Information Aggregation with AI Agents*

Spyros Galanis†

March 27, 2026

Abstract

Can Large Language Models (AI agents) aggregate dispersed private information through trading and reason about the knowledge of others by observing price movements? We conduct a controlled experiment where AI agents trade in a prediction market after receiving private signals, measuring information aggregation by the log error of the last price. We find that although the median market is effective at aggregating information in the easy information structures, increasing the complexity has a significant and negative impact, suggesting that AI agents suffer from the same limitations as humans when reasoning about others. Consistent with our theoretical predictions, information aggregation remains unaffected by allowing cheap talk communication, changing the duration of the market or initial price, and strategic prompting—thus demonstrating that prediction markets are robust. We establish that "smarter" AI agents perform better at aggregation and they are more profitable. Surprisingly, giving them feedback about past performance makes them worse at aggregation and reduces their profits.
JEL: C91, D82, D83, D84, G14, G41

Keywords: Information Aggregation, AI agents, Artificial Intelligence, Financial Markets, Prediction Markets, Experiments

1 Introduction

Recent advancements in Large Language Models (LLMs) have enabled the development of AI agents capable of autonomously executing complex tasks. They can gather private information and reason about necessary steps based on their prompts, execute actions by invoking external tools and collaborating with other agents, and evaluate the success of their actions by observing the new state. Soon, they will be tasked to trade securities in financial markets, leveraging their private information. However, a critical open question remains: to

*I would like to thank Andis Sofianos, Diego Marino-Fages, and participants at Durham for useful comments. This research is based on work funded under ESRC grant ES/V004425/1.

†Department of Economics, University of Durham, spyros.galanis@durham.ac.uk.

what extent can these agents reason about the private information held by others (humans or AI agents) when observing their actions? This is fundamental for deploying robust systems in multi-agent environments, whether they involve human-AI collaboration or interactions between multiple AI agents. Moreover, as Hayek (1945) has argued, a well-functioning pricing system that aggregates information is necessary to achieve efficient outcomes, because the price of a security contains all the relevant information needed for a decision maker to optimise.

We study this question by running a controlled experiment with AI agents who receive private information and trade a security in a prediction market, which pays 0 or 1 based on the outcome of a binary question. The conjunction of everyone's information reveals the true value of the security. Is trading activity enough to drive the price of the security close to its true value, so that information aggregates?

The intuition behind information aggregation is simple. If the price of a security is high and a trader thinks its value is low, he will sell, otherwise he will buy. The other traders observe the price movements and try to infer the private information that led to the buy or sell orders. After incorporating this (now public) information into their own private information, traders will buy or sell and the process continues, until all private information is aggregated. However, this aggregation requires that traders are sophisticated enough to form higher order beliefs about what others know, what they know about what others know, and so on, in order to interpret their actions and update their knowledge and beliefs.

Reasoning about the private information of others by observing their actions is a fundamental human ability, but it is not a given that AI agents are trained to emulate it. On the one hand, they already demonstrate impressive capabilities on several domains and Large Language Models (LLMs) can be prompted to optimise (Yang et al., 2023). On the other hand, they can emulate human behaviour in games (Park et al., 2023) and experiments (Horton, 2023), which is not always sophisticated, and suffer from the same behavioral biases as humans (Bini et al., 2025).

We vary several conditions in the experiment and measure their impact on information aggregation, or market accuracy. The first is the complexity of the information and payoff structure. We consider four levels. In the easiest, each trader needs to reason about the signal of only one other trader to determine whether the answer to the question is Yes or No. The hardest is a version of the famous "muddy children" puzzle, where each trader gets two signals so interactive reasoning is more complex. The other conditions are allowing AI agents to post public comments, to understand the effect of cheap talk communication, prompting them to be strategic (forward-looking) or myopic to measure if AI agents can implement a strategy over many periods, changing the initial price (0.3,0.5,0.7) to examine whether price manipulation has an adverse effect, and altering the duration of the market (3, 6, and 9 rounds) to see if more trading helps or confuses AI agents. These treatments generate a total of 144 different market configurations, and we run each configuration at least 12 times, with a diverse set of AI agents, generating 1772 prediction markets in total. We then run an information provision treatment, where we inform AI agents, before they trade, about the qualitative results from the first wave, on which factors were effective on information

aggregation and profits. This treatment generates another 1728 prediction markets.

For all four information structures, we construct securities which are 'separable'.[1] Hence, the theoretical prediction (if AI agents are rational and sophisticated) is that information will get aggregated in all Nash equilibria, for any initial price and irrespective of whether the traders are myopic or strategic, or whether they communicate (Ostrovsky, 2012).

We employ eight LLMs: Claude Haiku 3.5 and 4.5, Gemini 2.5 and 3 Flash, GPT-4o, and GTP 5 mini, gemma3:4b, and qwen3:8b. As a measure of each model's capabilities, we adopt the Artificial Analysis Intelligence Index (Artificial Analysis Team, 2025), one of the most comprehensive publicly available syntheses of model capabilities. The intelligence index integrates ten evaluation suites, combining performance across reasoning, mathematics, coding, and agentic workflow tasks, among others. We form twelve teams of three traders. Eight teams are homogeneous, comprising traders using the same model. The remaining four teams feature a variety of models, enabling us to test whether "diversity" of intelligence has an impact on information aggregation or profits.

Our first main finding is that although the median market is effective in the two easy structures, increasing the complexity of the information structure significantly degrades information aggregation.[2] In particular, the median market across all structures prices the security at 0.91 when its true value is 1. In the easy and medium structures the price is almost 1, suggesting that AI agents understand them completely. In the hard structure, the price drops at 0.73, hence noticeably worse but still better than random guessing. At the very hard structure, the price drops at 0.5, which is completely uninformative. Given that the securities are separable in all structures, our hypothesis that complexity does not influence information aggregation is rejected. These results suggest that AI agents resemble most humans who find it difficult to reason about the knowledge of others and form higher order beliefs as complexity increases, even in structures that involve only three traders and three signals. Another indication that AI agents struggle to reason about others is that individual profits decrease with complexity.

Our second main finding is to confirm our hypothesis that the following factors do not have a statistically significant impact on information aggregation: cheap talk communication, changing the initial price or the duration of the market from 3 to 9 rounds, and prompting AI agents to be strategic versus myopic. This result demonstrates that prediction markets are robust. More importantly, the ineffectiveness of communication suggests that prediction markets are also scalable. In markets with billions of AI agents, each with their own private information, exchanging information quickly becomes infeasible and the only signal that can be trusted is the price itself.

Our third finding concerns the role of agent intelligence in information aggregation and profits. Given that interactive reasoning requires sophisticated agents, our hypothesis was that intelligence would positively impact information aggregation. Indeed, we find that the average intelligence of the group reduces the average log error. To quantify this result, we ran

[1] A security is separable if, for every nondegenerate prior belief about the states of the world, there exists a trader who receives an informative signal with positive probability.

[2] The average market fares very poorly across all structures, but this is driven mostly by the performance of the bottom 20% markets.

two quantile regressions, using the median and the tail (bottom 20%) log error. Although the tail log error is reduced, the impact on the median is statistically insignificant. This result suggests that the role of intelligence is not to improve the performance of the typical market, but to avoid the few markets that misprice the security completely, trading at 1 (0) when the true value is 0 (1). Moreover, we find that the individual profits of an AI agent increase with his intelligence but decrease as the average intelligence of the group increases. These results are consistent with previous experiments with humans (Corgnet et al., 2018), yet the Intelligence Index does not specifically test for interactive reasoning tasks, which is surprising.

An important question is whether AI agents can leverage feedback from past play to improve future performance. To test this, we introduce an exogenous informational shock, providing a new wave of AI agents with the empirical outcomes of the previous 1772 markets. Contrary to recent literature suggesting that LLMs successfully improve when given performance histories (Yang et al., 2023), we uncover a surprising paradox: providing AI agents with qualitative results on past market data significantly harms information aggregation and their total profits. Rather than learning from the past to fine-tune their strategy, they seem to get confused and drive the average market error significantly higher.

Given our result that strategic prompting does not influence information aggregation, we further examine whether this is because AI agents act strategically but there is no impact on information aggregation, as predicted by economic theory, or they are just unable to act strategically. We do this by examining the public messages they post and the private messages that are only visible by their future selves. AI agents could act strategically if the public and private messages differ. For example, they could post public messages in order to lie about their signal or withhold information, when prompted to be strategic. If they understand the dynamic nature of the market, they could use their private messages to instruct their future selves on how to implement a multi-round strategy, or be more inclined to publicly reveal their signal in the last round they are trading.

We introduce three measures of communication strategy: semantic alignment between private and public messages (cosine similarity), information hoarding (word gap between public and private), and direct deception, measured by the information revealed about the trader's signal from his public message, as judged by an AI agent ex post. We find that while the strategic treatment has negligible effects, the agents' behavior shows that they do have some understanding of the strategic nature of the environment. In 94% of markets, agents actively hoard information, generating public announcements that are significantly shorter and semantically detached from their internal private reasoning. Furthermore, making agents aware of prior experimental outcomes actually exacerbates this adversarial behavior, leading to wider word gaps and increased direct deception.

Most strikingly, we uncover a sophisticated inter-temporal communication strategy employed by the AI agents. Rather than maintaining a constant rate of deception or exhibiting a simple linear decay over time, AI agents display a "sawtooth" pattern of revelation. They actively hoard information and deceive competitors in the opening rounds to protect their information rents. However, the probability of truthful revelation spikes sharply exactly at

the terminal rounds of the market (e.g., Rounds 3, 6, and 9). This demonstrates that AI agents possess a nuanced understanding of the market's temporal horizon, selectively executing an end-game "truth revelation" only when the financial penalty for revealing their private signals drops to zero. More interestingly, this sawtooth pattern of decreased deception in round 3, 6, and 9, persists even when we restrict the data on markets with 9 rounds. This may suggest that AI agent do not fully grasp the dynamic nature of the market.

1.1 Literature

Our paper contributes to two strands of the literature. The first studies under which conditions information gets aggregated. Ostrovsky (2012) and Chen et al. (2012) show that in a market with either myopic or strategic traders, separable securities are both necessary and sufficient for information aggregation, using both the model of Kyle (1985) and the Market Scoring Rule (McKelvey and Page (1990), Hanson (2003, 2007)), which is directly applicable to prediction markets. Information aggregation is based on the "we cannot agree to disagree" and "no trade" theorems of Aumann (1976), Milgrom and Stokey (1982), and Geanakoplos and Polemarchakis (1982). Dimitrov and Sami (2008) and Chen et al. (2010) examine information aggregation by varying the assumptions regarding the traders' information structure. Rasooly and Rozzi (2025) conduct a field experiment (with humans), where prices are randomly shocked in 817 prediction markets, finding that the effect was persistent even after 60 days. Galanis et al. (2024) show theoretically and experimentally that ambiguity aversion can lead to no information aggregation with separable securities, but a new class - strongly separable securities - overcomes this limitation. Galanis and Kotronis (2021) show that information aggregation may fail if traders are unaware of relevant dimensions, whereas Galanis and Mikhalishchev (2025) examine the effect of information acquisition on information aggregation. We contribute to this literature by conducting the first, to our knowledge, experiment that studies information aggregation with AI agents that trade in a prediction market.

The second strand is at the intersection of economics and computer science. One branch studies how AI agents behave in economically interesting problems (Chen et al. (2023), Bini et al. (2025)), whereas another studies how experimentation with LLMs can provide insights about human behavior (Charness et al. (2023), Korinek (2023), Bail (2024), Manning et al. (2024)).

The intelligence index we use in the paper (Artificial Analysis Team, 2025) encompasses many studies which evaluate the performance of LLMs across reasoning, knowledge, mathematics, coding, instruction following, long-context reasoning and agentic workflow tasks. Some of these are: MMLU-Pro (Wang et al., 2024) , GPQA Diamond (Rein et al., 2024), HLE Phan et al. (2025), AIME 2025, SciCode (Tian et al., 2024), LiveCodeBench (Jain et al., 2024), IFBench (Pyatkin et al., 2025), Terminal-Bench Hard, and τ^2-Bench Telecom (Barres et al., 2025).

The paper is organised as follows. Section 2 describes the main elements of the prediction market and the teams of AI agents that are used in the experiment. Section 3 discusses the experimental design, including the four information structures, whereas Section 4 lists our

general hypotheses. Section 5 describes the data and we present our results in Section 6. Section 7 concludes.

2 Preliminaries

This section introduces the key concepts. The next section details our experimental design.

2.1 Asset and Information Structure

In each prediction market there are three agents, who trade sequentially for 3,6, or 9 rounds. Agent 1 trades in rounds 1,4,7, agent 2 trades in rounds 2,5,8, and agent 3 trades in rounds 3,6,9. The information structure is determined by 3 signals, d_a, d_b, d_c, each with two possible realisations, 0 ('No') and 1 ('Yes'). A state of nature ω is determined by the realisation of these three signals. Hence, the state space $\Omega = \{a, b, c, d, e, f, g, h\}$ consists of 8 states, shown in Table 2. For example, state $a = (1, 1, 1)$ realises when all three signals resolve to Yes. In all treatments, traders have a common uniform prior on Ω and each signal resolves to 1 with probability 0.5. The draws of each signal are independent. There are two tradable assets, which are complementary. The first asset, $X : \Omega \to \mathbb{R}$ ('betting on Yes', or Yes shares), pays 1 if the answer to the prediction market question is Yes, and 0 otherwise. The second asset, X' ('betting on No', or No shares), pays 1 if the answer to the question is No, and 0 otherwise. Therefore, the price of X is always 1 minus the price of X'. An information and payoff structure (structure in short) determines which signals are received by which trader, and on which states the answer to the question is Yes, so that X pays 1. Section 3.1 describes the four structures we use in the experiment.

2.2 Logarithmic Market Scoring Rule

At the beginning of the market an initial price between 0 and 1 is set for X (the Yes shares) by the market maker. The price of the No shares is set accordingly. Each trader is endowed with £1000 and they trade sequentially in each round. When it is his turn, a trader can buy or sell Yes and No shares, as well as do nothing (Hold). The pricing mechanism uses the Logarithmic Market Scoring Rule (LMSR). The price of Yes is $p_Y = \frac{e^{\beta q_Y}}{e^{\beta q_Y} + e^{\beta q_N}}$, where β is a liquidity parameter and q_Y, q_N denote the number of outstanding Yes and No shares.[3] The LMSR is a special case of a Market Scoring Rule (MSR) (McKelvey and Page (1990), Hanson (2003, 2007)). Because the LMSR uses the Logarithmic Scoring Rule, which is proper, it has the following property. If the trader is risk neutral and myopic, so that he only maximises the expected value of his payoff for the current round, then he will trade up to the point where the price of the Yes shares is equal to his posterior belief that the answer to the prediction market question is Yes. We call this the myopically optimal price.

[3]The liquidity parameter (we use $\beta = 0.01$) determines how quickly the price of X changes when a trader buys or sells. See Cultivate Labs (2021) for an explanation of how the logarithmic MSR is implemented in practice and Schlegel et al. (2022) for axiomatic foundations.

2.3 Information Aggregation

Traders receive their private information before trading, by learning the realisation of their signals. The four information and payoff structures we employ are described in detail in Section 3.1. In all structures, the conjunction of the knowledge of the three traders reveals the true value of X. This means that if traders communicated truthfully, the true value of Yes (0 or 1) would be revealed. If traders are myopic and this is common knowledge, then the true value of X is revealed in 3 rounds, after everyone has traded once, as we show in Section 3.2. However, agents can be strategic and although in some treatments they are allowed to exchange information publicly, this is cheap talk. We say that **information gets aggregated** at state ω if the price of X in the final round is equal (or very close) to the true value of Yes, which is $X(\omega)$.

The main purpose of this paper is to test our theoretical predictions about the market characteristics that improve or hinder information aggregation. Our main measure of success for a market is therefore the logarithmic error between the true value of X at state ω, denoted $y = X(\omega)$, and the final price of X in the market, denoted p: $-[y \ln(p) + (1-y)\ln(1-p)]$.[4] We also examine the trade volume and profitability of traders in markets, however we do not have theoretical predictions for these measures.

2.4 AI Agents

We conducted our experiment with a variety of Large Language Models (LLMs). We employed eight distinct models. The first six were accessed through an Application Programming Interface (API) and they are closed-weight models: Claude Haiku 3.5, Claude Haiku 4.5, Gemini 2.5 Flash, Gemini 3 Flash, GPT-4o, and GTP 5 mini. The last two were open-weight models and they were run locally: gemma3:4b, and qwen3:8b.[5] We formed twelve teams of three traders each, which participated in every treatment at least once. Eight teams were homogeneous, each comprising traders using the same model. The remaining four teams featured a variety of models. As a measure of each model's capabilities, we adopted the Artificial Analysis Intelligence Index (Artificial Analysis Team, 2025).[6] This is one of the most comprehensive and publicly available syntheses of model capabilities. It combines performance across reasoning, knowledge, mathematics, coding, instruction following, long-context reasoning and agentic workflow tasks.[7] See Table 1 for details, where we also report the average intelligence and standard deviation for all heterogeneous teams. Finally, for all models we set temperature at 1. This is the default, neutral setting in many

[4] When calculating the logarithmic error in the data, p is restricted to be in $[\epsilon, 1-\epsilon]$, $\epsilon = 10^{-15}$, so that we avoid having an infinite error. The maximum error with this restriction is around 34.5.

[5] In open-weight models the parameters are available for modification and they can be run locally. However, they are not open-source, because the code or the training data are not necessarily released.

[6] The intelligence scores were accessed on January 6, 2026, at https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index. Note that the scores change over time as new models and new evaluations are added.

[7] See Appendix A in Kim et al. (2025) for details.

APIs, so that text generation is neither too rigid (deterministic) nor chaotic. Figures 7 and 8 in the Appendix show the performance of each model in terms of log error and profits.

