How AI Agents Mislead Investors:

Early Evidence from Automated Stock Recommendations*

Yongxin Xu[1] Yuhao Xuan[2] Gaoping Zheng[3]

Abstract:

Asking ChatGPT to provide stock recommendations from May 2024 to June 2025, we provide early evidence on how artificial intelligence (AI) agents mislead investors when autonomously assisting financial decision-making. Our analysis reveals three channels through which AI agent misinformation manifests: training-data reliance, web-searching bias, and representative heuristic reasoning. Model updates alleviate but do not eliminate misinformation, while prompting AI agents to revise their recommendations sometimes worsens it. Recommendations relying on training data underperform yet still attract investor attention, followed by less informed trading. Our study provides a systematic framework for evaluating AI agent reliability in automated investment advice and trading.

JEL classification codes: G11; G14; G18; G40; G41
***Keywords*:** Artificial intelligence; AI agent; Misinformation; Hallucinations; Financial advice; ChatGPT

[1] Xu is with the Monash Business School, Monash University. Email: yongxin.xu@monash.edu
[2] Xuan is with the Lingnan College, Sun Yat-sen University. Email: xuanyh5@mail.sysu.edu.cn
[3] Zheng is with the School of Economics, Finance and Marketing, RMIT University. E-mail: gaoping.zheng@rmit.edu.au

* We remain responsible for all errors and omissions.

1. Introduction

Since the emergence of ChatGPT, AI chatbots based on large language models (LLMs) have attracted enormous attention from investors and academia. Studies show that these AI chatbots, when provided with historical financial texts, can extract useful information to predict stock returns (e.g., Lopez-Lira and Tang 2025; Chen, Tang, Zhou, and Zhu 2025), replicate survey-based economic expectations (Bybee 2025), measure investor disagreement (Bhagwat, Cookson, Dim, and Niessner 2025), dissect corporate culture (Li, Mai, Shen, Yang, and Zhang 2025), analyze firm policies (Jha, Qian, Weber, and Yang 2024; Kim, Muhn, and Nikolaev 2024a,b), interpret earnings calls (Beckmann, Beckmeyer, Filippou, Menze, and Zhou 2024), and exhibit investor-like biases and preferences (Chen, Green, Gulen, and Zhou 2024; Fedyk, Kakhbod, Li, and Malmendier 2024; Ouyang, Yun, Zheng 2024). However, all these applications share a fundamental limitation: researchers must provide AI chatbots with data to analyze.

This limitation is rapidly disappearing. Artificial intelligence agents (AI agents) represent an evolution beyond AI chatbots, leveraging large language models (LLMs) with tool-using capabilities to autonomously take actions on users' behalf (Wang et al., 2024).[1] Unlike traditional AI chatbots, which analyze information provided by users and are bounded by their training data, AI agents can autonomously access real-time information and execute complex workflows, creating significant opportunities for automated investment advice and trading. However, this enhanced capability introduces new challenges that have already attracted regulatory scrutiny.[2] While prior studies have documented biases in traditional AI systems (e.g., Fedyk et al. 2024; Ouyang et al. 2024; Cao, Wang, and Xiang 2025; Dou, Goldstein, and Ji 2025), how systematic misinformation emerges when AI agents actively gather and process real-time information for financial decision-making remains largely unknown.

[1] According to a McKinsey report, "We are beginning an evolution from knowledge-based, gen-AI-powered tools—say, chatbots that answer questions and generate content—to gen AI–enabled "agents" that use foundation models to execute complex, multistep workflows across a digital world. In short, the technology is moving from thought to action." Source: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-agents-are-the-next-frontier-of-generative-ai

[2] https://www.investor.gov/introduction-investing/general-resources/news-alerts/alerts-bulletins/investor-alerts/artificial-intelligence-fraud; https://www.sec.gov/newsroom/press-releases/2024-36; https://www.sec.gov/newsroom/press-releases/2024-167

We address this gap by analyzing stock recommendations generated by ChatGPT using search tools across trading days from May 2024 to June 2025.[3] ChatGPT was among the first widely accessible LLMs that allowed registered users to use web search tools to obtain real-time information. Our data collection follows a structured protocol with six core questions: one initial question asking for stock recommendations, followed by five questions asking for updated recommendations. Since October 21, 2024, we added a question after the initial recommendation to verify whether it relied exclusively on internal training data. By collecting real-time recommendations and testing their subsequent performance, our method avoids the lookahead bias where pretrained language models' outputs are contaminated by future information in training data (Sarkar and Vafa 2024; Levy 2025) and the memorization problem where LLMs perfectly recall information from their training data (Lopez-Lira, Tang, and Zhu 2025).

We focus on stock recommendations for two reasons. First, investors are actively using LLMs for stock investments (e.g., Blankespoor, Croom, and Grant 2024; Bertomeu, Lin, Liu, and Ni 2025; Cheng, Lin, and Zhao 2025), making this an empirically important setting. Second, and more fundamentally, the process AI agents use to generate stock recommendations is representative of how they assist with other financial tasks requiring real-time information. AI agents must understand the user request, use tools to collect current information, analyze the gathered data, and provide actionable recommendations. Because this workflow generalizes across AI agent applications in finance, stock recommendations provide an ideal empirical setting for developing a systematic framework to detect AI agent misinformation.

Our comprehensive analyses reveal the following key findings. First, we identify three distinct channels through which AI agent misinformation manifests in initial stock recommendations: training-data reliance, web-searching bias, and representative heuristic reasoning. Second, model updates alleviate but do not eliminate misinformation, while prompting AI agents to revise their recommendations sometimes worsens it—revised recommendations exhibit similar or even more severe misinformation. Third, training-data recommendations underperform web-searched recommendations over a 5-day holding period yet attract similar investor attention spikes, suggesting investors cannot distinguish between

[3] Data collection is ongoing through December 2025.

the two information sources. Moreover, retail trading following training-data recommendations exhibits weaker return predictability, indicating this trading is less informed.

We now describe our findings in greater detail. The first channel of misinformation is training-data reliance, where ChatGPT uses internal training data to provide stock recommendations instead of using web-searched information. Specifically, we find that 29.7% of GPT-4 stock recommendations cite sources that were not actually searched, indicating reliance on internal training data. For the more advanced model GPT-4o, this percentage is even higher at 37.3%, with an average of 7.44 stocks per day (out of 20) citing unsearched sources compared to 5.87 for GPT-4. This pattern demonstrates that AI models performing better on general tasks do not necessarily provide better financial advice.

One concern regarding this measurement approach is that citing unsearched sources could reflect citation hallucination—where AI agents randomly cite sources regardless of whether they were searched—rather than genuine reliance on training data. Our further analyses do not support this citation hallucination explanation. First, we corroborate our findings using an alternative one-shot method where we ask AI agents in a single follow-up prompt to self-report their use of training data after generating recommendations. This method reveals consistent results: for GPT-4o, 50.3% of recommendation days were self-declared as using training data, closely matching the 37.3% identified through source verification. Notably, this scalable method reveals that training-data reliance persists even in ChatGPT-5, OpenAI's latest and most advanced model with substantially fewer hallucinations (untabulated). Second, we show later that recommendations citing unsearched sources systematically underperform those citing searched sources, with performance gaps of 15-35 basis points over five trading days. This systematic performance difference is difficult to reconcile with random citation errors and provides strong evidence that our measure captures genuine reliance on outdated training data rather than citation hallucination.

The second channel of misinformation is web-searching bias. When searching information online, GPT-4o disproportionately concentrates its searches on a small set of websites, primarily Yahoo Finance, Morningstar, and Investopedia. GPT-4 shows somewhat more diverse searching but still concentrates heavily on seven websites (Bing, Investopedia, Morningstar, NerdWallet, Fidelity Investments, Kiplinger, and Seeking Alpha). Since

concentrated search results severely limit the effectiveness of information searches (Xu, Xuan, and Zheng 2021), ChatGPT's concentrated search behavior limits the breadth and quality of information used to generate investment recommendations. Beyond harming individual investors through lower-quality recommendations, as AI adoption increases, the tendency for many investors to receive similar recommendations based on the same limited sources could exacerbate herding behavior, increase price volatility, and undermine efficient price discovery.

The third channel of misinformation is representative heuristic reasoning. When stating reasons for stock recommendations, GPT-4o focuses predominantly on firm fundamentals, such as "growth drivers and expansion" or "technology and innovation leadership." GPT-4o considers current valuation in only 5.1% of recommendations, suggesting it treats good firms as good investments—exhibiting the representative heuristic documented in behavioral finance (Tversky and Kahneman 1974; Lakonishok, Shleifer, and Vishny 1994; Barberis, Shleifer, and Vishny 1998). GPT-4 exhibits similar patterns, with valuation-related reasoning appearing in only 3.6% of recommendations. This result indicates that if investors use AI-provided reasoning as input for their own financial analysis, they may be systematically misled by the omission of valuation considerations.

We then examine whether prompting AI agents to revise their recommendations reduces misinformation. After collecting initial recommendations, we asked AI agents five times to update their stock lists by searching additional websites. These sequential prompts generate substantial changes in recommendations, as both the stocks and the ranking of stocks change considerably. Despite these substantial revisions, the three misinformation channels persist in updated recommendations. Training-data reliance remains prevalent, with GPT-4o's fifth update showing 40.5% of stocks citing unsearched sources (compared to 37.3% initially) and GPT-4 showing 48.2% (compared to 29.7% initially). Web-searching bias continues, with concentrated search patterns still dominating. Representative heuristic reasoning persists, with valuation considerations appearing in only 5.3% of GPT-4o's updated recommendations and 0.8% of GPT-4's updated recommendations (compared to 5.1% and 3.6% initially). These results demonstrate that simple prompting for self-revision cannot address the systematic misinformation embedded in AI agents' recommendation process.

A natural question is whether technological progress can mitigate these biases. During our sample period, ChatGPT has experienced two major model updates: in November 2024 and May 2025. These updates allow us to examine whether technological improvements reduce misinformation. We find that model upgrades reduce but do not eliminate training-data reliance. For example, even after the May 2025 update, 23.3% of GPT-4o stock recommendations still rely on internal training data rather than searched information. Web-searching bias initially increases after the November 2024 update but subsequently decreases following the May 2025 update. In contrast, representative heuristic reasoning persists, with valuation considerations remaining minimal (around 5%) across both updates. These results suggest that while some AI agent misinformation patterns evolve with model improvements, the core biases we identify persist across model generations.

The final part of our analysis examines the capital market consequences of misinformation generated by AI agents. Among the three misinformation channels we identify, training-data reliance represents the most unambiguous form of misinformation: using outdated internal knowledge instead of current market information[4] when explicitly prompted to search the web fundamentally contradicts the purpose of deploying AI agents for real-time financial decision-making. We therefore focus our main analysis on whether recommendations rely on training data versus web-searched information.

We find that training-data reliance systematically undermines investment performance. For GPT-4o, web-searched recommendations generate 5-day cumulative abnormal returns (i.e., market-adjusted returns) of 21.0 bps compared to only 6.3 bps for training-data recommendations. For GPT-4, the pattern is even more pronounced: web-searched recommendations generate 3.8 bps while training-data recommendations lose 30.9 bps.[5]

Despite the performance gap between training-data and web-searched recommendations, both types trigger short-lived spikes in investor attention measured by Google search volume and Twitter discussions (Da, Engelberg, and Gao 2011; Cookson, Lu, Mullins, and Niessner 2024). This similarity is particularly striking: while attention spikes following web-searched

[4] Before GPT-4 was removed after April 30, 2025, its training data cutoff was December 2023. Before January 29, 2025, GPT-4o had a training data cutoff of November 2023. After January 29, 2025, GPT-4o had a training data cutoff of June 2024.
[5] Supplementary analyses suggest that the other two misinformation channels also harm performance: concentrated searching leads to inferior outcomes for both models, while ignoring valuation produces worse returns for GPT-4o, though GPT-4's valuation-focused recommendations show an inconsistent pattern that may reflect its weaker reasoning capabilities.

recommendations could reflect current attention-grabbing events, training-data recommendations are based on information from months prior (training data cutoffs range from November 2023 to June 2024, while recommendations span May 2024 to June 2025). The fact that both types of recommendations generate comparable attention spikes suggests that investors cannot distinguish between recommendations based on current versus outdated information. This inability to identify AI agent misinformation has important implications: we find that retail trading is a weaker predictor of future returns after training-data recommendations than after web-searched recommendations, indicating that misinformed recommendations attract less informed trading activity.

Our study advances the understanding of how AI assists human financial decisions by distinguishing between AI as an expert versus AI as an agent. Previous studies examine AI as an expert that analyzes external information provided by humans using its internal knowledge base, finding that AI can provide useful insights for predicting stock price movements and firm activities (e.g., Lopez-Lira and Tang 2024; Chen, Tang, Zhou, and Zhu 2025; Chen, Kelly, and Xiu 2024; Jha, Qian, Weber, and Yang 2025; Kim, Muhn, and Nikolaev 2024a, 2024b; Beckmann, Beckmeyer, Filippou, Menze, and Zhou 2025). In contrast, we examine AI agents that autonomously use tools (e.g., web search engines) to collect and analyze real-time information. We document that this enhanced autonomy introduces systematic misinformation through training-data reliance, web-searching bias, and representative heuristic reasoning, highlighting significant risks for automated financial decision-making.

These findings have important implications for the growing literature on automated investment advisors. Traditional robo-advisors have demonstrated clear benefits, helping households make better wealth management decisions and improve welfare (e.g., Chak, Croxson, D'Acunto, Reuter, Rossi, and Shaw 2022; Reher and Sokolinski 2024; Rossi and Utkus 2024). Recent work by Pelster and Val (2024) extends this positive view to LLMs, conducting two queries with ChatGPT-4 (without web search capability at that time) and finding that higher-rated stocks generated modestly higher returns. Our systematic analysis across 14+ months reveals a more nuanced picture: while AI agents can generate recommendations that sometimes outperform market benchmarks, misinformation substantially undermines their performance.

Beyond these practical applications, this work contributes to the growing literature on how AI transforms information acquisition in financial markets. Evidence shows widespread AI adoption among market participants: Blankespoor, Croom, and Grant (2024) find that investors extensively use generative AI to gather and interpret financial or market data, believing it enhances their information processing capabilities, while Bertomeu, Lin, Liu, and Ni (2025) document reduced analyst forecast activity when ChatGPT access is restricted. This trend parallels broader concerns about AI-mediated information distortion, as Acemoglu, Ozdaglar, and Siderius (2025) demonstrate that AI-driven content recommendations and digital ads can create information bubbles and increase polarization. Our findings reveal that this optimism about AI-enhanced information processing may be misplaced: investors relying on AI agents are systematically exposed to training-data reliance, web-searching bias, and representative heuristic reasoning, potentially leading to widespread misinformed investment decisions.

Our work contributes to the emerging literature documenting various forms of AI bias in financial analysis. For example, a few studies document lookahead bias, where pretrained language models' outputs are contaminated by future information present in training data, leading to artificially inflated predictive performance (Sarkar and Vafa 2024; Levy 2025). Lopez-Lira, Tang, and Zhu (2025) further highlight the memorization problem where the so-called predictions directly come from LLMs' memory data. Glasserman and Lin (2023) also document a distraction effect, where LLMs mixed the general knowledge of the company mentioned with the sentiment of the underlying text. Cao, Wang, and Xiang (2025) find foreign bias, where US-based ChatGPT generates systematically more optimistic predictions about Chinese firms compared to China-based DeepSeek. Ross, Kim, and Lo (2024) demonstrate that LLMs exhibit economic behavior that falls between perfect rationality and typical human decision-making patterns. Chen, Green, Gulen, and Zhou (2024) document that LLMs display behavioral biases similar to human investors, particularly over-extrapolation from recent performance. While these studies focus on biases arising from existing training data, our findings reveal a distinct category of biases that emerge from AI agents' information-gathering and processing behaviors. This suggests that AI biases in finance operate through multiple complementary channels, requiring comprehensive frameworks to ensure reliable AI-assisted investment decisions.

We also contribute to understanding AI's broader impact on financial markets. Cheng, Lin, and Zhao (2025) document significant declines in trading volume during ChatGPT outages, while Dou, Goldstein, and Ji (2025) demonstrate that AI-powered algorithmic trading can facilitate market collusion and undermine efficiency. Our findings reveal a new channel through which AI may distort markets: training-data recommendations attract investor attention despite inferior performance and generate less informed retail trading. This pattern provides direct evidence that widespread AI agent adoption may harm market efficiency through misinformation propagation, as investors cannot distinguish between recommendations based on outdated training data versus current information.

Finally, from a technical perspective, our findings contribute to the AI hallucination literature, which focuses on LLMs generating plausible yet nonfactual content (Huang et al. 2025). Importantly, the three types of misinformation we identify (training-data reliance, web-searching bias, and representative heuristic reasoning) cannot be attributed to hallucination alone. These biases can occur even when AI agents report factually accurate information. For instance, training-data reliance may stem from AI agents privileging their extensive internal knowledge over limited external search results, regardless of factual accuracy. This distinction has important implications: while recent advances in AI technology, such as ChatGPT 5, dramatically reduce hallucination rates,[6] they may not address the systematic biases we document. Our findings suggest that simply getting smarter AI does not guarantee better financial decisions for humans. Understanding how AI agents collect and analyze information is crucial for ensuring reliable financial decision support.

2. Data collection and descriptive statistics

2.1 Data collection

Table 1 summarizes our data collection process. We ask GPT-4o and GPT-4, which are among the first LLMs that directly allow users to conduct web searching tasks. Our data collection from GPT 4 stopped on April 30, 2025, when OpenAI retired GPT 4. Through our data collection, we always ask six core questions: one question to ask LLMs to provide stock recommendations, followed by five questions to ask LLMs to update their recommendations. Since October 21, 2024, we have added two follow-up questions to check whether the first

[6] https://openai.com/index/introducing-gpt-5/

recommendation and the last recommendation are based on internal training data only. The prompts for these questions are provided in Table IA1. Table IA2 provides an example of GPT-4o's recommendation on November 13, 2024.

We conduct these inquiries after the closure of the US major stock exchanges and ask LLMs to recommend stocks for the next trading day. For example, we ask ChatGPT to provide a recommendation around 10 p.m. (New York time) on January 2nd, 2025. So the raw return of recommended stocks on January 3rd, 2025, is the return for day zero (i.e., D0), which is calculated as the closing price on January 3rd divided by the opening price on January 3rd, minus one. For the following trading days (e.g., D1 to D5), the raw return is calculated as the closing price of a day divided by the previous closing price, minus one. In our paper, the daily abnormal returns of recommended stocks are calculated as the daily raw return minus the corresponding value-weighted market return using CRSP data. This approach measures whether investors gain or lose compared to the counterfactual of simply investing in the market index.

In the data collection process, ChatGPT occasionally failed to provide complete responses (e.g., refusing to provide recommendations, providing fewer than 20 stocks, or not following the requested output format). In such cases, we resubmitted our request, but stopped after 10 unsuccessful attempts. Our sample does not include trading days from October 1-11, 2024, when ChatGPT's upgrade to a new version of GPT-4o temporarily prevented the model from responding to our prompts. To ensure ChatGPT generated stock recommendations based solely on our prompts rather than incorporating user-specific history, we disabled the "Reference saved memories" function throughout our data collection, which prevents ChatGPT from saving and using prior conversation memories when responding.

2.2 Descriptive statistics

We begin by examining whether ChatGPT successfully generates stock recommendations as requested. Table 2 Panel A shows that both models reliably provide tradeable stock recommendations. For initial recommendations, GPT-4o generates an average of 19.4 common stocks per day (out of 20 requested), with 0.4 American Depositary Receipts (ADRs). GPT-4 shows similar patterns, averaging 18.5 common stocks and 0.8 ADRs per day. Over-the-counter (OTC) stocks, non-stock recommendations (e.g., cryptocurrencies, funds, or industry categories) and untradeable stocks (e.g., not yet listed or already delisted) are rare for both

GPT-4o and GPT-4. These statistics demonstrate that AI agents successfully understand the task and provide actionable stock recommendations in nearly all cases.

Panel B lists the most frequently recommended stocks. Both models consistently favor large-cap technology and consumer stocks, with NVDA appearing in 95.9% of GPT-4o recommendations and 80.9% of GPT-4 recommendations, followed by MSFT, AMZN, TSLA, META, GOOGL, and AAPL. The overlap between GPT-4o and GPT-4 top recommendations suggests systematic preferences rather than random selection. Notably, the majority of frequently recommended stocks are S&P 100 constituents (shown in bold), indicating AI agents tend to recommend well-known, liquid large-cap stocks.

Figure A1 plots the number of URLs searched when generating initial stock recommendations over our sample period. Panel (a) shows GPT-4o's search behavior evolved substantially across three distinct phases: initially searching around 2 URLs per day, increasing to approximately 13 URLs following model updates, and later expanding to around 22 URLs per day. This pronounced upward trend demonstrates significant improvements in GPT-4o's information-gathering behavior over time, consistent with the two major model updates in November 2024 and May 2025. In contrast, Panel (b) shows GPT-4 maintained more stable search patterns, averaging around 4 URLs per day throughout the sample period.

Figure A2 presents word clouds showing the frequency of words used in AI agents' reasoning for stock recommendations. Both models predominantly use firm fundamental descriptors such as "growth," "technology," "innovation," "leadership," and "performance." Notably absent are valuation-focused terms like "undervalued" or "mispriced." This pattern is consistent with representative heuristic reasoning, where AI agents focus on company characteristics rather than investment opportunity metrics, effectively treating good firms as good investments regardless of valuation. The systematic absence of valuation considerations in the displayed reasoning provides early descriptive evidence of the misinformation channel we examine more rigorously in subsequent sections.

Taken together, these descriptive statistics reveal that ChatGPT is capable of providing seemingly sophisticated stock recommendations: it successfully identifies tradeable securities, searches multiple information sources (particularly for GPT-4o), and provides detailed reasoning emphasising business fundamentals. This apparent sophistication makes the

systematic misinformation we document in subsequent sections particularly concerning for investors who might rely on these seemingly reasonable recommendations.

3. AI agent misinformation

3.1 Training-data reliance

We first examine whether ChatGPT successfully uses web searching tools to collect real-time information for stock recommendations, or whether it relies on internal training data instead. To identify training-data reliance, we compare the websites ChatGPT searched after each inquiry with the sources cited for each stock recommendation. If a stock recommendation cites sources that were not actually searched, this indicates reliance on internal training data rather than real-time web searches. This method provides a conservative lower bound on training-data usage, as ChatGPT may also rely on its internal training data about websites that were actually searched, rather than using the current information retrieved from those searches.

