

Department of Economics

Information Aggregation with AI Agents

Spyros Galanis

Working Paper No. 2, 2026

Department Economics

Durham University Business School

Mill Hill Lane

Durham DH1 3LB, UK

Tel: +44 (0)191 3345200

<https://www.durham.ac.uk/business/about/departments/economics/>

© Durham University, 2025

Information Aggregation with AI Agents*

Spyros Galanis[†]

May 14, 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 negative impact, suggesting that AI agents may suffer from similar 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 are more profitable. Surprisingly, giving them feedback about past performance has no impact on aggregation.

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, partly because they are trained using Reinforcement Learning from Human Feedback (RLHF), 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. There are three traders and three signals, and we consider four levels of difficulty. 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 measure the effect of cheap talk communication, prompting them to be strategic (forward-looking) or myopic to measure whether 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 examine if more trading helps or confuses AI agents.¹ These treatments generate a total of 144 different

¹The initial price variable is measured as the initial log error between the initial price and the true value of the security.

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’.² 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 for our main analysis: Claude Haiku 3.5 and 4.5, Gemini 2.5 Flash and Gemini 3 Flash, GPT-4o and GPT-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 (Result 1) is that, although the median market is effective in the two easy structures, increasing the complexity of the information structure significantly degrades information aggregation.³ 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. In 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 may resemble some 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.

We conducted our baseline experiment in January 2026. However, given the rapid pace of innovation in artificial intelligence, model architectures are updated almost quarterly. To ensure the robustness of our findings, we conducted an out-of-sample extension of 576 markets in April 2026 using three newly released capability-frontier models: GPT-5.4, Claude Opus 4.6, and Gemini 3.1 Pro. As detailed in Section 6.5, comparing four frontier teams against our top four baseline teams yields no evidence that newer models improve information aggregation. Surprisingly, we find that models from the same company exhibit very similar failure modes across the two cohorts. This architectural persistence suggests that failure to reason interactively may be heavily dictated by proprietary alignment choices, such as how

²A 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.

³The average market fares very poorly across all structures, but this is driven mostly by the performance of the bottom 20% markets.

RLHF is implemented, rather than a universal limitation of AI scaling.

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, 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 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, using the same versions of models as in the first wave. Contrary to recent literature suggesting that LLMs successfully improve when given performance histories (Yang et al., 2023), we find that providing AI agents with qualitative results on past market data has no effect on information aggregation.

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 no effect, 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, our text analysis uncovers what initially appears to be a highly sophisticated inter-temporal communication strategy. Rather than maintaining a constant rate of deception, AI agents display a distinct “sawtooth” pattern of revelation: they actively hoard information and deceive competitors in the opening rounds to protect their information rents, but their probability of truthful revelation spikes sharply at specific intervals (Rounds 3, 6, and 9). At first glance, this mimics a sophisticated agent who selectively reveals more only when the financial penalty for honesty drops to zero. However, deeper analysis reveals a profound failure in dynamic planning. This sawtooth pattern persists even when we restrict our sample to 9-round markets. Agents prematurely drop their deception in Rounds 3 and 6, acting as if the game is ending, before reverting to high-deception baseline behaviors in the subsequent rounds. This suggests that AI agents do not fully grasp the dynamic nature of the market and they cannot plan ahead.

The inability to plan ahead aligns with recent findings in the computer science literature demonstrating that LLMs lack the capacity for autonomous multi-step planning and dynamic state tracking (Valmeekam et al., 2023; Kambhampati et al., 2024). Rather than executing a cohesive inter-temporal plan via backward induction, the models appear constrained by their architectural reliance on pattern matching and their inherent limitations in multi-step reasoning tasks (Dziri et al., 2023).

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 Mikhailishchev (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, and 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, and 7, agent 2 trades in rounds 2, 5, and 8, and agent 3 trades in rounds 3, 6, and 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, $X : \Omega \rightarrow \mathbb{R}$ (‘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 $p_Y = \frac{e^{\eta q_Y}}{e^{\eta q_Y} + e^{\eta q_N}}$, where η is a liquidity parameter and q_Y, q_N denote the number of outstanding Yes and No shares.⁴ 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.

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 : $-[y \ln(p) + (1 - y) \ln(1 - p)]$.⁵ We also examine the trade volume and profitability of traders in markets, however we do not have theoretical predictions for these measures.

⁴The liquidity parameter (we use $\eta = 0.01$) 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.

⁵When 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.

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 are closed-weight models: Claude Haiku 3.5, Claude Haiku 4.5, Gemini 2.5 Flash, Gemini 3 Flash, GPT-4o, and GPT 5 mini. The last two were open-weight models and they were run locally: gemma3:4b, and qwen3:8b.⁶ 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), accessed in January 2026.⁷ 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.⁸ 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 APIs, so that text generation is neither too rigid (deterministic) nor chaotic. Figures 8 and 9 show the performance of each model in terms of log error and profits.

In Section 6.5, we report the results from a third wave of 576 markets that we ran in April 2026, with three frontier models: GPT-5.4 (57), Claude Opus 4.6 (53), and Gemini 3.1 Pro (57), to check for robustness of our main results and in particular Result 1. The Artificial Intelligence Indices, accessed in April 2026, are reported in parentheses. They are not comparable with those accessed in January 2026 for the previous LLMs, as different tests and evaluations of models are added over time. Therefore, we keep the analysis separate.

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, calculates 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 there are in the market and what is the current round. Part

⁶In 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.

⁷The 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.

⁸See Appendix A in Kim et al. (2025) for details.

Table 1: AI Models, Team Compositions and Intelligence

Team Composition	Average Intelligence (Standard Deviation)
<i>Homogeneous Teams (3 Identical Agents)</i>	
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)
<i>Heterogeneous Teams (Mixed Agents)</i>	
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-weight models that were downloaded and run locally, whereas the other models were accessed through an API. All eight models have the same version number throughout the experiments.

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 the Supplemental Appendix 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

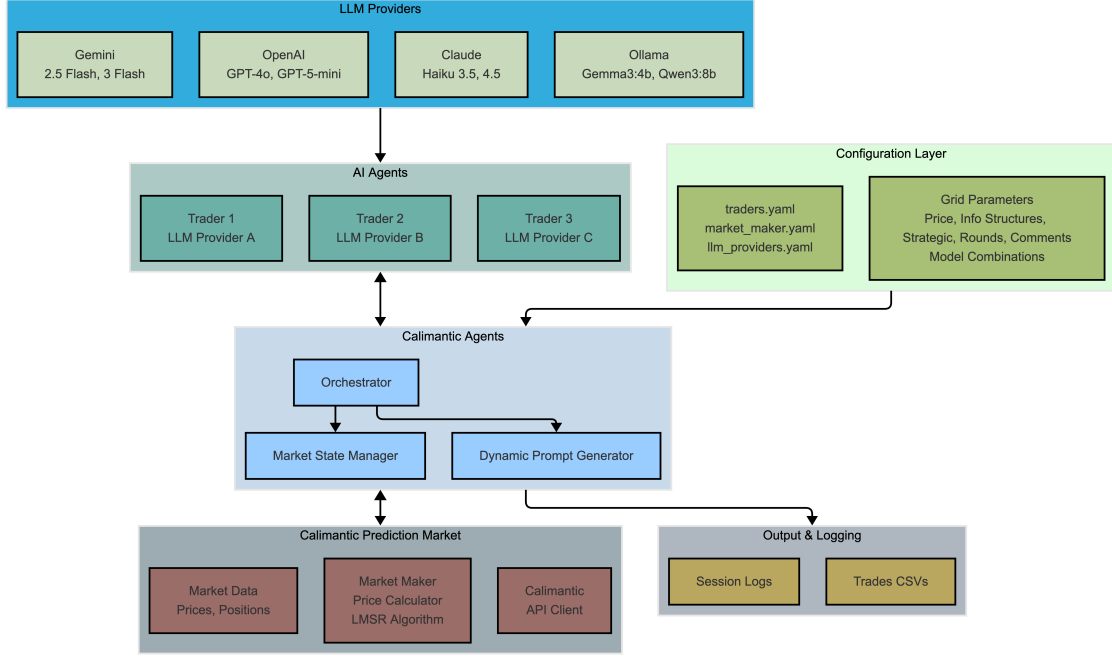


Figure 1: Prediction market platform

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 the Supplemental Appendix 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 t3s111o2ye2.⁹ 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 partition cells, depicted in the last part of Table 2. For example, the information partition of Trader 2 is $\{\{a, c\}, \{b, d\}, \{e, g\}, \{f, h\}\}$. 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.

⁹The 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
<u>t3s111y2, t3s110, t3s111</u>	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
<u>t3s111o2ye2</u>	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 0.75.¹⁰ 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 state 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 set shrinks. 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.

¹⁰Whether 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 no one made the connection.

It is described as follows. There are three girls who 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 price y_0 , and discount factor $\gamma \in (0, 1]$, in any Nash equilibrium, information gets aggregated.*

In Appendix [B](#), 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: $X = (\textcolor{blue}{1}, 1, 1, 0, 1, 0, 0, 0)$				
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: $X = (1, \textcolor{blue}{0}, 0, 0, 0, 0, 0, 0)$				
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: $X = (\textcolor{blue}{1}, 0, 0, 0, 0, 0, 0, 0)$				
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: $X = (\textcolor{blue}{0}, 1, 1, 0, 1, 0, 0, 0)$				
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 averaged 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 .¹¹ 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.*

¹¹See Galanis et al. (2024) for such an experiment with humans in a prediction market.

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 human participants. 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 6. *Information aggregation improves as the average intelligence of the group increases.*

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 C. 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 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 is as follows: Easy (990), Medium (1133), Hard (1210), Very Hard (1046). We therefore have the following hypothesis.

Hypothesis 7. *Trading volume is ordered from lowest to highest as follows: Easy ($t3s111y2$) < Very Hard ($t3s111o2ye2$) < Medium ($t3s1110$) < 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).

Hypothesis 8. *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 always 25, as shown in Table 15. We therefore have the following hypothesis.

Hypothesis 9. *Average profits are not correlated with the difficulty of the structure.*

[Corgnet et al. \(2018\)](#) show in a market experiment with human participants 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 profitability is driven by relative cognitive advantage: it is positively correlated with an agent’s own intelligence, but negatively correlated with the average intelligence of the market.*

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.¹² In Section 6.5, we report the results from a third wave of 576 markets that we ran in April 2026, with three frontier models (GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro) to check for robustness of our results, particularly Result 1.

