Measuring Misinformation in Financial Markets

Jianqing Fan, Qingfu Liu, Yang Song, Zilu Wang

August 11, 2024

Abstract

We propose a framework for measuring firm-level misinformation. By leveraging advanced machine learning and AI technologies and guided by the theory of the wisdom of select crowds, we transform and categorize hundreds of millions of unstructured texts into comparable information sets using two-tier (topic, entity)-classification using large language models, extract "truth" from each set of comparable information using the information reliability weighted consensus (average), and quantify the degree of misinformation based on divergence from the "truth", defined as the information reliability weighted average deviation from the consensus. Applying our framework to analyze 254.8 million textual materials, we validate its effectiveness in quantifying misinformation in several ways. We find that firms with weaker balance sheets and poorer governance structures exhibit higher misinformation, and misinformation spikes during major corporate events. We also demonstrate that misinformation significantly impacts investors' attention, trading volumes, stock returns, and risks.

Keywords: misinformation; machine learning; AI; large language models; topic-entity classification, divergence of information.

Fan is with Bendheim Center for Finance and Department of Operations Research and Financial Engineering, Princeton University; Email: jqfan@princeton.edu. Liu and Wang are with the School of Economics, Fudan University; Email: liuqf@fudan.edu.cn and zlwang22@m.fudan.edu.cn. Song is with the Foster School of Business, University of Washington; Email: songy18@uw.edu.

1 Introduction

Accurate and timely information is the cornerstone of financial markets (e.g., Fama, 1970; Grossman and Stiglitz, 1980). Yet, the rise of misleading or false information, often referred to as "misinformation," is becoming increasingly prevalent (e.g., Grinberg et al., 2019; Kartal and Tyran, 2022).[1] Such misinformation can possibly skew market perceptions, lead to poor decision-making, and cause significant market inefficiencies, thereby increasing financial risks. Despite its widespread presence, academic research on misinformation in financial markets remains limited. One major challenge lies in the difficulty of measuring misinformation due to factors such as the asymmetric nature and lack of disclosure of truthful information. Overcoming this challenge is crucial for understanding the impact of misinformation on market dynamics and investor behaviors.

Our study aims to address the urgent need to measure and analyze misinformation in financial markets. Leveraging the advancements in machine learning and AI, we propose a systematic framework for measuring firm-level misinformation. Our approach is inspired by two principles in discerning truth: information consistency and the wisdom of select crowds (Surowiecki, 2005; Mannes et al., 2014). Information consistency asserts that genuine information regarding the same subject should maintain coherence across diverse sources,[2] and when faced with conflicting information, the wisdom of select crowds highlights the power of prioritizing information from reliable sources.[3] The input to our framework is a massive firm-level textual corpus, which includes hundreds of millions of textual materials from various sources such as firm announcements, news reports, analyst reports, and social media posts. By integrating information from diverse sources, we capture collective perceptions to infer the underlying "truth" and quantify the degree of misinformation based on divergence

[1] Misinformation includes both unintentional errors and deliberately fabricated content.

[2] This information consistency principle is acknowledged across various fields, including psychology (Brashier and Marsh, 2020), computer science (Shu et al., 2017), and behavioral theory (Sadler, 2021).

[3] This principle also aligns with the "truth discovery" theory in the field of information science, which seeks to ascertain the veracity of information from conflicting sources (Li et al., 2016; Xu et al., 2021a).

from the "truth."

Specifically, our approach overcomes two primary challenges: (*i*) transforming and categorizing unstructured textual materials across diverse sources and formatting into comparable information and (*ii*) extracting the "truth" from each set of comparable information.

To overcome the first challenge, we develop a two-tier topic classification (TC) and entity extraction (EE) procedure using large language models (LLMs). Specifically, the TC procedure identifies the topics of the text (e.g., "financial performance" or "business operations"),[4] while the EE procedure further extracts specific entities[5] under a given topic (e.g., "operating income" under the topic of financial performance or "market share" under the topic of business operations). *Different information of the same company can be meaningfully compared only if they have identical topic-entity labels.* Subsequently, we also use LLMs to convert narrative texts into structured, quantitative vectors to quantify the underlying semantic insights.

To tackle the second challenge, we take a network approach. We construct information networks to identify the "truth" for each set of comparable information (i.e., information of the same company with identical topic-entity labels). Specifically, for a given comparable information set, each piece of information is treated as a node in the network, with connections among nodes measured by semantic similarity (Strogatz, 2001; Chandrasekaran and Mago, 2021). Recognizing that not all information contributes equally to understanding the truth, we assign higher "reliability" weights to information (nodes) from more trustworthy sources and information with higher logical coherence in sentence constructions. In this way, we prioritize the wisdom of select crowds. Consistent with the principle of information consistency, we identify an optimal representation of the "truth" in the information network, which is the consensus reinforced with reliability weights.

In the last step, we quantify the degree of misinformation for each set of comparable

[4]We selectively concentrate on eight topics that significantly impact both short-term and long-term firm values. The eight topics are business operations, financial performance, risk management, corporate governance, environmental responsibility, social responsibility, human capital, and R&D innovation.

[5]An "entity" here refers to the specific subject described in the text.

information by measuring deviations from the "truth" within the information set. With this, we calculate firm-level misinformation using a bottom-up approach across all the information sets regarding a firm. The framework for measuring misinformation is illustrated in the following chart.

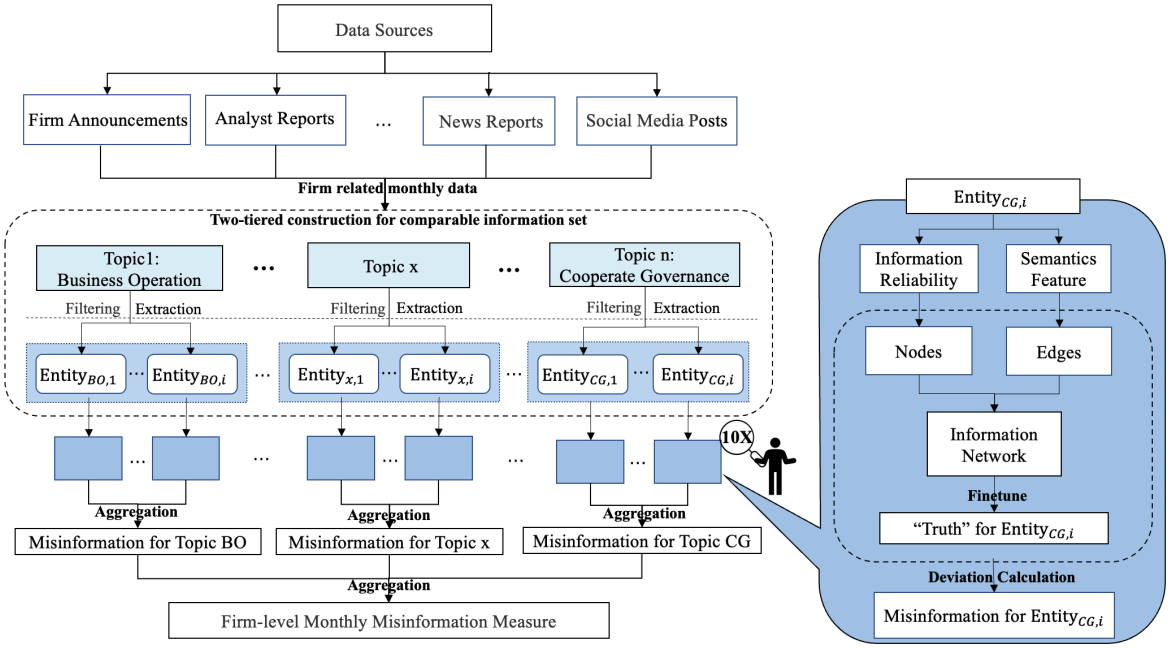


Figure 1 Conceptual Framework for Measuring Misinformation

We apply our method to measure monthly firm-level misinformation in the Chinese financial markets, the second-largest capital market in the world. Our study focuses on Chinese A-share companies from January 2015 to December 2022. After data processing, our textual corpus contains 254,826,060 articles,[6] pertaining to 3,994 public firms. While we study the Chinese markets due to the readily availability of this extensive corpus, our approach can be applied to other markets as well.

To start, we conduct several exercises to validate the misinformation measure (*MISI*). Our first validation exercise is based on the quality of corporate information disclosure. If our measure of misinformation is effective, firms with higher disclosure quality should ex-

[6]For simplicity, we refer to all firm announcements, news reports, analyst reports, and social media posts as articles.

hibit lower *MISI*. We find that a firm's *MISI* significantly negatively correlates with its information disclosure quality, whether assessed using regulatory or market-based metrics. Secondly, we analyze the occurrence of disclosed misinformation-related events, disclosed by either regulators or firms.[7] We find that the misinformation measure indeed registers high values during these events: in the time series, 89% of the disclosed events occur when a firm's *MISI* exceeds its average level, and cross-sectionally, firms with higher *MISI* consistently have more disclosed misinformation-related events. Both panel and logit regression models reinforce these findings. Thirdly, if our misinformation measure is valid, an increase in *MISI* should correspond with heightened public skepticism about a firm's authenticity (Dyck et al., 2008). We indeed observe that higher levels of *MISI* about a firm significantly correlate with increased public skepticism, evidenced by more discussions questioning the firm's authenticity.

After validating the effectiveness of our misinformation measure, we move on to explore the drivers of misinformation. We first examine whether and how firm characteristics are related to the degree of misinformation. We find that firms with weak balance sheets, high financial risks, and poor corporate governance structures have higher average levels of misinformation. Additionally, we find that as industries become more concentrated, the levels of misinformation generally increase. This finding suggests that industry concentration leads to a decrease in overall information transparency, fostering a conducive environment for the spread of misinformation.

Moreover, we examine whether the occurrence of major corporate events can drive misinformation. During significant corporate events, there is likely to be a surge in information speculation and dissemination. This creates an environment ripe for the spread of misinformation, as stakeholders may attempt to influence market perceptions and reactions. We find that major corporate actions, e.g., financing operations and mergers and acquisitions, indeed trigger the circulation of misinformation. This finding underscores the importance of

[7]Specifically, we identify a list of company events linked to misinformation based on company and regulatory announcements, comprising 1,340 events associated with 539 firms covered in our sample.

managing misinformation during major corporate activities.

In the last part of the paper, we explore the implications of misinformation on financial markets, focusing particularly on investor behaviors and stock return dynamics. To this end, we first investigate the relation between misinformation and investors' attention. Investors monitor information that could impact stock prices, making them particularly sensitive to news that appears to present new opportunities or risks, even if such news is misleading. We find that when a firm's misinformation level is higher, investors are more attracted to the company, as measured by their search intensity on the search engine Baidu.com (the Chinese equivalent of Google).[8] In addition, misinformation can significantly predict investors' attention levels over the next month.

Second, we examine whether higher levels of misinformation also predict higher trading volumes. For one, increased investor attention can translate into more trading activities. Additionally, higher levels of misinformation might prompt investors to react more frequently and aggressively in the market as they attempt to capitalize on perceived opportunities or mitigate potential risks. Our findings indicate that higher levels of misinformation are indeed associated with increased trading volume in the markets.

Third, as misinformation triggers investors' attention and trading activities, we also find a significant positive impulse of stock prices to the level of misinformation, followed by significant reversals. As a result, the degree of misinformation exhibits strong negative predictive power for future stock returns. For example, the top-decile firms with the highest misinformation significantly underperform the bottom-decile firms by about 84 basis points (bps) over the next month.

Last but not least, we document the strong explanatory and predictive power of misinformation on stock return volatility and crash risk. Intuitively, when genuine information about a company is obfuscated and misinformation circulates, the risk level of the company is likely to rise as well (e.g., Hutton et al., 2009). Consistent with this hypothesis, we find that

[8]For robustness, we also use the total search volume on 13 major Chinese stock trading apps.

the degree of misinformation is positively associated with contemporaneous return volatility and stock crash risk,[9] and it also strongly predicts these risk measures over the next month.