Table 3 Panel A presents our main results on training-data reliance at both the stock-recommendation level and the daily level. For GPT-4o, 37.3% of individual stock recommendations cite unsearched sources, indicating substantial reliance on training data. At the daily level, GPT-4o relies on training data for an average of 7.4 stocks per day (out of 20), with high variability (SD = 8.0 stocks). For GPT-4, 29.7% of stock recommendations cite unsearched sources, averaging 5.9 stocks per day (SD = 7.8 stocks). Paradoxically, the more advanced GPT-4o model exhibits greater training-data reliance than GPT-4, demonstrating that models performing better on general tasks do not necessarily provide better financial advice.

Panel B examines an alternative one-shot measure of training-data reliance where we ask ChatGPT in a single follow-up prompt whether its recommendations relied on training data. For GPT-4o, 50.3% of recommendation days were self-declared as using training data, closely matching the 37.3% stock-level reliance documented in Panel A using source verification. This consistency argues against the alternative explanation that our source-verification measure merely captures citation hallucination: if AI agents were randomly citing sources, self-declaration should not corroborate source verification. Moreover, in Section 5.1, we show that recommendations citing unsearched sources systematically underperform those citing searched sources, further ruling out the citation hallucination explanation.

GPT-4 self-reported training-data use in only 4.6% of days, substantially lower than the 29.7% stock-level reliance found through source verification. This discrepancy likely reflects measurement noise in the self-declaration method: because the confirmation prompt asks about all 20 recommendations collectively, AI agents may misclassify the entire portfolio when only some stocks rely on training data while others use web searches. Despite this measurement limitation, the high rate of self-declared training-data use by GPT-4o corroborates our primary finding that the more advanced model relies heavily on internal knowledge rather than real-time information searches. Notably, this scalable self-declaration method reveals that training-data reliance persists even in ChatGPT-5, OpenAI's latest and most advanced model (untabulated), suggesting this problem extends beyond the models in our main sample.

These results reveal a fundamental limitation of AI agents for financial decision-making: even when equipped with web search capabilities and explicitly prompted to gather current information, they frequently default to potentially outdated training data. Importantly, ChatGPT cites sources for all recommendations, even those not based on searched information, making it difficult for users to detect when recommendations rely on training data rather than current market conditions.

3.2 Web-searching bias

When ChatGPT does use web searches rather than relying solely on training data, how does it gather information online? We address this question by systematically analyzing the websites searched and cited in stock recommendations. This analysis excludes individual stock recommendations that rely solely on training data, focusing only on cases where AI agents actually conducted web searches.

Table 4 Panel A examines search behavior at both the daily level and individual stock recommendation level. For GPT-4o, AI agents search an average of 9.9 URLs per day when generating 20 stock recommendations. However, when we condition on days with at least one URL searched (excluding days with no web searching), GPT-4o accesses an average of 5.4 unique websites per day, with substantial variation (SD = 4.7 websites). At the individual stock level, GPT-4o cites an average of 1.4 sources per recommended stock. For GPT-4, the pattern shows less intensive searching: 3.7 URLs searched per day overall, 3.3 unique websites per day when searches occur, and 1.1 sources cited per stock. These statistics reveal that while

GPT-4o searches more extensively than GPT-4, both models rely on relatively few distinct information sources when generating recommendations.

Panel B lists the 20 most frequently searched websites and calculates search concentration using the Herfindahl-Hirschman Index (HHI) based on the top 10 websites. For GPT-4o, searches concentrate heavily on three websites: Yahoo Finance (3.1 searches per day), Morningstar (1.7 searches per day), and Investopedia (1.3 searches per day). The next most-searched website, The Motley Fool, drops sharply to only 0.3 searches per day, creating a clear concentration in the top three sources. For GPT-4, searches distribute somewhat more evenly across seven primary websites: Bing (0.8 searches per day), Investopedia (0.7), Morningstar (0.5), NerdWallet (0.3), Fidelity Investments (0.3), Kiplinger (0.3), and Seeking Alpha (0.2). The HHI of 0.153 in GPT-4o's search patterns is higher than the HHI of 0.128 in GPT-4's search patterns, suggesting that more advanced GPT-4o have a higher concentration when searching websites.

This concentrated search behavior represents a significant form of misinformation. Concentrated information sources severely limit the effectiveness of information searches by reducing the diversity of perspectives and increasing susceptibility to source-specific biases (Xu, Xuan, and Zheng 2021). By disproportionately relying on a small set of websites, ChatGPT's search patterns limit the breadth and quality of information used to generate investment recommendations.

3.3 Representative heuristic reasoning

Beyond examining which information sources AI agents use, we systematically analyze the reasoning they provide for stock recommendations. Understanding this reasoning is crucial because investors may rely on AI explanations to inform their own investment decisions, even if they do not directly follow the specific stock recommendations. We investigate whether AI agents exhibit representative heuristic bias—treating good firms as good investments without considering current valuation—a well-documented pattern in behavioral finance (Tversky and Kahneman 1974; Lakonishok, Shleifer, and Vishny 1994; Barberis, Shleifer, and Vishny 1998).

Table 5 Panel A examines basic characteristics of the reasoning provided at the individual stock recommendation level. Both models rarely fail to provide reasoning, with less than 1% of recommendations lacking explanations. The reasoning appears substantive: GPT-4o

provides an average of 12.7 words per stock explanation (SD = 4.1 words), while GPT-4 provides an average of 7.7 words (SD = 2.4 words). The sentiment of reasoning, measured using the Loughran and McDonald (2014) financial dictionary, is predominantly positive for both models: 65.8% positive for GPT-4o and 55.7% positive for GPT-4. These statistics suggest AI agents provide detailed, optimistic explanations for their recommendations.

To understand the content of this reasoning, we classify it into 12 categories using a prompt in Table IA3. Panels B and C present these classifications separately for recommendations relying on training data (Panel B) and recommendations based on web-searched information (Panel C).

Panel B shows that for training-data recommendations, both models focus predominantly on firm fundamentals. The most frequent reasoning themes are: "Growth drivers & expansion" (55.6% for GPT-4o, 38.5% for GPT-4), "Technology & innovation leadership" (39.9% for GPT-4o, 32.0% for GPT-4), "Market leadership & competitive position" (27.4% for GPT-4o, 16.2% for GPT-4), "Financial performance & metrics" (24.9% for GPT-4o, 15.4% for GPT-4), and "Sector & industry tailwinds" (21.7% for GPT-4o, 16.0% for GPT-4). Critically, "Valuation opportunity & price dislocation", reasoning that considers whether current prices are attractive relative to fundamentals, appears in only 5.1% of GPT-4o recommendations and 3.6% of GPT-4 recommendations. Panel C reveals nearly identical reasoning patterns for web-searched recommendations. "Growth drivers & expansion" (56.7% for GPT-4o, 40.4% for GPT-4), "Technology & innovation leadership" (41.1% for GPT-4o, 31.3% for GPT-4), "Market leadership & competitive position" (22.9% for GPT-4o, 15.2% for GPT-4), "Financial performance & metrics" (22.7% for GPT-4o, 14.7% for GPT-4), and "Sector & industry tailwinds" (23.1% for GPT-4o, 20.2% for GPT-4) again dominate the reasoning. "Valuation opportunity & price dislocation" remains rare: 4.8% for GPT-4o and 3.6% for GPT-4.

These results reveal two fundamental problems in AI agents' investment reasoning. First, by systematically ignoring valuation considerations and focusing almost exclusively on firm fundamentals, AI agents exhibit the same representative heuristic that leads human investors to underperform (Lakonishok, Shleifer, and Vishny 1994). If the market has already incorporated positive firm characteristics into stock prices, recommendations based solely on these characteristics will not generate positive returns. Second, the striking similarity between

reasoning patterns in Panels B and C suggests that investors cannot identify training-data reliance from AI-provided explanations alone, as the reasoning appears equally sophisticated regardless of information source.

4. Can AI agents reduce their own misinformation?

4.1 Do self-revisions change recommendations?

Having documented that initial recommendations suffer from three forms of misinformation, we now examine whether prompting AI agents to revise their recommendations addresses these problems. After collecting initial recommendations, we asked AI agents five times to update their stock lists by searching additional websites. We analyze whether these sequential updates generate meaningful changes in recommendations and whether they reduce misinformation.

Table 6 reports changes in stock recommendations across sequential updates. Panel A examines how many stocks differ from the initial recommendation. For GPT-4o, the median number of stock changes remains at zero for the first update, indicating that in more than half of cases, GPT-4o made no changes to its initial recommendations. Stock changes gradually increase with subsequent updates: mean changes rise from 1.3 in the first update to 1.9 stocks in the fifth update. This pattern reveals substantial inertia in GPT-4o's recommendations despite being explicitly prompted to update them. In contrast, GPT-4 shows more willingness to revise: the mean number of stock changes increases from 4.1 in the first update to 9.0 in the fifth update. Panel B examines ranking changes: how many stock positions differ from the initial recommendation, capturing both new stocks entering the list and existing stocks changing ranks. For GPT-4o, mean ranking changes increase from 4.3 for the first update to 5.9 in the fifth update. For GPT-4, ranking changes are far more dramatic: mean changes range from 10.3 to 15.8 positions across updates.

GPT-4's substantially higher revision rates suggest the more advanced GPT-4o model exhibits greater confidence (or stubbornness) in its initial recommendations. This is consistent with the more advanced model having stronger priors or higher confidence in its initial assessments.

4.2 Do self-revisions reduce misinformation?

We now examine whether the five sequential updates reduce the three channels of misinformation documented in Section 3. We focus on the fifth (final) updated recommendations, as these represent AI agents' most refined advice after being prompted multiple times to search additional websites and revise their stock lists.

Table 7 Panel A shows that training-data reliance remains prevalent even after five updates. For GPT-4o, 40.5% of individual stock recommendations in the fifth update cite unsearched sources, compared to 37.3% in initial recommendations (Table 3). At the daily level, GPT-4o relies on training data for an average of 8.1 stocks per day in the fifth update (SD = 7.9 stocks), slightly higher than the 7.4 stocks in initial recommendations. For GPT-4, the pattern is even more concerning: 48.2% of stock recommendations in the fifth update cite unsearched sources, substantially higher than the 29.7% in initial recommendations. At the daily level, GPT-4's training-data reliance increases from 5.9 stocks in initial recommendations to 9.5 stocks in the fifth update (SD = 8.2 stocks). These results demonstrate that prompting AI agents to update their recommendations not only fails to reduce training-data reliance but exacerbates it, particularly for GPT-4. Panel B examines the less-accurate self-declared measure, revealing a puzzling pattern: GPT-4o self-reports using training data in 31.0% of days for the fifth update compared to 50.3% initially, while GPT-4 reports 0.9% compared to 4.6% initially. This discrepancy between the source-verification measure (Panel A) and self-declaration measure (Panel B) suggests that AI agents may become less accurate in assessing their own information sources after multiple updates, though the noisiness of the daily-level self-declaration measure limits strong conclusions.

Table 8 examines whether sequential updates lead to more diverse information gathering. Panel A shows that while GPT-4o increases its search activity (from an average of 9.9 URLs per day initially to 12.3 URLs in the fifth update), this does not translate to meaningfully broader information sources: when conditioning on days with at least one URL searched, GPT-4o accesses an average of only 4.1 unique websites per day in the fifth update, substantially lower than the 5.4 websites in initial recommendations. For GPT-4, search activity actually declines: 3.3 URLs searched per day in the fifth update compared to 3.7 initially, with 2.8 unique websites accessed compared to 3.3 initially. Panel B reveals that concentrated search patterns not only persist but intensify. For GPT-4o, the top three websites—Yahoo Finance

(4.4 searches per day), Morningstar (1.7 searches per day), and Investopedia (1.4 searches per day)—continue to dominate, with search concentration (HHI = 0.196) substantially higher than initial recommendations (HHI = 0.153). For GPT-4, the concentration pattern shifts: the top five websites now stand out instead of top seven, with HHI increasing sharply from 0.128 to 0.208. These patterns demonstrate that AI agents' inherent bias toward some information sources not only persists but intensifies with sequential updates, suggesting that repeated prompting reinforces rather than diversifies their search behavior.

Representative heuristic reasoning remains unchanged in updated recommendations. Table 9 examines the reasoning provided in the fifth update. Panel A shows that both models continue to provide detailed explanations: GPT-4o averages 13.8 words per stock (SD = 4.3 words), close to the 12.7 words initially, while GPT-4 averages 8.1 words (SD = 2.0 words) compared to 7.7 words initially. The sentiment remains predominantly positive: 67.1% for GPT-4o and 60.6% for GPT-4, consistent with initial recommendations. Most critically, Panels B and C reveal that the fundamental flaw in reasoning persists. For training-data recommendations (Panel B), "Valuation opportunity & price dislocation" appears in only 5.3% of GPT-4o's fifth update compared to 5.1% initially, and in only 0.8% of GPT-4's fifth update compared to 3.6% initially. The reasoning continues to emphasize firm fundamentals: "Growth drivers & expansion" (61.1% for GPT-4o, 40.0% for GPT-4), "Technology & innovation leadership" (41.9% for GPT-4o, 39.3% for GPT-4), and "Market leadership & competitive position" (26.6% for GPT-4o, 21.0% for GPT-4). For web-searched recommendations (Panel C), the pattern is similar: "Valuation opportunity & price dislocation" appears in only 3.9% of GPT-4o's reasoning and 2.3% of GPT-4's reasoning, while firm fundamental themes continue to dominate.

Taken together, these results reveal that the three channels of misinformation persist, and sometimes intensify, in self-revised recommendations despite AI agents being explicitly prompted five times to search additional websites and update their stock lists. This suggests that the misinformation we document reflects fundamental limitations in how AI agents gather and process information for financial decision-making, rather than correctable oversights that can be addressed through simple iterative prompting.

4.3 Do model updates mitigate misinformation?

Having documented that self-revision fails to reduce misinformation, we now examine whether technological improvements can address these systematic biases. A natural expectation is that more advanced AI models should exhibit less misinformation—after all, if model updates improve reasoning capabilities and information processing, they should also improve the quality of financial recommendations. During our sample period, ChatGPT experienced two major model updates: in November 2024 and May 2025. We examine whether training-data reliance, web-searching bias, and representative heuristic reasoning diminish across these model generations. If the misinformation we document stems from correctable technical limitations rather than fundamental architectural issues, these model upgrades should substantially reduce or eliminate all three channels. Conversely, if misinformation persists despite these technological advances, this would suggest deeper structural problems in how AI agents gather and process financial information—problems that cannot be solved through incremental model improvements alone.

Table 10 examines whether model updates reduce training-data reliance. We split the sample into three periods based on major ChatGPT upgrades: before October 2024, from November 2024 to April 2025, and after May 2025. Panel A shows the ratio of individual stock recommendations relying on training data; Panel B shows the number of stocks per day (out of 20) relying on training data.

The results reveal that model updates alleviate but do not eliminate training-data reliance. For GPT-4o, the ratio of stocks citing unsearched sources drops substantially from 56.3% before October 2024 to 24.4% in the middle period and 23.3% after May 2025. At the daily level, GPT-4o's training-data reliance decreases from an average of 11.2 stocks per day to 4.9 stocks in the middle period and 4.7 stocks after May 2025. For GPT-4, training-data reliance also declines from 34.8% before October 2024 to 25.1% in the middle period, with the daily count dropping from 6.8 stocks to 5.0 stocks per day.

Despite these improvements, training-data reliance remains substantial even in the most recent period. After the May 2025 update, nearly one-quarter (23.3%) of GPT-4o's stock recommendations still cite unsearched sources, indicating persistent reliance on potentially outdated internal knowledge. This pattern demonstrates that while technological progress reduces the severity of this misinformation channel, it does not eliminate the fundamental

problem: AI agents continue to use internal training data instead of current information when explicitly prompted to search the web.

Table 11 examines whether model updates affect web-searching bias. Panel A reports search behavior metrics across the three time periods. Panel B shows the concentration of searches across the top 20 most-searched websites, with the Herfindahl-Hirschman Index (HHI) measuring search concentration based on the top 10 websites.

The results reveal a non-monotonic pattern in web-searching behavior. For GPT-4o, search intensity increases dramatically after the November 2024 update: the number of URLs searched per day jumps from 1.8 before October 2024 to 13.3 in the middle period, and further increases to 22.2 after May 2025. The number of unique websites accessed also rises substantially, from 3.0 before October 2024 to 10.6 after May 2025 (conditioning on days with searches). Despite this increased search activity, web-searching bias initially worsens before improving. The HHI increases from 0.146 before October 2024 to 0.240 in the middle period, indicating more concentrated search patterns despite searching more URLs. However, the May 2025 update reverses this trend, with HHI dropping to 0.093, suggesting more diversified information gathering.

Panel B illustrates this pattern through the evolution of frequently searched websites. Before October 2024, GPT-4o distributed searches relatively evenly across sources, with Bing (0.545 searches per day) as the most-used website. After the November 2024 update, searches concentrated heavily on Yahoo Finance (5.591 searches per day), Morningstar (2.574), and Investopedia (1.939), with a sharp drop to the fourth website. After May 2025, while Yahoo Finance (4.463) and Morningstar (3.902) remain dominant, the distribution becomes more balanced, with multiple websites receiving substantial search activity.

For GPT-4, the pattern is more stable. Search activity remains modest across all periods (3.3 to 4.1 URLs per day), and concentration stays relatively low (HHI ranging from 0.130 to 0.138). The most-searched websites are Investopedia, Bing, and Morningstar in both periods, without the dramatic concentration spike observed in GPT-4o.

These results demonstrate that model updates can both help and hurt web-searching bias, with GPT-4o experiencing an initial deterioration before improvement. Even in the most recent period, however, GPT-4o still concentrates searches on a few dominant sources (Yahoo

Finance, Morningstar, and Investopedia), suggesting that diversified information gathering remains a challenge for AI agents.

Table 12 examines whether model updates affect representative heuristic reasoning. Panel A analyzes basic characteristics of the reasoning provided. Panels B and C examine reasoning themes separately for training-data recommendations (Panel B) and web-searched recommendations (Panel C), as in Table 5.

Panel A shows that AI agents consistently provide reasoning across all periods, with failure rates near zero. The reasoning remains substantive in length, averaging 11-14 words for GPT-4o and 7-8 words for GPT-4 across periods. Sentiment remains predominantly positive throughout, ranging from 54% to 70% positive for both models. These basic characteristics show little change across model updates, suggesting stable presentation quality even as the underlying models evolve.

The critical question is whether model updates reduce representative heuristic bias by increasing attention to valuation considerations. Panels B and C reveal that valuation reasoning remains minimal across all periods. For training-data recommendations (Panel B), "Valuation opportunity & price dislocation" appears in only 5.2% of GPT-4o recommendations before October 2024, 5.0% in the middle period, and 4.2% after May 2025. For GPT-4, the pattern is similar: 2.4% initially, 5.0% in the middle period. For web-searched recommendations (Panel C), valuation considerations appear similarly infrequent: GPT-4o shows 2.5% before October 2024, 4.7% in the middle period, and 8.5% after May 2025; GPT-4 shows 3.2%, 4.0% across the two periods. Notably, the increase in valuation reasoning for GPT-4o's web-searched recommendations after May 2025 (from 4.7% to 8.5%) represents the only meaningful improvement observed, yet even this leaves valuation considerations appearing in fewer than one in ten recommendations.

In contrast, firm fundamental themes continue to dominate reasoning across all periods. "Growth drivers & expansion" consistently ranks first (ranging from 37.5% to 61.9%), followed by "Technology & innovation leadership" (ranging from 30.7% to 48.0%) and other fundamental characteristics. The persistence of these patterns across major model updates demonstrates that representative heuristic reasoning is deeply embedded in how AI agents

analyze investments, reflecting fundamental limitations that cannot be resolved through incremental technological improvements.

5. Market consequences of AI agent misinformation

5.1 Investment performance and training-data reliance

Having documented three channels of AI agent misinformation and shown that neither self-revision nor model updates eliminate these biases, we now examine their market consequences. Among the three misinformation channels we identify, training-data reliance represents the most unambiguous form of misinformation: using outdated internal knowledge instead of current market information, when explicitly prompted to search the web, fundamentally contradicts the purpose of deploying AI agents for real-time financial decision-making. We therefore focus our main analysis on whether recommendations rely on training data versus web-searched information. In supplementary analyses, we also examine whether web-searching bias and representative heuristic reasoning affect investment performance among web-searched recommendations.

Table 13 reports value-weighted abnormal returns for portfolios of stocks from initial recommendations, partitioned by whether AI agents relied on internal training data. Panel A uses our primary source-verification method; Panel B uses the alternative self-declaration method.

The results reveal that training-data reliance systematically undermines investment performance. Panel A shows that for GPT-4o, web-searched recommendations generate 5-day cumulative abnormal returns of 21.0 bps compared to only 6.3 bps for training-data recommendations. For GPT-4, the pattern is even more pronounced: web-searched recommendations generate 3.8 bps while training-data recommendations lose 30.9 bps. Interestingly, GPT-4o's training-data recommendations show initial Day 1 outperformance (16.0 bps versus 8.9 bps for web-searched), but this advantage quickly dissipates and reverses over Days 2-5. One explanation could be that AI recommendations attract an investor attention jump, as documented in Section 5.2, which creates temporary price pressure that subsequently reverses. Figure A3 visualizes these cumulative return patterns over the 5-day horizon.

Panel B, using the noisier self-declaration method, corroborates these findings. For GPT-4o, portfolios from days when AI agents self-declared training data use show 5-day returns of

-25.8 bps (SD = 179.2 bps) compared to 37.8 bps (SD = 231.1 bps) when they reported using web searches. For GPT-4, the pattern is even more stark: -57.8 bps versus 2.9 bps over Days 1-5. Figure A4 visualizes these self-declared patterns over the 5-day horizon.

These results demonstrate that the most egregious form of misinformation, relying on outdated training data, produces inferior returns. The finding that web-searched recommendations outperform training-data recommendations establishes that our framework can identify problematic AI recommendations. However, all portfolios exhibit extremely high volatility relative to mean returns. Standard deviations range from 179.2 bps to 372.8 bps across the various classifications, while mean 5-day returns range from -57.8 bps to 37.8 bps. This extreme volatility, present even in web-searched recommendations, suggests that investors should exercise caution when relying on AI agent stock recommendations.

