

Autonomous LLM Agents Trading in Decentralized Finance Markets

Wan Chu Cheong

London School of Economics and Political Science

August 2026

Abstract

This study examines how the introduction of autonomous large language model (LLM) agents changes the processing of corporate disclosures in financial markets. Using transaction-level data from financial prediction markets on Polymarket, where investors deploy autonomous agents to trade with real money, I investigate whether these agents expand market participants' information-processing capacity and how their trading affects market price discovery. I find that agent wallets make more accurate trading decisions than human wallets within the same market, and their relative accuracy improves following a major LLM model upgrade. The relative accuracy advantage varies systematically with the information being processed. It is greater when annual filings contain more tokens and longer sentences, indicating an advantage in processing larger volumes of textual information. The advantage is also stronger when year-over-year disclosure changes contain more litigious or forward-looking language, consistent with stronger semantic reasoning over qualitative financial information. In contrast, the advantage weakens when disclosures become more numerically intensive and when disclosure changes contain greater uncertainty or vagueness. Overall, agent order imbalance is associated with lower market price error, indicating that information processed by agents contributes to price discovery. These findings suggest that autonomous agents expand the capacity to process corporate disclosures and improve price formation, while their benefits are more limited when information requires greater judgment under ambiguity.

Keywords: information processing costs, behavioral bias, LLM agents

I appreciate the helpful comments and suggestions from Stefano Cascino, Alexandra Scherf, Xi Li, and Ane Tamayo. All errors are my own.

1. Introduction

Large language models (LLMs) have primarily been studied as tools that help investors extract and interpret information from unstructured financial text, including corporate disclosures and earnings-related narratives (van Binsbergen et al. 2023; Ecker et al. 2025). Recent advances fundamentally change this role. Autonomous LLM agents can continuously gather information, form predictions, and execute trades without direct human intervention (Turtel et al. 2025; Schoenegger et al. 2024). Their entry into financial markets therefore introduces a new class of capital market participant. Because processing corporate disclosures requires scarce attention and effort, market participants do not fully process all publicly available information (see Blankespoor et al. 2020 for a review). Autonomous agents can instead monitor disclosures continuously and process large quantities of firm-specific information at low marginal cost, potentially increasing the amount of public information reflected in trading decisions and market prices.

Whether this expanded processing capacity improves market outcomes, however, depends on the quality of the information agents extract. Autonomous agents may hallucinate, misinterpret information, or make reasoning errors that are translated directly into market positions without human review.¹ If multiple agents rely on similar models or information-processing strategies, these errors may also become correlated (Financial Stability Oversight Council 2023). I therefore examine how the participation of autonomous LLM agents changes the incorporation of public information into market prices. More specifically, I examine whether agents process corporate disclosure information that may otherwise receive limited attention from human traders, where the

¹ Cao et al. (2026) emphasize that reductions in information-processing costs do not eliminate the need for verification. AI-generated outputs may contain factual errors, hidden reasoning errors, or contextually inappropriate interpretations that can be difficult for users to detect. Verification therefore remains an important cost of using GenAI, particularly in judgment-intensive settings, where human expertise may complement AI by evaluating low-confidence outputs and identifying errors that require contextual judgment.

boundaries of this expanded processing capacity arise, and whether information reflected in agent trading contributes to price discovery.

I study these questions using firm-level financial prediction markets on Polymarket, where autonomous agents and human participants trade alongside one another before outcomes are known.² Prediction markets have traditionally aggregated information held by human participants, but autonomous systems are increasingly becoming another source of information reflected in market prices (Canny, 2026). Polymarket is well suited to this study because its technological infrastructure allows autonomous LLM agents to trade directly alongside human participants with real capital and without human review. In this setting, an LLM agent is an autonomous trading system that combines a language model with tools for information retrieval, market monitoring, and trade execution. The agent can collect public information, including corporate disclosures, news, and market prices, and use the LLM to interpret that information and assess the likelihood of a market outcome. It can then compare this assessment with the probability implied by market prices and execute trades directly through Polymarket’s application programming interface (API).

I focus on financial prediction markets tied to firm-level stock price outcomes, where binary contracts pay based on whether a firm’s stock price exceeds a specified threshold by a given date. Corporate disclosures contain information useful for forecasting future firm performance and stock

² In a prediction market, participants buy and sell contracts whose payoffs are tied to the outcome of a future event, and the contract price reflects the market's collective probability assessment; such markets aggregate dispersed information through prices and often produce forecasts that rival or exceed those of expert forecasters and polls (Wolfers and Zitzewitz, 2004; Arrow et al., 2008). On Polymarket, a participant who expects the outcome to occur buys a "Yes" contract and one who expects it not to occur buys a "No" contract; at resolution the winning contract pays one dollar and the losing contract pays zero. For example, a trader who buys a Yes contract at \$0.60 earns \$0.40 if the outcome occurs and loses the \$0.60 paid if it does not.

prices. These markets therefore provide a setting in which disclosure processing can directly inform trading decisions and price formation.³

I first show that wallets classified as LLM agents make trading decisions that are more closely aligned with eventual market outcomes. Relative to human wallets trading in the same markets, agent wallets are 12.1 percentage points more likely to finish with a net position on the side that ultimately proves correct. At the individual-trade level, agent wallets are also 7.7 percentage points more likely to place a trade in the direction of the outcome that ultimately occurs. This accuracy advantage also translates into economic gains: agent wallets earn higher returns on investment than human wallets within the same markets.

I then examine whether agent performance responds to advances in the underlying LLM technology. Around the November 13, 2025 release of GPT-5.1 on the OpenAI API, agent accuracy increases significantly relative to human-wallet accuracy. In the specification with wallet, market, and date fixed effects, the relative increase is 8.6 percentage points. This result provides complementary evidence that improvements in available LLM capabilities are associated with better autonomous trading decisions.

Prior research shows that investors incur substantial costs in acquiring and processing large volumes of corporate disclosure, which can limit the extent to which available information is incorporated into their decisions (Dyer et al., 2017; Blankespoor et al., 2020; Cohen et al., 2020). Autonomous LLM agents may alleviate this constraint by processing large quantities of textual information programmatically and at relatively low marginal cost. I therefore examine whether the relative accuracy associated with agent trading increases with the amount of disclosure available

³ Polymarket also hosts prediction markets on politics, sports, weather, and other nonfinancial events. Some of these markets more closely resemble traditional wagering settings and attract traders who derive utility from gambling or entertainment itself, rather than solely from trading on information (Thaler and Ziemba 1988; Dorn and Sengmueller 2009).

to process, measured by the number of tokens in the filing. The prediction is not mechanical. Longer disclosures increase the amount of information that can potentially be processed, but they also impose greater demands on LLM context processing, and prior research shows that model performance can deteriorate as relevant information becomes more difficult to retrieve and integrate from long inputs (Liu et al., 2024). I find that the relative accuracy of agent trading increases with annual-filing token length and with tokens per sentence. These findings suggest that, within the range of disclosures in my setting, autonomous agents can alleviate processing constraints associated with large volumes of textual information, potentially complementing the limited attention and processing capacity of human market participants.

Having established that autonomous agents appear to alleviate constraints associated with the volume of information to be processed, I next examine whether these benefits depend on the nature of the information itself. Following Cohen et al. (2020), I focus on year-over-year changes in disclosure text because additions and deletions are more likely to capture newly disclosed information. I find that the relative accuracy of agent trading is greater when year-over-year disclosure changes contain more litigious or forward-looking language, consistent with autonomous agents being particularly useful for processing qualitative textual information. In contrast, the relative advantage declines as annual filings become more numerically intensive, consistent with documented limitations of LLMs in numerical reasoning (Levy, 2026). It also weakens when disclosure changes contain greater uncertainty or vagueness and is substantially smaller in markets involving M&A rumors, where relevant information is less precise and traders must assess unverified future developments. Taken together, these findings suggest that the benefits of autonomous information processing are greatest when the primary friction is the cost

of processing textual information, but are more limited when the bottleneck shifts from processing capacity to numerical reasoning or judgment under ambiguity.

I further investigate whether the agent accuracy advantage is associated with lower exposure to behavioral bias. In binary prediction markets, low-priced contracts correspond to outcomes viewed as relatively unlikely and offer large state-contingent payoffs relative to their purchase prices. Prior research documents that traders tend to disproportionately favor such low-priced contracts, a pattern commonly referred to as the longshot bias (Thaler and Ziemba, 1988; Snowberg and Wolfers, 2010; Wolfers and Zitzewitz, 2004). I find that agent wallets allocate a significantly smaller share of their trades to low-priced contracts across several price thresholds. This evidence suggests that autonomous agents are less exposed to longshot trading behavior than human traders.