Notably, we also find that the influence of misinformation is stronger among firms that are held more by small and retail shareholders, suggesting small and retail investors are more susceptible to misinformation. In addition, the economic implications of misinformation are robust after controlling for commonly used disagreement measures. Overall, our framework provides a valuable tool for measuring misinformation, which can be applied to understanding the sources and impact of misinformation in financial markets.

Literature Our paper makes several contributions to the literature. Classic works have established the pivotal role of information in shaping investor behavior and market dynamics (e.g., Fama, 1970; Grossman and Stiglitz, 1980). While extensive research has analyzed the role of information in financial markets, very limited research has been devoted to studying misinformation, which has become increasingly prevalent in recent years.[10] Our study aims to bridge this gap by introducing a framework for systematically measuring misinformation at the firm level. Our research reveals that the circulation of misinformation significantly influences investors' attention, trading behaviors, and stock prices. To the best of our knowledge, we are the first to quantify and analyze misinformation for individual companies.

Second, our study contributes to the literature of financial textual analysis (e.g., Shiller, 2017; Gentzkow et al., 2019; Jiang et al., 2019; Loughran and McDonald, 2020; Chen et al., 2022; Fan et al., 2024). Existing research shows that analyst reports, social media posts, and news articles can significantly influence investor behavior and asset prices (e.g., Cohen et al., 2020; Beckmann et al., 2024). However, these sources are not without biases, which can skew perceptions and decision-making in financial markets (Qin et al., 2018; Fan et al., 2023).

[9] Following Kim et al. (2011), we use the negative conditional return skewness to measure crash risk.

[10] Some prior studies focus on explicit incidents related to misinformation. For example, Ahern and Sosyura (2015) and Schmidt (2020) analyze the effects of confirmed merger rumors. Clarke et al. (2020) and Kogan et al. (2023) employ an event study approach to investigate the fake news exposed by the Securities and Exchange Commission (SEC) investigations. Several theoretical papers, such as Andrei and Cujean (2017) and Pedersen (2022), have explored the impact of misinformation on pricing mechanisms in financial markets.

By employing advanced LLMs and an information network approach, our paper provides an empirical framework to assess the quality of information contained in textual materials. Relatedly, while previous textual analyses typically focus on document-level characteristics like sentiment (e.g., Chen et al., 2022, 2023; Fan et al., 2024), a few studies have shifted to extract and analyze specific topics from text materials (e.g., Bybee et al., 2023; Giglio et al., 2022; Li et al., 2023; Ross et al., 2024). We provide a methodology for multi-label topic classification and entity extraction that can be used for future research of textual analysis.

Our research is also related to the growing literature on machine learning and big data in finance (e.g., Kleinberg et al., 2018; Erel et al., 2021; Easley et al., 2021; Lyonnet and Stern, 2022; Kaniel et al., 2023; Chen et al., 2024; Van Binsbergen et al., 2024; Fan et al., 2024). For example, several papers demonstrate the power of ML, generative AI, and LLMs in information collection, processing, and extraction (e.g., Gentzkow et al., 2019; Giglio et al., 2022; Beckmann et al., 2024). Our paper integrates the strengths of these technologies to tackle the task of measuring misinformation, leveraging the understanding capabilities of generative AI for information labeling, the efficiency and speed of LLMs for information embedding, and the cost-effectiveness of few-shot frameworks for text classification and extraction.

We note that there is a strand of literature in computer science dedicated to misinformation detection (e.g., Shu et al., 2017; Wu et al., 2019; Zhou and Zafarani, 2020). However, these methods significantly diverge from our approach. They often rely on algorithms trained on specific datasets, where texts are pre-labeled as true or false, then assign probabilities of falsehood to individual texts based on these training datasets. Such techniques are not well-suited for the finance context, characterized by voluminous and subtly false information and the lack of specialized labels for training.

The rest of the paper proceeds as follows. Section 2 presents the methodology for measuring firm-level misinformation. Section 3 describes the data. Section 4 validates the misinformation measure. Section 5 examines the drivers of firm-level misinformation. Section 6 studies the implications of misinformation on investor attention, trading volume, stock

returns, and risks. Section 7 concludes. Technical details, additional analyses, and robustness checks are dedicated to the appendices.

2 Methodology for Measuring Misinformation

In this section, we develop the Misinformation Measurement via Truth Approximation (MMTA) framework. We overview the framework in Section 2.1, with details explained in Section 2.2.

2.1 Overview of the MMTA framework

Our framework for measuring misinformation is inspired by two principles in discerning truth: *information consistency* and *the wisdom of select crowds*. Rational individuals often gauge the truthfulness of information by comparing it against related knowledge and seeking verification through multiple channels. The greater the consistency of the information, the higher the likelihood of its truthfulness (Ji et al., 2023).[11] Furthermore, when different sources provide conflicting information, a rational actor will assign greater weights to information considered more reliable, prioritizing the wisdom of select crowds (Mannes et al., 2014; Xu et al., 2021a).[12]

Based on these two principles, we utilize advanced machine learning (ML) and artificial intelligence (AI) technologies to quantify firm-level misinformation on a monthly frequency from hundreds of millions of pieces of textual materials. In doing so, we tackle two primary challenges: first, converting and categorizing textual materials across diverse sources and

[11]This information consistency principle aligns with research on addressing AI hallucinations in natural language generation, where consistency is a key indicator of the authenticity of generated outputs. In the context of generative AI, "hallucination" refers to the phenomenon where the content generated is meaningless, unreal, or incorrect—a form of misinformation.

[12]This principle also aligns with the "truth discovery" theory in the field of information science, which seeks to ascertain the veracity of information from conflicting sources. Related surveys of truth discovery theory include Li et al. (2016); Xu et al. (2021a), and classic and recent works in this domain include Surowiecki (2005); Li et al. (2014); Ye et al. (2020); Burns et al. (2022).

formatting into comparable information, and second, extracting the "truth" from comparable information.

The first critical step involves systematically transforming noisy, unstructured textual materials into structured, comparable information. For instance, consider the following three segments regarding the same company:

- *The company emphasizes employee training and development, aiming to enhance overall skills and productivity.*
- *Our main product line has achieved significant growth.*
- *The company's marketing efforts have received widespread recognition.*

It is evident that Segment 1 is about human capital, and Segment 2 discusses business operations, precluding direct comparison. We thus assign an overarching topic label, denoted by o, to each segment (e.g., for Segment 1, o="human capital"; for Segment 2, o="business operations"). Although Segments 2 and 3 both fall into the broad topic of business operations, they focus on different specific entities and, therefore, cannot be directly compared. Here, an "entity" refers to the specific subject described within the text segment (denoted by e). For Segment 2, the entity is *product line*, and for Segment 3, it is *marketing*. Consequently, each segment in this example can be organized into a bivariate tuple (o, e): Segment 1 as (human capital, employees), Segment 2 as (business operations, product line), and Segment 3 as (business operations, marketing). It is evident that *only information with identical topic-entity labels of the same company can be meaningfully compared.*

Specifically, leveraging the recent development of LLMs, we develop a two-tier topic classification (TC) and entity extraction (EE) procedure to transform unstructured texts into comparable information sets. The TC procedure identifies the topics of the text, while the EE procedure captures entities underlying the identified topics. It is worth emphasizing that one piece of textual material (e.g., a quarterly report) can contain multiple topics, and

our algorithm can comprehensively identify all relevant topics and entities. After this step, unstructured textual information can be categorized by the topic-entity labels.

The second critical step involves extracting the "truth" from each comparable information set, i.e., information of the same company with identical topic-entity tuple (o, e). To achieve this, we first convert unstructured data into fixed-length embedding vectors using the ERNIE model (Enhanced Representation through Knowledge Integration), a BERT variant optimized by Baidu (the Chinese Google) for processing Chinese texts (Sun et al., 2020). In this way, each piece of information is represented as a tuple $(o, e, \mathbf{x})$, where o denotes the overarching topic, e identifies the specific entity, and $\mathbf{x}$ is the vector of semantic features that we use to quantify the meaning of the context.

Next, we conceptualize the extraction of truth as an optimization problem with two stages. In the first stage, we determine the appropriate "reliability" weight for each piece of information. This strategy acknowledges that not all information contributes equally to the understanding of the truth. We assign higher weights to information from more trustworthy sources and information with higher logical coherence in sentence constructions. In the second stage, we extract the optimal representation of truth within each comparable information set (denoted as $\mathbf{x}^*$). The optimization is set up to account for the reliability and quantity of information, therefore reflecting the principles of information consistency and the wisdom of select crowds.

With this, the falsity degree for a piece of information is calculated by the divergence between its semantic feature vector ($\mathbf{x}$) and the corresponding truth representation ($\mathbf{x}^*$) within its comparable information set. Then, we quantify the degree of misinformation at the firm level through a bottom-up integration process across all information sets regarding a firm.

2.2 MMTA: A Framework for Measuring Misinformation

In this section, we detail the MMTA framework, encompassing four key steps: (a) organizing raw textual materials into comparable topic-entity clusters; (b) converting narrative texts into structured, quantitative vectors to quantify semantic insights; (c) building information networks to identify the underlying "truth" in each cluster; and (d) measuring misinformation as deviations from the identified "truth" and aggregating these deviations to the firm level. We dedicate certain technical details to the appendix.

2.2.1 From clutter to clusters: categorizing raw text into comparable information set

As explained in Section 2.1, one primary challenge is making information across diverse sources and formatting comparable. To address this issue, we employ a two-tier strategy that effectively categorizes information, ensuring that each piece of information is assigned a specific topic-entity tuple (o, e). That is, each piece of information is given both a coarse-grained topic label (o) and a fine-grained entity label (e).

This process includes two steps: topic classification and entity extraction. In the first step, we perform topic classification.[13] Drawing on recent studies of firm value (e.g., Belo et al., 2022) and asset pricing (e.g., Feng et al., 2020; Leippold et al., 2022), we focus on eight key topics that are closely tied to both the short-term and long-term firm values: business operations (BO), financial performance (FP), risk management (RM), corporate governance (CG), environmental responsibility (ER), social responsibility (SR), human capital (HC), and R&D innovation (RD).

We use LLMs for topic classification following the approach outlined by Li et al. (2023).[14]

[13]Topic classification is a specific application of text classification, which assigns documents to one of a predefined set of labels—a classic task in natural language processing (NLP).

[14]Traditional NLP methods, such as lexicon-based approaches and Latent Dirichlet Allocation, often fail to effectively capture deeper semantic insights, especially in the financial sector (Fan et al., 2024; Ross et al., 2024). The increasing popularity of LLMs such as BERT and ERNIE in textual analysis is due to their enhanced ability to comprehend textual nuances (Chen et al., 2022, 2023).

Specifically, we utilize the ERNIE model, a BERT variant developed by Baidu for the Chinese language, to encode texts into quantifiable vectors. This process is commonly referred to as (contextualized) embeddings.[15] Given that financial texts are typically complex in content and cover multiple topics simultaneously, instead of directly fine-tuning the embeddings, we use them as inputs for the Unified Semantic Matching (USM) model to enhance the accuracy of topic classification. The USM model, primarily developed by Baidu, is an open-source and highly effective algorithm designed for diverse text-based semantic classification and information extraction.[16] In a nutshell, this model constructs a shared semantic space for user-defined labels and texts through Directed-Token-Linking (DTL) and then performs semantic matching. Although our implementation of multi-label topic classification is also a contribution to financial textual analysis, we leave the details in Appendix A.

After the initial step of topic classification, the second step aims to extract entities from texts within each topic cluster. Traditional entity extraction methods, typically dependent on manual fixed rules or lexicons, struggle to cope with the diverse and context-rich nature of financial texts (Etzioni et al., 2008; Li et al., 2020). Thus, we utilize LLMs to achieve our goal. In particular, we further fine-tune the USM model using the embeddings from the LLM (the ERNIE model) to extract entities. This refined algorithm allows us to unsupervisely identify entities by analyzing the language sequences, rather than relying on predefined entities.[17] The details of the entity extraction procedure are also in Appendix A.