We focus on the first wave because the statistics do not change when the second wave is added. 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) reveal 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.

5.1 Methodological Considerations

For our primary OLS specifications, we utilize CR2 cluster-robust standard errors clustered at the AI team composition level, featuring small-sample degrees of freedom adjustments (Pustejovsky and Tipton, 2018). This approach conservatively estimates treatment effects by accounting for intra-cluster correlation within specific AI teams. However, certain subsamples—such as the relative performance tests in Section 6.5 comparing April 2026 frontier

¹²With 12 teams of AI agents, the first wave should have consisted of $4 \times 3 \times 2 \times 2 \times 3 \times 12 = 1728$ 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

models against the January cohort—reduce the number of unique clusters to the point where cluster-robust inference becomes rank-deficient or severely underpowered. Rather than artificially inflating standard errors or engaging in stepwise variable omission, we address this by collapsing the longitudinal dimension of these sub-samples and employing exact small-sample statistical tests (Wilcoxon signed-rank test).

For our auxiliary quantile regressions in the main text, which assess typical market behavior (Q50) and worst-case tail risk (Q80), we report bootstrapped standard errors (B=200 resamples) rather than analytical cluster-robust standard errors. Estimating analytical clustered variance matrices for quantile regressions requires local error density estimation that is notoriously unstable and fragile in finite samples. Consequently, we employ standard XY-pair bootstrapping for these specific models. We emphasize that the primary purpose of these quantile specifications is to demonstrate the behavioral stability of the point estimates across the error distribution, rather than to serve as exact clustered inference.

Furthermore, because several of our core theoretical propositions predict a null result (specifically, the persistent failure of models to aggregate information), relying solely on conservative CR2 adjustments could theoretically favour our hypotheses by making it artificially easy to fail to reject the null. To stress-test these findings, we additionally estimate our OLS models using the unclustered HC1 standard errors, which we report in the Online Appendix. We find that our null results hold even under these strictly less conservative assumptions.

An alternative method for addressing the issue of a small number of clusters would be

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.

to significantly expand the array of LLMs employed, from 12 to more than 40. However, this approach introduces two severe limitations. First, we intentionally selected models exclusively from the leading frontier laboratories (Anthropic, Google, and OpenAI) alongside highly capable open-weight models. Expanding the baseline to a much larger quantity of models would have necessitated including peripheral, lower-tier agents. Had we done so, any observed breakdown in information aggregation could be easily dismissed as a trivial failure of basic competence. By artificially raising the capability floor and restricting our sample strictly to established models, we ensure that failures in information aggregation represent genuine limitations of frontier models. Second, expanding the sample to over 40 models would cause the number of simulated markets to jump from 4,000 to more than 13,000. Beyond significantly increasing the computational overhead, this exponential scaling would compromise our ability to conduct rigorous manual auditing and maintain the strict quality control required to monitor dynamic, multi-agent interactive simulations.

Finally, we deliberately refrain from estimating pooled regressions that combine the January and April cohorts. While pooling would ostensibly increase statistical power, doing so requires controlling for baseline model capabilities across generations. However, traditional standardized indices (e.g., MMLU, HumanEval) suffer from widespread training data contamination and fail to maintain reliable cross-generational consistency between early-2025 and mid-2026 models. Because these static benchmarks are non-stationary, utilizing them to construct a unified intelligence metric across differing vintages would introduce severe measurement error. By restricting our analysis to cohort-specific distributions and matched relative performance tests, we ensure our findings remain robust to the rapid deprecation of static AI benchmarks.

6 Results

We first report our results from the first wave (1772 markets), for information aggregation (Section 6.1), and trading volume and profits (Section 6.2), ignoring the information provision treatment. In Section 6.3, we describe the information provision treatment, utilising the full sample of 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. In Section 6.4, we examine how AI agents use the public and private comments, whereas in Section 6.5 we conduct a robustness check with the three frontier models from April 2026.

6.1 Information Aggregation

We present our main results for information aggregation 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 1.¹³ 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 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.¹⁴

In our OLS specification, we observe that as the structure becomes highly difficult, the average log error increases. The intercept (7.548) implies that the average baseline market in the easiest structure (t3s111y2) already suffers from severe mispricing. Using this baseline, the Medium structure shows no statistically significant increase in error ($\beta = 2.200, p > 0.1$). While the Hard ($\beta = 5.327, p < 0.05$) and Very Hard ($\beta = 6.658, p < 0.05$) structures generate massive penalties relative to the baseline, post-estimation Wald tests utilizing CR2 small-sample adjustments reveal that the means of adjacent highly complex structures are statistically indistinguishable. Accounting for finite-cluster penalties, we fail to reject the equality of coefficients between Medium and Hard ($F_{1,11} = 0.62, p = 0.448$), as well as between Hard and Very Hard ($F_{1,11} = 0.095, p = 0.764$).

This extreme variance suggests that the mean effect is driven by noisy, catastrophic outliers rather than a uniform shift in behavior. To disentangle typical market behavior from these outliers, we analyze the median ($\tau = 0.5$). The intercept of the median regression (≈ 0.018) implies that the typical baseline market converges almost perfectly to the truth (implied probability $\approx 98\%$). The Medium structure performs identically well. However, at the Hard structure, typical success drops drastically (implied probability $\approx 75\%, p < 0.01$). At the Very Hard structure, the median error penalty expands to 0.700 ($p < 0.001$), dropping the implied probability to $\approx 50\%$. Thus, while the Very Hard structure causes the average market to hallucinate wildly, it strictly causes the typical market to be no better than a coin toss.

¹³Recall that the log error between actual (y) and predicted (p) is $-(y \ln(p) + (1 - y) \ln(1 - p))$, where the predicted probability is bounded, $p \in [\epsilon, 1 - \epsilon]$, $\epsilon = 10^{-15}$, to avoid having an infinite log error.

¹⁴An 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.

Table 6: Information Aggregation Without Disclosure

	<i>Dependent variable:</i>		
	Mean (OLS)	Log Error Median (Q50)	Tail Risk (Q80)
	(1)	(2)	(3)
Constant	7.548*** (2.185)	0.018 (0.019)	3.170** (1.138)
Comments Allowed	0.191 (0.319)	0.001 (0.003)	0.000 (0.032)
Duration: 6 Rounds	−1.817 (1.172)	−0.001 (0.005)	−0.057 (0.177)
Duration: 9 Rounds	−0.627 (1.250)	−0.001 (0.005)	−0.028 (0.165)
Strategic Prompt	−0.053 (0.292)	−0.001 (0.003)	−0.000 (0.113)
Medium (t3s110)	2.200 (2.635)	−0.000 (0.003)	0.038 (1.808)
Hard (t3s111)	5.327* (2.317)	0.286** (0.106)	12.470 (7.451)
Very Hard (t3s111o2ye2)	6.658* (2.974)	0.700*** (0.084)	21.771*** (4.744)
Initial Error: Low	0.014 (0.120)	0.000 (0.004)	0.000 (0.069)
Initial Error: Medium	0.225 (0.186)	−0.000 (0.003)	−0.000 (0.063)
Average Intelligence	−0.221*** (0.059)	−0.0004 (0.0004)	−0.069** (0.025)
Intelligence SD	0.044 (0.227)	−0.0003 (0.0003)	−0.052* (0.026)
6 Rounds x Strategic	0.067 (0.261)	0.001 (0.004)	−0.000 (0.128)
9 Rounds x Strategic	−0.076 (0.385)	0.001 (0.004)	−0.028 (0.161)
Medium Struct x Initial Error (Low)	1.309 (1.279)	−0.000 (0.005)	0.019 (3.595)
Hard Struct x Initial Error (Low)	0.504 (0.682)	0.032 (0.124)	19.889 (12.267)
Very Hard x Initial Error (Low)	2.861 (1.623)	0.013 (0.167)	10.425* (5.186)
Medium Struct x Initial Error (Medium)	−0.725 (0.436)	0.000 (0.004)	0.019 (1.810)
Hard Struct x Initial Error (Medium)	−0.236 (0.715)	0.077 (0.149)	−9.945 (12.141)
Very Hard x Initial Error (Medium)	−0.337 (1.021)	−0.007 (0.083)	10.080 (9.342)
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 upper 20% tail of the log error distribution (the worst-performing markets). OLS model uses CR2 cluster-robust standard errors clustered by exact AI team composition, featuring small-sample degrees of freedom adjustments. Quantile models report standard bootstrapped standard errors due to small-cluster convergence constraints. 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.

In summary, we firmly reject the hypothesis that information successfully aggregates across all market structures. We find that as complexity increases, information aggregation strictly deteriorates—driven by catastrophic outliers in the average market and a fundamental collapse of accuracy in the typical market. Figure 2 visually confirms this tiered deterioration across both the mean and median log error.

Result 1. *Information aggregation deteriorates strictly at higher levels of complexity. While extreme outliers obscure the mean effects, the typical (median) market follows a distinct, tiered collapse: Easy ($t3s111y2$) $\approx$ Medium ($t3s110$) $<$ Hard ($t3s111$) $<$ Very Hard ($t3s111o2ye2$).*

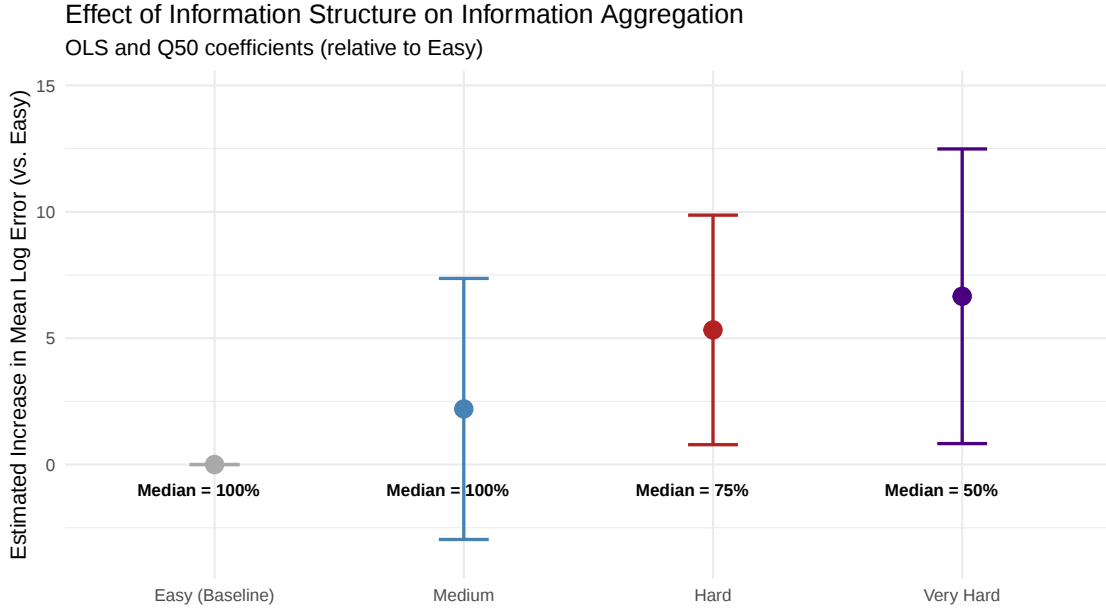


Figure 2: **The Complexity Effect.** Point estimates and 95% CR2 confidence intervals represent the OLS mean marginal effect of the structure. Text labels indicate the Q50 median marginal effects, displayed as implied probabilities ($p = e^{-\log \text{error}}$). As the structure becomes more complex, the information aggregation deteriorates.