Tables IA4 and IA5 in the Internet Appendix examine whether the other two misinformation channels—web-searching bias and representative heuristic reasoning—also affect investment performance among web-searched recommendations. Table IA4 shows that for GPT-4o, concentrated search days generate stronger Day 1 returns (27.7 bps versus 5.1 bps for diverse search days) but lower 5-day performance (16.8 bps versus 21.9 bps). Standard deviations remain extremely high for both groups (231.0 bps for concentrated searches versus 221.0 bps for diverse searches). For GPT-4, diverse search days consistently outperform over the 5-day horizon (16.2 bps versus -4.4 bps), though both exhibit substantial volatility (standard deviations of 489.5 bps and 266.0 bps, respectively). These patterns suggest that web-searching bias harms investment performance.

Table IA5 shows that for GPT-4o, valuation-focused recommendations generate higher 5-day cumulative returns (42.6 bps versus 15.4 bps for non-valuation reasoning), suggesting that representative heuristic reasoning undermines investment performance. For GPT-4, the patterns are less clear: valuation-focused recommendations underperform over the 5-day horizon (-13.4 bps versus 7.6 bps), with volatility remaining extreme for both groups. This inconsistency may reflect GPT-4's weaker reasoning capabilities, as GPT-4 cites valuation considerations in only 3.6% of web-searched recommendations compared to 4.8% for GPT-4o.[7]

[7] OpenAI-reported results show GPT-4o improves over GPT-4 on reasoning-heavy benchmarks, including MATH

The consistently high volatility across all categories reinforces that AI agent recommendations carry substantial risk regardless of the specific misinformation channel. Combined with our training-data findings, our results demonstrate that all three misinformation channels we identify consistently harm investment performance across both models, with one exception: GPT-4's valuation-focused recommendations show an inconsistent pattern that may reflect its weaker reasoning capabilities. These findings validate our framework for detecting problematic AI recommendations.

5.2 Investor attention

Having established that training-data recommendations underperform, we examine whether these flawed recommendations nevertheless attract investor attention. If investors cannot distinguish between recommendations based on outdated training data versus current information, both types should generate similar attention patterns despite their performance differences.

Figures 1 and 2 plot investor attention around recommendation dates, measured by Google search volume index (SVI) and Twitter discussions, respectively. Both figures partition stocks by whether AI agents relied on training data versus web-searched information. Day 0 represents the trading day when recommendations are made, and the x-axis shows trading days relative to the recommendation date.

The results reveal striking patterns. Figure 1 shows that Google search volume spikes sharply following recommendations, peaking around Day 1 for both training-data and web-searched recommendations across both models. For GPT-4o, median SVI increases from approximately 32 at Day -2 to 40 at Day 1 for training-data recommendations (Panel a), and from approximately 33 at Day -2 to 41 at Day 1 for web-searched recommendations (Panel c). For GPT-4, the pattern is similar: median SVI rises from approximately 28 at Day -2 to 36 at Day 1 for training-data recommendations (Panel b), and from approximately 30 at Day -2 to 36 at Day 1 for web-searched recommendations (Panel d). Importantly, the magnitude of attention spikes appears comparable between training-data and web-searched

(mathematical problem-solving: 76.6% vs 42.5%) and GPQA (graduate-level multi-step reasoning: 53.6% vs 35.7%). Source: https://openai.com/index/hello-gpt-4o/

recommendations, suggesting investors cannot distinguish between the two information sources based on the recommendations themselves.

Figure A5 in the Internet Appendix confirms these patterns using regression analysis. Controlling for market capitalization, news coverage, news sentiment, stock returns, volatility, and turnover with event*firm fixed effects, we find statistically significant increases in search volume on Days 0 to 2 relative to Day -1 for both training-data and web-searched recommendations across both models. The coefficient magnitudes align closely with the raw patterns in Figure 1, demonstrating that the investor attention spikes are robust to controlling for other determinants of search activity.

Figure 2 shows similar patterns for social media discussions on Twitter. Both training-data and web-searched recommendations trigger substantial increases in Twitter posts, peaking around Day 1. For GPT-4o, median Twitter discussions increase from approximately 190 posts at Day -2 to 320 posts at Day 1 for training-data recommendations (Panel a), and from approximately 120 posts at Day -2 to 195 posts at Day 1 for web-searched recommendations (Panel c). For GPT-4, median discussions rise from approximately 180 posts at Day -2 to 320 posts at Day 1 for training-data recommendations (Panel b), and from approximately 100 posts at Day -2 to 165 posts at Day 1 for web-searched recommendations (Panel d). The attention spikes appear substantial for both types of recommendations, though training-data recommendations generate somewhat larger increases in absolute terms. Regression analysis in Figure A6 of the Internet Appendix corroborates these Twitter activity patterns. After controlling for market capitalization, news coverage, news sentiment, stock returns, volatility, and turnover with event*firm fixed effects, Twitter discussion volumes remain significantly elevated on Days 0 to 2 relative to Day -1 for both recommendation types across both models, with coefficient estimates matching the magnitudes observed in Figure 2.

Notably, both Google search volume and Twitter discussions decline relatively quickly after Day 1, returning toward baseline levels by Days 4-5. The combination of comparable attention spikes across both information sources, despite the performance differences documented in Section 5.1, suggests that investors cannot distinguish training-data recommendations from web-searched recommendations. This inability to identify misinformed AI recommendations has important implications for market efficiency and investor protection.

5.3 Retail trading and return predictability

The attention spikes documented in Section 5.2 raise a natural question: do they translate into actual trading activity, and if so, is this trading informed? We examine the return predictability of retail trading around recommendation dates to assess whether the trading activity attracted by AI recommendations contains information about future returns.

Figure 3 plots the return predictability of retail trading for stocks from initial recommendations, partitioned by whether AI agents relied on training data versus web-searched information. For each trading day relative to the recommendation date, we estimate the coefficient from a cross-sectional regression of day-t stock returns on day-t-1 retail trading, controlling for stock characteristics. Control variables include the natural logarithm of market capitalization on day t-1, the standard deviation of daily returns over trading days t-20 to t-1, average daily share turnover over trading days t-20 to t-1, the stock return on day t-1, the cumulative daily return over trading days t-6 to t-2, and the cumulative daily return over trading days t-16 to t-7. We also control for day fixed effects. We measure retail trading using marketable retail order imbalance (MROIB), following Barber, Huang, Jorion, Odean, and Schwarz (2024). Day 0 represents the trading day when recommendations are made. Dots represent coefficient estimates; vertical lines denote 90% confidence intervals.

The results reveal that retail trading following training-data recommendations exhibits weaker return predictability. For both models, web-searched recommendations show stronger predictability at certain horizons compared to training-data recommendations. For GPT-4o (Panels a versus c), the difference is most apparent on Days 2 and 4: training-data recommendations generate significant negative predictability, while web-searched recommendations show no significant predictability at these horizons. For GPT-4 (Panels b versus d), web-searched recommendations exhibit significantly positive predictability on Day 1, while training-data recommendations show near-zero predictability. This consistent pattern indicates that misinformation from training-data reliance attracts less informed trading activity.

Combined with the attention patterns (Section 5.2) and performance results (Section 5.1), these findings reveal that AI agent misinformation generates a cascade of market consequences: training-data recommendations attract similar attention despite inferior performance and stimulate trading that lacks predictive power for future returns.

6. Conclusion

This paper provides early evidence on AI agent misinformation when assisting financial decision-making. Using daily stock recommendations collected from ChatGPT between May 2024 and June 2025, we identify three distinct channels through which AI agent misinformation manifests: training-data reliance, web-searching bias, and representative heuristic reasoning. Model updates alleviate but do not eliminate these biases, while prompting AI agents to revise their recommendations sometimes worsens misinformation. We demonstrate that recommendations relying on training data systematically underperform web-searched recommendations over a 5-day horizon, yet both attract similar investor attention spikes, suggesting investors cannot distinguish between the two information sources. Moreover, retail trading following training-data recommendations exhibits weaker return predictability, indicating this trading is less informed.

Our findings reveal that the enhanced capabilities of AI agents—specifically their ability to autonomously gather and process real-time information—paradoxically introduce new forms of systematic bias rather than improving investment advice quality. This distinction between AI as an expert (analyzing human-provided information) and AI as an agent (autonomously collecting information) has important implications for understanding when and how AI can reliably assist financial decision-making.

Our findings carry significant implications for financial regulators and policymakers concerned with investor protection and market integrity. First, existing regulatory frameworks for investment advisors may need updating to address AI agents' unique misinformation channels. Unlike human advisors or traditional robo-advisors that follow predetermined rules, AI agents exhibit systematic biases in information gathering and processing that are difficult to detect without specialized frameworks like ours. Regulators should consider requiring disclosure when financial advice involves AI agents with autonomous information-gathering capabilities, distinguishing these systems from simpler AI tools. Second, beyond disclosure requirements, financial literacy programs should educate investors about the specific limitations of AI agent advice, particularly the three misinformation channels we identify. Critically, the distinction between reducing hallucination rates (a technical AI problem that recent model improvements have addressed) and addressing systematic misinformation (a

behavioral AI problem that model updates alone cannot solve) is essential for both investor education and effective regulation.

This study opens several avenues for future research. First, while we focus on stock recommendations, the three misinformation channels we identify (training-data reliance, web-searching bias, and representative heuristic reasoning) likely affect AI agents across other financial domains. Future work should examine whether similar patterns emerge in AI agent assistance for portfolio allocation, risk management, credit analysis, or corporate financial decision-making. Understanding the generalizability of misinformation across financial contexts would help develop comprehensive frameworks for AI agent reliability.

Second, our finding that AI agents cannot effectively self-correct their misinformation suggests the need for external validation mechanisms. Future research could explore what types of interventions, such as human oversight, alternative AI architectures, or hybrid human-AI systems, can mitigate the misinformation we document. Additionally, as AI technology continues evolving rapidly, longitudinal studies tracking how misinformation channels change across model generations would provide valuable insights for both researchers and practitioners.

Finally, comparative studies examining different LLMs (e.g., ChatGPT vs. DeepSeek, ChatGPT vs. Gemini) would shed light on whether different AI agents exhibit the same types of misinformation channels we identify but to varying degrees, or whether fundamentally different misinformation patterns emerge across models. Understanding how misinformation manifests across different AI systems is crucial for developing robust frameworks for reliable financial decision support.

References

Acemoglu, D., Ozdaglar, A., and Siderius, J., 2025. AI and social media: A political economy perspective. NBER Working Paper 33892. https://www.nber.org/papers/w33892

Barber, B. M., Huang, X., Jorion, P., Odean, T., and Schwarz, C. 2024. A (sub) penny for your thoughts: Tracking retail investor activity in TAQ. The Journal of Finance, 79(4), 2403-2427.

Barberis, N., Shleifer, A., and Vishny, R. 1998. A model of investor sentiment. Journal of Financial Economics, 49(3), 307-343.

Beckmann, L., Beckmeyer, H., Filippou, I., Menze, S., and Zhou, G., 2025. Unusual financial communication: ChatGPT, earnings calls, and financial markets. Working paper. https://ssrn.com/abstract=4699231.

Bertomeu, J., Lin, Yu., Liu, Y., and Ni, Z., 2025. The impact of generative AI on information processing: Evidence from the ban of ChatGPT in Italy. Journal of Accounting and Economics 80(1), 101782.

Bhagwat, V., Cookson, J. A., Dim, C., and Niessner, M., 2025. The market's mirror: Revealing investor disagreement with LLMs. Working paper. https://ssrn.com/abstract=5375473

Blankespoor, E., Croom, J., and Grant, S. M., 2024. Generative AI and investor processing of financial information. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5053905

Bybee, L., 2025. The ghost in the machine: Generating beliefs with large language models. Working paper. https://lelandbybee.com/files/LLM.pdf

Cao, S., Wang, C. C. Y., Xiang, Y., 2025. When LLMs go abroad: Foreign bias in AI financial predictions. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5440116

Chak, I., Croxson, K., D'Acunto, F., Reuter, J., Rossi, A., and Shaw, J. M., 2022. Improving household debt management with robo-advice. NBER Working Paper 30616. https://www.nber.org/system/files/working_papers/w30616/w30616.pdf

Chen, Green, Gulen, and Zhou, 2024. What does ChatGPT make of historical stock returns? Extrapolation and miscalibration in LLM stock return forecasts. Working paper. https://arxiv.org/abs/2409.11540

Chen, Y., Kelly, B., and Xiu, D., 2024. Expected returns and large language models. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4416687

Chen, J., Tang, G., Zhou, G., and Zhu, W., 2025. ChatGPT and Deepseek: Can they predict the stock market and macroeconomy? Working paper. https://arxiv.org/abs/2502.10008

Cheng, Q., Lin, P., and Zhao, Y., 2025. Does generative AI facilitate investor Trading? Early evidence from ChatGPT outages. Journal of Accounting and Economics, 101821.

Cookson, J. A., Lu, R., Mullins, W., and Niessner, M., 2024. The social signal. Journal of Financial Economics, 158, 103870.

Da, Z., Engelberg, J., and Gao, P., 2011. In search of attention. Journal of Finance, 66(5), 1461-1499.

Dou, W. W., Goldstein, I., and Ji, Y., 2025. AI-powered trading, algorithmic collusion, and price efficiency. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4452704

Fedyk, A., Kakhbod, A., Li, P., and Malmendier, U., 2024. AI and perception biases in investments: An experimental study. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4787249

Glasserman, P., and Lin, C., 2023. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. Working paper. https://arxiv.org/abs/2309.17322

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... and Liu, T. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55.

Jha, M., Qian, J., Weber, M., and Yang, B., 2025. ChatGPT and corporate policies. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4521096

Kim, A. G., Muhn, M., and Nikolaev, V. V., 2024a. Bloated disclosures: Can ChatGPT help investors process information? Working paper. https://ssrn.com/abstract=4425527.

Kim, A. G., Muhn, M., and Nikolaev, V. V., 2024b. From transcripts to insights: Uncovering corporate risks using generative AI. Working paper. https://ssrn.com/abstract= 4593660.

Lakonishok, J., Shleifer, A., and Vishny, R. W. 1994. Contrarian investment, extrapolation, and risk. The Journal of Finance, 49(5), 1541-1578.

Levy, B., 2025. Caution ahead: Numerical reasoning and look-ahead bias in AI models. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5082861

Li, K., Mai, F., Shen, R., Yang, C., and Zhang, T., 2025. Dissecting corporate culture using generative AI. Review of Financial Studies forthcoming.

Lopez-Lira, A., and Tang, Y., 2025. Can ChatGPT forecast stock price movements? Return predictability and large language models. Working paper. https://ssrn.com/abstract= 4412788.

Lopez-Lira, A., Tang, Y., and Zhu, M., 2025. The memorization problem: Can we trust LLMs' economic forecasts? Working paper. https://arxiv.org/abs/2504.14765

Ouyang, S., Yun, H., and Zheng, X., 2024. AI as decision-maker: Ethics and risk preferences of LLMs. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4851711

Pelster, M., and Val, J., 2024. Can ChatGPT assist in picking stocks? Financial Research Letters 59, 104786.

Reher, M., and Sokolinski, S., 2024. Robo advisors and access to wealth management. Journal of Financial Economics 155, 103829.

Ross, J., Kim, Y., and Lo, A. W., 2024. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. Working paper. https://arxiv.org/abs/2408.02784

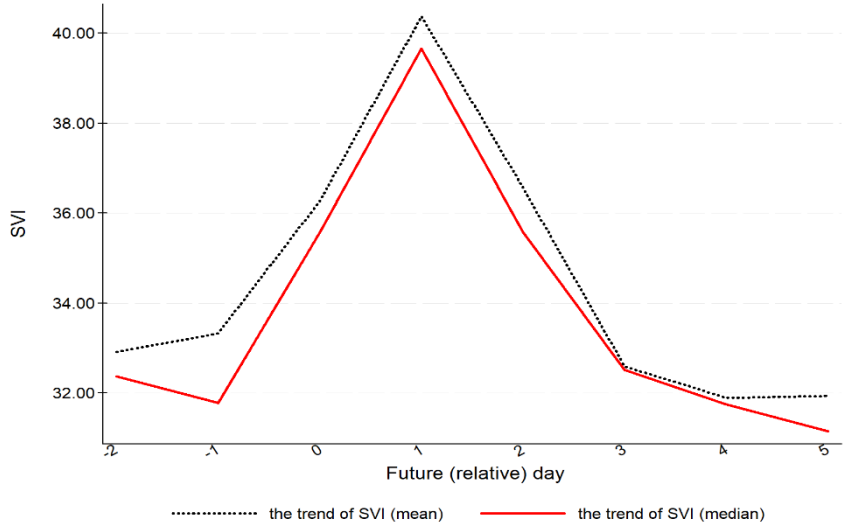
Rossi, A. G., and Utkus, S., 2024. The diversification and welfare effects of robo-advising. Journal of Financial Economics 157, 103869.

Sarkar, S. K., and Vafa, K., 2024. Lookahead bias in pretained language models. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4754678

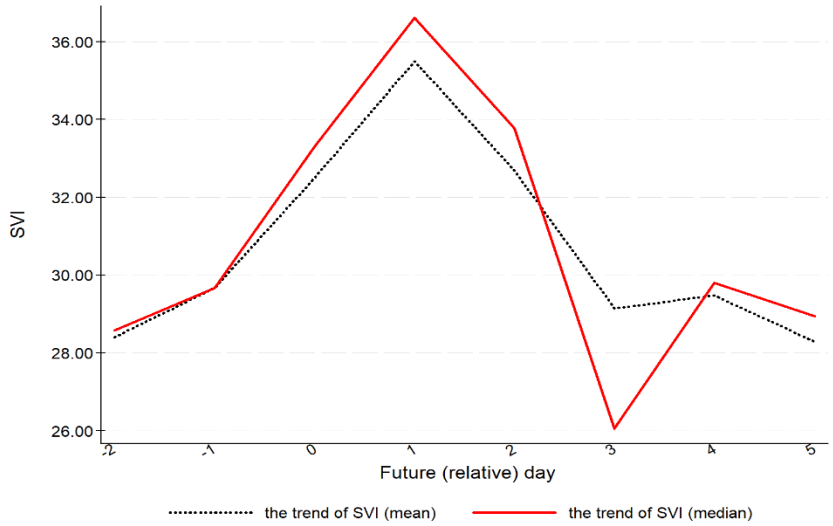
Tversky, A., and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. Science, 185(4157), 1124-1131.

Figure 1: Google search volume around recommendation dates

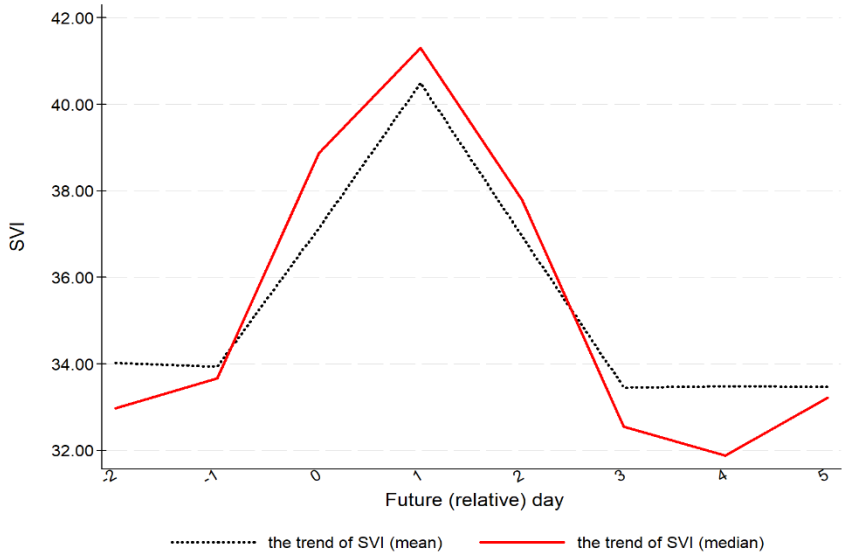
This figure plots the Google search volume index (SVI) for stocks from initial recommendations, partitioned by whether AI agents relied on internal training data. For each trading day, we first calculate the average SVI across all stocks within each category (training-data versus web-searched). We then compute the median and mean of these daily averages across all recommendation dates. The solid line shows the median SVI; the dotted line shows the mean SVI. Day 0 represents the trading day when AI agents provide stock recommendations. The x-axis shows trading days relative to the recommendation date. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. Panels (a) and (b) examine stocks relying on training data; Panels (c) and (d) examine stocks relying on web-searched information. Panels (a) and (c) show GPT-4o results; Panels (b) and (d) show GPT-4 results.