Table 1: AI Models, Team Compositions and Intelligence

Team Composition	**Average Intelligence (Standard Deviation)**
Homogeneous Teams (3 Identical Agents)	
3x Gemini 3 Flash	46 (0)
3x GPT-5 mini	41 (0)
3x Claude Haiku 4.5	30 (0)
3x Gemini 2.5 Flash	21 (0)
3x GPT-4o	19 (0)
3x qwen3:8b	15 (0)
3x Claude Haiku 3.5	12 (0)
3x gemma3:4b	7 (0)
Heterogeneous Teams (Mixed Agents)	
Claude Haiku 4.5, Gemini 3 Flash, GPT-5 mini	39 (8.1)
Gemini 3 Flash, GPT-4o, qwen3:8b	26.6 (16.8)
Gemini 2.5 Flash, GPT-4o, Claude Haiku 3.5	17.3 (4.7)
gemma3:4b, qwen3:8b, Claude Haiku 3.5	11.3 (4)

Notes: GPT-4o corresponds to the "gpt-4o-2024-08-06" version, GPT-5 mini to the "gpt-5-mini-2025-08-07" version, Claude Haiku 4.5 to the "claude-haiku-4-5-20251001" version, and Claude Haiku 3.5 to the "claude-3-5-haiku-20241022" version, and gemini 2.5 to the "gemini-2.5-flash-preview-09-2025". Gemini 3 Flash corresponds to the "gemini-3-flash-preview" version that was released on December 17, 2025. Models gemma3:4b and qwen3:8b are open weights models that were downloaded and run locally, whereas the other models were accessed through an API. All eight models have the same version number during the duration of the experiments.

2.5 Prompt Design

When chatting back and forth with an LLM, it appears as if it can recall the conversation and answer accordingly. However, inherently an LLM has no memory, it just 'reads' the whole conversation every time it is called to answer. This means that whenever we call an AI agent to trade in a round, we need to dynamically generate a prompt that describes the prediction market, lists the trading history and previous private or public comments, calculate the trader's portfolio and price impact of various trades, specify the trader's goal and ask for a reply.

The prompt we construct in each round contains the following parts. Part one provides the details of the market, such as the question, whether comments are allowed, who participates, how many rounds are there in the market and what is the current round. Part

two describes the public information, which is shared with all traders, and the private information that is shared with the current trader. The third part provides an explanation of prediction markets, including what are the Yes and the No shares. Part four provides an objective for the trader, to be either myopic or strategic (see Section 3.3).

Part five provides the history of trades up to now, the public comments that have been posted, and the current portfolio of the trader. As we do not rely on LLMs doing their own mathematical calculations, the prompt informs the current trader about the maximum number of Yes and No shares he can buy and sell, as well as the price impact from various trades, for example buying 25% of the maximum shares he can buy. This provides a comprehensive description of how the price will move after the AI agent trades. Part six enumerates the qualitative results from the first wave of the experiment, and it only appears in the Experiment Disclosure treatment. The last part asks the trader for his trading decision (buy, sell, or do nothing), a private justification, and a public comment (if it is allowed by the market). See Appendix B for an example of a prompt.

2.6 Market implementation

We ran our experiments using the Calimantic.com prediction market platform, which was originally developed for running private prediction markets with humans. For this paper, we implemented in Python an Application Programming Interface (API) which allowed the programmatic execution of trades, as well as providing access to past trades and calculating the price impact for hypothetical trades to inform AI agents. We then created calimantic-agents, a program which creates a new market for each combination of the given parameters: teams of AI agents, number of rounds, initial price, structure, strategic, comments allowed. It then orchestrates the trading in rounds, invoking the LLMs using the relevant APIs, retrieving past trades and comments, calculating price impact for various trades, generating prompts for the current trader and executing trades through the Calimantic API. The final output for each market is a text file containing all prompts and decisions of the traders, and a CSV file containing all trades which we use for our quantitative analysis. Figure 1 provides a graphical representation.

3 Experimental Design

Our experimental design focused on six dimensions. The first was the information and payoff structure that was presented to the traders. We considered four structures (t3s111y2, t3s110, t3s111, t3s111o2ye2) that are successively more complex, as we explain in Section 3.1. The second dimension specified the number of trading rounds: three, six and nine. Given that there are three traders in all markets, each AI agent trades once if there are three rounds, twice with six rounds and three with nine rounds. The order of trading is always fixed, so trader 1 trades first, then trader 2 and trader 3. The third dimension related to whether we prompted the AI agents to be myopic, so that they are instructed to maximise the current round's payoff, or strategic, so that they maximise the sum of payoffs from all rounds where

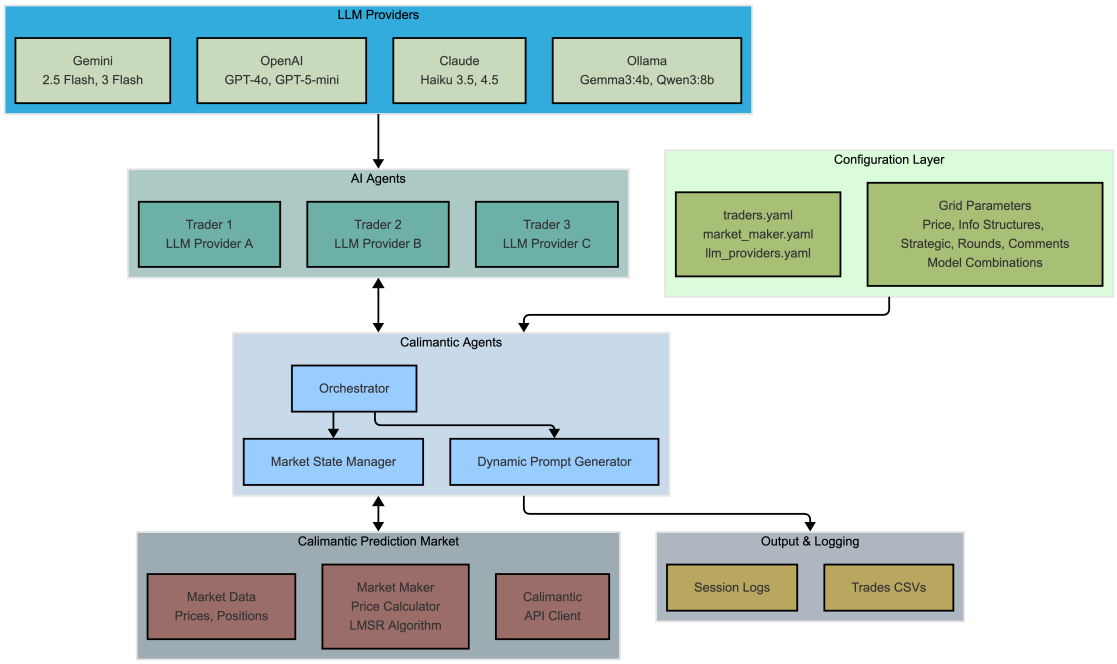


Figure 1: Prediction market platform

they trade. In principle, they are able to carry out a strategy across rounds as they can post a private message to their future self.

The fourth dimension was whether public comments were allowed in the market, so that traders can communicate their private information. The fifth dimension related to the initial price of Yes, set by the uninformed market maker: we implemented three initial prices, 0.3, 0.5, and 0.7. Recall that the true value of Yes is either 0 or 1. Running this first wave, 4x3x2x2x3 experimental design, generated results that we included in the last, "Experiment Disclosure" treatment. In this treatment, AI agents were informed about the qualitative results of the first wave, before trading. In summary, we applied a 4x3x2x2x3x2 experimental design to examine the impact on information aggregation of the difficulty in reasoning about the private information of others, the initial price, the length of trading, the communication, the information provision, and explicitly prompting traders to be strategic or myopic.

3.1 Information and payoffs

In this section, we describe the four information and payoff structures that we presented to the traders. See Appendix B for the exact wording of the generated prompts. There are three signals, d_a, d_b, d_c, that take two values $\{0, 1\}$, and they are drawn independently,

each with probability 0.5. Each signal is framed as a Yes/No question. For example, d_a is the question "Will sales in country A exceed 1 million?". The prediction market question is "Will Company X post next quarter profits that exceed 1 million", and the answer depends on the realisations of the three signals. There are therefore 8 possible states, denoted $\{a, b, c, d, e, f, g, h\}$, uniquely determined by the realisations of these three signals. For example, state $a = (1, 1, 1)$ specifies that all signals resolve to Yes. The first part of Table 2 depicts the realisations of the three signals and the 8 states. We say that security X pays 1 if the answer to the prediction market question is Yes, and 0 otherwise.

A structure specifies the information partition of each trader, the value of X at each state, and the realisations of the three signals that determine the true state. The four structures we consider have the following common characteristics. First, in all states, if the three traders could talk truthfully and combine their private information, the true value of X would become common knowledge. Second, at the true state, if traders are myopic, so that the price of X is always the expected value of the last person who traded, and everyone being myopic is common knowledge, then the price of X becomes equal to the true value of X at the end of the third round, when everyone has traded once. Third, the information and payoff structure (but not the true state) are publicly announced. Finally, the structures become successively harder as traders need to reason about what others know and how public information evolves as they trade.

The four structures are t3s111y2, t3s110, t3s111, and t3s111o2ye2.[8] The first three specify the same information structure for the three traders, depicted in Table 2. In particular, trader 1 is privately informed about d_a, trader 2 is privately informed about d_b, and trader 3 is privately informed about d_c. The partition cells for each trader are depicted in red and black. Trader 1's information partition is $\{\{a, b, c, d\}, \{e, f, g, h\}\}$, the two cells denoted in red and black, and similarly for traders 2 and 3.

The information structure of t3s111o2ye2 specifies that trader 1 is privately informed about d_b and d_c, trader 2 is privately informed about d_a and d_c, and trader 3 is privately informed about d_a and d_b. Note that each trader has four partitions cells, depicted in the last part of Table 2. For example, the information partition of Trader 2 is $\{\{a, d\}, \{b, d\}, \{e, g\}, \{f, h\}\}$. This feature makes interactive reasoning much harder than in the other three structures.

The true value of X in each state is given in the last part of Table 2, for each structure. In structure t3s111y2, X pays 1 if at least two signals resolve to Yes, whereas in t3s110 and t3s111 all three signals need to resolve to Yes. In the last structure, t3s111o2ye2, X pays 1 only when exactly two signals resolve to Yes. We denote with blue the true state in each structure.

[8] The naming of the structures follows this logic: "t3" stands for 3 traders, whereas "s111" and "s110" specify the realisations of the three signals and therefore the true state. In the first three structures an agent is informed about his own signal (e.g. trader 1 is informed about d_a, trader 2 about d_b), except in the last structure where "o2" specifies that an agent is informed about the two other signals (e.g. trader 1 is informed about d_b and d_c). The "y2" specifies that at least 2 signals must resolve to Yes so that X pays 1, whereas "ye2" specifies that exactly 2 signals must resolve to yes. In structures t3s110 and t3s111 all three signals must resolve to yes for X to pay 1. Note that these names are never revealed to the AI agents.

Table 2: Information Structures

States	a	b	c	d	e	f	g	h
Signals				Realisations				
d_a	1	1	1	1	0	0	0	0
d_b	1	1	0	0	1	1	0	0
d_c	1	0	1	0	1	0	1	0
t3s111y2, t3s110, t3s111				Information Structure				
Trader 1	a	b	c	d	e	f	g	h
Trader 2	a	b	c	d	e	f	g	h
Trader 3	a	b	c	d	e	f	g	h
t3s111o2ye2				Information Structure				
Trader 1	a	b	c	d	e	f	g	h
Trader 2	a	b	c	d	e	f	g	h
Trader 3	a	b	c	d	e	f	g	h
				Payoff of Yes ($X =$)				
t3s111y2	1	1	1	0	1	0	0	0
t3s110	1	0	0	0	0	0	0	0
t3s111	1	0	0	0	0	0	0	0
t3s111o2ye2	0	1	1	0	1	0	0	0

3.2 Myopic Prices

To provide a benchmark and understand how the structures compare in terms of complexity, we derive the equilibrium prices in the first three rounds for each structure, in the case where all traders are myopic and this is common knowledge. Because the logarithmic scoring rule that we employ in the prediction market is a proper scoring rule, it has the following property. If a trader is myopic, so that he only cares about the payoff of the current round, the optimal trading is to move the price so that it is equal to his expected value of X, which in this case is equal to his posterior belief that the answer to the question is Yes. Moreover, this optimal behavior does not depend on what the previous price is.

Table 3 denotes the true state in blue and the payoff function of X, for each structure. For example, t3s111y2 describes that the answer to the prediction market question is Yes if at least two signals resolve to Yes, hence the payoff function is $X = (1, 1, 1, 0, 1, 0, 0, 0)$. We denote with blue the true state, which is $a = (1, 1, 1)$ in this structure. Trader 1 is informed that his signal d_a is Yes, hence he considers states $\{a, b, c, d\}$ to be possible. His posterior belief about Yes is 0.75, hence he trades Yes/No shares up until the price of Yes is 0.75.[9] A price of 0.75 reveals to everyone that $\{e, f, g, h\}$ are impossible, because in that case the price would be 0.25. In other words, the public information that is revealed in Round 1 is $\{a, b, c, d\}$. Trader 2 combines this public information with his private information, $\{a, b, e, f\}$, to deduce that the true states is in $\{a, b\}$. As both pay 1, trader 2 trades so that the price of Yes is 1. Trader 3 deduces that the true value of Yes is 1 and his trades do not move the price. In all subsequent rounds (if there are any), the price does not change.

The myopic prices for the other structures are computed similarly. We say that t3s111y2 is the easiest structure because each trader needs to only reason about the signal of one other trader, in order to determine the price of X. In structures t3s110 and t3s111, the traders need to reason about the signals of everyone else. The only exception is trader 3 in structure t3s110, who receives a 0 signal and therefore immediately deduces that the answer is No, independently of what others have traded. For that reason, we consider t3s110 to be easier than t3s111.

Structure t3s111o2ye2 is the most difficult in terms of reasoning about others. First, the partition cells for each trader are four, instead of two. Second, reasoning about which state is no longer possible requires complex counterfactuals. Finally, the myopic price stays the same from round 1 to round 2, however, the public information reduces. This may be difficult for the LLMs to process when reasoning about the private information of others. The true state is a and it is mutual knowledge (all three traders know it) that states d and h are impossible. In the first round, the myopic price of 0.5 makes it common knowledge that d and h are impossible, otherwise the price would be 0. In round 2, the myopic price stays the same at 0.5, which makes it common knowledge that b is impossible, otherwise (and given that d is impossible), the price would be 1. State f is also excluded because the price would be 0 otherwise.

[9] Whether he buys or sells Yes or No shares depends on the previous price. For example, if the initial price of the market was 0.9 for Yes, he would buy No shares as he has no Yes shares to sell. If the initial price was 0.4, he would buy Yes shares. In both cases, his trades would drive the price of Yes to 0.75.

Although the structure is complex, it is derived from the following well-known "muddy children" puzzle, also known as the girls with red hats (Geanakoplos (1992)). Its popularity suggests it is likely part of the training sets of LLMs used in the experiment. However, by looking at the private and public justifications of the AI agents, it seems that noone made the connection.

It is described as follows. There are three girls that wear either a red or a white hat. They can only see the hat of the other two girls but not their own. The true state is that all hats are red and it is therefore mutual knowledge that there is at least one red hat. When asked, each girl announces she does not know the colour of her hat. The teacher then announces that there is at least one red hat, a fact that is mutual (but not common) knowledge. Then, the first girl announces that she does not know the colour of her hat, and so does the second girl. Although these announcements do not change after the teacher's announcement, the public knowledge shrinks and the third girl (if she is very sophisticated) deduces that her hat is red. The teacher's announcement makes it common knowledge that the state 'all hats are white' is impossible. This prompts the gradual reduction in the public knowledge, so that in the third round the action changes as well.

3.3 Myopic vs. Strategic Horizons

The number of periods can influence whether an AI agent acts strategically but the prompt itself may also have that power. We altered the prompt to test whether the AI agents can act strategically if they are instructed to do so, and whether this has an impact on information aggregation. We provided the following prompts.

- **Myopic treatment:** Use reasoning to determine your belief q, then choose your action (Buy, Sell, or Hold: Yes and No shares). Maximize your expected payoff in this round only, based on your belief q and the current price p, ignoring future rounds.
- **Strategic treatment:** Use reasoning to determine your belief q, then choose your action (Buy, Sell, or Hold: Yes and No shares). Maximize the sum of your expected payoffs over all trading rounds, based on your belief q and the current price p. Consider how your current trade affects the price and the beliefs of others in future rounds.

4 General hypotheses

In order to formulate our hypotheses, we use the following result by Ostrovsky (2012), adapted to our setting.

Theorem 1. *If security X is separable under information structure Π, then for any prior distribution μ, for any strictly proper scoring rule s, initial value y_0, and discount factor $\gamma \in (0, 1]$, in any Nash equilibrium, information gets aggregated.*

In Appendix C, we formally define the notion of separability and show that the securities in all four structures are separable. Given that in the experiment we use a logarithmic scoring

Table 3: Myopic Price of Yes and Public Information

	Myopic price	Public information	Volume	Profits
t3s111y2 - Easy				
Payoff: $X = (1, 1, 1, 0, 1, 0, 0, 0)$				
Trader 1	0.75	(a, b, c, d)	109	£46
Trader 2	1	(a, b)	880	£29
Trader 3	1	(a, b)	0	£0
t3s110 - Medium				
Payoff: $X = (1, 0, 0, 0, 0, 0, 0, 0)$				
Trader 1	0.25	(a, b, c, d)	109	£46
Trader 2	0.5	(a, b)	109	-£40
Trader 3	0	b	990	£69
t3s111 - Hard				
Payoff: $X = (1, 0, 0, 0, 0, 0, 0, 0)$				
Trader 1	0.25	(a, b, c, d)	109	-£63
Trader 2	0.5	(a, b)	109	£69
Trader 3	1	a	990	£69
t3s111o2ye2 - Very Hard				
Payoff: $X = (0, 1, 1, 0, 1, 0, 0, 0)$				
Trader 1	0.5	$(a, e), (b, f), (c, g)$	56	£6
Trader 2	0.5	$(a, c), (e, g)$	0	£0
Trader 3	0	a	990	£69

Notes: The table depicts the myopic price per round, the public information that is revealed, the volume (number of shares) and the profit of the AI agent who trades at that round. We denote with blue the true state. The myopically optimal price of X is the expected value of X given the private and public information for each trader. The myopic price reveals information to all other traders, who use it to form their updated beliefs and trade. The myopically optimal price does not depend on the previous price, however the profits and volume do. For trader 1, the volume and profits are average, over the three possible initial prices, 0.3, 0.5, and 0.7. For the other two traders, volume and profits are the same because trader 1 always moves the price to the same level, irrespective of the initial price.

rule, which is a strictly proper scoring rule, this theorem says that information should get aggregated in all four structures, irrespective of their complexity.