While Figure 2 demonstrates the magnitude of the final aggregation error, Figure 3 illustrates the dynamic mechanism behind this failure. By plotting the empirical price paths against the theoretical myopic equilibrium derived in Section 3.2, we observe a striking breakdown in AI reasoning. In the Easy and Medium structures, the High-Intelligence agents tightly track the myopically optimal prices. However, in the Hard and Very Hard structures, this tracking collapses. Rather than following the theoretical price to 1 or 0, the smartest agents anchor their prices near 0.5, a trajectory consistent with difficulty in higher-order reasoning. Conversely, the Low-Intelligence (red) agents have a similar price trajectory that converges to 0.75 in both Hard and Very Hard, even though the true value is 1 in the former

and 0 in the latter. This demonstrates that any apparent ‘success’ in complex structures is an artifact of noisy trading rather than sophisticated interactive reasoning.

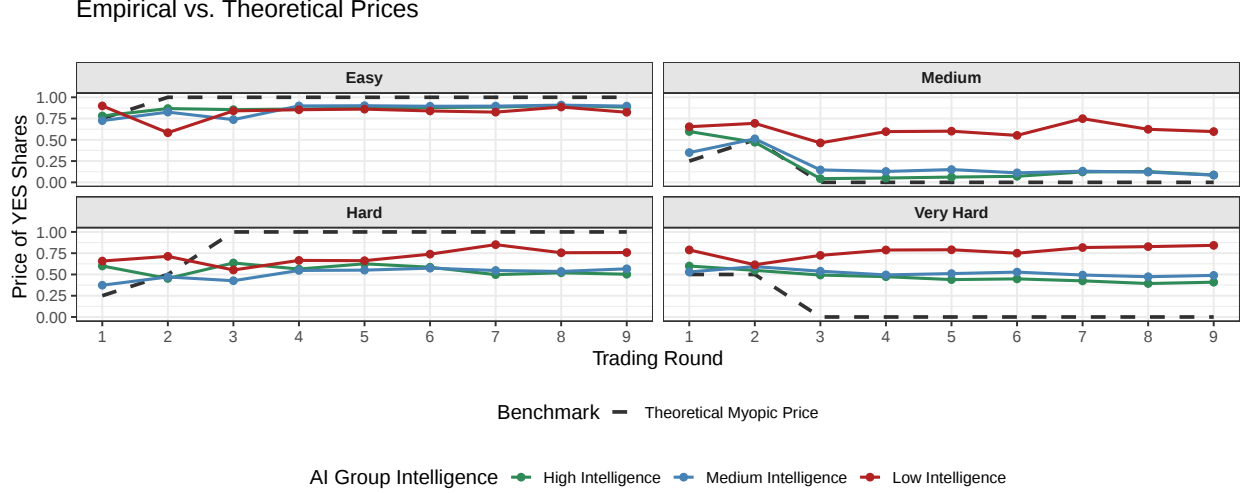


Figure 3: Myopically Optimal and Actual Prices Across Structures. Smarter markets tightly track the myopically optimal prices in the Easy and Medium structures, but revert to 0.5 in the Hard and Very Hard. Low-intelligence markets perform better in the Hard structure but this seems more an artifact of noisy trading rather than sophisticated interactive reasoning.

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 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.

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

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

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. 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 ($\beta = 0.01, p > 0.9$), 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.*

Interestingly, this robustness does not hold in the most complex environment. The interaction between the Very Hard structure and the low initial log error (Very Hard $\times$ Initial Error (Low)) generates a large, positive, and statistically significant coefficient in both the OLS mean regression ($\beta = 2.86, p < 0.1$) 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 increases, which is consistent with the theory. 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 means that an increase from the lower intelligence scores (7) teams to the highest (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 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, preventing low-intelligence agents from confidently driving the security price to the opposite direction.

Figure 5 maps how AI agents, grouped in three intelligence tiers, discover the true value of the security. The probability of the true state is the price of Yes if the answer is Yes

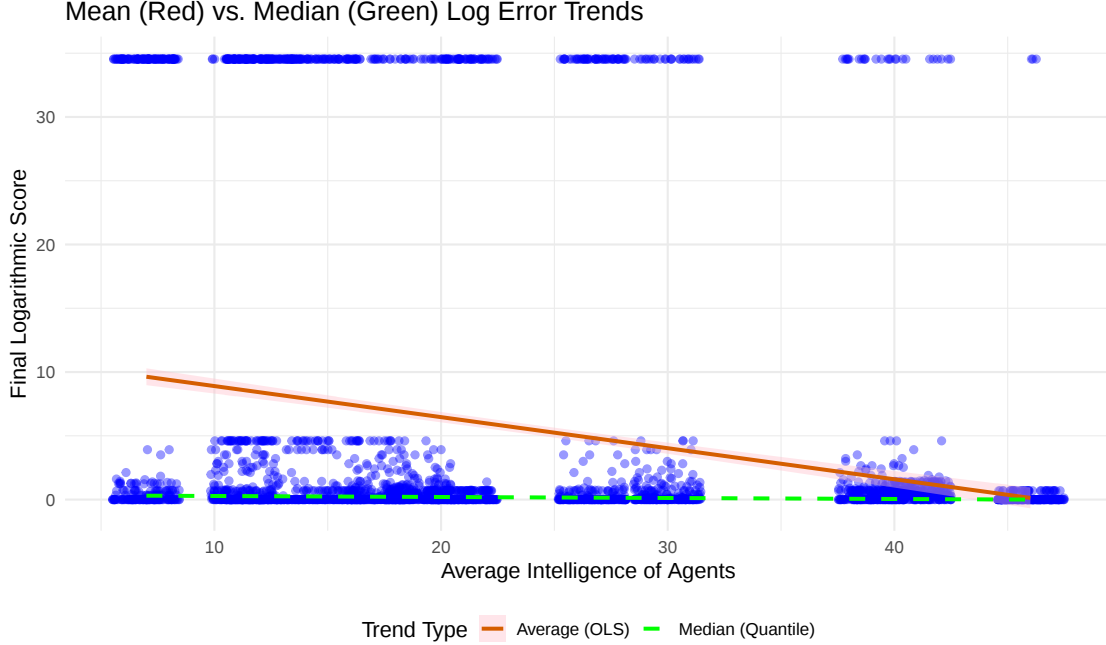


Figure 4: **The Intelligence Effect.** As the AI agents become smarter, information aggregation (mean log error) improves, but the median log error is unaffected. Points are jittered horizontally to show density of markets with the same intelligence.

and the price of No if the answer is No. The High Intelligence cohort (green line) exhibits rapid informational efficiency, sharply converging toward the true state and subsequently stabilizing. In contrast, the Low Intelligence cohort (red line) exhibits a ‘sawtooth’ trajectory, indicating noisy, inefficient belief updating.

While higher intelligence improves performance on average, the quantile regressions reveal that this is driven primarily by tail 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.*

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 broadly aligns with myopic theoretical predictions, though with notable statistical caveats. First, while point estimates suggest volume increases from the Easy baseline to the Very Hard structure, this difference is completely statistically insignificant ($p > 0.1$). Second, both the Hard and Medium structures generate marginally significantly more volume than the Easy baseline ($p < 0.1$). However, while the theoretical expectation predicts that volume in the Hard market should exceed the Medium market, our point estimates reverse this order. To rigorously test this, we execute a post-estimation Wald test utilizing our CR2 cluster-robust variance-covariance matrix. The test fails to reject the null hypothesis of equality between the two coefficients ($F_{1,1752} = 0.94, p = 0.332$), indicating that trading volume between the Hard and Medium structures is statistically indistinguishable.

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

Second, we firmly reject the hypothesis that individual profits are positive. In our OLS specification, the constant is negative and statistically significant, indicating that baseline expected profits are strictly negative, so that on average, participation in the market generates financial losses. See Figure 9 for a breakdown of profits by AI model, where the only model with positive profits is Gemini 3 Flash. However, an analysis of the trading sequence reveals a striking microstructure dynamic. At first glance, the coefficients for trading second or third appear statistically indistinguishable from the baseline first-mover ($p > 0.1$). Yet, the point estimates suggest a massive divergence: trading second carries a severe -42.1 penalty, while trading third yields a 77.4 premium. To test the relative advantage of acting last, we perform a post-estimation Wald test utilizing our cluster-robust standard errors. The test overwhelmingly rejects the equality of the second and third trading positions ($F_{1,5293} = 48.30, p < 0.001$). This confirms a highly specific late-mover advantage: while neither late position is statistically distinguishable from the first mover, the final agent to act extracts a massive, highly significant financial advantage over the intermediate agent. See Figure 6 for an illustration.

Result 8. *Market participation generates losses rather than profits. However, trading last (third) yields a highly significant financial advantage over trading in the middle (second).*

We also reject the hypothesis that average profits are uncorrelated with structural complexity. As shown in Table 7, the negative coefficient observed in the Medium ($\beta = -30.5, p > 0.1$) treatment is not statistically significant relative to the Easy baseline. However, as complexity increases further, the profit penalties become severe. To confirm the strict monotonicity of this decline, we execute post-estimation Wald tests utilizing our CR2 cluster-robust variance-covariance matrix. These tests reject the equality of coefficients between the Medium and Hard structures ($F_{1,5293} = 15.28, p < 0.001$), as well as between the Hard and Very Hard structures ($F_{1,5293} = 5.86, p = 0.015$). This confirms that once complexity crosses a certain threshold, average profits strictly and monotonically decrease. This is theoretically consistent with Result 1: the deterioration of information aggregation

equates to an increasing divergence between the final market price and the true value of the security, mechanically resulting in lower average profits.