On average, there are 26.34 entities identified per month for a given company under a given topic. Examples of extracted entities are shown in Figure 2, which presents the high-frequency entity word cloud for the eight pre-determined topics based on the textual

[15]Embeddings are dense vector representations of text that capture the complex semantic relationships inherent in texts.

[16]When we started our project in February 2023, the USM model ranked first in the FewCLUE (Few-Shot Chinese Language Understanding Evaluation), a benchmark for assessing Chinese language comprehension.

[17]To address variations in language styles and terminologies across diverse information sources, we also utilize a text clustering algorithm, the hierarchical clustering (Hastie et al., 2009), which is adept at organizing the semantic relationships between words. This method allows us to amalgamate terms like "profit" and "earnings"—linguistically distinct yet semantically analogous—into a unified entity for nuanced analysis.

(a) Financial Performance

(b) Risk Management

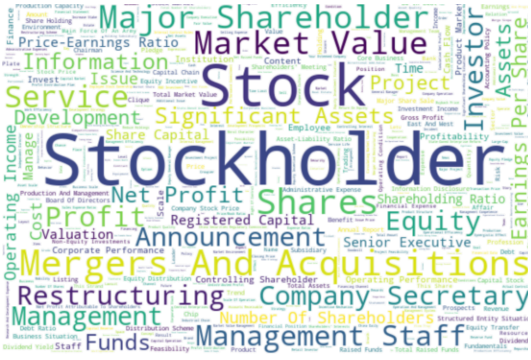
(c) Environmental Responsibility

(d) Human Capital

(e) R&D Innovation

(f) Business Operations

(g) Social Responsibility



(h) Corporate Governance

Figure 2 Examples of Extracted Topics and Entities

Notes: This figure displays examples of entities extracted from the 2018 data in the FT corpus. It presents eight word clouds, each representing the high-frequency entities associated with a specific topic. The size of each entity word within these clouds reflects its frequency.

materials in 2018.[18] The size of each word in these clouds corresponds to its frequency. The figure clearly demonstrates the effectiveness of our entity extraction algorithm. For example, under the topic of "financial performance," entities like "operating income," "net profit," and "earnings per share" emerge prominently. Under the topic of "human capital," entities like "talents" and "employee" are highly frequent. To further validate our results, we follow the approach of Sautner et al. (2023) and invite ten doctoral students in finance to collaboratively assess the reasonableness of the extracted entities.

2.2.2 Dissecting the narrative: quantifying semantic insights

Transforming complex narratives into quantifiable features is a significant challenge in textual analysis (Fan et al., 2024). Traditional models like TF-IDF and lexicons, while useful for providing keyword insights, often miss the intricate relationships of words and overlook deeper semantic connections (Kogan and Meursault, 2021). In contrast, LLMs that rely on the Transformer architecture have been proven to effectively encode textual data into vectors (embeddings), capturing rich semantic features (Chen et al., 2022; Li et al., 2023).

To ensure a comprehensive understanding of textual information, we extract semantic features based on ERNIE, the BERT variant optimized for processing Chinese texts. Similar to BERT, ERNIE translates the composite meaning of the entire input sequence into a 768-dimensional numeric vector associated with the classification (CLS) token. Following Kim and Nikolaev (2023), we use the embeddings associated with the CLS tokens as our semantic feature vector.[19]

[18]We employ the word cloud visualization technique to display these results, utilizing the wordcloud library in Python.

[19]A well-known issue with such LLMs is the tendency of their vector representations to converge into a narrow semantic space, leading to a "collapse" where distinct narratives appear misleadingly similar (Su et al., 2021). To address this issue, we utilize the ERNIE model optimized within the CoSENT framework, which employs a ranking loss mechanism to enhance the distinctiveness of text representations. This optimization does not alter the architecture of the model; instead, it refines embeddings to ensure that they emphasize subtle disparities and contrasts in narratives, thus leading to a more accurate analysis. For details, see https://github.com/bojone/CoSENT.

2.2.3 Finding the truth: building information networks

After structuring textual materials into clusters based on topics and entities, complete with vector representations, our next goal is to unearth underlying "truth" from the vectors of semantic features. We take an unsupervised approach with three steps: (*i*) constructing a network to quantify the similarity of information within a given comparable information set, (*ii*) measuring information reliability through assessing source credibility and content coherence, and (*iii*) extracting truth from each of these information networks, balancing reliability and quantity of information.

Constructing information network For each topic-entity pair (o, e) of a given company, we construct an undirected information network, $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, for the corresponding information set (articles) I associated with (o, e). Here, $\mathcal{V}$ represents the nodes, each of which corresponds to an individual piece of information in information set I. $\mathcal{E}$ denotes the set of edges. For example, the edge between information i and j is established based on the cosine similarity between the vector semantic representation $\mathbf{x}_i$ and $\mathbf{x}_j$ of information i and j:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \times \|\mathbf{x}_j\|_2}. \tag{1}$$

Cosine similarity is extensively used in text analysis to measure the closeness of textual content in high-dimensional spaces (Chandrasekaran and Mago, 2021; Fan et al., 2023). We also define the edge weight matrix $\mathbf{B} = (b_{ij})_{|I| \times |I|}$ by

$$b_{ij} = 1 + s(\mathbf{x_i}, \mathbf{x}_j) = 1 + \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \times \|\mathbf{x}_j\|_2}. \tag{2}$$

Consequently, more positively related information carries higher weight on their edge.

Measuring reliability of information We measure the reliability of each piece of information based on the credibility of the information source and the coherence of the information

content. In doing so, we can prioritize more credible information.

Information sources vary significantly in credibility. For example, retail investors and analysts possess different levels of knowledge and insights, while official media sources are typically more reliable than unofficial ones. We quantify this aspect by assigning a credibility weight to each information source (e.g., a particular news agency).[20]

Recognizing the dynamic nature of credibility, we introduce a penalty mechanism to adjust the weight of each information source based on its propensity to disseminate misinformation. Specifically, a source's credibility weight is recalibrated at the end of each month, contingent on its average misinformation measure in the preceding month. This is done using an exponential decay formula for adjustment:

$$c_{s,t} = c_{s,t-1} \times e^{-\alpha MISI_{s,t-1}}. \tag{3}$$

Here, $c_{s,t}$ represents the credibility weight of information source s in month t. $MISI_{s,t-1}$ denotes the average misinformation measure calculated from all articles disseminated by source s in month $t-1$; a higher value indicates greater misinformation.[21] The parameter α is a positive constant determining the rate of decay due to misinformation, set here at 0.1. To initialize source credibility, we use ChatGPT-4 to obtain initial source weights, following the approach of Yang and Menczer (2023).[22] It is worth noting that our results are robust without taking this dynamic adjustment approach.

Additionally, we assess the content reliability by evaluating the logical coherence of sentence constructions, inferred through their formation probability. This step helps mitigate the impact of texts that are grammatically incorrect or do not conform to common sense. To

[20]The information sources include firm announcements, 13 different news outlets, analyst reports, and social media posts. Detailed descriptions are provided in Section 3.

[21]See Section 2.2.4 for the exact calculation of *MISI*.

[22]Specifically, we use ChatGPT-4 to rate the credibility of each source. Take the social media platform, GUBA, as an example; we use the following prompt:"Rate the website's credibility: rate GUBA on a scale between 0 and 1, where 0 means very low credibility and 1 means very high credibility. The assistant returns the rating -1 when the assistant has no knowledge of the website, otherwise the assistant should provide the best estimation."

achieve this, we use the generative pre-trained GPT model of Radford et al. (2019).[23] The rationale behind using this model lies in its proven efficacy in understanding and generating language, making it well suited for assessing sentence coherence. The output of this algorithm is a score of content coherence (denoted by l).

We then integrate the content coherence and the source credibility to assign an initial reliability weight for each piece of information, assuming equal importance for both factors:

$$w_i = c_{s,t} + l_i. \tag{4}$$

Here, w_i represents the reliability weight of information i at time t, which is from source s. For notational simplicity, we omit subscripts t and s in the weight w_i. $c_{s,t}$ represents the credibility of source s at time t, calculated based on equation (3), and l_i is the content coherence score for information i. Both $c_{s,t}$ and l_i are numerical values ranging from 0 to 1, where higher values indicate greater reliability.

To assign a final reliability weight to each piece of information, we perform one extra step of refinement using the constructed information network. This refinement step balances two goals: first, the refined weights align closely with the initial reliability weights defined in equation (4); second, assigning similar weights to information that exhibits similar semantic proximity within the network, so that information conveying similar meanings shares similar reliability assessments. The objective of this refinement is to mitigate the influence of source-specific biases, a problem highlighted in Qin et al. (2018) and Fan et al. (2023).

Specifically, following Garza and Schaeffer (2019), this refinement step is formalized via

[23]It is important to note that the GPT model mentioned here is distinct from the ChatGPT-4 referred to earlier. ChatGPT-4, based on a large-scale GPT model, is a fine-tuned interactive application product, conceptualized as an AI tool, whereas the underlying GPT model is a pre-trained algorithmic model based on the Transformer architecture.

the following the optimization problem:

$$\min_{\widetilde{\mathbf{w}}} \mathcal{J}(\widetilde{\mathbf{w}}) = \min_{\widetilde{\mathbf{w}}} \frac{\eta}{2}||\widetilde{\mathbf{w}} - \mathbf{w}||_{\mathcal{F}}^2 + \frac{1}{2}\left(\sum_{i,j=1} b_{ij} \left|\left| \frac{1}{\sqrt{f_i}}\widetilde{w}_i - \frac{1}{\sqrt{f_j}}\widetilde{w}_j \right|\right|^2\right). \tag{5}$$

Here, $\mathbf{w}$ denotes the vector of initial reliability weights (defined in equation (4)) for all information in set I associated with topic-entity (o, e), and $\widetilde{\mathbf{w}}$ represents the refined reliability weights. The Frobenius norm $\|\cdot\|_{\mathcal{F}}$ measures the distance between the refined and initial weights, and $f_i = \sum_{k=1}^{|I|} b_{ik}$ with the edge weight, b_{ik}, defined in equation (2). The regularization parameter η balances the trade-off between maintaining the proximity of initial and updated weights and the semantic similarity among the texts. To get the idea, suppose that information i and information j have identical semantic representation ($\mathbf{x}_i$=$\mathbf{x}_j$). Then, $f_i = f_j$ and b_{ij} achieves the maximum value, so that the optimization process tends to assign more similar values to $\widetilde{w}_i$ and $\widetilde{w}_j$.

While we take the steps to dynamically adjust the credibility scores of information sources and refine reliability weights to mitigate the influence of inherent source-specific biases, we also verify that our results in the following sections are quantitatively robust, albeit slightly weaker, under a simplified procedure without these steps.

Extracting the truth Next, we turn to extract "truth" within the information set associated with a given topic-entity pair of a company. Our method transforms this task into an optimization problem designed to account for the quantity and reliability of the information, reflecting the principles of information consistency and the wisdom of select crowds.

Consider, for example, the scenario involving evaluating a company's financial performance. If all information sources uniformly suggest high performance, this consensus strongly indicates its veracity. However, contrasting scenarios, such as an authoritative source claiming "high performance" while numerous social media discussions suggest "low performance," present a more complex picture. In such a case, despite the generally lower credibility of

social media, the overwhelming volume of these discussions could suggest the truth is more aligned with "low performance."

Our goal is to balance information quality and volume. In particular, we calculate the "truth" $\mathbf{x}^*$ by

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i \in I} \widetilde{w}_i \cdot d(\mathbf{x}_i, \mathbf{x}) = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i \in I} \widetilde{w}_i \cdot (1 - s(\mathbf{x}_i, \mathbf{x})), \tag{6}$$

where $\widetilde{w}_i$ represents the refined reliability weight for information i in equation (5), and $d(\mathbf{x}_i, \mathbf{x}) = 1 - s(\mathbf{x}_i, \mathbf{x})$ measures the semantic difference between information $\mathbf{x}_i$ and $\mathbf{x}$. The vector $\mathbf{x}^*$ symbolizes our approximation of the truth, which is extracted from the consensus weighted by information reliability. As one can see, more reliable information carries higher weights in the optimization process. This results in an information reliability-weighted consensus (6).