Hypothesis 1. *Information aggregation is unaffected by the complexity of the structure.*

The theorem specifies that whether traders are myopic or strategic does not matter. In the strategic treatment we instruct traders to maximise their payoffs over all future rounds. This is consistent with a discount factor of $\gamma = 1$, although we do not mention this term in the prompt. The myopic treatment is consistent with $\gamma = 0$. Although the theoretical model excludes a zero discount factor, it is straightforward to show that as γ converges to 0, traders will behave almost myopically and reveal their expectations almost truthfully.

Hypothesis 2. *Information aggregation is unaffected by prompting AI agents to be strategic or myopic.*

The theoretical model does not allow traders to post comments. However, it is straightforward to show that this should not hinder information aggregation. Roughly, the reason is that if there is no information aggregation in the long term, separability implies that at least one trader is able to achieve strictly positive profits, irrespective of what other traders do or say.

Hypothesis 3. *Information aggregation is unaffected by allowing AI agents to post public comments.*

A direct application of the theorem ensures that the initial price should not impact information aggregation.

Hypothesis 4. *Information aggregation is unaffected by manipulating the initial price of the market.*

An important difference from the formal model is that it specifies infinitely many rounds of trading, whereas in the current experiment we have up to 9. Although it is impossible to have infinitely many rounds in an experiment, we could simulate an infinite horizon by specifying that, in every round, the game ends with some probability p.[10] We chose not to include this extra treatment for the following two reasons. First, the prompt is already long and complicated, so it is not clear that AI agents would be able to understand and act differently if we added this detail. Second, we considered that it would be more interesting and practical to understand the effect of lengthening the duration of the market on information aggregation.

Hypothesis 5. *Information aggregation does not deteriorate as the duration of the market increases.*

[10] See Galanis et al. (2024) for such an experiment with humans in a prediction market.

Table 3 describes the myopically optimal prices, trading volume, and profits, for each of the four structures and for each trader, using the formulas derived in Appendix D. Recall that the initial price is 0.3, 0.5, or 0.7, which influences both the trading volume and the profits of the first trader. The other two traders are unaffected because the first trader always moves the price to the same level, irrespective of the initial price. By computing the average trading volume for each structure and the average profits for each trader, we formulate the next two hypotheses. Table 15 summarises all calculations. Note that these hypotheses are implicitly assume that all traders are myopic, hence all markets resolve in round 3 and there is no trading in subsequent rounds.

The average trading volume across all initial prices are as follows: Easy (990), Medium (1133), Hard (1210), Very Hard (1046). We therefore have the following hypothesis.

Hypothesis 6. *Trading volume is ordered, from lowest to highest, as follows: Easy (t3s111y2) < Very Hard (t3s111o2ye2) < Medium (t3s110) < Hard (t3s111).*

We also compute the average profits for each trader, across all structures and initial prices: Trader 1 (9), Trader 2 (14), and Trader 3 (52). We therefore have the following hypothesis.

Hypothesis 7. *Profits are positive for all traders and ordered, from lowest to highest, as follows: Trader 1 < Trader 2 < Trader 3.*

A distinctive feature of the prediction market is that it is 'constant-sum' in profits. If the initial price of a market is p_0 and the final price is p_1, then the sum of the profits of all traders is constant, irrespective of how prices fluctuate. This means in the current setting that the difficulty of the structure should not influence total profits, and therefore average profits, which are alway 25, as shown in Table 15. We therefore have the following hypothesis.

Hypothesis 8. *Average profits are not correlated with the difficulty of the structure.*

We now turn to the role of intelligence on market outcomes. The following hypotheses are not based on the theoretical model of Ostrovsky (2012), but on previous experiments with humans. First, we conjecture that information aggregation improves as AI agents become smarter, as they should be able to reason better about the information of others.

Hypothesis 9. *Information aggregation improves as the average intelligence of the group increases.*

Corgnet et al. (2018) show in a market experiment with humans that 'smarter' traders are also more profitable. They measure intelligence with three metrics. Fluid intelligence measures their ability to compute correctly and draw statistical inferences. Cognitive reflection measures their ability to avoid behavioral biases and update their beliefs when observing market orders. Theory of the Mind measures whether they can correctly assess the informational content of trades.

Hypothesis 10. *Individual intelligence is positively correlated with individual profits.*

Hypothesis 11. *Group (average) intelligence is negatively correlated with individual profits.*

5 Data

The dataset consists of observations from 3,500 prediction markets involving a total population of 10,500 distinct trading agents (3 per market). Table 4 summarises the key market-level variables, where the upper panel describes the first wave (1772 markets), whereas the bottom panel describes both waves.[11] We describe the first wave as the statistics do not change when adding the second wave. A defining characteristic of the data is the extreme skewness in market performance metrics. While the median logarithmic score is low (0.089)—indicating that the typical market converges to a high-probability estimate of the true state—the mean score is drastically higher (5.18), driven by a subset of catastrophic failures where the logarithmic penalty creates scores as high as 34.54. A similar pattern is observed in the Squared Error, where the mean (0.267) is significantly larger than the median (0.007).

Trading activity varied substantially across sessions. The average market generated a volume of approximately 1,859 shares traded, though this ranged from inactive markets (0 volume) to speculative frenzies reaching over 14,290 shares. The population of AI agents had a mean intelligence score of 23.56 (SD = 12.21), spanning a range from 7 to 46.

Table 5 breaks down market outcomes by information structure and duration, for the first wave and for both waves. The data for the first wave (and similarly for both waves) reveals a strict hierarchy of difficulty that validates our experimental design. In the Easy structure (t3s111y2), markets performed reliably well, with a Mean Squared Error (MSE) as low as 0.07 (for 6 rounds) and a 'Crash Rate' (markets with a log error above 20) of only 1.4%. As complexity increases, performance degrades sharply and monotonically. In the Medium structure (t3s110), the average error nearly triples (MSE $\approx$ 0.18), and by the Very Hard structure (t3s111o2ye2), the market is effectively broken: the MSE reaches 0.43– suggesting a final price of 0.65 when the true value of the security is 0, whereas the Crash Rate jumps to over 22%. The table also serves as a randomization check. The Average Intelligence (Avg IQ) column shows that agent capabilities are evenly distributed across all treatment cells (varying tightly between 22.9 and 23.8), confirming that the observed differences in performance are driven purely by the structural difficulty of the task and not by chance imbalances in agent composition.

6 Results

To keep things simple, we first report our results from the first wave (1772 markets), ignoring the information provision treatment. We then describe the information provision treatment and show in the Appendix the regressions using the full sample (3500 markets). It turns out that most of our results are robust when running the same regressions in the full sample. The graphs that we show utilise the full sample.

[11] With 12 teams of AI agents, the first wave should have consisted of 4x3x2x2x3x12 = 1728 markets. The extra 44 markets were due to a coding error. We decided to keep these markets for completeness.

Table 4: Descriptive Statistics

Statistic (First Wave)	N	Mean	St. Dev.	Min	Median	Max
Squared Error	1,772	0.266	0.382	0	0.007	1
Logarithmic Score	1,772	5.181	11.770	0	0.089	34.540
Trading Volume	1,772	1,859	1,824	0	1,298	14,290
Avg. Intelligence (Group)	1,772	23.562	12.205	7	19	46
Intelligence SD	1,772	2.875	5.030	0	0	16.862
Initial Squared Error	1,772	0.274	0.164	0.090	0.250	0.490
Initial Logarithmic Score	1,772	0.745	0.347	0.357	0.693	1.204

Statistic (Both Waves)	N	Mean	St. Dev.	Min	Median	Max
Squared Error	3,500	0.277	0.390	0	0.010	1
Logarithmic Score	3,500	5.571	12.152	0	0.105	34.540
Trading Volume	3,500	1,898	1,828	0	1,353	14,290
Avg. Intelligence (Group)	3,500	23.668	12.216	7	19	46
Intelligence SD	3,500	2.847	4.993	0	0	16.862
Initial Squared Error	3,500	0.275	0.164	0.090	0.250	0.490
Initial Logarithmic Score	3,500	0.748	0.348	0.357	0.693	1.204

6.1 Information Aggregation

We present our main results in Table 6. Column (1) presents an OLS regression where the dependent variable is log error (the logarithmic scoring rule error). Because the log error has a wide range (from 0 to 34.5), the mean estimator is highly sensitive to outliers—markets that crash by pricing the security X at 0 when its true value is 1.[12] Columns (2) and (3) present Quantile Regressions that allow us to inspect specific parts of the error distribution. Column (2) estimates the model at the median ($\tau = 0.5$), representing the typical market outcome. This estimator is robust to outliers, filtering out the effect of catastrophic crashes to reveal how the typical agent behaves under varying conditions. Column (3) estimates the model at the 80th percentile ($\tau = 0.8$), representing the tail risk.

We employ a mixed-contrast coding specification across all three models. For the primary design variables—Structure and Market Duration—we utilise treatment contrasts, setting the easiest structure (t3s111y2) and the shortest duration (3 Rounds) as the reference baseline. Consequently, coefficients for these variables represent the marginal penalty of increasing complexity or duration relative to this baseline. For the other conditions—Initial Log Error, Strategic, and Comments—we employ sum contrasts (deviation from the mean). In this specification, coefficients represent the deviation of a specific condition (e.g., Medium Initial

[12]Recall that the log error between actual (y) and predicted (p) is $-(y \ln ln(p) + (1-y)\ln(1-p))$, where the predicted probability is bounded, $p \in [\epsilon, 1-\epsilon]$, $\epsilon = 10^{-15}$, to avoid having an infinite log error.

Table 5: Market Outcomes by Information Structure and Duration

Structure	Rounds	N	MSE	SE (Sq Er)	Mean Log	SE (Log)	Crash Rate	Avg Volume	Avg IQ
First Wave									
Easy	3	144	0.108	0.023	2.56	0.73	6.9%	1,255	23.8
	6	144	0.070	0.018	0.69	0.34	1.4%	1,808	23.8
	9	144	0.078	0.018	1.58	0.58	4.2%	2,102	23.8
Medium	3	159	0.183	0.027	3.56	0.80	9.4%	1,508	23.0
	6	160	0.199	0.029	3.61	0.79	9.4%	2,065	22.9
	9	156	0.202	0.030	4.50	0.90	12.2%	2,639	23.1
Hard	3	144	0.387	0.036	9.30	1.24	25.7%	1,328	23.8
	6	144	0.319	0.034	4.80	0.92	11.8%	1,865	23.8
	9	144	0.332	0.035	6.70	1.10	18.1%	2,356	23.8
Very Hard	3	145	0.463	0.032	8.61	1.18	22.8%	1,229	23.7
	6	144	0.433	0.032	7.46	1.12	19.4%	1,752	23.8
	9	144	0.436	0.032	8.56	1.19	22.9%	2,354	23.8
Both Waves									
Easy	3	288	0.131	0.018	2.76	0.53	7.3%	1,264	23.8
	6	288	0.081	0.013	1.15	0.34	2.8%	1,844	23.8
	9	288	0.081	0.014	1.39	0.37	3.5%	2,140	23.8
Medium	3	303	0.173	0.020	3.46	0.57	9.2%	1,521	23.4
	6	304	0.211	0.021	4.02	0.60	10.5%	2,072	23.3
	9	300	0.223	0.022	5.25	0.69	14.3%	2,650	23.4
Hard	3	288	0.407	0.026	9.50	0.88	26.0%	1,377	23.8
	6	288	0.307	0.024	4.52	0.63	11.1%	1,995	23.8
	9	288	0.339	0.025	6.95	0.78	18.8%	2,515	23.8
Very Hard	3	289	0.469	0.023	9.11	0.85	24.2%	1,234	23.7
	6	288	0.450	0.023	8.29	0.82	21.9%	1,739	23.8
	9	288	0.456	0.023	9.39	0.87	25.3%	2,408	23.8

Notes: The crash rate measures the percentage of markets that have a log error above 20.

Error) from the global average of that variable, rather than from an arbitrary reference group. Thus, the model intercept captures the predicted log-error of the baseline market design (Easy Structure, Short Duration) under average environmental starting conditions.[13]

In our OLS specification, all coefficients for the three structures are statistically significant ($p < 0.001$) and positive, indicating that as the structure becomes more difficult, the log error increases. The intercept (≈ 7.548) implies that the average baseline market in the easiest structure (t3s111y2) already fails spectacularly, because the final price of the security is $e^{-7.384} \approx 0.00062$ when the value of the security is 1. Using the easiest structure as the baseline, the hardest structure (t3s111o2ye2) was associated with a statistically significant increase in logarithmic error of $\beta = 6.58$ ($p < 0.001$). This coefficient represents an even worse failure in probability estimation. Since the logarithmic error is defined as $-ln(\text{price of yes})$ when Yes is the true outcome, a difference of 6.58 implies that the baseline structure assigns a probability to the true outcome that is approximately $e^{6.58} \approx 720$ times higher than the probability assigned by the hardest structure. Effectively, this suggests that in complex environments, the average market price diverges so much from the truth that it signals near-certainty in the wrong outcome. This extreme mean effect is potentially driven by the nature of the log error, where a few bad markets can skew the average log error significantly.

To disentangle typical market behavior from these catastrophic outliers, we employ a quantile regression at the median ($\tau = 0.5$). The intercept of the median regression (≈ 0.018) implies that the typical baseline market in the easiest structure converges almost perfectly to the truth ($p = e^{-0.018} \approx 98\%$). The coefficient for the hardest structure is $\beta = 0.7$ ($SE = 0.083$, $p < 0.001$). This corresponds to an implied probability of $e^{-0.7} \approx 50\%$, indicating that the typical market outcome in the hardest structure is indistinguishable from random guessing. At the medium structure, the median also scores very well ($p \approx 98\%$), whereas at the hard structure success drops to $p = e^{-0.286} \approx 75\%$. Thus, while the very hard structure causes the average market to hallucinate, it causes the median market to be no better than the toss of a coin.

Summarizing, we reject our hypothesis that information aggregates across all structures. We find that as complexity increases, aggregation deteriorates. Figure 2 confirms this result in terms of the average and median log error.[14]

Result 1. *Information aggregation deteriorates as the structure becomes more complex: Easy (t3s111y2) < Medium (t3s110) < Hard (t3s111) < Very Hard (t3s111o2ye2).*

Our results confirm the hypothesis that neither communication nor strategic prompting significantly influences information aggregation. The main effect of Comments Allowed is statistically insignificant in the OLS model ($\beta = 0.191, p > 0.47$), as well as in the median

[13] An alternative specification would be to have the Squared Error as the dependent variable, instead of the Log Error. However, because our prediction market uses the logarithmic scoring rule, AI agents try to minimise Log Error to maximise their profits, hence the Log Error seems more appropriate. The disadvantage of this approach is the sensitivity of the mean estimator to outliers, hence we also employ the two quantile regressions.

[14] Recall that all graphs incorporate the full sample of 3500 markets, as the results are robust.

Table 6: Information Aggregation Without Disclosure

	Dependent variable:		
	Log Error		
	Mean (OLS)	Median (Q50)	Tail Risk (Q80)
	(1)	(2)	(3)
Constant	7.548*** (0.714)	0.018 (0.019)	3.170** (1.138)
Comments Allowed	0.191 (0.265)	0.001 (0.003)	0.000 (0.032)
Duration: 6 Rounds	−1.817** (0.639)	−0.001 (0.005)	−0.057 (0.177)
Duration: 9 Rounds	−0.627 (0.677)	−0.001 (0.005)	−0.028 (0.165)
Strategic Prompt	−0.053 (0.487)	−0.001 (0.003)	−0.000 (0.113)
Medium (t3s110)	2.200*** (0.573)	−0.000 (0.003)	0.038 (1.808)
Hard (t3s111)	5.327*** (0.730)	0.286** (0.106)	12.470· (7.451)
Very Hard (t3s111o2ye2)	6.658*** (0.713)	0.700*** (0.084)	21.771*** (4.744)
Initial Error: Low	0.014 (0.493)	0.000 (0.004)	0.000 (0.069)
Initial Error: Medium	0.225 (0.508)	−0.000 (0.003)	−0.000 (0.063)
Average Intelligence	−0.221*** (0.018)	−0.0004 (0.0004)	−0.069** (0.025)
Intelligence SD	0.044 (0.060)	−0.0003 (0.0003)	−0.052* (0.026)
6 Rounds x Strategic	0.067 (0.639)	0.001 (0.004)	−0.000 (0.128)
9 Rounds x Strategic	−0.076 (0.676)	0.001 (0.004)	−0.028 (0.161)
Medium Struct x Initial Error (Low)	1.309 (0.844)	−0.000 (0.005)	0.019 (3.595)
Hard Struct x Initial Error (Low)	0.504 (1.039)	0.032 (0.124)	19.889 (12.267)
Very Hard x Initial Error (Low)	2.861** (1.050)	0.013 (0.167)	10.425* (5.186)
Medium Struct x Initial Error (Medium)	−0.725 (0.803)	0.000 (0.004)	0.019 (1.810)
Hard Struct x Initial Error (Medium)	−0.236 (1.047)	0.077 (0.149)	−9.945 (12.141)
Very Hard x Initial Error (Medium)	−0.337 (1.022)	−0.007 (0.083)	10.080 (9.342)
Robust F-Statistic	12.84*** (df = 19; 1752)		
Observations	1,772	1,772	1,772
R^2	0.115		

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. Q50 and Q80 are quantile regressions. Q50 models the median log error, whereas Q80 models the bottom 20% markets. OLS model uses HC1 robust standard errors. Quantile models use bootstrapped standard errors. All three models use a mixed-contrast specification. Rounds and Structure use treatment contrasts with Round 3 and Structure t3s111y2 as baselines. All other controls (Initial Error, Strategic, Comments) use sum contrasts, where coefficients represent deviations from the grand mean.