Result 9. *Individual profits decrease as the structure becomes more complex: Easy ($t3s111y2$) $\approx$ Medium ($t3s110$) $>$ Hard ($t3s111$) $>$ Very Hard ($t3s111o2ye2$).*

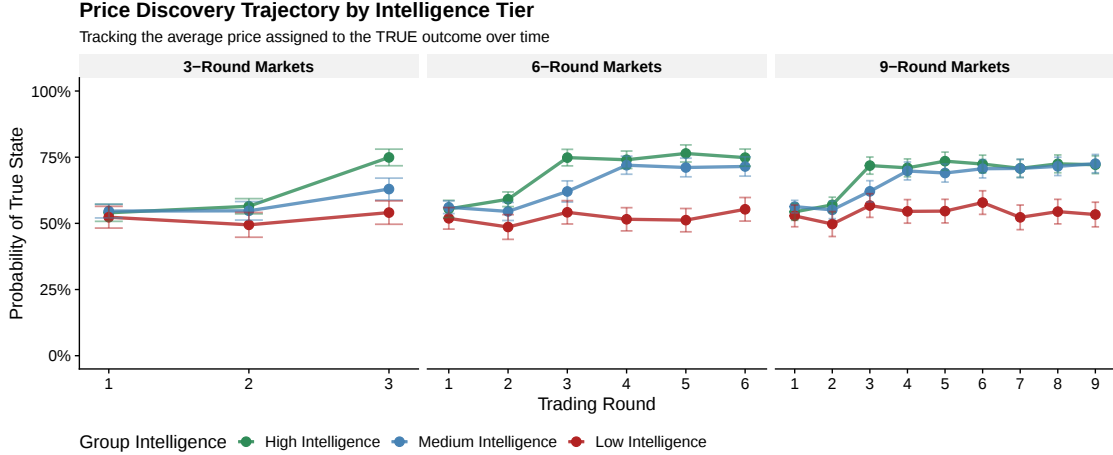


Figure 5: **The Price Discovery Trajectory.** The High Intelligence cohort of AI agents is always closer to the truth, dominating the Medium and Low Intelligence cohorts in 3, 6, and 9 round markets.

We also observe a substitution effect between communication and trading: allowing public comments reduces trading volume ($\beta = -85.0, p < 0.05$), suggesting that verbal coordination reduces the need for costly signalling through price mechanisms.

The analysis of individual agent profitability (Table 7, Column 2) reveals that financial success in these markets is primarily driven by relative rather than absolute cognitive advantage. We find a strong positive effect of individual intelligence on profit ($\beta = 14.08, p < 0.001$), coupled with a nearly equal and opposite negative effect from the average intelligence of the market ($\beta = -11.02, p < 0.001$). This confirms that agents benefit significantly from their own cognitive capacity, but suffer from the intelligence of their competitors.

Result 10. *Individual profitability is driven by relative cognitive advantage: an agent’s profits increase strictly with their own intelligence, but decrease significantly as the average intelligence of their competitors rises.*

6.3 Information Provision

Can AI agents improve their market performance by receiving feedback from historical play? Yang et al. (2023) propose an information provision technique that leverages LLMs as iterative optimizers: by updating an LLM’s prompt with a history of past actions and their

Table 7: Trading Volume and Profits Without Disclosure

	<i>Dependent variable:</i>	
	Trading Volume	Individual Profits
	(1)	(2)
Constant	1,795.988 (1,064.073)	−80.719* (31.683)
Comments Allowed	−85.077* (42.516)	−2.723 (4.048)
Duration: 6 Rounds	543.398*** (108.871)	2.447 (20.007)
Duration: 9 Rounds	1,037.835*** (192.429)	−5.396 (17.292)
Strategic Prompt	65.396 (36.642)	0.491 (2.876)
Medium (t3s110)	308.391 (174.854)	−30.548 (33.429)
Hard (t3s111)	128.183 (74.893)	−69.955* (32.035)
Very Hard (t3s111o2ye2)	55.372 (266.568)	−98.904* (42.381)
Initial Error: Low	−180.641** (68.906)	−10.767** (3.809)
Initial Error: Medium	−14.704 (34.607)	−4.386 (2.691)
Average Intelligence	−27.925 (24.667)	−11.019*** (2.252)
Individual Intelligence		14.079*** (2.150)
Intelligence SD	22.347 (41.617)	0.606 (2.473)
Trades Second		−42.127 (28.229)
Trades Third		77.390 (55.932)
6 Rounds x Strategic	−9.810 (55.547)	1.725 (4.474)
9 Rounds x Strategic	−87.983 (62.172)	3.680 (4.496)
Medium Struct x Initial Error (Low)	273.357 (158.440)	−21.173 (19.587)
Hard Struct x Initial Error (Low)	87.880 (70.669)	−13.015 (10.054)
Very Hard x Initial Error (Low)	414.759 (264.776)	−32.672 (23.473)
Medium Struct x Initial Error (Med)	41.568 (67.534)	4.394 (4.099)
Hard Struct x Initial Error (Med)	62.984 (55.047)	11.491 (8.982)
Very Hard x Initial Error (Med)	−238.720 (401.800)	−1.778 (11.232)
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 utilize CR2 cluster-robust standard errors clustered by exact AI team composition, featuring small-sample degrees of freedom adjustments. Rounds, Information Structure, and Trader Position (Model 2 only) use treatment contrasts, with ‘Round 3’, ‘t3s111y2’ (Easy), 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.

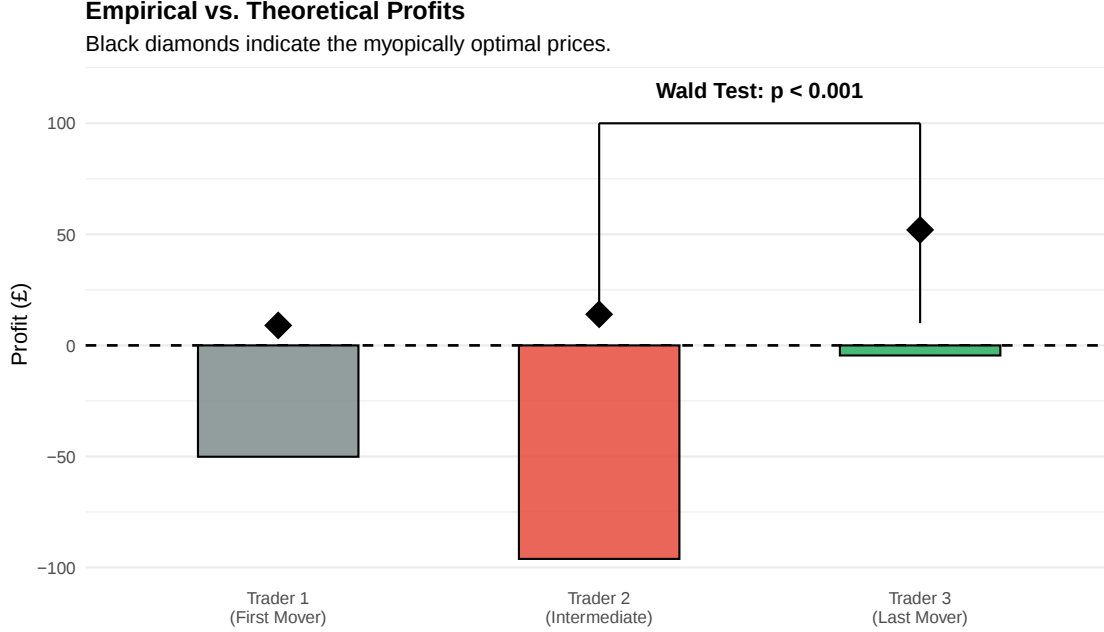


Figure 6: **The Trading Position Effect.**

relative success, the model self-corrects, outperforming human-designed prompts by up to 50%.

Building on this mechanism, we postulate that providing AI agents with aggregated historical data about market dynamics will enable them to form more accurate predictions, thereby strictly improving information aggregation. However, the anticipated effect on individual profitability is ambiguous. Because this informational shock is provided uniformly across all agents in the market, no single trader gains an asymmetric informational advantage. We formalize these hypotheses below:

Hypothesis 11. *The effect of information provision on information aggregation is positive.*

Hypothesis 12. *The effect of information provision on individual profits is zero.*

To test these hypotheses, we introduce an information provision treatment (e.g., [Berg et al. \(1995\)](#), [Cai et al. \(2009\)](#)). Specifically, we use the empirical results from our initial baseline sample (1,772 markets) to construct an exogenous informational shock. This summary information is then 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. This temporal separation guarantees that the information provided in the second phase is strictly exogenous to the error term of the new markets.

Crucially, to prevent data leakage, we utilize identical LLM model versions across both the baseline and treatment samples via a stateless API. LLMs lack endogenous memory

of past interactions; every API call, even within sequential rounds of the same market, instantiates a completely blank version of the model. Thus, the only mechanism through which the models experience “memory” of the first 1,772 markets is through our explicit, exogenous text prompt.

Below is the text appended to the prompt in the Experiment Disclosure treatment.¹⁵ We intentionally adopted a neutral tone, providing only qualitative historical data without attempting to steer the AI agents on how to execute their trading strategies based on this data:

=== 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.

¹⁵The qualitative summaries provided in the information treatment were constructed using provisional analysis (employing HC1 robust standard errors) immediately following the first wave of the experiment. While our final, more conservative CR2 clustered specifications render some adjacent structural boundaries statistically indistinguishable (e.g., the effect of Easy and Medium on market accuracy), the directional point estimates for the treatment parameters remain identical. Consequently, the information provided represents a valid, directionally accurate heuristic regarding the historical market environment. The provisional analysis is reported in the Online Appendix.