2.2.4 Measuring misinformation: aggregating deviations from the truth

In the last step, we calculate firm-level misinformation using a bottom-up approach. To this end, we first calculate the misinformation measure for each piece of information i as the deviation from the identified "truth" within its information set:

$$MISI_i^{piece} = d(\mathbf{x}_i, \mathbf{x}^*) = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}^*}{\|\mathbf{x}_i\|_2 \times \|\mathbf{x}^*\|_2}.$$

Here, $\mathbf{x}_i$ and $\mathbf{x}^*$ represent the vector of semantic features for information i and the corresponding "truth," respectively.

With this, the misinformation measure at the topic level is calculated as

$$MISI^{topic} = \sum_{e \in E^o} MISI_e^{entity} = \sum_{e \in E^o} \frac{1}{|\mathcal{I}^e|} \sum_{i \in \mathcal{I}^e} MISI_i^{piece},$$

where E^o is the set of entities associated with topic o, and $|\mathcal{I}^e|$ is the number of pieces of

information associated with entity e. It is an aggregated deviation for a topic. Then, the misinformation measure at the firm level is calculated as

$$MISI^{firm} = \frac{1}{\sum_{o\in\mathcal{O}} |\mathcal{I}^o|} \sum_{o\in\mathcal{O}} |\mathcal{I}^o| \cdot MISI_o^{topic}.$$

Here, $\mathcal{O}$ is the set of the eight pre-determined topics relevant to the firm, and $|\mathcal{I}^o|$ is the total number of pieces of information for topic o, which is used to weigh the contribution of each topic to the firm-level *MISI*.[24]

3 Data and the Misinformation Measure

In this section, we detail the massive dataset that we use to measure firm-level misinformation. We also provide descriptive statistics of the misinformation measure.

3.1 Dataset

We construct a comprehensive firm-level textual corpus (FT corpus) spanning the period from January 1, 2015, to December 31, 2022. This corpus includes multiple datasets, encompassing firm disclosures, analyst reports, news reports, and social media posts, further categorized by their sources:

- **Firm Disclosures:** This dataset consists of firm announcements, annual and quarterly reports, and investor Q&A sessions, collected from the China Stock Market and Accounting Research (CSMAR) database and the WIND database. It covers 5,000 firms with a total of 1,071,464 records.

- **News Reports:** We gather over 3,327,580 news reports related to more than 4,500 firms from the Huike database and CSMAR database. These news reports are from 13

[24]In untabulated exercises, we find that our results are also robust if we take a simple average across topics.

major media sources, which we classify as official, semi-official, and unofficial, following An et al. (2022) and Hong et al. (2023).[25]

- **Analyst Reports:** This collection comprises 568,468 analyst reports covering roughly 4,500 firms. The data are sourced from the WIND database, the CSMAR database, and manually collected from the Luobo Investment Research website.

- **Social Media:** The dataset includes 285,080,647 posts from GUBA, which is China's largest retail investor stock market forum.

For notational simplicity, we refer to all firm disclosures, news reports, analyst reports, and social media posts as articles. In preparing the FT corpus, we exclude articles without a verifiable publication date and articles that do not reference a specific company. Consistent with Sautner et al. (2023), we also exclude firm-month observations with articles totaling fewer than 100 in that month to ensure data reliability. Our final sample consists of 254,826,060 articles, pertaining to 3,994 publicly-traded companies. As we highlighted in Section 2, our algorithm can comprehensively identify all relevant topics within a single article if the article (e.g., a quarterly report) contains multiple topics.

Panel A of Table 1 presents descriptive statistics for the total number of articles per firm and their distribution from various sources. The first row represents the overall article volume at the firm level, with an average of 62,792 articles per firm. This figure ranges dramatically, from a minimum of 326 to a maximum of 302,285. The second through fifth rows detail the volume from four different sources. Firm disclosures and analyst reports represent more specialized information sources, with relatively low average article counts of of 235.86 and 50.49, respectively. Social media dominates in volume, with an average of 61,746.82 articles and a high degree of variability.

[25]The 13 media sources of news reports include two governmental newspapers: People' s Daily and Xinhua News Agency; five mainstream economic and financial newspapers: 21st Century Business Herald, China Business News, Economic Observer, China Securities Journal, and Securities Times; three online financial platforms: Sina Finance, NetEase Finance, and Phoenix Finance; and three financial information websites: Financial Community, Hexun, and WinShang.

[Insert Table 1 here]

In Appendix B, we establish the validity of our FT corpus through Information Gain Analysis (Ash, 1990; Fedyk and Hodson, 2023).[26] This analysis quantifies the incremental information each new data source brings to our corpus, and it demonstrates the stability and comprehensiveness of our FT corpus.

We also collect a comprehensive dataset of firm characteristics from the CSMAR database. The firm-specific characteristics include the logarithm of total assets (Size), fixed asset ratio (Tangibility), current debt to asset ratio (Debt), revenue growth rate (RevGrowth), financial volatility risk (FinRisk), financing constraint risk (FinConstraint), the proportion of independent directors (IndBoard), supervisory board size (SupBoard), the proportion of shares held by institutional investors (Institution%), and stock ownership concentration (ShareConc). Descriptive statistics for these variables are presented in Panel B of Table 1, with detailed calculations explained in Appendix Table C.1.

3.2 A first look at the misinformation measure

Utilizing the FT corpus and the methodology proposed in Section 2, we compute the degree of misinformation (*MISI*) for each firm on a monthly basis.

Panel A of Table 2 presents summary statistics of the firm-level and firm-topic-level *MISI*. The average firm-level *MISI* is 0.14 with a standard deviation of 0.11. Notably, misinformation is relatively high in topics such as financial performance (FP), corporate governance (CG), and business operations (OB). This aligns with our expectations, as these areas are critical to a company's core operations, attract significant market attention, and are frequent targets for fraudulent activities (Dimmock et al., 2018; Clarke et al., 2020). Conversely, the average level of misinformation is relatively low for environmental responsibility (ER), human capital (HC), and social responsibility (SR), where even the 75th percentile

[26] It is important to note that, while our corpus currently encompasses six text-based data groups, the methodology is adaptable, allowing for integration with multi-modal data.

is zero. There is also substantial variation in misinformation metrics at both the firm and firm-topic levels.

[Insert Table 2 here]

Panel B of Table 2 also shows the correlations of *MISI* across various topics. It is clear that firm-level *MISI* predominantly relates to misinformation in the areas of FP, CG, and OB, with a correlation of 0.61, 0.67, and 0.56, respectively. This suggests that firm-level misinformation is mostly driven by misinformation about financial performance, corporate governance, and business operations.

Panel C of Table 2 further shows that the misinformation measure is quite persistent. Firm-level misinformation has an autoregressive coefficient of 0.53. Among the eight topics, misinformation related to corporate governance is the most persistent, followed by misinformation on risk management and business operations. On the contrary, misinformation pertaining to social responsibility is the least persistent.

4 Validating the Misinformation Measure

In this section, we conduct several exercises to validate the misinformation measure. We find that *MISI* negatively correlates with the quality of corporate information disclosure and spikes when the disclosed misinformation-driven events occur. In addition, we document an uptick in public skepticism about a firm when misinformation of the firm increases. All these results support the effectiveness of our misinformation measure.

4.1 Misinformation and disclosure quality

Timely and accurate disclosure of corporate information is crucial in mitigating information asymmetry and enhancing the overall informational environment (Cookson and Niessner, 2020; Huang et al., 2021). Companies with high-quality disclosure practices can

curb the spread and impact of misleading content. Therefore, if our *MISI* is an effective measure of misinformation, firms with high disclosure quality should exhibit lower *MISI*.

To measure firm disclosure quality, we utilize the quality grades of corporate information disclosure generated and released annually by the Shanghai Stock Exchange and Shenzhen Stock Exchange.[27] These grades are labeled A, B, C, and D, corresponding to excellent, good, fair, and poor, respectively. We define an indicator variable *IDQ* that equals one if a firm scores an "excellent" grade of A or a "good" grade of B, and it equals zero otherwise. We run panel regressions to investigate the relationship between *IDQ* and *MISI*. Since the disclosure grades are updated annually, we aggregate the monthly *MISI* into annual metrics to align with these updates. Specifically, for each year, we calculate the average, median, and maximum *MISI*, each serving as dependent variables separately in our regression analyses.

The results, as shown in Table 3, consistently reveal a significant negative correlation between firm-level disclosure quality and measures of misinformation across all model specifications. Columns (1) to (3) detail a significant negative relation between misinformation and the average monthly *MISI*. Columns (4) to (6) and Columns (7) to (9) show similar results when the dependent variables are changed to the median *MISI* and the maximum *MISI* for a given year, respectively. These findings remain robust after accounting for firm fixed effects, year fixed effects, and industry fixed effects.

[Insert Table 3 here]

For robustness, we also employ a market-based measure of disclosure quality proposed by Kim and Verrecchia (2001), which can be calculated using stock prices and trading volumes. The results using this alternative measure are presented in Table C.2 in Appendix C, which further substantiates the validity of our misinformation measure.

[27] We collect this data from the CSMAR database.

4.2 Validation using disclosed misinformation-related events

In this section, we further validate our misinformation measure by analyzing the occurrence of recorded misinformation-related events, which are disclosed by either regulators or firms. We find that the misinformation measure indeed registers high values during these events.

To see this, we construct a dataset of disclosed company events associated with misinformation. Our event data is sourced from two primary channels: company announcements for rumor clarification from the CSMAR dataset and penalty announcements issued by the China Securities Regulatory Commission (CSRC). For both sources, we apply a keyword filter for terms associated with misinformation, such as "falsified records," "major discrepancies," "misleading statements," and "inconsistencies with facts." Only events with a clearly defined occurrence month are retained, and we also manually check these events to ensure accuracy. We are able to identify 1,340 events associated with 539 firms that are covered in our FT corpus.[28]

We begin our analysis by counting the occurrence of the disclosed misinformation events. We find that 89% of the events occur when a firm's *MISI* is above the historical average level of the firm's misinformation measure, and more than 70% of the events take place when a firm's *MISI* is above the 90th percentile of the firm's historical *MISI* score. We further categorize firms into three equal groups each month based on their *MISI* measure and count the number of misinformation events for each group. Figure 3 shows the number of events for each group from January 2015 to December 2022. It is evident that groups with higher *MISI* consistently have more disclosed misinformation events.

To control for other confounding factors, we also estimate a panel regression where we

[28]We note that firms without recorded events do not necessarily indicate that they are free of misinformation-related events. One reason is that while constructing this dataset, we include only events with clear occurrence dates, leaving out those with ambiguous timings. Moreover, the events currently disclosed may represent only a part of the sample, with many misinformation-related incidents remaining undetected or officially unrevealed (Fan et al., 2023).