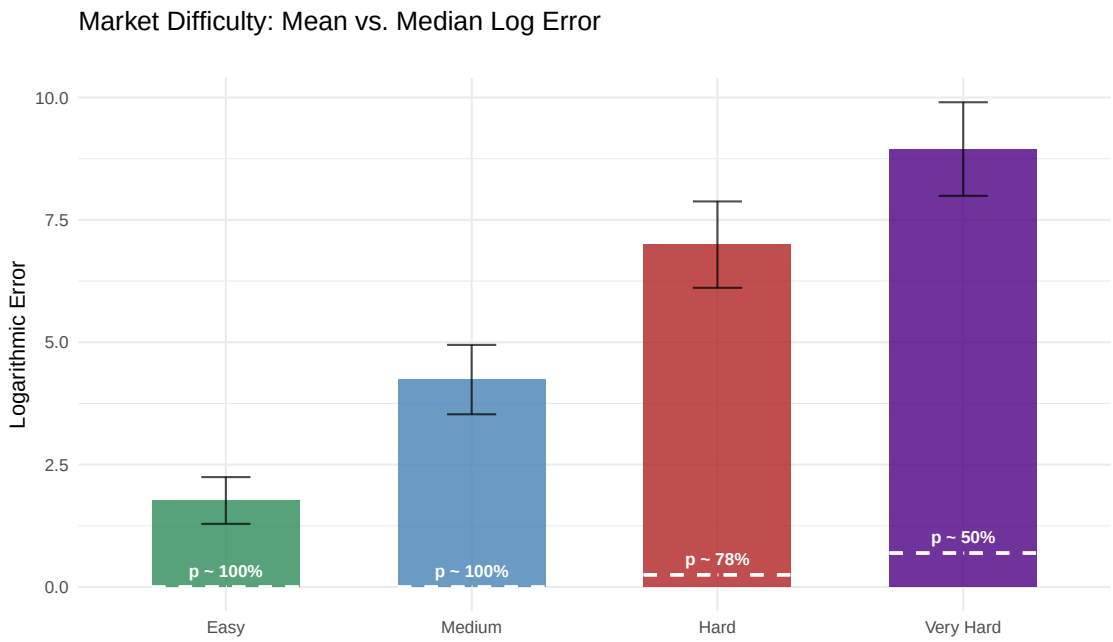


Figure 2: **The Complexity Effect.** As the structure becomes more complex, the information aggregation deteriorates. The dashed line, $p(True)$, indicates the median price of the security that paid 1.

and tail-risk quantile regressions. Similarly, the Strategic Prompt coefficient is statistically indistinguishable from zero ($\beta = -0.053, p > 0.9$). This lack of significance holds even when controlling for interaction effects with duration, suggesting that the core aggregation mechanism is robust to these variations.[15]

Result 2. *Information aggregation is unaffected by allowing AI agents to post public comments.*

Result 3. *Information aggregation is unaffected by prompting AI agents to be strategic or myopic.*

Communication and strategic prompting may fail to influence information aggregation for at least two reasons. The first is that highly sophisticated AI agents behave according to the theoretical predictions. The second is that they are not very sophisticated, so they fail to understand the prompt, or they cannot plan ahead and implement a strategic plan.

[15]To ensure that our null results regarding communication and strategic prompting were not due to low statistical power, we conducted a sensitivity analysis. With a sample size of N=1,772 and a model specification with 19 predictors, our design possesses 80% power to detect an effect size (f^2) as small as 0.012 at the 5% significance level. This corresponds to a minimum detectable increase in $\mathbb{R}^2$ of approximately 0.012. In other words, if strategic behavior or communication explained even 1.2% of the variance in market outcomes, our analysis would have detected it. The fact that these coefficients remain statistically insignificant suggests that their influence on information aggregation is not merely unobserved, but practically negligible.

To understand their thought process better, we analyse their public comments and private justifications in Section 6.4.

The initial log error is the log error between the initial price and the true value of X. The low initial log error is when the initial price is 0.7 and the true value of X is 1, or the initial price is 0.3 and the true value is 0, whereas the medium initial log error is when the initial price is 0.5. We observe no significant main effect of the low initial log error on market performance ($\beta = 0.1, p > 0.9$), and similarly for the medium initial log error. This suggests that under average conditions, the markets are highly robust: information aggregation is independent of the initial price.

Result 4. *Information aggregation is unaffected by manipulating the initial price of the market.*

It is interesting to note that this robustness does not hold in the most complex environment. The interaction between the hardest information structure and the low initial log error (Very Hard × Initial Error (Low)) reveals a critical mechanism of market failure. We observe a large, positive, and statistically significant coefficient in both the OLS mean regression ($\beta = 2.86, p < 0.01$) and the 80th percentile tail regression ($\beta = 10.42, p < 0.05$). This positive coefficient indicates a 'signal inversion' effect: in the most complex environment, markets that begin with a more accurate price signal (lower initial error) paradoxically end with significantly worse performance than those starting with higher error. The magnitude of this effect in the tail model ($\beta = 10.42$) is very high, suggesting that for fragile markets, a low initial error does not act as an anchor for truth but rather as a catalyst for hallucination. Crucially, this effect is entirely absent in the median regression ($\beta = 0.01, p > 0.9$), suggesting that the 'typical' market ignores the initial price, which is consistent with the theory.

Our results support the hypothesis that information aggregation does not deteriorate when the number of rounds goes from 3 to 9 (Figure 3). However, we do find a significant beneficial effect of moderate duration, from 3 to 6. In Model 1 (OLS), increasing duration from 3 to 6 rounds significantly reduces log error ($\beta = -1.81, p < 0.01$). This improvement diminishes and becomes insignificant at 9 rounds ($\beta = -0.62, p > 0.3$). Furthermore, the quantile regressions (Models 2 and 3) show no effect of duration on the median or 80th percentile.

Result 5. *Information aggregation does not deteriorate as the duration of the market increases.*

Our analysis strongly supports the hypothesis that average intelligence improves information aggregation (Figure 4). In the OLS specification, Average Intelligence is a highly significant predictor of market success ($\beta = -0.221, p < 0.001$), with smarter teams achieving consistently lower log errors. This imeans that an increase from the dumbest groups (Average Intelligence =7) to the smartest (46) reduces the logarithmic error penalty by 8.6 points. Because logarithmic scoring scales exponentially, an 8.6 reduction in error implies that the smartest agents assign a probability to the true outcome that is roughly e^8.6 (over

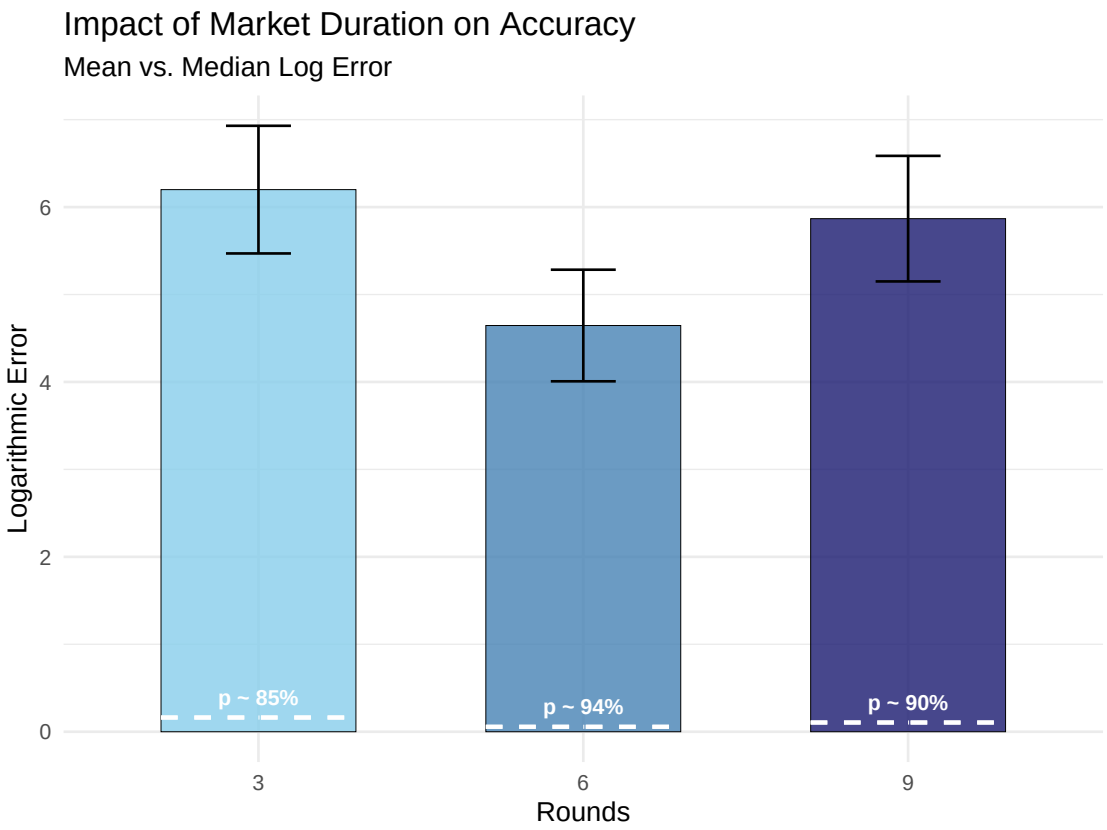


Figure 3: **The Duration Effect.** Although there is an improvement in information aggregation from 3 to 6 rounds, there is no statistically significant difference between 3 and 9 rounds.

5,000 times) larger than the probability assigned by the least intelligent agents. This massive multiplier illustrates intelligence primarily acting as a safeguard against catastrophic market mispricing (where low-intelligence agents confidently drive the correct asset price to near-zero).

While higher intelligence improves performance on average, the quantile regressions reveal that this is driven primarily by risk mitigation rather than universal optimisation. We find no significant effect of intelligence on the median market ($\beta \approx 0, p > 0.5$), suggesting that typical markets reach an accuracy ceiling regardless of marginal gains in cognition. However, intelligence plays a critical role in the tail of the distribution (Q80), where it significantly reduces error ($\beta = -0.069, p < 0.01$). This implies that average intelligence acts as a form of crash protection: improving the bottom markets rather than the median. The Intelligence SD (Standard Deviation) measures whether increasing the diversity of the group improves information aggregation. We found an effect only in the tail of the distribution (Q80), where it significantly reduces error ($\beta = -0.052, p < 0.05$).

Result 6. *Information aggregation improves as the average intelligence increases.*

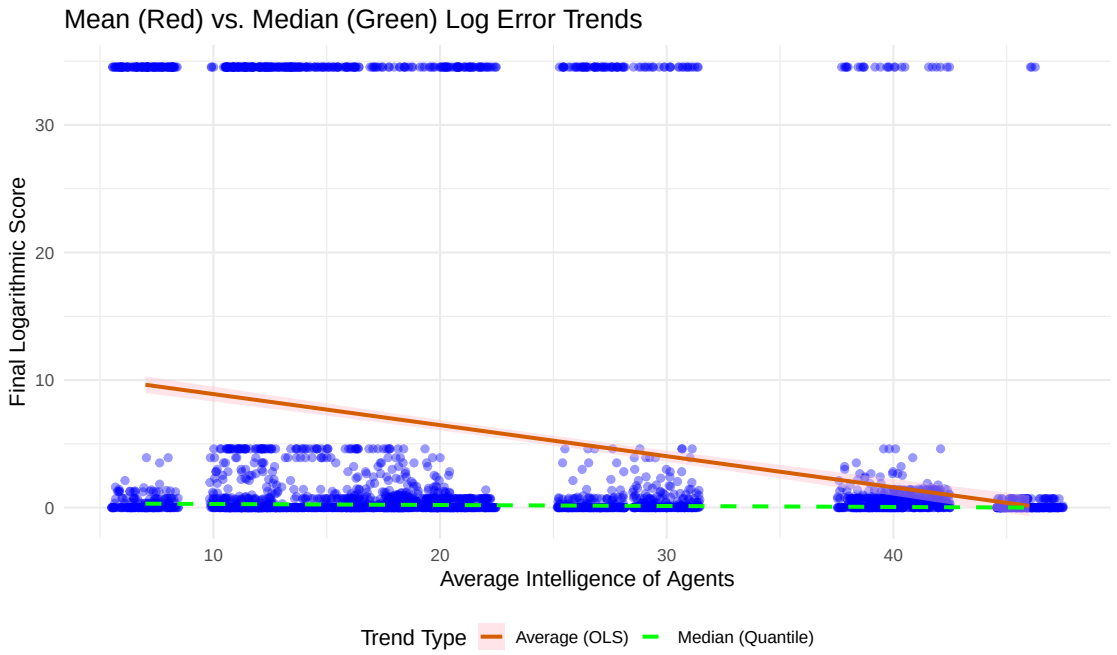


Figure 4: **The Intelligence Effect.** As the AI agents become smarter, information aggregation (mean log error) improves, but the median log error is unaffected. Points are jittered horizontally to show density of markets with the same intelligence.

6.2 Trading Volume and Profits

In Table 7, we report two regressions, on trading volume and individual profits, using the same independent variables as in the previous models on information aggregation.

We first find that trading volume across structures almost follows the myopic theoretical predictions. The only difference is that two highest volume markets are Hard (t3s111) $<$ Medium (t3s110), instead of Medium (t3s110) $<$ Hard (t3s111).

Result 7. *Trading volume is ordered, from lowest to highest, as follows: Easy (t3s111y2) $<$ Very Hard (t3s111o2ye2) $<$ Hard (t3s111) $<$ Medium (t3s110).*

Second, we reject our hypothesis that profits are positive and increase with the order of trading (Figure 5). Relative to the first mover, agents who trade second suffer a significant profitability penalty, rejecting our hypothesis ($\beta = -42.1, p < 0.001$). However, agents who trade third enjoy a massive profitability premium ($\beta = 77.3, p < 0.001$), which is consistent with our prediction that the trader 3 has the highest profits. The averages in Figure 5 show that all traders have negative profits, rejecting our hypothesis that all traders make strictly positive profits. See Figure 7 for a breakdown of profits by AI model, where the only model with positive profits is Gemini 3 Flash.

Result 8. *Profits are negative for all traders and ordered, from lowest to highest, as follows: Trader 2 $<$ Trader 1 $<$ Trader 3.*

We also reject the hypothesis that average profits are not correlated with the complexity of the structure. The negative coefficients observed in the Medium ($\beta = -35.85, p < 0.05$), Hard ($\beta = -71.16, p < 0.001$), and Very Hard ($\beta = -101.64, p < 0.001$) treatments show that as complexity increases, average profits decrease. This is not surprising given Result 1, because the deterioration of information aggregation is equivalent to an increasing difference between final price and true value of the security, which in turn implies lower total and average profits.

Result 9. *Average profits decrease as the structure becomes more complex: Easy (t3s111y2) < Medium (t3s110) < Hard (t3s111) < Very Hard (t3s111o2ye2).*

We also find a strong negative relationship between average intelligence and volume ($\beta = -27.9, p < 0.001$), supporting the view that higher cognitive capacity allows agents to agree on the equilibrium price with fewer transactions.[16] Conversely, intelligence standard deviation is positively associated with volume ($\beta = 22.3, p < 0.001$), suggesting that trading activity is fuelled by cognitive heterogeneity—possibly because of the arbitrage opportunities created when smarter AI agents exploit the mispricing of less capable traders in order to increase their profits.

Furthermore, we observe a substitution effect between communication and trading: allowing public comments reduces trading volume ($\beta = -85.0, p < 0.05$), implying that verbal coordination reduces the need for costly signalling through price mechanisms. Finally, the interaction between the Very Hard structure and low log initial error—previously identified as leading to deteriorating information aggregation—is associated with a massive surge in volume ($\beta = 414, p < 0.05$). This indicates that the catastrophic mispricing of the security is not accompanied by market paralysis, but by excessive trading.

The analysis of individual agent profitability (Table 7, Column 2) reveals that financial success in these markets is driven by relative rather than absolute cognitive advantage. We find a strong positive effect of individual intelligence on profit ($\beta = 14, p < 0.001$), but a nearly equal and opposite negative effect of average intelligence ($\beta = -11, p < 0.001$). This confirms that agents benefit from their own intelligence but suffer from the intelligence of their competitors. In highly intelligent markets, prices correct too quickly for any single agent to capture significant arbitrage rents; the ideal condition for individual profit is therefore high personal capability situated within a low-capability environment.

Result 10. *Individual intelligence is positively correlated with individual profits.*

Result 11. *Group (average) intelligence is negatively correlated with individual profits.*

6.3 Information Provision

Can AI agents improve their performance by receiving feedback from past play? Yang et al. (2023) propose an information provision technique that leverages LLMs as optimisers. The

[16] Recall that in all four structures, the true value of the security can be revealed in three rounds if everyone plays the myopic best and this is common knowledge, as we showed in Section 3.2.