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 presents the mixed-contrast regression models for information aggregation across the full, pooled sample of 3,500 markets, introducing a dummy variable for the Experiment Disclosure treatment. Strikingly, we firmly reject Hypothesis 11. The provision of historical market data yields no statistically significant improvement in market accuracy across the mean ($\beta = 2.075, p > 0.1$), the median ($\beta = -0.00004, p > 0.1$), or the tail-risk distributions. Furthermore, the interaction term between Disclosure and Average Intelligence is completely null. This indicates a behavioral limitation: even the most highly capable LLMs in our sample failed to internalize the explicit strategic feedback to improve their probability estimations.

Result 11. *Information provision has a null effect on information aggregation.*

Table 9 replicates this analysis for market activity and individual profitability. Consistent with Hypothesis 12, the disclosure treatment has no statistically observable impact on either trading volume ($\beta = 171.9, p > 0.1$) or individual profits ($\beta = -21.05, p > 0.1$). Providing all participants with the exact same strategic heuristics did not generate a measurable advantage for any individual trader.

Result 12. *Information provision has a null effect on individual profits.*

Finally, analyzing the pooled sample of 3,500 markets confirms that all primary findings from the first wave of experiments are highly robust. The baseline market mechanics remain unchanged: strategic prompting, communication, initial price, and market duration have no significant impact on information aggregation. Conversely, the structural complexity of the market continues to significantly hinder aggregation, while higher average intelligence strictly improves it.

Table 8: Information Aggregation With Disclosure

	<i>Dependent variable:</i>		
	Mean (OLS)	Log Error Median (Q50)	Tail Risk (Q80)
	(1)	(2)	(3)
Constant	7.196*** (2.061)	0.018* (0.008)	5.638*** (0.954)
Comments Allowed	0.318 (0.258)	0.001 (0.001)	0.014 (0.038)
Duration: 6 Rounds	−1.550 (1.118)	−0.001 (0.001)	−0.107 (0.112)
Duration: 9 Rounds	−0.328 (1.183)	−0.001 (0.001)	−0.019 (0.086)
Strategic Prompt	−0.007 (0.231)	−0.001 (0.001)	−0.007 (0.064)
Medium (t3s110)	2.772 (3.002)	−0.00001 (0.001)	0.075 (0.971)
Hard (t3s111)	5.222* (2.225)	0.206* (0.093)	11.809 (6.577)
Very Hard (t3s111o2ye2)	7.192* (3.332)	0.697*** (0.024)	20.126*** (5.325)
Initial Error: Low	0.123 (0.198)	0.00000 (0.001)	0.008 (0.046)
Initial Error: Medium	0.234 (0.236)	−0.00001 (0.001)	−0.003 (0.048)
Experiment Disclosure	2.075 (1.963)	−0.00004 (0.008)	0.980 (4.638)
Average Intelligence	−0.218*** (0.058)	−0.0004* (0.0002)	−0.123*** (0.021)
Intelligence SD	−0.012 (0.223)	−0.0003** (0.0001)	−0.102*** (0.020)
6 Rounds x Strategic	0.083 (0.294)	0.001 (0.001)	0.021 (0.093)
9 Rounds x Strategic	−0.182 (0.286)	0.001 (0.001)	−0.014 (0.084)
Medium Struct x Init. Error (Low)	0.657 (0.786)	−0.00000 (0.001)	0.003 (1.897)
Hard Struct x Init. Error (Low)	0.120 (0.474)	0.045 (0.097)	−8.149 (11.199)
Very Hard x Init. Error (Low)	1.665 (1.222)	0.008 (0.047)	9.638* (4.721)
Medium Struct x Init. Error (Med)	−0.399 (0.378)	0.00001 (0.001)	0.006 (0.957)
Hard Struct x Init. Error (Med)	0.187 (0.594)	−0.077 (0.130)	18.273 (11.427)
Very Hard x Init. Error (Med)	−0.321 (0.975)	−0.004 (0.024)	8.923 (6.687)
Disclosure x Av. Intelligence	−0.053 (0.061)	0.00000 (0.0002)	−0.023 (0.109)
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 upper 20% tail of the log error distribution (the worst-performing markets). OLS model uses CR2 cluster-robust standard errors clustered by exact AI team composition, featuring small-sample degrees of freedom adjustments. 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

	<i>Dependent variable:</i>	
	Trading Volume	Individual Profits
	(1)	(2)
Constant	1,757.638 (1,057.684)	−77.335* (30.791)
Comments Allowed	−36.269 (29.026)	−3.552 (2.747)
Duration: 6 Rounds	563.399*** (104.689)	3.359 (17.416)
Duration: 9 Rounds	1,080.922*** (209.860)	−12.727 (19.250)
Strategic Prompt	37.028 (21.233)	0.345 (2.338)
Medium (t3s110)	316.783* (156.405)	−35.852 (38.603)
Hard (t3s111)	212.977** (67.246)	−71.160* (32.793)
Very Hard (t3s111o2ye2)	43.866 (267.233)	−101.644* (45.614)
Initial Error: Low	−145.678* (66.395)	−12.168*** (3.042)
Initial Error: Medium	44.294 (56.922)	−5.193 (3.645)
Experiment Disclosure	171.986 (137.059)	−21.052 (14.316)
Average Intelligence	−28.182 (24.757)	−11.057*** (1.855)
Individual Intelligence		14.091*** (1.787)
Intelligence SD	24.293 (40.089)	1.154 (2.467)
Trades Second		−39.316 (32.790)
Trades Third		74.604 (53.060)
6 Rounds x Strategic	26.283 (34.055)	0.524 (3.928)
9 Rounds x Strategic	−45.471 (41.863)	5.343 (3.828)
Medium Struct x Initial Error (Low)	166.553 (110.724)	−9.165 (10.672)
Hard Struct x Initial Error (Low)	69.433 (64.362)	−6.278 (7.583)
Very Hard x Initial Error (Low)	349.973 (261.708)	−18.174 (15.738)
Medium Struct x Initial Error (Med)	−28.194 (77.289)	2.954 (4.152)
Hard Struct x Initial Error (Med)	93.975 (73.295)	2.619 (8.980)
Very Hard x Initial Error (Med)	−252.898 (441.683)	1.676 (11.363)
Disclosure x Average Intelligence	−3.394 (5.256)	0.493 (0.445)
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 CR2 cluster-robust standard errors clustered by exact AI team composition, featuring small-sample degrees of freedom adjustments. 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.

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. However, this null result may also be driven by the inability of LLMs to execute dynamic strategic plans (Kambhampati et al., 2024). In our experiment, a strategic agent could coordinate an inter-temporal plan by leaving instructions for its future self in the private comments. Alternatively, it could manipulate competitors by deceiving them in public comments, either by lying outright or by hoarding information (revealing less in public than in private).

To investigate these behavioral mechanics, we construct three measures of communication strategy. First, we compute the **cosine similarity** between an agent’s private and public messages, a standard natural language processing metric for semantic alignment ranging from 0 (divergent) to 1 (identical) (Manning and Schutze, 1999).¹⁶ Second, we measure **information hoarding** via the word gap (the number of words in the private minus the public message). Across all markets, private messages average 83.5 words while public messages average only 40, and private messages are strictly longer in 94% of markets (Figure 7). Finally, we measure explicit **deception**. Using an LLM-as-a-judge methodology, we task our most capable model in the first wave, Gemini 3 Flash, with evaluating in all rounds and markets whether the public comment truthfully reveals the agent’s private signal (Truthful Revelation), avoids disclosing the signal (Information Withholding), or states the opposite of the true signal (Lying).¹⁷ We estimate an ordered logit model across these three hierarchical categories, where positive coefficients indicate a shift toward withholding or lying.

Table 10 presents the pooled models for cosine similarity (OLS, Column 1), information hoarding (OLS, Column 2), and deception (Ordered Logit, Column 3). In Tables 12, 13, and 14, we report the same regressions when splitting the sample by market duration (3, 6, and 9 rounds).

The strategic prompt has a statistically insignificant effect on both cosine similarity and only marginally reduces information hoarding by less than one word ($\beta = -0.788, p < 0.1$). As demonstrated in Tables 12, 13, and 14, even this weak, marginal effect on the word gap fails to survive consistently when splitting the sample by market duration (3, 6, 9 rounds), except for the information hoarding case when the duration is 9 rounds. This confirms a strictly null behavioral response: AI agents either cannot, or choose not to, execute complex deceptive strategies when explicitly prompted to be strategic. Individual intelligence similarly provides mixed results, correlating with less explicit lying but marginally more information hoarding. Moreover, information hoarding and deception increase as the structure becomes more complex.

Strikingly, the Experiment Disclosure treatment triggers a significant response. Disclosing the past history induces agents to heavily increase the gap between private and public messages (+14.8 words, $p < 0.001$) and deceive more (+0.86, $p < 0.001$). Recall that one of

¹⁶Each 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).

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

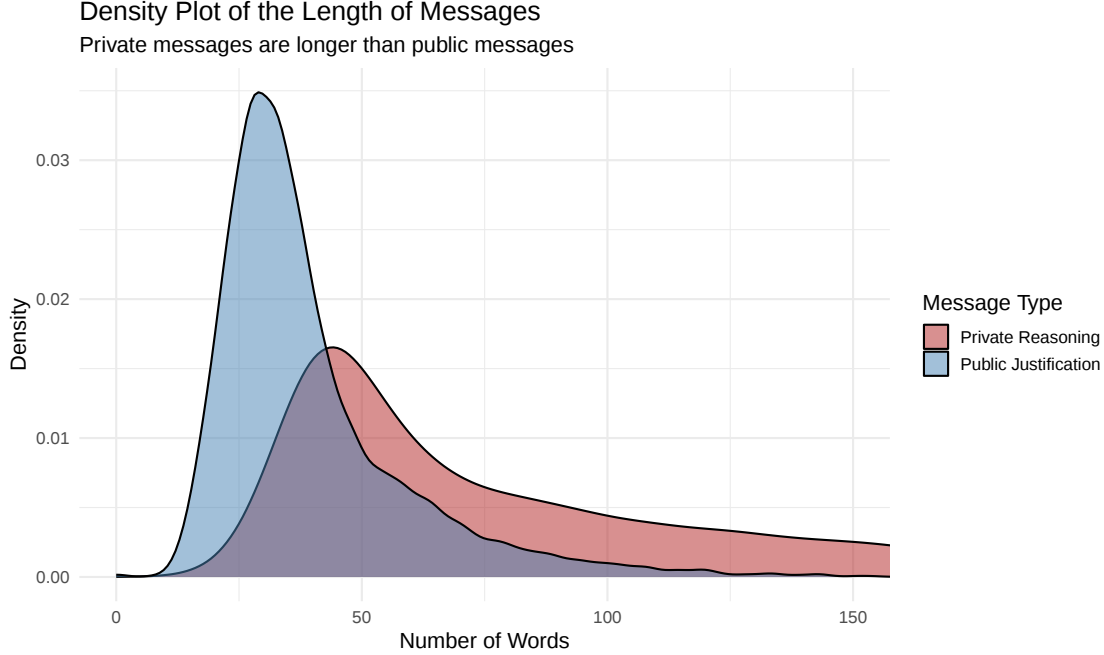


Figure 7: Length of Private and Public Messages

the qualitative results we disclosed was that public comments have no effect on information aggregation.