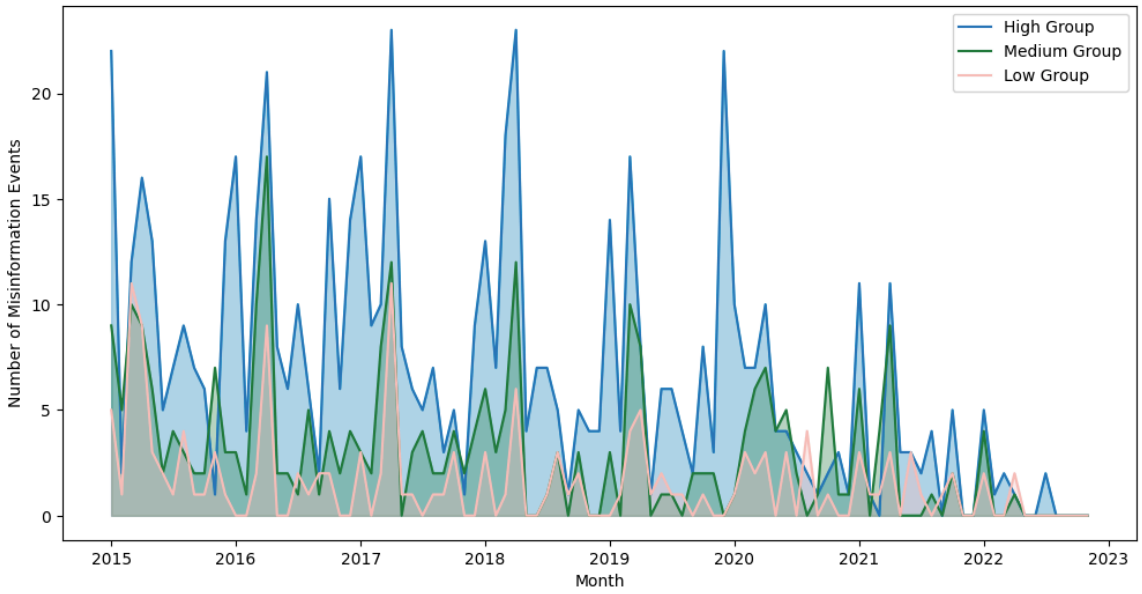


Figure 3 Monthly Misinformation Events by Group

Notes: This figure shows the event coverage across firms with different misinformation measures. Firms are categorized into three equal groups each month according to their *MISI*. The number of misinformation events indicates the occurrence of the disclosed misinformation incidents within each group.

regress an indicator for misinformation-related events on our misinformation measure:

$$\begin{aligned} Event_{i,t} = \alpha_0 + \alpha_1 MISI_{i,t} + \alpha_2 Firm\ Characteristics \\ + Firm\ FE + Month\ FE + Ind \times Year\ FE + \epsilon_{i,t}. \end{aligned} \tag{7}$$

Here, $Event_{i,t}$ indicates whether a firm i has a disclosed misinformation-related event in month t. *Firm Characteristics* is a comprehensive set of control variables related to governance, balance sheet, and financial risk attributes that may influence misinformation events.

Table 4 presents the regression results, where columns (1) to (4) progressively include firm fixed effects, month fixed effects, industry-by-year fixed effects, and firm characteristics. The analysis confirms that *MISI* is highly correlated with the occurrence of misinformation events, even after controlling for a long list of firm characteristics.

[Insert Table 4 here]

For robustness, we also employ a logit regression model for the analysis, shown in Columns (5) to (8) of Table 4. For example, Column (8) displays an estimated coefficient of

6.37 for *MISI*, which is highly significant, indicating that each unit increase in *MISI* increases the logarithm odds of the event by approximately 6.37. Converting this into an odds ratio, we find that a one standard deviation increase in *MISI* corresponds to an odds ratio of 2.02 ($e^{6.37\times0.11} \approx 2.02$), which is also economically significant.

4.3 Validation based on public response

Drawing on the insights of Dyck et al. (2008) and Joe et al. (2009), we posit that if our misinformation measure is effective, an increase in *MISI* should correspond with heightened public skepticism about a firm's authenticity. In other words, a significant amount of public discussions questioning a firm's veracity strongly indicate the presence of misinformation.

To validate our measure from this perspective, we construct a lexicon focused on terms related to the concept of falsehood.[29] We use keyword matching against this lexicon across the FT corpus (excluding firm announcements and firm quarterly and annual reports) to measure public skepticism:

$$PS_{i,t} = \frac{1}{|\mathcal{I}_{i,t}|} \sum_{p \in \mathcal{I}_{i,t}} \delta(p, \mathcal{B}). \tag{8}$$

Here, $\mathcal{I}_{i,t}$ is the information set for firm i at time t, and $|\mathcal{I}_{i,t}|$ represents the number of pieces of information. $\mathcal{B}$ represents the "falsehood" lexicon set, and $\delta(p, \mathcal{B})$ is a binary function defined as:

$$\delta(p, \mathcal{B}) = \begin{cases} 1 & \text{if any word in } p \text{ matches any keyword in } \mathcal{B}, \\ 0 & \text{if otherwise.} \end{cases} \tag{9}$$

[29]The lexicon, developed with the assistance of ChatGPT-4 and further refined through manual review, primarily captures expressions that convey doubt and skepticism about factual accuracy. The prompt used is, "I need a lexicon of core words that the public uses to express doubt and skepticism about the accuracy of information; please provide relevant key terms." Key words in the lexicon include "swindler," "deception," "cheat," "fraudulent," "sham," "hoax," "untruth," "forgery," "fabrication," "mislead," "impersonate," "false report," "rumor," "misconvey," "spread falsehoods," "untruthful," "fake news," "exaggerate," "overstate," "deceive," "conceal," "misunderstand," "counterfeit," "misreport," "slander," "defame," "misguide," "misconception," "forge," "pretend," "false advertising," and "fake reviews."

In addition to the ratio, we also use $\log(\sum_{p\in\mathcal{I}_{i,t}} \delta(p,\mathcal{B}))$ as an alternative measure of public skepticism.

Table 5 reports that higher levels of misinformation about a firm significantly correlate with increased public skepticism about the firm, evidenced by more discussions questioning the firm's authenticity. As public skepticism is a natural response to the circulation of misinformation, this analysis further supports the validity of our misinformation measure.

[Insert Table 5 here]

In summary, based on the evidence from corporate information disclosure quality, the occurrence of disclosed misinformation-related events, and public skepticism about information authenticity, we find strong support for the effectiveness of the misinformation measure.

5 What Drives Misinformation?

In this section, we explore firm characteristics that may drive misinformation. We find that firms with weak balance sheets, poor governance structures, and firms from highly concentrated industries have higher average levels of misinformation. We also show that misinformation tends to increase around major corporate events.

Firm characteristics and misinformation We first examine whether and how firm characteristics are related to the degree of misinformation. Following Fan et al. (2023) and Li et al. (2024), we focus on firm characteristics related to balance sheet, financial risk, and corporate governance structure.

Specifically, we estimate the following regression specification at a quarterly frequency:

$$MISI_{i,q} = \alpha_0 + \alpha_1 Firm\ characteristics_{i,q} + FE + \epsilon_{i,q}. \tag{10}$$

Here, $Firm\ characteristics_{i,q}$ denotes a set of firm-specific characteristics in quarter q, while

$MISI_{i,q}$ represents the average monthly misinformation in quarter q. The regression is estimated based on quarterly data due to the availability of firm characteristics. We also include firm fixed effects, quarter fixed effects, and industry-by-year fixed effects to capture the industry-wide trends. Table 6 reports the results.

There are a few takeaways from this table. First, firms with weaker balance sheets generally exhibit a higher degree of misinformation. Specifically, firms with a higher debt-to-asset ratio and slower revenue growth tend to have higher *MISI*. In addition, firms with higher financial volatility risk and those facing more financing constraints also have higher levels of misinformation on average. Lastly, firms with weaker governance structures tend to have a higher degree of misinformation. This is reflected by the negative coefficients of the proportion of independent directors (IndBoard), the supervisory board size (SupBoard), and institutional ownership (Institution%).

[Insert Table 6 here]

Industry concentration and misinformation We further analyze the relationship between the degree of misinformation and industry concentration. Intuitively, more concentrated industries are more opaque and likely to be associated with higher levels of misinformation, while more competitive industries undergo more careful scrutiny and should observe lower levels of misinformation on average. In other words, as industries become more concentrated, overall information transparency is likely to diminish, creating an environment that incentivizes the dissemination of misinformation.

Following Leippold et al. (2022), we use the Guidelines for Industry Classification of Listed Companies issued by the China Securities Regulatory Commission (CSRC) in 2012, which identifies 90 industries. We calculate the average *MISI* for companies within each industry. At a glance, we find that heavy industries, such as "Non-metallic Mineral Mining" and "Oil and Natural Gas Extraction," and financial sectors, such as "Capital Market Services," "Currency Financial Services," and "Insurance Industry," exhibit higher levels

of misinformation. Indeed, these industries in China are dominated by a few state-owned enterprises.

To comprehensively test this hypothesis, we measure industry concentration using the Herfindahl-Hirschman Index (HHI) based on firm total assets, a common measure of market concentration. We then regress a firm's *MISI* on its industry HHI, with results detailed in Columns (5) and (6) of Table 6. The HHI coefficients consistently show significant positive values, indicating positive correlations between industry concentration and the level of misinformation.

Major corporate events and misinformation Having established the relationship between firm characteristics and misinformation, we now examine whether the occurrence of major corporate events can drive misinformation. We hypothesize that misinformation will increase around major corporate events due to the heightened attention these events attract from market participants and media. Indeed, during significant corporate events, there is likely to be a surge in information dissemination and speculation. This creates an environment ripe for the spread of misinformation, as stakeholders may attempt to influence market perceptions and reactions. Additionally, the complexity and uncertainty surrounding these events can lead to misunderstandings and the unintentional spread of inaccurate information.

Specifically, we gather data on corporate events from the CSMAR database, focusing on corporate financing operations, corporate splits, and mergers and acquisitions. To assess the impact of these corporate actions on the level of misinformation, we estimate the following regression model:

$$MISI_{i,t} = \alpha_0 + \alpha_1 Actions_{i,t} + \alpha_2 Controls + FE + \epsilon_{i,t}, \tag{11}$$

where $Actions_{i,t}$ denotes the number of events by firm i in month t. We also include the list of firm characteristics as controls. Table 7 depicts the results.

[Insert Table 7 here]

The results in Panel A of Table 7 indicate a significant positive relationship between the occurrence of major corporate events and the degree of misinformation, which is highly statistically significant. The findings are robust to various controls and fixed effects. In other words, corporate actions such as financing operations, corporate splits, and mergers and acquisitions trigger the spread of misinformation in the markets. In Panel B, we also expand the corporate events to include equity pledges, legal disputes, and changes in senior management. The results remain robust.

In summary, firms with weak balance sheets and poor governance structures and firms in less competitive industries tend to exhibit higher degrees of misinformation. Moreover, levels of misinformation increase significantly in the presence of major corporate events, calling for monitoring and managing misinformation during these periods.

6 Misinformation, Investor Behavior, and Stock Returns

The flow of information plays a critical role in shaping investor behavior and market dynamics. Consequently, misinformation is also expected to significantly impact investors' perceptions and reactions. In this section, we show that the circulation of misinformation significantly influences investor attention, trading volume, and stock returns. We also find that misinformation has strong explanatory and predictive power for stock return volatility and crash risk. Notably, all these effects are stronger among firms that are held more by retail and small investors, suggesting that they are more susceptible to misinformation relative to large and institutional investors.

Misinformation, investor attention, and trading volume We first explore the relationship between misinformation and investors' attention. Investors monitor information

that could impact stock prices, making them particularly sensitive to news that appears to present new opportunities or risks, even if such news is misleading. Furthermore, the possible uncertainty and ambiguity created by misinformation can compel investors to seek additional information.

We use each company's search volume on the search enging Baidu.com as a proxy for attention, following the idea of Da et al. (2011). Panel A of Table 8 reports the connection between *MISI* and investors' attention. It is clear that when a firm's misinformation level is higher, investors are more attracted to the company. In addition, misinformation can significantly predict investor attention levels over the next month. These findings indicate that the circulation of misinformation is capable of drawing increased attention. For robustness, we also use the search volume of investors on 13 major stock trading and stock broker apps to measure investors' attention. The results, shown in Table C.3 in Appendix C, lend further support to our conclusion.

[Insert Table 8 here]

We further examine whether higher levels of misinformation are also associated with higher trading volume. There are several reasons to expect this relationship. First, increased investor attention can translate into more trading activities. Additionally, higher levels of misinformation might prompt investors to react more frequently and aggressively in the market as they attempt to capitalize on perceived opportunities or mitigate potential risks.