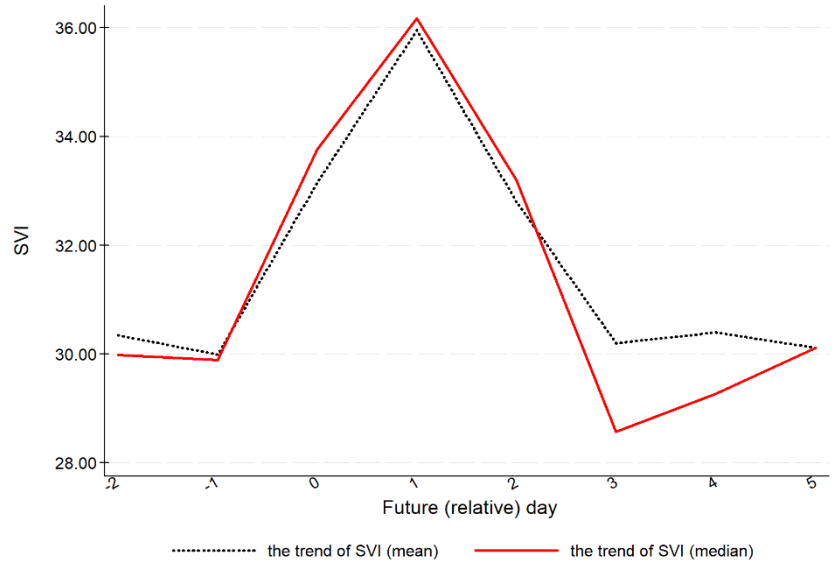
(a) GPT-4o (training)



(b) GPT-4 (training)



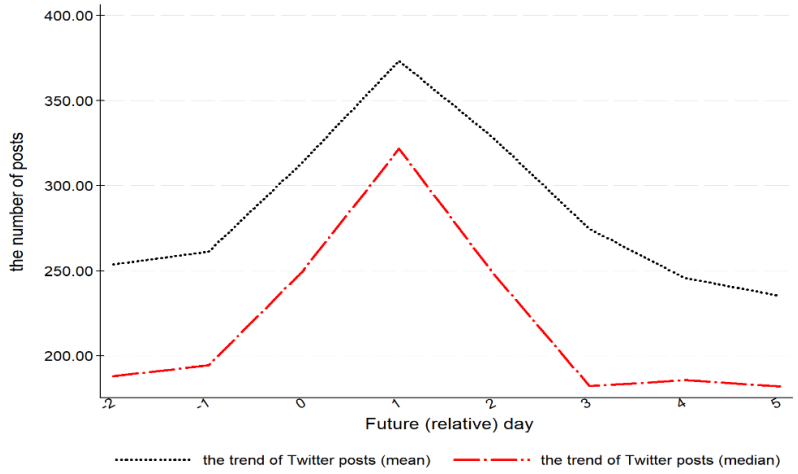
(c) GPT-4o (non-training)



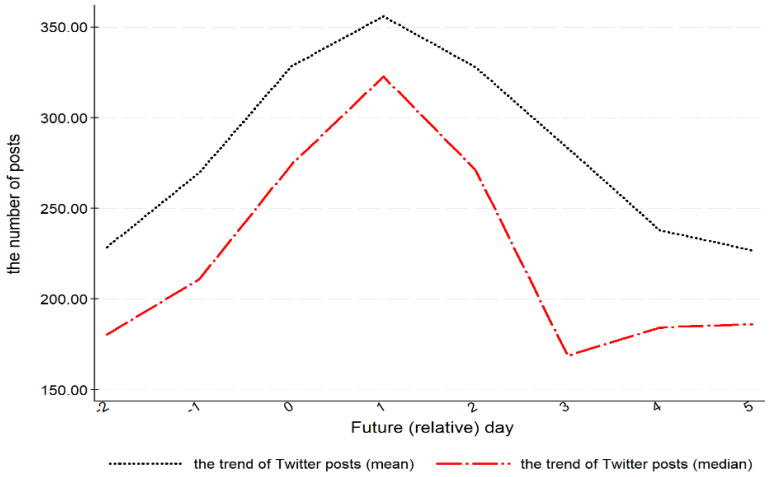
(d) GPT-4 (non-training)

Figure 2: Social media attention around recommendation dates

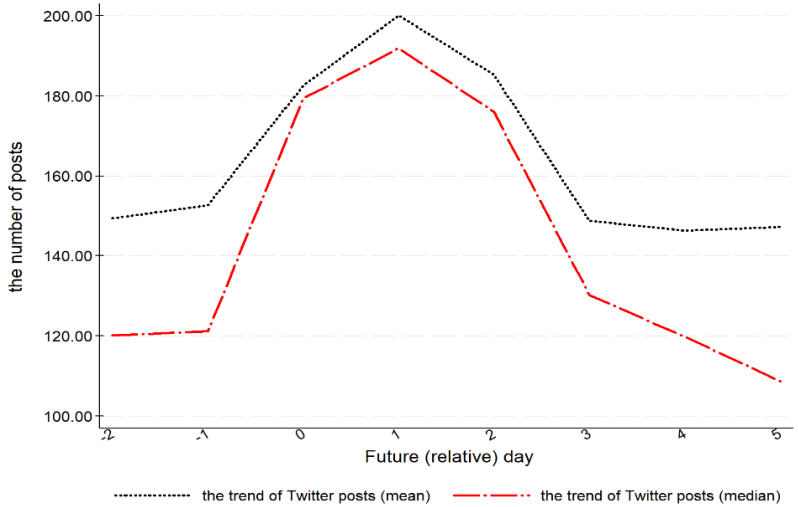
This figure plots social media attention for stocks from initial recommendations, partitioned by whether AI agents relied on internal training data. We measure social media attention as the number of discussions on Twitter ("X" after July 2023). For each trading day, we first calculate the average number of discussions across all stocks within each category (training-data versus web-searched). We then compute the median and mean of these daily averages across all recommendation dates. The solid line shows the median number of discussions; the dotted line shows the mean. Day 0 represents the trading day when AI agents provide stock recommendations. The x-axis shows trading days relative to the recommendation date. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. Panels (a) and (b) examine stocks relying on training data; Panels (c) and (d) examine stocks relying on web-searched information. Panels (a) and (c) show GPT-4o results; Panels (b) and (d) show GPT-4 results.



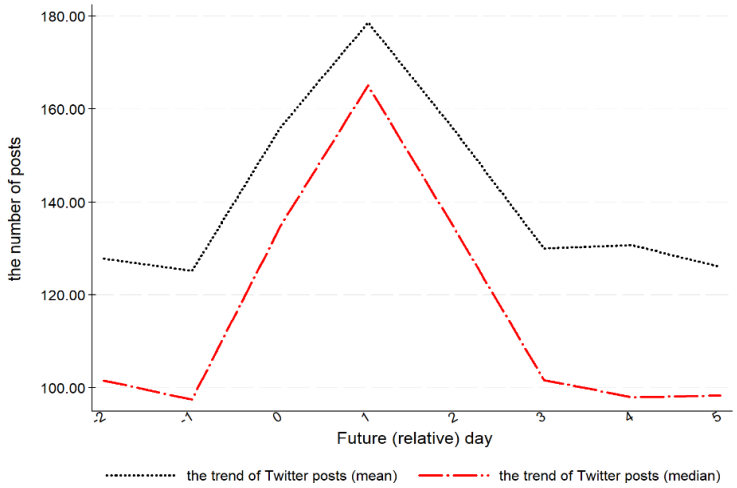
(a) GPT-4o (training)



(b) GPT-4 (training)



(c) GPT-4o (non-training)



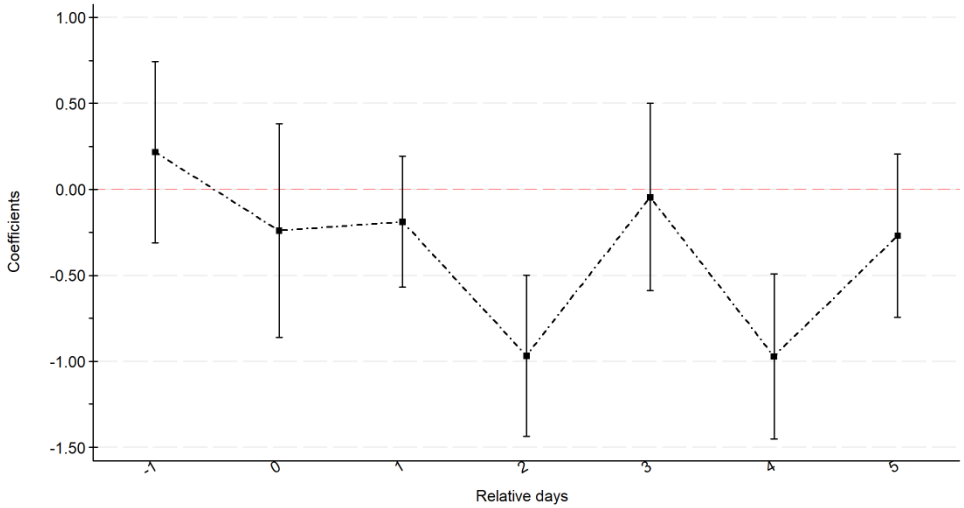
(d) GPT-4 (non-training)

Figure 3: Return predictability of retail trading around recommendation dates

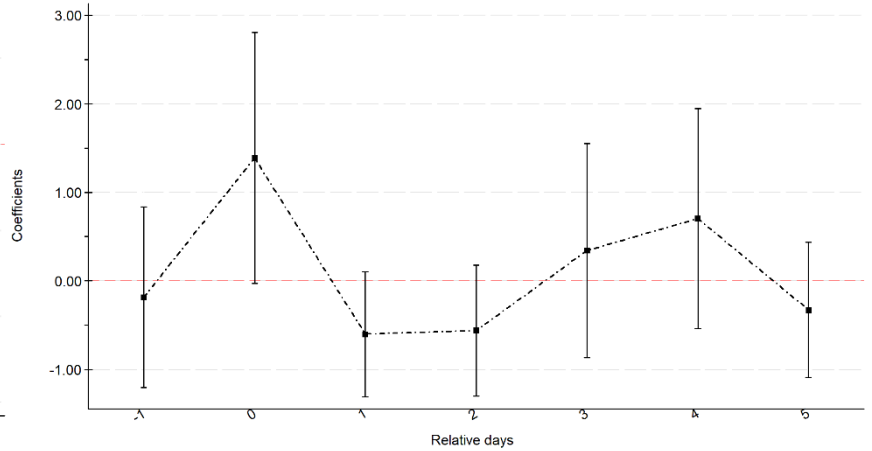
This figure plots the return predictability of retail trading for stocks from initial recommendations, partitioned by whether AI agents relied on internal training data. For each trading day relative to the recommendation date, we estimate the coefficient from a cross-sectional regression of day-t stock excess returns on day-t-1 retail trading, controlling for stock characteristics. We measure retail trading using marketable retail order imbalance (*MROIB*), calculated as buy share volumes minus sell share volumes for a stock on day t-1, scaled by the sum of buy and sell share volumes (Barber et al., 2024). Control variables include the natural logarithm of market capitalization on day t-1, the standard deviation of daily returns over trading days t-20 to t-1, average daily share turnover over trading days t-20 to t-1, the stock return on day t-1, the cumulative daily return over trading days t-6 to t-2, and the cumulative daily return over trading days t-16 to t-7. We also control for day fixed effects. Standard errors are clustered by firm. The regression model is specified as follows:

$$Abnormal\ return_{i,t} = \beta_1 MROIB_{i,t-1} + Control\ variables + \mu_t + \varepsilon_{i,t}.$$

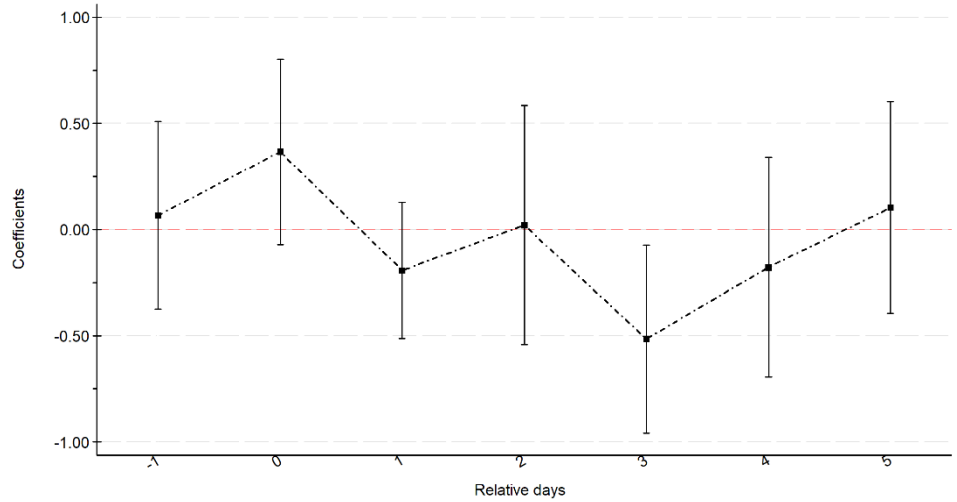
Day 0 represents the trading day when AI agents provide stock recommendations. The x-axis shows trading days relative to the recommendation date. Dots represent coefficient estimates on *MROIB*; vertical lines denote 90% confidence intervals. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. Panels (a) and (b) examine stocks relying on training data; Panels (c) and (d) examine stocks relying on web-searched information. Panels (a) and (c) show GPT-4o results; Panels (b) and (d) show GPT-4 results.



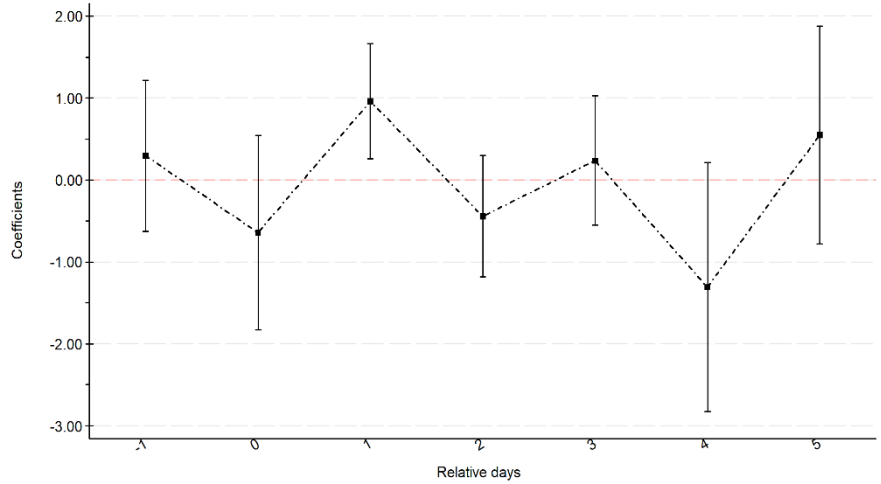
(a) GPT-4o (training)



(b) GPT-4 (training)



(c) GPT-4o (non-training)



(d) GPT-4 (non-training)

Table 1: Data collection summary

This table summarizes our data collection process. We use GPT-4o and GPT-4, which are among the first LLMs allowing registered users to conduct web searching. Data collection from GPT-4 ended on April 30, 2025, when OpenAI retired GPT-4. Our data collection follows a structured protocol with six core questions: one initial question asking for stock recommendations, followed by five questions asking for updated recommendations. Since October 21, 2024, we added a question after the initial recommendation to verify whether it relied on internal training data. Since November 13, 2024, we added another question after the 5th update to verify whether the final recommendation relied on internal training data. The prompts for all questions are provided in the Internet Appendix.

Model	Start date	End date	# of trading days	# of trading days with recommendations
GPT-4o	28/05/2024	30/06/2025	273	266
GPT-4	28/05/2024	29/04/2025	232	230

Table 2: Summary statistics of stocks from initial recommendations

This table summarizes characteristics of stocks from initial recommendations. Panel A reports daily statistics at the recommendation-set level (N = number of days). Variables are defined as follows: "Missing" indicates days when AI agents failed to provide 20 stock recommendations; "Common stock" counts standard publicly traded company stocks; "ADR" counts American Depositary Receipts; "OTC" counts over-the-counter traded stocks; "Untradeable stocks" counts stocks that cannot be traded (e.g., not yet listed or already delisted); "Non-stock" counts recommendations that are not individual stocks (e.g., cryptocurrencies, funds, bonds, industries, or stock categories like "Small-Caps"). Our subsequent descriptive statistics for AI agent misinformation focus on recommendations classified as Common stock, ADR, or OTC, as these represent tradeable equity securities. Panel B lists the 40 most frequently recommended stocks. Bolded stocks are S&P 100 constituents. Sector classifications from Global Industry Classification Standard (GICS) are shown in brackets. Recommendation frequency (in parentheses) equals the number of days the stock was recommended divided by total query trading days.

Panel A: Daily stock recommendations

	N	p25	Mean	Median	p75	S.D.
GPT-4o						
Missing	266	0.000	0.004	0.000	0.000	0.061
Common stock	266	20.000	19.398	20.000	20.000	1.609
ADR	266	0.000	0.417	0.000	0.000	0.984
OTC	266	0.000	0.049	0.000	0.000	0.233
Untradeable stocks	266	0.000	0.015	0.000	0.000	0.150
Non-stock	266	0.000	0.060	0.000	0.000	0.319
GPT-4						
Missing	230	0.000	0.013	0.000	0.000	0.114
Common stock	230	18.000	18.517	19.000	20.000	2.523
ADR	230	0.000	0.839	0.000	1.000	1.256
OTC	230	0.000	0.183	0.000	0.000	0.577
Untradeable stocks	230	0.000	0.083	0.000	0.000	0.320
Non-stock	230	0.000	0.330	0.000	0.000	1.255

Panel B: Top 40 stocks recommended by ChatGPT

GPT-4o	GPT-4
NVDA (95.9%)	**NVDA** (80.9%)
[*Information Technology*]	[*Information Technology*]
MSFT (91.4%)	**AMZN** (78.7%)
[*Information Technology*]	[*Consumer Discretionary*]
AMZN (91.4%)	**TSLA** (71.3%)
[*Consumer Discretionary*]	***[****Consumer Discretionary****]***
TSLA (86.5%)	**MSFT** (67.0 %)
[*Consumer Discretionary*]	[*Information Technology*]
META (86.5%)	**AAPL** (62.2%)
[*Communication Services*]	[*Information Technology*]
GOOGL (85.0%)	**GOOGL** (60.0%)
[*Communication Services*]	[*Communication Services*]
AAPL (82.7%)	**PFE** (43.5%)
[*Information Technology*]	[*Health Care*]
JNJ (66.2%)	**META** (42.6%)
[*Health Care*]	[*Communication Services*]
V (64.7%)	**JNJ** (39.6%)
[*Financials*]	[*Health Care*]
PG (64.3%)	**KO** (30.9%)
[*Consumer Staples*]	[*Consumer Staples*]
JPM (56.0%)	**XOM** (27.8%)
[*Financials*]	[*Energy*]
BRKB (46.2%)	**BA** (26.5%)
[*Financials*]	[*Industrials*]
MA (44.7%)	**GS** (26.5%)
[*Financials*]	[*Financials*]
HD (44.4%)	**V** (26.5%)
[*Consumer Discretionary*]	[*Financials*]
PFE (43.2%)	**WMT** (24.8%)
[*Health Care*]	[*Consumer Staples*]
UNH (43.2%)	**CVX** (23.5%)
[*Health Care*]	[*Energy*]
XOM (41.4%)	**UNH** (20.4%)
[*Energy*]	[*Health Care*]
KO (36.1%)	**AVGO** (19.6%)
[*Consumer Staples*]	[*Information Technology*]
WMT (34.6%)	**PG** (19.1%)
[*Consumer Staples*]	[*Consumer Staples*]
DIS (32.0%)	**BRKB** (18.3%)
[*Communication Services*]	[*Financials*]
AVGO (30.5%)	**INTC** (17.4%)
[*Information Technology*]	[*Information Technology*]
NFLX (26.7%)	**NFLX** (16.1%)
[*Communication Services*]	[*Communication Services*]
BAC (24.4%)	**MA** (16.1%)
[*Financials*]	[*Financials*]

PEP (24.8%)	**NKE** (16.1 %)
[*Consumer Staples*]	[*Consumer Discretionary*]
INTC (22.9%)	**MCD** (15.7%)
[*Information Technology*]	[*Consumer Discretionary*]
COST (21.8%)	**JPM** (15.7 %)
[*Consumer Staples*]	[*Financials*]
LLY (19.2%)	**ADBE** (14.3%)
[*Health Care*]	[*Information Technology*]
AMD (16.9%)	**AMD** (13.9%)
[*Information Technology*]	[*Information Technology*]
ADBE (15.8%)	MRNA (13.9%)
[*Information Technology*]	[*Health Care*]
PLTR (15.4%)	**T (13.9%)**
[*Information Technology*]	[*Communication Services*]
CRM (15.0%)	DAL (13.0%)
[*Information Technology*]	[*Industrials*]
CVX (12.8%)	**GE (13.0%)**
[*Energy*]	[*Industrials*]
C (9.0%)	**LLY (12.6%)**
[*Financials*]	[*Health Care*]
GOOG (7.5%)	**BLK (11.7%)**
[*Communication Services*]	[*Financials*]
ABBV (7.5%)	**PLTR (11.3%)**
[*Health Care*]	[*Information Technology]*
PYPL (7.1%)	TSM (11.3%)
[*Financials*]	[*Information Technology*]
TSM (7.1%)	CCL (10.0%)
[*Information Technology*]	[*Consumer Discretionary*]
NKE (6.8%)	MU (9.6%)
[*Consumer Discretionary*]	[*Information Technology*]
MCD (6.8%)	**DIS (9.1%)**
[*Consumer Discretionary*]	[*Communication Services*]
CRWD (6.4%)	**HD (9.1%)**
[*Information Technology*]	[*Consumer Discretionary*]

Table 3: Training-data reliance for initial recommendations

This table examines training-data reliance in the initial recommendations. Panel A identifies stocks where AI agents did not cite any external sources from their web searches, indicating reliance on internal training data. The first row shows the ratio at the stock-recommendation level based on the indicator for each stock; the second row shows the count per daily recommendation (out of 20 stocks). Panel B shows a less accurate, lower-cost alternative measure at the daily level: cases where AI agents voluntarily stated they used "training data" when asked to confirm their information sources. This self-declared measure is a lower-cost but noisier alternative to the source-verification method used in Panel A. The confirmation prompt asks about all 20 recommendations collectively, so AI agents may misclassify the whole portfolio when only some rely on training data while others use web-searched information. The lower number of observations in Panel B reflects that we added the confirmation prompt at a later date in our data collection.

Panel A: Stock recommendations relying on training data

	N	p25	Mean	Median	p75	S.D.
Ratio of stocks with training-data reliance						
GPT-4o	5,303	0.000	0.373	0.000	1.000	0.484
GPT-4	4,561	0.000	0.297	0.000	1.000	0.457
Number of stocks per day with training-data reliance (out of 20)						
GPT-4o	266	0.000	7.444	4.000	16.000	7.981
GPT-4	230	0.000	5.891	0.000	13.000	7.833

Panel B: Self-declared training-data reliance at the daily level

	N	p25	Mean	Median	p75	S.D.
GPT-4o	165	0.000	0.503	1.000	1.000	0.502
GPT-4	129	0.000	0.039	0.000	0.000	0.194

Table 4: Web-searching bias in initial recommendations

This table examines web-searching bias in the initial recommendations. Panel A shows search behavior metrics at the daily level (first two rows, where each day has 20 stock recommendations) and at the individual stock recommendation level (third row). "URLs searched" counts total individual web pages accessed; "unique websites" counts distinct website domains used (e.g., multiple Yahoo Finance articles count as multiple URLs but one unique source). The second row conditions on days when at least one URL was searched, excluding days when AI agents provided recommendations without any web searching. The third row reports the number of sources cited per individual stock recommendation. Panel B lists the 20 most frequently searched websites in our whole sample period and their search frequency per day (shown in parentheses) separately for GPT-4o and GPT-4. Search concentration (HHI) is the Herfindahl index calculated using the top 10 most-searched websites' shares of total citations: ∑(each of top 10 websites' share of all citations)². Higher values indicate more concentrated search patterns. This analysis excludes individual stock recommendations relying solely on training data. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches.