Finally, by modelling the current round as a categorical variable, the regression uncovers a pattern which is consistent with end-of-horizon disclosure, but also with simpler round-number heuristics. Rather than exhibiting a simple linear decay in deception and information hoarding over time, there is a “sawtooth” pattern. 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 ($p < 0.001$). 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. However, when we split the models by duration of 3, 6, and 9 rounds (Tables 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 or 6, even though it ends in round 9, suggesting that it may not fully grasp the dynamic horizon of the market.

6.5 Robustness with Frontier AI Models

The rapid pace of innovation in AI is evidenced by the frequent release of new models every few months. To test whether the failure of information aggregation in complex environments (Result 1) is a transient artifact of older AI capabilities, we conducted in April 2026 an out-

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

	<i>Dependent variable:</i>		
	Cosine Similarity (OLS)	Word Gap (OLS)	Deception (Ordered Logit)
	(1)	(2)	(3)
Constant	0.324*** (0.043)	33.880* (15.423)	
Strategic Prompt	0.003 (0.002)	−0.788 (0.416)	−0.006 (0.036)
Round 2	0.030* (0.014)	−13.194 (7.874)	−0.398*** (0.074)
Round 3	0.004 (0.013)	−25.105* (9.969)	−0.665*** (0.076)
Round 4	0.030* (0.012)	−12.247** (3.723)	−0.205** (0.079)
Round 5	0.040** (0.015)	−23.619** (7.966)	−0.160 (0.082)
Round 6	0.005 (0.016)	−30.972** (9.860)	−0.560*** (0.091)
Round 7	0.038 (0.020)	−18.893*** (5.068)	−0.201* (0.101)
Round 8	0.047** (0.015)	−26.708** (8.346)	−0.081 (0.099)
Round 9	0.033 (0.019)	−32.644*** (9.907)	−0.487*** (0.112)
Duration: 6 Rounds	0.002 (0.005)	−0.468 (0.906)	−0.036 (0.093)
Duration: 9 Rounds	−0.002 (0.005)	−1.072 (0.866)	−0.017 (0.094)
Medium (t3s110)	−0.022** (0.007)	−1.213 (1.063)	1.198*** (0.108)
Hard (t3s111)	0.002 (0.006)	2.417* (1.099)	0.907*** (0.124)
Very Hard (t3s111o2ye2)	0.001 (0.008)	5.666*** (1.423)	1.369*** (0.103)
Experiment Disclosure	0.010 (0.018)	14.858*** (3.174)	0.861*** (0.074)
Initial Error	0.002 (0.012)	2.867 (2.227)	0.268 (0.208)
Individual Intelligence	0.001 (0.002)	0.844 (0.484)	−0.004 (0.002)
Observations	10,620	10,620	13,189
R ²	0.027	0.156	

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. Columns 1 and 2 (OLS) report CR2 cluster-robust standard errors in parentheses, clustered by exact AI team composition with small-sample degrees of freedom adjustments. Column 3 (Ordered Logit) reports CR1 cluster-robust standard errors clustered at the market level to account for intra-market correlation across competing agents and repeated communication rounds.

of-sample robustness check using a cohort of frontier LLMs. This subsample comprises 576 markets played across all four structures (Easy, Medium, Hard, and Very Hard). We employ four distinct teams: three homogeneous markets populated by GPT-5.4, Claude Opus 4.6, and Gemini 3.1 Pro, and a heterogeneous team containing one agent of each model. We benchmark these four frontier teams against the top four teams from our original January 2026 baseline.

Because the vintage treatment (April versus January) varies strictly at the team level, and because our sub-sample contains a small number of clusters ($G=8$), standard cluster-robust inference or Wild Cluster Bootstrapping can suffer from severe power limitations and coarse p-values. Therefore, following standard econometric recommendations for inference with a small number of treated groups (Bertrand et al., 2004; Donald and Lang, 2007), we adopt a highly conservative aggregation approach. We collapse the repeated market-level observations by calculating the mean logarithmic error for each of the eight teams. This absorbs all arbitrary intra-cluster correlation, resulting in a purely cross-sectional dataset ($N=8$) upon which we can conduct exact small-sample statistical tests.¹⁸

We conduct a Welch’s two-sample t -test (which does not assume equal variances) to compare the aggregated market efficiency of the two cohorts. We find no evidence that the newer models improve information aggregation. In fact, the point estimate indicates that the April 2026 frontier models generated a higher mean logarithmic error ($\mu_{April} = 2.30$) than the top January models ($\mu_{Jan} = 1.31$). Due to the high variance across the models, however, this apparent degradation is statistically insignificant ($t=0.707$, $p=0.523$). Figures 10 and 11 display the mean squared errors across all four structures for the baseline and frontier models, respectively.

These results provide strong empirical validation for our primary specification. Scaling up to the April capability frontier does not systematically resolve the bounded rationality or coordination frictions observed in the prediction markets. Consequently, we conclude that our findings are structurally robust and not merely an artifact of the specific LLM vintage.

We next examine whether the failure to aggregate information in the Very Hard structure differs across LLMs. Table 11 reports the results of a Wilcoxon signed-rank tests evaluating whether the error distributions for each team are statistically distinguishable from a mean absolute error of 0.5, or random guessing.¹⁹ We define three classes: hedging, hallucination and convergence. Hedging means that the p-value fails to reject the null hypothesis of random guessing. Hallucination occurs when the null hypothesis is rejected but with a larger absolute error, whereas convergence happens when the rejection is with a smaller absolute error.

We first observe that around 70% of the teams price the security at 0.5, including all frontier teams. Second, more than half of all teams are hedging. Third, the classes are preserved within models of the same company, suggesting architectural persistence across the January and April cohorts. Models from OpenAI (5.4, 5 Mini, and 4o) and Anthropic (Opus 4.6 and Haiku 4.5) maintain hedging. In contrast, Gemini 3 Flash and 3.1 Pro exhibit

¹⁸We exclude the information provision treatment markets in the January cohort, as the April cohort did not receive such a treatment.

¹⁹Recall that the median market priced the security at 0.5, when the true value is 0 (Table 6, Q50).

convergence, although the less advanced Gemini 2.5 Flash is hedging.

Table 11: Behavioral Failure Modes Across AI Cohorts (Very Hard Structure)

Team Composition	N	Median Error	Mean Error	p-value	Result
Panel A: April 2026 Frontier Cohort					
Gemini 3.1 Pro (x3)	36	0.5	0.2878	0.000	Convergence
GPT 5.4 Pro (x3)	36	0.5	0.4739	0.371	Pure Hedging
Opus 4.6 (x3)	36	0.5	0.5206	0.287	Pure Hedging
Opus 4.6/Gemini 3.1 Pro/GPT 5.4 Pro	36	0.5	0.6122	0.032	Hallucination
Panel B: January 2026 Baseline Cohort					
Gemini 3 Flash/GPT-4o/Qwen 3	36	0.165	0.3806	0.039	Convergence
Gemini 3 Flash (x3)	36	0.255	0.2506	0.000	Convergence
Haiku 3.5/Gemma 3/Qwen 3	36	0.49	0.5992	0.349	Pure Hedging
Gemini 2.5 Flash (x3)	36	0.5	0.4722	0.424	Pure Hedging
Haiku 4.5 (x3)	36	0.5	0.4825	0.725	Pure Hedging
GPT-5 Mini (x3)	36	0.5	0.4939	0.932	Pure Hedging
Haiku 4.5/Gemini 3 Flash/GPT-5 Mini	36	0.5	0.5122	1.000	Pure Hedging
Qwen 3 (x3)	37	0.5	0.5195	0.686	Pure Hedging
GPT-4o (x3)	36	0.59	0.5556	0.297	Pure Hedging
Haiku 3.5/Gemini 2.5 Flash/GPT-4o	36	0.76	0.6867	0.004	Hallucination
Gemma 3 (x3)	36	1	0.8636	0.000	Hallucination
Haiku 3.5 (x3)	36	1	0.9972	0.000	Hallucination

Two-sided Wilcoxon signed-rank test against a completely uninformative price ($H_0 := 0.5$). We report median and mean absolute errors. Results document the persistence of cognitive failure modes across both the baseline and frontier capability cohorts.

How can we explain the inability of state-of-the-art models to solve the Very Hard structure, and that in some cases they perform worse than earlier models? On the one hand, newer models should perform better because they have bigger context windows, so they can process more information, and they are more competent at mathematical operations, such as Bayesian updating. On the other hand, they are more aggressively trained using Reinforcement Learning from Human Feedback (RLHF). This technique trains LLMs to choose answers that would be graded highly by humans. Although RLHF significantly improves the responses in many tasks (Ouyang et al., 2022), a growing literature documents that RLHF leads to unintended behavioral characteristics that may worsen their performance in other tasks.

Kadavath et al. (2022) show that RLHF makes LLMs significantly worse at accurately reporting the probability that they know the answer to a question, if the task is new to them. In many cases they are under-confident. Xiong et al. (2023) show that fine-tuned LLMs tend to be overconfident when verbalizing their confidence, potentially imitating human patterns when expressing confidence. Tao et al. (2025) create a large-scale dataset of hedging

expressions with human-annotated confidence scores, and show that most modern LLMs underperform when expressing their uncertainty through hedging language.

These results suggest AI agents might struggle to articulate their uncertainty and accurately communicate the probability they assign to each signal being true. This is important not only for their publicly accessible messages but also for private messages intended solely for their future selves. Furthermore, they must interpret and quantify the messages of others into probabilities regarding signals. Consistent application of RLHF could explain why models from the same company perform similarly in information aggregation even though one is significantly more advanced in other tasks.