Finally, I examine whether autonomous agent trading contributes to market price discovery. I construct separate order imbalance measures for agent and human wallets. I find that stronger agent order imbalance is associated with lower subsequent market price error, after controlling for human order imbalance in the same market and time window. This evidence moves beyond individual trading accuracy and suggests that autonomous agents contribute to the incorporation of information into market prices.

My study contributes to the accounting literature on LLMs and financial information processing by studying autonomous LLM agents as a new class of market participant that independently processes corporate disclosures and trades on the resulting information. This paper shifts the focus from human-in-the-loop AI assistance to autonomous delegation. Whereas existing applications generally leave investors responsible for evaluating AI output and deciding whether to act (Ecker et al., 2026; Moss et al., 2026), the agents in my setting use their own information

assessments to make and execute trading decisions. This setting allows me to examine whether the gains from scalable information processing persist when decision making is delegated to autonomous agents without contemporaneous human verification. The findings suggest that autonomous agents can do more than replicate information processing already undertaken by human investors. By lowering the marginal cost of monitoring and analyzing public information, agents bring into trading information that would otherwise receive limited human attention.

The Polymarket setting also helps address an important empirical challenge in this literature. Because many prior studies evaluate model outputs using retrospective prediction exercises, their findings may be vulnerable to look-ahead bias or other forms of information leakage (Levy, 2026; Lopez-Lira and Tang, 2026; van Binsbergen et al., 2023; Zhang et al., 2025).⁴ In my setting, LLM agents and human traders compete in real time in the same unresolved markets. Trades are timestamped and executed before outcomes are known, and performance is measured using the realized payoffs of pre-resolution positions. This design therefore allows me to examine LLM trading risks without relying on retrospectively generated predictions, thereby reducing concerns about look-ahead bias.

Second, the study contributes to the literature on investor protection and less sophisticated investors. Prior research shows that individual investors face information-processing constraints and behavioral biases (Hirshleifer and Teoh, 2003; Barber and Odean, 2008; Blankespoor et al., 2020). Delegating investment decisions to autonomous LLM agents may mitigate some of these frictions, but the benefits are not uniform across disclosure settings. As LLM-based investment

⁴ Retrospective prediction exercises require researchers to reconstruct the information set that would have been available at the time of prediction. This reconstruction is difficult because financial databases often contain data that are updated, restated, revised, or observed only after the prediction date. Look-ahead bias arises when information that was not actually available at the time of prediction inadvertently enters the model input, benchmark construction, or evaluation procedure. For example, in an earnings-forecasting setting, a model intended to predict two-year-ahead earnings would be contaminated if it used one-year-ahead realized earnings as an input, because that variable would not have been known when the forecast was made (Zhang et al., 2025).

tools become increasingly accessible to individual investors (Ecker et al., 2025), this trade-off has implications for investor protection and financial regulation.

2. Institutional Background, Related Literature and Hypothesis Development

2.1. Institutional Background

2.1.1. Financial prediction markets

Polymarket is a large blockchain-based prediction market platform where participants trade contracts tied to future events. This study focuses on firm-level financial prediction markets that ask whether a publicly traded firm's stock price will exceed a specified threshold by a given date. Each market has two contracts corresponding to the outcomes Yes and No. A trader buying the Yes contract is betting that the firm's stock price will exceed the specified threshold, whereas a trader buying the No contract is betting that it will not. Contract prices range from 0 to 1 and represent the market-implied probability of the corresponding outcome. At settlement, the contract associated with the realized outcome pays \$1, while the other pays zero. Traders may hold their contracts until settlement or unwind their positions before the market is resolved.

2.1.2. LLM Agent Frameworks and Adoption

Polymarket is particularly useful for studying autonomous LLM agents because it allows computer programs to access market information and place trades directly through its application programming interface (API). Human traders typically place trades through Polymarket's web interface. In contrast, LLM agents can retrieve market information and execute trades directly through the API without human intervention.

An LLM trading agent on Polymarket is more than a language model that generates text. Publicly available agent frameworks and developer documentation describe systems that use

LLMs to collect and interpret market data and public information and to interact automatically with external search tools and trading platforms (Polymarket, 2026; OpenAI, 2026; Ferrag et al., 2026; Xia et al., 2026). Recent activity on Polymarket illustrates the broader information-to-trading process enabled by such automated systems. Industry reports describe AI-powered trading strategies that use “ensemble probability models trained on news and social data” to estimate real-world event probabilities and identify contracts that appear undervalued relative to those estimates (Okoth, 2026). In a firm-level prediction market, an agent can scan available markets and process information such as contract prices, recent news, earnings announcements, and other corporate disclosures. Based on this information, the agent predicts whether a firm’s share price is likely to exceed the threshold specified in a market. When the agent predicts a Yes outcome but considers the prevailing price of the Yes contract low relative to its own assessment of the outcome probability, it views the contract as underpriced and purchases it. The agent can repeat this process across many markets and direct its trading activity toward those with the largest estimated mispricing.

This process resembles the work of a human analyst because both evaluate public information, assess its implications for future outcomes, compare those assessments with market prices, and trade when they believe prices do not fully incorporate available information. However, LLM agents can perform these tasks continuously across a much larger set of markets at relatively low marginal cost. Their ability to process information at scale may therefore increase the likelihood of identifying mispricing that would be costly or time-consuming for individual human traders to detect. When agents trade on information that has not yet been incorporated into market prices, their trades may also convey that information to the market and contribute to subsequent price discovery. Figure 1 illustrates this process.

2.2. *Related Literature and Hypothesis Development*

The entry of autonomous LLM agents may affect financial markets by changing which public information is processed and incorporated into prices. Processing corporate disclosures requires investors to devote attention to identifying relevant filings, extracting useful information, and assessing its implications for firm value. These costs can cause investors to disregard publicly available accounting information or incorporate it into prices only gradually (Blankespoor et al., 2019; Blankespoor et al., 2020). Consistent with this view, prior research documents that information contained in annual reports, including changes in disclosure language, can predict future returns, suggesting that economically relevant public information is not always processed immediately (Cohen et al., 2020).

Autonomous LLM agents can alter this processing constraint. Unlike human market participants, whose attention and processing capacity are scarce, autonomous agents can continuously monitor information sources and process corporate disclosures across many firms and markets at relatively low marginal cost. Consequently, information that is publicly available but costly for individual investors to identify or analyze may become economically feasible for agents to process. Agent participation can therefore expand the effective set of public information reflected in trading, even without changing the underlying information available to the market.

This expanded processing capacity can have consequences for price discovery. When an autonomous agent identifies information that has not yet been fully incorporated into a prediction-market price, trading on that information should move the market-implied probability toward the value implied by the underlying information. As similar information is incorporated through repeated agent trading, market prices should more closely reflect the ultimately realized outcome. Thus, the potential contribution of autonomous agents arises not simply from making predictions

more cheaply, but from reducing the processing frictions that can prevent existing public information from entering prices.

At the same time, autonomous processing need not eliminate all information frictions. Human investors face limited attention and incur costs to become aware of, acquire, and process corporate disclosures, which can lead them to disregard value-relevant public information or incorporate it only gradually (Hirshleifer and Teoh, 2003; Blankespoor et al., 2019; You and Zhang, 2009; Miller, 2010). These frictions are particularly important when disclosures are lengthy or complex. Autonomous LLM agents may alleviate some of these constraints by continuously monitoring and processing public information at low marginal cost. However, their contribution to price discovery ultimately depends on whether they can successfully interpret the information they acquire. The benefits of expanded processing capacity may therefore be greater when the primary friction is the amount of information to be processed, but more limited when the information itself is ambiguous or requires substantial contextual judgment.

Although autonomous agents can process information at low marginal cost, lower processing costs do not imply that disclosures are processed effectively. LLMs can struggle to retrieve relevant information from long contexts (Liu et al., 2024), generate unreliable outputs under uncertainty (Farquhar et al., 2024), and exhibit miscalibrated confidence in their predictions (Li et al., 2025). These limitations may become particularly consequential in autonomous settings, where errors in information retrieval or interpretation can propagate through subsequent reasoning and actions without human review (Zhou et al., 2024). Given these competing considerations, I state the hypothesis in the null form:

H1: Trading by autonomous LLM agents does not contribute to the incorporation of information into prediction-market prices.

3. Data, Sample, and Variable Construction

3.1. Data Sources and Sample Selection

To construct the sample, I begin with the full universe of Polymarket markets over the sample period. Based on each market's question, description, and resolution rules, I identify markets tied to publicly traded firms. I then retain firm-level financial prediction markets that ask whether a company's stock price will exceed a specified threshold by a given date. I exclude markets related to non-firm outcomes, such as politics, sports, and macroeconomic events, as well as firm-level markets that do not involve stock-price outcomes. Each retained market is matched to the corresponding publicly traded firm. This procedure yields 7,555 financial prediction markets between May 2024 and April 2026.

