When Machines Disagree: Evidence from Large Language Models

Si Cheng Lin Hu Kun Li

February 24, 2026

Abstract

Using six leading large language models to extract sentiment from news headlines and predict stock returns, we document substantial variation across model providers and investment horizons. This model disagreement reflects systematic differences in causal reasoning rather than noise. Stocks with higher cross-provider and cross-horizon dispersions earn lower future returns, especially among firms with more opaque information environments and greater operating uncertainty. Model disagreement amplifies post-earnings-announcement drift for small firms, while delaying the immediate price reaction to earnings announcements for large firms. Model disagreement also increases both overall and retail trading volume. Evidence from exogenous ChatGPT outages reinforces our conclusions. Overall, our findings highlight the growing role of AI in shaping investor beliefs, stock price dynamics, and price informativeness.

Keywords: Large language model, generative AI, model disagreement, sentiment, investment horizon, return prediction, price efficiency

JEL Code: G10, G11, G12, G14, C53

Contact Information: Cheng: scheng24@syr.edu, Syracuse University; Hu: lin.hu@anu.edu.au, Australian National University; Li: kun.li@anu.edu.au, Australian National University.

1 Introduction

Generative AI tools such as ChatGPT, Copilot, and Gemini have advanced at an unprecedented pace in recent years. Shortly after its November 2022 release as a research preview, ChatGPT became the fastest app ever to reach 100 million monthly active users—a milestone achieved within just two months.[1] As of 2025, ChatGPT has approximately 160 million daily active users and processes around 2.5 billion prompts from users worldwide every day.[2]

Generative AI, often powered by large language models (LLMs), is capable of generating human-like text, interpreting complex data, and performing a wide range of cognitive tasks with increasing accuracy. A growing body of research documents its potential to enhance financial decision-making by summarizing complex corporate disclosures (Kim et al., 2024a; Wong et al., 2025), extracting nuanced and hard-to-measure information and facilitating trading (Bai et al., 2023; Bernard et al., 2024; Kim et al., 2024c; Chang et al., 2025; Cheng et al., 2025; Jha et al., 2025; Lyonnet et al., 2025), and forecasting earnings and returns (Lopez-Lira and Tang, 2023; Chen et al., 2024; Kim et al., 2024b; Bertomeu et al., 2025; Chen et al., 2025).

Despite the rapid development and widespread adoption of generative AI tools, whether and how these tools shape investor beliefs and behaviors, and subsequently affect stock price dynamics and price informativeness, remains an open question. While most prior studies focus on a single LLM (e.g., ChatGPT) or include a few models for comparison or robustness checks, we bridge this gap by examining six leading LLM providers—ChatGPT, Copilot, Gemini, Claude, LLaMA, and Mistral—which collectively account for over 90% of usage. In addition, unlike prior work primarily investigating whether LLMs can extract valuable financial signals, our focus lies on the potential disagreement among LLMs. Specifically, do LLMs exhibit significant disagreement when processing the same information? If so,

[1] See, Perez (2025), "ChatGPT doubled its weekly active users in under 6 months, thanks to new releases".
[2] See, Barr (2025), "ChatGPT is crushing Google in the AI race. Unless you look at the data differently." and Silberling (2025), "ChatGPT users send 2.5 billion prompts a day".

what drives this disagreement? How does it affect stock returns, price informativeness, and trading activity?

Ex ante, the extent to which these LLMs disagree with one another remains an empirical question. On one hand, these models are pre-trained on enterprise-scale platforms with a vast amount of text data and billions of parameters, and thus may extract similar signals from financial information. On the other hand, unlike scientific facts, financial content is often open to divergent interpretations based on specific priors and assumptions. Therefore, different LLMs may reach different conclusions when processing the same information. Despite their superior ability to process complex information, conflicting signals from LLMs may amplify information uncertainty in financial markets and potentially distort asset prices. Given the growing adoption of generative AI for processing financial information and informing investment decisions (e.g., Blankespoor et al., 2024; Chang et al., 2025), our findings have important implications for investors, practitioners, and policy makers.

We analyze a comprehensive dataset of news headlines for U.S. common stocks from January 2023 to December 2024, during which generative AI tools became widely adopted. Each LLM is prompted separately to assess how a given headline might affect the corresponding firm's stock price and to generate a set of recommendations. These include a categorical response (good, bad, or unknown) and a continuous sentiment score ranging from 0 to 100 across three prediction horizons: the next day, the next week, and the next month.[3] This prompt design allows us to capture two types of model disagreement. First, for each news at each prediction horizon, we compute *cross-provider dispersion*, defined as the standard deviation of LLM-based news sentiment across providers. Second, we calculate *cross-horizon dispersion* by first computing the standard deviation of sentiment across the three prediction horizons for each news-provider pair, and then averaging these values across providers for each news.

As a starting point, we find substantial cross-provider dispersion across all sentiment measures and prediction horizons, with the highest dispersion observed for next-week

[3]LLMs are instructed to assign sentiment scores as follows: 0–20 for strongly negative sentiment, 21–40 for negative, 41–60 for neutral/mixed, 61–80 for positive, and 81–100 for strongly positive.

predictions. Notably, this dispersion accounts for 12% to 39% of the sample average—even for a relatively simple task of classifying news as good, bad, or unknown. In addition, we document significant cross-horizon dispersion across all sentiment measures. Since all LLMs receive the same news headlines and prompts, these findings suggest that LLMs not only differ in their interpretation of public signals but also adjust their sentiment assessments based on the investment horizon.

Next, we examine the determinants of model disagreement. Unlike human readers, who often struggle to process complex information effectively, LLMs exhibit less disagreement on complex news for next-day and next-week predictions, as measured by word count, the Fog index, and the proportion of complex words. In contrast, cross-provider dispersion increases with news complexity when making next-month predictions.

We then leverage the explanations accompanying each LLM's sentiment assessment to analyze the causal reasoning providers articulate when justifying their classifications. Using directed acyclic graphs (DAGs) to extract cause–effect chains from these explanations, we show that provider disagreement is not merely noise. Instead, disagreement among LLM providers is systematically linked to differences in causal reasoning, specifically in (i) the causes and effects extracted from news, (ii) the mapping of specific cause and effect categories to impacts, and (iii) the clustering and cross-horizon stability of providers' reasoning styles.

Turning to the return predictions, we first confirm and extend prior work by demonstrating return predictability across a broader set of LLMs. However, a substantial portion (if not all) of the sentiment information appears to be incorporated into stock prices on the first trading day, regardless of the forecast horizon. Focusing on next-day predictions, a one-standard-deviation increase in the average sentiment rank $AvgRank$ is associated with a 0.13% increase in next-day returns.

Next, we examine how model disagreement affects future stock returns. Consistent with Miller (1977), when differences of opinion are combined with short-sale constraints, equity prices tend to reflect the views of more optimistic investors, resulting in lower future returns.

We find that cross-provider and cross-horizon dispersions are not significantly related to next-day returns, but are negatively associated with returns over the remainder of the week. Specifically, a one-standard-deviation increase in cross-provider dispersion $StdRank$ and cross-horizon dispersion $StdRankHor$ is associated with a 0.12% and 0.14% decline in next-week returns, excluding the first day. The negative relation between cross-provider dispersion and stock returns also persists over the monthly horizon. In addition, the return predictability associated with model disagreement is more pronounced for firms with more opaque information environments and greater operating uncertainty.

To strengthen causal inference, we exploit ChatGPT outages as exogenous shocks to cross-provider dispersion. Given ChatGPT's dominant market share among AI tools, unexpected outages are likely to attenuate the return predictability of cross-provider dispersion. Consistent with this prediction, the return predictability associated with cross-provider dispersion $StdRank$ is reduced by 67% during outage periods. Overall, LLM-based model disagreement provides a novel proxy for heterogeneous beliefs in the market and has meaningful asset pricing implications.

Furthermore, we investigate whether model disagreement on the announcement day affects the speed at which earnings information is incorporated into stock prices. For small firms with market capitalizations below the NYSE median breakpoint, LLM-based horizon disagreement in news headlines amplifies price underreaction to earnings news, contributing to the post-earnings-announcement drift. In contrast, for large firms, which typically receive greater investor attention and exhibit higher transparency, both cross-provider and cross-horizon dispersion appear to delay the immediate price reaction on the announcement day. These findings suggest that model disagreement can impede the incorporation of earnings information into prices and, in turn, reduce price informativeness.

Finally, we analyze whether model disagreement affects trading activity. We find that both cross-provider and cross-horizon dispersion are associated with elevated trading volume. A one-standard-deviation increase in cross-provider dispersion $StdRank$ and cross-horizon dispersion $StdRankHor$ is associated with a 4.64% and 7.04% increase in next-day abnormal trading volume, respectively. The effect persists for at least a week.

Focusing specifically on retail trades, both types of dispersion are related to higher retail buy and sell volumes, while their impact on retail order imbalance is largely insignificant.

Collectively, these findings suggest that as generative AI tools become increasingly integrated into financial information processing, model disagreement may serve as an additional source of belief dispersion in the market, amplifying information uncertainty and reducing price efficiency.

This paper contributes to several strands of the literature. First, the paper relates to recent work on the role of AI in information processing in financial markets. While prior studies highlight the capabilities of individual models to summarize complex corporate disclosures (Kim et al., 2024a; Wong et al., 2025), extract nuanced and hard-to-measure information and facilitate trading (Bai et al., 2023; Bernard et al., 2024; Kim et al., 2024c; Chang et al., 2025; Cheng et al., 2025; Jha et al., 2025; Lyonnet et al., 2025), forecast earnings and returns (Lopez-Lira and Tang, 2023; Chen et al., 2024; Kim et al., 2024b; Bertomeu et al., 2025; Chen et al., 2025), and simulate economic forecasts (Bybee, 2025; Hansen et al., 2025), we focus on the disagreement among LLMs when processing the same information. Using six leading LLM providers to assess news sentiment, we document substantial cross-provider dispersion across all sentiment measures and prediction horizons. Although generative AI tools significantly reduce investors' information processing costs, our findings suggest that users should exercise independent judgment rather than fully rely on AI-generated recommendations.

Second, our findings contribute to the literature on heterogeneous beliefs in financial markets, particularly when disagreement arises from differential interpretation of public signals (e.g., Harris and Raviv, 1993; Kandel and Pearson, 1995; Banerjee and Kremer, 2010; Li et al., 2022). Prior empirical work has identified a range of sources of investor and analyst disagreement, including culture and language (Chang et al., 2014), partisanship (Cookson et al., 2020), investment philosophies (Cookson and Niessner, 2020), geographic proximity (Gerken and Painter, 2023), and investment horizons (Cookson et al., 2024). We examine a novel aspect of investor disagreement: variation in how LLMs interpret the same publicly available news, both across model providers and across investment horizons.

Our findings highlight a potential unintended consequence that the widespread use of generative AI may amplify systematic belief dispersion among market participants, leading to possible price distortions. Moreover, a contemporaneous study by Bali et al. (2025) proposes a statistical model of heterogeneous beliefs where investors are represented by different machine learning model specifications. We instead focus on disagreement among six leading LLMs, capturing the real-world adoption of generative AI tools by investors.

The remainder of this paper proceeds as follows. Section 2 outlines the LLMs and empirical methodology. Section 3 describes the data and main variables. Section 4 analyzes the determinants of model disagreement. Section 5 studies the asset pricing implications. Section 6 examines the relation between model disagreement and trading activity. A brief conclusion follows in Section 7.

2 Large Language Models and Prompt

LLMs are now an integral part of the information environment faced by investors. Because their design and deployment influence how financial news is framed and interpreted, understanding their behavior is essential for studying market-relevant information flows. This section introduces the major LLM providers and describes the structured prompt we employ to generate consistent, comparable forecasts across models.

2.1 LLM Providers

Table 1: Top 5 Generative AI Chatbots in the US (2025)

Rank	Chatbot	Description	Market Share	Backend LLM
1	ChatGPT	General-purpose AI chatbot	59.7%	GPT-3.5, GPT-4
2	Microsoft Copilot	General-purpose AI assistant	14.4%	GPT-4
3	Google Gemini	General-purpose AI assistant	13.5%	Gemini
4	Perplexity AI	Accuracy-focused AI search engine	6.2%	Mistral 7B, Llama 2
5	Claude AI	Business-focused AI assistant	3.2%	Claude 3

To ensure relevance to real-world user behavior, we focus on LLMs that underpin the most widely used generative AI chatbots. Table 1 reports the market share of leading AI assistants in the United States as of 2025, measured by user base.[4] The top platforms—ChatGPT, Microsoft Copilot (via Azure OpenAI), Google Gemini, and Claude—collectively account for nearly 90% of usage. These systems rely on proprietary models developed by OpenAI, Google DeepMind, and Anthropic, and are embedded in widely accessible interfaces such as ChatGPT, Bing, Google Search, and Claude.ai. Their widespread availability makes them especially relevant for retail users, who interact with these tools through web-based assistants rather than programmatic APIs. For instance, Blankespoor et al. (2024) find that nearly half of surveyed investors already use generative AI, primarily to gather and interpret financial or market data. We focus on these providers to capture the dominant closed-model environments shaping real-time information access for investors.

To complement these major systems, we also incorporate open-weight models from Meta and Mistral. While their direct market share is smaller, they play an increasingly important role in the generative AI ecosystem. These models are often integrated into emerging applications such as Perplexity (which blends OpenAI, Mistral, and LLaMA backends) and Brave's Leo chatbot (based on LLaMA). Open models are particularly common in privacy-sensitive or developer-led deployments, including browser extensions, community-run bots, and trading platforms frequented by individual investors.

Together, the six selected LLM providers represent a spectrum of design philosophies, ranging from fully closed, highly aligned systems to lightweight, open-access alternatives. All serve as plausible proxies for the AI technologies shaping the beliefs and behaviors of investors. By covering both market-dominant and emerging models, our analysis captures the heterogeneous yet realistic information environments encountered by everyday users of generative AI. Internet Appendix Section A provides detailed summaries of each model used.

[4]Market share estimates are based on a May 2025 report by First Page Sage, available at https://firstpagesage.com/seo-blog/generative-ai-statistics/ (accessed July 2025).

2.2 Prompt

To extract structured forecasts from the chatbot, we implement a custom prompt designed to emulate a professional financial analyst's rapid-reaction framework. This prompt instructs the chatbot to assume the role of a seasoned financial expert and assess how a given news headline about a specified company might affect its stock price over three distinct horizons: the next day, the next week, and the next month. To ensure output consistency, the prompt enforces strict formatting rules and categorical response conventions. We also set the model temperature to zero to maximize result reproducibility and minimize the likelihood of hallucinations. The prompt used is as follows:

> Start fresh with following instructions:
> *You are a financial expert with extensive stock recommendation experience.*
> Analyze how this news headline ‘‘`[HEADLINE]`’’ about `[COMPANY]` could affect its stock price.
> Important: Even if the headline refers to a future event, provide your analysis based on typical patterns and expected outcomes. Do not ask for clarification or wait for actual results.
> *You MUST provide your analysis in the exact table format below, regardless of timing or data availability:*

Timeframe	Impact	Sentiment Score (0–100)	Explanation (< 30 words)
Next Day	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]
Next Week	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]
Next Month	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]

The response must strictly adhere to the following constraints:

- The `Impact` column must use *exactly one* of three labels: GOOD, BAD, or UNKNOWN. Combinations or other variations (e.g., “somewhat good”, “neutral”) are explicitly disallowed.

- The `Sentiment Score` is an independent measure of sentiment intensity, on a 0–100 scale: 0–20 for strongly negative sentiment, 21–40 negative, 41–60 neutral/mixed, 61–80 positive, and 81–100 strongly positive.

- The `Timeframe Analysis` reflects immediate market reaction (next day), short-term analyst and investor response (next week), and expected medium-run adjustment to fundamentals (next month).

Compared to existing prompts in the literature (e.g., Lopez-Lira and Tang, 2023; Chen et al., 2025), which focus on categorical recommendation outputs (YES/NO/UNKNOWN) over a single prediction horizon, our prompt generates a rich set of outputs with greater granularity across multiple time horizons. Its strict categorical format helps isolate model-driven variation in directional predictions from changes in sentiment intensity. This design allows us to capture and analyze response heterogeneity across providers and investment horizons. In addition, the prompt's format and language are designed to be easily interpretable by non-expert retail investors and compatible with spreadsheet-based workflows or automated trading systems.

2.3 Illustrative Prompt Outcomes

To illustrate the interpretive variation introduced by LLMs under our structured prompt, we present an example using the headline *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*, with the company explicitly specified as *"AMC Entertainment Holdings Inc."* in the input.[5] This announcement, while financially significant, is open to divergent interpretations depending on assumptions about market signaling, investor psychology, and corporate strategy.

Table 2 presents chatbot responses from six leading LLM providers across three prediction horizons. Disagreement emerges immediately: although four models interpret the news as `GOOD` for the next-day impact—citing strategic repositioning, improved liquidity, or

[5]The link to the news is here.

enhanced financial flexibility, two models (OpenAI and Google) classify the same headline as BAD, pointing to concerns about revenue loss or divestment.

As the forecast horizon extends, two types of divergence become apparent. First, cross-provider dispersion persists and sometimes intensifies. At the one-month horizon, OpenAI and Google revise their stance to GOOD, emphasizing potential long-term efficiencies and improvements, while Mistral reverses its initially positive classification to BAD, citing doubts about long-term partnership stability and growth prospects. Meanwhile, the remaining three models adopt a more agnostic stance, shifting to UNKNOWN and highlighting execution risks and expansion challenges.

Second, we observe temporal variation within individual models. For instance, Meta initially assigns a GOOD label with high sentiment for the next day, flips to BAD the following week, and then shifts to UNKNOWN by the next month. This evolution reflects an internal reassessment as the model weighs near-term optimism against growing strategic ambiguity. Similar temporal reversals are evident in other models, including Mistral and OpenAI.

Overall, this example demonstrates the dual diagnostic power of our tabular prompt: it enables both cross-model comparison and intra-model temporal tracking. Disagreements across models likely arise from differences in how they weigh divestiture signals, market expectations, and strategic framing. Temporal shifts within models, by contrast, reveal their dynamic interpretive capacity—offering researchers a way to treat LLMs not just as static classifiers, but as agents capable of time-sensitive belief updating.

3 Data and Main Variables

This section describes the data sources, sample construction, and main variables we compile for the analysis. We bring together standard financial datasets with a large collection of news headlines and chatbot-generated sentiment measures to capture how LLMs interpret market-relevant information. Our setup enables us to quantify both average sentiment and the extent of disagreement across providers and horizons, forming the basis for the empirical tests that follow.