6.6 Robustness to Alternative Inference Specifications

To ensure that our main findings are driven by underlying behavioral realities rather than the conservative variance penalties of our chosen econometric framework, we re-estimated our global specifications utilizing alternative standard error structures. The corresponding robustness tables are reported in the Supplementary Appendix. For market-level outcomes, we relaxed the inference model to unclustered HC1 standard errors. For individual-level outcomes, we applied CR1 cluster-robust standard errors clustered at the unique market level, capturing local intra-market dependencies while removing the broader panel-level team penalty.

Crucially, the primary null effects observed for communication, strategic prompting, initial price, and extended market duration (9 rounds) remain statistically insignificant even under these less restrictive inference specifications. Conversely, when the team-level clustering penalty is removed, the 6-round duration, experiment disclosure, and the Medium information structure cross the threshold into statistical significance in the information aggregation OLS. Similarly, the order of trading demonstrates a statistically significant effect on individual profits under market-level clustering. Because these alternative specifications assume independence across separate market sessions, ignoring the broader intra-team correlation inherent in our multi-round, agent-based panel design, they naturally produce smaller standard errors. The fact that these specific variables achieve significance under localized or unclustered errors, but are appropriately attenuated under our primary CR2 specification, confirms that their effects are highly marginal and heavily sensitive to intra-team behavioral persistence.

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. While it is tempting to anthropomorphize highly capable AI models and assume they inherently possess these higher-order interactive reasoning skills, empirical evidence remains scarce.

To our knowledge, this paper provides the first experimental evidence on information aggregation and interactive reasoning among autonomous AI agents within a prediction market framework. Consistent with economic theory, we find that cheap talk communication, initial price, market duration, and strategic prompting have no significant effect on market accuracy. However, increasing the complexity of the information and payoff structure strictly degrades it. While the median AI team successfully solves the easy and medium structures, they completely fail to aggregate information in the hard and very hard environments. This suggests that current LLMs do not yet possess hyper-rational interactive reasoning capabilities.

Within our main sample, the impact of the intelligence index in our main experiment is significant and positive: higher average market intelligence clearly enhances information aggregation. However, a robustness check comparing the top four teams from the January cohort to four state-of-the-art teams from the April cohort reveals a stagnation at the frontier, yielding no subsequent improvement in accuracy. A plausible explanation is that the newer cohort underwent more intensive training using Reinforcement Learning from Human Feedback (RLHF). This alignment training may inadvertently introduce human-like cognitive biases or hinder the models’ ability to effectively communicate uncertainty. Interestingly, LLMs across cohorts but from the same parent company perform similarly, suggesting persistence on how RLHF is applied.

At the individual level, we find that an agent’s profitability is dictated by relative, rather than absolute, cognitive advantage. While higher individual intelligence yields financial gains, an increase in the average intelligence of the team strictly erodes them. Furthermore, the order of trading matters, as trading third is strictly more profitable than trading second, confirming a structural late-mover advantage.

Crucially, our experiments reveal limitations in how LLMs process feedback from past play. Contrary to recent literature suggesting LLMs can self-improve via iterative prompting, we find that providing agents with information about past play fails to improve market accuracy. Rather than optimizing their pricing strategies, the main effect is for agents to actively hoard information and increase deception.

Finally, text analysis of the agents’ private and public messages shows that they struggle to act strategically and they are unable to design and execute a dynamic plan. They reveal more information as the game approaches the end, indicating that they have some understanding about its dynamic nature. However, they also exhibit a “sawtooth” pattern of revelation that spikes in rounds 3, 6, and 9, even when the game ends in 9 rounds, suggesting that they may rely on token-matching heuristics instead of backward induction reasoning.

Ultimately, our findings suggest a nuanced frontier for trading with AI agents. While they demonstrate competence in basic market mechanics and scale predictably with model intelligence, their inability to act strategically and autonomously execute multi-step interactive reasoning indicates a significant distance until they become the sophisticated agents

usually assumed in economic theory.

A Appendix

Table 12: Robustness Check: Cosine Similarity by Market Duration

	<i>Dependent variable:</i>		
	Cosine Similarity (Public, Private)		
	3-Round	6-Round	9-Round
	(1)	(2)	(3)
Constant	0.334*** (0.057)	0.314*** (0.044)	0.333*** (0.042)
Strategic Prompt	−0.007 (0.008)	−0.006 (0.007)	−0.006 (0.007)
Round 2	0.025 (0.014)	0.033* (0.015)	0.032* (0.013)
Round 3	0.005 (0.016)	0.005 (0.014)	0.003 (0.014)
Round 4		0.029 (0.015)	0.032** (0.011)
Round 5		0.042** (0.016)	0.039** (0.014)
Round 6		0.009 (0.017)	0.003 (0.018)
Round 7			0.039* (0.018)
Round 8			0.047*** (0.013)
Round 9			0.032 (0.019)
Medium (t3s110)	−0.017 (0.014)	−0.015 (0.010)	−0.029*** (0.008)
Hard (t3s111)	−0.006 (0.010)	0.012 (0.010)	−0.001 (0.007)
Very Hard (t3s111o2ye2)	0.001 (0.015)	0.005 (0.010)	−0.001 (0.007)
Experiment Disclosure	0.004 (0.019)	0.011 (0.019)	0.011 (0.018)
Initial Error	0.009 (0.026)	0.001 (0.016)	0.001 (0.012)
Individual Intelligence	0.001 (0.002)	0.001 (0.002)	0.001 (0.002)
Observations	1,776	3,552	5,292
R ²	0.015	0.034	0.029

Notes: ·p<0.1; *p<0.05; **p<0.01; ***p<0.001. CR2 cluster-robust standard errors, clustered by AI team composition, 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

	<i>Dependent variable:</i>		
	Word Gap		
	3-Round Markets	6-Round Markets	9-Round Markets
	(1)	(2)	(3)
Constant	30.445 (17.813)	35.355* (15.763)	31.111* (14.134)
Strategic Prompt	1.365 (1.192)	0.227 (0.960)	2.564* (1.290)
Round 2	−13.385 (8.313)	−12.328 (8.341)	−13.802 (7.461)
Round 3	−24.149* (10.091)	−25.950* (10.454)	−24.945* (9.743)
Round 4		−12.730** (3.883)	−11.833** (4.002)
Round 5		−23.211** (8.601)	−24.144** (7.531)
Round 6		−31.712** (10.433)	−30.467** (9.379)
Round 7			−18.861*** (5.144)
Round 8			−26.700*** (8.090)
Round 9			−32.712*** (9.620)
Medium (t3s110)	−2.959 (2.270)	−3.839* (1.759)	1.157 (1.207)
Hard (t3s111)	0.958 (2.769)	−0.100 (1.270)	4.582** (1.554)
Very Hard (t3s111o2ye2)	3.865 (2.826)	4.459*** (1.302)	7.073*** (1.788)
Experiment Disclosure	12.640*** (2.191)	15.753*** (3.654)	14.981*** (3.707)
Initial Error	2.330 (2.853)	2.005 (4.383)	3.618* (1.491)
Individual Intelligence	1.050 (0.578)	0.813 (0.486)	0.796 (0.452)
Observations	1,776	3,552	5,292
R ²	0.143	0.152	0.159

Notes: $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. CR2 cluster-robust standard errors, clustered by AI team composition, 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

	<i>Dependent variable:</i>		
	3-Round	Deception 6-Round	9-Round
	(1)	(2)	(3)
Strategic Prompt	−0.078 (0.133)	0.012 (0.113)	0.039 (0.113)
Round 2	−0.426*** (0.126)	−0.364** (0.130)	−0.388** (0.126)
Round 3	−0.628*** (0.130)	−0.696*** (0.134)	−0.645*** (0.131)
Round 4		−0.278* (0.120)	−0.114 (0.115)
Round 5		−0.145 (0.125)	−0.161 (0.124)
Round 6		−0.510*** (0.137)	−0.615*** (0.140)
Round 7			−0.183 (0.125)
Round 8			−0.062 (0.124)
Round 9			−0.482*** (0.140)
Medium (t3s110)	1.703*** (0.242)	0.917*** (0.172)	1.270*** (0.163)
Hard (t3s111)	1.650*** (0.268)	0.693*** (0.199)	0.853*** (0.190)
Very Hard (t3s111o2ye2)	1.994*** (0.230)	1.093*** (0.160)	1.401*** (0.157)
Experiment Disclosure	0.693*** (0.139)	0.790*** (0.117)	0.971*** (0.118)
Initial Error	0.552 (0.401)	0.113 (0.323)	0.284 (0.332)
Individual Intelligence	0.013** (0.004)	−0.009* (0.004)	−0.007* (0.004)
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, ordered as Truth < Withhold < Lie. Positive coefficients indicate a shift toward higher propensities of withholding or lying. All specifications employ a mixed-contrast framework: treatment contrasts are used for Round, Structure, and Experiment Disclosure, with Round 1, Easy (t3s111y2), and No Disclosure serving as the respective omitted baselines. The Strategic Prompt indicator utilizes sum contrasts (representing deviations from the grand mean), while the Initial Error continuous variable is mean-centered. Standard errors reported in parentheses are CR1 cluster-robust standard errors clustered at the unique market level ('slug') to account for intra-market correlation across competing agents and repeated communication rounds.

B 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 with $n = |I|$. 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 $\text{Supp}(\mu)$ 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 $\omega \in \text{Supp}(\mu)$,
- (ii) $E_\mu[X|\Pi_i(\omega)] = v$ for all $i = 1, \dots, n$ and $\omega \in \text{Supp}(\mu)$.

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 $\lambda_i : \Pi_i \rightarrow \mathbb{R}$ for $i = 1, \dots, n$ 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.