In Panel B of Table 8, we regress contemporaneous and future trading volume on the misinformation measure, controlling for firm characteristics and the occurrence of corporate events as detailed in Table 7. Our findings indicate that higher levels of misinformation are indeed associated with increased trading activities in the markets.

It is worth noting that our results barely change after controlling for the common measures of disagreement. To see this, we employ the commonly-used indicators of disagreement: investor disagreement and analyst disagreement (Huang et al., 2021). Following Diether et al.

(2002) and Cookson and Niessner (2020), investor disagreement is calculated through the second moment of textual semantic stances on social media.[30] Following Jiang and Sun (2014) and Li and Li (2021), we use monthly forecasts of current-year earnings-per-share (EPS) by financial analysts to calculate analyst disagreement. One can see that our estimates change little when these additional controls are included.

Misinformation and stock returns As misinformation triggers investors' attention and trading activities, we expect a significant impulse of stock prices to misinformation, followed by reversals. For example, due to short-selling restrictions in the Chinese stock market, actors might disseminate positive misinformation to artificially inflate stock prices, which can lead to negative price reversals.

Specifically, we investigate the influence of misinformation on contemporaneous and future stock returns through regression models, where we also control investor disagreement and analyst disagreement. We use both panel regressions with fixed effects and the standard Fama-MacBeth procedure. Panels A and B of Table 9 present the impact of misinformation on contemporaneous and subsequent stock returns, respectively.

[Insert Table 9 here]

The regression analysis demonstrates that the degree of misinformation is significantly positively correlated with contemporaneous stock returns. The inclusion of controls and disagreement measures does not diminish the significance of misinformation. Shifting the focus to the return predictability of misinformation in Panel B of Table 9, *MISI* shows significant predictive power on future returns based on panel regressions with fixed effects and the Fama-MacBeth regressions. For example, column (3) of Panel B presents a coefficient of -0.0411 that is highly statistically significant. In terms of magnitude, a one standard deviation increase in misinformation predicts a negative return of 45 basis points (bps) over the next month.

[30]We use China's largest social media investing platform, GUBA, where positive posts are scored as 1, neutral posts as 0, and negative posts as -1, with data sourced from the CSMAR database.

To further test return predictability, we form a long-short portfolio by sorting firms into deciles based on their misinformation in the previous month. We find that the long-short portfolio delivers an average return spread of 84 bps over the next month, which is significant at the 99% confidence level.

Misinformation, return volatility, and stock crash risk When genuine information is obfuscated and misinformation about a firm circulates, the risk level of its stock is likely to rise as well (Hutton et al., 2009; Xu et al., 2021b). We find that this is indeed the case in the data.

Specifically, we measure return volatility based on daily returns in a given month. To measure firm-specific crash risk, we use the negative conditional return skewness (*NCSKEW*) following Kim et al. (2011).[31] We then estimate the relationship between misinformation, return volatility, and stock crash risk in the contemporaneous month and over the next month.

Panel A of Table 10 studies the contemporaneous risk measures, demonstrating that the degree of misinformation is significantly associated with stock return volatility and crash risk. Panel B further shows that misinformation significantly predicts return volatility and crash risk over the next month. Notably, after controlling for firm characteristics, investor and analyst disagreement, and various fixed effects, the level of misinformation retains statistically significant coefficients for both risk measures. These results underscore the strong

[31]To calculate *NCSKEW*, for each firm i on day d in month t, we estimate firm-specific daily returns, $D_{i,d}$, via the following regression:

$$r_{i,d} = \alpha + \beta_{1i} r_{m,d-2} + \beta_{2i} r_{m,d-1} + \beta_{3i} r_{m,d} + \beta_{4i} r_{m,d+1} + \beta_{5i} r_{m,d+2} + \epsilon_{i,d}, \tag{12}$$

where $r_{i,d}$ and $r_{m,d}$ represent the return of stock i and the market index on day d, respectively. The firm-specific daily return $D_{i,d}$ is calculated as $\ln(1+\epsilon_{i,d})$, where $\epsilon_{i,d}$ is the residual in equation (12). *NCSKEW* is then calculated as:

$$NCSKEW_{i,t} = -\frac{n(n-1)^{1/2} \sum_d D_{i,d}^3}{(n-2)\left(\sum_d D_{i,d}^2\right)^{3/2}},$$

where n is the number of observations for firm i in month t.

influence of misinformation on firm-level risks, as reflected by return volatility and stock crash risk.

[**Insert Table 10 here**]

Misinformation and small shareholders Retail and small investors are more susceptible to misinformation relative to large and institutional investors. Thus, we expect that the influence of misinformation is more salient among firms whose investors are more likely to be retail investors.

To this see, we follow Leippold et al. (2022) to proxy a firm's investor base by the average market capitalization per shareholder.[32] Firms with lower average market cap per shareholder are more likely to be held by small and retail investors. As in Leippold et al. (2022), we then classify firms into two groups based on the bottom 30% threshold and study the influence of misinformation on these two groups of firms. Results are displayed in Table 11, where the key variable of interests is the interaction of misinformation and an indicator of whether a firm has the bottom 30% average market cap per shareholder. Overall, we find that when firms have more small and retail investors, misinformation has stronger effects on investors' attention level, trading volume, return dynamics, and stock volatility. In Appendix Table C.4, we also use the proportion of institutional investors to classify firms, and we obtain similar results.[33]

[**Insert Table 11 here**]

In summary, the results in this section highlight that misinformation has important implications for financial markets: The circulation of misinformation significantly impacts investors' attention, trading volume, stock return, volatility, and crash risk, particularly among firms with a substantial presence of retail investors.

[32]We gather data from CSMAR on the number of outstanding A-share shareholders. We then compute the average market capitalization per shareholder on a monthly basis.

[33]We calculate the percentage of non-institutional investors (one minus the proportion of institutional investors) for each firm on a monthly basis. We then classify firms into two groups each month using the bottom 30% threshold.

7 Conclusion

Measuring and understanding misinformation, an increasingly critical issue in financial markets, presents a significant challenge to researchers. We tackle this challenge by introducing a systematic framework for quantifying firm-level misinformation, drawing on the principles of information consistency and collective wisdom. In doing so, we leverage the latest advancements in ML and AI and analyze 254.8 million pieces of textual materials.

We demonstrate that our framework offers an effective tool for identifying and quantifying misinformation. Our validation exercises demonstrate that the misinformation measure negatively correlates with the quality of corporate information disclosure and accurately reflects the disclosed misinformation-related events. Additionally, this misinformation measure is corroborated by the measure of public skepticism, further indicating its effectiveness.

Our exploration of the drivers of misinformation reveals that firms with weaker balance sheets and poorer governance structures, as well as those in more concentrated industries, exhibit higher levels of misinformation. Moreover, our analysis shows that significant corporate events, such as financing operations, corporate splits, and mergers and acquisitions, can drive the spread of misinformation. This finding underscores the importance of monitoring and managing misinformation during periods of heightened corporate activities.

We also demonstrate that misinformation has significant implications for financial markets. Specifically, we find that the circulation of misinformation strongly influences investor behaviors and stock return dynamics. Additionally, the degree of misinformation has strong predictive power for stock return volatility and crash risk. Our study also suggests that small and retail investors are more susceptible to misinformation relative to large investors. Overall, our framework provides a valuable tool for measuring misinformation, enabling researchers, practitioners, and regulators to better understand the sources and impact of misinformation in financial markets.

References

Ahern, K. R. and Sosyura, D. (2015). Rumor has it: Sensationalism in financial media. *Review of Financial Studies*, 28(7):2050–2093.

Allen, J., Arechar, A. A., Pennycook, G., and Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393.

An, Y., Jin, H., Liu, Q., and Zheng, K. (2022). Media attention and agency costs: Evidence from listed companies in China. *Journal of International Money and Finance*, 124:102609.

Andrei, D. and Cujean, J. (2017). Information percolation, momentum and reversal. *Journal of Financial Economics*, 123(3):617–645.

Ash, R. B. (1990). Information theory. *Dover Publications*.

Beckmann, L., Beckmeyer, H., Filippou, I., Menze, S., and Zhou, G. (2024). Unusual financial communication-Evidence from ChatGPT, earnings calls, and the stock market. *Available at SSRN*.

Belo, F., Gala, V. D., Salomao, J., and Vitorino, M. A. (2022). Decomposing firm value. *Journal of Financial Economics*, 143(2):619–639.

Brashier, N. M. and Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, 71:499–515.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. (2022). Discovering latent knowledge in language models without supervision. *ArXiv preprint arXiv:2212.03827*.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2023). Business news and business cycles. *Journal of Finance, Forthcoming*.

Chandrasekaran, D. and Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys*, 54(2):1–37.

Chen, J., Tang, G., Zhou, G., and Zhu, W. (2023). ChatGPT, stock market predictability and links to the macroeconomy. *Available at SSRN*.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Chen, S.-S. and Wang, Y. (2012). Financial constraints and share repurchases. *Journal of Financial Economics*, 105(2):311–331.

Chen, Y., Kelly, B. T., and Xiu, D. (2022). Expected returns and large language models. *Available at SSRN*.

Clarke, J., Chen, H., Du, D., and Hu, Y. J. (2020). Fake news, investor attention, and market reaction. *Information Systems Research*, 32(1):35–52.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75(3):1371–1415.

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

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

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

Dimmock, S. G., Gerken, W. C., and Graham, N. P. (2018). Is fraud contagious? Coworker influence on misconduct by financial advisors. *Journal of Finance*, 73(3):1417–1450.

Dyck, A., Volchkova, N., and Zingales, L. (2008). The corporate governance role of the media: Evidence from Russia. *Journal of Finance*, 63(3):1093–1135.

Easley, D., López de Prado, M., O' Hara, M., and Zhang, Z. (2021). Microstructure in the machine age. *Review of Financial Studies*, 34(7):3316–3363.

Erel, I., Stern, L. H., Tan, C., and Weisbach, M. S. (2021). Selecting directors using machine learning. *Review of Financial Studies*, 34(7):3226–3264.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

Fan, J., Liu, Q., Wang, B., and Zheng, K. (2023). Unearthing financial statement fraud: Insights from news coverage analysis. *Available at SSRN*.

Fan, J., Xue, L., and Zhou, Y. (2024). How much can machines learn finance from chinese text data? *Management Science*.

Fedyk, A. and Hodson, J. (2023). When can the market identify old news? *Journal of Financial Economics*, 149(1):92–113.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370.

Garza, S. E. and Schaeffer, S. E. (2019). Community detection with the label propagation algorithm: A survey. *Physica A: Statistical Mechanics and its Applications*, 534:122058.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14:337–368.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 US presidential election. *Science*, 363(6425):374–378.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. *Springer*.

Hong, Z., Liu, Q., Tse, Y., and Wang, Z. (2023). Black mouth, investor attention, and stock return. *International Review of Financial Analysis*, 90:102921.

Huang, D., Li, J., and Wang, L. (2021). Are disagreements agreeable? Evidence from information aggregation. *Journal of Financial Economics*, 141(1):83–101.

Hutton, A. P., Marcus, A. J., and Tehranian, H. (2009). Opaque financial reports, r2, and crash risk. *Journal of Financial Economics*, 94(1):67–86.

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

Jiang, F., Lee, J., Martin, X., and Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1):126–149.

Jiang, H. and Sun, Z. (2014). Dispersion in beliefs among active mutual funds and the cross-section of stock returns. *Journal of Financial Economics*, 114(2):341–365.

Joe, J. R., Louis, H., and Robinson, D. (2009). Managers' and investors' responses to media exposure of board ineffectiveness. *Journal of Financial and Quantitative Analysis*, 44(3):579–605.

Kaniel, R., Lin, Z., Pelger, M., and Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1):94–138.

Kartal, M. and Tyran, J.-R. (2022). Fake news, voter overconfidence, and the quality of democratic choice. *American Economic Review*, 112(10):3367–3397.

Kim, A. and Nikolaev, V. V. (2023). Context-based interpretation of financial information. *Chicago Booth Research Paper*, (23-08).