3.1 Data and Sample Selection

We follow Lopez-Lira and Tang (2023) to compile a dataset of news headlines for common stocks listed on the NYSE, NYSE MKT (formerly AMEX), and NASDAQ. We obtain daily and monthly stock data from the Center for Research in Security Prices (CRSP). Quarterly and annual financial statement data come from the COMPUSTAT database. Analyst forecast data are obtained from the Institutional Brokers' Estimate System (I/B/E/S). The intraday millisecond trade and quote data come from the NYSE Trade and Quote (TAQ) database. For comparison, we also incorporate news sentiment measures from RavenPack.

The sample period spans January 2023 to December 2024, during which generative AI tools became widely adopted. The final sample includes 133,867 headlines covering 3,895 unique stocks, with an average of 2,447 stocks per month. Approximately 80% (106,751) of the headlines are classified as overnight news released either before 9:00 a.m. or after 4:00 p.m. on a trading day, while the remaining 20% are classified as intraday news.

3.2 LLM Disagreement

We begin by converting the chatbot responses into numerical values and construct three LLM-based news sentiment measures: (i) *News Score*: equals 1 if $Impact = \texttt{BAD}$, 2 if $Impact = \texttt{UNKNOWN}$, and 3 if $Impact = \texttt{GOOD}$; (ii) *Sentiment Rank*: equals 1 if $Sentiment\ Score \leq$

20, 2 if $20 < Sentiment\ Score \leq 40$, 3 if $40 < Sentiment\ Score \leq 60$, 4 if $60 < Sentiment\ Score \leq 80$, and 5 if $Sentiment\ Score > 80$; and (iii) *Sentiment Score*: the raw sentiment score output directly from the chatbot, ranging from 0 to 100.

Next, for each news at each prediction horizon (next-day, next-week, and next-month), we compute the average *News Score*, *Sentiment Rank*, and *Sentiment Score* across six LLM providers, denoted as $AvgNews$, $AvgRank$, and $AvgSent$, respectively. We also calculate the cross-provider dispersions, defined as the standard deviations of *News Score*, *Sentiment Rank*, and *Sentiment Score* across providers, and denote them as $StdNews$, $StdRank$, and $StdSent$.

Finally, we calculate cross-horizon dispersions. For each news-provider pair, we compute the standard deviations of *News Score*, *Sentiment Rank*, and *Sentiment Score* across three prediction horizons. We then average these standard deviations across providers for each news to obtain the cross-horizon dispersions, denoted as $StdNewsHor$, $StdRankHor$, and $StdSentHor$. Internet Appendix Table A.1 provides a detailed definition for each variable.

Table 3, Panels A to C, report summary statistics for the LLM-based news sentiment measures and their cross-provider dispersions for next-day, next-week, and next-month predictions, respectively. Several findings are worth noting. First, the average sentiment measures ($AvgNews$, $AvgRank$, and $AvgSent$) are highest for next-day predictions, followed by next-month and then next-week predictions. This pattern suggests greater optimism in the immediate term and increased uncertainty in the intermediate horizon.

Second, we find substantial cross-provider dispersion across all sentiment measures and prediction horizons, with the highest dispersion occurring for the next-week predictions. For instance, the average $StdSent$ is 8.392 for next-day, 14.561 for next-week, and 8.254 for next-month predictions, corresponding to 13%, 31%, and 15% of the sample average, respectively. Turning to the *News Score*, which takes only three possible values—`BAD`, `UNKNOWN`, and `GOOD`, the average $StdNews$ is 0.297 for next-day, 0.705 for next-week, and 0.338 for next-month predictions. These figures represent 12%, 39%, and 15% of the

respective sample means. Moreover, $StdNews$ at the 75th percentile is 0.548, 0.983, and 0.516 for the next-day, next-week, and next-month predictions, respectively.[6]

Panel D further reports significant cross-horizon dispersion across all sentiment measures. The average $StdNewsHor$, $StdRankHor$, and $StdSentHor$ are 0.628, 0.642, and 14.845, respectively. Since all LLMs receive the same news headlines and prompts, these results suggest that LLMs not only differ in their interpretation of public signals but also adjust their sentiment assessments based on the investment horizon. This further motivates our analysis of how LLM disagreement affects asset prices, price informativeness, and trading activity.

4 What Drives News Sentiment Dispersion?

A central question in our study is why large language models diverge in their interpretation of the same news. Disagreement could reflect informational complexity, variation in model training and alignment, or systematic differences in reasoning styles. In this section, we explore these possibilities by relating disagreement to observable news and firm characteristics and by analyzing the causal explanations that models provide alongside their predictions.

4.1 News and Firm Characteristics

We begin by examining the determinants of model disagreement. Although state-of-the-art LLMs have demonstrated strong performance in processing firm-specific news (e.g., Lopez-Lira and Tang, 2023; Chen et al., 2024), the sources of disagreement across models remain underexplored. On the one hand, disagreement may increase with informational complexity, as more complex news are subject to a wider range of interpretations. For

[6]To put this in perspective, in a simplified scenario with three LLMs assigning sentiment scores of 50, 50, and 75 (or 50, 50, and 65), the corresponding $StdSent$ is 14.434 (8.660). If three LLMs classify the news as `BAD`, `UNKNOWN`, and `GOOD` (or `UNKNOWN`, `UNKNOWN`, and `GOOD`), the corresponding $StdNews$ is 1.000 (0.577).

example, Loughran and Mcdonald (2014) find that low 10-K readability is associated with higher post-filing return volatility, potentially due to persistent uncertainty.

However, standard readability measures used in the corporate disclosure literature, such as file size, word count, the Fog index, and the frequency of complexity words (e.g., Li, 2008; Loughran and Mcdonald, 2014; Loughran and McDonald, 2023), are designed for human readers. In contrast, LLMs, which are trained on vast textual datasets and characterized by extensive parameterization, may be better equipped to process complex information and capture the true sentiment, thereby reducing disagreement.

To proceed, we estimate the following panel regression at the daily news level:

$$Disp_{i,n,t} = \alpha + \beta_1 News_{i,n,t} + \beta_2 C_{i,t-1} + \varepsilon_{i,n,t}, \tag{1}$$

where $Disp_{i,n,t}$ refers to a list of LLM-based news sentiment dispersion measures for news n related to stock i on day t, proxied by $StdRank$ for next-day, next-week, and next-month predictions across providers, and by $StdRankHor$ for cross-horizon dispersion.[7] $News_{i,n,t}$ refers to a list of news characteristics, including *Complexity*, *AvgRank*, and *Overnight*. Vector C stacks all other stock-level control variables, namely, the *Log(Market Cap)*, *Book-to-Market*, *Profitability*, *Investment*, *1M Return*, *Momentum*, *NumAna*, and *AnaDisp*. Internet Appendix Table A.1 provides a detailed definition for each variable. We also include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels to account for the time-series and cross-sectional correlations in model disagreement.

We report the results in Table 4. Models 1–3 focus on cross-provider dispersion for next-day predictions. First, in contrast to human readers, high news complexity (*Complexity*)—a composite measure based on word count, the Fog index, and the percentage of complex words—is associated with lower cross-provider dispersion. This suggests that LLMs may overcome barriers that typically hinder human information processing and may even benefit from longer, more sophisticated text. Second, we find greater disagreement among LLMs

[7] For brevity, we focus on news sentiment as proxied by *Sentiment Rank*. Unreported results are robust to alternative measures, including *News Score* and *Sentiment Score*.

when processing more negative news (i.e., low *AvgRank*) and overnight news. These findings remain robust after controlling for additional firm characteristics, including analyst forecast dispersion.

Models 4–6 and 7–9 examine cross-provider dispersion for next-week and next-month predictions, respectively. At the weekly horizon, model disagreement continues to decline with news complexity, whereas at the monthly horizon, it is positively associated with news complexity. In contrast to the next-day results, cross-provider dispersion increases with more positive news at both the weekly and monthly horizons. As shown in Models 10–12, cross-horizon dispersion decreases with news complexity but is higher for news released overnight. Finally, cross-provider and cross-horizon dispersions are not systematically related to specific firm characteristics.

Overall, we find that LLMs disagree less on complex news when making short-term predictions (next-day and next-week) but disagree more on complex news when making longer-term predictions (next-month). These distinct features of LLM information processing may have important implications for financial markets, especially as investors increasingly rely on generative AI tools to interpret financial information (e.g., Blankespoor et al., 2024; Chang et al., 2025).

4.2 Causal Reasoning Analysis

In addition to the news and firm characteristics examined in the previous tests, we leverage the explanations accompanying each LLM's sentiment assessment to analyze the causal reasoning that providers articulate when justifying their classifications. This analysis sheds light on whether model disagreement reflects systematic differences in how models interpret news or whether it is better understood as random variation. Specifically, we ask whether differences in impact labels (good, bad, or unknown) are associated with differences in stated reasoning and whether providers exhibit stable and distinct reasoning profiles.

We proceed in four steps. First, we illustrate qualitative differences across providers by comparing selected cause–effect chains for the AMC headline discussed earlier. Second, we

quantify the similarity of extracted causes and effects across providers for the same headline and horizon, contrasting cases with identical versus differing impact labels. Third, we analyze provider-level patterns in cause and effect categories and their mapping to impact labels. Finally, we compare providers in terms of overall reasoning styles and stability across horizons.

In each year, we select 1,000 news headlines with the highest cross-provider dispersion, as measured by $StdNews$. For these headlines, we use the model *Explanation* together with the associated *News Score* introduced earlier and, following Bybee (2025), apply a prompt that elicits a concise cause–effect chain in structured form, enabling automated parsing. The prompt used is as follows:

> Identify the causal statement explaining a [Impact]: [Explanation].
>
> Output ONLY a JSON object with this exact format: {"cause":" ≤ 5 words>","intermediate_cause":" ≤5 words or ">","effect":" ≤5 words>"}
>
> Rules:
>
> - cause: the initial trigger or reason
> - intermediate_cause: optional middle step (use empty string if none)
> - effect: the final outcome or result
> - Keep each field to 5 words or less
> - Use clear, concise language

Model explanations generated in response to this prompt were processed to extract directed acyclic graphs (DAGs) that capture the stated cause and effect. Table 5 illustrates the extracted cause–effect data for the AMC headline (*"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*) across providers and horizons. For each provider, the table reports the assigned impact label (`GOOD`, `BAD`, or `UNKNOWN`) along with the corresponding cause and effect. Within the same horizon, providers may agree on the impact label but diverge in their reasoning. For instance, among those assigning `GOOD`

for the next day, Anthropic, Azure OpenAI, and Meta identify the cause as a licensing or strategic shift, while Mistral emphasizes the stake sale. Even when labels align, providers may differ in their identified effects: Anthropic, Azure OpenAI, and Meta specify a positive market reaction, while Mistral points to positive short-term financials.

Across horizons, both impact labels and cause–effect chains vary for the same provider, reflecting adjustments in interpretation over time. For example, Anthropic, Azure OpenAI, and Meta assign `BAD` for the next week, citing financial concern and analyst skepticism as potential causes, whereas Mistral assigns `UNKNOWN`, attributing it to the licensing shift.

We next assess whether model disagreement is systematically associated with differences in causal reasoning. For each headline–horizon pair, we form all provider pairs ($6 \times 5 = 30$) and compute cosine similarities for causes and effects separately using sentence embeddings. Provider pairs are then grouped by whether they assign the same or different impact labels.

Figure 1 shows the resulting distributions. The average similarity in causes (effects) is 0.391 (0.370) when providers agree on the impact label and 0.295 (0.250) when they disagree, with the differences significant at the 1% level. These findings suggest that model disagreement in impact labeling stems from variation in the underlying reasoning processes of LLM providers rather than from random noise.

We then study how providers map specific causes (in cause–effect chains) to impact classifications. The previously extracted causes are grouped into six primary topic categories: *corporate governance*, *cost/liquidity*, *demand/market*, *macroeconomic*, *regulation/policy*, and *technological*. These categories are formed through an automated keyword-based classification system that maps extracted cause text to predefined taxonomic groups, where each category contains a curated list of relevant keywords. For example, *corporate governance* includes terms such as *management*, *leadership*, *board*, *executive*, *strategy*, *decision*, and *corporate*. The classification algorithm assigns causes to the category with the highest keyword match count, representing an unsupervised rule-based approach that relies on predefined heuristics rather than training data, ensuring consistent categorization across all provider responses while capturing the main drivers of market events.

Figure 2 reports, for each provider and cause category, the distribution of impact labels conditional on that cause, pooled across prediction horizons:

$$\Pr(\text{Impact} \in \{\texttt{GOOD}, \texttt{BAD}, \texttt{UNKNOWN}\} \mid \text{Cause} = c,\ \text{Provider} = p), \tag{2}$$

The figure shows clear cross-provider differences within the same cause category, with red denoting GOOD impact, green denoting BAD, and grey denoting UNKNOWN. For example, subfigure (a) shows that in the case of *corporate governance*, Anthropic and Google more frequently assign the UNKNOWN label, while Azure OpenAI and OpenAI more often assign GOOD. Subfigure (f) shows that when the cause is *technological*, Google and Mistral are more likely to assign GOOD, while Anthropic tends to assign BAD. Similar variation arises in other categories, suggesting that cross-provider disagreement can be partly explained by differences in how cause types are mapped into impact labels.

A parallel exercise examines how effects (in cause–effect chains) map to impact classifications across topic categories. We apply the same keyword-based classification procedure to classify effects into six primary categories: *competitive position*, *innovation/capacity*, *market/valuation*, *profitability/cost*, *revenue/growth*, and *risk/volatility*. We then repeat the analysis in Equation (2), replacing *Cause* with *Effect*.

Figure 3 reports, for each provider and effect topic, the conditional distribution of impact labels. Once again, clear cross-provider differences emerge. For example, subfigure (a) shows that when effects relate to *competitive position*, Google and OpenAI more frequently classify the impact as GOOD, Anthropic and Mistral lean toward BAD, and Meta and Azure OpenAI tend to assign UNKNOWN.

Notably, the topic categories themselves do not inherently indicate positive or negative news and should be sentiment-neutral. Thus, systematic leanings toward certain labels may reflect the differing priors of individual LLMs, leading to divergent sentiment assessments even when they extract similar news topics as causes and effects.

Finally, we compare providers in terms of overall reasoning styles and stability across horizons. Figure 4 presents three complementary panels. Subfigure (a) shows a principal

component analysis (PCA) projection of providers in the reasoning space, where proximity indicates greater similarity in extracted cause–effect chains. Here, Google, Meta, and Mistral cluster closely, Azure OpenAI and OpenAI form a separate group, and Anthropic stands apart.

Figure 4, subfigure (b) reports cause stability, plotting the average correlation and divergence of causes across horizons. Correlation is calculated as the Pearson correlation coefficient between cause category distributions across adjacent time horizons (Pearson, 1895), while divergence is measured as the Jensen-Shannon divergence between these distributions, capturing the degree of distributional shift (Lin, 1991). Most providers display high cause correlation, but Anthropic shows greater divergence, while Azure OpenAI and Google exhibit nearly identical causes across horizons.

Figure 4, subfigure (c) presents the corresponding stability metrics for effects. Effect stability varies more widely: Mistral and Google display relatively high effect correlation and low divergence, whereas Azure OpenAI, OpenAI, and Anthropic exhibit lower effect correlation and higher divergence, indicating greater shifts in their consequence framing across horizons.

Taken together, these exercises suggest that provider disagreement is not merely noise. Rather, disagreement among LLM providers is systematically linked to differences in causal reasoning, specifically in (i) the causes and effects extracted from news, (ii) the mapping of specific cause and effect categories to impacts, and (iii) the clustering and cross-horizon stability of providers' reasoning styles.

5 Asset Pricing Implications

Having established the extent and drivers of disagreement across LLM providers, we now turn to the asset pricing implications. If model disagreement reflects heterogeneity in how investors interpret news, it may affect both the speed and direction with which information is incorporated into prices. In this section, we examine whether LLM-based sentiment measures and their dispersions predict stock returns, interact with firm and news characteristics, and influence market reactions around earnings announcements.

5.1 Stock Return Predictability

In this subsection, we first examine whether LLMs can predict stock returns across different horizons. Specifically, we estimate the following daily firm-level panel regression:

$$R_{i,t+1} = \alpha + \beta_1 NewsSent_{i,t} + \varepsilon_{i,t+1}, \tag{3}$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, and $NewsSent_{i,t}$ refers to a list of LLM-based average news sentiment measures, including $AvgNews$, $AvgRank$, and $AvgSent$. For overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, we enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). For intraday news released between 9 a.m. and 4 p.m. on day t, we enter the position at the close of day t and exit at the close of the next trading day $t+1$ (close-to-close return). If a firm has multiple overnight or intraday news on a given day, we compute the average sentiment for each category separately at the firm-day level.

In addition to the next-day return ($R_{i,t+1}$), we also consider cumulative returns over the periods $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), corresponding to the weekly and monthly prediction horizons. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. We also include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels.

Table 6 presents the results, with Panels A to C corresponding to next-day, next-week, and next-month predictions, respectively. As shown in Panel A, all three LLM-based sentiment measures significantly predict next-day returns (Models 1–3), while the return predictability does not persist beyond the one-day horizon (Models 4–9). A one-standard-deviation increase in $AvgNews$ ($AvgRank$, $AvgSent$) is associated with a 0.13%, 0.13%, and 0.14% increase in next-day returns, respectively.[8] These findings confirm and extend prior work by demonstrating return predictability across a broader set of LLMs, beyond ChatGPT and LLaMA (e.g., Lopez-Lira and Tang, 2023; Chen et al., 2024).[9]

In Panel B, consistent with the weekly prediction horizon, we find significant return predictability both on the next day (Models 1–3) and over the subsequent week excluding the first day (Models 4–6), with the effect being more pronounced for the next day. In Panel C, despite the longer monthly horizon, return predictability remains concentrated on the next day (Models 1–3).

Collectively, these findings suggest that the information embedded in news text is not immediately and fully incorporated into market prices, potentially due to limits to arbitrage and limited investor attention (e.g., Miller, 1977; Sims, 2003; Peng, 2005; Kozak et al., 2018). This delayed adjustment creates an opportunity to predict short-term stock returns using publicly available news. However, a substantial portion (if not all) of the sentiment information appears to be incorporated into stock prices on the first trading day, regardless of the forecast horizon. This pattern aligns with the notion that stock prices are forward-looking and that stock market is relatively efficient.

Next, we examine how model disagreement affects future stock returns. Miller (1977) argues that, in the presence of heterogeneous beliefs about fundamental values and impediments to arbitrage, such as short-sale constraints, stock prices tend to reflect the views of more optimistic investors. Consistent with this view, Diether et al. (2002) find that stocks

[8]The impact of stock return is computed as $0.213 \times 0.594 = 0.13\%$, where 0.213 is the regression coefficient of $AvgNews$ in Model 1, and 0.594 is the standard deviation of $AvgNews$ (Table 3, Panel A).