Panel A: Information sources used

	N	p25	Mean	Median	p75	S.D.
Number of URLs searched per day						
GPT-4o	266	3.000	9.921	11.000	16.000	8.110
GPT-4	230	2.000	3.683	4.000	5.000	2.370
Number of unique websites per day (only days with URLs searched)						
GPT-4o	215	3.000	5.372	4.000	5.000	4.692
GPT-4	194	2.000	3.314	3.000	4.000	1.204
Number of sources cited per recommended stock						
GPT-4o	5,303	1.000	1.429	1.000	2.000	0.620
GPT-4	4,561	1.000	1.118	1.000	1.000	0.331

Panel B: Concentration in top 20 most-searched websites

GPT-4o	GPT-4
Top 10 Search concentration HHI= 0.153	Top 10 search concentration HHI= 0.128
Yahoo Finance (3.128)	Bing (0.843)
Morningstar (1.741)	Investopedia (0.704)
Investopedia (1.312)	Morningstar (0.474)
The Motley Fool (0.312)	NerdWallet (0.296)
U.S. News (0.308)	Fidelity Investments (0.278)
Seeking Alpha (0.293)	Kiplinger (0.274)
The Wall Street Journal (0.233)	Seeking Alpha (0.196)
Bing (0.226)	Nasdaq (0.077)
Forbes (0.226)	TradingView (0.057)
NerdWallet (0.214)	TipRanks (0.057)
MarketWatch (0.117)	InvestorPlace (0.048)
Reuters (0.117)	Investing.com (0.039)
Investing.com (0.113)	Yahoo Finance (0.035)
Fidelity Investments (0.102)	Stock Analysis (0.026)
Insider Monkey (0.098)	Stock-screener.org (0.026)
Kiplinger (0.098)	Investor's Business Daily (0.022)
Nasdaq (0.094)	Invezz (0.022)
Investor's Business Daily (0.086)	Benzinga (0.017)
Tastylive (0.075)	The Wall Street Journal (0.017)
Bankrate (0.071)	Barchart / StatMuse / The Motley Fool / WallStreetZen (0.013)

Table 5: Representative heuristic reasoning in initial recommendations

This table examines representative heuristic reasoning in the initial recommendations. Panel A analyzes reasoning characteristics at the individual stock recommendation level. "Ratio of recommendations without reasoning" indicates the proportion where AI agents did not provide any reasoning. "Word count" measures reasoning length. "Sentiment" is calculated as (number of positive words - number of negative words) / (total positive and negative words), using the Loughran and McDonald (2014) financial dictionary. Panels B and C show the reasoning themes, with percentages indicating how often each theme appears (calculated as the number of recommendations citing that theme divided by the total number of recommendations). Panel B analyzes recommendations relying on training data; Panel C analyzes recommendations relying on web-searched information. Categories can overlap as recommendations may cite multiple themes.

Panel A: Characteristics of reasoning provided

	N	p25	Mean	Median	p75	S.D.
Ratio of recommendations without reasoning provided						
GPT-4o	5,303	0.000	0.004	0.000	0.000	0.060
GPT-4	4,561	0.000	0.000	0.000	0.000	0.000
Word count of reasoning						
GPT-4o	5,284	10.000	12.749	12.000	15.000	4.119
GPT-4	4,561	6.000	7.733	7.000	9.000	2.410
Sentiment of reasoning						
GPT-4o	5,284	0.000	0.658	1.000	1.000	0.543
GPT-4	4,561	0.000	0.557	1.000	1.000	0.569

Panel B: Reason classifications for training-data recommendations

GPT-4o	GPT-4
Growth drivers & expansion (55.6%)	Growth drivers & expansion (38.5%)
Technology & innovation leadership (39.9%)	Technology & innovation leadership (32.0%)
Market leadership & competitive position (27.4%)	Other/Multiple themes (19.3%)
Financial performance & metrics (24.9%)	Market leadership & competitive position (16.2%)
Sector & industry tailwinds (21.7%)	Sector & industry tailwinds (16.0%)
Other/Multiple themes (8.5%)	Financial performance & metrics (15.4%)
Strategic initiatives & corporate actions (6.9%)	Recovery & turnaround themes (7.2%)
Income & shareholder returns (6.0%)	Strategic initiatives & corporate actions (5.1%)
Valuation opportunity & price dislocation (5.1%)	**Valuation opportunity & price dislocation** (3.6%)
Recovery & turnaround themes (4.5%)	Income & shareholder returns (3.5%)
Analyst actions & rating changes (0.8%)	Analyst actions & rating changes (0.7%)
Event-driven catalysts (0.4%)	Event-driven catalysts (0.1%)

Panel C: Reason classifications for web-searched recommendations

GPT-4o	GPT-4
Growth drivers & expansion (56.7%)	Growth drivers & expansion (40.4%)
Technology & innovation leadership (41.1%)	Technology & innovation leadership (31.3%)
Sector & industry tailwinds (23.1%)	Sector & industry tailwinds (20.2%)
Market leadership & competitive position (22.9%)	Other/Multiple themes (18.0%)
Financial performance & metrics (22.7%)	Market leadership & competitive position (15.2%)
Other/Multiple themes (9.3%)	Financial performance & metrics (14.7%)
Strategic initiatives & corporate actions (6.7%)	Recovery & turnaround themes (6.4%)
Income & shareholder returns (6.3%)	Strategic initiatives & corporate actions (5.4%)
Recovery & turnaround themes (4.8%)	**Valuation opportunity & price dislocation (3.6%)**
Valuation opportunity & price dislocation (4.8%)	Analyst actions & rating changes (2.1%)
Analyst actions & rating changes (1.6%)	Income & shareholder returns (1.9%)
Event-driven catalysts (0.2%)	Event-driven catalysts (0.5%)

Table 6: Changes in stock recommendations across sequential updates

This table reports changes in stock recommendations across five sequential updates for GPT-4o and GPT-4. Each day, we first collected initial recommendations, then requested five updates using the prompt in the internet appendix. Panel A shows how many stocks (out of 20) differ from the initial recommendation. Panel B shows how many stocks have different rankings from the initial recommendation. A ranking change occurs when either a different stock occupies the same rank position, or the same stock occupies a different rank position. The number of observations varies because GPT occasionally failed to provide updates.

Panel A: Number of stocks different from the initial recommendation

	N	p25	Mean	Median	p75	S.D.
1st update for GPT-4o	266	0.000	1.335	0.000	1.000	3.317
2nd update for GPT-4o	265	0.000	1.528	0.000	1.000	3.399
3rd update for GPT-4o	265	0.000	1.664	0.000	1.000	3.539
4th update for GPT-4o	266	0.000	1.812	0.000	1.000	3.722
5th update for GPT-4o	266	0.000	1.947	0.000	2.000	3.780
1st update for GPT-4	230	0.000	4.135	2.000	7.000	5.292
2nd update for GPT-4	230	0.000	6.161	5.000	11.000	5.870
3rd update for GPT-4	229	2.000	7.764	7.000	13.000	6.285
4th update for GPT-4	227	3.000	8.555	9.000	14.000	6.386
5th update for GPT-4	227	3.000	9.031	9.000	15.000	6.435

Panel B: Number of ranking changes from the initial recommendation

	N	p25	Mean	Median	p75	S.D.
1st update for GPT-4o	266	0.000	4.312	0.000	8.000	6.785
2nd update for GPT-4o	265	0.000	5.023	0.000	11.000	7.136
3rd update for GPT-4o	265	0.000	5.185	0.000	11.000	7.168
4th update for GPT-4o	266	0.000	5.628	0.000	12.000	7.453
5th update for GPT-4o	266	0.000	5.974	0.000	14.000	7.654
1st update for GPT-4	230	0.000	10.291	13.000	18.000	8.182
2nd update for GPT-4	230	6.000	13.122	17.000	19.000	7.699
3rd update for GPT-4	229	14.000	15.031	18.000	20.000	6.645
4th update for GPT-4	227	15.000	15.767	18.000	20.000	6.142
5th update for GPT-4	227	16.000	15.833	19.000	20.000	6.220

Table 7: Training-data reliance in final updated recommendations

This table examines training-data reliance in the fifth (final) updated recommendations. In Panel A, we classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. The first row shows the ratio at the stock-recommendation level based on the indicator for each stock; the second row shows the count per daily recommendation (out of 20 stocks). Panel B shows a less accurate, lower-cost alternative measure at the daily level: cases where AI agents voluntarily stated they used "training data" when asked to confirm their information sources. This self-declared measure is a lower-cost but noisier alternative to the source-verification method used in Panel A. The confirmation prompt asks about all 20 recommendations collectively, so AI agents may misclassify the whole portfolio when only some rely on training data while others use web-searched information. The lower number of observations in Panel B reflects that we added the confirmation prompt at a later date in our data collection.

Panel A: Stock recommendations relying on training data

	N	p25	Mean	Median	p75	S.D.
Ratio of stocks with training-data reliance						
GPT-4o	5,318	0.000	0.405	0.000	1.000	0.491
GPT-4	4,518	0.000	0.482	0.000	1.000	0.500
Number of stocks per day with training-data reliance (out of 20)						
GPT-4o	266	0.000	8.105	6.000	17.000	7.885
GPT-4	227	0.000	9.595	9.000	19.000	8.181

Panel B: Self-declared training-data reliance at the daily level

	N	p25	Mean	Median	p75	S.D.
GPT-4o	155	0.000	0.310	0.000	1.000	0.464
GPT-4	112	0.000	0.009	0.000	0.000	0.094

Table 8: Web-searching bias in final updated recommendations

This table examines web-searching bias in the fifth (final) updated recommendations. Panel A shows search behavior metrics at the daily level (first two rows, where each day has 20 stock recommendations) and at the individual stock recommendation level (third row). "URLs searched" counts total individual web pages accessed; "unique websites" counts distinct website domains used (e.g., multiple Yahoo Finance articles count as multiple URLs but one unique source). The second row conditions on days when at least one URL was searched, excluding days when AI agents provided recommendations without any web searching. The third row reports the number of sources cited per individual stock recommendation. Panel B lists the 20 most frequently searched websites and their search frequency per day (shown in parentheses) separately for GPT-4o and GPT-4. Search concentration (HHI) is the Herfindahl index calculated using the top 10 most-searched websites' shares of total citations: ∑(each of top 10 websites' share of all citations)². Higher values indicate more concentrated search patterns. This analysis excludes individual stock recommendations relying solely on training data. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches.

Panel A: Information sources used

	N	p25	Mean	Median	p75	S.D.
Number of URLs searched per day						
GPT-4o	266	2.000	12.289	10.000	12.000	18.579
GPT-4	227	2.000	3.374	3.000	5.000	2.742
Number of unique websites per day (only days with URLs searched)						
GPT-4o	216	3.000	4.125	3.000	4.000	2.609
GPT-4	176	2.000	2.807	3.000	3.000	0.905
Number of sources cited per recommended stock						
GPT-4o	5,318	1.000	2.043	2.000	2.000	1.352
GPT-4	4,518	1.000	1.238	1.000	1.000	0.459

Panel B: Concentration in top 20 most-searched websites

GPT-4o	GPT-4
Top 10 search concentration HHI=0.196	Top 10 search concentration HHI= 0.208
Yahoo Finance (4.353)	Investopedia (1.083)
Morningstar (2.237)	Bing (0.765)
Investopedia (2.120)	Fidelity Investments (0.630)
The Wall Street Journal (0.609)	Seeking Alpha (0.296)
Seeking Alpha (0.534)	Morningstar (0.178)
MarketWatch (0.440)	Stock Analysis (0.087)
Reuters (0.305)	NerdWallet (0.057)
Bing (0.267)	The Wall Street Journal (0.048)
Fidelity Investments (0.222)	MarketWatch (0.039)
Bloomberg (0.195)	Kiplinger (0.030)
Forbes (0.124)	Investing.com (0.026)
The Motley Fool (0.117)	TradingView (0.013)
NerdWallet (0.113)	Investor's Business Daily (0.009)
U.S. News (0.102)	Nasdaq (0.009)
Kiplinger (0.098)	Wall Street Zen (0.009)
Investor's Business Daily (0.049)	Bloomberg (0.004)
Investing.com (0.038)	Business Insider (0.004)
BigCharts (0.038)	FinanceCharts (0.004)
MarketBeat (0.023)	Finviz (0.004)
Zacks Investment Research (0.023)	InvestorPlace / Kavout / Reuters / TheStreet / ThinkAdvisor (0.004)

Table 9: Representative heuristic reasoning in final updated recommendations
This table examines representative heuristic reasoning in the fifth (final) updated recommendations. Panel A analyzes reasoning characteristics at the individual stock recommendation level. "Ratio of recommendations without reasoning" indicates the proportion where AI agents did not provide any reasoning. "Word count" measures reasoning length. "Sentiment" is calculated as (number of positive words - number of negative words) / (total positive and negative words), using the Loughran and McDonald (2014) financial dictionary. Panels B and C show the reasoning themes, with percentages indicating how often each theme appears (calculated as the number of recommendations citing that theme divided by the total number of recommendations). Panel A reports the results for stock recommendations based on training data, while Panel B reports the results for stock recommendations based on web-searched information. Categories can overlap as recommendations may cite multiple themes.

Panel A: Characteristics of reasoning provided

	N	p25	Mean	Median	p75	S.D.
Ratio of recommendations without reasoning provided						
GPT-4o	5,318	0.000	0.000	0.000	0.000	0.000
GPT-4	4,518	0.000	0.000	0.000	0.000	0.021
Word count of reasoning						
GPT-4o	5,318	11.000	13.794	13.000	16.000	4.311
GPT-4	4,516	7.000	8.104	8.000	9.000	2.020
Sentiment of reasoning						
GPT-4o	5318	0.000	0.671	1.000	1.000	0.537
GPT-4	4516	0.000	0.606	1.000	1.000	0.533

Panel B: Reason classifications for training-data recommendations

GPT-4o	GPT-4
Growth drivers & expansion (61.1 %)	Growth drivers & expansion (40.0%)
Technology & innovation leadership (41.9%)	Technology & innovation leadership (39.3%)
Market leadership & competitive position (26.6%)	Market leadership & competitive position (21.0%)
Financial performance & metrics (25.7%)	Other/Multiple themes (17.7 %)
Sector & industry tailwinds (23.4%)	Sector & industry tailwinds (14.9%)
Strategic initiatives & corporate actions (9.2%)	Financial performance & metrics (13.7%)
Other/Multiple themes (6.4%)	Strategic initiatives & corporate actions (7.0%)
Income & shareholder returns (6.0%)	Recovery & turnaround themes (5.8%)
Valuation opportunity & price dislocation (5.3%)	Income & shareholder returns (2.3%)
Recovery & turnaround themes (4.8%)	**Valuation opportunity & price dislocation** (0.8%)
Analyst actions & rating changes (0.7%)	Analyst actions & rating changes (0.2%)
Event-driven catalysts (0.4%)	Event-driven catalysts (0.2%)

Panel C: Reason classifications for web-searched recommendations

GPT-4o	GPT-4
Growth drivers & expansion (59.0%)	Technology & innovation leadership (44.0%)
Technology & innovation leadership (44.5%)	Growth drivers & expansion (40.5%)
Sector & industry tailwinds (24.2%)	Market leadership & competitive position (21.3%)
Financial performance & metrics (23.5%)	Sector & industry tailwinds (17.1%)
Market leadership & competitive position (22.1%)	Other/Multiple themes (15.1%)
Other/Multiple themes (8.8%)	Financial performance & metrics (14.1%)
Strategic initiatives & corporate actions (7.0%)	Strategic initiatives & corporate actions (7.2%)
Income & shareholder returns (6.0%)	Recovery & turnaround themes (4.7%)
Recovery & turnaround themes (4.9%)	**Valuation opportunity & price dislocation** (2.3%)
Valuation opportunity & price dislocation (3.9%)	Income & shareholder returns (2.1%)
Analyst actions & rating changes (1.8%)	Analyst actions & rating changes (0.3%)
Event-driven catalysts (0.3%)	Event-driven catalysts (0.1%)

Table 10: Training-data reliance for initial recommendations and ChatGPT upgrades
This table examines how training-data reliance in initial recommendations changes following ChatGPT model upgrades. We classify a recommendation as relying on training data if the websites cited as information sources for this recommendation were not actually searched, indicating the information came from internal training data rather than real-time web searches. The first row shows the ratio at the stock-recommendation level based on the indicator for each stock; the second row shows the count per daily recommendation (out of 20 stocks). We split the sample into three periods based on major ChatGPT upgrades: Before October 2024, From November 2024 to April 2025, and After May 2025. We are not able to conduct similar analysis for self-declared training-data reliance, which was only collected from November 2024 to April 2025.

	N	p25	Mean	Median	p75	S.D.
Ratio of stocks with training-data reliance						
Before Oct 2024						
GPT-4o	2,183	0.000	0.563	1.000	1.000	0.496
GPT-4	2,166	0.000	0.348	0.000	1.000	0.476
From Nov 2024 to Apr 2025						
GPT-4o	2,300	0.000	0.244	0.000	0.000	0.430
GPT-4	2,395	0.000	0.251	0.000	1.000	0.434
After May 2025						
GPT-4o	820	0.000	0.233	0.000	0.000	0.423
Number of stocks per day with training-data reliance (out of 20)						
Before Oct 2024						
GPT-4o	110	0.000	11.164	12.500	20.000	8.989
GPT-4	110	0.000	6.845	0.000	15.000	8.604
From Nov 2024 to Apr 2025						
GPT-4o	115	0.000	4.878	2.000	8.000	6.226
GPT-4	120	0.000	5.017	0.000	9.500	6.975
After May 2025						
GPT-4o	41	1.000	4.659	3.000	6.000	5.136

Table 11: Web-searching bias in initial recommendations and ChatGPT upgrades

This table examines whether web-searching bias in initial recommendations changes following ChatGPT model upgrades. Panel A shows search behavior metrics at the daily level (first two rows, where each day has 20 stock recommendations) and at the individual stock recommendation level (third row). "URLs searched" counts total individual web pages accessed; "unique websites" counts distinct website domains used (e.g., multiple Yahoo Finance articles count as multiple URLs but one unique source). The second row conditions on days when at least one URL was searched, excluding days when AI agents provided recommendations without any web searching. The third row reports the number of sources cited per individual stock recommendation. Panel B lists the 20 most frequently searched websites and their search frequency per day (shown in parentheses) separately for GPT-4o and GPT-4. Search concentration (HHI) is the Herfindahl index calculated using the top 10 most-searched websites' shares of total citations: ∑(each of top 10 websites' share of all citations)². Higher values indicate more concentrated search patterns. We split the sample into three periods based on major ChatGPT upgrades: Before October 2024, From November 2024 to April 2025, and After May 2025. This analysis excludes individual stock recommendations relying solely on training data. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches.

Panel A: Information sources used

	N	p25	Mean	Median	p75	S.D.
Number of URLs searched per day						
Before Oct 2024						
GPT-4o	110	0.000	1.818	2.000	3.000	1.883
GPT-4	110	2.000	3.273	3.000	5.000	2.449
From Nov 2024 to Apr 2025						
GPT-4o	115	11.000	13.296	12.000	17.000	3.086
GPT-4	120	3.000	4.058	4.000	5.000	2.239
After May 2025						
GPT-4o	41	18.000	22.195	21.000	25.000	5.591
Number of unique websites per day (only days with URLs searched)						
Before Oct 2024						
GPT-4o	59	2.000	2.983	3.000	4.000	0.956
GPT-4	87	2.000	3.115	3.000	4.000	1.050
From Nov 2024 to Apr 2025						
GPT-4o	115	3.000	4.730	4.000	5.000	2.552
GPT-4	107	3.000	3.477	3.000	4.000	1.298
After May 2025						
GPT-4o	41	4.000	10.610	5.000	18.000	7.755
Number of sources cited per recommended stock						
Before Oct 2024						
GPT-4o	2183	1.000	1.662	2.000	2.000	0.541
GPT-4	2166	1.000	1.113	1.000	1.000	0.335
From Nov 2024 to Apr 2025						
GPT-4o	2300	1.000	1.253	1.000	2.000	0.620
GPT-4	2395	1.000	1.122	1.000	1.000	0.327
After May 2025						
GPT-4o	820	1.000	1.301	1.000	2.000	0.618

Panel B: Concentration in top 20 most-searched websites

Before Oct 2024		*From Nov 2024 to Apr 2025*		*After May 2025*
GPT-4o	*GPT-4*	*GPT-4o*	*GPT-4*	*GPT-4o*
HHI= 0.146	HHI= 0.138	HHI= 0.240	HHI= 0.130	HHI= 0.093
Bing	Bing	Yahoo Finance	Investopedia	Yahoo Finance
(0.545)	(0.791)	(5.591)	(0.933)	(4.463)
NerdWallet	Morningstar	Morningstar	Bing	Morningstar
(0.273)	(0.509)	(2.574)	(0.892)	(3.902)
Investopedia	Investopedia	Investopedia	Morningstar	Investopedia
(0.264)	(0.455)	(1.939)	(0.442)	(2.366)
Seeking Alpha	NerdWallet	Seeking Alpha	Fidelity Investments	U.S. News
(0.118)	(0.436)	(0.522)	(0.317)	(1.293)
Nasdaq	Kiplinger	The Motley Fool	Kiplinger	The Motley Fool
(0.091)	(0.291)	(0.357)	(0.258)	(1.000)
MarketBeat	Fidelity Investments	Wall Street Journal	Seeking Alpha	Forbes (0.780)
(0.073)	(0.236)	(0.348)	(0.192)	
Morningstar	Seeking Alpha	Reuters	NerdWallet	Insider Monkey
(0.064)	(0.200)	(0.270)	(0.167)	(0.537)
Investing.com	TipRanks	U.S. News	Nasdaq	Wall Street Journal
(0.055)	(0.091)	(0.252)	(0.092)	(0.537)
InvestorPlace	InvestorPlace	Forbes	TradingView	Investing.com
(0.055)	(0.045)	(0.243)	(0.083)	(0.463)
Yahoo Finance	Nasdaq	MarketWatch	Yahoo Finance	NerdWallet
(0.055)	(0.045)	(0.243)	(0.067)	(0.439)
Business Insider	Stock Analysis	Fidelity Investments	Investing.com	FinanceCharts
(0.045)	(0.027)	(0.191)	(0.058)	(0.415)
Kiplinger	Stock-screener.org	Bloomberg	InvestorPlace	Investor's Business Daily
(0.027)	(0.027)	(0.096)	(0.050)	(0.366)
Money Crashers	TradingView	Kiplinger	Investor's Business Daily	Bankrate
(0.027)	(0.027)	(0.087)	(0.042)	(0.341)
Kavout	Benzinga	NerdWallet	Invezz	StatMuse

(0.018)	(0.018)	(0.078)	(0.042)	(0.341)
TradingView	Investing.com	Investor's Business Daily	The Wall Street Journal	Tastylive
(0.018)	(0.018)	(0.070)	(0.033)	(0.341)
Barchart	MarketBeat	Tastylive	Barchart	Kiplinger
(0.009)	(0.018)	(0.052)	(0.025)	(0.317)
Fidelity Investments	Wall Street Zen	Bankrate	StatMuse	Zacks Investment Research
(0.009)	(0.018)	(0.043)	(0.025)	(0.317)
GoBankingRates	ABG Analytics	Investing.com	Stock Analysis	ABG Analytics (0.268)
(0.009)	(0.009)	(0.043)	(0.025)	
Livemint	Moneyzine	YouTube	Stock-screener.org	Nasdaq
(0.009)	(0.009)	(0.043)	(0.025)	(0.268)
Stock Analysis / The Motley Fool / The Stock Watcher / Wall Street Zen / ABG Analytics / AOL		Business Insider / Insider Monkey / Nasdaq	The Motley Fool / TipRanks	Barron's / Stock-screener.org
(0.009)		(0.035)	(0.025)	(0.220)

Table 12: Representative heuristic reasoning in initial recommendations and ChatGPT upgrades

This table examines whether representative heuristic reasoning in the initial recommendations changes following ChatGPT model upgrades. Panel A analyzes reasoning characteristics at the individual stock recommendation level. "Ratio of recommendations without reasoning" indicates the proportion where AI agents did not provide any reasoning. "Word count" measures reasoning length. "Sentiment" is calculated as (number of positive words - number of negative words) / (total positive and negative words), using the Loughran and McDonald (2014) financial dictionary. Panels B and C show the reasoning themes, with percentages indicating how often each theme appears (calculated as the number of recommendations citing that theme divided by the total number of recommendations). Panel B analyzes recommendations relying on training data; Panel C analyzes recommendations relying on web-searched information. Categories can overlap as recommendations may cite multiple themes. We split the sample into three periods based on major ChatGPT upgrades: Before October 2024, From November 2024 to April 2025, and After May 2025.