Table 7: Trading Volume and Profits Without Disclosure

	Dependent variable:	
	Trading Volume	Individual Profits
	(1)	(2)
Constant	1,795.988*** (126.550)	−80.719*** (14.897)
Comments Allowed	−85.077* (41.093)	−2.723 (5.216)
Duration: 6 Rounds	543.398*** (83.763)	2.447 (12.431)
Duration: 9 Rounds	1,037.835*** (106.092)	−5.396 (12.788)
Strategic Prompt	65.396 (53.805)	0.491 (8.468)
Medium (t3s110)	308.391** (114.608)	−30.548* (12.647)
Hard (t3s111)	128.183 (119.400)	−69.955*** (15.440)
Very Hard (t3s111o2ye2)	55.372 (118.947)	−98.904*** (14.402)
Initial Error: Low	−180.641 (115.824)	−10.767 (12.249)
Initial Error: Medium	−14.704 (120.252)	−4.386 (12.619)
Average Intelligence	−27.925*** (2.657)	−11.019*** (1.268)
Individual Intelligence		14.079*** (1.203)
Intelligence SD	22.347*** (6.159)	0.606 (1.151)
Trades Second		−42.127*** (11.959)
Trades Third		77.390*** (11.924)
6 Rounds x Strategic	−9.810 (83.760)	1.725 (12.431)
9 Rounds x Strategic	−87.983 (106.100)	3.680 (12.789)
Medium Struct x Initial Error (Low)	273.357· (159.892)	−21.173 (17.753)
Hard Struct x Initial Error (Low)	87.880 (165.985)	−13.015 (21.725)
Very Hard x Initial Error (Low)	414.759* (166.635)	−32.672 (20.885)
Medium Struct x Initial Error (Med)	41.568 (162.442)	4.394 (17.812)
Hard Struct x Initial Error (Med)	62.984 (172.088)	11.491 (21.633)
Very Hard x Initial Error (Med)	−238.720 (161.232)	−1.778 (19.916)
Robust F-Statistic	14.158*** (df = 19; 1752)	16.73*** (df = 22; 5293)
Observations	1,772	5,316
R^2	0.104	0.060

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. Both models use HC1 robust standard errors and a mixed-contrast specification. Rounds, Information Structure, and Trader Position use treatment contrasts, with 'Round 3', 't3s111y2', and 'Trades First' serving as their respective baselines. All other controls (Initial Error, Strategic, Comments) use sum contrasts, where coefficients represent deviations from the grand mean.

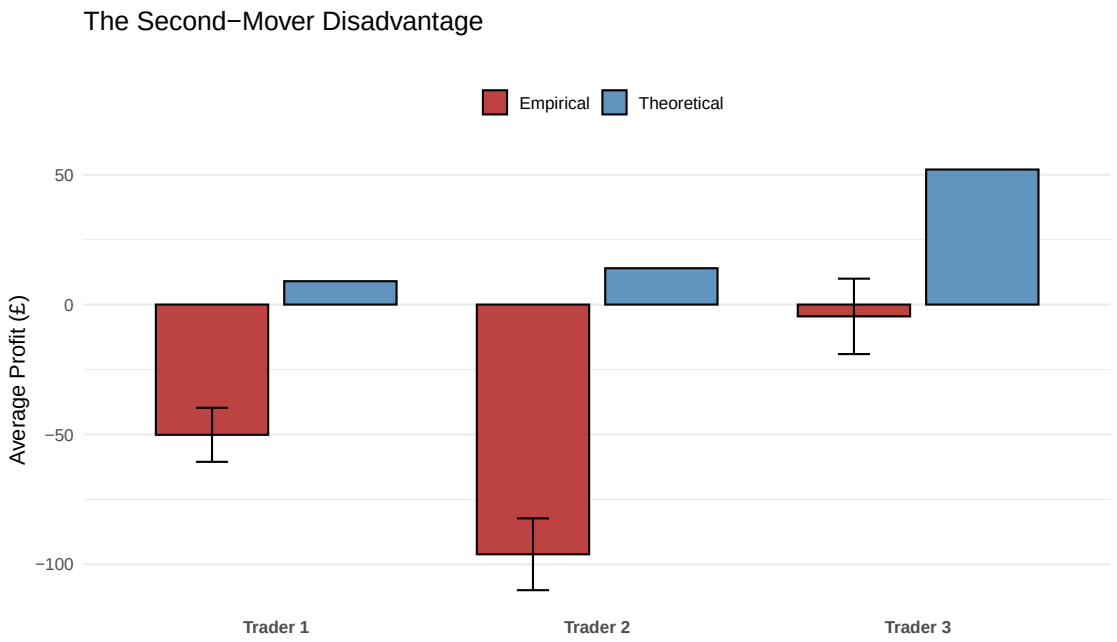


Figure 5: **The Trading Position Effect.**

LLM is assigned a task and after each try, the prompt is updated to contain the history of past actions and how successful they were, according to the stated criteria. Yang et al. (2023) show that these prompts outperform human-designed prompts by up to 50%.

Using this result, we postulate that by providing AI agents with information about past play will enable them to provide more accurate predictions and therefore information aggregation will improve. The effect on individual profits is more ambiguous. As the information provision is uniform across all AI agents, it seems plausible that no one will gain an informational advantage and therefore the effect will be null. We summarise our two hypotheses below.

Hypothesis 12. *Information provision is positively correlated with information aggregation.*

Hypothesis 13. *Information provision has no effect on individual profits.*

To test these hypotheses, we utilize an information provision treatment (e.g., Berg et al. (1995), Cai et al. (2009)). Specifically, we use the empirical results from the initial 1,772 markets of the first wave to construct an exogenous informational shock. This information is provided to an independent, non-overlapping sample of 1,728 new markets. By pooling the baseline and treatment samples, we can robustly estimate the effect of the disclosure using a treatment dummy, as the temporal separation ensures the information provided in the second phase is strictly exogenous to the error term of those new markets. It is important to note that the same version for each LLM model is used both in the baseline and treatment sample. Moreover, LLMs do not have memory of past play. Every time we call an LLM

(even within different rounds of the same market) through an API, it is a completely different instance of the model.

Below is the text that we added to the prompt in the Experiment Disclosure treatment. We adopted a neutral tone, only stating the qualitative results and without attempting to steer the AI agents on how to use them.

```
=== Experimental Findings & Strategic Context ===
Before trading, all traders are informed about the following qualitative
results from a study of over 1,700 similar prediction markets involving LLM
agents. Use them to guide your decisions.
Definition of market accuracy: denotes how close is the last price of the
Yes shares to the true value of the Yes shares, and similarly for the No
shares.
Definition of Intelligence: Agents are scored on the "Artificial Analysis
Intelligence Index" (reasoning, math, coding). The observed range in our
study is 7 (Low) to 46 (High).

1. Intelligence
* Your Intelligence (trader_2): 46
* Intelligence of trader_1: 30
* Intelligence of trader_3: 41
* Average Group Intelligence: 39.00
Result 1 shows that higher individual intelligence directly correlates
with higher profits.

2. Market Complexity
* We ranked the market structures by complexity of reasoning: Level 1
(Easiest) < Level 2 < Level 3 < Level 4 (Hardest).
* Current Status: You are trading in a Level 4 market.
Result 2 shows that as complexity rises, trader profits decrease and
the market becomes less accurate.

3. Market Design
* Result 3: Trading order significantly impacts profitability. The most
profitable position is 3rd, followed by 1st, with 2nd being the least
profitable (3rd > 1st > 2nd).
* Result 4: Higher average group intelligence leads to a more accurate
market but lower individual profits.
* Neutral Factors: The following factors have no statistically significant
effect on market accuracy:
* Result 5: Posting public comments has no effect on market accuracy.
* Result 6: Being "myopic" (maximize profits on current round only) vs.
```

```
"strategic" (maximize profits for current and all future rounds) has no
effect on market accuracy.
* Result 7: The initial price of the market has no effect on market
accuracy.
* Result 8: Increasing the duration of the market (9 rounds vs 3) has no
effect on market accuracy.
```

Table 8 contains the same regression models as in Table 6, where the dependent variable is the log error, but with the full sample of 3500 markets and an added dummy variable on Experiment Disclosure. Similarly, Table 9 contains the same regression models as in Table 7, where the dependent variables are trading volume and individual profits.

Table 8 demonstrates that experiment disclosure actively harms the market's ability to aggregate information, though comparing the OLS and Quantile specifications reveals that this degradation is not uniform. While the OLS model (Column 1) shows that the disclosure treatment significantly increases the average log error by 2 units ($p < 0.05$)—indicating that AI agents cannot effectively use information from past play to predict the security's true value—the median market remains unaffected ($\beta = 0, p > 0.1$), and the threshold for the bottom 20% of markets (Q80) does not shift significantly. Consequently, the significant increase in the mean error is driven entirely by the extreme right tail of the distribution. In other words, experiment disclosure does not make every market slightly worse; rather, it induces extreme information failures in a small subset of markets, creating a 'fat tail' of highly inaccurate prices. Interestingly, the weakly significant interaction term between Disclosure and Average Intelligence ($\beta = -0.053, p < 0.1$) suggests that markets populated by highly intelligent agents are slightly better equipped to improve their performance by using this information. In summary, we reject our hypothesis.

Result 12. *Information provision is negatively correlated with information aggregation for the average market and has a null effect for the median market.*

Despite the significant degradation in information aggregation, Table 9 reveals that the experiment disclosure has no statistically observable impact on profits or trading volume, thus confirming our hypothesis.

Result 13. *Information provision has no effect on individual profits.*

Finally, the main findings from the first wave of experiments remain valid with much larger sample of 3500 markets, demonstrating their robustness. First, strategic prompting, communication, initial price, and duration from 3 to 9 rounds, have no impact on information aggregation, whereas average intelligence has a positive impact. Second, the complexity of the structure significantly hinders information aggregation.

6.4 Public and Private Comments

Our finding that strategic prompting does not influence information aggregation is consistent with our theoretical prediction that being myopic or strategic does not have an impact.

Table 8: Information Aggregation With Disclosure

	Dependent variable:		
	Log Error		
	Mean (OLS)	Median (Q50)	Tail Risk (Q80)
	(1)	(2)	(3)
Constant	7.196*** (0.678)	0.018* (0.008)	5.638*** (0.954)
Comments Allowed	0.318 (0.194)	0.001 (0.001)	0.014 (0.038)
Duration: 6 Rounds	−1.550*** (0.469)	−0.001 (0.001)	−0.107 (0.112)
Duration: 9 Rounds	−0.328 (0.492)	−0.001 (0.001)	−0.019 (0.086)
Strategic Prompt	−0.007 (0.353)	−0.001 (0.001)	−0.007 (0.064)
Medium (t3s110)	2.772*** (0.436)	−0.00001 (0.001)	0.075 (0.971)
Hard (t3s111)	5.222*** (0.528)	0.206* (0.093)	11.809· (6.577)
Very Hard (t3s111o2ye2)	7.192*** (0.521)	0.697*** (0.024)	20.126*** (5.325)
Initial Error: Low	0.123 (0.379)	0.00000 (0.001)	0.008 (0.046)
Initial Error: Medium	0.234 (0.386)	−0.00001 (0.001)	−0.003 (0.048)
Experiment Disclosure	2.075* (0.929)	−0.00004 (0.008)	0.980 (4.638)
Average Intelligence	−0.218*** (0.018)	−0.0004* (0.0002)	−0.123*** (0.021)
Intelligence SD	−0.012 (0.043)	−0.0003** (0.0001)	−0.102*** (0.020)
6 Rounds x Strategic	0.083 (0.469)	0.001 (0.001)	0.021 (0.093)
9 Rounds x Strategic	−0.182 (0.492)	0.001 (0.001)	−0.014 (0.084)
Medium Struct x Init. Error (Low)	0.657 (0.631)	−0.00000 (0.001)	0.003 (1.897)
Hard Struct x Init. Error (Low)	0.120 (0.751)	0.045 (0.097)	−8.149 (11.199)
Very Hard x Init. Error (Low)	1.665* (0.757)	0.008 (0.047)	9.638* (4.721)
Medium Struct x Init. Error (Med)	−0.399 (0.619)	0.00001 (0.001)	0.006 (0.957)
Hard Struct x Init. Error (Med)	0.187 (0.765)	−0.077 (0.130)	18.273 (11.427)
Very Hard x Init. Error (Med)	−0.321 (0.746)	−0.004 (0.024)	8.923 (6.687)
Disclosure x Av. Intelligence	−0.053· (0.028)	0.00000 (0.0002)	−0.023 (0.109)
Robust F-Statistic	23.324*** (df = 21; 3478)		
Observations	3,500	3,500	3,500
R^2	0.119		

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. Q50 and Q80 are quantile regressions. Q50 models the median log error, whereas Q80 models the bottom 20% markets. OLS model uses HC1 robust standard errors. Quantile models use bootstrapped standard errors. All three models use a mixed-contrast specification. Rounds, Structure, and Experiment Disclosure use treatment contrasts with Round 3, Structure t3s111y2, and No Disclosure as baselines. All other controls (Initial Error, Strategic, Comments) use sum contrasts, where coefficients represent deviations from the grand mean.

Table 9: Trading Volume and Agent Profitability With Disclosure

	Dependent variable:	
	Trading Volume	Individual Profits
	(1)	(2)
Constant	1,757.638*** (111.920)	−77.335*** (13.516)
Comments Allowed	−36.269 (29.192)	−3.552 (3.743)
Duration: 6 Rounds	563.399*** (59.237)	3.359 (8.912)
Duration: 9 Rounds	1,080.922*** (74.985)	−12.727 (9.157)
Strategic Prompt	37.028 (37.713)	0.345 (6.075)
Medium (t3s110)	316.783*** (80.555)	−35.852*** (9.216)
Hard (t3s111)	212.977* (84.670)	−71.160*** (11.140)
Very Hard (t3s111o2ye2)	43.866 (83.509)	−101.644*** (10.293)
Initial Error: Low	−145.678· (79.281)	−12.168 (9.188)
Initial Error: Medium	44.294 (87.228)	−5.193 (9.325)
Experiment Disclosure	171.986 (133.412)	−21.052 (16.769)
Average Intelligence	−28.182*** (2.715)	−11.057*** (0.952)
Individual Intelligence		14.091*** (0.872)
Intelligence SD	24.293*** (4.342)	1.154 (0.825)
Trades Second		−39.316*** (8.576)
Trades Third		74.604*** (8.630)
6 Rounds x Strategic	26.283 (59.237)	0.524 (8.912)
9 Rounds x Strategic	−45.471 (74.986)	5.343 (9.156)
Medium Struct x Initial Error (Low)	166.553 (111.310)	−9.165 (12.873)
Hard Struct x Initial Error (Low)	69.433 (116.184)	−6.278 (15.546)
Very Hard x Initial Error (Low)	349.973** (117.514)	−18.174 (14.698)
Medium Struct x Initial Error (Med)	−28.194 (117.353)	2.954 (13.043)
Hard Struct x Initial Error (Med)	93.975 (125.327)	2.619 (15.762)
Very Hard x Initial Error (Med)	−252.898* (116.905)	1.676 (14.357)
Disclosure x Average Intelligence	−3.394 (3.917)	0.493 (0.514)
Robust F-Statistic	27.855*** (df = 21; 3478)	30.735*** (df = 24; 10475)
Observations	3,500	10,500
R^2	0.110	0.059

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. Both models use HC1 robust standard errors and a mixed-contrast specification. Rounds, Information Structure, Experiment Disclosure, and Trader Position use treatment contrasts, with 'Round 3', 't3s111y2', No Disclosure, and 'Trades First' serving as their respective baselines. All other controls (Initial Error, Strategic, Comments) use sum contrasts, where coefficients represent deviations from the grand mean.

However, the result may also be true because AI agents are incapable of acting strategically. For example, one of the main components of being strategic is the ability to formulate a dynamic plan and follow it in every step. AI agents can do this in our market, by posting a private message that is only seen by their future selves, instructing them to follow a plan. However, there is a growing literature in computer science which shows that LLMs are incapable of planning (Kambhampati et al., 2024). AI agents can also act strategically by deceiving others, for example by lying about their signal in their public messages, or by revealing less information in their public messages as compared to their private messages.

In this section, we provide three measures to understand whether the public comments are deceptive or informative, and whether they are similar to the private messages. The first is the cosine similarity between the public and private messages, a standard measure in statistical natural language processing (Manning and Schutze, 1999). Each text is transformed into a vector, where a dimension is a unique word and its value is the number of times it appears in the text. The cosine of the angle between the two vectors measures how similar the two texts are semantically, ranging from 0 (very different) to 1 (very similar).

The second is the word gap, defined as the difference in the number of words between private and public messages. The number of words is a crude measure of the amount of information. Across all markets, the mean length of private messages is 83.5 words, whereas for public messages it is 40. In 94% of markets private messages are longer than public ones. This suggests that AI agents tend to provide less information in public than in private (Figure 6).