[9]Unreported results show similar next-day return predictability across each of the six LLM providers individually.

with greater dispersion in analyst earnings forecasts earn lower future returns. Motivated by this literature, we estimate the following daily firm-level panel regression:

$$R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1}, \quad (4)$$

where $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $AvgRank_{i,t}$ is the average of sentiment ranks across providers. Since our earlier findings show that return predictability is strongest in the immediate term, we focus on sentiment ranks derived from next-day predictions. To compare LLM-based measures with existing sentiment analysis, we also include $CSS_{i,t}$, the composite sentiment score from Ravenpack. In addition to the next-day return ($R_{i,t+1}$), we also consider cumulative returns over the periods $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), defined as in Equation (3). Finally, we include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels.

Table 7 reports the results. As expected, the average sentiment rank predicts returns only for the next day. Specifically, a one-standard-deviation increase in $AvgRank$ is associated with a 0.09% increase in next-day returns (Model 4). In the longer term, as shown in Models 9–12, we also observe evidence of return reversals over the one-month horizon.

Notably, cross-provider and cross-horizon dispersions are not significantly related to next-day returns, but are negatively associated with returns over the remainder of the week, consistent with Miller (1977) (Models 4–8). A one-standard-deviation increase in $StdRank$ and $StdRankHor$ is associated with a 0.12% and 0.14% decline in next-week returns, excluding the first day (Model 8). The negative relation between cross-provider dispersion and stock returns persists over the monthly horizon: a one-standard-deviation increase in $StdRank$ is associated with a 0.33% decline in next-month returns, excluding the first day (Model 12).

We conduct three sets of robustness checks. First, our findings remain robust in the subsample of overnight news (Internet Appendix Table A.2, Panel A). Second, since all

LLMs in our sample are trained on data prior to November 2023, we focus on subperiods of 2024 as an out-of-sample test to mitigate potential look-ahead bias. Because our analysis does not focus on the return predictability of LLMs, look-ahead bias is unlikely to be a significant concern in this context. On the contrary, if LLMs incorporated forward-looking information, they would likely converge more in their responses, thereby weakening our findings. Nevertheless, our main findings remain unchanged (Internet Appendix Table A.2, Panel B). Third, we examine alternative LLM-based sentiment measures constructed from *News Score* and *Sentiment Score* (Internet Appendix Table A.3). Our main results continue to hold, although with weaker statistical significance. Overall, LLM-based model disagreement provides a novel proxy for heterogeneous beliefs in the market and has meaningful asset pricing implications.

5.2 ChatGPT Outages

To further strengthen causal inference, we exploit ChatGPT outages as exogenous shocks to cross-provider dispersion. Given ChatGPT's dominant market share among AI tools, unexpected outages are likely to attenuate the return predictability of cross-provider dispersion: when ChatGPT is unavailable, investors cannot rely on it, and the dispersion we compute from all six LLM providers becomes a noisier proxy for the true dispersion in market beliefs.[10]

We identify ChatGPT outage events using official reports from OpenAI, which provide the date, start and end times, outage type, and severity.[11] We require outages to coincide

[10] Because cross-horizon dispersion is first computed within each provider across prediction horizons and then averaged across providers, it is less likely to be affected by ChatGPT outages. We therefore focus on cross-provider dispersion in this analysis.

[11] See, https://status.openai.com/history.

with the timing of news releases, mapping overnight news to overnight outages and intraday news to intraday outages. We then estimate the following daily firm-level panel regression:

$$R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRank_{i,t} \times Outage_{i,t} + \beta_3 StdRankHor_{i,t} + \beta_4 AvgRank_{i,t} + \beta_5 CSS_{i,t} + \beta_6 Outage_{i,t} + \varepsilon_{i,t+1}, \quad (5)$$

where $Outage_{i,t}$ is a dummy variable that equals 1 if a ChatGPT outage occurs and 0 otherwise. All other variables are defined as in Equation (4). In addition to the next-day return ($R_{i,t+1}$), we also consider cumulative returns over the periods $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), defined as in Equation (3). Finally, we include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels.

Table 8 presents the results. First, although cross-provider dispersion is not unconditionally related to next-day returns (Table 7, Model 1), it is negatively associated with next-day returns in the absence of ChatGPT outages, as indicated by the negative β_1 coefficient (Model 1). This effect is offset during outage periods, as reflected by the positive β_2 coefficient (Model 1).

Second, while the negative relation between cross-provider dispersion and stock returns persists over weekly and monthly horizons (Models 4–9), ChatGPT outages attenuate return predictability over short horizons within a week, as indicated by the positive β_2 coefficient (Models 1–6). This pattern is consistent with the temporary nature of ChatGPT outages. As shown in Model 6, in the absence of outages, a one-standard-deviation increase in $StdRank$ is associated with a 0.16% decline in next-week returns, excluding the first day. During ChatGPT outage periods, this return predictability is reduced by 67%. This substantial attenuation indicates that cross-provider dispersion is significantly affected when ChatGPT is unavailable. Overall, these findings further strengthen the causal link between LLM-based model disagreement and future stock returns.

5.3 Heterogeneity Analysis

In this subsection, we examine whether the previously documented relation between model disagreement and future stock returns varies with news and firm characteristics. According to Miller (1977), stocks with greater dispersion of opinions are more likely to be overpriced, resulting in lower future returns. Consequently, this effect should be stronger when overall news sentiment is more positive, reflecting the presence of more optimistic investors in the market. Moreover, the impact of model disagreement on future returns is expected to be more pronounced for firms with more opaque information environments, where heightened information frictions make it more difficult to resolve uncertainty. Similarly, the effect may intensify when information processing costs are higher—for instance, for firms with greater operating uncertainty or in the presence of more complex news.

We test these hypotheses using the following daily firm-level panel regression:

$$\begin{aligned} R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 StdRank_{i,t} \times Char_{i,t} \\ + \beta_4 StdRankHor_{i,t} \times Char_{i,t} + \beta_5 Char_{i,t} + \varepsilon_{i,t+1}, \end{aligned} \tag{6}$$

where $Char_{i,t}$ denotes a set of moderating variables, including proxies for (i) overall news sentiment: *AvgRank*, the average of sentiment ranks across providers; (ii) the quality of the information environment: *Log(Market Cap)*, the logarithm of the market capitalization; *NumAna*, the number of analysts covering the firm; and *AnaDisp*, analyst forecast dispersion; (iii) operating uncertainty: *ROAVOL*, earnings volatility, and *IVOL*, idiosyncratic return volatility; and (iv) news complexity: *Complexity*, a composite measure based on word count, the Fog index, and the percentage of complex words. For brevity, sentiment ranks are based on next-day predictions. All other variables are defined as in Equation (4). In addition to the next-day return ($R_{i,t+1}$), we also consider cumulative returns over the periods $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), defined as in Equation (3). Finally, we include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels.

Table 9 presents the results, with Panel A focusing on the next-day return ($R_{i,t+1}$), and Panels B and C presenting cumulative returns over the periods $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. As shown in Panel A, while cross-provider dispersion $StdRank$ is not unconditionally associated with next-day returns (Table 7, Model 3), it predicts lower next-day returns, particularly for firms with more positive news, smaller market capitalizations, and lower analyst coverage (Models 1–3).

In Panels B and C, cross-provider dispersion $StdRank$ exhibits a persistent effect on returns over weekly and monthly horizons, particularly for smaller firms, firms with high analyst forecast dispersion, and those with high idiosyncratic return volatility (Models 2, 4, and 6). Furthermore, cross-horizon dispersion $StdRankHor$ is associated with lower future returns for firms with high earnings volatility and idiosyncratic return volatility (Models 5–6). Finally, model disagreement shows no effect on future returns through news complexity across any horizon (Panels A–C, Model 7).

Overall, disagreement among LLMs contributes to heterogeneous beliefs in the stock market and has important implications for asset pricing. Consistent with Miller (1977), when differences of opinion are combined with short-sale constraints, equity prices tend to reflect the views of more optimistic investors, resulting in lower future returns. Our findings are more pronounced for firms with more positive news sentiment, more opaque information environments, and greater operating uncertainty. While generative AI tools substantially reduce information processing costs for investors, a potential unintended consequence is increased systematic belief dispersion among market participants, leading to possible price distortions.

5.4 Model Disagreement and Earnings Announcement Returns

While corporate earnings announcements are important value-relevant information events, an extensive literature has documented post-earnings-announcement drift, suggesting that investors underreact to earnings news due to limited attention (e.g., Bernard and Thomas, 1989; Hirshleifer and Teoh, 2003; DellaVigna and Pollet, 2009; Hirshleifer et al., 2009; Ben-

Rephael et al., 2017). In this subsection, we investigate whether model disagreement on the announcement day affects the speed at which earnings information is incorporated into stock prices. To proceed, we estimate the following quarterly firm-level panel regression:

$$\begin{aligned} R_{i,t} = \alpha &+ \beta_1 SUE_{i,t} + \beta_2 SUE_{i,t} \times StdRank_{i,t} + \beta_3 SUE_{i,t} \times StdRankHor_{i,t} \\ &+ \beta_4 SUE_{i,t} \times AnaDisp_{i,t-1} + \beta_5 StdRank_{i,t} + \beta_6 StdRankHor_{i,t} \\ &+ \beta_7 AnaDisp_{i,t-1} + \beta_8 C_{i,t-1} + \varepsilon_{i,t}, \end{aligned} \tag{7}$$

where $R_{i,t}$ is stock i's return on earnings announcement day t (close-to-close return). $SUE_{i,t}$ is the standardized unexpected earnings (SUE), defined as the difference between the actual earnings for the quarter and the average of the most recent analyst forecasts, divided by the standard deviation of those forecasts. $AnaDisp_{i,t-1}$ is the standard deviation of analyst forecasts divided by the absolute value of the mean forecast. Vector C stacks all other stock-level control variables, namely, the *AvgRank*, *Log(Market Cap)*, *Book-to-Market*, *Profitability*, *Investment*, and *Momentum*. All other variables are defined as in Equations (1) and (4). We include all news released between the close of day $t-1$ and the close of day t, with sentiment ranks based on next-day predictions. While β_1 coefficient captures the unconditional earnings response, the variables of interest are the β_2 and β_3 coefficients, which reflect the incremental impact of model disagreement.

In addition to the next-day return ($R_{i,t+1}$), we also consider cumulative returns over $t+1$ to $t+30$ ($R_{i,t+1:t+30}$) and $t+1$ to $t+40$ ($R_{i,t+1:t+40}$), respectively. In these cases, we enter the position at the market close of day t and exit at the close of trading day $t+30$ or $t+40$, respectively. Finally, we include quarter and day-of-the-week fixed effects and cluster standard errors at both the firm and calendar day levels, following Ben-Rephael et al. (2017).

Table 10 reports the results. Given that limited attention and model disagreement are likely to have a more pronounced impact on smaller firms, we analyze two subsamples. Panel A focuses on small firms with market capitalizations below the NYSE median breakpoint, while Panel B reports similar statistics for large firms above this threshold. Several findings

are worth noting. First, the post-earnings-announcement drift persists in our sample period across both subsamples. While stock prices respond to the earnings surprise on the announcement day (Model 1), there is a delayed reaction extending up to 40 days (Models 6 and 11), as evidenced by the significantly positive β_1 coefficient.

Second, for small firms, as shown in Panel A, the β_2 and β_3 coefficients are statistically insignificant on the announcement day (Models 2–5), indicating that neither cross-provider dispersion $StdRank$ nor cross-horizon dispersion $StdRankHor$ affects the immediate price reaction on the announcement day. However, the β_3 coefficient is significantly positive across all specifications during the post-announcement window, while the baseline post-earnings-announcement drift (captured by the β_1 coefficient) becomes insignificant or even negative (Models 8–10 and 13–15). Our results are robust to controlling for alternative proxies of divergent opinions, such as analyst dispersion (captured by the β_4 coefficient). These findings suggest that the well-documented post-earnings-announcement drift may be partially driven by heterogeneous beliefs among investors with varying investment horizons.

Turning to large firms in Panel B, although they typically receive greater investor attention and exhibit higher transparency, both cross-provider dispersion $StdRank$ and cross-horizon dispersion $StdRankHor$ appear to delay the incorporation of information on the announcement day, as indicated by the negative β_2 and β_3 coefficients (Models 4–5). Furthermore, model disagreement does not exhibit any incremental impact on the post-earnings-announcement drift for large firms (Models 7–10 and 12–15).

Overall, LLM-based horizon disagreement in news headlines amplifies price underreaction to earnings news, contributing to the post-earnings-announcement drift among small firms. In contrast, for large firms, both cross-provider and cross-horizon dispersion appear to delay the immediate price reaction on the announcement day. These findings suggest that model disagreement may hinder the incorporation of information following earnings announcements and reduce price informativeness.

6 Model Disagreement and Trading Volume

An extensive literature examines the effect of differences of opinion on trading volume (e.g., Harris and Raviv, 1993; Kandel and Pearson, 1995; Bamber et al., 1997; Banerjee and Kremer, 2010). While our earlier results show that LLM-based model disagreement is associated with future stock returns, in this section we investigate its impact on trading volume. Specifically, we estimate the following daily firm-level panel regression:

$$Vol_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1}, \quad (8)$$

where $Vol_{i,t+1}$ is the logarithm of the trading volume for stock i on day $t+1$. All other variables are defined as in Equation (4). We restrict the sample to overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. In addition to the next-day trading volume ($Vol_{i,t+1}$), we also consider cumulative trading volume over the periods $t+2$ to $t+5$ ($Vol_{i,t+2:t+5}$, in logarithms) and $t+2$ to $t+20$ ($Vol_{i,t+2:t+20}$, in logarithms). Finally, we include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels.

Table 11, Panel A reports the results. As expected, cross-provider and cross-horizon dispersions are significantly related to future trading volume. Specifically, a one-standard-deviation increase in $StdRank$ and $StdRankHor$ is associated with a 2.24% and 2.04% increase in next-day trading volume, respectively (Model 4, scaled by the sample standard deviation of the corresponding trading volume measure).[12] The economic magnitude diminishes over longer horizons but remains statistically significant: a one-standard-deviation increase in $StdRank$ and $StdRankHor$ is associated with a 0.95% and 0.47% increase in next-week trading volume, excluding the first day (Model 8). Moreover, the effect of

[12] The effects of $StdRank$ and $StdRankHor$ on trading volume are computed as $0.123 \times 0.367/2.018 = 2.24\%$ and $0.246 \times 0.167/2.018 = 2.04\%$, where 0.123 and 0.246 are the regression coefficients of $StdRank$ and $StdRankHor$ in Model 4, 0.367 is the standard deviation of $StdRank$ (Table 3, Panel A), 0.167 is the standard deviation of $StdRankHor$ (Table 3, Panel D), and 2.018 is the standard deviation of Vol_{t+1} (Table 3, Panel E).

cross-provider dispersion $StdRank$ persists over the monthly horizon, with a one-standard-deviation increase in $StdRank$ associated with a 0.72% increase in next-month trading volume, excluding the first day (Model 12).

While controlling for firm fixed effects captures within-firm variation in trading volume over time, we also explicitly examine abnormal trading volume. For stock i on day t, the abnormal trading volume ($AbVol_{i,t}$) is defined as the difference between log volume on day t and the firm's average log volume from $t-140$ to $t-20$ trading days, following Cookson et al. (2024). We then compute the daily average of abnormal trading volume over weekly and monthly horizons, excluding the first day.

Table 11, Panel B reports the results. Our findings remain consistent but exhibit larger economic magnitudes. For instance, a one-standard-deviation increase in $StdRank$ and $StdRankHor$ is associated with a 4.64% and 7.04% increase in next-day abnormal trading volume, respectively (Model 4, scaled by the sample standard deviation of the corresponding trading volume measure). Consistent with earlier results, the effect of $StdRank$ ($StdRankHor$) persists over the monthly (weekly) horizon. As shown in Model 12 (Model 8), a one-standard-deviation increase in $StdRank$ ($StdRankHor$) is associated with a 1.98% (3.30%) increase in next-month (next-week) abnormal trading volume, excluding the first day.

Our findings remain valid across two additional robustness checks: (i) the out-of-sample period in 2024 (Internet Appendix Table A.4), and (ii) alternative LLM-based sentiment measures based on next-week and next-month predictions (Internet Appendix Table A.5). Overall, LLM-based model disagreement is associated with higher trading volume, and the effect persists for up to one month, suggesting that the use of generative AI in financial information processing may further amplify trading activity in financial markets.

While generative AI is increasingly adopted by all types of investors, we next focus on the trading behavior of retail investors. On the one hand, less sophisticated retail investors may benefit disproportionately from low-cost, easily accessible generative AI tools for financial information processing. On the other hand, more sophisticated institutional

investors are likely better positioned to consolidate diverse resources and leverage their existing advantages with the aid of generative AI. Thus, whether the effect is stronger for retail trading remains an empirical question.

To investigate this, we follow the algorithm proposed by Barber et al. (2024) to identify retail trades from TAQ data. This approach exploits two key institutional features of retail trading. First, most marketable equity orders initiated by retail investors take place off-exchange and are either filled from a broker's inventory or routed to wholesalers (Battalio et al., 2016). Accordingly, we restrict our analysis to off-exchange trades, which are designated with the exchange code "D" in TAQ. Second, retail traders typically receive a small price improvement, i.e., a small fraction of a cent, relative to the national best bid or offer (NBBO) (Boehmer et al., 2021). Building on this, Barber et al. (2024) modify the Boehmer et al. (2021) algorithm by signing trades using quoted spread midpoints. Specifically, for subpenny trades (i.e., trades not executed at a round penny), we classify a trade as a retail buy (sell) transaction if the execution price is greater (less) than the quote midpoint, while trades executed between 40% and 60% of the NBBO remain unsigned.

We then estimate the following daily firm-level panel regression:

$$BuyVol_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1}, \quad (9)$$

where $BuyVol_{i,t+1}$ is the logarithm of the retail buy volume for stock i on day $t+1$. All other variables are defined as in Equation (4). We restrict the sample to overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. In addition to retail buy volume ($BuyVol_{i,t+1}$), we also examine retail sell volume ($SellVol_{i,t+1}$, in logarithms) and retail order imbalance ($OIBS_{i,t+1}$). Finally, we include firm and calendar day fixed effects and cluster standard errors at both the firm and calendar day levels.

As a robustness check, we replace retail buy volume with abnormal retail buy volume. For stock i on day t, the abnormal retail buy volume ($AbBuyVol_{i,t}$) is defined as the difference between log retail buy volume on day t and the firm's average log retail buy

volume from $t-140$ to $t-20$ trading days. Similarly, retail sell volume is replaced with abnormal retail sell volume ($AbSellVol$), and retail order imbalance is replaced with abnormal retail order imbalance ($AbOIBS$), both defined relative to their respective firm-level averages over the $t-140$ to $t-20$ trading-day window.