Panel A: Characteristics of reasoning provided

	N	p25	Mean	Median	p75	S.D.
Ratio of recommendations without reasoning provided						
Before Oct 2024						
GPT-4o	2,183	0.000	0.000	0.000	0.000	0.000
GPT-4	2,166	0.000	0.000	0.000	0.000	0.000
From Nov 2024 to Apr 2025						
GPT-4o	2,300	0.000	0.000	0.000	0.000	0.000
GPT-4	2,395	0.000	0.000	0.000	0.000	0.000
After May 2025						
GPT-4o	820	0.000	0.023	0.000	0.000	0.151
Word count of reasoning						
Before Oct 2024						
GPT-4o	2,183	9.000	11.053	11.000	13.000	3.411
GPT-4	2,166	6.000	7.494	7.000	9.000	2.257
From Nov 2024 to Apr 2025						
GPT-4o	2,300	11.000	14.252	13.000	17.000	4.013
GPT-4	2,395	6.000	7.948	8.000	10.000	2.521
After May 2025						
GPT-4o	801	10.000	13.059	12.000	15.000	4.429
Sentiment of reasoning						
Before Oct 2024						
GPT-4o	2,183	0.000	0.695	1.000	1.000	0.525
GPT-4	2,166	0.000	0.574	1.000	1.000	0.559
From Nov 2024 to Apr 2025						
GPT-4o	2,300	0.000	0.663	1.000	1.000	0.532
GPT-4	2,395	0.000	0.542	1.000	1.000	0.577
After May 2025						
GPT-4o	801	0.000	0.544	1.000	1.000	0.604

Panel B: Reason classifications for training-data recommendations

GPT-4o	*GPT-4*	*GPT-4o*	*GPT-4*	*GPT-4o*
Before Oct 2024		*From Nov 2024 to Apr 2025*		*After May 2025*
Growth drivers & expansion (57.2%)	Growth drivers & expansion (39.2%)	Growth drivers & expansion (55.3%)	Growth drivers & expansion (37.5%)	Growth drivers & expansion (46.1%)
Technology & innovation leadership (42.0%)	Technology & innovation leadership (32.5%)	Technology & innovation leadership (37.6%)	Technology & innovation leadership (31.2%)	Financial performance & metrics (35.6%)
Market leadership & competitive position (32.7%)	Other/Multiple themes (18.5%)	Sector & industry tailwinds (34.6%)	Other/Multiple themes (20.4%)	Technology & innovation leadership (33.5%)
Financial performance & metrics (20.0%)	Financial performance & metrics (17.0%)	Financial performance & metrics (32.1%)	Market leadership & competitive position (18.3%)	Sector & industry tailwinds (21.5%)
Sector & industry tailwinds (15.8%)	Sector & industry tailwinds (16.3%)	Market leadership & competitive position (21.0%)	Sector & industry tailwinds (15.6%)	Market leadership & competitive position (12.0%)
Other/Multiple themes (8.6%)	Market leadership & competitive position (14.5%)	Strategic initiatives & corporate actions (9.4%)	Financial performance & metrics (13.5%)	Other/Multiple themes (12.0%)
Income & shareholder returns (6.6%)	Recovery & turnaround themes (6.0%)	Other/Multiple themes (7.3%)	Recovery & turnaround themes (8.6%)	Recovery & turnaround themes (6.8%)
Strategic initiatives & corporate actions (5.9%)	Strategic initiatives & corporate actions (5.6%)	**Valuation opportunity & price dislocation (5.0%)**	**Valuation opportunity & price dislocation (5.0%)**	Strategic initiatives & corporate actions (5.8%)
Valuation opportunity & price dislocation (5.2%)	Income & shareholder returns (2.8%)	Income & shareholder returns (5.0%)	Income & shareholder returns (4.7%)	Income & shareholder returns (5.2%)
Recovery & turnaround themes (4.4%)	**Valuation opportunity & price dislocation (2.4%)**	Recovery & turnaround themes (3.9%)	Strategic initiatives & corporate actions (4.5%)	**Valuation opportunity & price dislocation (4.2%)**
Event-driven catalysts (0.4%)	Analyst actions & rating changes (1.2%)	Analyst actions & rating changes (1.2%)	Analyst actions & rating changes (0.2%)	Analyst actions & rating changes (3.1%)
Analyst actions & rating changes (0.2%)	Event-driven catalysts (0.3%)			Event-driven catalysts (1.6%)

Panel C: Reason classifications for web-searched recommendations

GPT-4o	*GPT-4*	*GPT-4o*	*GPT-4*	*GPT-4o*
Before Oct 2024		*From Nov 2024 to Apr 2025*		*After May 2025*
Growth drivers & expansion (61.9%)	Growth drivers & expansion (40.2%)	Growth drivers & expansion (57.3%)	Growth drivers & expansion (40.6%)	Growth drivers & expansion (47.4%)
Technology & innovation leadership (48.0%)	Technology & innovation leadership (30.7%)	Technology & innovation leadership (38.4%)	Technology & innovation leadership (31.7%)	Technology & innovation leadership (38.5%)
Market leadership & competitive position (26.1%)	Sector & industry tailwinds (21.7%)	Sector & industry tailwinds (27.8%)	Sector & industry tailwinds (19.0%)	Financial performance & metrics (22.7%)
4 Financial performance & metrics (22.2%)	Other/Multiple themes (18.5%)	Financial performance & metrics (23.1%)	Other/Multiple themes (17.6%)	Sector & industry tailwinds (18.8%)
Sector & industry tailwinds (17.5%)	Market leadership & competitive position (16.6%)	Market leadership & competitive position (22.7%)	Financial performance & metrics (15.7%)	Market leadership & competitive position (18.7%)
Other/Multiple themes (7.6%)	Financial performance & metrics (13.5%)	Other/Multiple themes (9.8%)	Market leadership & competitive position (14.0%)	Other/Multiple themes (10.5%)
Strategic initiatives & corporate actions (5.5%)	Recovery & turnaround themes (6.1%)	Strategic initiatives & corporate actions (7.5%)	Recovery & turnaround themes (6.7%)	**Valuation opportunity & price dislocation (8.5%)**
Income & shareholder returns (5.0%)	Strategic initiatives & corporate actions (4.7%)	Income & shareholder returns (7.2%)	Strategic initiatives & corporate actions (5.9%)	Recovery & turnaround themes (6.9%)
Recovery & turnaround themes (4.1%)	**Valuation opportunity & price dislocation (3.2%)**	**Valuation opportunity & price dislocation (4.7%)**	**Valuation opportunity & price dislocation (4.0%)**	Strategic initiatives & corporate actions (6.4%)
Valuation opportunity & price dislocation (2.5%)	Analyst actions & rating changes (2.3%)	Recovery & turnaround themes (4.5%)	Analyst actions & rating changes (2.0%)	Income & shareholder returns (6.1%)
Analyst actions & rating changes (0.5%)	Income & shareholder returns (2.1%)	Analyst actions & rating changes (1.5%)	Income & shareholder returns (1.7%)	Analyst actions & rating changes (3.4%)
Event-driven catalysts (0.3%)	Event-driven catalysts (0.6%)	Event-driven catalysts (0.2%)	Event-driven catalysts (0.4%)	Event-driven catalysts (0.3%)

Table 13: Training-data reliance and future performance of stocks from the initial recommendations

This table reports value-weighted abnormal returns for portfolios of stocks from initial recommendations, partitioned by whether AI agents relied on internal training data. Panel A uses source verification; Panel B uses self-declaration. Abnormal return equals the raw return minus the corresponding CRSP market return. Average abnormal returns are reported above, with standard deviations in parentheses below. Because recommendations are made after market close, investors can only purchase stocks at the opening price the next day. Therefore, the Day 1 return is calculated as the closing price on Day 1 minus the opening price on Day 1, divided by the opening price on Day 1. The bottom rows report cumulative abnormal returns for Days 1-5. In Panel A, we classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. In Panel B, we classify all stocks from a given day as relying on training data if ChatGPT self-declared that it used internal training data to generate that day's stock recommendations.

Panel A: Stock recommendations citing unsearched websites

	GPT-4o		GPT-4	
	Training	Web-searched	Training	Web-searched
Day 1	0.160	0.089	-0.161	0.030
	(1.631)	(1.552)	(1.930)	(2.146)
Day 2	-0.019	-0.034	0.110	-0.065
	(1.073)	(1.100)	(1.368)	(1.381)
Day 3	0.046	-0.009	-0.043	0.076
	(1.055)	(1.053)	(1.380)	(1.551)
Day 4	-0.085	0.105	-0.068	-0.152
	(1.091)	(1.052)	(1.388)	(1.500)
Day 5	-0.040	0.059	-0.147	0.149
	(1.098)	(1.117)	(1.077)	(1.467)
Days 1-5	0.063	0.210	-0.309	0.038
	(2.428)	(2.222)	(3.272)	(3.547)

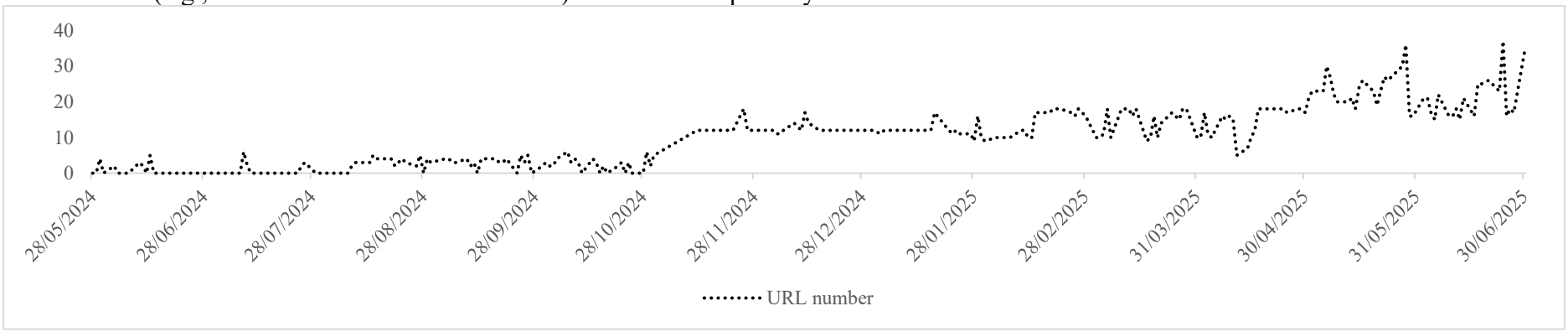
Panel B: Self-declared training data reliance at the daily level

	GPT-4o		GPT-4	
	Training	Web-searched	Training	Web-searched
Day 1	0.027	0.126	-0.487	0.088
	(1.177)	(1.930)	(1.695)	(2.200)
Day 2	-0.019	0.031	-0.156	-0.093
	(0.721)	(0.696)	(0.497)	(1.156)
Day 3	-0.096	0.040	-0.228	0.006
	(0.763)	(1.059)	(1.263)	(1.471)
Day 4	-0.206	0.169	0.266	-0.112
	(0.675)	(0.944)	(0.461)	(1.175)
Day 5	0.036	0.012	0.027	0.141
	(0.677)	(0.798)	(0.373)	(1.235)
Days 1-5	-0.258	0.378	-0.578	0.029
	(1.792)	(2.311)	(1.839)	(3.728)

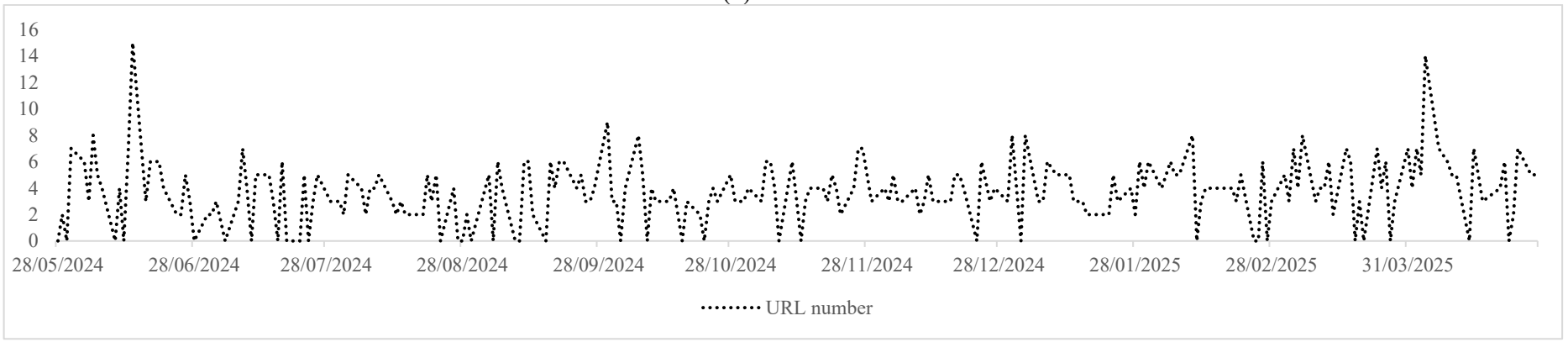
Internet Appendix

Figure A1. Number of URLs searched to generate initial stock recommendations

This figure plots the number of URLs AI agents searched when generating daily stock recommendations over the sample period. Panel (a) shows GPT-4o results from May 28, 2024 to June 30, 2025; Panel (b) shows GPT-4 results from May 28, 2024 to April 29, 2025. Each point represents the count of distinct URLs searched on a given day for the initial (non-updated) recommendations of 20 stocks. Note that multiple URLs from the same website (e.g., different Yahoo Finance articles) are counted separately.

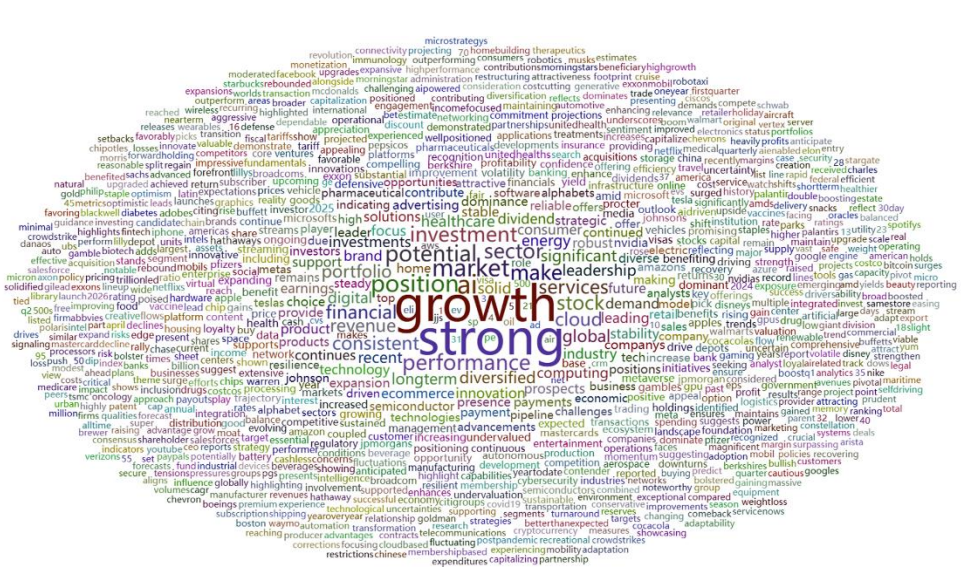


(a) GPT-4o

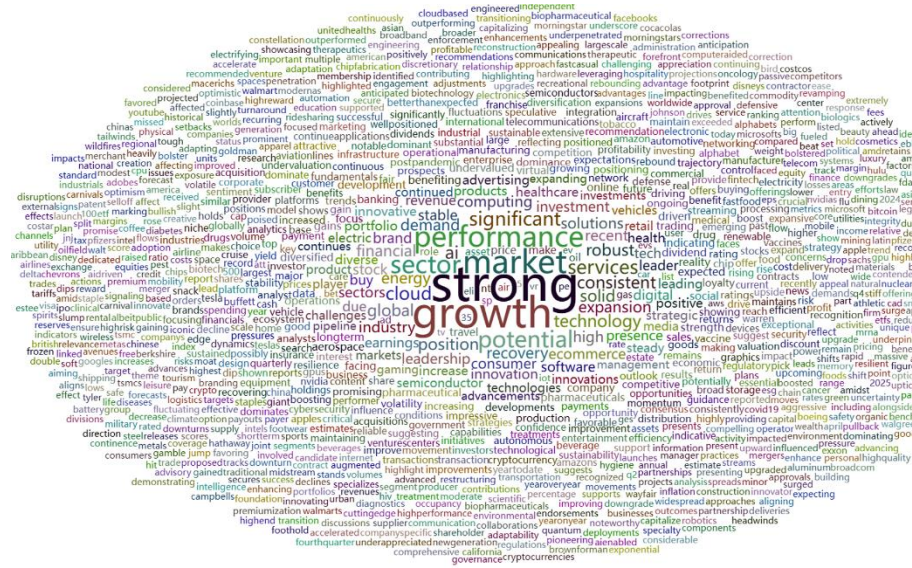


(b) GPT-4

Figure A2: The word cloud for the reasons ChatGPT uses to generate initial recommendations
This figure displays word clouds showing the frequency of words used in AI agents' reasoning for initial stock recommendations across the entire sample period. Word size reflects usage frequency. Panel (a) shows GPT-4o; Panel (b) shows GPT-4.



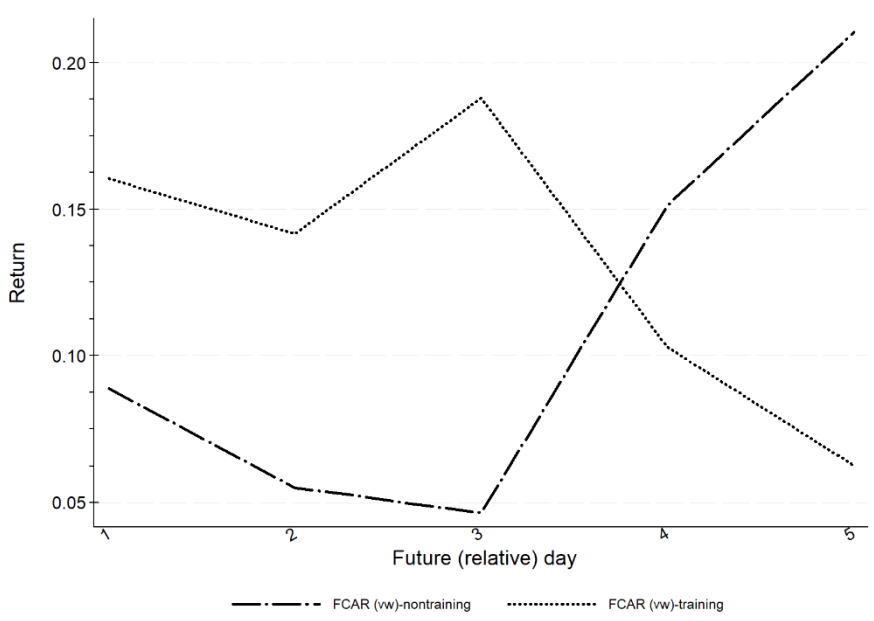
(a) GPT-4o



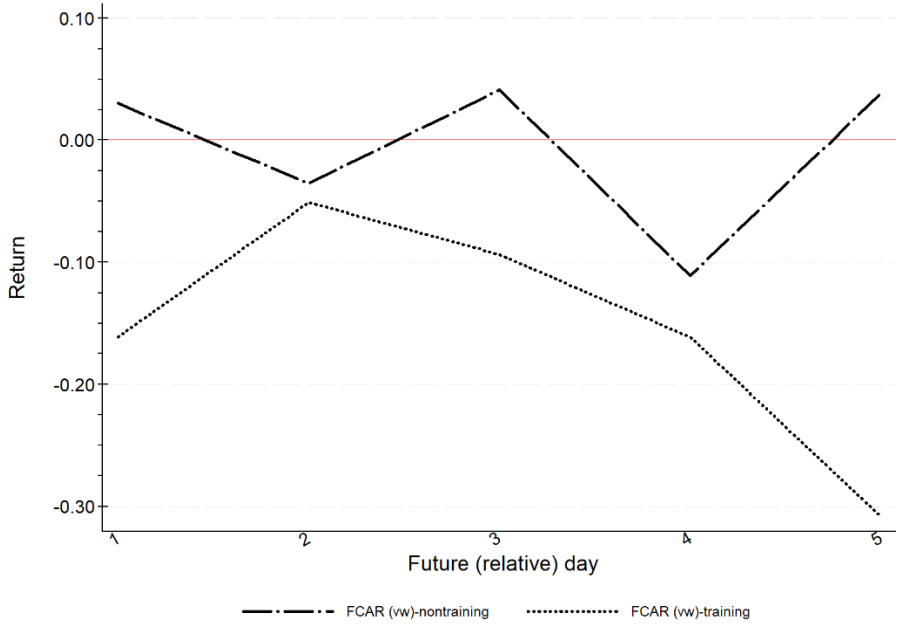
(b) GPT-4

Figure A3: Training-data reliance and future performance of stocks from initial recommendations

This figure plots cumulative value-weighted abnormal returns for portfolios constructed from initial recommendations, separated by information source at the individual stock level. Each day, we classify each of the approximately 20 recommended stocks and create two value-weighted portfolios: one containing stocks where AI agents cited web-searched sources (solid line) and one containing stocks where AI agents relied on training data, i.e., all cited sources were not actually searched (dotted line). Lines show the mean return across all daily portfolios. Day 1 represents the first trading day after the recommendation. Returns are cumulative from Day 1. Panel (a) shows GPT-4o results; Panel (b) shows GPT-4 results.