For these markets, I collect transaction-level data from Polymarket. Each observation records the pseudonymous wallet address, market contract, trade direction, execution price, position size in USDC, and timestamp precise to the second. The financial-market sample contains 1,083,560 trades with aggregate volume of approximately \$54.1 million USDC. I supplement these records with each market's resolution outcome, which allows me to determine whether individual trades and wallet-level net positions are directed toward the eventual winning outcome.

I match each financial prediction market to the corresponding public firm and to the most recent 10-K and 10-Q filings available on SEC EDGAR before the market closes. I also use LSEG Workspace to identify M&A rumor events occurring before market resolution.

3.2. Agent Classification

Following prior research that infers machine activity from observable behavioral patterns (Cao et al., 2023),⁵ I identify autonomous trading using two jointly observed signals in wallet-level trading behavior. Unlike equity markets, which operate during specified trading hours, Polymarket operates continuously, allowing participants to trade at any time of day. Autonomous AI agents are specifically designed to exploit this continuous trading environment. For example, industry evidence describes an autonomous AI agent that executes trading strategies on Polymarket continuously around the clock, including during periods when human users are sleeping or otherwise occupied (Canny, 2026; Olas. 2026; Pearl. 2026). I exploit this feature together with external OpenAI API outages to identify wallets whose trading behavior exhibits both persistent around-the-clock activity and dependence on LLM infrastructure. The two signals are jointly informative. Sustained activity across the full 24-hour cycle is difficult to maintain through manual interaction, while an abrupt cessation of trading during an external API outage provides an independent operational signal that the wallet relies on LLM infrastructure.

I focus on OpenAI API outages rather than disruptions to other OpenAI services because autonomous agents interact with LLMs programmatically through APIs. Polymarket’s open-source agent framework requires an OpenAI API key and integrates LLM calls with market information retrieval and trade execution (Polymarket, 2026). In contrast, outages affecting the ChatGPT web interface may disrupt human users without impairing programmatic agent operation. I obtain OpenAI API outage incidents from OpenAI’s public status-history archive, which reports the timing and affected components of service disruptions, and combine overlapping incidents into

⁵ Cao et al. (2023) infer machine readership from observable digital behavior rather than directly observing whether a filing is processed by an algorithm. Using SEC EDGAR log files, they classify an IP address as a machine visitor when it downloads filings from more than 50 unique firms in a day, and use firms’ historical machine-download activity as a proxy for expected machine readership.

common outage intervals. The primary *Agent* indicator equals one for wallets that exhibit trading activity in all 24 UTC hours over their complete trading history across all Polymarket markets and whose trading activity abruptly ceases entirely during the identified OpenAI API outage intervals.

To assess whether the classification instead captures conventional algorithmic trading, I examine market-making activity among classified agent. Market making represents a prominent form of algorithmic trading, where automated systems continuously post and update orders to supply liquidity (Hendershott et al., 2011; Menkveld, 2013; Brogaard et al., 2014).⁶ Polymarket separately identifies market-making activity through its maker-rebate program, which compensates wallets when their orders resting on the order book are executed. I use these rebate records to identify wallets engaged in market making.⁷ Among the 57,670 wallets in the full sample, 0.163% receive maker rebates, whereas none of the 8,227 wallets classified as agents do. Thus, the agent classification does not appear to capture algorithmic trading bot.

3.3. Descriptive Statistics

Table 1 reports descriptive statistics for the main variables. Panel B describes the 7,555 financial prediction markets. The average market attracts 48.2 wallets and 128.8 represented trades, with mean trading volume of 6,521.9 USDC. Market activity is right-skewed: the median market

⁶ Prior research identifies several forms of algorithmic trading in equity markets, including high-frequency market making and latency arbitrage. These strategies are typically used by sophisticated trading firms that invest in specialized infrastructure, such as locating their computers close to an exchange's servers. This infrastructure allows them to receive market information and execute trades within fractions of a second. They may also benefit from exchange-specific fee and rebate programs (Hendershott et al., 2011; Menkveld, 2013; Brogaard et al., 2014). Polymarket does not provide the same institutional environment, making these algorithmic trading strategies inapplicable in this setting.

⁷ Polymarket operates a maker-rebate program in eligible markets, including but not limited to Sports, Finance, Politics, and Economics. When a trader places a limit order that rests on the order book and another trader subsequently executes against that order, the resting trader is the “maker” and the executing trader is the “taker.” Polymarket charges a fee to the taker and returns a portion of the collected fees to qualifying makers as rebates. I identify market makers based on whether a wallet receives any maker rebate over its complete Polymarket trading history across all markets, rather than only within the financial prediction markets studied in the main analysis.

has 36 active wallets and 72 represented trades, compared with maxima of 2,083 wallets and 6,599 trades. The average market remains open for 9.4 days.

Agents participate broadly across the financial prediction markets. The average market contains 8.6 agent wallets, and 94.5% of markets contain at least one wallet classified as an agent. Agent wallets account for 16.7% of wallets in the average market, 14.0% of shares traded, and 19.0% of USDC trading volume. Thus, agent participation is widespread but does not dominate trading activity in the typical market.

Panel C reports summary statistics at the wallet-market level. *Agent* equals one for 17.7% of the 364,344 wallet-market observations. Mean *Net Position Accuracy* is 0.535, indicating that slightly more than half of wallet-market positions are on the eventual winning outcome. Wallets are generally not specialized in financial prediction markets: *Finance Trade Share* has a mean of 0.082 and a median of 0.038, indicating that most wallet activity occurs in other Polymarket categories. Longshot trading is also common. On average, 37.7%, 42.5%, and 45.6% of a wallet's trades within a market are executed at prices below 0.10, 0.30, and 0.50, respectively.

For observations matched to SEC filings, the average 10-K has a *10-K Log N Tokens* value of 10.804 and contains 32.6 tokens per sentence. Numerical information accounts for 33.5% of tokens in 10-K filings on average. The remaining disclosure characteristics exhibit meaningful cross-sectional variation, providing variation for the tests of how agent performance changes with the information-processing demands of corporate disclosures.

4. Empirical Results

4.1. Baseline Results: Agent Trading Performance

The baseline analysis compares the trading accuracy of LLM-agent wallets and human wallets within the same financial prediction markets. I estimate the following wallet-market-level regression:

$$Accuracy_{im} = \alpha + \beta Agent_i + \gamma'X_i + \mu_m + \varepsilon_{im} \quad (1)$$

where i indexes wallets and m indexes markets. The dependent variable, *Accuracy*, is *Net Position Accuracy*, which equals one if wallet i 's terminal net position in market m is directed toward the contract that ultimately resolves to one dollar, and zero otherwise. *Agent* is an indicator equal to one for wallets classified as autonomous LLM agents using the criteria described in Section 3.2. The coefficient of interest, β , captures the difference in trading accuracy between agent and human wallets within the same market.

The vector X_i contains two wallet-level controls. *Log(Lifespan Days)* captures trading tenure and is defined as the natural logarithm of one plus the number of calendar days between a wallet's first and last trades on Polymarket. *Finance Trade Share* measures specialization in financial prediction markets and is defined as the fraction of the wallet's complete Polymarket trading activity conducted in the financial markets examined in this study. Market fixed effects, μ_m , absorb characteristics common to all wallets trading in the same market, including the underlying firm, the stock-price threshold, market resolution date, and prediction difficulty. Standard errors are two-way clustered at the wallet and market levels.

Columns (1) and (2) of Table 2 report the wallet-market accuracy results. Column (1) includes market fixed effects without wallet-level controls. The coefficient on *Agent* is 0.108 ($t = 3.09$). Column (2) adds *Log(Lifespan Days)* and *Finance Trade Share*. The coefficient on *Agent*

increases to 0.121 ($t = 3.52$). Thus, agent wallets are 12.1 percentage points more likely than human wallets trading in the same market to hold a terminal net position on the eventual winning side after controlling for wallet trading tenure and specialization in financial markets.

Columns (3) and (4) repeat the analysis at the trade level using *Trade-Level Accuracy* as the dependent variable. *Trade-Level Accuracy* equals one when an individual trade is directed toward the eventual winning outcome. The coefficient on *Agent* is 0.060 ($t = 2.51$) in column (3) and increases to 0.077 ($t = 3.12$) after including the wallet-level controls in column (4). The higher accuracy of agent wallets is therefore evident both in their terminal market positions and in the direction of individual trades.