C 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 η 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

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

The cost of buying these Yes shares is

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

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:

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

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

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:

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

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:

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

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:

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

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

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

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

D Supplementary 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.
- 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.
- Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

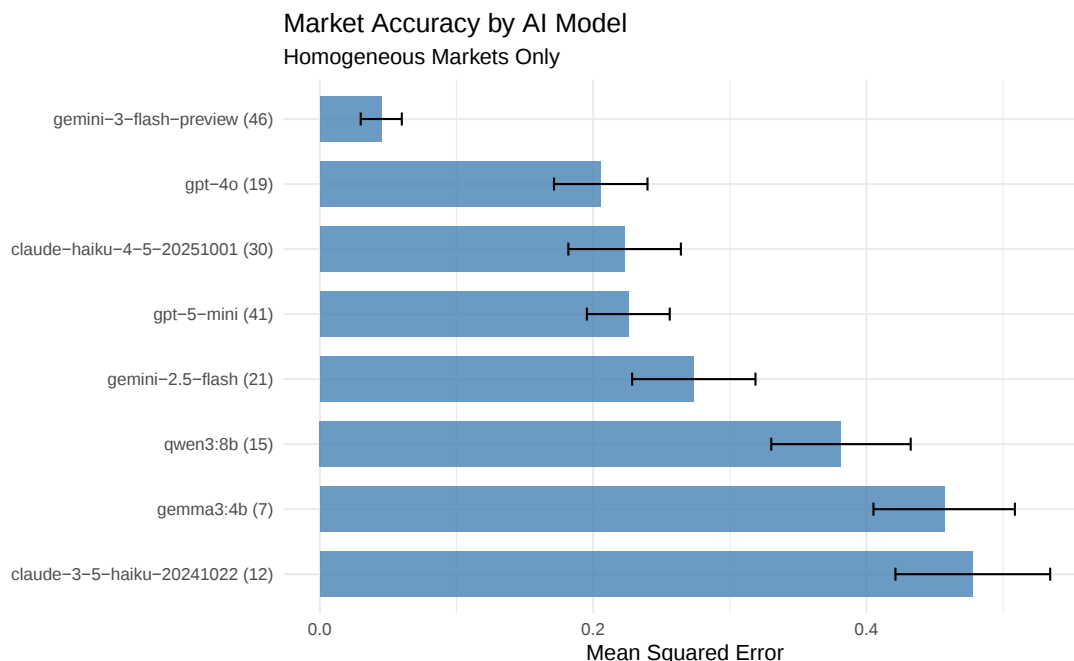


Figure 8: Mean Squared Error by AI Model

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.

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.

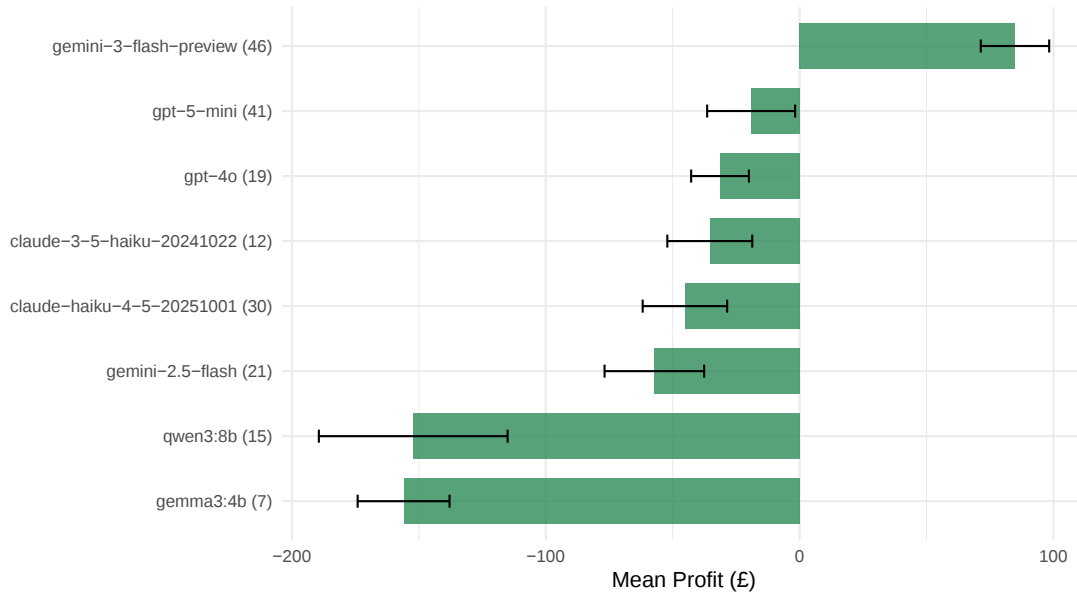


Figure 9: Average Profits by AI Model

- 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.
- Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233.
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., et al. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in neural information processing systems*, 36:70293–70332.
- 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.

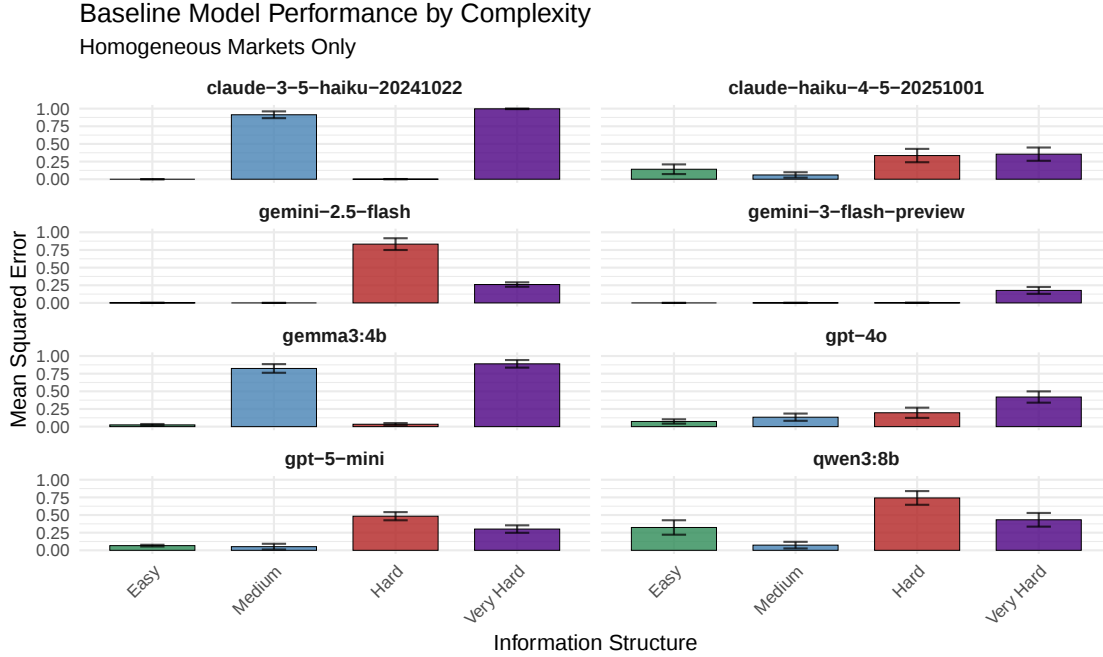


Figure 10: Mean Squared Error by Structure and Model

- Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119.
- 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*.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- 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*.

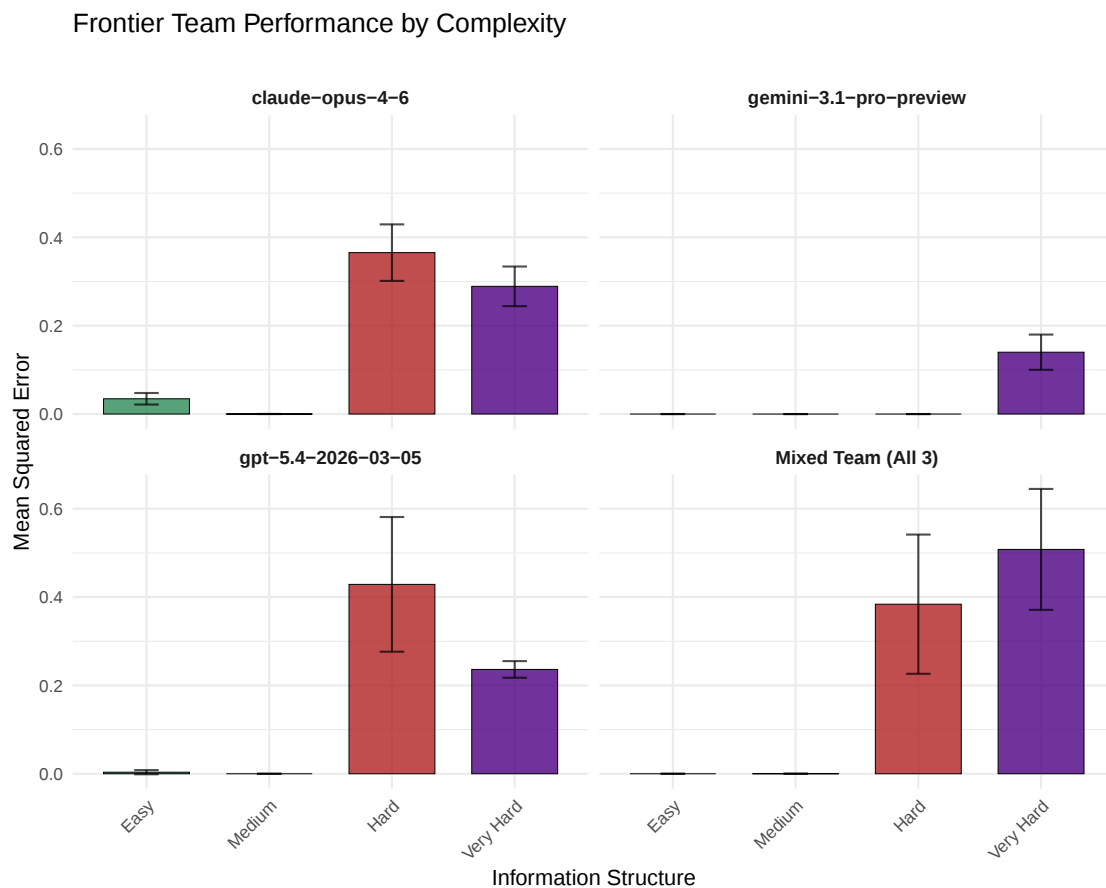


Figure 11: Mean Squared Error by Structure and Frontier Model

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.

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.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- 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*.
- Pustejovsky, J. E. and Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683.
- 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*.
- 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*.
- Tao, L., Yeh, Y.-F., Kai, B., Dong, M., Huang, T., Lamb, T. A., Yu, J., Torr, P. H., and Xu, C. (2025). Can large language models express uncertainty like human? *arXiv preprint arXiv:2509.24202*.
- 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.

- Valmeekam, K., Marquez, M., Sreedharan, S., and Kambhampati, S. (2023). On the planning abilities of large language models-a critical investigation. *Advances in neural information processing systems*, 36:75993–76005.
- 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.
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. (2023). Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.
- 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*.