Table 12 reports the results, with Panel A for retail trading and Panel B for abnormal retail trading. Both cross-provider and cross-horizon dispersions are significantly related to retail buy and sell volume on the next day, although their effect on order imbalance is insignificant. As shown in Panel A, a one-standard-deviation increase in $StdRank$ and $StdRankHor$ is associated with a 1.62% and 2.27% increase in next-day retail buy volume, respectively (Model 4), and a 1.77% and 2.05% increase in next-day retail sell volume, respectively (Model 8), with all effects scaled by the sample standard deviation of the corresponding trading volume measure.

The economic magnitude further increases when using abnormal trading volume. As shown in Panel B, a one-standard-deviation increase in $StdRank$ and $StdRankHor$ is associated with a 2.52% and 5.40% increase in next-day abnormal retail buy volume, respectively (Model 4), and a 3.05% and 5.59% increase in next-day abnormal retail sell volume, respectively (Model 8), with all effects scaled by the sample standard deviation of the corresponding trading volume measure.

In addition to next-day trading volume on day $t+1$, Internet Appendix Table A.6 extends the analysis to trading over $t+2$ to $t+5$ and $t+2$ to $t+20$ days and focuses on abnormal trading volume. We do not find significant evidence for cross-provider dispersion, while cross-horizon dispersion continues to increase abnormal retail trading volume over longer horizons, especially for retail sell volume.

7 Conclusion

We employ six leading LLMs to assess news sentiment regarding the expected impact on stock prices over three horizons: the next day, next week, and next month. First, we find substantial cross-provider dispersion across all sentiment measures and prediction horizons, with the highest dispersion observed for next-week predictions. This dispersion accounts for 12% to 39% of the sample average. We also document significant cross-horizon dispersion across all sentiment measures. Second, LLMs exhibit less disagreement on complex news for next-day and next-week predictions but greater disagreement for next-month predictions. Moreover, this disagreement is systematically linked to differences in causal reasoning rather than noise. Third, while average news sentiment strongly predicts next-day returns, cross-provider and cross-horizon dispersions are negatively associated with returns over the remainder of the week, especially for firms with more opaque information environments and greater operating uncertainty. Our identification strategy exploiting exogenous ChatGPT outages further reinforces the causal link between LLM-based model disagreement and future stock returns. Fourth, LLM-based horizon disagreement amplifies price underreaction to earnings news for small firms, contributing to the post-earnings-announcement drift. In contrast, for large firms, both cross-provider and cross-horizon dispersions appear to delay the immediate price reaction on the announcement day. Finally, both types of dispersion are associated with elevated overall and retail trading volume for at least one week.

Our findings provide timely evidence on the role of generative AI in financial information processing. While investors benefit from a wide range of affordable and accessible generative AI tools to analyze complex information, an unintended consequence may be the amplification of systematic belief dispersion among market participants, thereby increasing information uncertainty and reducing price efficiency. These results also highlight the importance of exercising independent judgment rather than relying uncritically on AI-generated recommendations.

References

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2009). High idiosyncratic volatility and low returns: International and further U.S. evidence. *Journal of Financial Economics*, 91(1):1–23.

Bai, J. J., Boyson, N. M., Cao, Y., Liu, M., and Wan, C. (2023). Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A. *SSRN Electronic Journal*.

Bali, T. G., Kelly, B. T., Moerke, M., and Rahman, J. A. (2025). Machine forecast disagreement. *SSRN Electronic Journal*.

Bamber, L. S., Barron, O. E., and Stober, T. L. (1997). Trading volume and different aspects of disagreement coincident with earnings announcements. *The Accounting Review*, 72(4):575–597.

Banerjee, S. and Kremer, I. (2010). Disagreement and learning: Dynamic patterns of trade. *The Journal of Finance*, 65(4):1269–1302.

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.

Battalio, R., Corwin, S. A., and Jennings, R. (2016). Can brokers have it all? On the relation between make-take fees and limit order execution quality. *The Journal of Finance*, 71(5):2193–2238.

Ben-Rephael, A., Da, Z., and Israelsen, R. D. (2017). It depends on where you search: Institutional investor attention and underreaction to news. *The Review of Financial Studies*, 30(9):3009–3047.

Bernard, D., Blankespoor, E., de Kok, T., and Toynbee, S. (2024). Using GPT models to measure the complexity of business transactions. *SSRN Electronic Journal*.

Bernard, V. L. and Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27:1–36.

Bertomeu, J., Lin, Y., 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.

Blankespoor, E., Croom, J., and Grant, S. (2024). Generative AI and investor processing of financial information. *SSRN Electronic Journal*.

Boehmer, E., Jones, C. M., Zhang, X., and Zhang, X. (2021). Tracking retail investor activity. *The Journal of Finance*, 76(5):2249–2305.

Bybee, J. L. (2025). The ghost in the machine: Generating beliefs with large language models. *Working Paper*.

Chang, A., Dong, X., Martin, X., and Zhou, C. (2025). AI (ChatGPT) democratization and trading inequality. *SSRN Electronic Journal*.

Chang, Y.-C., Hong, H., Tiedens, L., Wang, N., and Zhao, B. (2014). Does diversity lead to diverse opinions? Evidence from languages and stock markets. *Working Paper*.

Chen, J., Tang, G., Zhou, G., and Zhu, W. (2025). ChatGPT and Deepseek: Can they predict the stock market and macroeconomy? *SSRN Electronic Journal*.

Chen, Y., Kelly, B. T., and Xiu, D. (2024). Expected returns and large language models. *SSRN Electronic Journal*.

Cheng, Q., Lin, P., and Zhao, Y. (2025). Does generative AI facilitate investor trading? early evidence from ChatGPT outages. *Journal of Accounting and Economics*, page 101821.

Cookson, J. A., Dim, C., and Niessner, M. (2024). Disagreement on the horizon. *SSRN Electronic Journal*.

Cookson, J. A., Engelberg, J. E., and Mullins, W. (2020). Does partisanship shape investor beliefs? Evidence from the COVID-19 pandemic. *The Review of Asset Pricing Studies*, 10(4):863–893.

Cookson, J. A. and Niessner, M. (2020). Why don't we agree? Evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228.

DellaVigna, S. and Pollet, J. M. (2009). Investor inattention and Friday earnings announcements. *The Journal of Finance*, 64(2):709–749.

Diether, K. B., Malloy, C. J., and Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, 57(5):2113–2141.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Francis, J., LaFond, R., Olsson, P. M., and Schipper, K. (2004). Costs of equity and earnings attributes. *The Accounting Review*, 79(4):967–1010.

Gerken, W. C. and Painter, M. O. (2023). The value of differing points of view: Evidence from financial analysts' geographic diversity. *The Review of Financial Studies*, 36(2):409–449.

Green, J., Hand, J. R. M., and Zhang, X. F. (2017). The characteristics that provide independent information about average U.S. monthly stock returns. *The Review of Financial Studies*, 30(12):4389–4436.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A. (2025). Simulating the survey of professional forecasters. *SSRN Electronic Journal*.

Harris, M. and Raviv, A. (1993). Differences of opinion make a horse race. *The Review of Financial Studies*, 6(3):473–506.

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64(5):2289–2325.

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

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Jha, M., Qian, J., Weber, M., and Yang, B. (2025). ChatGPT and corporate policies. *SSRN Electronic Journal*.

Kandel, E. and Pearson, N. D. (1995). Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy*, 103(4):831–872.

Kim, A. G., Muhn, M., and Nikolaev, V. V. (2024a). Bloated disclosures: Can chatGPT help investors process information? *SSRN Electronic Journal*.

Kim, A. G., Muhn, M., and Nikolaev, V. V. (2024b). Financial statement analysis with large language models. *SSRN Electronic Journal*.

Kim, A. G., Muhn, M., and Nikolaev, V. V. (2024c). From transcripts to insights: Uncovering corporate risks using generative AI. *SSRN Electronic Journal*.

Kozak, S., Nagel, S., and Santosh, S. (2018). Interpreting factor models. *The Journal of Finance*, 73(3):1183–1223.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3):221–247.

Li, S. Z., Maug, E., and Schwartz-Ziv, M. (2022). When shareholders disagree: Trading after shareholder meetings. *The Review of Financial Studies*, 35(4):1813–1867.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *SSRN Electronic Journal*.

Loughran, T. and Mcdonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671.

Loughran, T. and McDonald, B. (2023). Measuring firm complexity. *Journal of Financial and Quantitative Analysis*, 59(6):2487–2514.

Lyonnet, V., Shams, A., and Zhang, S. (2025). What do early-stage investors ask? An LLM analysis of expert calls. *SSRN Electronic Journal*.

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *The Journal of Finance*, 32(4):1151–1168.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Peng, L. (2005). Learning with information capacity constraints. *Journal of Financial and Quantitative Analysis*, pages 307–329.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Wong, T. J., Yi, Y., Yu, G., Zhang, S., and Zhang, T. (2025). Enhancing investor engagement with AI-summarized disclosures. *Working Paper*.

Table 2: Chatbot Responses to AMC's Stake Sale in Saudi Arabia Joint Venture

This table summarizes the chatbot responses from six providers across three prediction horizons for the news headline *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Timeframe	Provider	Impact	Sentiment	Explanation
Next Day	Anthropic	GOOD	70	Shift to licensing partnership seen as a positive strategic move.
	Azure OpenAI	GOOD	70	Market may react positively to strategic shift and potential for increased revenue.
	Google	BAD	35	Initial negative reaction due to loss of direct revenue stream.
	Meta	GOOD	75	AMC's focus on licensing partnership may lead to short-term optimism.
	Mistral	GOOD	75	The sale may provide cash, improving short-term financials.
	OpenAI	BAD	40	Investors may react negatively to the divestment news.
Next Week	Anthropic	BAD	40	Investors may be concerned about the financial impact of the joint venture sale.
	Azure OpenAI	BAD	40	Analysts might express concerns over the implications of reduced ownership and control.
	Google	UNKNOWN	50	Analysts will assess licensing deal's profitability; mixed reactions possible.
	Meta	BAD	30	Analysts may question the impact of this shift on AMC's revenue streams.
	Mistral	UNKNOWN	50	Uncertainty over future earnings potential from the licensing model.
	OpenAI	UNKNOWN	50	Analysts will assess the implications of the licensing shift.
Next Month	Anthropic	UNKNOWN	50	Long-term implications depend on the success of the new licensing model.
	Azure OpenAI	UNKNOWN	50	Long-term effects depend on execution of licensing strategy and market response.
	Google	GOOD	70	If licensing proves more efficient/profitable, stock could rebound.
	Meta	UNKNOWN	50	The long-term effects of this change depend on how it affects AMC's global expansion plans.
	Mistral	BAD	30	Long-term partnership stability and growth prospects are uncertain.
	OpenAI	GOOD	70	Potential for improved cash flow and reduced risk could enhance long-term outlook.

Table 3: Summary Statistics

This table presents summary statistics for the data used in the paper during the period from 2023 to 2024. Panel A reports the means, standard deviations, medians, and quantile distributions of average LLM-based news sentiment and its cross-provider dispersion for next-day predictions. Panels B and C report similar statistics for next-week and next-month predictions, respectively. Panel D reports summary statistics for cross-horizon dispersion. Panel E reports summary statistics for other stock characteristics. Internet Appendix Table A.1 provides a detailed definition for each variable.

	Mean	Std.Dev.	P10	P25	Median	P75	P90
Panel A: News Sentiment from Next-Day Predictions							
AvgNews	2.582	0.594	1.500	2.500	3.000	3.000	3.000
AvgRank	3.578	0.635	2.500	3.500	4.000	4.000	4.000
AvgSent	65.155	13.652	42.944	60.833	71.667	75.000	75.889
StdNews	0.297	0.356	0.000	0.000	0.000	0.548	0.837
StdRank	0.345	0.367	0.000	0.000	0.408	0.548	0.974
StdSent	8.392	7.417	0.000	2.041	5.164	13.934	19.235
Panel B: News Sentiment from Next-Week Predictions							
AvgNews	1.789	0.507	1.000	1.333	1.833	2.333	2.333
AvgRank	2.781	0.516	2.000	2.333	2.833	3.333	3.333
AvgSent	46.299	10.136	32.500	37.500	45.000	55.833	59.167
StdNews	0.705	0.318	0.000	0.516	0.816	0.983	1.033
StdRank	0.722	0.305	0.389	0.516	0.816	0.983	1.033
StdSent	14.561	5.238	6.055	10.328	16.330	18.348	20.595
Panel C: News Sentiment from Next-Month Predictions							
AvgNews	2.204	0.275	2.000	2.000	2.167	2.500	2.583
AvgRank	3.183	0.293	2.833	3.000	3.167	3.500	3.500
AvgSent	55.781	7.249	48.333	50.000	54.167	64.167	65.833
StdNews	0.338	0.244	0.000	0.000	0.408	0.516	0.548
StdRank	0.375	0.250	0.000	0.000	0.408	0.548	0.548
StdSent	8.254	5.499	0.000	4.082	10.206	13.197	14.024
Panel D: Cross-Horizon Dispersion							
StdNewsHor	0.628	0.180	0.359	0.526	0.648	0.763	0.859
StdRankHor	0.642	0.167	0.455	0.526	0.622	0.789	0.859
StdSentHor	14.845	3.349	10.871	12.240	14.031	17.961	19.641
Panel E: Other Stock Characteristics							
R_{t+1}	-0.089	5.958	-4.681	-1.783	-0.007	1.568	4.203
$R_{t+2:t+5}$	0.177	10.008	-7.569	-3.160	0.000	3.011	7.286
$R_{t+2:t+20}$	0.568	19.341	-16.521	-7.221	-0.067	6.931	16.126
Vol_{t+1}	-0.527	2.018	-3.169	-1.701	-0.403	0.767	1.903
$Vol_{t+2:t+5}$	0.828	1.932	-1.724	-0.317	0.930	2.064	3.169
$Vol_{t+2:t+20}$	2.414	1.883	-0.081	1.298	2.517	3.616	4.698
$AbVol_{t+1}$	0.239	0.870	-0.563	-0.232	0.132	0.584	1.161
$AbVol_{t+2:t+5}$	0.134	0.698	-0.484	-0.218	0.060	0.392	0.842
$AbVol_{t+2:t+20}$	0.084	0.576	-0.392	-0.173	0.034	0.282	0.632
$BuyVol_{t+1}$	-4.111	2.204	-6.778	-5.375	-4.028	-2.772	-1.480
$SellVol_{t+1}$	-4.031	2.151	-6.627	-5.263	-3.960	-2.726	-1.466
$OIBS_{t+1}$	-3.401	29.329	-35.919	-16.416	-2.540	10.098	27.085
$AbBuyVol_{t+1}$	0.308	1.079	-0.695	-0.238	0.229	0.799	1.467
$AbSellVol_{t+1}$	0.300	1.048	-0.672	-0.236	0.217	0.774	1.446
$AbOIBS_{t+1}$	0.003	0.294	-0.316	-0.130	0.003	0.134	0.314
CSS	0.018	0.071	0.000	0.000	0.000	0.040	0.100
Log(Market Cap)	7.404	2.607	3.971	5.651	7.473	9.084	10.752
NumAna	7.780	8.594	0.000	1.000	5.000	12.000	20.000
AnaDisp	0.182	1.245	0.007	0.016	0.040	0.102	0.267
ROAVOL	0.087	1.544	0.002	0.007	0.017	0.039	0.097
IVOL	2.943	4.668	0.868	1.209	1.942	3.440	5.704
Complexity	0.499	0.152	0.295	0.387	0.494	0.606	0.706

Table 4: Determinants of News Sentiment Dispersion

This table presents the results of the following daily news-level panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$Disp_{i,n,t} = \alpha + \beta_1 News_{i,n,t} + \beta_2 C_{i,t-1} + \varepsilon_{i,n,t},$$

where $DISP_{i,n,t}$ refers to a list of LLM-based news sentiment dispersion measures for news n related to stock i on day t, proxied by $StdRank$ for next-day (Models 1–3), next-week (Models 4–6), and next-month (Models 7–9) predictions across providers, and by $StdRankHor$ for cross-horizon dispersion (Models 10–12). $News_{i,n,t}$ refers to a list of news characteristics, including *Complexity*, *AvgRank*, and *Overnight*. Vector C stacks all other stock-level control variables, namely, the *Log(Market Cap)*, *Book-to-Market*, *Profitability*, *Investment*, *1M Return*, *Momentum*, *NumAna*, and *AnaDisp*. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Dep. Var. =	StdRank (Next Day)			StdRank (Next Week)			StdRank (Next Month)			StdRankHor		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Complexity	-1.243***	-1.285***	-1.247***	-0.333***	-0.351***	-0.375***	1.608***	1.562***	1.611***	-0.921***	-0.915***	-0.956***
	(-16.46)	(-15.75)	(-13.44)	(-5.39)	(-5.30)	(-4.96)	(23.79)	(22.09)	(20.04)	(-24.80)	(-23.24)	(-21.62)
AvgRank	-0.431***	-0.431***	-0.436***	0.496***	0.494***	0.506***	0.194***	0.194***	0.219***	0.003	0.001	0.006***
	(-109.87)	(-104.71)	(-100.70)	(77.51)	(78.04)	(74.53)	(23.91)	(23.56)	(21.03)	(1.12)	(0.43)	(2.62)
Overnight	0.265***	0.252***	0.227***	0.174***	0.175***	0.213***	-0.123***	-0.099***	-0.111***	0.053***	0.052***	0.066***
	(12.12)	(10.26)	(8.53)	(8.62)	(8.10)	(8.76)	(-6.93)	(-5.27)	(-5.10)	(5.14)	(4.77)	(5.21)
Log(Market Cap)		0.136***	0.082*		-0.226***	-0.195***		-0.194***	-0.227***		-0.040***	-0.068***
		(3.88)	(1.81)		(-7.70)	(-4.70)		(-6.43)	(-6.02)		(-2.74)	(-3.49)
Book-to-Market		0.018	0.043		0.009	0.039		0.034***	0.076*		-0.015**	-0.004
		(1.38)	(1.28)		(0.82)	(1.23)		(3.16)	(1.90)		(-2.36)	(-0.27)
Profitability		0.000	-0.000		0.003	0.006*		0.000	0.004		0.001	0.001*
		(0.16)	(-0.18)		(0.89)	(1.77)		(0.14)	(1.41)		(1.14)	(1.77)
Investment		-0.011	-0.013		-0.025	-0.025		-0.027**	-0.022		0.015***	0.013**
		(-0.62)	(-0.63)		(-1.63)	(-1.25)		(-2.14)	(-1.53)		(2.72)	(2.01)
1M Return		0.066*	0.142**		0.043*	0.026		0.078*	0.170***		-0.012	-0.005
		(1.89)	(2.37)		(1.71)	(0.45)		(1.78)	(3.43)		(-0.91)	(-0.21)
Momentum		-0.041**	-0.002		0.073***	0.060**		0.065***	0.057**		-0.011	-0.010
		(-2.06)	(-0.09)		(3.66)	(2.22)		(3.38)	(2.48)		(-1.30)	(-0.89)
NumAna		0.001	-0.003		-0.002	-0.007		-0.004	0.000		-0.001	0.005
		(0.22)	(-0.43)		(-0.85)	(-1.33)		(-1.27)	(0.08)		(-0.49)	(1.56)
AnaDisp			0.002			-0.018***			-0.010**			-0.006***
			(0.49)			(-3.04)			(-2.21)			(-2.66)
Obs	133,602	110,268	83,562	133,602	110,268	83,562	133,602	110,268	83,562	133,602	110,268	83,562
R-squared	0.665	0.665	0.672	0.556	0.557	0.556	0.512	0.513	0.493	0.684	0.680	0.666