Column (5) examines whether the accuracy advantage also translates into higher investment returns. Using *Return on Investment* as the dependent variable and including the wallet-level controls and market fixed effects, the coefficient on *Agent* is positive at 0.095 ($t = 2.50$). Thus, the superior directional accuracy of agent wallets is accompanied by higher realized returns.

4.2. GPT-5.1 API Release and LLM Agent Accuracy

To provide evidence on whether advances in LLM technology translate into better autonomous trading decisions, I examine changes in agent accuracy around the November 13, 2025 release of GPT-5.1 on the OpenAI API (OpenAI 2025). I focus on the GPT-5.1 release for several reasons. First, GPT-5.1 represents a discrete frontier-model upgrade rather than a minor model snapshot, pricing change, or platform modification. Second, no other major OpenAI model release occurs within the seven-day window surrounding the event, reducing concerns about overlapping technology shocks. Third, the release occurs during a period of substantial Polymarket activity, providing sufficient observations for both agent and human wallets before and after the event across a broad set of markets, as shown in Panel A of Table 3.

I compare agent and human trades in a narrow window around the release, using trades placed during the seven calendar days before the release and the seven calendar days after the release. I estimate the following regression specification:

$$Accuracy_{imt} = \alpha + \beta_1 Agent_i + \beta_2 Post_t + \beta_3 (Agent_i \times Post_t) + \Gamma X_i + \mu_i + \lambda_m + \delta_t + \varepsilon_{imt} \quad (2)$$

where i , m , and t index wallets, markets, and dates, respectively. *Accuracy* is *Trade-Level Accuracy*, *Agent* equals one for wallets classified as LLM agents, *Post* equals one for trades placed after the November 13, 2025 GPT-5.1 release. X contains *Log(Lifespan Days)* and *Finance Trade Share*. The terms μ_i , λ_m , and δ_t denote wallet, market, and date fixed effects, respectively. The coefficient of interest is β_3 , which captures the change in agent accuracy after the release relative to the contemporaneous change in human-wallet accuracy. In specifications with wallet fixed effects, *Agent* and the wallet-level controls are absorbed; with UTC-date fixed effects, *Post* is absorbed. Standard errors are two-way clustered by wallet and market.

Table 3 reports the results. In column (1), the coefficient on *Agent* is -0.025 ($t = -0.67$), indicating no significant difference in accuracy between agent and human wallets before the GPT-5.1 release. The coefficient on *Post GPT-5.1* is -0.018 ($t = -0.80$), suggesting that human-wallet accuracy does not change significantly after the release. In contrast, the coefficient on *Agent* $\times$ *Post GPT-5.1* is 0.158 ($t = 3.14$), indicating that agent accuracy increases by 15.8 percentage points more than human-wallet accuracy after the release. Column (2) adds wallet, market, and date fixed effects, which absorb the main effects of *Agent* and *Post GPT-5.1*. The interaction remains positive and statistically significant, with a coefficient of 0.086 ($t = 4.21$). Thus, after accounting for time-invariant wallet characteristics, persistent differences across markets, and common daily shocks,

agent wallets experience an 8.6 percentage-point larger increase in trade-level accuracy than human wallets.

These results provide complementary evidence to the cross-sectional tests on disclosure characteristics. Rather than comparing agents and humans across information environments, this analysis uses a discrete change in the LLM technology available through the API and asks whether the relative performance of agent wallets changes at the same time. The findings show a marked increase in agent accuracy relative to human accuracy immediately following the GPT-5.1 API release. This pattern is consistent with improvements in API-accessible LLM capabilities being incorporated into at least some autonomous trading systems.

4.3. The Capabilities and Boundaries of Autonomous Disclosure Processing

4.3.1. Disclosure Length and Numerical Intensity

I first examine whether the accuracy advantage of LLM agents varies with the amount of textual information that must be processed. Greater disclosure length creates competing predictions for relative agent performance. Longer filings increase investors' information-processing demands and may therefore increase the relative advantage of LLM agents, which can process large volumes of text programmatically (Blankespoor et al. 2020). At the same time, evidence from the LLM literature suggests that model performance can deteriorate as input length increases because relevant information becomes more difficult to identify and use effectively within long contexts (Liu et al. 2024).

For each market, I use the most recent 10-K publicly available by the market closing date and retain the full primary filing document. To capture the volume of information that LLM agents may process, I measure disclosure length in tokens. Tokens are the units in which textual inputs are transmitted to and processed by LLM APIs; the number of input tokens determines the amount

of text included in a request, contributes to API usage costs, and is constrained by the model’s context window. Token counts therefore provide a natural measure of the information-processing demands faced by autonomous LLM agents. *10-K Log N Tokens* is the natural logarithm of one plus the total number of filing tokens and captures overall disclosure length. *10-K Tokens per Sentence* is the total number of tokens divided by the number of sentences and captures sentence-level processing burden. All disclosure measures are standardized to have a mean of zero and a standard deviation of one.

Table 4 presents the results. In column (1), the coefficient on *Agent × 10-K Log N Tokens* is positive and statistically significant at 0.014 ($t = 2.48$). Because the disclosure measure is standardized, this estimate implies that a one-standard-deviation increase in filing length is associated with a 1.4 percentage-point increase in the accuracy advantage of LLM agents relative to human traders. Column (2) yields a similar result for *10-K Tokens per Sentence*: the interaction coefficient is 0.011 ($t = 2.12$), indicating that the relative agent advantage increases by 1.1 percentage points for a one-standard-deviation increase in sentence-level processing burden.

These results indicate that the relative advantage of LLM agents increases with the amount of textual information to be processed. However, processing more text is distinct from reasoning over numerical information embedded in that text. I therefore next examine whether the LLM agent accuracy advantage extends to numerically intensive disclosures. This test serves two purposes. First, the ability to process large volumes of text does not necessarily imply an advantage in numerical reasoning. Levy (2026) shows that unaided LLMs perform poorly when financial tasks require identifying, combining, and reasoning over multiple numerical inputs, particularly when the relevant numbers are embedded in full financial statements. Although autonomous LLM

agents can mitigate these limitations by generating code or using external computational tools, whether they do so effectively in practice remains an empirical question.

Second, the numerical-intensity test helps distinguish LLM-based automation from conventional algorithmic trading. Traditional algorithmic trading systems commonly rely on programmed quantitative rules and structured market data to identify trading opportunities, manage orders, and respond rapidly to changes in prices and liquidity (Chaboud et al., 2014). In contrast, a defining feature of LLM-based agents is their ability to process and reason over unstructured textual information. A pattern in which the identified agents perform relatively better as textual processing demands increase, but less well as disclosures become more numerically intensive, would therefore be more consistent with LLM-based automation than with purely quantitative algorithmic trading.

I measure numerical content using *10-K Numerical Intensity*, defined as the proportion of tokens containing numerical information in the most recent 10-K. The measure is standardized before constructing its interaction with *Agent*. Column (3) of Table 4 presents the results. The coefficient on $Agent \times 10\text{-}K \text{ Numerical Intensity}$ is negative and statistically significant at -0.010 ($t = -2.43$), indicating that a one-standard-deviation increase in numerical intensity is associated with a 1.0 percentage-point reduction in the accuracy advantage of LLM agents relative to human traders. Combined with the results showing a stronger agent advantage for longer textual disclosures, the evidence is consistent with the identified wallets exhibiting a comparative advantage characteristic of LLM-based information processing rather than conventional quantitative automation.

4.3.2. *Litigious and Forward-Looking Disclosure*

I next examine whether the accuracy advantage of LLM agents is greater when disclosures contain richer textual information. I focus on litigious and forward-looking language because both convey economically relevant information through narrative disclosure. Litigious language captures information related to legal claims, litigation, regulation, and other legal exposures (Loughran and McDonald 2011), while forward-looking language conveys management's expectations, plans, and assessments of future firm outcomes and has been shown to contain information relevant to investors and analysts (Li 2010; Bozanic et al. 2018).

Following Cohen et al. (2020), I focus on year-over-year changes in 10-K filings because additions and deletions are more likely to capture incremental information than text that is repeated from prior disclosures. For each Polymarket market, I select the firm's most recent filing available by the market closing date and compare it with the prior filing of the same form. I use sequence matching to identify words and sentences added or removed relative to the prior-year filing.

Using this changed content, I construct two measures of qualitative disclosure. *Litigiousness* is the fraction of changed words classified as litigious by the Loughran–McDonald Master Dictionary. *Forward-Looking Content* is the fraction of changed sentences containing at least one forward-looking term or phrase from the list developed by Bozanic et al. (2018), including language related to expectations, forecasts, intentions, and plans.