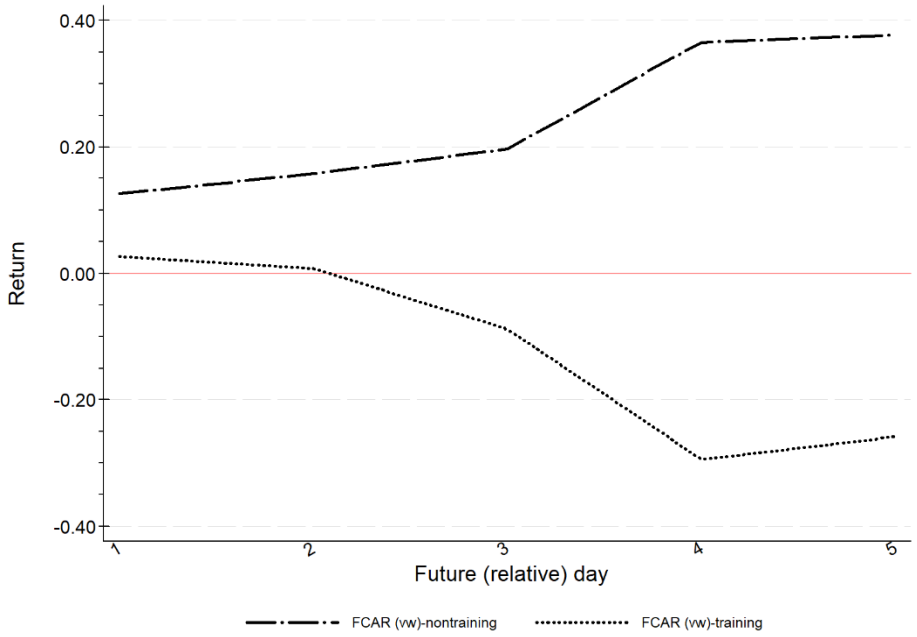


(a) GPT-4o

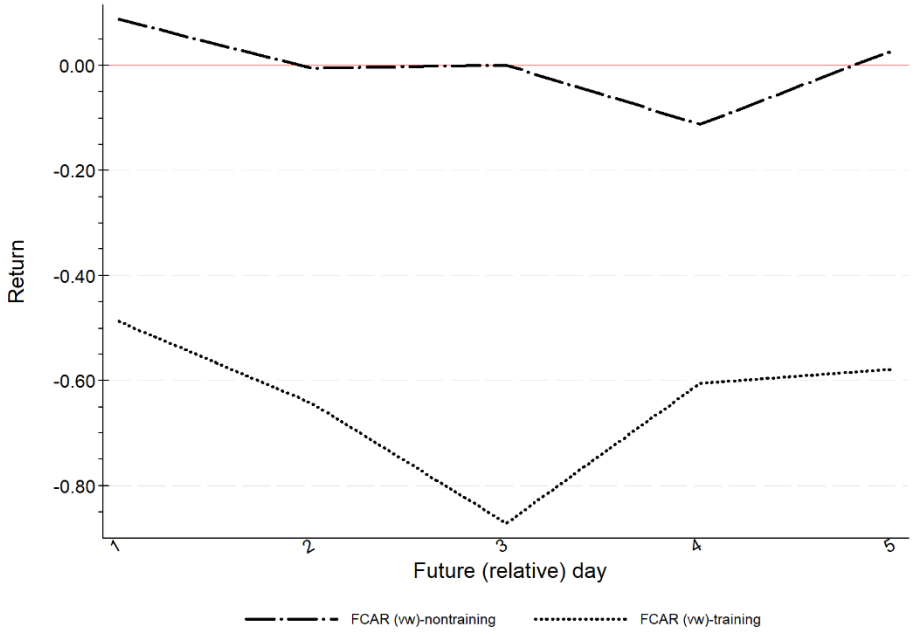


(b) GPT-4

Figure A4: Training-data reliance (self-declared) and the future performance of stocks from initial recommendations
This figure plots cumulative value-weighted abnormal returns for portfolios of stocks from initial recommendations, using an alternative measure of training-data reliance based on AI agents' self-declarations. After receiving the initial 20 stock recommendations, we asked AI agents whether they used training data for those recommendations. The solid line shows the mean return for portfolios from days when AI agents reported using searched information; the dotted line shows the mean return for portfolios from days when AI agents reported using training data. This self-declared measure is a lower-cost but noisier alternative to the source-verification method used in Figure 5. The confirmation prompt asks about all 20 recommendations collectively, so on each day, all stocks are classified into the same category based on the day-level response. AI agents may misclassify the whole portfolio when only some rely on training data while others use web-searched information. Day 1 represents the first trading day after the recommendation. Returns are cumulative from Day 1. Panel (a) shows GPT-4o results; Panel (b) shows GPT-4 results.



(a) GPT-4o



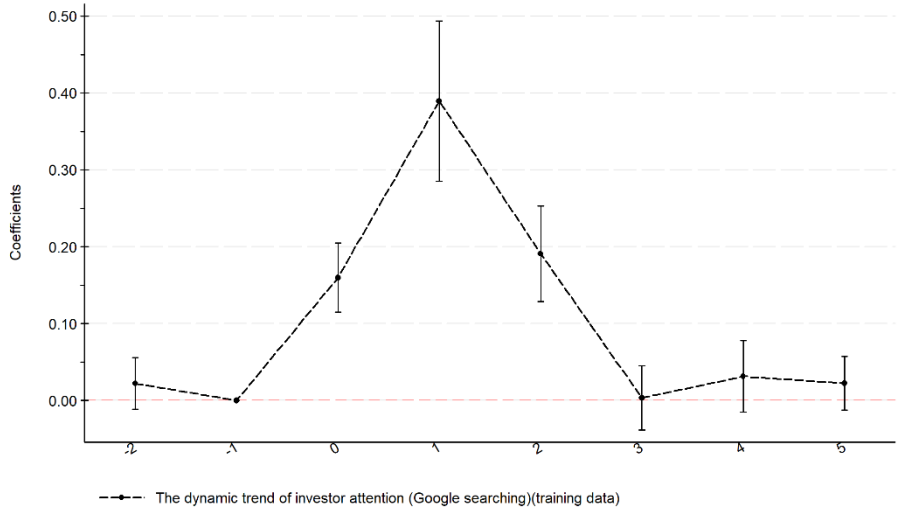
(b) GPT-4

Figure A5: Robustness test for Figure 1

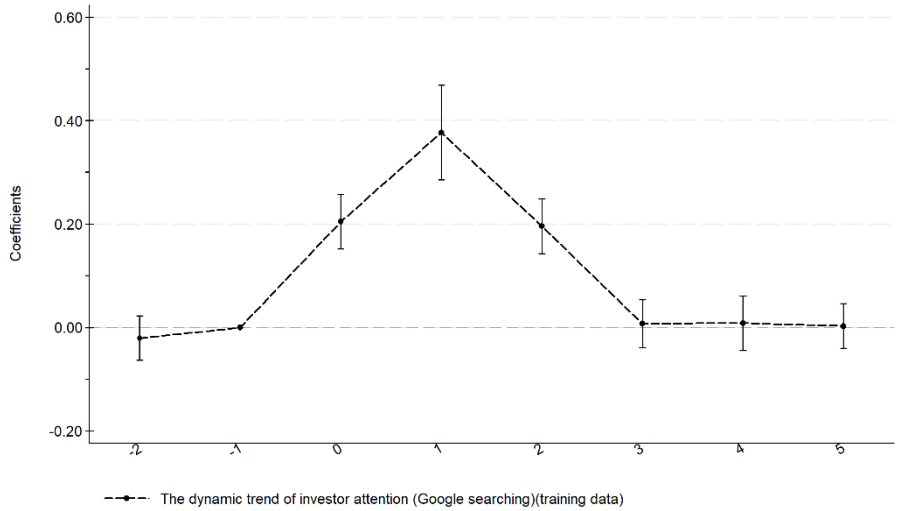
This figure plots coefficient estimates from regressions of Google search volume index (SVI) on relative trading days, partitioned by whether AI agents relied on internal training data. We treat each trading day with recommendations as an event. Within each event *k*, for each recommended stock *i*, we construct a subsample containing observations from two days before to five days after the recommendation day. We then stack these subsamples and regress the natural logarithm of SVI on dummies for relative days using the following model:

$$Ln(SVI)_{k,i,t} = \beta_1 Day(-2)_{k,i,t} + \beta_2 Day(0)_{k,i,t} + \beta_3 Day(1)_{k,i,t} + \beta_4 Day(2)_{k,i,t} + \beta_5 Day(3)_{k,i,t} + \beta_6 Day(4)_{k,i,t} + \beta_7 Day(5)_{k,i,t} + Control\ variables + event * firm\ FE + \varepsilon_{k,i,t},$$

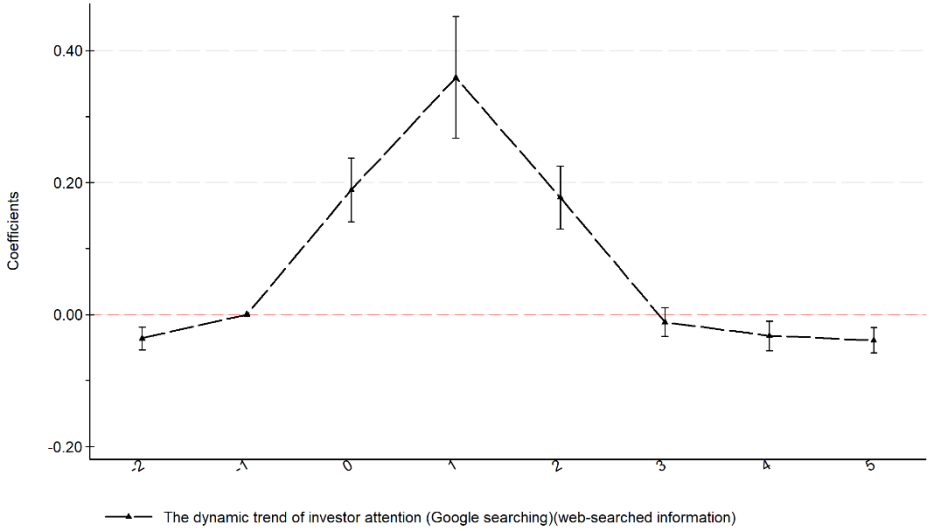
wherein control variables include the natural logarithm of market capitalization on day t-1, the natural logarithm of *Dow Jones Newswire* count on day t-1, the sentiment of *Dow Jones Newswire* on day t-1 (average event sentiment score from the Ravenpack database), standard deviation of daily returns over trading days t-20 to t-1, average daily share turnover over trading days t-20 to t-1, the stock return on day t, the cumulative daily return over trading days t-5 to t-1, and the cumulative daily return over trading days t-15 to t-6. We also control for event*firm fixed effects and use robust standard errors clustered by firm. Day 0 represents the trading day when AI agents provide stock recommendations. The x-axis shows trading days relative to the recommendation date. Dots represent coefficient estimates; vertical lines denote 90% confidence intervals. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. Panels (a) and (b) examine stocks relying on training data; Panels (c) and (d) examine stocks relying on web-searched information. Panels (a) and (c) show GPT-4o results; Panels (b) and (d) show GPT-4 results.



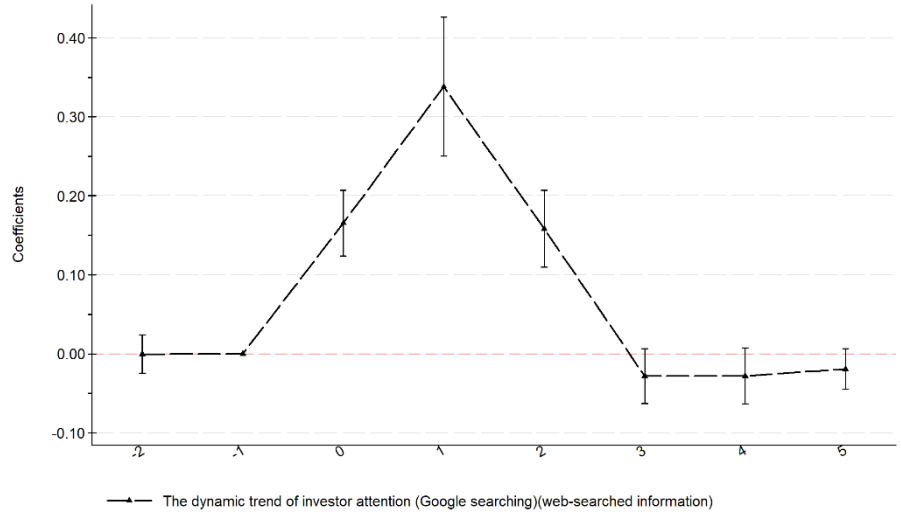
(a) GPT-4o (training)



(b) GPT-4 (training)



(c) GPT-4o (non-training)



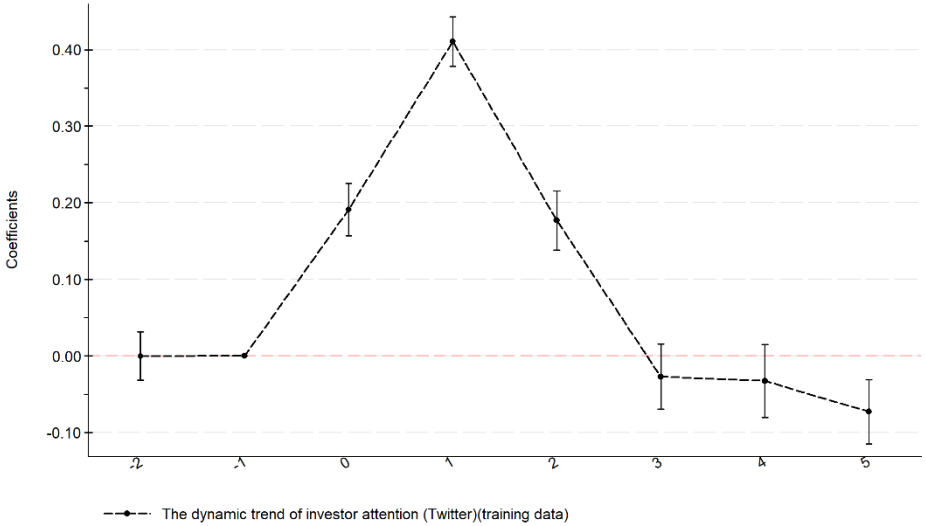
(d) GPT-4 (non-training)

Figure A6: Robustness test for Figure 2

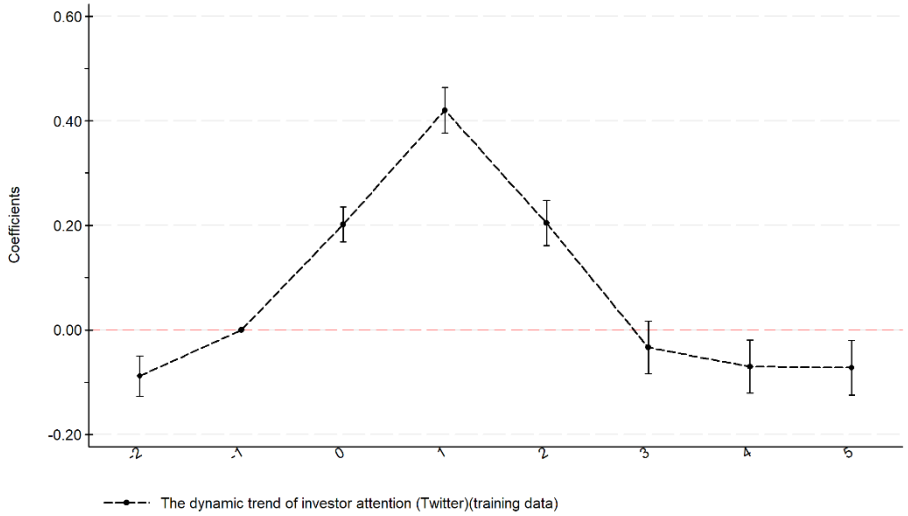
This figure plots coefficient estimates from regressions of investor Twitter discussions for stocks from initial recommendations on relative trading days, partitioned by whether AI agents relied on internal training data. We treat each trading day with recommendations as an event. Within each event *k*, for each recommended stock *i*, we construct a subsample containing observations from two days before to five days after the recommendation day. We then stack these subsamples and regress the natural logarithm of the number of Twitter posts related to stocks on dummies for relative days using the following model:

$$Ln(Tweets)_{k,i,t} = \beta_1 Day(-2)_{k,i,t} + \beta_2 Day(0)_{k,i,t} + \beta_3 Day(1)_{k,i,t} + \beta_4 Day(2)_{k,i,t} + \beta_5 Day(3)_{k,i,t} + \beta_6 Day(4)_{k,i,t} + \beta_7 Day(5)_{k,i,t} + Control\ variables + event * firm\ FE + \varepsilon_{k,i,t},$$

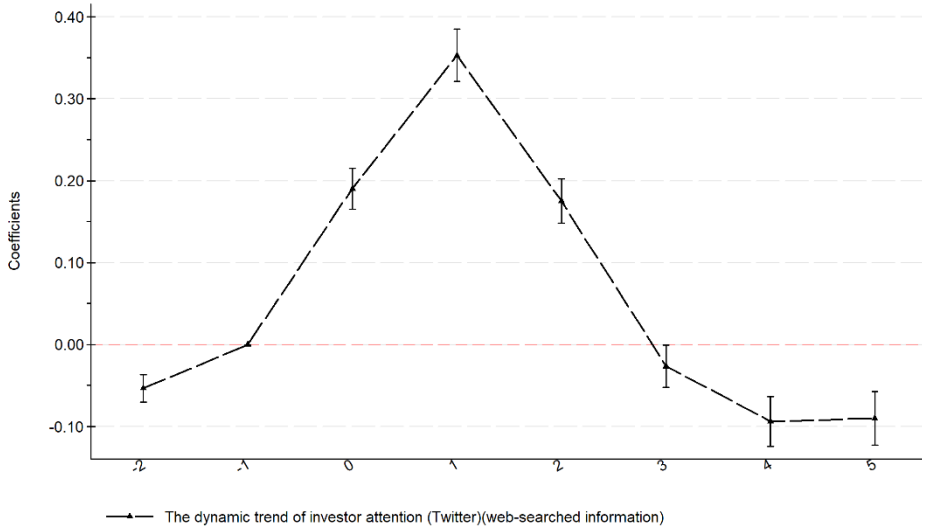
wherein control variables include the natural logarithm of market capitalization on day t-1, the natural logarithm of *Dow Jones Newswire* count on day t-1, the sentiment of *Dow Jones Newswire* on day t-1 (average event sentiment score from the Ravenpack database), standard deviation of daily returns over trading days t-20 to t-1, average daily share turnover over trading days t-20 to t-1, the stock return on day t, the cumulative daily return over trading days t-5 to t-1, and the cumulative daily return over trading days t-15 to t-6. We also control for event*firm fixed effects and use robust standard errors clustered by firm. Day 0 represents the trading day when AI agents provide stock recommendations. The x-axis shows trading days relative to the recommendation date. Dots represent coefficient estimates; vertical lines denote 90% confidence intervals. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for that stock were not actually searched, indicating the information came from internal training data rather than real-time web searches. Panels (a) and (b) examine stocks relying on training data; Panels (c) and (d) examine stocks relying on web-searched information. Panels (a) and (c) show GPT-4o results; Panels (b) and (d) show GPT-4 results.