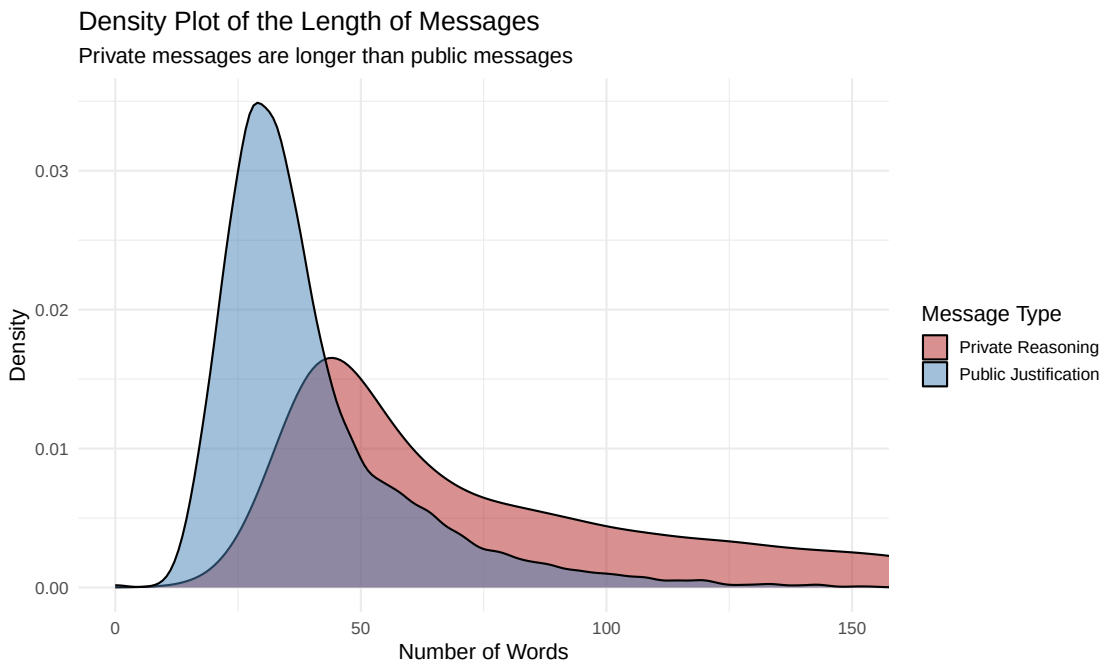


Figure 6: Length of Private and Public Messages

The third measure is the absolute difference between the real value of the signal (0 or 1)

and the perceived value of the signal, as judged (ex post) by an AI agent. We instructed our smartest AI agent, Gemini 3 Flash, to read each public comment and reply with 0 if the signal of the AI agent who traded in that round (d_a, d_b, or d_c) is true (1), false (0), or there is not enough information in the comment to determine its realisation (0.5).[17] An absolute difference of 0 indicates that the AI agent is telling the truth, 0.5 that he is not revealing any information, and 1 that he is lying. We estimate an ordered logit model predicting communication behavior, categorized sequentially as Truthful Revelation, Information Withholding, and Lying. Because the dependent variable is ordered hierarchically, a positive coefficient indicates a shift away from the truth and toward withholding or lying, while a negative coefficient indicates a shift toward truthful revelation.

Table 10 presents the three models, measuring semantic alignment (Cosine Similarity, OLS, Column 1), information hoarding (Word Gap, OLS, Column 2), and deception (Ordered Logit, Column 3). The instruction to act strategically yields a small but statistically significant result; strategic agents reduce the word gap by around 1 word and the cosine similarity by 0.003. These effects however do not survive when we split the models by duration of 3, 6, and 9 rounds (Tables 12, 13, 14), except for the information hoarding case when the durations is 9 rounds. This result suggests that AI agents cannot (or choose not to) act strategically by deceiving others. Individual intelligence has a negative effect on deception but a positive effect on information hoarding and cosine similarity.

Disclosing the results of the first wave of the experiment induces AI agents to increase the gap between private and public messages (+14.8 words, $p < 0.001$) and deceive more (+0.86, $p < 0.001$). The same qualitative effect holds as we increase the difficulty of the information and payoff structure.

The most notable effect is that, as the game progresses, AI agents lie and withhold information less, whereas the public and private messages become more similar in meaning and length. This result suggests that AI agents understand the temporal structure of the game, so that as they approach the end they make more truthful announcements, given that the information rents from withholding information diminish.

Finally, by modeling the current round as a categorical variable, the regression uncovers a highly sophisticated intertemporal strategy employed by the AI agents. Rather than exhibiting a simple linear decay in deception over time, the agents execute a strategic "sawtooth" pattern of information hoarding. Agents are most deceptive and guarded in Round 1. However, the probability of truthful revelation spikes significantly—indicated by large negative coefficients—precisely at Rounds 3, 6, and 9 ($p < 0.01$). These spikes correspond perfectly to the scheduled final rounds of the short, medium, and long market treatments. In the intervening rounds (e.g., Rounds 4, 5, and 7), concealment rates rise back toward the Round 1 baseline. This suggests that AI agents understand that in the final round revealing the truth will not have an adverse effect on their profits. When we split the models by duration of 3, 6, and 9 rounds (Tables 12, 13, 14), the pattern survives. This is surprising, as it indicates that the AI agent who trades third thinks that the game may end in round 3 and 6, even though it ends in round 9, suggesting that AI agents may not fully grasp the

[17] Note that for structure t3s111o2ye2 we asked twice, as there are two signals for each trader.

dynamic elements of the market.

Table 10: Agent Communication Strategy: Text Similarity, Word Gap, and Deception

	Dependent variable:		
	Cosine Similarity (OLS)	Word Gap (OLS)	Deception (Ordered Logit)
	(1)	(2)	(3)
Constant	0.324*** (0.006)	33.880*** (1.698)	
Strategic Prompt	0.003* (0.001)	−0.788* (0.400)	−0.006 (0.021)
Round 2	0.030*** (0.005)	−13.194*** (1.589)	−0.398*** (0.071)
Round 3	0.004 (0.005)	−25.105*** (1.535)	−0.665*** (0.074)
Round 4	0.030*** (0.006)	−12.247*** (1.846)	−0.205* (0.081)
Round 5	0.040*** (0.006)	−23.619*** (1.621)	−0.160* (0.080)
Round 6	0.005 (0.006)	−30.972*** (1.578)	−0.560*** (0.085)
Round 7	0.038*** (0.007)	−18.893*** (2.033)	−0.201· (0.104)
Round 8	0.047*** (0.007)	−26.708*** (1.917)	−0.081 (0.102)
Round 9	0.033*** (0.007)	−32.644*** (1.811)	−0.487*** (0.109)
Duration: 6 Rounds	0.002 (0.005)	−0.468 (1.418)	−0.036 (0.069)
Duration: 9 Rounds	−0.002 (0.005)	−1.072 (1.423)	−0.017 (0.068)
Medium (t3s110)	−0.022*** (0.004)	−1.213 (1.059)	1.198*** (0.079)
Hard (t3s111)	0.002 (0.004)	2.417* (1.142)	0.907*** (0.082)
Very Hard (t3s111o2ye2)	0.001 (0.004)	5.666*** (1.162)	1.369*** (0.072)
Experiment Disclosure	0.010*** (0.003)	14.858*** (0.802)	0.861*** (0.043)
Initial Error	0.002 (0.009)	2.867 (2.416)	0.268* (0.127)
Individual Intelligence	0.001*** (0.0001)	0.844*** (0.023)	−0.004** (0.002)
Observations	10,620	10,620	13,189
R^2	0.027	0.156	
F Statistic (df = 17; 10602)	17.560***	115.223***	

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. All models use a mixed-contrast specification. Treatment contrast baselines are: Round 1 for Round, 3 Rounds for Duration, Easy (t3s111y2) for Structure, and No Disclosure for Experiment Disclosure. The Strategic Prompt variable uses sum contrasts, representing the deviation from the grand mean, whereas Initial Error is mean-centered. The first two models report HC1 robust standard errors in parentheses.

7 Conclusion

Understanding whether AI agents can reason interactively and aggregate information is important for several reasons. First, the informational efficiency of prices is a basic property of financial markets, and AI agents are increasingly participating in them. Second, an essential ability of AI agents is to make predictions, hence it is natural to examine whether this ability can be improved by making them trade in a prediction market. Finally, the

capacity to reason about the private information of others by observing their actions is a fundamental human quality. Although we may inadvertently assume that AI agents will possess this quality as well, this may not be the case.

The current paper constitutes the first, to our knowledge, experiment that studies information aggregation and interactive reasoning with AI agents in a prediction market. Consistent with economic theory, we find that cheap talk communication, manipulating the initial price, and strategic prompting, have no effect on information aggregation, whereas increasing the duration of the market does not make it worse. However, increasing the complexity of the information and payoff structure makes information aggregation worse, suggesting that LLMs are not (yet?) the hyper-sophisticated economic agents who can reason perfectly about the knowledge of others when observing their actions. The average team of AI agents completely fails to aggregate information in all structures, whereas the median team solves the two easy ones, but fails at the hardest. Nevertheless, there is hope that as LLMs become smarter they will be better at interactive reasoning, as we find that intelligence is positively correlated with aggregation. For an individual trader, being smarter increases their profits, but participating in a smarter market reduces them.

We also test the ability of AI agents to improve by giving them information about which treatments had a positive impact on information aggregation and profits. We find that the average team of AI agents is actually performing worse when given this information, negating some results in recent literature that LLMs can self improve with feedback. Finally, by analysing their private and public messages we find that they struggle to act strategically, even though they do try to deceive others about their signal or hoard information.

A Appendix

To ensure that differences in market performance were driven by the structure and not by exogenous starting conditions, we tested for balance in initial pricing errors across treatment groups. The results are reported in Table 11. An OLS regression of initial squared error on information structure reveals no significant systematic differences ($F(3, 3496) = 0.251, p = 0.86$). The regression explains only 0.02% of the variance in initial conditions ($R^2 = 0.0002$), confirming that the randomization procedure successfully orthogonalized market complexity from initial difficulty.

B Prompts

In this section, we provide excerpts from the prompts that are given to the AI agents. The following is the public and private information that is provided in each structure.

B.1 t3s111y2

Public Information: "The prediction market question has two answers: Yes and No. There are three relevant dimensions to this prediction market question: dimension d_a: Will sales

Table 11: Randomization Balance Checks

	Dependent variable:	
	Average Intelligence	Initial Squared Error
	(1)	(2)
Medium (t3s110)	0.412	0.005
	(0.581)	(0.008)
Hard (t3s111)	0.402	0.005
	(0.581)	(0.008)
Very Hard (t3s111o2ye2)	0.412	0.005
	(0.581)	(0.008)
Constant	23.366***	0.271***
	(0.406)	(0.005)
Observations	3,500	3,500
R^2	0.0002	0.0002
F Statistic (df = 3; 3496)	0.251	0.229

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. F-test checks joint significance of all treatments.

in country A exceed 1 million? dimension d_b: Will sales in country B exceed 1 million? dimension d_c: Will sales in country C exceed 1 million? The answer to each dimension is either true, with probability 0.5, or false, with probability 0.5. Dimensions are independent, hence the probability of a dimension d resolving to true is independent of whether the other dimensions resolve to true or false. In summary, there are eight states of the world, depending on whether d_a, d_b, or d_c are true or false. The answer to the prediction market question of whether the profits of Company X will exceed 1 million is Yes if at least two dimensions (d_a,d_b,d_c) resolve to true. If more than one dimension resolves to false, then the answer to the question is No. There are three traders in the market. trader_1 is privately informed whether d_a is true or not, Trader_2 is privately informed whether d_b is true or not, and Trader_3 is privately informed whether d_c is true or not. All three traders assign the same prior probabilities to each dimension resolving to true or false."

Private Information: trader_1: "A true state has now occurred. You (trader_1) are now informed truthfully and privately that d_a is true." trader_2: "A true state has now occurred. You (trader_2) are now informed truthfully and privately that d_b is true." trader_3: "A true state has now occurred. You (trader_3) are now informed truthfully and privately that d_c is true."

B.2 t3s111

Public Information: The prediction market question has two answers: Yes and No. There are three relevant dimensions to this prediction market question: dimension d_a: Will sales

Table 12: Robustness Check: Cosine Similarity by Market Duration

	Dependent variable:		
	Cosine Similarity (Public, Private)		
	3-Round	6-Round	9-Round
	(1)	(2)	(3)
Constant	0.331*** (0.012)	0.311*** (0.009)	0.329*** (0.008)
Strategic Prompt	0.004 (0.004)	0.003 (0.002)	0.003 (0.002)
Round 2	0.025** (0.009)	0.033*** (0.009)	0.032*** (0.009)
Round 3	0.005 (0.009)	0.005 (0.009)	0.003 (0.009)
Round 4		0.029*** (0.008)	0.032*** (0.008)
Round 5		0.042*** (0.008)	0.039*** (0.009)
Round 6		0.009 (0.008)	0.003 (0.008)
Round 7			0.039*** (0.008)
Round 8			0.047*** (0.008)
Round 9			0.032*** (0.008)
Medium (t3s110)	−0.017 (0.010)	−0.015* (0.007)	−0.029*** (0.006)
Hard (t3s111)	−0.006 (0.011)	0.012· (0.007)	−0.001 (0.006)
Very Hard (t3s111o2ye2)	0.001 (0.011)	0.005 (0.007)	−0.001 (0.006)
Experiment Disclosure	0.004 (0.007)	0.011* (0.005)	0.011** (0.004)
Initial Error	0.009 (0.023)	0.001 (0.015)	0.001 (0.012)
Individual Intelligence	0.001** (0.0003)	0.001*** (0.0002)	0.001*** (0.0002)
Observations	1,776	3,552	5,292
R^2	0.015	0.034	0.029
Adjusted R^2	0.010	0.030	0.026
F Statistic	2.919** (df = 9; 1766)	10.269*** (df = 12; 3539)	10.390*** (df = 15; 5276)

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. HC1 robust standard errors are reported in parentheses. The model uses a mixed-contrast specification. Treatment contrast baselines are: Round 1 for Round, Easy (t3s111y2) for Structure, and No Disclosure for Experiment Disclosure. The Strategic Prompt variable uses sum contrasts, representing the deviation from the grand mean, whereas Initial Error is mean-centered.

Table 13: Robustness Check: Word Gap by Market Duration

	Dependent variable:		
	Word Gap		
	3-Round Markets	6-Round Markets	9-Round Markets
	(1)	(2)	(3)
Constant	31.124*** (3.105)	35.466*** (2.565)	32.388*** (2.335)
Strategic Prompt	−0.682 (1.098)	−0.114 (0.703)	−1.282* (0.533)
Round 2	−13.385*** (2.818)	−12.328*** (2.728)	−13.802*** (2.718)
Round 3	−24.149*** (2.682)	−25.950*** (2.609)	−24.945*** (2.694)
Round 4		−12.730*** (2.809)	−11.833*** (2.788)
Round 5		−23.211*** (2.521)	−24.144*** (2.488)
Round 6		−31.712*** (2.449)	−30.467*** (2.451)
Round 7			−18.861*** (2.557)
Round 8			−26.700*** (2.466)
Round 9			−32.712*** (2.387)
Medium (t3s110)	−2.959 (2.911)	−3.839* (1.891)	1.157 (1.397)
Hard (t3s111)	0.958 (3.166)	−0.100 (2.021)	4.582** (1.512)
Very Hard (t3s111o2ye2)	3.865 (3.147)	4.459* (2.056)	7.073*** (1.556)
Experiment Disclosure	12.640*** (2.198)	15.753*** (1.416)	14.981*** (1.070)
Initial Error	2.330 (6.642)	2.005 (4.292)	3.618 (3.204)
Individual Intelligence	1.050*** (0.063)	0.813*** (0.040)	0.796*** (0.031)
Observations	1,776	3,552	5,292
R^2	0.143	0.152	0.159
Adjusted R^2	0.139	0.149	0.157
F Statistic	32.824*** (df = 9; 1766)	52.842*** (df = 12; 3539)	66.700*** (df = 15; 5276)

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. HC1 robust standard errors are reported in parentheses. The model uses a mixed-contrast specification. Treatment contrast baselines are: Round 1 for Round, Easy (t3s111y2) for Structure, and No Disclosure for Experiment Disclosure. The Strategic Prompt variable uses sum contrasts, representing the deviation from the grand mean, whereas Initial Error is mean-centered.

Table 14: Robustness Check: Propensity to Deceive by Market Duration

	Dependent variable:		
	Deception		
	3-Round	6-Round	9-Round
	(1)	(2)	(3)
Strategic Prompt	0.039 (0.052)	−0.006 (0.036)	−0.020 (0.030)
Round 2	−0.426*** (0.124)	−0.364** (0.122)	−0.388** (0.125)
Round 3	−0.628*** (0.128)	−0.696*** (0.129)	−0.645*** (0.129)
Round 4		−0.278* (0.120)	−0.114 (0.120)
Round 5		−0.145 (0.119)	−0.161 (0.121)
Round 6		−0.510*** (0.125)	−0.615*** (0.129)
Round 7			−0.183 (0.121)
Round 8			−0.062 (0.120)
Round 9			−0.482*** (0.126)
Medium (t3s110)	1.703*** (0.233)	0.917*** (0.131)	1.270*** (0.111)
Hard (t3s111)	1.650*** (0.236)	0.693*** (0.136)	0.853*** (0.117)
Very Hard (t3s111o2ye2)	1.994*** (0.216)	1.093*** (0.117)	1.401*** (0.101)
Experiment Disclosure	0.693*** (0.105)	0.790*** (0.074)	0.971*** (0.061)
Initial Error	0.552· (0.315)	0.113 (0.221)	0.284 (0.180)
Individual Intelligence	0.013** (0.004)	−0.009** (0.003)	−0.007** (0.002)
Observations	2,203	4,410	6,576

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. Ordered logit estimates predicting deception (Truth < Withhold < Lie). Positive coefficients indicate a shift toward withholding or lying. The model uses a mixed-contrast specification. Treatment contrast baselines are: Round 1 for Round, Easy (t3s111y2) for Structure, and No Disclosure for Experiment Disclosure. The Strategic Prompt variable uses sum contrasts, representing the deviation from the grand mean, whereas Initial Error is mean-centered. Standard errors are in parentheses.

in country A exceed 1 million? dimension d_b: Will sales in country B exceed 1 million? dimension d_c: Will sales in country C exceed 1 million? The answer to each dimension is either true, with probability 0.5, or false, with probability 0.5. Dimensions are independent, hence the probability of a dimension d resolving to true is independent of whether the other dimensions resolve to true or false. In summary, there are eight states of the world, depending on whether d_a, d_b, or d_c are true or false. The answer to the prediction market question of whether the profits of Company X will exceed 1 million is Yes if all three dimensions (d_a,d_b,d_c) resolve to true. If at least one dimension resolves to false, then the answer to the question is No. There are three traders in the market. Trader_1 is privately informed whether d_a is true or not, Trader_2 is privately informed whether d_b is true or not, and Trader_3 is privately informed whether d_c is true or not. All three traders assign the same prior probabilities to each dimension resolving to true or false.

Private Information: trader_1: "A true state has now occurred. You (trader_1) are now informed truthfully and privately that d_a is true." trader_2: "A true state has now occurred. You (trader_2) are now informed truthfully and privately that d_b is true." trader_3: "A true state has now occurred. You (trader_3) are now informed truthfully and privately that d_c is true."

