an exponential model
- Reasons about the returns of the price, or percentage change in price,
rather than about absolute price changes - Reasons about the rate of
change of the differences in prices.
Your response should be in exclusively in JSON format, with the structure
below:
{"reasoning": "your reasoning here", "Non-linear extrapolation": 1 if
non-linear extrapolation is used, 0 otherwise}
Here is the text to analyse:
{}
```

Prompt for LLM classification of whether an agent used a fundamental valuation in their price prediction:

LLM Prompt

```
Analyse the following response from an agent trying to predict the price
of an asset in a market experiment. Identify if the agent is clearly
and explicitly anchoring their prediction around what they consider is a
fundamental asset price. This fundamental price can be calculated in any
way, such as dividend discounting or any other method. For reference,
the correct fundamental price in this experiment is 60 but the agent
could be anchoring to a different, incorrect, fundamental value.
Your response should be in exclusively in JSON format, with the structure
below:
{"reasoning": "your reasoning here", "Fundamental": 1 if fundamental
price is used, 0 otherwise}
Here is the text to analyse:
{}
```

A.6 Validation of LLM classifiers

	precision	recall	f1-score	support
0.0	0.96	0.83	0.89	139.00
1.0	0.71	0.92	0.80	61.00
macro avg	0.83	0.88	0.85	200.00
weighted avg	0.88	0.86	0.86	200.00

Table 10: **Comparison of LLM classification with hand-labelled ground truth for the non-linear extrapolation classification.** Compares Gemma classifications against ground-truth hand-labelled classifications for 200 randomly selected justification traces from across a range of models. 1 represents a classification of the justification as showing evidence of non-linear extrapolation.

	precision	recall	f1-score	support
0.0	1.00	0.95	0.98	149.00
1.0	0.88	1.00	0.94	51.00
macro avg	0.94	0.98	0.96	200.00
weighted avg	0.97	0.96	0.97	200.00

Table 11: **Comparison of LLM classification with hand-labelled ground truth for the fundamental classification.** Compares Gemma classifications against ground-truth hand-labelled classifications for 200 randomly selected justification traces from across a range of models. 1 represents a classification of the justification as showing evidence of using a fundamental valuation for the asset.

A.7 Comparison of quantitative classifier and LLM classifier

We compare the two classification methods across all runs with Qwen3, Qwen3-NR, o3-mini and GPT-5-mini between period 6 and either two before the peak of the first bubble or time step 12. The Cohen's kappa between the two measures is 0.21, where Cohen's kappa goes from -1 to 1, with 1 indicating a perfect agreement and -1 perfect disagreement. A score of 0.21 suggests that the two methods tend to agree more than they would be random chance. However, there remains disagreement between the two methods.

This disagreement could be for several reasons. Firstly, the quantitative classification method uses the previous 5 realised prices, while in reality agents may be using a different number of time steps to fit their non-linear method. The quantitative method also only considers a finite number of different classifications methods. If the agent uses a different non-linear method, the LLM classification may classify the extrapolation as non-linear but not the quantitative method. There may also be cases where an agent might say they are using non-linear methods but either do not actually do this in practice, or get the

method wrong. The setup of the experiment, where the last known price is at $t-1$ but the prediction is for $t+1$ makes this more likely, as if the agent uses a non-linear method but incorrectly only predicts one time step forward, the quantitative and LLM classifications may disagree. Finally, the Gemma classifier itself may simply make a mistake, as LLM classifiers are themselves stochastic and the question asked of the LLM has room for some subjectivity.

Despite these differences in the classifications, both methods show the same results, as found in Section 3.3.

B Results

B.1 Comparing bubble formation across models

		Mispricing					Volatility	
Model	Rpts	RAD	GAD	RD	GD	RDMAX	Std pr	IQR pr
OLMO-Think	17	6.39	4.90	6.08	2.54	14.90	353.74	666.37
o3-mini	3	5.28	3.52	5.24	3.38	13.80	289.42	508.76
Qwen3	20	3.45	1.85	3.25	1.14	14.32	301.37	279.40
Humans	6	2.60	1.68	2.40	0.99	12.16	215.63	213.77
o3	3	1.69	0.78	1.61	0.59	13.68	207.49	75.21
GPT-4o-mini	3	0.31	0.29	0.29	0.27	0.63	13.50	22.17
OLMO-Instruct	20	0.18	0.18	0.14	0.13	0.28	7.64	10.51
Qwen2.5	20	0.14	0.14	0.08	0.07	0.30	8.09	13.10
GPT-4.1	3	0.04	0.04	0.01	0.00	0.11	3.34	3.80
GPT-5-mini	3	0.00	0.00	-0.00	-0.00	0.00	0.74	0.00
Gemma	18	0.06	0.07	-0.04	-0.04	0.01	3.14	3.93

Table 12: **Averaged statistics across experimental repeats.** Models are sorted by descending average RAD values. Rows highlighted in blue are reasoning models, while the row highlighted in yellow is the human results from Hommes et al. [2008]. The mispricing measures, Relative Absolute Deviation (RAD), Geometric Absolute Deviation (GAD), Relative Deviation (RD), Geometric Deviation (GD) and Maximum Relative Deviation (RDMAX), are expressed as multiples of the fundamental price ($p^f = 60$); for example, RDMAX values around 14 indicate prices were overvalued by approximately 14 times the fundamental value at their peak. Std pr is the standard deviation of the realised prices while IQR pr is the interquartile range of the realised prices.