Table 5 presents the results. The coefficient on $Agent \times Litigiousness$ is positive and statistically significant at 0.017 ($t = 3.47$), indicating that a one-standard-deviation increase in *Litigiousness* is associated with a 1.7 percentage-point increase in the accuracy advantage of LLM agents relative to human wallets. Similarly, the coefficient on $Agent \times Forward-Looking Content$ is positive at 0.010 ($t = 1.69$), indicating that a one-standard-deviation increase in forward-looking

content is associated with a 1.0 percentage-point increase in the relative accuracy advantage of LLM agents. Overall, the results suggest that LLM agents exhibit a greater relative accuracy advantage when disclosures contain more litigious or forward-looking information, consistent with an advantage in processing and integrating qualitative information.

4.3.3. The Limits of LLM Agents in Judgment-Intensive Settings

I examine whether the LLM agent accuracy advantage weakens when interpreting information requires greater judgment. Unlike forward-looking or litigious language, uncertain and vague disclosure may provide less precise signals about firm outcomes and place greater weight on contextual interpretation. I also examine M&A rumors as a complementary setting in which firm value depends on judgment-intensive assessments of whether a potential transaction will occur and on its likely implications for the firm.

I construct *Uncertainty* and *Vagueness* from year-over-year changes in firms' 10-K and 10-Q filings. *Uncertainty* is the fraction of changed words classified as uncertain by the Loughran–McDonald Master Dictionary, while *Vagueness* is the fraction of changed sentences containing at least one uncertainty word. I complement the disclosure-based tests with a judgment-intensive event setting. *M&A Rumor* equals one when LSEG Workspace classifies a “Significant M&A – Rumor” event for the firm and the event begins between the Polymarket market’s opening and closing dates.

Table 6 presents the results. Column (1) examines *Uncertainty*. The coefficient on *Agent* $\times$ *Uncertainty* is negative and statistically significant at -0.012 ($t = -2.50$), indicating that a one-standard-deviation increase in *Uncertainty* is associated with a 1.2 percentage-point reduction in the accuracy advantage of LLM agents relative to human wallets. Column (2) presents similar evidence for *Vagueness*. The coefficient on *Agent* $\times$ *Vagueness* is -0.009 ($t = -1.73$), indicating

that the relative agent advantage declines by 0.9 percentage points for a one-standard-deviation increase in *Vagueness*. Column (3) examines *M&A Rumor*. The coefficient on *Agent* $\times$ *M&A Rumor* is negative and statistically significant at -0.099 ($t = -2.27$), indicating that the accuracy advantage of LLM agents is 9.9 percentage points smaller in markets involving M&A rumors.

These findings suggest that the information-processing advantage of LLM agents is more limited when the relevant information is ambiguous or when evaluating firm outcomes requires greater judgment about uncertain developments.

4.4. Behavioral Bias and Longshot Trading

I examine whether autonomous LLM agents exhibit lower exposure to behavioral longshot bias. In prediction markets, low-priced contracts correspond to outcomes viewed as relatively unlikely and offer large state-contingent payoffs relative to their purchase prices. Prior research documents that traders tend to overpay for such contracts, a pattern commonly referred to as the longshot bias (Thaler and Ziemba, 1988; Snowberg and Wolfers, 2010). This behavior has been linked to preferences for lottery-like payoffs and other behavioral motives (Tversky and Kahneman, 1992; Barberis and Huang, 2008; Dorn and Sengmueller, 2009; Kumar, 2009). Because autonomous LLM agents do not derive utility from gambling or entertainment, their trading decisions may be less affected by these preferences. However, agents may still trade low-probability contracts when they assess them to be mispriced, so lower longshot trading is not mechanical.

To examine this possibility, I construct *Fraction Longshot* as the fraction of a wallet's trades in a market executed at an effective price below 0.10, 0.30, or 0.50. These thresholds capture progressively broader sets of low-priced contracts. I regress each measure on *Agent*, with market fixed effects and, in the controlled specifications, *Log(Lifespan Days)* and *Finance Trade Share*.

Table 7 presents the results. The coefficient on *Agent* is negative across all three longshot thresholds. In the controlled specifications, agent wallets allocate 9.6, 9.9, and 10.3 percentage points fewer trades to contracts priced below 0.10, 0.30, and 0.50, respectively, with *t*-statistics of -2.58 , -2.78 , and -2.98 . Thus, agent wallets are systematically less likely than human wallets trading in the same market to trade low-priced contracts. These results are consistent with a lower propensity among LLM agents to exhibit longshot demand.

4.5. Agent Trading and Market Price Accuracy

The preceding analyses show that autonomous LLM agents make informative trading decisions and that their relative performance varies with the characteristics of the information being processed. I next examine whether the information reflected in agent trading is incorporated into market prices. Specifically, I test whether agent order imbalance is associated with lower subsequent squared price error. This analysis moves beyond the accuracy of individual trading decisions and evaluates whether autonomous agent participation contributes to market-level price formation.

I use *Squared Price Error*, equivalent to the Brier score for a binary outcome, to measure prediction-market price accuracy.⁸ For market m and time window t , *Squared Price Error* is defined as:

$$SPE_{m,t} = (p_{m,t} - p_m^{outcome})^2 \quad (3)$$

⁸ Prediction-market prices can be interpreted as market-implied probabilities, so a continuous measure of the distance between the market price and the realized outcome is more informative than a binary indicator of whether the market favors the winning contract. Squaring the pricing error also places disproportionately greater weight on large errors, particularly confident predictions that ultimately prove incorrect. For example, if a market ultimately resolves YES, a price of 0.90 produces a squared error of 0.01, whereas a price of 0.60 produces a squared error of 0.16. Thus, although the absolute pricing error is four times larger in the latter case, its squared error is sixteen times larger. This feature is desirable because large deviations from the realized outcome represent more substantial failures of price accuracy.

where $p_{m,t}$ is the market price observed at time t and $p_m^{outcome} \in \{0,1\}$ is the realized payoff of the contract. Lower values of $SPE_{m,t}$ indicate that the market price is closer to the realized outcome and therefore more accurate. For market m and time window t , order imbalance is defined as net buy volume scaled by total trading volume within the same market-time window:

$$OIB_{m,t} = \frac{B_{m,t} - S_{m,t}}{B_{m,t} + S_{m,t}} \quad (4)$$

I construct separate OIB measures for agent and human wallets within each market-time window. I test whether agent trading contributes to subsequent market price accuracy by estimating:

$$SPE_{m,t+1} = \alpha + \beta_A AgentOIB_{m,t} + \beta_H HumanOIB_{m,t} + \mu_m + \lambda_{dw} + \varepsilon_{m,t+1} \quad (5)$$

where the unit of observation is a market-time-window pair. The regression is estimated separately using 15-minute, 30-minute, and 1-hour windows. For each market m and time window t , I measure Agent OIB and Human OIB using trades executed during window t . μ_m denotes market fixed effects, and λ_{dw} denotes date-by-time-window fixed effects. The latter absorb common shocks affecting all markets during the same intraday interval on a given calendar date.⁹

Table 8 presents the results. Across all three time windows, higher Agent OIB is associated with significantly lower *Squared Price Error* at $t + 1$. The coefficients are -0.053 , -0.046 , and -0.043 for the 15-minute, 30-minute, and 1-hour specifications, respectively, and are statistically significant at the 1% level. Thus, conditional on human order imbalance within the same market and time window, stronger agent order imbalance is associated with lower subsequent price error.

⁹ For example, in the 15-minute specification, all market observations between 10:00 and 10:15 UTC on a given date share the same fixed effect.

4.6. Robustness to Alternative Active-Hour Cutoffs

Table 9 assesses whether the baseline accuracy result depends on the strict 24-hour activity requirement. I successively lower the minimum number of active UTC hours to 23, 22, 21, and 20. The coefficient on *Agent* remains positive and statistically significant at the 1% level across all alternative definitions, ranging from 0.159 ($t = 4.84$) to 0.184 ($t = 5.78$). These results indicate that the higher directional accuracy of agent wallets is not driven by the specific 24-hour cutoff.

5. Conclusion

This study examines how autonomous LLM agents change financial information processing when investors delegate decisions to systems that independently acquire information, form predictions, and execute trades. Using transaction-level data from firm-level financial prediction markets on Polymarket, I identify autonomous LLM-agent wallets based on around-the-clock trading activity and abrupt disruptions to their trading during OpenAI API outages. This setting allows me to compare agents and human traders operating in the same unresolved markets with real capital at stake.

I find that LLM-agent wallets make more accurate trading decisions than human wallets, both in their terminal market positions and at the individual-trade level, and earn higher returns on investment. The relative performance of agents also improves following the release of GPT-5.1 on the OpenAI API, providing complementary evidence that advances in underlying LLM capabilities translate into better autonomous trading decisions.