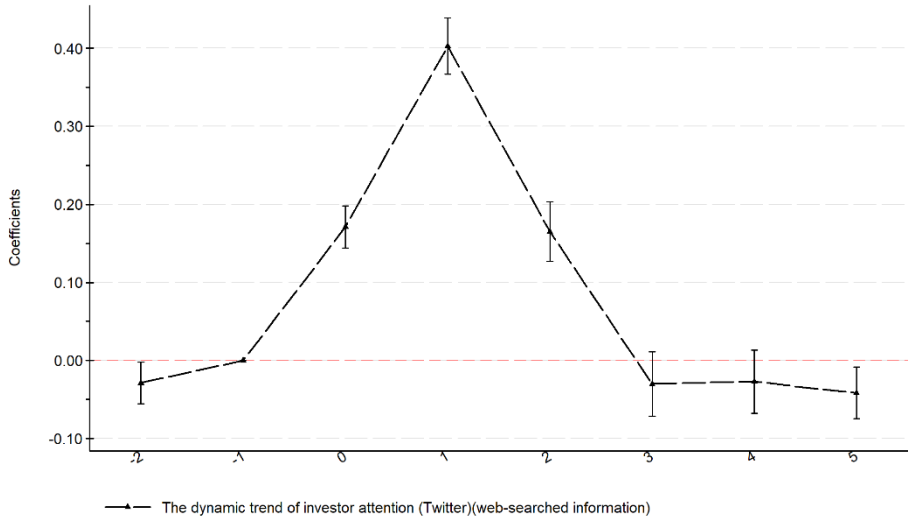
(a) GPT-4o (training)



(b) GPT-4 (training)



(c) GPT-4o (non-training)



(d) GPT-4 (non-training)

Table IA1: Prompt templates

Initial prompt
Assume you are an investor in the US who wants to plan for your investment in the US stock market. You have internet access, so your information set for making your plan includes both macro-level information and recent stock price performance. Let's think step by step. Step 1: Remember that you have access to the Internet. Please choose the top 20 stocks you would like to buy on the following trading day, using the browser tool to gather the latest information from relevant financial websites. Pay special attention to the financial websites that are most commonly used by US investors, such as Yahoo Finance, Bloomberg, Reuters, Morningstar, the Wall Street Journal, Investopedia, Seeking Alpha, MarketWatch, CNBC, Fidelity Investments, etc. Note that you should check well-regarded financial websites, as many as possible, including those listed above. But if you do find only one website, do not pretend you checked more than one website. Step 2: Rank these 20 stocks from 1 to 20 (i.e., give 1 to the stock you are most likely to buy and 20 to the stock you are least likely to buy). List these stocks from rank 1 to rank 20. Step 3: For all 20 stocks you choose, provide each stock with an explanation to support your choice. If you do not know the answer, you can simply say "I don't know." Step 4: Provide the output of your analysis in the YAML format. Specifically, output your analysis of the 20 stocks you chose in the following YAML format: - stock name: [List the stock name, e.g., 'Nvidia', 'Amazon'] - stock symbol: [List of stock symbol, e.g., 'NVDA', 'AMZN'] - ranking: [ranking of a stock, e.g., '1', '20'] - reasoning: "Your explanation for choosing the ranking" - information source: [List the information sources that you actually searched and used, e.g., 'Morningstar', 'Wall Street Journal', 'my training data']

Self-declaration of training data reliance for the initial recommendation (Since October 21, 2024**)**
Did you make the recommendations based on the information you searched? Or did you just make the recommendations based on your training data, without searching the internet? Explain how you did it step by step. Provide the output of your analysis in the following YAML format: - Answer: [the answer of your results, e.g., 'My analysis is based on my searching results.', 'My recommendation is based on my trading data.']

Subsequent prompts (repeated five times)
Please update your list of the top 20 stocks by searching more websites (e.g., Yahoo Finance, Bloomberg, Reuters, Morningstar, the Wall Street Journal, Investopedia, Seeking Alpha, MarketWatch, CNBC, Fidelity Investments, etc.) If your analyses of some stocks remain the same, you do not have to make changes to these stocks. Provide the output of your updated analysis in the YAML format. Specifically, output your analysis of the 20 stocks you chose in the following YAML format: - stock name: [List the stock name, e.g., 'Nvidia', 'Amazon'] - stock symbol: [List of stock symbol, e.g., 'NVDA', 'AMZN'] - ranking: [ranking of a stock, e.g., '1', '20'] - reasoning: "Your explanation for choosing the ranking" - information source: [List the information sources that you actually used in your analysis, e.g., 'Morningstar', 'Wall Street Journal', 'my training data']

Self-declaration of training data reliance for the fifth (final) updated recommendation (Since November 13, 2024**)**
Did you make the recommendations based on the information you searched? Or did you just make the recommendations based on your training data, without searching the internet? Explain how you did it step by step. Provide the output of your analysis in the following YAML format: - Answer: [the answer of your results, e.g., 'My analysis is based on my searching results.', 'My recommendation is based on my trading data.']

Table IA2: One example of GPT-4o's recommendation for stocks to be bought on November 13, 2024

20 stocks recommended for the initial prompt

```
- stock name: 'Apple Inc.'
  stock symbol: 'AAPL'
  ranking: '1'
  reasoning: "Apple continues to demonstrate strong financial performance,
driven by robust sales of its iPhone lineup and growth in services. Its
consistent innovation and loyal customer base make it a compelling
investment."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Microsoft Corporation'
  stock symbol: 'MSFT'
  ranking: '2'
  reasoning: "Microsoft's dominance in cloud computing through Azure, along
with its diversified product portfolio, positions it well for sustained
growth. Its strong financials and strategic acquisitions bolster its market
position."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'NVIDIA Corporation'
  stock symbol: 'NVDA'
  ranking: '3'
  reasoning: "NVIDIA leads in graphics processing units (GPUs) and is
capitalizing on the AI boom. Its products are integral to various high-
growth sectors, including gaming, data centers, and autonomous vehicles."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Amazon.com, Inc.'
  stock symbol: 'AMZN'
  ranking: '4'
  reasoning: "Amazon's e-commerce dominance and expansion into cloud
computing with AWS provide multiple revenue streams. Its continuous
innovation in logistics and technology supports long-term growth prospects."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Alphabet Inc. (Class A)'
  stock symbol: 'GOOGL'
  ranking: '5'
  reasoning: "Alphabet's leadership in online advertising, coupled with
ventures into cloud computing and autonomous vehicles, offers diversified
growth opportunities. Its strong balance sheet supports ongoing innovation."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Meta Platforms, Inc.'
  stock symbol: 'META'
  ranking: '6'
  reasoning: "Meta's vast user base across its social media platforms and
investments in the metaverse position it for future growth. Its advertising
model continues to generate substantial revenue."
  information source: ['Morningstar', 'Wall Street Journal']
```

```
- stock name: 'Tesla, Inc.'
  stock symbol: 'TSLA'
  ranking: '7'
  reasoning: "Tesla's leadership in electric vehicles and advancements in
battery technology drive its growth. Its expansion into energy solutions and
autonomous driving technology adds to its long-term potential."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Berkshire Hathaway Inc. (Class B)'
  stock symbol: 'BRK.B'
  ranking: '8'
  reasoning: "Berkshire Hathaway's diversified holdings across various
industries and strong management under Warren Buffett make it a stable
investment choice. Its financial strength allows for strategic
acquisitions."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Johnson & Johnson'
  stock symbol: 'JNJ'
  ranking: '9'
  reasoning: "Johnson & Johnson's diversified healthcare products and strong
pharmaceutical pipeline provide steady revenue. Its consistent dividend
payments and financial stability appeal to long-term investors."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Visa Inc.'
  stock symbol: 'V'
  ranking: '10'
  reasoning: "Visa's global payment network and the shift towards digital
transactions position it well for continued growth. Its strong profit
margins and strategic partnerships enhance its market presence."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Procter & Gamble Co.'
  stock symbol: 'PG'
  ranking: '11'
  reasoning: "Procter & Gamble's portfolio of leading consumer brands and
global reach provide stable cash flows. Its focus on innovation and
efficiency supports sustained profitability."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'JPMorgan Chase & Co.'
  stock symbol: 'JPM'
  ranking: '12'
  reasoning: "JPMorgan's strong financial performance and leadership in the
banking sector make it a solid investment. Its diversified services and
global presence contribute to its resilience."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Walmart Inc.'
  stock symbol: 'WMT'
  ranking: '13'
```

```
  reasoning: "Walmart's extensive retail network and investments in e-
commerce enhance its competitive position. Its focus on cost leadership and
supply chain efficiency supports its profitability."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Mastercard Incorporated'
  stock symbol: 'MA'
  ranking: '14'
  reasoning: "Mastercard's global payment processing network and the trend
towards cashless transactions drive its growth. Its strategic initiatives in
digital payments and cybersecurity strengthen its market position."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'The Home Depot, Inc.'
  stock symbol: 'HD'
  ranking: '15'
  reasoning: "Home Depot's leadership in the home improvement retail sector
and strong same-store sales growth reflect its operational excellence. Its
focus on professional customers and e-commerce expansion supports future
growth."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Pfizer Inc.'
  stock symbol: 'PFE'
  ranking: '16'
  reasoning: "Pfizer's robust pharmaceutical pipeline and success with
COVID-19 vaccines highlight its innovation capabilities. Its strong cash
flows and dividend yield make it attractive to income-focused investors."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Coca-Cola Company'
  stock symbol: 'KO'
  ranking: '17'
  reasoning: "Coca-Cola's extensive beverage portfolio and global brand
recognition provide stable revenue streams. Its strategic acquisitions and
focus on healthier options align with changing consumer preferences."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'PepsiCo, Inc.'
  stock symbol: 'PEP'
  ranking: '18'
  reasoning: "PepsiCo's diversified product offerings in beverages and
snacks, along with its global distribution network, support consistent
growth. Its investments in healthier products and sustainability initiatives
enhance its brand image."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'AbbVie Inc.'
  stock symbol: 'ABBV'
  ranking: '19'
  reasoning: "AbbVie's strong portfolio of immunology and oncology drugs,
along with its acquisition of Allergan, bolster its revenue base. Its focus
on research and development supports long-term growth prospects."
```

```
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Intel Corporation'
  stock symbol: 'INTC'
  ranking: '20'
  reasoning: "Intel's leadership in semiconductor manufacturing and
investments in next-generation technologies position it for future growth.
Its focus on expanding production capacity addresses supply chain
challenges."
  information source: ['Morningstar', 'Wall Street Journal']
```

Self-declaration of training data reliance for the initial recommendation

```
- Answer: "My recommendations are based on my training data. While I initiated
the search process, I did not use specific up-to-date information from the
internet to finalize the stock rankings and reasoning. Instead, the analysis
reflects general insights from my training, which includes historical
patterns, common investment rationales, and publicly known information up to
2023."
```

20 stocks recommended for the 5th update

```
- stock name: 'NVIDIA Corporation'
  stock symbol: 'NVDA'
  ranking: '1'
  reasoning: "NVIDIA's leadership in AI and graphics processing has driven
its market capitalization to over $3 trillion. The company's recent
inclusion in the Dow Jones Industrial Average and strong financial
performance underscore its growth trajectory."
  information source: ['Reuters', 'MarketWatch', 'Wall Street Journal']

- stock name: 'Tesla, Inc.'
  stock symbol: 'TSLA'
  ranking: '2'
  reasoning: "Tesla's stock surged following the U.S. presidential election,
with expectations of reduced regulations and potential government roles for
CEO Elon Musk. The company's innovation in electric vehicles and energy
solutions continues to drive investor confidence."
  information source: ['Investopedia', 'MarketWatch', 'Reuters']

- stock name: 'JPMorgan Chase & Co.'
  stock symbol: 'JPM'
  ranking: '3'
  reasoning: "JPMorgan's shares rose significantly post-election, driven by
prospects of deregulation and improved business sentiment. As a leading
financial institution, it stands to benefit from a pro-business
environment."
  information source: ['MarketWatch', 'Wall Street Journal']

- stock name: 'Amazon.com, Inc.'
```

```
  stock symbol: 'AMZN'
  ranking: '4'
  reasoning: "Amazon's dominance in e-commerce and cloud computing, coupled
with continuous innovation, positions it well for sustained growth. Its
diverse revenue streams and global reach make it a compelling investment."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Apple Inc.'
  stock symbol: 'AAPL'
  ranking: '5'
  reasoning: "Apple's strong financial performance, driven by robust sales
of its iPhone lineup and growth in services, continues to make it a solid
investment. Its consistent innovation and loyal customer base are key
strengths."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Microsoft Corporation'
  stock symbol: 'MSFT'
  ranking: '6'
  reasoning: "Microsoft's dominance in cloud computing through Azure, along
with its diversified product portfolio, positions it well for sustained
growth. Its strong financials and strategic acquisitions bolster its market
position."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Meta Platforms, Inc.'
  stock symbol: 'META'
  ranking: '7'
  reasoning: "Meta's vast user base across its social media platforms and
investments in the metaverse position it for future growth. Its advertising
model continues to generate substantial revenue."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Alphabet Inc. (Class A)'
  stock symbol: 'GOOGL'
  ranking: '8'
  reasoning: "Alphabet's leadership in online advertising, coupled with
ventures into cloud computing and autonomous vehicles, offers diversified
growth opportunities. Its strong balance sheet supports ongoing innovation."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Berkshire Hathaway Inc. (Class B)'
  stock symbol: 'BRK.B'
  ranking: '9'
  reasoning: "Berkshire Hathaway's diversified holdings across various
industries and strong management under Warren Buffett make it a stable
investment choice. Its financial strength allows for strategic
acquisitions."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Johnson & Johnson'
  stock symbol: 'JNJ'
  ranking: '10'
```

```
  reasoning: "Johnson & Johnson's diversified healthcare products and strong
pharmaceutical pipeline provide steady revenue. Its consistent dividend
payments and financial stability appeal to long-term investors."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Visa Inc.'
  stock symbol: 'V'
  ranking: '11'
  reasoning: "Visa's global payment network and the shift towards digital
transactions position it well for continued growth. Its strong profit
margins and strategic partnerships enhance its market presence."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Procter & Gamble Co.'
  stock symbol: 'PG'
  ranking: '12'
  reasoning: "Procter & Gamble's portfolio of leading consumer brands and
global reach provide stable cash flows. Its focus on innovation and
efficiency supports sustained profitability."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Mastercard Incorporated'
  stock symbol: 'MA'
  ranking: '13'
  reasoning: "Mastercard's global payment processing network and the trend
towards cashless transactions drive its growth. Its strategic initiatives in
digital payments and cybersecurity strengthen its market position."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'The Home Depot, Inc.'
  stock symbol: 'HD'
  ranking: '14'
  reasoning: "Home Depot's leadership in the home improvement retail sector
and strong same-store sales growth reflect its operational excellence. Its
focus on professional customers and e-commerce expansion supports future
growth."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Pfizer Inc.'
  stock symbol: 'PFE'
  ranking: '15'
  reasoning: "Pfizer's robust pharmaceutical pipeline and success with
COVID-19 vaccines highlight its innovation capabilities. Its strong cash
flows and dividend yield make it attractive to income-focused investors."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Coca-Cola Company'
  stock symbol: 'KO'
  ranking: '16'
  reasoning: "Coca-Cola's extensive beverage portfolio and global brand
recognition provide stable revenue streams. Its strategic acquisitions and
focus on healthier options align with changing consumer preferences."
  information source: ['Morningstar', 'Wall Street Journal']
```

```
- stock name: 'PepsiCo, Inc.'
  stock symbol: 'PEP'
  ranking: '17'
  reasoning: "PepsiCo's diversified product offerings in beverages and
snacks, along with its global distribution network, support consistent
growth. Its investments in healthier products and sustainability initiatives
enhance its brand image."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'AbbVie Inc.'
  stock symbol: 'ABBV'
  ranking: '18'
  reasoning: "AbbVie's strong portfolio of immunology and oncology drugs,
along with its acquisition of Allergan, bolster its revenue base. Its focus
on research and development supports long-term growth prospects."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Intel Corporation'
  stock symbol: 'INTC'
  ranking: '19'
  reasoning: "Intel's leadership in semiconductor manufacturing and
investments in next-generation technologies position it for future growth.
Its focus on expanding production capacity addresses supply chain
challenges."
  information source: ['Morningstar', 'Wall Street Journal']

- stock name: 'Micron Technology, Inc.'
  stock symbol: 'MU'
  ranking: '20'
  reasoning: "Micron topped the list of the 20 biggest winners in the S&P
500 this earnings season, demonstrating significant revenue growth and
improved profit margins, indicating strong pricing power and operational
efficiency."
  information source: ['MarketWatch']
```

Self-declaration of training data reliance for the initial recommendation

```
- Answer: "My recommendations are based on a combination of my training data
and the information I searched online. I used online resources such as Yahoo
Finance, Bloomberg, Reuters, Morningstar, and MarketWatch to update and
validate the analysis of the stocks. Any rankings or reasoning that remained
unchanged reflect consistency between my training data and the latest
information gathered from these sources."
```

Table IA3: Prompt for reasoning classification

Analyze these stock recommendation reasons and categorize them using the criteria below.

Output format: Provide only the category numbers (e.g., "1, 4, 7"). Include ALL applicable categories - most reasoning statements will fit multiple categories.

Categories:

1. Growth drivers & expansion Future growth prospects, market expansion, new markets, TAM growth, demand trends, scaling opportunities. *Examples: "strong growth prospects," "expansion into emerging markets," "secular growth in AI," "market expansion opportunities"*

2. Technology & innovation leadership AI capabilities, technological advancement, R&D pipeline, digital transformation, software/platform leadership, semiconductor innovation. *Examples: "AI leadership," "cloud computing dominance," "innovative technology platform," "digital transformation leader"*

3. Market leadership & competitive position Market dominance, competitive moats, brand strength, market share gains, industry leadership, competitive advantages. *Examples: "leading market position," "dominant player," "strong competitive moat," "market share leader"*

4. Financial performance & metrics Earnings growth, revenue trends, margin expansion, cash flow strength, profitability improvements, financial beats/misses. *Examples: "strong earnings growth," "revenue beat," "margin expansion," "robust cash generation"*

5. Sector & industry tailwinds Industry-specific growth, sector rotation, regulatory benefits, commodity cycles, macro industry trends. *Examples: "renewable energy sector growth," "healthcare sector strength," "industry consolidation benefits"*

6. Strategic initiatives & corporate actions M&A activity, partnerships, strategic deals, business model changes, restructuring, new product launches. *Examples: "strategic acquisition," "partnership announcement," "business restructuring," "new product launch"*

7. Valuation opportunity & price dislocation Undervaluation, attractive entry points, price corrections, value opportunities, multiple compression/expansion. *Examples: "undervalued after selloff," "attractive valuation," "trading below fair value," "good entry point"*

8. Income & shareholder returns Dividend yield, dividend growth, share buybacks, capital return policies, REIT distributions. *Examples: "attractive dividend yield," "share repurchase program," "dividend increase," "strong capital returns"*

9. Recovery & turnaround themes Business recovery, cyclical upturns, turnaround situations, post-crisis rebounds, operational improvements. *Examples: "recovery play," "turnaround story," "cyclical upturn," "operational improvements"*

10. Analyst actions & rating changes Analyst upgrades/downgrades, price target changes, recommendation changes, consensus revisions. *Examples: "analyst upgrade," "price target increase," "buy rating," "consensus improvement"*

11. Event-driven catalysts Specific upcoming events, regulatory approvals, clinical trial results, earnings guidance, contract wins. *Examples: "FDA approval pending," "earnings guidance update," "major contract win," "regulatory decision"*

12. Other/Multiple themes Complex reasoning spanning many categories or unclear/vague statements that don't fit clearly above.

Key Instructions:

- **Most statements will have 2-4 categories** - include all that apply
- Technology companies may often include categories 1, 2, and 3 together
- Financial performance (category 4) often combines with other fundamental themes
- If reasoning mentions specific sector context, include category 5
- For vague or very general statements, use category 12

Table IA4: Web-searching bias and future performance of stocks from the initial recommendations

This table reports value-weighted abnormal returns for portfolios of stocks from initial recommendations that relied on web-searched information, partitioned by search concentration. We classify each day's recommendation task based on ChatGPT's search concentration pattern for that day. Abnormal return equals the raw return minus the corresponding CRSP market return. Average abnormal returns are reported above, with standard deviations in parentheses below. Because recommendations are made after market close, investors can only purchase stocks at the opening price the next day. Therefore, the Day 1 return is calculated as the closing price on Day 1 minus the opening price on Day 1, divided by the opening price on Day 1. The bottom rows report cumulative abnormal returns for Days 1-5. We classify a day as having concentrated searching if ChatGPT only cited the top 3 websites for GPT-4o (led by Yahoo Finance at 3.128 searches per day, with a large drop to 0.312 for the 4th website, Motley Fool) or the top 7 websites for GPT-4 (led by Bing at 0.843 searches per day, with a drop to 0.077 for the 8th website, Nasdaq). Other days are considered as having diverse searching. This analysis excludes recommendations relying solely on training data. We classify a recommendation as relying on training data if the websites cited as information sources for this recommendation were not actually searched, indicating the information came from internal training data rather than real-time web searches.

	GPT-4o		GPT-4	
	Concentrated	Diverse	Concentrated	Diverse
Day 1	0.277	0.051	-0.005	0.098
	(1.178)	(1.618)	(1.374)	(3.198)
Day 2	-0.005	-0.040	0.038	-0.288
	(1.094)	(1.105)	(1.235)	(1.629)
Day 3	-0.028	-0.004	0.007	0.215
	(0.996)	(1.066)	(1.070)	(2.234)
Day 4	-0.039	0.134	-0.125	-0.208
	(0.781)	(1.100)	(1.221)	(1.961)
Day 5	-0.038	0.078	0.040	0.345
	(0.639)	(1.191)	(1.161)	(1.935)
Days 1-5	0.168	0.219	-0.044	0.162
	(2.310)	(2.210)	(2.660)	(4.895)

Table IA5: Representative heuristic reasoning and future performance of stocks from initial recommendations

This table reports value-weighted abnormal returns for portfolios of stocks from initial recommendations that relied on web-searched information, partitioned by whether AI agents used valuation-focused reasoning. Abnormal return equals the raw return minus the corresponding CRSP market return. Average abnormal returns are reported above, with standard deviations in parentheses below. Because recommendations are made after market close, investors can only purchase stocks at the opening price the next day. Therefore, the Day 1 return is calculated as the closing price on Day 1 minus the opening price on Day 1, divided by the opening price on Day 1. The bottom rows report cumulative abnormal returns for Days 1-5. We separate stocks into two groups: those where AI agents cited "Valuation opportunity & price dislocation" as a reasoning theme versus those citing other reasoning themes (e.g., growth drivers & expansion, technology & innovation leadership). We classify reasoning using the prompt in Table IA3. Reasoning classifications are defined in Table 5. This analysis excludes recommendations relying solely on training data. We classify each individual stock recommendation as relying on training data if the websites cited as information sources for this recommendation were not actually searched, indicating the information came from internal training data rather than real-time web searches.

	GPT-4o		GPT-4	
	Value-focused reasoning	Non-value-focused reasoning	Value-focused reasoning	Non-value-focused reasoning
Day 1	0.169	0.086	-0.993	0.038
	(1.553)	(1.550)	(7.757)	(2.162)
Day 2	0.089	-0.056	1.315	-0.034
	(1.722)	(1.116)	(13.941)	(1.399)
Day 3	-0.096	-0.043	-0.615	0.080
	(1.586)	(1.131)	(3.806)	(1.602)
Day 4	-0.067	0.134	-0.682	-0.153
	(1.616)	(1.201)	(4.232)	(1.531)
Day 5	0.330	0.034	0.841	0.144
	(1.237)	(1.133)	(6.484)	(1.509)
Days 1-5	0.426	0.154	-0.134	0.076
	(3.827)	(2.264)	(13.613)	(3.503)