Kim, J.-B., Li, Y., and Zhang, L. (2011). Corporate tax avoidance and stock price crash risk: Firm-level analysis. *Journal of Financial Economics*, 100(3):639–662.

Kim, O. and Verrecchia, R. E. (2001). The relation among disclosure, returns, and trading volume information. *The Accounting Review*, 76(4):633–654.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293.

Kogan, S. and Meursault, V. (2021). Corporate disclosure: Facts or opinions? *Available at SSRN*.

Kogan, S., Moskowitz, T. J., and Niessner, M. (2023). Social media and financial news manipulation. *Review of Finance*, 27(4):1229–1268.

Kremer, I., Mansour, Y., and Perry, M. (2014). Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012.

Leippold, M., Wang, Q., and Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145(2):64–82.

Li, D. and Li, G. (2021). Whose disagreement matters? Household belief dispersion and stock trading volume. *Review of Finance*, 25(6):1859–1900.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys*, 50(6):1–45.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Li, K., Mai, F., Shen, R., Yang, C., and Zhang, T. (2023). Dissecting corporate culture using generative AI–Insights from analyst reports. *Available at SSRN*.

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., and Han, J. (2014). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1187–1198.

Li, Q., Shan, H., Tang, Y., and Yao, V. (2024). Corporate climate risk: Measurements and responses. *Review of Financial Studies*, 37(6):1778–1830.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016). A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16.

Lou, J., Lu, Y., Dai, D., Jia, W., Lin, H., Han, X., Sun, L., and Wu, H. (2023). Universal information extraction as unified semantic matching. *Proceedings of the AAAI conference on Artificial Intelligence*.

Loughran, T. and McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, 12:357–375.

Lyonnet, V. and Stern, L. H. (2022). Venture capital (mis) allocation in the age of AI. *Available at SSRN*.

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, 107(2):276.

Pedersen, L. H. (2022). Game on: Social networks and markets. *Journal of Financial Economics*, 146(3):1097–1119.

Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535.

Qin, B., Strömberg, D., and Wu, Y. (2018). Media bias in China. *American Economic Review*, 108(9):2442–2476.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ross, L., Horn, J., Pilanci, M., Luo, K., and Zhou, G. (2024). Bottom up vs top down: What does firm 10-K tell us? *Available at SSRN*.

Sadler, E. (2021). A practical guide to updating beliefs from contradictory evidence. *Econometrica*, 89(1):415–436.

Sautner, Z., Van Lent, L., Vilkov, G., and Zhang, R. (2023). Firm-level climate change exposure. *Journal of Finance*, 78(3):1449–1498.

Schmidt, D. (2020). Stock market rumors and credibility. *Review of Financial Studies*, 33(8):3804–3853.

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.

Su, J., Cao, J., Liu, W., and Ou, Y. (2021). Whitening sentence representations for better semantics and faster retrieval. *ArXiv preprint arXiv:2103.15316*.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Surowiecki, J. (2005). The wisdom of crowds. *Anchor*.

Van Binsbergen, J. H., Bryzgalova, S., Mukhopadhyay, M., and Sharma, V. (2024). (Almost) 200 years of news-based economic sentiment. *National Bureau of Economic Research*.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Wu, L., Morstatter, F., Carley, K. M., and Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.

Xu, F., Sheng, V. S., and Wang, M. (2021a). A unified perspective for disinformation detection and truth discovery in social sensing: A survey. *ACM Computing Surveys*, 55(1):1–33.

Xu, Y., Xuan, Y., and Zheng, G. (2021b). Internet searching and stock price crash risk: Evidence from a quasi-natural experiment. *Journal of Financial Economics*, 141(1):255–275.

Yang, K.-C. and Menczer, F. (2023). Large language models can rate news outlet credibility. *ArXiv preprint arXiv:2304.00228*.

Ye, C., Wang, H., Zheng, K., Kong, Y., Zhu, R., Gao, J., and Li, J. (2020). Constrained truth discovery. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):205–218.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Table 1 Summary Statistics

	Mean	SD	Min	p25	p50	p75	Max	N
	Panel A: textual corpus							
Total Sources	62791.59	54825.86	326.00	26590.00	47882.50	81088.00	302285.00	3,994
Firm Disclosures	235.86	245.71	7.00	100.00	153.00	275.00	1495.00	3,994
News	651.71	1255.17	17.00	187.00	289.00	472.00	7662.00	3,994
Analyst Reports	50.49	78.10	0.00	1.00	16.00	65.00	388.00	3,994
Social Media	61746.82	53981.32	103.00	26136.00	47042.50	80033.00	292522.00	3,994
	Panel B: firm-level characteristics							
Size	22.34	1.44	18.93	21.35	22.11	23.06	26.95	104,205
Tangibility	0.19	0.15	0.00	0.07	0.16	0.27	0.67	105,350
Debt	0.81	0.18	0.22	0.71	0.86	0.95	1.00	103,275
RevGrowth	0.14	0.63	-0.83	-0.15	0.04	0.25	3.84	103,093
FinRisk	0.02	0.03	0.00	0.00	0.01	0.02	0.23	94,397
FinConstraint	1.03	0.08	0.83	0.98	1.02	1.07	1.26	101,426
IndBoard	0.36	0.11	0.00	0.33	0.36	0.43	0.56	106,721
SupBoard	3.49	1.12	0.00	3.00	3.00	3.00	15.00	104,724
Institution %	42.73	24.82	0.40	21.80	43.47	62.76	92.56	104,568
ShareCon	3.03	4.25	0.35	0.87	1.54	3.18	27.68	104,192

Note: This table presents the summary statistics for the corpus used to measure misinformation and the variables used in different regression analyses. For each variable, we report the number of observations, mean, standard deviation, minimum, 25th, median, 75th percentiles, and maximum. Panel A provides descriptive statistics for the total number of articles per firm and their distribution from various sources. Panel B offers descriptive statistics for the core variables used in subsequent analyses, encompassing firm-level characteristics. They are the logarithm of total assets (Size), fixed asset ratio (Tangibility), current debt to asset ratio (Debt), revenue growth rate (RevGrowth), financial volatility risk (FinRisk), financing constraint risk (FinConstraint), the proportion of independent directors (IndBoard), supervisory board size (SupBoard), the proportion of shares held by institutional investors (Institution%), and stock ownership concentration (ShareConc). See Appendix Table C.1 for additional details.

Table 2 Summary Statistics for Misinformation Measures

	Panel A: firm-level and topic-level *MISI*							
	Mean	Std.Dev	Min	p25	Median	p75	Max	N
Firm-Level	0.14	0.11	0.00	0.07	0.12	0.18	4.87	313,482
FP	0.22	0.20	0.00	0.07	0.19	0.32	7.59	302,491
CG	0.15	0.23	0.00	0.03	0.10	0.19	18.45	310,731
OB	0.09	0.10	0.00	0.01	0.07	0.14	5.37	303,839
RM	0.05	0.11	0.00	0.00	0.00	0.05	9.30	293,526
RD	0.03	0.10	0.00	0.00	0.00	0.03	14.12	235,586
ER	0.02	0.08	0.00	0.00	0.00	0.00	4.31	175,800
HC	0.02	0.09	0.00	0.00	0.00	0.00	3.27	159,037
SR	0.01	0.05	0.00	0.00	0.00	0.00	2.74	120,451

	Panel B: correlation matrix of *MISI* by topic								
	Firm-Level	FP	CG	OB	RM	RD	ER	HC	SR
Firm-Level	1.00								
FP	0.61	1.00							
CG	0.67	0.28	1.00						
OB	0.56	0.40	0.24	1.00					
RM	0.39	0.35	0.31	0.27	1.00				
RD	0.17	0.08	0.06	0.40	0.07	1.00			
ER	0.18	0.13	0.12	0.17	0.22	0.03	1.00		
HC	0.09	0.04	0.06	0.18	0.05	0.11	0.01	1.00	
SR	0.12	0.05	0.04	0.12	0.07	0.03	0.05	0.08	1.00

	Panel C: persistence of *MISI* by topic								
	Firm-Level	FP	CG	OB	RM	RD	ER	HC	SR
Coefficient	0.53***	0.23***	0.56***	0.32***	0.32***	0.30***	0.17***	0.10***	0.08***
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.05)	(0.02)	(0.01)	(0.01)

Note: This table presents descriptive statistics for the misinformation measure (*MISI*). Panel A displays the degree of misinformation at the firm level and by topic. Panel B shows the average correlation coefficients for each pair of topics. Panel C details the persistence of the *MISI*, specifically through regression coefficients from period t to $t+1$ at the firm level and by topic. Abbreviations used are as follows: FP - Financial Performance, CG - Corporate Governance, OB - Operational Business, RM - Risk Management, RD - R&D Innovation, ER - Environmental Responsibility, HC - Human Capital, and SR - Social Responsibility. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 3 Misinformation and Firm Disclosure Quality

	Dep. Variable: $MISI^{avg}$			Dep. Variable: $MISI^{med}$			Dep. Variable: $MISI^{max}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IDQ	-0.0161***	-0.0052***	-0.0052***	-0.0157***	-0.0051***	-0.0052***	-0.0221***	-0.0066**	-0.0065**
	(0.0020)	(0.0014)	(0.0014)	(0.0019)	(0.0014)	(0.0014)	(0.0038)	(0.0032)	(0.0032)
Constant	0.1584***	0.1494***	0.1494***	0.1484***	0.1397***	0.1397***	0.3018***	0.2890***	0.2890***
	(0.0016)	(0.0012)	(0.0012)	(0.0016)	(0.0011)	(0.0011)	(0.0031)	(0.0026)	(0.0026)
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Time FE	NO	YES	YES	NO	YES	YES	NO	YES	YES
Industry FE	NO	NO	YES	NO	NO	YES	NO	NO	YES
N	22,332	22,332	22,332	22,332	22,332	22,332	22,332	22,332	22,332
R-squared	0.288	0.585	0.586	0.304	0.590	0.591	0.224	0.418	0.420

Note: This table presents an analysis of misinformation and firm disclosure quality. In these regressions, the dependent variables are the misinformation measures (*MISI*), and the independent variables are indicators of firm disclosure quality (IDQ), with a value of one representing good quality and zero indicating moderate quality. Columns (1) to (3) use average misinformation measures ($MISI^{avg}$) in a given year as dependent variables, whereas Columns (4) to (6) use median values ($MISI^{med}$). Columns (7) to (9) use the maximum misinformation measure in a given year ($MISI^{max}$) as the dependent variable. The results in columns (1) to (9) are based on annual frequency. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 4 ***MISI*** **and Disclosed Misinformation-Related Events**

	Dep. Variable: *Event*							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MISI	0.0130***	0.0239***	0.0229***	0.0221***	3.51***	6.62***	6.63***	6.37***
	(0.0015)	(0.0019)	(0.0019)	(0.0019)	(0.28)	(0.33)	(0.33)	(0.35)
Constant	0.0013***	-0.0002	-0.0001	-0.0221***	-7.76***	-6.53***	-7.23***	-9.66***
	(0.0002)	(0.0003)	(0.0003)	(0.0048)	(0.11)	(0.24)	(0.91)	(1.32)
Control	NO	NO	NO	YES	NO	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES
Month FE	NO	YES	YES	YES	NO	YES	YES	YES
Industry ×Year FE	NO	NO	YES	YES	NO	NO	YES	YES
N	313,479	313,479	308,266	280,965	313,482	277,939	255,884	232,207
R-squared	0.031	0.034	0.036	0.036	-	-	-	-

Note: This table shows the relationship between the misinformation measure and the disclosed misinformation-related events. Regressions are estimated at the firm-month level. The dependent variable, $Event_{i,t}$, is an indicator of whether firm i has a disclosed misinformation-related event in month t. Columns (1) to (4) employ a panel regression model, progressively including firm fixed effects, month fixed effects, industry × year fixed effects, and firm characteristics as controls. Firm characteristics include the logarithm of total assets (Size), fixed asset ratio (Tangibility), current debt ratio (Debt), revenue growth rate (RevGrowth), financial volatility risk (FinRisk), financing constraint risk (FinConstraint), the proportion of independent directors (IndBoard), supervisory board size (SupBoard), the proportion of shares held by institutional investors (Institution%), and stock ownership concentration (ShareConc). Columns (5) to (8) present results from the logit regression model, which also incorporates these controls progressively. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 5 Misinformation and Public Skepticism