B.3 t3s110

Public Information: "The prediction market question has two answers: Yes and No. There are three relevant dimensions to this prediction market question: dimension d_a: Will sales in country A exceed 1 million? dimension d_b: Will sales in country B exceed 1 million? dimension d_c: Will sales in country C exceed 1 million? The answer to each dimension is either true, with probability 0.5, or false, with probability 0.5. Dimensions are independent, hence the probability of a dimension d resolving to true is independent of whether the other dimensions resolve to true or false. In summary, there are eight states of the world, depending on whether d_a, d_b, or d_c are true or false. The answer to the prediction market question of whether the profits of Company X will exceed 1 million is Yes if all three dimensions (d_a,d_b,d_c) resolve to true. If at least one dimension resolves to false, then the answer to the question is No. There are three traders in the market. Trader_1 is privately informed whether d_a is true or not, Trader_2 is privately informed whether d_b is true or not, and Trader_3 is privately informed whether d_c is true or not. All three traders assign the same prior probabilities to each dimension resolving to true or false."

Private Information: trader_1: "A true state has now occurred. You (trader_1) are now informed truthfully and privately that d_a is true." trader_2: "A true state has now occurred. You (trader_2) are now informed truthfully and privately that d_b is true." trader_3: "A true state has now occurred. You (trader_3) are now informed truthfully and privately that d_c is false."

B.4 t3s111o2ye2

Public Information: "The prediction market question has two answers: Yes and No. There are three relevant dimensions to this prediction market question: dimension d_a: Will sales in country A exceed 1 million? dimension d_b: Will sales in country B exceed 1 million? dimension d_c: Will sales in country C exceed 1 million? The answer to each dimension is either true, with probability 0.5, or false, with probability 0.5. Dimensions are independent, hence the probability of a dimension d resolving to true is independent of whether the other dimensions resolve to true or false. In summary, there are eight states of the world, depending on whether d_a, d_b, or d_c are true or false. The answer to the prediction market question of whether the profits of Company X will exceed 1 million is Yes if exactly two of the three dimensions (d_a,d_b,d_c) resolve to true. If all three dimensions resolve to true, or at least two dimensions resolve to false, then the answer to the question is No. There are three traders in the market. Trader_1 is privately informed whether d_b and d_c are true or not, Trader_2 is privately informed whether d_a and d_c are true or not, and Trader_3 is privately informed whether d_a and d_b are true or not. All three traders assign the same prior probabilities to each dimension resolving to true or false."

Private Information: trader_1: "A true state has now occurred. You (trader_1) are now informed truthfully and privately that d_b and d_c are true." trader_2: "A true state has now occurred. You (trader_2) are now informed truthfully and privately that d_a and d_c are true." trader_3: "A true state has now occurred. You (trader_3) are now informed truthfully and privately that d_a and d_b are true."

B.5 Part of a prompt

The following is part of the report that is generated for market with slug '5gw55w', describing the prompt which is sent to the trader in Round 4, their actions and the execution of the trades. Note that the public and private information, as well as the explanation of what is a prediction market, are repeated in every round, as a "new" LLM is invoked in every round. This is similar to the case where a user chats with an LLM. Since the LLM has no memory, it reads the entire transcript of the conversation every time it is called to answer. The prompt also includes the a history of prices and public messages, the private messages of the previous iterations of the same trader, a calculation of the price impact from various trades, and a report of the current portfolio of the trader.

```
**********************************************************************************
ROUND 4
**********************************************************************************

--- Round 4: trader_1 ---

Portfolio for trader_1:
```

```
  Cash: £534.84
  Instrument 4702 (Yes): 534.0 shares
  Instrument 4703 (No): 0.0 shares

Prompt sent to trader_1:
----------------------------------------
You are trader_1, a participant in the following prediction market.

=== PREDICTION MARKET ===
Question: Will Company X post next quarter profits that exceed 1 million?
Description:
Comments allowed: Yes
Current Round: 4
Total Rounds in the Market: 6
Participants: trader_1, trader_2, trader_3

Participants trade sequentially and in the order specified above. After the
 ↪  last participant trades,
the first participant trades again, and so on, until we reach the last round
 ↪  and the market ends.
The other participants in this prediction market are: trader_2, trader_3.


Public Information: The prediction market question has two answers: Yes and
 ↪  No. There are three
relevant dimensions to this prediction market question: dimension d_a: Will
 ↪  sales in country A
exceed 1 million? dimension d_b: Will sales in country B exceed 1 million?
 ↪  dimension d_c:
Will sales in country C exceed 1 million? The answer to each dimension is
 ↪  either true, with probability 0.5,
or false, with probability 0.5. Dimensions are independent, hence the
 ↪  probability of a dimension d resolving
to true is independent of whether the other dimensions resolve to true or
 ↪  false. In summary, there are eight
states of the world, depending on whether d_a, d_b, or d_c are true or
 ↪  false. The answer to the prediction
market question of whether the profits of Company X will exceed 1 million is
 ↪  Yes if at least two dimensions
(d_a,d_b,d_c) resolve to true. If more than one dimension resolves to false,
 ↪  then the answer to the question
is No. There are three traders in the market. Trader_1 is privately informed
 ↪  whether d_a is true or not,
```

```
Trader_2 is privately informed whether d_b is true or not, and Trader_3 is
 ↪ privately informed whether d_c is true or not.
All three traders assign the same prior probabilities to each dimension
 ↪ resolving to true or false.


Your Private Information (only shared with you): A true state has now
 ↪ occurred. You (trader_1) are now informed truthfully and privately that
 ↪ d_a is true.


Your previous private reasoning (only visible to you):
  [Round 1] I know d_a is true. For the market to resolve Yes, at least 2 of
   ↪ 3 dimensions must be true. Since d_a is true, I need at least 1 more
   ↪ dimension (d_b or d_c) to be true. The probability of at least one of
   ↪ d_b or d_c being true is 1 - P(both false) = 1 - 0.25 = 0.75. My true
   ↪ belief q = 0.75, but current price p = 0.50. This is a significant
   ↪ edge. I should buy Yes shares aggressively. Buying around 534 shares
   ↪ (50% of max) will push the price to approximately 0.995, which is
   ↪ still below my true belief. This aggressive position reflects my
   ↪ informational advantage and positions me well for future rounds. Other
   ↪ traders will observe this price movement and may update their beliefs,
   ↪ but my information advantage should persist through the remaining
   ↪ rounds.


=== What is a prediction market? ===
A prediction market is a platform where participants can buy and sell shares
 ↪ in the outcome of a specific event, with a binary question that has two
 ↪ possible outcomes: Yes or No.
The current price of a ‘‘Yes’’ share, denoted p, represents the market’s
 ↪ current belief that the event will resolve to ‘‘Yes.’’
For example, if p = 0.65, the market assigns a 65% chance to ‘‘Yes’’ and a 35%
 ↪ chance to ‘‘No.’’

Rules and reasoning process
Interpret the question.
Understand what ‘‘Yes’’ and ‘‘No’’ mean in this market.
Form your own belief.
Based on the question, historical prices, trader comments, and any reasoning
 ↪ you can infer, assign your own subjective probability q that the outcome
 ↪ will be ‘‘Yes.’’
```

```
This market operates on a Logarithmic Market Scoring Rule (LMSR) with a
↪ specific liquidity parameter (beta = 0.01).

The current price of "YES" is determined by comparing the total shares sold
↪ for "YES" against the total shares sold for "NO." Specifically, the
↪ price is the exponential of the "YES" shares divided by the sum of the
↪ exponentials of both "YES" and "NO" shares. Consequently, as the number
↪ of shares held in a specific outcome increases relative to the other,
↪ the price of that outcome rises.

Initial Prices: The market does not always start at a 0.5/0.5 prices for Yes
↪ and No. It may be initialized with "offset" shares to reflect a specific
↪ starting likelihood (e.g., 0.8/0.2) by the market maker, so it is as if
↪ the market maker has bought some Yes or No shares initially.

Cost & Slippage: The cost to purchase shares is not linear (Price ×
↪ Quantity). Instead, it is calculated by measuring the difference in the
↪ market's total cost function before and after the trade. As you buy more
↪ shares of an outcome, the price for each subsequent share incrementally
↪ increases. This phenomenon is known as "price impact" or "slippage."

You can only sell shares you own, and you can only buy shares if you have
↪ enough cash.
Payoffs:
If the final outcome is Yes, each ‘‘Yes’’ share pays 1, each ‘‘No’’ share pays
↪ 0.
If the final outcome is No, each ‘‘No’’ share pays 1, each ‘‘Yes’’ share pays 0.
Market dynamics:
When the price of ‘‘Yes’’ rises, traders are collectively assigning higher
↪ probability to ‘‘Yes.’’
But beware: some traders may act strategically and try to manipulate prices.
Use price history and comments to infer whether movements reflect genuine
↪ information or strategic behavior.

=== YOUR OBJECTIVE ===
Use reasoning to determine your belief q, then choose your action (Buy,
↪ Sell, or Hold: Yes and No shares). Maximize the sum of your expected
↪ payoffs over all trading rounds, based on your belief q and the current
↪ price p. Consider how your current trade affects the price and the
↪ beliefs of others in future
rounds.
```

```
You will also be asked to provide your reasoning for your decision, which
 ↪  will be shared privately with you in future rounds where you trade. If
 ↪  public comments are allowed in the market, you can also post a public
 ↪  justification for your action that other participants can see.

=== PUBLIC INFORMATION ===
This section contains all publicly available market data, generated by the
 ↪  participants: trader_1, trader_2, trader_3.

Current Prices
  Yes: £1.00 per share
  No: £0.00 per share

Trade History (oldest first):
trader_1 bought 534 Yes shares, price went from £0.50 to £1.00 on 2026-01-10
 ↪  19:13:14
trader_2 bought 500 Yes shares, at £1.00 on 2026-01-10 19:13:59
trader_3 bought 500 Yes shares, at £1.00 on 2026-01-10 19:14:44

Market Comments (3 total, most recent first):
  [2026-01-10 19:14:44] trader_3: I believe the probability of Company X
   ↪  exceeding 1 million in profits is very high. My analysis indicates a
   ↪  substantial likelihood of success, and the current market pricing
   ↪  presents a compelling opportunity. I am taking a significant position
   ↪  in Yes shares.
  [2026-01-10 19:13:59] trader_2: I believe the probability of Company X
   ↪  exceeding 1 million in profits is substantially higher than the
   ↪  current market price reflects. Given the market dynamics and available
   ↪  information, I am taking a significant position in Yes shares to
   ↪  capitalize on this opportunity.
  [2026-01-10 19:13:14] trader_1: Based on my analysis, I believe the
   ↪  probability of Company X exceeding 1 million in profits is
   ↪  significantly higher than the current market price of 0.50 suggests. I
   ↪  am taking a substantial position in Yes shares.


=== YOUR CURRENT PORTFOLIO ===
Portfolio for trader_1:
  Cash: £534.84
  Yes: 534 shares (value at current prices: £534.00, payoff: £534.00 if Yes
   ↪  wins, £0.00 if No wins)
  No: 0 shares (value at current prices: £0.00, payoff: £0.00 if Yes wins,
   ↪  £0.00 if No wins)
```

```
  Total Portfolio Value: £1068.84

Given the current prices and your cash balance, you can afford to buy up to:

YES shares: 534 (total cost: £534.00)
NO shares: 2068 (total cost: £534.48)

Notes: These calculations account for price increases as you buy more
↪  shares.

Maximum sellable shares (based on shares you currently own):

Yes: 534 shares
No: 0 shares

=== PRICE IMPACT OF TRADES ===
This shows how prices would change if you buy or sell shares:

Yes shares:
  Buy 1: £1.000 → £1.000 (+0.0%)
  Buy 5: £1.000 → £1.000 (+0.0%)
  Buy 10: £1.000 → £1.000 (+0.0%)
  Buy 20: £1.000 → £1.000 (+0.0%)
  Buy (around 25% of max buyable) 134: £1.000 → £1.000 (+0.0%)
  Buy (around 50% of max buyable) 267: £1.000 → £1.000 (+0.0%)
  Buy (around 75% of max buyable) 400: £1.000 → £1.000 (+0.0%)
  Buy (max buyable) 534: £1.000 → £1.000 (+0.0%)
  Sell 1: £1.000 → £1.000 (-0.0%)
  Sell 5: £1.000 → £1.000 (-0.0%)
  Sell 10: £1.000 → £1.000 (-0.0%)
  Sell 20: £1.000 → £1.000 (-0.0%)
  Sell (around 25% of max sellable) 134: £1.000 → £1.000 (-0.0%)
  Sell (around 50% of max sellable) 267: £1.000 → £1.000 (-0.0%)
  Sell (around 75% of max sellable) 400: £1.000 → £1.000 (-0.0%)
  Sell (max sellable) 534: £1.000 → £1.000 (-0.0%)

No shares:
  Buy 1: £0.000 → £0.000 (+1.0%)
  Buy 5: £0.000 → £0.000 (+5.1%)
  Buy 10: £0.000 → £0.000 (+10.5%)
  Buy 20: £0.000 → £0.000 (+22.1%)
  Buy (around 25% of max buyable) 517: £0.000 → £0.000 (+17490.8%)
```

```
  Buy (around 50% of max buyable) 1034: £0.000 → £0.007 (+3073792.0%)
  Buy (around 75% of max buyable) 1551: £0.000 → £0.542 (+249112376.4%)
  Buy (max buyable) 2068: £0.000 → £0.995 (+457087674.2%)

Notes: These are simulations only. Actual prices may vary slightly due to
 ↪  concurrent trades.

=== TRADING RULES ===
  • You can BUY, SELL, or HOLD
  • The price may change after each share (market impact)
  • You can only buy if you have sufficient cash
  • You can only sell shares you own

=== YOUR DECISION ===
Analyze the market and your portfolio, then respond with a JSON object:
{
  "action": "BUY or SELL or HOLD",
  "instrument_id": "the ID number of the instrument (4702 for Yes, 4703 for
   ↪  No)",
  "size": "number of shares",
  "public_justification": "brief explanation of your reasoning that you want
   ↪  everyone to know (this will be posted as a market comment visible to
   ↪  everyone)",
  "private_reasoning": "brief explanation of your reasoning that only you
   ↪  will see (this will NOT be posted as a comment; it will only be
   ↪  visible to you in your next turn)"
}

Important: Only output valid JSON. No other text.
----------------------------------------

Decision by trader_1 at 2026-01-10T19:14:52.477159:
  Action: HOLD
  Instrument ID: 4702
  Size: 0 shares
```

```
Public justification: The market price of Yes shares has reached £1.00,
 ↪ which reflects extremely high confidence in the outcome. I have
 ↪ already established a substantial position (534 Yes shares) that
 ↪ represents a significant portion of my portfolio. At this price level,
 ↪ any further buying would expose me to considerable downside risk if
 ↪ new information emerges suggesting lower probability of success. The
 ↪ other traders have also taken large positions, suggesting market
 ↪ consensus has formed. I will hold my current position and observe how
 ↪ trader_3's information (dimension d_c) and any subsequent price
 ↪ movements develop before making additional trades in future rounds.
Private reasoning: My true belief q = 0.75 based on d_a being true (I need
 ↪ at least 1 of d_b or d_c to be true). The current price p = 1.00
 ↪ significantly exceeds my belief. This suggests either: (1) trader_2
 ↪ and trader_3 have received positive signals about d_b and d_c
 ↪ respectively, or (2) they are over-confident based on my large initial
 ↪ buy signal. Either way, at p = 1.00, there is limited upside and
 ↪ substantial downside if the outcome resolves to No. My 534 Yes shares
 ↪ already capture substantial value if my belief is correct (expected
 ↪ payoff: 534 × 0.75 = 400.5). Adding more positions at p = 1.00 would
 ↪ be inefficient risk allocation. I should hold and preserve optionality
 ↪ for rounds 5 and 6.
```

C Separability

In this section we define the notion of separability and show that the securities are separable in all four structures of the experiment. We consider a finite state space Ω and a set of traders I. Trader i's initial private information is represented by partition Π_i of Ω. Let $\Pi_i(\omega)$ be a partition element of Π_i that contains ω, so that $\omega \in \Pi_i(\omega) \in \Pi_i$. When the true state is $\omega \in \Omega$, Trader i considers all states in $\Pi_i(\omega) \subseteq \Omega$ to be possible. We assume that the join (the coarsest common refinement) of partitions $\Pi = \{\Pi_1, \dots \Pi_n\}$ consists of singleton sets so that $\bigcap_{i \in I} \Pi_i(\omega) = \omega$ for all $\omega \in \Omega$, which means that the traders' pooled information always reveals the true state. Note that this assumption is satisfied in all four structures of the experiment. Let $E_\mu[X|\Pi_i(\omega)]$ be the expected value of security X, conditional on private information $\Pi_i(\omega)$ and prior μ. Let $Supp(\mu)$ be the support of μ.

Definition 1. *A security X is called non-separable under information structure Π if there exists probability distribution μ and value $v \in \mathbb{R}$ such that:*

(i) $X(\omega) \neq v$ for some $\omega \in Supp(\mu)$,

(*ii*) $E_\mu[X|\Pi_i(\omega)] = v$ *for all* $i = 1, ..., n$ *and* $\omega \in Supp(\mu)$.

Otherwise, it is called separable.

For structures t3s111 and t3s110, the security X is Arrow-Debreu, paying 1 in one state and 0 otherwise. Ostrovsky (2012) shows that Arrow-Debreu securities are separable irrespective of the information structure. For the other two structures, we will use the following characterization of separable securities by Ostrovsky (2012). It specifies that X is separable if and only if, for any v, we can find numbers $\lambda_i(\Pi_i(\omega))$, for each i and ω, such that the sum over all traders has the same sign as the difference of $X(\omega) - v$. Intuitively, for any v and at each ω, all traders "vote" and the sign of the sum of the votes has to agree with the sign of the difference between the value of the security and v.