Table 5: Reasoning for the AMC Headline Across Horizons and Providers

This table summarizes the cause–effect chain, based on the explanations provided by each LLM provider, across three prediction horizons for the news headline *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Timeframe	Provider	Impact	Cause → Effect
Next Day	Anthropic	GOOD	Licensing shift → Positive market reaction
	Azure OpenAI	GOOD	Strategic shift → Positive market reaction
	Google	BAD	Loss of revenue stream → Negative market reaction
	Meta	GOOD	Licensing shift → Positive market reaction
	Mistral	GOOD	Stake sale → Positive short-term financials
	OpenAI	BAD	Divestment news → Negative investor reaction
Next Week	Anthropic	BAD	Financial concern → Negative investor sentiment
	Azure OpenAI	BAD	Financial concern → Negative investor sentiment
	Google	UNKNOWN	Licensing shift → Mixed market reaction
	Meta	BAD	Analyst skepticism → Negative investor sentiment
	Mistral	UNKNOWN	Licensing shift → Mixed market reaction
	OpenAI	UNKNOWN	Licensing shift → Mixed market reaction
Next Month	Anthropic	UNKNOWN	Licensing shift → Mixed market reaction
	Azure OpenAI	UNKNOWN	Licensing shift → Mixed market reaction
	Google	GOOD	Licensing shift → Positive market reaction
	Meta	UNKNOWN	Licensing shift → Long-term effects
	Mistral	BAD	Uncertain outlook → Investor caution
	OpenAI	GOOD	Improved cash flow → Positive long-term outlook

Table 6: News Sentiment and Stock Returns

Panel A, Models 1–3 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t+1} = \alpha + \beta_1 NewsSent_{i,t} + \varepsilon_{i,t+1},$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, and $NewsSent_{i,t}$ refers to a list of LLM-based average news sentiment measures for next-day predictions, including $AvgNews$, $AvgRank$, and $AvgSent$. For overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, we enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). For intraday news released between 9 a.m. and 4 p.m. on day t, we enter the position at the close of day t and exit at the close of the next trading day $t+1$ (close-to-close return). Models 4–6 and 7–9 replace $R_{i,t+1}$ with cumulative returns over $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. Panels B and C report similar statistics for next-week and next-month predictions. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	$R_{i,t+1}$			$R_{i,t+2:t+5}$			$R_{i,t+2:t+20}$		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Panel A: Stock Returns Regressed on Lagged News Sentiment from Next-Day Predictions									
AvgNews	0.213***			0.035			-0.161		
	(6.04)			(0.53)			(-1.28)		
AvgRank		0.201***			0.024			-0.180	
		(5.62)			(0.35)			(-1.50)	
AvgSent			0.010***			0.001			-0.008
			(5.83)			(0.38)			(-1.40)
Obs	109,637	109,637	109,637	109,504	109,504	109,504	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.156	0.156	0.156	0.173	0.173	0.173
Panel B: Stock Returns Regressed on Lagged News Sentiment from Next-Week Predictions									
AvgNews	0.291***			0.161**			0.081		
	(5.84)			(2.19)			(0.53)		
AvgRank		0.287***			0.166**			0.061	
		(5.95)			(2.26)			(0.40)	
AvgSent			0.015***			0.009**			0.003
			(5.63)			(2.22)			(0.33)
Obs	109,637	109,637	109,637	109,504	109,504	109,504	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.156	0.156	0.156	0.173	0.173	0.173
Panel C: Stock Returns Regressed on Lagged News Sentiment from Next-Month Predictions									
AvgNews	0.254***			-0.035			-0.262		
	(2.87)			(-0.24)			(-0.93)		
AvgRank		0.226***			0.001			-0.249	
		(2.91)			(0.01)			(-0.95)	
AvgSent			0.010***			-0.001			-0.010
			(2.97)			(-0.12)			(-0.87)
Obs	109,637	109,637	109,637	109,504	109,504	109,504	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.156	0.156	0.156	0.173	0.173	0.173

Table 7: News Sentiment Dispersion and Stock Returns

Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, $AvgRank_{i,t}$ is the average of sentiment ranks across providers, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. Sentiment ranks are based on next-day predictions. For overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, we enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). For intraday news released between 9 a.m. and 4 p.m. on day t, we enter the position at the close of day t and exit at the close of the next trading day $t+1$ (close-to-close return). Models 5–8 and 9–12 replace $R_{i,t+1}$ with cumulative returns over $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	$R_{i,t+1}$				$R_{i,t+2:t+5}$				$R_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	-0.151		-0.143	-0.140	-0.374**		-0.336**	-0.329*	-0.943***		-0.904***	-0.896***
	(-1.26)		(-1.20)	(-1.16)	(-2.21)		(-1.99)	(-1.93)	(-3.10)		(-2.92)	(-2.87)
StdRankHor		-0.209	-0.179	-0.181		-0.926***	-0.854**	-0.859**		-1.072	-0.883	-0.888
		(-1.01)	(-0.87)	(-0.88)		(-2.61)	(-2.41)	(-2.42)		(-1.57)	(-1.27)	(-1.27)
AvgRank	0.131**	0.200***	0.134**	0.142**	-0.148	0.020	-0.134	-0.118	-0.616***	-0.183	-0.600***	-0.584***
	(2.03)	(5.59)	(2.08)	(2.16)	(-1.25)	(0.29)	(-1.15)	(-0.95)	(-3.32)	(-1.52)	(-3.23)	(-2.94)
CSS				-0.133				-0.266				-0.265
				(-0.44)				(-0.51)				(-0.31)
Obs	109,637	109,637	109,637	109,637	109,504	109,504	109,504	109,504	106,873	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.092	0.156	0.156	0.156	0.156	0.173	0.173	0.173	0.173

Table 8: News Sentiment Dispersion and Stock Returns

Models 1–3 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRank_{i,t} \times Outage_{i,t} + \beta_3 StdRankHor_{i,t} + \beta_4 AvgRank_{i,t} + \beta_5 CSS_{i,t} + \beta_6 Outage_{i,t} + \varepsilon_{i,t+1},$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $Outage_{i,t}$ is a dummy variable that equals 1 if a ChatGPT outage occurs and 0 otherwise, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, $AvgRank_{i,t}$ is the average of sentiment ranks across providers, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. Sentiment ranks are based on next-day predictions. For overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, we enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). For intraday news released between 9 a.m. and 4 p.m. on day t, we enter the position at the close of day t and exit at the close of the next trading day $t+1$ (close-to-close return). Models 4–6 and 7–9 replace $R_{i,t+1}$ with cumulative returns over $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	$R_{i,t+1}$			$R_{i,t+2:t+5}$			$R_{i,t+2:t+20}$		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
StdRank	-0.230*	-0.223*	-0.219	-0.479***	-0.445**	-0.438**	-1.098***	-1.063***	-1.055***
	(-1.70)	(-1.65)	(-1.62)	(-2.59)	(-2.41)	(-2.35)	(-3.56)	(-3.39)	(-3.34)
StdRank × Outage	0.214*	0.217*	0.217*	0.280*	0.292*	0.292*	0.416	0.427	0.427
	(1.92)	(1.94)	(1.94)	(1.70)	(1.77)	(1.77)	(1.08)	(1.11)	(1.11)
StdRankHor		-0.189	-0.192		-0.868**	-0.872**		-0.896	-0.901
		(-0.93)	(-0.94)		(-2.45)	(-2.46)		(-1.28)	(-1.29)
AvgRank	0.132**	0.135**	0.143**	-0.148	-0.133	-0.117	-0.615***	-0.599***	-0.582***
	(2.04)	(2.09)	(2.17)	(-1.25)	(-1.14)	(-0.94)	(-3.32)	(-3.22)	(-2.93)
CSS			-0.130			-0.268			-0.279
			(-0.43)			(-0.52)			(-0.33)
Outage	-0.269***	-0.270***	-0.270***	-0.161	-0.166	-0.165	0.136	0.131	0.131
	(-2.73)	(-2.75)	(-2.75)	(-1.29)	(-1.33)	(-1.33)	(0.45)	(0.44)	(0.44)
Obs	109,637	109,637	109,637	109,504	109,504	109,504	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.156	0.156	0.156	0.173	0.173	0.173

Table 9: News Sentiment Dispersion and Stock Returns: Heterogeneity Analysis

Panel A presents the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 StdRank_{i,t} \times Char_{i,t} + \beta_4 StdRankHor_{i,t} \times Char_{i,t} + \beta_5 Char_{i,t} + \varepsilon_{i,t+1},$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, and $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks. Sentiment ranks are based on next-day predictions. $Char_{i,t}$ refers to a list of stock characteristics, including *AvgRank*, *Log(Market Cap)*, *NumAna*, *AnaDisp*, *ROAVOL*, *IVOL*, and *Complexity*. For overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, we enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). For intraday news released between 9 a.m. and 4 p.m. on day t, we enter the position at the close of day t and exit at the close of the next trading day $t+1$ (close-to-close return). Panels B and C replace $R_{i,t+1}$ with cumulative returns over $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Char =	AvgRank Model 1	Log(Market Cap) Model 2	NumAna Model 3	AnaDisp Model 4	ROAVOL Model 5	IVOL Model 6	Complexity Model 7
Panel A: Stock Returns ($R_{i,t+1}$) Regressed on Lagged Sentiment Rank Dispersion and Firm Characteristics							
StdRank	1.152**	-1.155***	-0.341**	0.025	-0.220**	-0.605**	-0.334*
	(2.22)	(-3.23)	(-2.39)	(0.19)	(-2.15)	(-2.39)	(-1.78)
StdRankHor	-1.178	-0.684	-0.421	0.125	-0.222	-0.450	-0.883*
	(-0.89)	(-0.92)	(-1.55)	(0.60)	(-1.13)	(-0.50)	(-1.73)
StdRank × Char	-0.354***	0.148***	0.025***	0.004	-0.041	0.105	0.431
	(-2.60)	(4.01)	(4.60)	(0.10)	(-0.36)	(1.30)	(1.21)
StdRankHor × Char	0.241	0.037	0.024*	-0.173	-0.123	0.042	1.407
	(0.72)	(0.46)	(1.83)	(-1.21)	(-0.63)	(0.13)	(1.55)
AvgRank	0.212	0.230***	0.139**	0.161**	0.085	0.071	0.131**
	(1.10)	(3.55)	(2.16)	(2.57)	(1.38)	(1.16)	(2.05)
Char		-1.241***	-0.016*	0.070	0.087	0.176	-0.763
		(-9.52)	(-1.76)	(0.93)	(0.52)	(0.64)	(-1.23)
Obs	109,637	109,635	109,637	78,274	98,122	106,495	109,637
R-squared	0.092	0.097	0.092	0.112	0.097	0.128	0.092
Panel B: Stock Returns ($R_{i,t+2:t+5}$) Regressed on Lagged Sentiment Rank Dispersion and Firm Characteristics							
StdRank	1.361	-1.596***	-0.439**	-0.270*	-0.264	0.154	-0.477
	(1.44)	(-3.55)	(-2.35)	(-1.89)	(-1.53)	(0.40)	(-1.60)
StdRankHor	-2.799	-1.504	-1.097**	-0.825**	-0.737**	0.792	-0.757
	(-0.99)	(-1.31)	(-2.45)	(-2.53)	(-1.97)	(0.87)	(-0.90)
StdRank × Char	-0.461*	0.195***	0.012	-0.075*	-0.031	-0.188*	0.296
	(-1.84)	(3.75)	(1.47)	(-1.70)	(-0.48)	(-1.74)	(0.53)
StdRankHor × Char	0.480	0.041	0.028	0.059	-0.367***	-0.634**	-0.195
	(0.64)	(0.31)	(1.17)	(0.41)	(-2.99)	(-2.09)	(-0.13)
AvgRank	-0.136	0.063	-0.133	-0.135*	-0.086	-0.122	-0.138
	(-0.36)	(0.53)	(-1.13)	(-1.72)	(-0.70)	(-1.09)	(-1.18)
Char		-2.761***	-0.011	-0.041	0.305***	0.812***	0.078
		(-11.42)	(-0.69)	(-0.44)	(2.69)	(3.84)	(0.08)
Obs	109,504	109,348	109,504	78,184	97,949	106,330	109,504
R-squared	0.156	0.165	0.156	0.179	0.181	0.168	0.156
Panel C: Stock Returns ($R_{i,t+2:t+20}$) Regressed on Lagged Sentiment Rank Dispersion and Firm Characteristics							
StdRank	0.272	-3.389***	-1.043***	-0.319	-0.867***	0.403	-1.255**
	(0.16)	(-3.63)	(-2.70)	(-1.08)	(-2.90)	(0.88)	(-2.11)
StdRankHor	0.656	0.508	-1.599*	-0.566	-0.040	2.041	-2.997**
	(0.18)	(0.22)	(-1.77)	(-0.87)	(-0.06)	(1.62)	(-2.02)
StdRank × Char	-0.334	0.418***	0.014	-0.343*	-0.198**	-0.475***	0.808
	(-0.72)	(4.07)	(0.77)	(-1.77)	(-2.00)	(-3.59)	(0.68)
StdRankHor × Char	-0.414	-0.322	0.087*	-0.144	-0.454**	-1.208***	4.182
	(-0.45)	(-1.21)	(1.82)	(-0.30)	(-2.57)	(-2.92)	(1.60)
AvgRank	-0.134	0.033	-0.601***	-0.351**	-0.450**	-0.513***	-0.604***
	(-0.25)	(0.18)	(-3.24)	(-1.99)	(-2.47)	(-2.71)	(-3.24)
Char		-10.501***	-0.068*	0.248	0.456***	1.728***	-2.359
		(-19.57)	(-1.76)	(0.84)	(3.46)	(5.10)	(-1.35)
Obs	106,873	106,720	106,873	77,140	95,619	103,871	106,873
R-squared	0.173	0.206	0.173	0.224	0.184	0.185	0.173

Table 10: News Sentiment Dispersion and Earnings Announcement Returns

Panel A, Models 1–5 present the results of the following quarterly panel regressions with quarter and day-of-the-week fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t} = \alpha + \beta_1 SUE_{i,t} + \beta_2 SUE_{i,t} \times StdRank_{i,t} + \beta_3 SUE_{i,t} \times StdRankHor_{i,t} + \beta_4 SUE_{i,t} \times AnaDisp_{i,t-1} + \beta_5 StdRank_{i,t} + \beta_6 StdRankHor_{i,t} + \beta_7 AnaDisp_{i,t-1} + \beta_8 C_{i,t-1} + \varepsilon_{i,t},$$

where $R_{i,t}$ is stock i's return on earnings announcement day t (close-to-close return). $SUE_{i,t}$ is the standardized unexpected earnings (SUE), defined as the difference between the actual earnings for the quarter and the average of the most recent analyst forecasts, divided by the standard deviation of those forecasts. $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $AnaDisp_{i,t-1}$ is the standard deviation of analyst forecasts divided by the absolute value of the mean forecast. We include all news released between the close of day $t-1$ and the close of day t, with sentiment ranks based on next-day predictions. Vector C stacks all other stock-level control variables, namely, the *AvgRank*, *Log(Market Cap)*, *Book-to-Market*, *Profitability*, *Investment*, and *Momentum*. Models 6–10 and 11–15 replace $R_{i,t}$ with cumulative returns over $t+1$ to $t+30$ ($R_{i,t+1:t+30}$) and $t+1$ to $t+40$ ($R_{i,t+1:t+40}$), respectively. In these cases, we enter the position at the market close of day t and exit at the close of trading day $t+30$ or $t+40$, respectively. Panel A focuses on small firms with market capitalizations below the NYSE median breakpoint, while Panel B reports similar statistics for large firms above this threshold. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Earnings Announcement Returns Regressed on Lagged SUE and Sentiment Rank Dispersion (Small Firms)