The relative advantage of LLM agents varies systematically with the information being processed. The agent accuracy advantage increases with disclosure length and sentence length and

is stronger when year-over-year disclosure changes contain more litigious or forward-looking language. In contrast, the advantage declines as disclosures become more numerically intensive and when disclosure changes contain greater uncertainty or vagueness. It is also substantially smaller in markets involving M&A rumors, where traders must assess unverified information and uncertain future developments. Together, these results suggest that autonomous LLM agents have a comparative advantage in processing large volumes of textual financial information, but that this advantage becomes more limited when tasks require numerical reasoning or greater judgment under ambiguity. Agent wallets also allocate fewer trades to low-probability contracts, consistent with lower exposure to behavioral longshot demand.

Finally, agent order imbalance predicts lower subsequent price error, suggesting that information reflected in agent trading contributes to market price discovery. Overall, the findings suggest that autonomous LLM agents can expand investors' capacity to process corporate disclosures and contribute to price discovery. At the same time, the benefits of delegation are not uniform across information-processing tasks. Human judgment remains relevant when financial information is ambiguous and difficult to verify.

References

- Arrow, K.J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J.O., Levmore, S., Litan, R., Milgrom, P., Nelson, F.D. and Neumann, G.R., 2008. The promise of prediction markets. *Science*, 320(5878), pp.877-878.
- Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A. and Hovy, E., 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8), pp.852-863.
- Barber, B.M. and Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2), pp.785-818.
- Barberis, N. and Huang, M., 2008. Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, 98(5), pp.2066–2100.
- Blankespoor, E., deHaan, E. and Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2-3), p.101344.
- Bozanic, Z., Roulstone, D.T. and Van Buskirk, A., 2018. Management earnings forecasts and other forward-looking statements. *Journal of Accounting and Economics*, 65(1), pp.1-20.
- Brogaard, J., Hendershott, T. and Riordan, R., 2014. High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8), pp.2267–2306.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, pp.1877-1901.
- Cao, S., Jiang, W., Yang, B. and Zhang, A.L., 2023. How to talk when a machine is listening: Corporate disclosure in the age of AI. *The Review of Financial Studies*, 36(9), pp.3603-3642.
- Cao, S.S., Chen, W.X., Ma, G. and Srinivasan, S., 2026. Generative AI in capital markets: Information production, dissemination, and processing. *Journal of Accounting Research*, 64(3), pp.1427-1450.
- Chaboud, A.P., Chiquoine, B., Hjalmarsson, E. and Vega, C., 2014. Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69(5), pp.2045–2084.
- Cohen, L., Malloy, C. and Nguyen, Q., 2020. Lazy prices. *The Journal of Finance*, 75(3), pp.1371-1415.

Dorn, D. and Sengmueller, P., 2009. Trading as entertainment? *Management Science*, 55(4), pp.591–603.

Dyer, T., Lang, M. and Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2-3), pp.221-245.

Ecker, F., Li, X., Li, Y. and Wu, F., 2025. How stock market participants use generative artificial intelligence: Evidence from user-platform interaction data. Available at SSRN.

Farquhar, S., Kossen, J., Kuhn, L. and Gal, Y., 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), pp.625-630.

Ferrag, M.A., Tihanyi, N. and Debbah, M., 2026. From LLM reasoning to autonomous AI agents: A comprehensive review. *IEEE Access*, 14, pp.84237–84285.

Financial Stability Oversight Council, 2023. Annual Report. U.S. Department of the Treasury, Washington, D.C.

Hendershott, T., Jones, C.M. and Menkveld, A.J., 2011. Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1), pp.1–33.

Hirshleifer, D. and Teoh, S.H., 2003. Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1-3), pp.337-386.

Hirshleifer, D., Lim, S.S. and Teoh, S.H., 2011. Limited investor attention and stock market misreactions to accounting information. *Review of Asset Pricing Studies*, 1(1), pp.35–73.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P., 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), pp.1-38.

Kumar, A., 2009. Who gambles in the stock market? *The Journal of Finance*, 64(4), pp.1889–1933.

Levy, B., 2026. Caution ahead: Numerical reasoning and look-ahead bias in AI models. *Journal of Accounting Research*, 64(3), pp.1139–1188.

Olas. 2026. “Introducing Polystrat, an AI Agent That Trades Polymarket on Autopilot.”

Pearl. 2026. “Polystrat | AI Agent That Trades Polymarket on Autopilot.”

Li, C., Zhou, H., Glavaš, G., Korhonen, A. and Vulić, I., 2025, July. Large language models are miscalibrated in-context learners. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 11575-11596).

- Li, F., 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), pp.221-247.
- Li, F., 2010. The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), pp.1049-1102.
- Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. and Liang, P., 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, pp.157-173.
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y. and Tang, J., 2024. AgentBench: Evaluating LLMs as agents. In: The Twelfth International Conference on Learning Representations.
- Lopez-Lira, A., and Y. Tang. 2026. “Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models.” *Journal of Financial Economics*, forthcoming.
- Loughran, T. and McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35-65.
- Menkveld, A.J., 2013. High frequency trading and the new market makers. *Journal of Financial Markets*, 16(4), pp.712–740.
- Moss, A., Wegner, J. and Zechman, S.L., 2026. AI meets DIY: The Impact of Human Intervention on AI-Assisted Investing. Available at SSRN 6803678.
- Okoth, L. 2026. “Arbitrage Bots Dominate Polymarket With Millions in Profits as Humans Fall Behind.” *Yahoo Finance*.
- OpenAI, 2026. *OpenAI Agents SDK documentation*.
- OpenAI. 2025. *Introducing GPT-5.1 for Developers*.
- Polymarket, 2024. *Polymarket Agents*. GitHub repository.
- Polymarket, 2026. *Introduction: Overview of the Polymarket APIs*. Polymarket Documentation.
- Schoenegger, P., Tuminauskaite, I., Park, P.S., Bastos, R.V.S. and Tetlock, P.E., 2024. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45), p.eadp1528.
- Snowberg, E. and Wolfers, J., 2010. Explaining the favorite–long shot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, 118(4), pp.723-746.

- Thaler, R.H. and Ziemba, W.T., 1988. Anomalies: Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives*, 2(2), pp.161-174.
- Turtel, B., Franklin, D. and Schoenegger, P., 2025. LLMs can teach themselves to better predict the future. arXiv preprint arXiv:2502.05253.
- Tversky, A. and Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), pp.297–323.
- van Binsbergen, J.H., Han, X. and Lopez-Lira, A., 2023. Man versus machine learning: The term structure of earnings expectations and conditional biases. *The Review of Financial Studies*, 36(6), pp.2361-2396.
- Canny, W., and AI Boost. 2026. “AI Agents Are Quietly Rewriting Prediction Market Trading.” *CoinDesk*, March 15, 2026.
- Wolfers, J. and Zitzewitz, E., 2004. Prediction markets. *Journal of Economic Perspectives*, 18(2), pp.107-126.
- Xia, Y., You, P., Wang, T., Liu, F., Qi, H., Wu, X. and Zhang, S., 2026. Agentic Trading: When LLM Agents Meet Financial Markets. arXiv preprint arXiv:2605.19337.
- Zhang, Y., Zhu, Y. and Linnainmaa, J.T., 2025. Man versus machine learning revisited. *The Review of Financial Studies*, 38(12), pp.3768-3790.
- Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D. and Alon, U., 2024, May. Webarena: A realistic web environment for building autonomous agents. In International Conference on Learning Representations (Vol. 2024, pp. 15585-15606).

Appendix A
Variable Definitions

Variable	Definition
10-K Log N Tokens	Natural logarithm of the token count in the most recent 10-K available for the market.
10-K Numerical Intensity	Share of tokens containing numerical content in the most recent 10-K.
10-K Toks Per Sent	Token count divided by sentence count in the most recent 10-K available for the market.
Agent	Indicator equal to one if a wallet trades in all 24 UTC hours over its complete Polymarket history and its trading activity abruptly ceases during identified OpenAI API outage intervals.
Agent OIB	Net buy volume of agent wallets scaled by total trading volume in the same market-time window.
Finance Trade Share	Fraction of the wallet's complete Polymarket trades placed in the financial prediction markets used in the paper.
Forward-Looking Content	Forward-looking-language intensity in the year-over-year changed filing text.
Fraction Longshot	Fraction of a wallet's trades in a market executed at an effective price below 0.10, 0.30, or 0.50.
Human OIB	Net buy volume of human wallets scaled by total trading volume in the same market-time window.
Litigiousness	Litigious-language intensity in text added to or removed from the latest filing relative to the comparable prior-year filing.
Log(Lifespan Days)	Natural logarithm of one plus the calendar days between the wallet's first and last trades in its complete Polymarket history.
<i>M&A Rumor</i>	Indicator equal to one if LSEG Workspace identifies an M&A rumor involving the firm between the market opening and closing dates, and zero otherwise.
Net Position Accuracy	Indicator equal to one when a wallet's terminal net position in a market is directed toward the contract that ultimately pays one dollar.
Post GPT-5.1	Indicator equal to one for trades placed during the seven calendar days after the November 13, 2025 release of GPT-5.1 on the OpenAI API, and zero for trades placed during the seven calendar days before the release.
Return on Investment	Realized profit and loss of a wallet in a market divided by the total capital deployed by the wallet in that market.
Squared Price Error	Squared difference between the prediction-market price and the realized binary outcome. Lower values indicate greater market price accuracy.