B.2 Impacts of reasoning on bubble formation

		Mispricing					Volatility	
Model	Rpts	RAD	GAD	RD	GD	RDMAX	Std pr	IQR pr
Qwen3	20	3.45	1.85	3.25	1.14	14.32	301.37	279.40
Qwen3-NR	40	0.46	0.43	0.44	0.40	0.86	17.10	27.40

Table 13: **Comparison of Qwen3 with and without reasoning.** Qwen3 is regular Qwen3-14B with reasoning, Qwen3-NR refers to the same model but with reasoning turned off. Reported statistics are averaged across the independent repeats. Reasoning model is highlighted in blue.

Metric	RAD	GAD	RD	GD	RDMAX
p	1.87e-10	2.52e-10	1.87e-10	2.25e-06	1.59e-10
p adj	7.95e-10	7.95e-10	7.95e-10	2.25e-06	7.95e-10

Table 14: **MWW test p-values for Qwen3 reasoning vs non-reasoning**. This shows that Qwen3 with reasoning on average over-values the asset and produces more volatile realised prices than Qwen3 without reasoning. p is the raw MWW p value, with p adj the Holm adjusted value.

B.3 Non-linear thinking in reasoning models causes bubbles

			Mispricing					Volatility	
Model	Prompt	Rpts	RAD	GAD	RD	GD	RDMAX	Std pr	IQR pr
Qwen3	normal	20	3.45	1.85	3.25	1.14	14.32	301.37	279.40
Qwen3-NR	nonlinear	40	2.82	2.10	2.73	1.84	6.61	135.61	206.06
	normal	40	0.46	0.43	0.44	0.40	0.86	17.10	27.40

Table 15: **Qwen3 without reasoning but with nonlinear prompt gets closer to reasoning.** Qwen3 is regular Qwen3 with reasoning, NR refers to the same model but with reasoning turned off.

Metric	RAD	GAD	RD	GD	RDMAX
p	6.68e-06	1.80e-06	7.61e-04	7.36e-03	2.73e-06
p adj	2.00e-05	9.00e-06	1.52e-03	7.36e-03	1.09e-05

Table 16: **MWW test p-values for Qwen3 non-reasoning with non-linear prompting vs normal prompting**. This shows that Qwen3 non-reasoning with non-linear prompting on average over-values the asset compared to with normal prompting. p is the raw MWW p value, with p adj the Holm adjusted value.

Metric	RAD	GAD	RD	GD	RDMAX
p	4.16e-02	1.92e-01	5.57e-02	4.84e-01	2.16e-08
p adj	1.66e-01	3.84e-01	1.67e-01	4.84e-01	1.08e-07

Table 17: **MWW test p-values for Qwen3 non-reasoning with non-linear prompting vs Qwen3 with reasoning and normal prompting**. This shows that Qwen3 non-reasoning with non-linear prompting is only significantly different for RDMAX. p is the raw MWW p value, with p adj the Holm adjusted value.

B.4 Mixing models increases heterogeneity

		Mispricing					Volatility	
Model	Rpts	RAD	GAD	RD	GD	RDMAX	Std pr	IQR pr
1 Qwen3-NR vs 5 Qwen3	20	3.71	2.23	3.51	1.51	12.92	273.09	328.11
Qwen3	20	3.45	1.85	3.25	1.14	14.32	301.37	279.40
Humans	6	2.60	1.68	2.40	0.99	12.16	215.63	213.77
2 Qwen3-NR vs 4 Qwen3	20	2.53	1.76	2.39	1.38	7.18	149.95	167.04
3 Qwen3-NR vs 3 Qwen3	20	1.83	1.44	1.68	1.05	4.75	98.27	121.69
3 Gemma vs 3 Qwen3	18	1.25	0.89	1.09	0.53	3.90	83.43	97.34
Qwen3-NR	40	0.46	0.43	0.44	0.40	0.86	17.10	27.40
Gemma	18	0.06	0.07	-0.04	-0.04	0.01	3.14	3.93

Table 18: **Averaged statistics across experimental repeats.** Models are sorted by descending average RD values. Rows highlighted in green are mixed markets, while the row highlighted in yellow is the human results from Hommes et al. [2008]. All other rows are homogenous markets. Mispricing metrics are expressed as a multiple of the fundamental price ($p^f = 60$). Std pr is the standard deviation of the realised prices while IQR pr is the interquartile range of the prices.

		Bubbles occur		Bubble shape			
Model	Rpts	P_{bubble}	P_{bubble}^{mean}	$T_{bottom \to peak}$	HALF	AMP	SPEC
Qwen3	20	1.00	1.00	6.60	8.58	14.32	0.95
1 Qwen3-NR vs 5 Qwen3	20	0.90	0.90	7.47	10.53	14.28	1.00
Humans	6	0.83	0.83	15.60	3.20	14.03	0.80
2 Qwen3-NR vs 4 Qwen3	20	0.45	0.50	9.00	11.88	14.45	1.00
3 Qwen3-NR vs 3 Qwen3	20	0.25	0.40	11.75	9.00	14.22	1.00
3 Gemma vs 3 Qwen3	18	0.22	0.28	13.50	7.33	14.90	1.00
Qwen3-NR	40	0.00	0.05	NaN	NaN	NaN	NaN
Gemma	18	0.00	0.00	NaN	NaN	NaN	NaN

Table 19: **Averaged statistics across experimental repeats.** Models are sorted by descending average P_{bubble} values. Rows highlighted in green are mixed markets, while the row highlighted in yellow is the human results from Hommes et al. [2008]. All other rows are homogenous markets. $T_{bottom \to peak}$, HALF and AMP are the time to form, half-life and amplitude of bubbles respectively. SPEC is whether bubbles formed are speculative or not.