	$R_{i,t}$					$R_{i,t+1:t+30}$					$R_{i,t+1:t+40}$				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15
SUE	0.521***	0.399***	0.808***	0.735***	0.719***	0.213**	0.092	-0.397	-0.885*	-0.881*	0.224**	0.077	-0.582	-1.205**	-1.214**
	(9.27)	(4.41)	(3.77)	(3.34)	(3.29)	(2.32)	(0.54)	(-1.27)	(-1.93)	(-1.91)	(2.06)	(0.39)	(-1.50)	(-1.99)	(-1.99)
SUE × StdRank		0.175		0.069	0.052		0.281		0.606	0.606		0.328		0.766	0.742
		(0.96)		(0.43)	(0.32)		(0.75)		(1.42)	(1.44)		(0.78)		(1.53)	(1.49)
SUE × StdRankHor			-0.525	-0.459	-0.491			0.890*	1.319**	1.324**			1.172*	1.721**	1.693**
			(-1.52)	(-1.38)	(-1.50)			(1.69)	(2.20)	(2.20)			(1.83)	(2.22)	(2.19)
SUE × AnaDisp					0.188*					-0.037					0.138
					(1.89)					(-0.25)					(0.94)
StdRank		3.411*		3.605**	3.654**		-0.585		-0.715	-0.698		0.415		0.548	0.683
		(1.85)		(1.97)	(1.99)		(-0.20)		(-0.24)	(-0.24)		(0.10)		(0.14)	(0.17)
StdRankHor			-1.876	-2.905	-2.948			1.095	0.464	0.544			-3.822	-4.999	-4.762
			(-0.58)	(-0.91)	(-0.92)			(0.18)	(0.07)	(0.09)			(-0.56)	(-0.72)	(-0.69)
AnaDisp					0.083					-0.199					-0.620
					(0.23)					(-0.29)					(-0.81)
AvgRank		7.781***	6.181***	7.840***	7.797***		2.510	2.936**	2.554	2.563		3.882	3.905**	4.093	4.064
		(7.16)	(8.79)	(7.19)	(7.14)		(1.24)	(2.01)	(1.23)	(1.23)		(1.41)	(2.24)	(1.47)	(1.45)
Log(Market Cap)	0.114	0.213	0.160	0.193	0.187	-0.290	-0.272	-0.254	-0.276	-0.277	-1.705**	-1.667**	-1.702**	-1.715**	-1.727**
	(0.32)	(0.58)	(0.44)	(0.51)	(0.50)	(-0.45)	(-0.43)	(-0.39)	(-0.42)	(-0.42)	(-2.09)	(-2.03)	(-2.06)	(-2.05)	(-2.06)
Book-to-Market	0.295	0.244	0.222	0.238	0.237	-1.322***	-1.376***	-1.321***	-1.370***	-1.364***	-1.689***	-1.749***	-1.696***	-1.749***	-1.728***
	(1.05)	(0.89)	(0.83)	(0.88)	(0.86)	(-2.98)	(-3.08)	(-2.93)	(-3.04)	(-3.04)	(-3.39)	(-3.43)	(-3.36)	(-3.40)	(-3.39)
Profitability	0.028	0.054	0.049	0.049	0.050	0.020	0.029	0.036	0.033	0.034	0.031	0.045	0.046	0.043	0.045
	(0.51)	(0.93)	(0.82)	(0.87)	(0.87)	(0.19)	(0.28)	(0.34)	(0.31)	(0.32)	(0.24)	(0.35)	(0.35)	(0.33)	(0.35)
Investment	-0.522	-0.596	-0.584	-0.610	-0.595	-0.911	-0.928	-0.909	-0.892	-0.890	-1.001	-1.033	-1.000	-0.989	-0.960
	(-1.32)	(-1.37)	(-1.27)	(-1.39)	(-1.39)	(-1.44)	(-1.42)	(-1.40)	(-1.38)	(-1.37)	(-1.02)	(-1.01)	(-0.98)	(-0.98)	(-0.97)
Momentum	-0.093	-0.259	-0.242	-0.256	-0.268	-0.234	-0.300	-0.309	-0.306	-0.304	0.078	-0.014	-0.019	-0.020	-0.032
	(-0.32)	(-0.87)	(-0.82)	(-0.86)	(-0.90)	(-0.29)	(-0.36)	(-0.36)	(-0.36)	(-0.36)	(0.10)	(-0.02)	(-0.02)	(-0.02)	(-0.04)
Obs	2,936	2,936	2,936	2,936	2,936	2,935	2,935	2,935	2,935	2,935	2,935	2,935	2,935	2,935	2,935
R-squared	0.041	0.064	0.063	0.065	0.066	0.104	0.105	0.105	0.106	0.106	0.114	0.116	0.116	0.117	0.117

Table 10 (continued)

Panel B: Earnings Announcement Returns Regressed on Lagged SUE and Sentiment Rank Dispersion (Large Firms)

	$R_{i,t}$					$R_{i,t+1:t+30}$					$R_{i,t+1:t+40}$				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15
SUE	0.517***	0.483***	0.635***	0.756***	0.765***	0.110***	0.080	0.258	0.231	0.248	0.123***	0.114	0.388**	0.439*	0.452*
	(15.55)	(11.41)	(4.87)	(5.33)	(5.38)	(2.86)	(1.12)	(1.49)	(1.08)	(1.16)	(2.73)	(1.34)	(2.09)	(1.91)	(1.95)
SUE × StdRank		-0.105		-0.188*	-0.194*		0.083		0.038	0.035		0.020		-0.076	-0.063
		(-1.02)		(-1.85)	(-1.86)		(0.49)		(0.22)	(0.19)		(0.11)		(-0.38)	(-0.32)
SUE × StdRankHor			-0.291	-0.392*	-0.407**			-0.242	-0.219	-0.232			-0.429	-0.472	-0.448
			(-1.48)	(-1.95)	(-2.01)			(-0.89)	(-0.76)	(-0.80)			(-1.42)	(-1.49)	(-1.42)
SUE × AnaDisp					0.056					0.008					-0.208
					(0.46)					(0.04)					(-1.07)
StdRank		3.088***		3.334***	3.368***		-0.452		-0.201	-0.156		0.196		0.727	0.720
		(4.37)		(4.66)	(4.67)		(-0.52)		(-0.23)	(-0.18)		(0.18)		(0.67)	(0.67)
StdRankHor			0.103	-0.911	-1.104			-1.746	-1.733	-2.092			-3.461	-3.621	-3.928*
			(0.08)	(-0.68)	(-0.82)			(-0.86)	(-0.84)	(-1.02)			(-1.59)	(-1.65)	(-1.80)
AnaDisp					0.703					1.434***					1.510**
					(1.57)					(2.95)					(2.38)
AvgRank		4.634***	3.449***	4.674***	4.690***		0.027	0.154	0.090	0.111		0.214	0.090	0.346	0.347
		(10.16)	(11.39)	(10.13)	(10.10)		(0.06)	(0.50)	(0.19)	(0.24)		(0.40)	(0.29)	(0.64)	(0.65)
Log(Market Cap)	-0.110	-0.048	-0.049	-0.054	-0.024	0.101	0.104	0.097	0.097	0.155	0.254	0.257	0.243	0.242	0.299
	(-1.43)	(-0.63)	(-0.63)	(-0.70)	(-0.33)	(0.60)	(0.61)	(0.57)	(0.57)	(0.90)	(1.43)	(1.43)	(1.36)	(1.36)	(1.64)
Book-to-Market	-0.260**	-0.226*	-0.222*	-0.221*	-0.237**	0.339	0.341	0.347	0.346	0.312	0.152	0.153	0.164	0.164	0.125
	(-2.28)	(-1.92)	(-1.94)	(-1.87)	(-2.06)	(1.51)	(1.51)	(1.54)	(1.53)	(1.37)	(0.61)	(0.61)	(0.66)	(0.66)	(0.50)
Profitability	0.024	0.029	0.030	0.031	0.031	0.055**	0.055**	0.056**	0.056**	0.055**	0.015	0.015	0.016	0.017	0.016
	(0.74)	(1.00)	(1.04)	(1.06)	(1.10)	(2.41)	(2.40)	(2.46)	(2.43)	(2.54)	(0.53)	(0.53)	(0.60)	(0.62)	(0.61)
Investment	0.093	-0.006	-0.021	-0.004	-0.015	0.694	0.686	0.688	0.687	0.665	0.864	0.861	0.860	0.864	0.840
	(0.29)	(-0.02)	(-0.07)	(-0.01)	(-0.05)	(1.20)	(1.19)	(1.19)	(1.19)	(1.17)	(1.43)	(1.42)	(1.42)	(1.43)	(1.39)
Momentum	-0.883**	-1.269***	-1.238***	-1.285***	-1.257***	2.102***	2.091***	2.068***	2.072***	2.130***	3.095***	3.080***	3.051***	3.040***	3.104***
	(-2.32)	(-3.28)	(-3.20)	(-3.33)	(-3.27)	(3.28)	(3.24)	(3.21)	(3.21)	(3.32)	(3.54)	(3.51)	(3.48)	(3.46)	(3.56)
Obs	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788	4,788
R-squared	0.083	0.133	0.129	0.134	0.135	0.201	0.201	0.201	0.201	0.203	0.247	0.247	0.248	0.248	0.249

Table 11: News Sentiment Dispersion and Trading Volume

Panel A, Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$Vol_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $Vol_{i,t+1}$ is the logarithm of the trading volume for stock i on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. We only include overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. Models 5–8 and 9–12 replace $Vol_{i,t+1}$ with the logarithm of the trading volume over $t+2$ to $t+5$ ($Vol_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($Vol_{i,t+2:t+20}$), respectively. Panel B presents similar regressions with calendar-day fixed effects, replacing trading volume with abnormal trading volume ($AbVol$), defined as the difference between log volume on day t and the firm's average log volume from $t-140$ to $t-20$ trading days. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Trading Volume Regressed on Lagged Sentiment Rank Dispersion

	$Vol_{i,t+1}$				$Vol_{i,t+2:t+5}$				$Vol_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.140***		0.130***	0.123***	0.053***		0.051***	0.050***	0.037***		0.036***	0.037***
	(7.58)		(7.03)	(6.68)	(3.60)		(3.43)	(3.35)	(2.79)		(2.68)	(2.81)
StdRankHor		0.266***	0.239***	0.246***		0.064**	0.053*	0.054*		0.029	0.021	0.019
		(7.13)	(6.45)	(6.66)		(2.01)	(1.68)	(1.72)		(1.02)	(0.75)	(0.68)
AvgRank	0.042***	-0.026***	0.037***	0.021*	0.004	-0.021***	0.003	0.001	0.006	-0.012***	0.006	0.010
	(3.87)	(-4.23)	(3.42)	(1.92)	(0.48)	(-4.29)	(0.36)	(0.07)	(0.74)	(-2.64)	(0.67)	(1.17)
CSS				0.295***				0.049				-0.082**
				(5.53)				(1.14)				(-2.17)
Obs	86,042	86,042	86,042	86,042	85,897	85,897	85,897	85,897	83,832	83,832	83,832	83,832
R-squared	0.846	0.846	0.846	0.846	0.882	0.882	0.882	0.882	0.905	0.905	0.905	0.905

Panel B: Abnormal Trading Volume Regressed on Lagged Sentiment Rank Dispersion

	$AbVol_{i,t+1}$				$AbVol_{i,t+2:t+5}$				$AbVol_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.137***		0.120***	0.110***	0.048***		0.042***	0.037**	0.034***		0.032***	0.031**
	(7.30)		(6.39)	(5.87)	(3.23)		(2.79)	(2.50)	(2.89)		(2.70)	(2.57)
StdRankHor		0.382***	0.356***	0.367***		0.142***	0.134***	0.138***		0.050*	0.043	0.045
		(9.70)	(9.05)	(9.30)		(4.48)	(4.18)	(4.31)		(1.84)	(1.58)	(1.63)
AvgRank	0.046***	-0.019***	0.039***	0.015	0.011	-0.012**	0.008	-0.002	0.018**	0.002	0.017**	0.014*
	(3.91)	(-2.74)	(3.27)	(1.24)	(1.13)	(-2.00)	(0.84)	(-0.18)	(2.27)	(0.36)	(2.14)	(1.66)
CSS				0.431***				0.187***				0.062*
				(7.57)				(4.00)				(1.67)
Obs	86,120	86,120	86,120	86,120	85,975	85,975	85,975	85,975	83,915	83,915	83,915	83,915
R-squared	0.048	0.049	0.049	0.050	0.040	0.040	0.041	0.041	0.031	0.031	0.031	0.031

Table 12: News Sentiment Dispersion and Retail Trading Volume

Panel A, Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$BuyVol_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $BuyVol_{i,t+1}$ is the logarithm of the retail buy volume for stock i on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. We only include overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. Models 5–8 and 9–12 replace $BuyVol_{i,t+1}$ with the logarithm of the retail sell volume ($SellVol_{i,t+1}$) and retail order imbalance ($OIBS_{i,t+1}$), respectively. Panel B presents similar regressions with calendar-day fixed effects, where retail buy volume is replaced with abnormal retail buy volume ($AbBuyVol$), defined as the difference between log retail buy volume on day t and the firm's average log retail buy volume from $t-140$ to $t-20$ trading days. Similarly, retail sell volume is replaced with abnormal retail sell volume ($AbSellVol$), and retail order imbalance is replaced with abnormal retail order imbalance ($AbOIBS$), both defined relative to their respective firm-level averages over the $t-140$ to $t-20$ trading-day window. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Retail Trading Volume Regressed on Lagged Sentiment Rank Dispersion

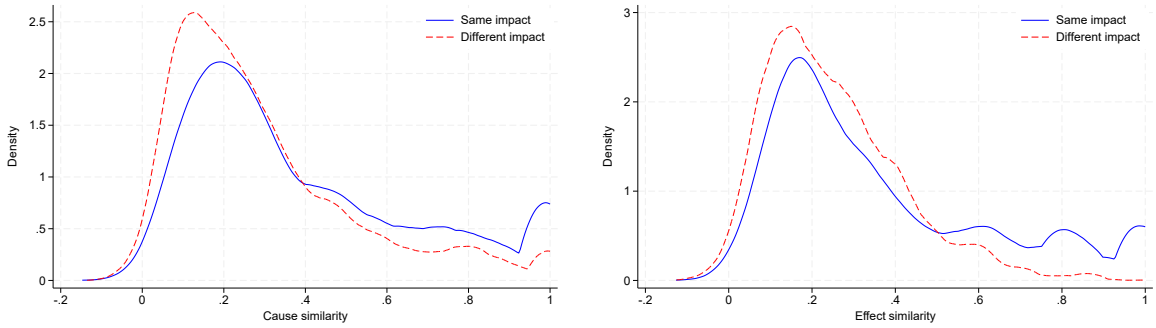
	$BuyVol_{i,t+1}$				$SellVol_{i,t+1}$				$OIBS_{i,t+1}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.120***		0.108***	0.097***	0.125***		0.114***	0.104***	-0.139		-0.176	-0.248
	(4.97)		(4.41)	(3.97)	(5.38)		(4.86)	(4.48)	(-0.27)		(-0.34)	(-0.48)
StdRankHor		0.309***	0.286***	0.299***		0.277***	0.253***	0.264***		0.802	0.839	0.923
		(6.12)	(5.63)	(5.89)		(5.84)	(5.30)	(5.54)		(0.70)	(0.73)	(0.81)
AvgRank	0.058***	0.000	0.053***	0.026*	0.046***	-0.015**	0.040***	0.018	0.682**	0.750***	0.665**	0.489*
	(4.26)	(0.03)	(3.81)	(1.86)	(3.29)	(-1.99)	(2.89)	(1.24)	(2.49)	(4.05)	(2.44)	(1.65)
CSS				0.484***				0.420***				3.223*
				(6.59)				(6.01)				(1.90)
Obs	76,121	76,121	76,121	76,121	76,145	76,145	76,145	76,145	76,435	76,435	76,435	76,435
R-squared	0.792	0.792	0.792	0.792	0.794	0.794	0.794	0.794	0.089	0.089	0.089	0.089

Panel B: Abnormal Retail Trading Volume Regressed on Lagged Sentiment Rank Dispersion

	$AbBuyVol_{i,t+1}$				$AbSellVol_{i,t+1}$				$AbOIBS_{i,t+1}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.103***		0.087***	0.074***	0.114***		0.098***	0.087***	-0.001		-0.001	-0.001
	(4.25)		(3.58)	(3.05)	(4.78)		(4.08)	(3.63)	(-0.14)		(-0.13)	(-0.26)
StdRankHor		0.353***	0.334***	0.349***		0.360***	0.339***	0.351***		-0.000	-0.000	0.001
		(6.62)	(6.24)	(6.51)		(7.32)	(6.86)	(7.11)		(-0.03)	(-0.02)	(0.05)
AvgRank	0.048***	-0.001	0.041***	0.010	0.037**	-0.018**	0.030**	0.003	0.008***	0.008***	0.008***	0.006**
	(3.15)	(-0.18)	(2.71)	(0.66)	(2.45)	(-2.20)	(1.98)	(0.19)	(2.86)	(4.48)	(2.85)	(2.07)
CSS				0.546***				0.477***				0.029*
				(6.83)				(6.17)				(1.71)
Obs	75,912	75,912	75,912	75,912	75,934	75,934	75,934	75,934	76,226	76,226	76,226	76,226
R-squared	0.034	0.035	0.035	0.036	0.035	0.035	0.035	0.036	0.011	0.011	0.011	0.011

Figure 1: Causes and Effects Similarity

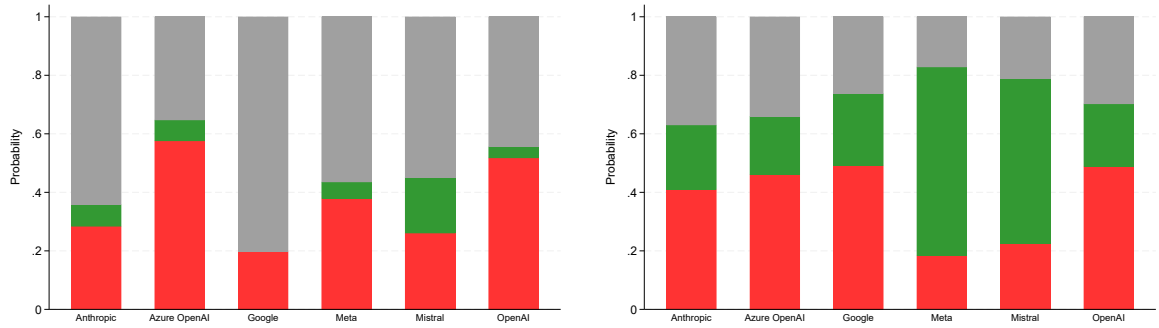
For each headline–horizon pair, we form all provider pairs ($6 \times 5 = 30$) and compute cosine similarities for causes and effects separately using sentence embeddings, with causes and effects extracted from model explanations of news headlines. Provider pairs are then grouped by whether they assign the same impact label (blue line) or different labels (red line). Subfigures (a) and (b) show the density distributions of cosine similarity scores for causes and effects, respectively.



(a) Causes **(b)** Effects

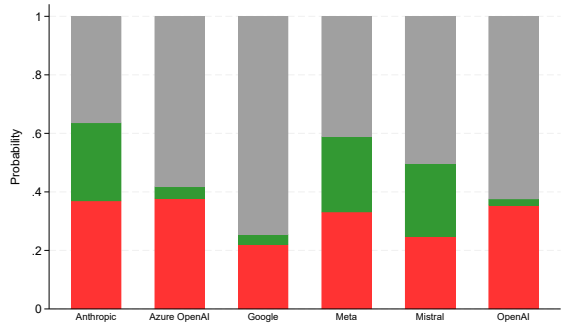
Figure 2: Cause Impact Mapping across Topics

This figure reports, for each provider and cause category, the distribution of impact labels (`GOOD`, `BAD`, `UNKNOWN`) assigned to extracted causes, pooled across prediction horizons. Causes are grouped into six primary topic categories: corporate governance, cost/liquidity, demand/market, macroeconomic, regulation/policy, and technological, displayed in subfigures (a) to (f). Categories are formed through an automated keyword-based classification system that maps cause text into predefined taxonomic groups. Colors denote impact types, with red indicating `GOOD`, green indicating `BAD`, and gray indicating `UNKNOWN`.