Variable	Definition
Trade-Level Accuracy	Indicator equal to one when an individual trade is directed toward the contract that ultimately pays one dollar; sales are converted to equivalent opposite-side purchases.
Uncertainty	Uncertainty-language intensity in the year-over-year changed filing text.
Vagueness	Vagueness-language intensity in the year-over-year changed filing text.

Figure 1

Workflow of an LLM Agent in Financial prediction markets

This figure illustrates the decision process of an LLM agent operating in Polymarket. The agent collects publicly available information, including market prices, news, corporate disclosures, and other relevant information sources. The information is processed by the LLM to form an assessment of the likelihood that a market event will occur. The agent then compares this assessment with the probability implied by market prices. When the difference between the two is sufficiently large, the agent executes a trade through the prediction market platform.

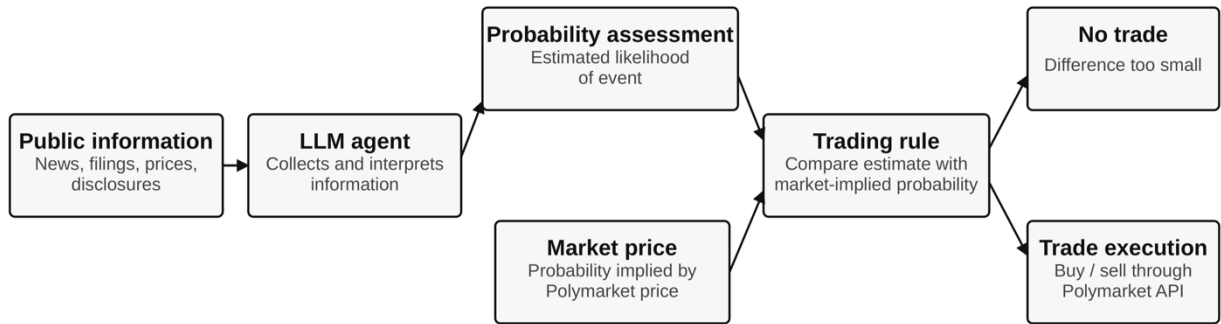


Table 1

Sample Construction and Summary Statistics

This table reports the construction and characteristics of the sample used in the main analyses. Panel A describes the construction of the financial prediction market sample. The initial sample includes all Polymarket markets closed between May 2024 and April 2026. *Markets Remaining* reports the number of markets retained after each selection stage, and *Markets Excluded* reports the number excluded at that stage. Panel B reports market-level characteristics for the 7,555 financial prediction markets in the final sample, including wallet participation, trading activity, agent participation, and market duration. Panel C reports summary statistics for the variables used in the main wallet-market analyses. *Agent, Net Position Accuracy*, wallet-level controls, and longshot trading measures are reported for the full wallet-market sample. Disclosure characteristics are reported for the subsample of wallet-market observations matched to SEC EDGAR filings. Variable definitions are provided in Appendix A.

Panel A: Sample Construction

Sample-Selection Stage	Markets Remaining	Markets Excluded
Markets closed between May 2024 and April 2026	638,776	—
Identify markets potentially related to publicly traded firms based on market questions and resolution criteria	14,031	624,745
Retain markets linked to publicly traded firms in SEC Analytics	9,121	4,910
Exclude firm-level markets that do not involve stock-price outcomes	7,555	1,566

Panel B: Market-Level Summary Statistics

Variable	N	Mean	SD	Min	p25	Median	p75	Max
Number of Wallets	7,555	52.450	64.034	2	25	38	60	2,089
Number of Agent Wallets	7,555	9.497	11.245	0	3	6	12	133
Any Agent Wallet	7,555	0.950	0.218	0.000	1.000	1.000	1.000	1.000
Agent Wallet Share	7,555	0.170	0.094	0.000	0.103	0.167	0.229	0.531
Number of Trades	7,555	143.423	232.031	2	43	79	158	6,926
Total Shares Traded	7,555	14,332.378	35,385.516	10.000	3,507.785	7,816.191	16,802.225	1,720,020.268
Total USDC Trading Volume	7,555	7,157.402	17,690.077	5.000	1,744.629	3,905.083	8,377.442	859,303.031
Agent Share of Shares Traded	7,555	0.141	0.146	0.000	0.029	0.086	0.211	0.835
Agent Share of USDC Trading Volume	7,555	0.188	0.253	0.000	0.017	0.068	0.256	0.999
Market Duration (Days)	7,555	9.419	10.659	0.142	4.426	7.042	7.123	364.702

Panel C. Wallet-Market Summary Statistics

Variable	N	Mean	SD	Min	p25	Median	p75	Max
Agent	364,344	0.177	0.382	0.000	0.000	0.000	0.000	1.000
Finance Trade Share	364,344	0.082	0.123	0.000	0.013	0.038	0.096	1.000
Fraction Longshot ($p < 0.10$)	364,344	0.377	0.459	0.000	0.000	0.000	1.000	1.000
Fraction Longshot ($p < 0.30$)	364,344	0.425	0.461	0.000	0.000	0.000	1.000	1.000
Fraction Longshot ($p < 0.50$)	364,344	0.456	0.462	0.000	0.000	0.400	1.000	1.000
Log(Lifespan Days)	364,344	5.101	0.955	0.000	4.691	5.198	5.746	7.138
Net Position Accuracy	364,344	0.535	0.499	0.000	0.000	1.000	1.000	1.000
M&A Rumor	361,974	0.006	0.080	0.000	0.000	0.000	0.000	1.000
10-K Log N Tokens	311,526	10.804	0.301	10.286	10.601	10.832	10.969	11.538
10-K Numerical Intensity	311,526	0.335	0.092	0.119	0.295	0.352	0.404	0.600
10-K Tokens per Sentence	311,526	32.596	1.980	26.984	31.909	32.369	33.296	38.191
Forward-Looking Content	311,526	0.102	0.025	0.044	0.097	0.107	0.119	0.180
Litigiousness	311,526	0.015	0.008	0.002	0.007	0.013	0.022	0.029
Uncertainty	311,526	0.013	0.004	0.003	0.010	0.013	0.016	0.029
Vagueness	311,526	0.314	0.068	0.158	0.248	0.327	0.360	0.548

Table 2**Trading Accuracy of LLM Agents**

This table reports baseline regressions comparing the trading performance of LLM-agent and human wallets. Columns (1) and (2) examine Net Position Accuracy, which equals one if the wallet's net position at market close is in the eventual winning outcome, and zero otherwise. Columns (3) and (4) examine Trade-Level Accuracy, which equals one if an individual trade is placed on the eventual winning outcome, and zero otherwise. Column (5) examines Return on Investment. Agent equals one for wallets classified as autonomous LLM agents. Columns (2), (4), and (5) include Log(Lifespan Days) and Finance Trade Share as wallet-level controls. Log(Lifespan Days) measures wallet trading tenure, and Finance Trade Share measures the share of a wallet's complete Polymarket trading activity conducted in the financial prediction markets examined in this study. All specifications include market fixed effects. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Net Position Accuracy		Trade-Level Accuracy		Return on Investment
	(1)	(2)	(3)	(4)	(5)
Agent	0.108*** (3.09)	0.121*** (3.52)	0.060** (2.51)	0.077*** (3.12)	0.095** (2.50)
Log(Lifespan Days)		0.029*** (2.82)		0.020** (2.32)	0.043*** (4.60)
Finance Trade Share		0.509*** (5.23)		0.195*** (4.10)	0.631*** (7.15)
Market Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	364,344	364,344	1,083,560	1,083,560	364,344
Adj. R-squared	0.136	0.149	0.014	0.018	0.132