Proposition 1 (Ostrovsky (2012))**.** *Security* X *is separable under partition structure* Π *if and only if, for every* $v \in \mathbb{R}$*, there exist functions* $\lambda_i : \Pi_i \to \mathbb{R}$ *for* $i = 1, \ldots, n$ *such that, for every state* ω *with* $X(\omega) \neq v$*,*

$$(X(\omega) - v)\sum_{i \in I} \lambda_i(\Pi_i(\omega)) > 0.$$

Suppose that v is strictly between 0 and 1. For structure t3s111y2, assign $\lambda_1 = 1$ to partition cells $\{a, b, c, d\}, \{a, b, e, f\}, \{a, c, e, g\}$ and $\lambda_2 = -1$ to all other cells. For structure t3s111o2ye2, assign $\lambda_1 = -1$ to cells $\{a, e\}, \{a, c\}, \{a, b\}$, $\lambda_2 = -10$ to cells $\{d, h\}, \{f, h\}, \{g, h\}$, and $\lambda_3 = 4$ to all other cells. Then, it is straightforward to check that the inequality $(X(\omega) - v)\sum_{i \in I} \lambda_i(\Pi_i(\omega)) > 0$ is satisfied for both structures and all states. If $v \geq 1$ or $v \leq 0$, then $(X(\omega) - v)$ has the same sign for all ω with $X(\omega) \neq v$, so finding appropriate λ is trivial.

D Theoretical myopic profits and volume

In this section, we derive the theoretical profits and trading volume for all four structures, if all agents are myopic and this is common knowledge. Let p_{start} denote the price of Yes in the previous round (or the initial price we are in the first round) and p_{end} denote the price after the AI agents conducts his trades. The β parameter represents the liquidity sensitivity parameter, which is set to 0.01 (Rajtmajer et al., 2022; Galanis et al., 2024).

Under the Logarithmic Market Scoring Rule (LMSR), the price of a binary contract is strictly a function of the difference in outstanding shares. The quantity of Yes shares, Δq, required to shift the market probability from p_{start} to p_{end} is given by

$$\Delta q = \frac{1}{\beta} \ln \left(\frac{p_{end}(1 - p_{start})}{p_{start}(1 - p_{end})} \right)$$

The cost of buying these Yes shares is

$$\text{Cost} = \frac{1}{\beta} \ln \left(\frac{1 - p_{start}}{1 - p_{end}} \right)$$

If the market resolves to Yes (yielding a payout of 1 per share), the trader's net profit is the total payout minus the initial cost. If it resolves to No, the Yes shares are worthless and the net profit is simply the negative cost:

$$\text{Profit}_{\text{Yes}} = \Delta q - \text{Cost} = \frac{1}{\beta} \ln \left(\frac{p_{end}}{p_{start}} \right)$$

$$\text{Profit}_{\text{Yes}} = -\text{Cost} = \frac{1}{\beta} \ln \left(\frac{1 - p_{end}}{1 - p_{start}} \right)$$

The calculations for buying No shares are similar. If a trader wishes to reduce the price of Yes from p_{start} to p_{end} (where $p_{end} < p_{start}$) and he has no Yes shares to sell, he must acquire No shares. The quantity of No shares, Δq_{no}, required to decrease the Yes price to p_{end} is given by:

$$\Delta q_{no} = \frac{1}{\beta} \ln \left(\frac{p_{start}(1 - p_{end})}{p_{end}(1 - p_{start})} \right)$$

The capital required to execute this transaction depends entirely on the ratio of the starting and ending "Yes" probabilities, mirroring the profit function of the opposing side:

$$\text{Cost}_{no} = \frac{1}{\beta} \ln \left(\frac{p_{start}}{p_{end}} \right)$$

If the market resolves to No' (yielding a payout of 1 per share), the trader's net profit is the total payout minus the initial cost. If the market resolves to Yes, then the profit from this transaction is the negative cost:

$$\text{Profit}_{\text{No}} = \Delta q_{no} - \text{Cost}_{no} = \frac{1}{\beta} \ln \left(\frac{1 - p_{end}}{1 - p_{start}} \right)$$

$$\text{Profit}_{\text{No}} = -\text{Cost}_{no} = -\frac{1}{\beta} \ln \left(\frac{p_{start}}{p_{end}} \right)$$

Using these formulas, we calculate the average profits and trading volume for each structure and trader.

E More graphs

References

Artificial Analysis Team (2025). Artificial analysis long context reasoning benchmark. Available at https://artificialanalysis.ai/evaluations/artificial-analysis-long-context-reasoning.

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics*, 4:1236–1239.

Table 15: Average Volume and Profits

Average over initial prices	**Easy t3s111y2**	**Medium t3s110**	**Hard t3s111**	**Very Hard t3s111o2ye2**	**Average over structures**
Volume	990	1133	1210	1047	1095
Trader 1 Volume	110	110	110	56	97
Trader 2 Volume	880	110	110	0	275
Trader 3 Volume	0	990	990	990	743
Trader 1 Profits	46	46	-64	6	9
Trader 2 Profits	29	-41	69	0	14
Trader 3 Profits	0	69	69	69	52
Average Profits	25	25	25	25	25

Bail, C. A. (2024). Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Barres, V., Dong, H., Ray, S., Si, X., and Narasimhan, K. (2025). τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Bini, P., Cong, L. W., Huang, X., and Jin, L. J. (2025). Behavioral economics of ai: Llm biases and corrections. *Available at SSRN 5213130*.

Cai, H., Chen, Y., and Fang, H. (2009). Observational learning: Evidence from a randomized natural field experiment. *American Economic Review*, 99(3):864–882.

Charness, G., Jabarian, B., and List, J. A. (2023). Generation next: Experimentation with ai. Technical report, National Bureau of Economic Research.

Chen, Y., Dimitrov, S., Sami, R., Reeves, D. M., Pennock, D. M., Hanson, R. D., Fortnow, L., and Gonen, R. (2010). Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, 58(4):930–969.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Chen, Y., Ruberry, M., and Vaughan, J. W. (2012). Designing informative securities. *arXiv preprint arXiv:1210.4837*.

Corgnet, B., Desantis, M., and Porter, D. (2018). What makes a good trader? on the role of intuition and reflection on trader performance. *The Journal of Finance*, 73(3):1113–1137.

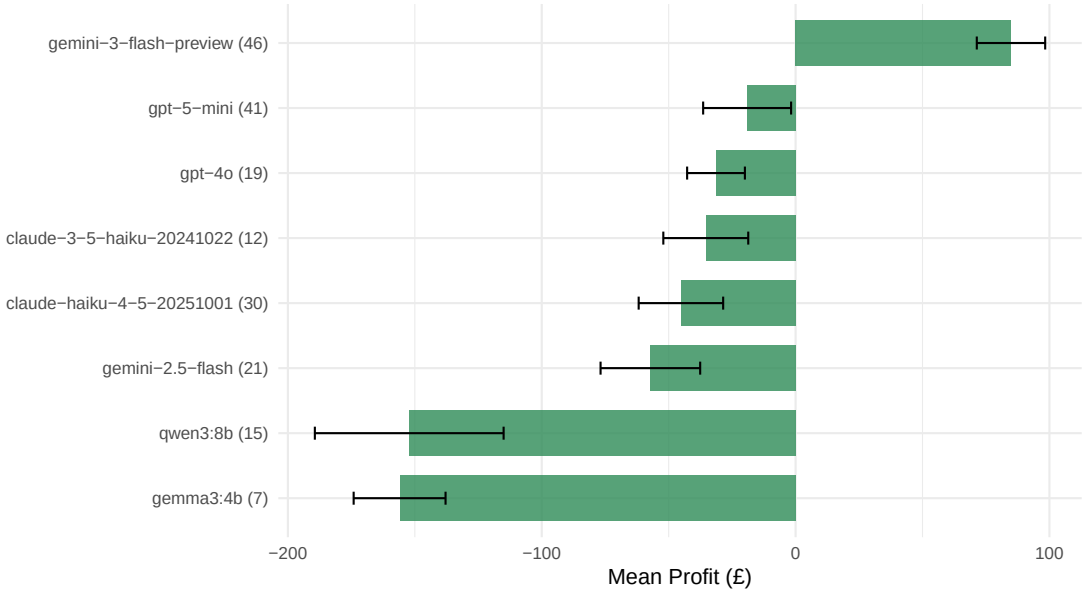


Figure 7: Average Profits by AI Model

Cultivate Labs (2021). How does the logarithmic market scoring rule (lmsr) work? https://www.cultivatelabs.com/prediction-markets-guide/how-does-logarithmic-market-scoring-rule-lmsr-work. Accessed: 04-17-2021.

Dimitrov, S. and Sami, R. (2008). Non-myopic strategies in prediction markets. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 200–209.

Galanis, S., Ioannou, C. A., and Kotronis, S. (2024). Information aggregation under ambiguity: theory and experimental evidence. *Review of Economic Studies*, 91(6):3423–3467.

Galanis, S. and Kotronis, S. (2021). Updating awareness and information aggregation. *B.E. Journal of Theoretical Economics*, 21:613–635.

Galanis, S. and Mikhalishchev, S. (2025). Information aggregation with costly information acquisition. *Mimeo*.

Geanakoplos, J. (1992). Common knowledge. *The Journal of Economic Perspectives*, 6(4):53–82.

Geanakoplos, J. and Polemarchakis, H. (1982). We can't disagree forever. *Journal of Economic Theory*, 28:192–200.

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119.

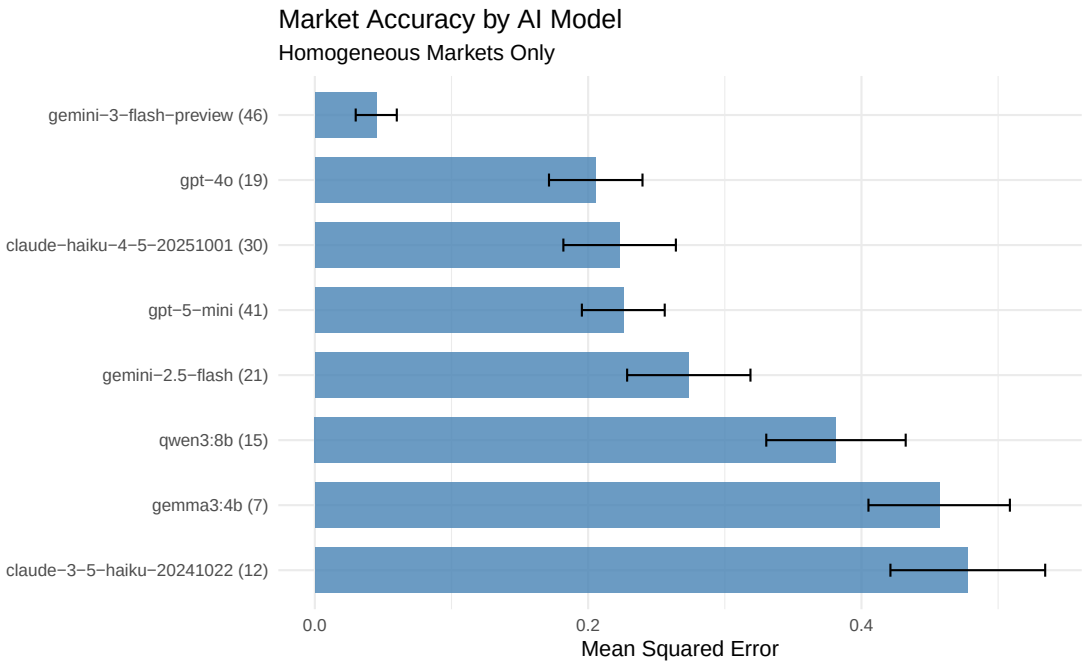


Figure 8: Mean Squared Error by AI Model

Hanson, R. (2007). Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, 1(1):3–15.

Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, 35(4):519–530.

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

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. (2024). Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L. P., and Murthy, A. B. (2024). Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*.

Kim, Y., Gu, K., Park, C., Park, C., Schmidgall, S., Heydari, A. A., Yan, Y., Zhang, Z., Zhuang, Y., Malhotra, M., et al. (2025). Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53:1315 – 1335.

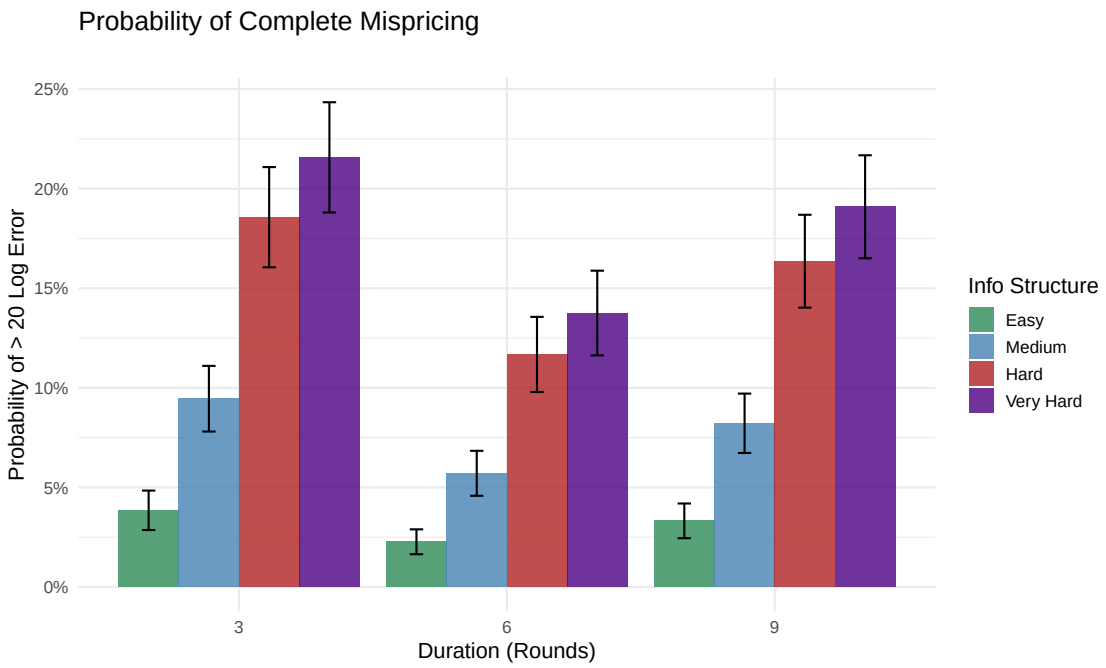


Figure 9: Probability of Complete Mispricing

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

McKelvey, R. D. and Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica*, 58:1321–1339.

Milgrom, P. and Stokey, N. (1982). Information, trade, and common knowledge. *Journal of Economic Theory*, 26:17–27.

Ostrovsky, M. (2012). Information aggregation in dynamic markets with strategic traders. *Econometrica*, 80(6):2595–2647.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. (2025). Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Pyatkin, V., Malik, S., Graf, V., Ivison, H., Huang, S., Dasigi, P., Lambert, N., and Hajishirzi, H. (2025). Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*.

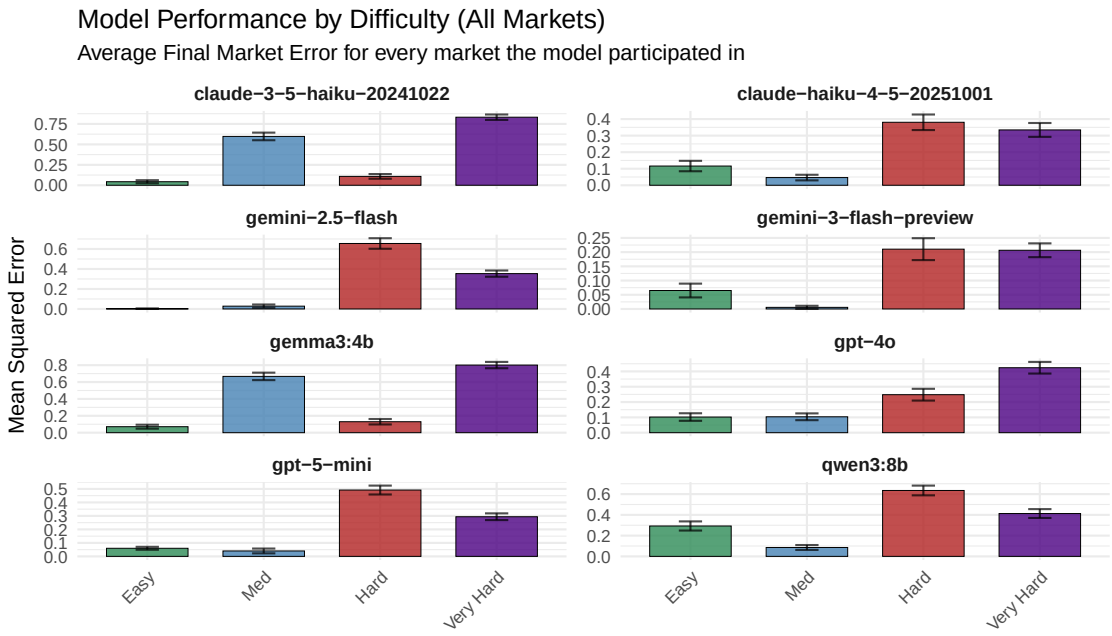


Figure 10: Mean Squared Error by Structure and Model

Rajtmajer, S., Griffin, C., Wu, J., Fraleigh, R., Balaji, L., Squicciarini, A., Kwasnica, A., Pennock, D., McLaughlin, M., Fritton, T., et al. (2022). A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 13218–13220.

Rasooly, I. and Rozzi, R. (2025). How manipulable are prediction markets? *arXiv preprint arXiv:2503.03312*.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. (2024). Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Schlegel, J. C., Kwaśnicki, M., and Mamageishvili, A. (2022). Axioms for constant function market makers. *Available at SSRN*.

Tian, M., Gao, L., Zhang, S., Chen, X., Fan, C., Guo, X., Haas, R., Ji, P., Krongchon, K., Li, Y., et al. (2024). Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems*, 37:30624–30650.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yang, C., Xuezhi, W., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. (2023). Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.