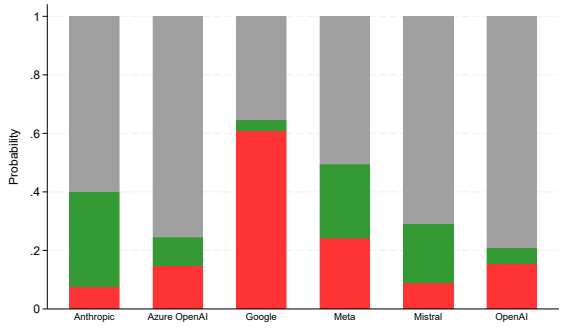


(a) Corporate Governance

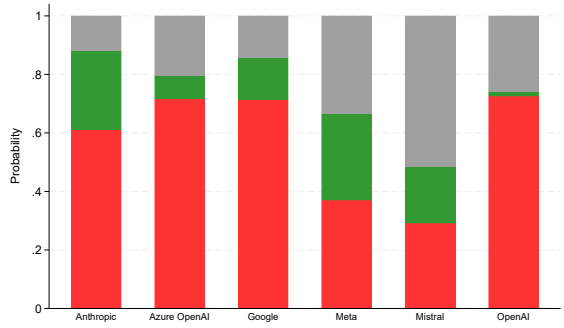
(b) Cost/Liquidity



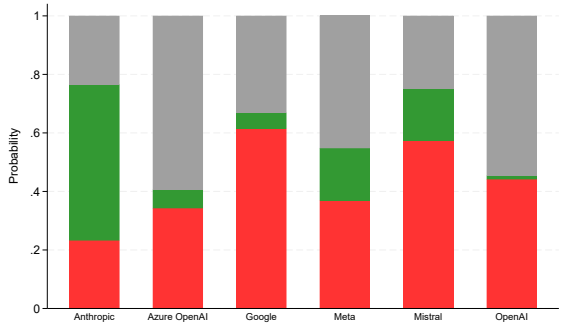
(c) Demand/Market



(d) Macroeconomic



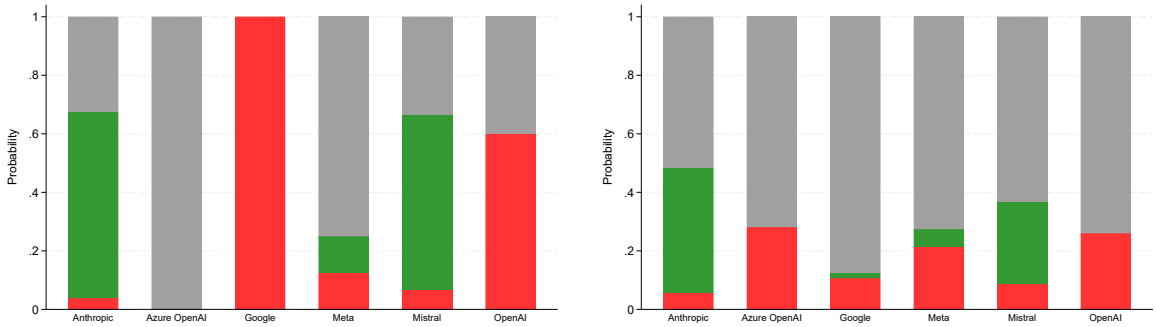
(e) Regulation/Policy



(f) Technological

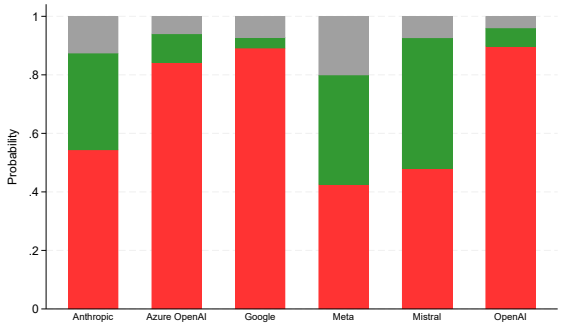
Figure 3: Effect Impact Mapping across Topics

This figure reports, for each provider and effect category, the distribution of assigned impact labels (`GOOD`, `BAD`, `UNKNOWN`), pooled across prediction horizons. Effects are grouped into six primary categories: competitive position, innovation/capacity, market/valuation, profitability/cost, revenue/growth, and risk/volatility, displayed in subfigures (a) to (f). Categories are formed through an automated keyword-based classification system that maps cause text into predefined taxonomic groups. Colors denote impact types, with red indicating `GOOD`, green indicating `BAD`, and gray indicating `UNKNOWN`.

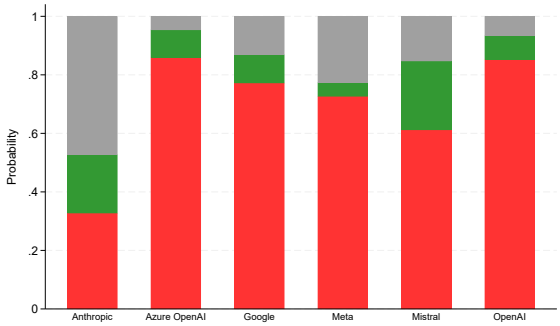


(a) Competitive Position

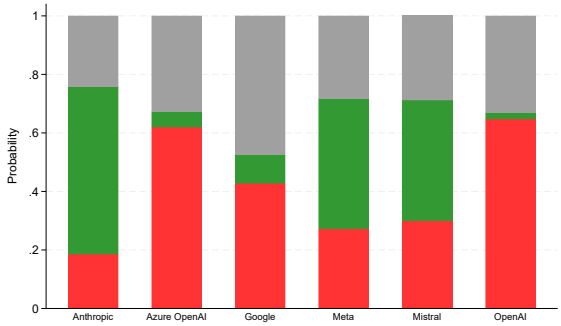
(b) Innovation/Capacity



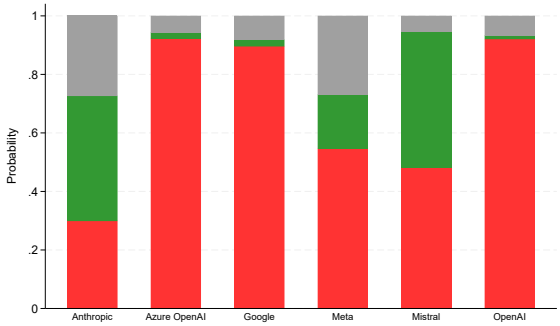
(c) Market/Valuation



(d) Profitability/Cost



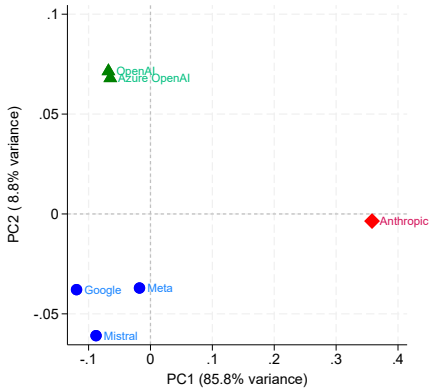
(e) Revenue/Growth



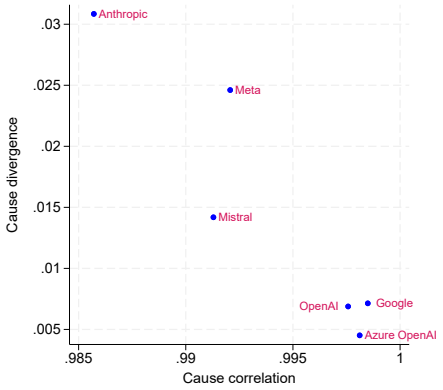
(f) Risk/Volatility

Figure 4: Provider Comparison

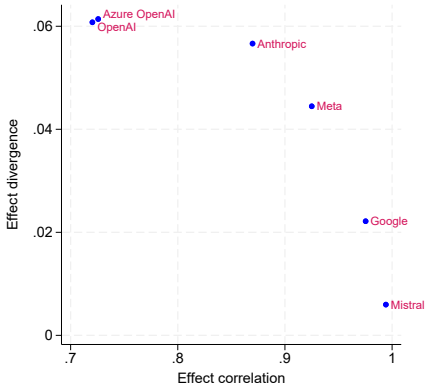
This figure compares LLM providers in terms of overall reasoning styles and temporal stability. Subfigure (a) presents a principal component analysis (PCA) projection of providers in the reasoning space, where proximity reflects similarity in extracted cause–effect chains. Subfigures (b) and (c) report stability metrics for causes and effects, respectively, using two complementary measures: Pearson correlation coefficients, which capture consistency in category distributions across adjacent horizons, and Jensen–Shannon divergence, which captures distributional shifts.



(a) Clustering



(b) Cause Stability



(c) Effect Stability

Internet Appendix for

"When Machines Disagree: Evidence from Large Language Models"

This Internet Appendix provides an overview of the large language models, variable definitions, and supplementary empirical results. Most of these results are discussed in the paper.

Section A. Large Language Models

Section B. Data Description

- Table A.1: Variable Definitions

Section C. Additional Analyses

- Table A.2: News Sentiment Dispersion and Stock Returns: Subsamples (Robustness of Table 7)
- Table A.3: News Sentiment Dispersion and Stock Returns: Alternative Sentiment Measures (Robustness of Table 7)
- Table A.4: News Sentiment Dispersion and Trading Volume: Subsample (Robustness of Table 11)
- Table A.5: News Sentiment Dispersion and Trading Volume: Alternative Prediction Horizons (Robustness of Table 11)
- Table A.6: News Sentiment Dispersion and Retail Trading Volume over Longer Horizons (Robustness of Table 12)

A Large Language Models

OpenAI (`gpt-4o-mini`) is the leading commercial LLM provider by API market share. The GPT-4o-mini model, with approximately 8 billion parameters, represents a lightweight version of OpenAI's flagship GPT-4o, optimized for low-latency, multimodal use cases while retaining strong reasoning ability. Its design balances cost-efficiency with general-purpose capabilities across text, image, and audio modalities. GPT-4o was trained on publicly available and licensed data up to October 2023, with performance tuned via human feedback. While the full technical details remain undisclosed, GPT-4o models are known for high alignment and robustness but face criticism for limited transparency and overreliance on post-training safety layers. OpenAI's models are accessible both directly and via Microsoft's Azure OpenAI platform.

Azure OpenAI (`gpt-4o-mini`) mirrors the OpenAI model, with approximately 8 billion parameters, but is hosted through Microsoft Azure's infrastructure. This integration enables access to OpenAI models with enterprise-grade compliance, data governance, and Azure-native deployment. While the underlying architecture is identical to OpenAI's `gpt-4o-mini`, users frequently observe differences in outputs due to variations in system prompting, deployment environments, and post-processing layers. These discrepancies can lead to subtle but meaningful divergences in behavior, even when querying the same model version. We include Azure OpenAI specifically because it closely approximates the model environment used in Microsoft's web-based Copilot tools–such as those embedded in Bing and Microsoft 365–which are among the most widely deployed generative AI applications targeting retail users. Thus, despite its enterprise-facing infrastructure, Azure plays a central role in shaping the AI-generated information ecosystem available to everyday investors.

Google (`gemini-1.5-flash`) develops the Gemini family (formerly Bard), aiming to integrate large-scale AI directly into Google Search and Workspace. Gemini 1.5 Flash, estimated at roughly 32 billion parameters, was released in mid-2024 as a high-speed

variant tuned for retrieval-augmented generation, document summarization, and real-time chat. While less powerful than Gemini 1.5 Pro, Flash models benefit from Google's internal infrastructure, including proprietary TPU training and integration with Google's search index. Gemini 1.5 models are trained on data through November 2023. Market adoption is strong in consumer-facing tools, though developer usage lags behind OpenAI due to platform limitations and inconsistent API latency.

Anthropic (`claude-3-haiku-20240307`) offers the Claude series, which emphasizes safety and alignment through a Constitutional AI approach. The Claude 3 Haiku variant, released in March 2024, has approximately 20 billion parameters, and is the smallest and fastest in the Claude 3 family, designed for cost-effective inference with competitive accuracy. Trained up to August 2023, it performs well in tasks requiring careful reasoning and moderation. Claude models are particularly favored for enterprise deployments in sensitive environments, although users occasionally note limitations in instruction-following depth compared to GPT-4-class models. Anthropic is backed by major cloud providers but offers its models primarily through its own APIs and select partnerships.

Meta (`llama3`) is the leading provider of open-weight LLMs. LLaMA 3, released in April 2024, comes in 8-billion and 70-billion-parameter variants, designed for widespread research and on-device deployment. Trained on 15 trillion tokens with a data cutoff in March 2023, LLaMA 3 offers strong performance on standard NLP tasks with low hardware requirements. While it lacks out-of-the-box alignment safeguards, its openness allows developers full control over fine-tuning and deployment. Meta's strategy contrasts with closed providers by fostering an open ecosystem; however, concerns over misuse persist due to minimal usage gating.

Mistral (`mistral`) is a French AI startup known for releasing compact, open-source models trained from scratch. The Mistral model (released late 2023) is a 7-billion parameter dense transformer trained on publicly available data up to July 2023, offering high

performance-to-cost efficiency in local and cloud environments. It outperforms earlier open models like LLaMA 2 and has been widely adopted in privacy-sensitive applications. Mistral's philosophy emphasizes lightweight, reproducible models, though their smaller size makes them less competitive in long-context or complex reasoning tasks.

B Data Description

Table A.1: Variable Definitions

Variables	Definitions
A. LLM-based News Sentiment and Dispersions	
AvgNews	For each news at each prediction horizon, the average *News Score* across all LLM providers, where *News Score* equals 1 if $Impact = \texttt{BAD}$, 2 if $Impact = \texttt{UNKNOWN}$, and 3 if $Impact = \texttt{GOOD}$. *Impact* is an output from the chatbot.
AvgRank	For each news at each prediction horizon, the average *Sentiment Rank* across all LLM providers for each news, where *Sentiment Rank* equals 1 if $Sentiment\ Score \leq 20$, 2 if $20 < Sentiment\ Score \leq 40$, 3 if $40 < Sentiment\ Score \leq 60$, 4 if $60 < Sentiment\ Score \leq 80$, and 5 if $Sentiment\ Score > 80$. *Sentiment Score* is an output from the chatbot.
AvgSent	For each news at each prediction horizon, the average *Sentiment Score* across all LLM providers for each news, where *Sentiment Score* is an output from the chatbot, ranging from 0 to 100.
StdNews	For each news at each prediction horizon, the standard deviation of *News Score* across all LLM providers, where *News Score* is defined as in *AvgNews*.
StdRank	For each news at each prediction horizon, the standard deviation of *Sentiment Rank* across all LLM providers, where *Sentiment Rank* is defined as in *AvgRank*.
StdSent	For each news at each prediction horizon, the standard deviation of *Sentiment Score* across all LLM providers, where *Sentiment Score* is defined as in *AvgSent*.
StdNewsHor	For each news-provider pair, first compute the standard deviation of *News Score* across three prediction horizons, where *News Score* is defined as in *AvgNews*. Then, average these standard deviations across providers for each news.
StdRankHor	For each news-provider pair, first compute the standard deviation of *Sentiment Rank* across three prediction horizons, where *Sentiment Rank* is defined as in *AvgRank*. Then, average these standard deviations across providers for each news.
StdSentHor	For each news-provider pair, first compute the standard deviation of *Sentiment Score* across three prediction horizons, where *Sentiment Score* is defined as in *AvgSent*. Then, average these standard deviations across providers for each news.
B. News Characteristics	
Word Count	The total number of words in the news headline.
Fog	The Gunning-Fog readability index, calculated as $Fog = (words\ per\ sentence + percent\ of\ complex\ words) \times 0.4$, where complex words are those with three or more syllables, following Li (2008).
%Complexity	The number of complex words divided by the total number of words in the news headline, where complex words are identified using the complexity lexicon of Loughran and McDonald (2023).

Table A.1 (continued)

Complexity	Each month, we rank all news based on *Word Count*, *Fog*, and *%Complexity* (as defined above), and compute their respective percentile ranks (normalized between 0 and 1). We then calculate the average of these percentile ranks.
Overnight	A dummy variable that equals 1 if the news is released before 9:00 a.m. or after 4:00 p.m. on a trading day, and 0 otherwise.
C. Other Stock Characteristics	
Log(Market Cap)	The logarithm of the market capitalization, calculated as the number of common shares outstanding times share price, as reported in CRSP.
Book-to-Market	The book value of equity divided by market capitalization at fiscal year-end. The book value of equity equals the stockholders equity (COMPUSTAT annual item SEQ) plus deferred tax and investment tax credit (item TXDITC, imputing zero if missing) minus the book value of the preferred stock. Depending on availability, we use the redemption value (item PSTKRV), liquidation value (item PSTKL), carrying value (item PSTK), or zero in that order as the book value of the preferred stock.
Profitability	Profits divided by book equity, where profits equals revenues (COMPUSTAT annual item REVT) minus cost of goods (item COGS) minus selling, general, and administrative expense (item XSGA, imputing zero if missing) minus interest expense (item XINT, imputing zero if missing), and book equity is defined as in *Book-to-Market*, following Fama and French (2015).
Investment	Year-over-year fraction change in book assets (COMPUSTAT annual item AT), following Fama and French (2015).
1M Return	Monthly stock return, as reported in CRSP.
Momentum	In a given month t, computed as the cumulative return from month $t-11$ to month $t-1$, following Jegadeesh and Titman (1993).
NumAna	The number of analysts covering the firm, as reported in I/B/E/S.
AnaDisp	The standard deviation of analyst forecasts divided by the absolute value of the mean forecast, as reported in I/B/E/S.
CSS	The composite sentiment score, ranging between -1 and 1, as reported in RavenPack.
ROAVOL	The standard deviation of ROA over 16 quarters, where ROA is calculated as income before extraordinary items (COMPUSTAT quarterly item IBQ) divided by one-quarter lagged total assets (item ATQ), following Francis et al. (2004) and Green et al. (2017).
IVOL	The standard deviation of residuals from the Fama-French three-factor model (Fama and French, 1993), estimated using daily returns within a month, following Ang et al. (2009). Specifically, we regress daily stock excess returns on the market, size, and book-to-market factor returns and obtain the residuals.

Table A.1 (continued)

SUE	The difference between the actual earnings for the quarter and the average of the most recent analyst forecasts, divided by the standard deviation of those forecasts, as reported in I/B/E/S.
Retail OIBS	The retail order imbalance for share volume for stock i on day d is calculated as follows: $OIBS_{i,d} = \frac{SB_{i,d}-SS_{i,d}}{SB_{i,d}+SS_{i,d}}$, where $SB_{i,d}$ and $SS_{i,d}$ are the buy and sell volumes from retail orders, respectively, following Barber et al. (2024).