Table 3

GPT-5.1 API Release and LLM Agent Accuracy

This table examines changes in LLM agent accuracy around the November 13, 2025 release of GPT-5.1 on the OpenAI API. The sample includes trades placed during the seven days before and seven days after the release. Panel A reports the number of trades placed by agent and human wallets before and after the release. Panel B reports trade-level regressions in which the dependent variable is Trade-Level Accuracy. Post GPT-5.1 equals one for trades placed after November 13, 2025. The coefficient of interest is Agent $\times$ Post GPT-5.1, which captures the change in agent accuracy after the release relative to the contemporaneous change in human-wallet accuracy. Column (1) includes Agent, Post GPT-5.1, their interaction, Log(Lifespan Days), and Finance Trade Share. Column (2) includes wallet, market, and date fixed effects; Agent, Post GPT-5.1, and the wallet-level controls are absorbed and therefore not separately estimated. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

Panel A. Sample Composition

Sample Composition	Before	After
Human trades	21,658	35,695
Agent trades	4,445	7,293

Panel B. Changes in Trade-Level Accuracy Around the GPT-5.1 Release

	Trade-Level Accuracy	
	(1)	(2)
Agent	-0.025 (-0.67)	
Post GPT-5.1	-0.018 (-0.80)	
Agent $\times$ Post GPT-5.1	0.158*** (3.14)	0.086*** (4.21)
Log(Lifespan Days)	-0.006 (-0.59)	
Finance Trade Share	0.065 (1.43)	
Constant	0.500*** (9.01)	
Wallet Fixed Effects	No	Yes
Market Fixed Effects	No	Yes
Date Fixed Effects	No	Yes
Observations	69,091	66,502
Adj. R-squared	0.007	0.247

Table 4

Textual and Numerical Disclosure Characteristics and LLM Agent Performance

This table reports regressions of Net Position Accuracy on the Agent indicator and characteristics of 10-K disclosures. Column (1) examines 10-K Log N Tokens, column (2) examines 10-K Tokens per Sentence, and column (3) examines 10-K Numerical Intensity. All specifications include Log(Lifespan Days) and Finance Trade Share as controls and market fixed effects. The main effects of the disclosure characteristics are absorbed by market fixed effects. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Net Position Accuracy		
	(1)	(2)	(3)
Agent	0.124*** (3.37)	0.123*** (3.36)	0.123*** (3.35)
Agent × 10-K Log N Tokens	0.014** (2.48)		
Agent × 10-K Tokens per Sentence		0.011** (2.12)	
Agent × 10-K Numerical Intensity			-0.010** (-2.43)
Log(Lifespan Days)	0.018 (1.61)	0.018 (1.61)	0.018 (1.61)
Finance Trade Share	0.512*** (5.01)	0.512*** (5.01)	0.512*** (5.01)
Market Fixed Effects	Yes	Yes	Yes
Observations	311,526	311,526	311,526
Adj. R-squared	0.147	0.147	0.147

Table 5

Litigious and Forward-Looking Disclosure and LLM Agent Performance

This table reports regressions of Net Position Accuracy on Agent and its interactions with measures of litigious and forward-looking disclosure. Column (1) examines Litigiousness, and column (2) examines Forward-Looking Content. All specifications include Log(Lifespan Days) and Finance Trade Share as controls and market fixed effects. The main effects of the disclosure characteristics are absorbed by market fixed effects. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Net Position Accuracy	
	(1)	(2)
Agent	0.123*** (3.35)	0.123*** (3.35)
Agent × Litigiousness	0.017*** (3.47)	
Agent × Forward-Looking Content		0.010* (1.69)
Log(Lifespan Days)	0.018 (1.61)	0.018 (1.61)
Finance Trade Share	0.512*** (5.01)	0.512*** (5.01)
Market Fixed Effects	Yes	Yes
Observations	311,526	311,526
Adj. R-squared	0.147	0.147

Table 6

Uncertainty, Vagueness, M&A Rumors, and LLM Agent Performance

This table reports regressions of Net Position Accuracy on Agent and its interactions with measures of uncertainty, vagueness, and M&A rumors. Columns (1), (2), and (3) examine Uncertainty, Vagueness, and M&A Rumor, respectively. M&A Rumor is an indicator equal to one if the firm is subject to an M&A rumor, and zero otherwise. All specifications include Log(Lifespan Days) and Finance Trade Share as controls and market fixed effects. Because Uncertainty, Vagueness, and M&A Rumor vary at the market level, their main effects are absorbed by the market fixed effects. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Net Position Accuracy		
	(1)	(2)	(3)
Agent	0.123*** (3.36)	0.122*** (3.34)	0.122*** (3.53)
Agent × Uncertainty	-0.012** (-2.50)		
Agent × Vagueness		-0.009* (-1.73)	
Agent × M&A Rumor			-0.099** (-2.27)
Log(Lifespan Days)	0.018 (1.61)	0.018 (1.60)	0.029*** (2.80)
Finance Trade Share	0.512*** (5.01)	0.512*** (5.01)	0.510*** (5.23)
Market Fixed Effects	Yes	Yes	Yes
Observations	311,526	311,526	361,974
Adj. R-squared	0.147	0.147	0.150

Table 7

Longshot Trading and LLM Agent Behavior

This table reports regressions of longshot trading propensity on the Agent indicator. The dependent variable is the fraction of a wallet's trades in a market executed at an effective price below 0.10 in columns (1) and (2), below 0.30 in columns (3) and (4), and below 0.50 in columns (5) and (6). Columns (2), (4), and (6) include Log(Lifespan Days) and Finance Trade Share as controls. All specifications include market fixed effects. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Fraction Longshot					
	p<0.10		p<0.30		p<0.50	
	(1)	(2)	(3)	(4)	(5)	(6)
Agent	-0.082** (-2.18)	-0.096*** (-2.58)	-0.086** (-2.39)	-0.099*** (-2.78)	-0.090*** (-2.61)	-0.103*** (-2.98)
Log(Lifespan Days)		-0.032*** (-3.51)		-0.030*** (-3.20)		-0.028*** (-2.94)
Finance Trade Share		-0.546*** (-6.16)		-0.470*** (-5.13)		-0.420*** (-4.53)
Market Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	364,344	364,344	364,344	364,344	364,344	364,344
Adj. R-squared	0.162	0.181	0.148	0.162	0.143	0.154

Table 8**Agent and Human Order Imbalance and Market Price Accuracy**

This table reports market-time-window regressions of Squared Price Error on Agent OIB and Human OIB, which enter each regression jointly. Columns (1), (2), and (3) aggregate trades over 15-minute, 30-minute, and 1-hour windows, respectively. Lower Squared Price Error indicates a market price closer to the realized outcome. Agent OIB and Human OIB are defined as each group's net buy volume scaled by total trading volume within the corresponding market-time window. All specifications include market fixed effects and date-by-time-window fixed effects corresponding to the aggregation interval. Standard errors are two-way clustered by market and time window, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Squared Price Error		
	15-Minute Window	30-Minute Window	1-Hour Window
	(1)	(2)	(3)
Agent OIB	-0.053*** (-16.35)	-0.046*** (-16.51)	-0.043*** (-15.95)
Human OIB	-0.056*** (-17.66)	-0.048*** (-18.04)	-0.045*** (-17.61)
Market Fixed Effects	Yes	Yes	Yes
Date-by-Time-Window Fixed Effects	Yes	Yes	Yes
Observations	56,738	59,027	61,272
Adj. R-squared	0.571	0.565	0.558

Table 9

Robustness to Alternative Active-Hour Cutoffs

This table reports regressions of Net Position Accuracy on alternative definitions of the Agent indicator based on minimum active-hour requirements. Columns (1) through (4) require activity in at least 23, 22, 21, and 20 distinct UTC hours, respectively. All Agent definitions additionally require zero trades during core-LLM API outage intervals and at least one trade outside those intervals. All specifications include Log(Lifespan Days) and Finance Trade Share as controls and market fixed effects. Standard errors are two-way clustered at the wallet and market levels, and t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variable definitions are provided in Appendix A.

	Net Position Accuracy			
	Active Hours ≥ 23	Active Hours ≥ 22	Active Hours ≥ 21	Active Hours ≥ 20
	(1)	(2)	(3)	(4)
Agent	0.159*** (4.84)	0.174*** (5.39)	0.179*** (5.59)	0.184*** (5.78)
Log(Lifespan Days)	0.032*** (3.13)	0.033*** (3.29)	0.034*** (3.41)	0.035*** (3.50)
Finance Trade Share	0.510*** (5.35)	0.506*** (5.39)	0.504*** (5.41)	0.506*** (5.44)
Market Fixed Effects	Yes	Yes	Yes	Yes
Observations	364,344	364,344	364,344	364,344
Adj. R-squared	0.157	0.161	0.163	0.164