C Additional Analyses

Table A.2: News Sentiment Dispersion and Stock Returns: Subsamples

Panel A, Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, $AvgRank_{i,t}$ is the average of sentiment ranks across providers, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. We only include overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. We enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). Models 5–8 and 9–12 replace $R_{i,t+1}$ with cumulative returns over $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. Panel B reports similar statistics for the subperiod covering the year 2024 only. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	$R_{i,t+1}$				$R_{i,t+2:t+5}$				$R_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Panel A: Stock Returns Regressed on Lagged Sentiment Rank Dispersion (Overnight News)												
StdRank	-0.150		-0.143	-0.140	-0.502**		-0.478**	-0.461**	-1.049***		-1.000***	-0.995***
	(-0.99)		(-0.96)	(-0.94)	(-2.54)		(-2.42)	(-2.31)	(-2.92)		(-2.74)	(-2.72)
StdRankHor		-0.198	-0.169	-0.171		-0.665	-0.567	-0.587		-1.323*	-1.118	-1.124
		(-0.83)	(-0.73)	(-0.75)		(-1.62)	(-1.39)	(-1.43)		(-1.70)	(-1.41)	(-1.42)
AvgRank	0.116	0.188***	0.119	0.125	-0.223	0.021	-0.211	-0.168	-0.617***	-0.104	-0.592***	-0.578**
	(1.49)	(4.74)	(1.55)	(1.56)	(-1.59)	(0.27)	(-1.53)	(-1.14)	(-2.76)	(-0.77)	(-2.65)	(-2.46)
CSS				-0.099				-0.807				-0.249
				(-0.23)				(-1.05)				(-0.24)
Obs	86,036	86,036	86,036	86,036	85,980	85,980	85,980	85,980	83,989	83,989	83,989	83,989
R-squared	0.103	0.103	0.103	0.103	0.161	0.161	0.161	0.161	0.180	0.180	0.180	0.180
Panel B: Stock Returns Regressed on Lagged Sentiment Rank Dispersion (Year 2024)												
StdRank	-0.217*		-0.230*	-0.220*	-0.492**		-0.454*	-0.459*	-0.861*		-0.884*	-0.856*
	(-1.71)		(-1.79)	(-1.72)	(-2.20)		(-1.94)	(-1.93)	(-1.82)		(-1.80)	(-1.73)
StdRankHor		0.086	0.182	0.180		-0.727	-0.537	-0.536		-0.046	0.317	0.311
		(0.30)	(0.63)	(0.63)		(-1.39)	(-0.98)	(-0.97)		(-0.05)	(0.32)	(0.31)
AvgRank	0.063	0.157***	0.068	0.092	-0.209*	-0.047	-0.223*	-0.235*	-0.693**	-0.341*	-0.685**	-0.620**
	(0.87)	(3.16)	(0.93)	(1.24)	(-1.69)	(-0.55)	(-1.82)	(-1.74)	(-2.59)	(-1.94)	(-2.59)	(-2.19)
CSS				-0.414				0.203				-1.102
				(-0.96)				(0.23)				(-0.92)
Obs	52,469	52,469	52,469	52,469	52,302	52,302	52,302	52,302	49,858	49,858	49,858	49,858
R-squared	0.127	0.127	0.127	0.127	0.182	0.182	0.182	0.182	0.195	0.195	0.195	0.195

Table A.3: News Sentiment Dispersion and Stock Returns: Alternative Sentiment Measures

Panel A, Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$R_{i,t+1} = \alpha + \beta_1 StdNews_{i,t} + \beta_2 StdNewsHor_{i,t} + \beta_3 AvgNews_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $R_{i,t+1}$ is stock i's return on day $t+1$, $StdNews_{i,t}$ is the standard deviation of LLM-based news scores across providers, $StdNewsHor_{i,t}$ is the cross-horizon dispersion of news scores, $AvgNews_{i,t}$ is the average of news scores across providers, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. News scores are based on next-day predictions. For overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, we enter the position at the market open and exit at the close of day $t+1$ (open-to-close return). For intraday news released between 9 a.m. and 4 p.m. on day t, we enter the position at the close of day t and exit at the close of the next trading day $t+1$ (close-to-close return). Models 5–8 and 9–12 replace $R_{i,t+1}$ with cumulative returns over $t+2$ to $t+5$ ($R_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($R_{i,t+2:t+20}$), respectively. In these cases, we enter the position at the market close of day $t+1$ and exit at the close of trading day $t+5$ or $t+20$, respectively. Panel B reports similar statistics that replace news scores with sentiment scores. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	$R_{i,t+1}$				$R_{i,t+2:t+5}$				$R_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Panel A: Stock Returns Regressed on Lagged News Score Dispersion												
StdNews	-0.155		-0.136	-0.133	-0.281		-0.203	-0.193	-0.694**		-0.609*	-0.597*
	(-1.63)		(-1.42)	(-1.36)	(-1.57)		(-1.18)	(-1.10)	(-2.30)		(-1.89)	(-1.84)
StdNewsHor		-0.213	-0.132	-0.134		-0.671*	-0.551	-0.555		-0.965*	-0.613	-0.618
		(-1.18)	(-0.73)	(-0.74)		(-1.89)	(-1.61)	(-1.62)		(-1.67)	(-0.99)	(-1.00)
AvgNews	0.146***	0.210***	0.152***	0.157***	-0.087	0.025	-0.061	-0.044	-0.464**	-0.173	-0.435**	-0.414**
	(2.81)	(5.88)	(2.98)	(2.92)	(-0.72)	(0.36)	(-0.54)	(-0.36)	(-2.46)	(-1.36)	(-2.30)	(-2.06)
CSS				-0.083				-0.253				-0.296
				(-0.27)				(-0.49)				(-0.35)
Obs	109,637	109,637	109,637	109,637	109,504	109,504	109,504	109,504	106,873	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.092	0.156	0.156	0.156	0.156	0.173	0.173	0.173	0.173
Panel B: Stock Returns Regressed on Lagged Sentiment Score Dispersion												
StdSent	-0.009		-0.008	-0.008	-0.014*		-0.011	-0.010	-0.035**		-0.031*	-0.031*
	(-1.60)		(-1.53)	(-1.48)	(-1.69)		(-1.28)	(-1.23)	(-2.25)		(-1.95)	(-1.90)
StdSentHor		-0.009	-0.006	-0.006		-0.052***	-0.048***	-0.048***		-0.065*	-0.054	-0.054
		(-0.92)	(-0.65)	(-0.65)		(-3.13)	(-2.92)	(-2.92)		(-1.93)	(-1.56)	(-1.56)
AvgSent	0.006**	0.010***	0.007**	0.007**	-0.005	0.004	-0.000	0.000	-0.023**	-0.004	-0.018*	-0.017*
	(2.33)	(5.74)	(2.52)	(2.57)	(-0.89)	(1.52)	(-0.06)	(0.07)	(-2.56)	(-0.67)	(-1.88)	(-1.68)
CSS				-0.137				-0.240				-0.304
				(-0.45)				(-0.46)				(-0.36)
Obs	109,637	109,637	109,637	109,637	109,504	109,504	109,504	109,504	106,873	106,873	106,873	106,873
R-squared	0.092	0.092	0.092	0.092	0.156	0.156	0.156	0.156	0.173	0.173	0.173	0.173

Table A.4: News Sentiment Dispersion and Trading Volume: Subsample

Panel A, Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$Vol_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $Vol_{i,t+1}$ is the logarithm of the trading volume for stock i on day $t+1$, $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. We only include overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. Models 5–8 and 9–12 replace $Vol_{i,t+1}$ with the logarithm of the trading volume over $t+2$ to $t+5$ ($Vol_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($Vol_{i,t+2:t+20}$), respectively. Our analysis is restricted to the subperiod spanning the year 2024. Panel B presents similar regressions with calendar-day fixed effects, replacing trading volume with abnormal trading volume ($AbVol$), defined as the difference between log volume on day t and the firm's average log volume from $t-140$ to $t-20$ trading days. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Trading Volume Regressed on Lagged Sentiment Rank Dispersion (Year 2024)

	$Vol_{i,t+1}$				$Vol_{i,t+2:t+5}$				$Vol_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.084***		0.072***	0.065***	0.006		0.002	0.001	0.025		0.027*	0.027*
	(3.82)		(3.18)	(2.88)	(0.36)		(0.13)	(0.04)	(1.63)		(1.78)	(1.76)
StdRankHor		0.234***	0.210***	0.219***		0.071*	0.070*	0.072*		-0.036	-0.045	-0.045
		(4.42)	(3.90)	(4.08)		(1.79)	(1.73)	(1.79)		(-1.07)	(-1.33)	(-1.32)
AvgRank	0.031**	0.007	0.037***	0.020	0.009	0.011	0.011	0.007	0.018*	0.005	0.016*	0.016
	(2.40)	(0.84)	(2.88)	(1.49)	(0.89)	(1.57)	(1.09)	(0.66)	(1.93)	(0.81)	(1.77)	(1.64)
CSS				0.363***				0.087				0.014
				(4.64)				(1.51)				(0.29)
Obs	41,247	41,247	41,247	41,247	41,102	41,102	41,102	41,102	39,180	39,180	39,180	39,180
R-squared	0.874	0.874	0.874	0.874	0.912	0.912	0.912	0.912	0.934	0.934	0.934	0.934

Panel B: Abnormal Trading Volume Regressed on Lagged Sentiment Rank Dispersion (Year 2024)

	$AbVol_{i,t+1}$				$AbVol_{i,t+2:t+5}$				$AbVol_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.105***		0.083***	0.074***	0.024		0.011	0.007	0.044**		0.040**	0.037**
	(4.10)		(3.18)	(2.84)	(1.08)		(0.52)	(0.34)	(2.45)		(2.24)	(2.09)
StdRankHor		0.422***	0.396***	0.407***		0.213***	0.210***	0.214***		0.076*	0.063	0.066*
		(7.14)	(6.67)	(6.84)		(4.42)	(4.31)	(4.39)		(1.92)	(1.60)	(1.66)
AvgRank	0.044***	0.023**	0.057***	0.036**	0.023	0.025***	0.030**	0.020	0.037***	0.022***	0.039***	0.034***
	(2.89)	(2.32)	(3.69)	(2.29)	(1.65)	(2.86)	(2.14)	(1.45)	(3.12)	(3.00)	(3.27)	(2.73)
CSS				0.422***				0.181**				0.106*
				(4.75)				(2.56)				(1.84)
Obs	41,343	41,343	41,343	41,343	41,201	41,201	41,201	41,201	39,281	39,281	39,281	39,281
R-squared	0.055	0.056	0.057	0.057	0.042	0.043	0.043	0.043	0.033	0.033	0.033	0.033

Table A.5: News Sentiment Dispersion and Trading Volume: Alternative Prediction Horizons

Panel A, Models 1–4 present the results of the following daily panel regressions with calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$AbVol_{i,t+1} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $AbVol_{i,t+1}$ is the abnormal trading volume for stock i on day $t+1$, defined as the difference between log volume on day t and the firm's average log volume from $t-140$ to $t-20$ trading days. $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. We only include overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-week predictions. Models 5–8 and 9–12 replace $AbVol_{i,t+1}$ with the abnormal trading volume over $t+2$ to $t+5$ ($AbVol_{i,t+2:t+5}$) and $t+2$ to $t+20$ ($AbVol_{i,t+2:t+20}$), respectively. Panel B reports similar statistics for next-month predictions. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	$AbVol_{i,t+1}$				$AbVol_{i,t+2:t+5}$				$AbVol_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Panel A: Abnormal Trading Volume Regressed on Lagged Sentiment Rank Dispersion from Next-Week Predictions												
StdRank	0.213***		0.184***	0.177***	0.097***		0.087***	0.084***	0.068***		0.064***	0.064***
	(12.82)		(10.24)	(9.79)	(7.43)		(5.95)	(5.74)	(6.32)		(5.30)	(5.25)
StdRankHor		0.346***	0.192***	0.168***		0.141***	0.068*	0.060		0.081***	0.028	0.026
		(7.88)	(4.09)	(3.58)		(4.09)	(1.80)	(1.59)		(2.82)	(0.87)	(0.82)
AvgRank	-0.162***	-0.021**	-0.129***	-0.145***	-0.064***	-0.001	-0.052***	-0.058***	-0.025***	0.018***	-0.020**	-0.021**
	(-13.09)	(-2.08)	(-8.68)	(-9.61)	(-6.24)	(-0.15)	(-4.20)	(-4.64)	(-3.08)	(2.66)	(-2.07)	(-2.18)
CSS				0.373***				0.130***				0.025
				(6.70)				(2.85)				(0.69)
Obs	86,120	86,120	86,120	86,120	85,975	85,975	85,975	85,975	83,915	83,915	83,915	83,915
R-squared	0.050	0.049	0.050	0.051	0.041	0.040	0.041	0.041	0.032	0.031	0.032	0.032
Panel B: Abnormal Trading Volume Regressed on Lagged Sentiment Rank Dispersion from Next-Month Predictions												
StdRank	0.099***		0.141***	0.132***	0.073***		0.090***	0.087***	0.017		0.026**	0.025**
	(5.07)		(7.33)	(6.91)	(4.99)		(6.15)	(5.90)	(1.40)		(2.14)	(2.08)
StdRankHor		0.455***	0.504***	0.485***		0.170***	0.202***	0.195***		0.097***	0.106***	0.105***
		(10.53)	(11.84)	(11.48)		(4.90)	(5.85)	(5.67)		(3.29)	(3.58)	(3.56)
AvgRank	-0.028*	0.069***	0.045***	0.025	-0.018	0.026**	0.011	0.004	0.026**	0.046***	0.042***	0.040***
	(-1.77)	(4.29)	(2.68)	(1.48)	(-1.47)	(2.02)	(0.79)	(0.26)	(2.57)	(4.32)	(3.80)	(3.65)
CSS				0.250***				0.091**				0.017
				(4.51)				(1.97)				(0.47)
Obs	86,120	86,120	86,120	86,120	85,975	85,975	85,975	85,975	83,915	83,915	83,915	83,915
R-squared	0.047	0.049	0.050	0.050	0.040	0.040	0.041	0.041	0.031	0.032	0.032	0.032

Table A.6: News Sentiment Dispersion and Retail Trading Volume over Longer Horizons

Panel A, Models 1–4 present the results of the following daily panel regressions with firm and calendar day fixed effects, as well as their corresponding t-statistics with standard errors clustered at the firm and calendar day levels:

$$AbBuyVol_{i,t+2:t+5} = \alpha + \beta_1 StdRank_{i,t} + \beta_2 StdRankHor_{i,t} + \beta_3 AvgRank_{i,t} + \beta_4 CSS_{i,t} + \varepsilon_{i,t+1},$$

where $AbBuyVol_{i,t+2:t+5}$ is the abnormal retail buy volume for stock i over days $t+2$ to $t+5$. Abnormal retail buy volume on day t is defined as the difference between log retail buy volume on day t and the firm's average log retail buy volume from $t-140$ to $t-20$ trading days. $StdRank_{i,t}$ is the standard deviation of LLM-based sentiment ranks across providers, $StdRankHor_{i,t}$ is the cross-horizon dispersion of sentiment ranks, and $CSS_{i,t}$ is the composite sentiment score from Ravenpack. We only include overnight news released before 9 a.m. on trading day $t+1$ or after 4 p.m. on the previous day t, with sentiment ranks based on next-day predictions. Models 5–8 and 9–12 replace $AbBuyVol_{i,t+2:t+5}$ with abnormal retail sell volume ($AbSellVol_{i,t+2:t+5}$) and abnormal retail order imbalance ($AbOIBS_{i,t+2:t+5}$), respectively. Panel B reports similar statistics for abnormal trading activity over days $t+2$ to $t+20$, replacing $AbBuyVol_{i,t+2:t+5}$, $AbSellVol_{i,t+2:t+5}$, and $AbOIBS_{i,t+2:t+5}$ with $AbBuyVol_{i,t+2:t+20}$, $AbSellVol_{i,t+2:t+20}$, and $AbOIBS_{i,t+2:t+20}$, respectively. Internet Appendix Table A.1 provides a detailed definition for each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Abnormal Retail Trading Volume Regressed on Lagged Sentiment Rank Dispersion

	$AbBuyVol_{i,t+2:t+5}$				$AbSellVol_{i,t+2:t+5}$				$AbOIBS_{i,t+2:t+5}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.015		0.010	0.006	0.022		0.014	0.010	0.001		0.001	0.002
	(0.76)		(0.52)	(0.29)	(1.11)		(0.73)	(0.52)	(0.17)		(0.37)	(0.50)
StdRankHor		0.098***	0.096**	0.101***		0.155***	0.152***	0.157***		-0.013*	-0.014**	-0.014**
		(2.61)	(2.53)	(2.65)		(4.19)	(4.06)	(4.19)		(-1.94)	(-1.97)	(-2.04)
AvgRank	-0.006	-0.013*	-0.008	-0.018	-0.010	-0.020***	-0.013	-0.023*	0.004**	0.004***	0.004**	0.005***
	(-0.45)	(-1.92)	(-0.60)	(-1.40)	(-0.82)	(-3.18)	(-1.08)	(-1.84)	(2.33)	(3.52)	(2.50)	(2.88)
CSS				0.184***				0.173***				-0.018*
				(2.82)				(2.94)				(-1.85)
Obs	76,459	76,459	76,459	76,459	76,458	76,458	76,458	76,458	76,523	76,523	76,523	76,523
R-squared	0.032	0.032	0.032	0.033	0.035	0.035	0.035	0.035	0.011	0.011	0.011	0.011

Panel B: Abnormal Retail Trading Volume Regressed on Lagged Sentiment Rank Dispersion

	$AbBuyVol_{i,t+2:t+20}$				$AbSellVol_{i,t+2:t+20}$				$AbOIBS_{i,t+2:t+20}$			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
StdRank	0.013		0.012	0.010	0.012		0.008	0.008	0.001		0.001	0.001
	(0.85)		(0.75)	(0.63)	(0.77)		(0.51)	(0.48)	(0.47)		(0.62)	(0.65)
StdRankHor		0.032	0.029	0.031		0.085***	0.083***	0.084***		-0.007	-0.007	-0.007
		(0.99)	(0.90)	(0.97)		(2.79)	(2.71)	(2.73)		(-1.45)	(-1.50)	(-1.51)
AvgRank	-0.003	-0.009*	-0.003	-0.008	-0.007	-0.013**	-0.009	-0.010	0.003**	0.002***	0.003**	0.003**
	(-0.26)	(-1.65)	(-0.32)	(-0.72)	(-0.69)	(-2.47)	(-0.87)	(-0.97)	(2.41)	(3.17)	(2.52)	(2.49)
CSS				0.078				0.021				-0.003
				(1.44)				(0.44)				(-0.45)
Obs	75,420	75,420	75,420	75,420	75,390	75,390	75,390	75,390	75,427	75,427	75,427	75,427
R-squared	0.028	0.028	0.028	0.028	0.027	0.027	0.027	0.027	0.014	0.014	0.014	0.014