	Dep. Variable: $\log(PS^{count})$				Dep. Variable: PS^{ratio}			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MISI	4.9499***	3.8591***	3.4624***	3.3371***	0.0010***	0.0045***	0.0040***	0.0039***
	(0.0531)	(0.0515)	(0.0430)	(0.0430)	(0.0002)	(0.0003)	(0.0003)	(0.0003)
Constant	4.7445***	4.8945***	4.9885***	4.0289***	0.0049***	0.0044***	0.0045***	0.0004
	(0.0073)	(0.0071)	(0.0060)	(0.1393)	(0.0000)	(0.0000)	(0.0000)	(0.0009)
Control	NO	NO	NO	YES	NO	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES
Month FE	NO	YES	YES	YES	NO	YES	YES	YES
Industry ×Year FE	NO	NO	YES	YES	NO	NO	YES	YES
N	312,285	312,285	307,078	296,472	312,285	312,285	307,078	296,472
R-squared	0.422	0.538	0.569	0.584	0.113	0.143	0.146	0.148

Note: This table shows the relationship between misinformation and public skepticism. Regressions are estimated at the firm-month level. The dependent variables in Columns (1) to (4) are the logarithm of total public skepticism count (PS^{count}), while Columns (5) to (8) use the ratio of public skepticism (PS^{ratio}) defined in equation (8). These control variables include Size, Debt, Tangibility, and RevGrowth. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 6 Misinformation, Firm characteristics, and Industry Concentration

	Dep. Variable: *MISI*					
	(1)	(2)	(3)	(4)	(5)	(6)
Size	0.0140***			0.0170***		
	(0.0012)			(0.0014)		
Tangibility	-0.0017			0.0069		
	(0.0065)			(0.0067)		
Debt	0.0124***			0.0128***		
	(0.0035)			(0.0035)		
RevGrowth	-0.0009***			-0.0006*		
	(0.0003)			(0.0003)		
FinRisk		0.0850***		0.1005***		
		(0.0125)		(0.0124)		
FinConstraint		0.1199***		0.0380***		
		(0.0130)		(0.0132)		
IndBoard			-0.0179***	-0.0162***		
			(0.0035)	(0.0044)		
SupBoard			-0.0025**	-0.0032***		
			(0.0011)	(0.0012)		
Institution%			-0.0003***	-0.0005***		
			(0.0001)	(0.0001)		
ShareCon			0.0004**	0.0005***		
			(0.0002)	(0.0002)		
HHI					1.67***	1.35***
					(0.33)	(0.26)
Constant	-0.1833***	0.0156	0.1670***	-0.2566***	-0.26***	0.14***
	(0.0285)	(0.0134)	(0.0049)	(0.0307)	(0.04)	(0.00)
Firm FE	YES	YES	YES	YES	YES	YES
Quarter FE	YES	YES	YES	YES	NO	YES
Industry ×Year FE	YES	YES	YES	YES	NO	NO
N	101,028	94,140	103,895	93,481	97,691	97,691
R-squared	0.446	0.453	0.444	0.459	0.181	0.439

Note: This table illustrates the relationship between misinformation, firm characteristics, and industry concentration. As the independent variables of interest are measured quarterly, all regressions are conducted at the firm-quarter level. The dependent variable is the average degree of firm-level misinformation per quarter. The set of firm characteristics includes the logarithm of total assets (Size), fixed asset ratio (Tangibility), current debt ratio (Debt), revenue growth rate (RevGrowth), financial volatility risk (FinRisk), financing constraint risk (FinConstraint), the proportion of independent directors (IndBoard), supervisory board size (SupBoard), the proportion of shares held by institutional investors (Institution%), and stock ownership concentration (ShareConc). HHI is the Herfindahl-Hirschman index of a firm's industry, calculated based on firm total assets. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 7 Misinformation and Major Corporate Events

	Panel A: financing-related actions			
	Dep. Variable: *MISI*			
	(1)	(2)	(3)	(4)
$Actions^{Fin}$	0.0038***	0.0023***	0.0020***	0.0022***
	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Constant	0.1348***	0.1359***	0.1376***	0.0116
	(0.0001)	(0.0001)	(0.0001)	(0.0110)
Controls	NO	NO	NO	YES
Firm FE	YES	YES	YES	YES
Month FE	NO	YES	YES	YES
Industry ×Year FE	NO	NO	YES	YES
N	311,601	311,601	306,548	296,089
R-squared	0.133	0.323	0.330	0.339
	Panel B: including other corporate actions			
	Dep. Variable: *MISI*			
	(5)	(6)	(7)	(8)
$Actions^{Com}$	0.0012***	0.0017***	0.0013***	0.0015***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Constant	0.1359***	0.1352***	0.1371***	0.0118
	(0.0002)	(0.0001)	(0.0001)	(0.0110)
Controls	NO	NO	NO	YES
Firm FE	YES	YES	YES	YES
Month FE	NO	YES	YES	YES
Industry ×Year FE	NO	NO	YES	YES
N	311,601	311,601	306,548	296,089
R-squared	0.131	0.323	0.330	0.339

Note: This table presents an analysis of major corporate events and misinformation. Regressions are estimated at the firm-month level. Panel A investigates the impact of corporate financing-related events ($Actions^{Fin}$), specifically equity financing (including rights issues and seasoned equity offerings), equity transactions, and mergers and acquisitions. These control variables include Size, Debt, Tangibility, and RevGrowth. Panel B extends the analysis to a broader set of corporate events ($Actions^{Com}$), adding equity pledges, legal disputes, and changes in senior management. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 8 Misinformation, Investor Attention, and Trading Volume

	Panel A: investor attention					
	$Attention_t$			$Attention_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	0.8104***	0.8101***	0.8070***	0.0474***	0.0458***	0.0430***
	(0.0289)	(0.0290)	(0.0288)	(0.0074)	(0.0074)	(0.0074)
$Attention_t$				0.6715***	0.6716***	0.6732***
				(0.0045)	(0.0045)	(0.0046)
InvestorDIS		-0.0114	0.0035		-0.0588***	-0.0631***
		(0.0258)	(0.0257)		(0.0129)	(0.0129)
AnalystDIS		-0.0002	-0.0004		0.0002	0.0003
		(0.0005)	(0.0005)		(0.0002)	(0.0002)
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES
Industry × Year FE	YES	YES	YES	YES	YES	YES
N	165,162	165,109	162,702	161,622	161,581	159,267
R-squared	0.760	0.760	0.764	0.868	0.868	0.868
	Panel B: trading volume					
	$Trading\ Volume_t$			$Trading\ Volume_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	1.3715***	1.3907***	1.3110***	0.2654***	0.2479***	0.1089***
	(0.0443)	(0.0448)	(0.0441)	(0.0165)	(0.0162)	(0.0139)
$Trading\ Volume_t$				0.5565***	0.5587***	0.6398***
				(0.0048)	(0.0048)	(0.0040)
InvestorDIS		0.8194***	0.4655***		-0.5761***	-0.3520***
		(0.0504)	(0.0472)		(0.0267)	(0.0230)
AnalystDIS		0.0006***	0.0006***		0.0004	0.0002
		(0.0002)	(0.0002)		(0.0003)	(0.0002)
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES
Industry × Year FE	YES	YES	YES	YES	YES	YES
N	300,669	300,570	288,604	300,602	300,503	288,537
R-squared	0.678	0.679	0.711	0.812	0.812	0.832

Note: This table shows the relationship between misinformation and attention and trading volume. Regressions are estimated at the firm-month level. Panel A explores investor attention, measured as the logarithm of each company's search volume on Baidu, and Panel B studies trading volume, measured as the logarithm of each company's total trading volume. The control variables include investor disagreement (*InvestorDIS*), analyst disagreement (*AnalystDIS*), Size, Debt, Tangibility, Revenue Growth, and Actions (defined in Table 7). For the attention and trading volume in subsequent periods (columns 4-6 of Panels A and B), we control for the respective measures from the current period. We omit the intercepts to save space. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 9 Misinformation and Stock Returns

	Panel A: contemporaneous returns					
	Panel Regression			Fama-Macbeth		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	0.0385***	0.0361***	0.0290***	0.0464***	0.0398**	0.0280*
	(0.0040)	(0.0040)	(0.0038)	(0.0160)	(0.0157)	(0.0146)
InvestorDIS		-0.0980***	-0.0779***	-0.1221***	-0.0727**	
		(0.0053)	(0.0044)	(0.0463)	(0.0290)	
AnalystDIS		0.0002	0.0002		0.0285***	0.0395***
		(0.0001)	(0.0001)	(0.0083)	(0.0143)	
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	-	-	-
Month FE	YES	YES	YES	-	-	-
Industry × Year FE	YES	YES	YES	-	-	-
N	299,042	298,981	288,422	299,051	298,990	288,460
R-squared	0.251	0.252	0.289	0.005	0.009	0.039
	Panel B: future returns					
	Panel Regression			Fama-Macbeth		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	-0.0429***	-0.0435***	-0.0411***	-0.0245***	-0.0235***	-0.0200***
	(0.0025)	(0.0025)	(0.0026)	(0.0057)	(0.0058)	(0.0059)
$Return_t$	-0.0469***	-0.0471***	-0.0520***	-0.0245**	-0.0246**	-0.0360***
	(0.0031)	(0.0031)	(0.0032)	(0.0109)	(0.0109)	(0.0111)
InvestorDIS		-0.0215***	-0.0246***		-0.0080	-0.0086
		(0.0048)	(0.0048)		(0.0111)	(0.0131)
AnalystDIS		0.0003	0.0003*		0.0006	0.0008
		(0.0002)	(0.0002)		(0.0095)	(0.0083)
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	-	-	-
Month FE	YES	YES	YES	-	-	-
Industry × Year FE	YES	YES	YES	-	-	-
N	298,975	298,914	288,355	298,982	298,921	288,391
R-squared	0.304	0.304	0.305	0.017	0.019	0.047

Note: This table shows the relationship between misinformation and stock returns. Regressions are estimated at the firm-month level. Panel A explores current returns. Specifically, Columns (1) to (3) employ a two-way fixed effects regression. Standard errors are clustered at the firm level. Columns (4) to (6) utilize the Fama-Macbeth regression method, where standard errors are computed using Newey-West adjustment with six lags. The control variables include investor disagreement (*InvestorDIS*), analyst disagreement (*AnalystDIS*), Size, Debt, Tangibility, Revenue Growth, and Actions (defined in Table 7). For Panel B, we control for the current period's return ($Return_t$). Due to space constraints, the constant coefficients are not displayed. Standard errors are shown in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 10 Misinformation, Return Volatility, and Stock Crash Risk

	Panel A: contemporaneous risk					
	$NCSKEW_t$			$Volatility_t$		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	0.2110***	0.2104***	0.2045***	0.0276***	0.0274***	0.0212***
	(0.0140)	(0.0140)	(0.0143)	(0.0013)	(0.0013)	(0.0009)
InvestorDIS		-0.0388	-0.0456*		-0.0012	0.0001
		(0.0269)	(0.0273)		(0.0015)	(0.0007)
AnalystDIS		0.0007***	0.0007***		0.0001***	0.0001***
		(0.0002)	(0.0002)		(0.0000)	(0.0000)
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES
Industry × Year FE	YES	YES	YES	YES	YES	YES
N	297,671	297,593	286,342	300,277	300,181	288,295
R-squared	0.051	0.051	0.052	0.104	0.105	0.255
	Panel B: future risk					
	$NCSKEW_{t+1}$			$Volatility_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	0.0576***	0.0564***	0.0516***	0.0077***	0.0076***	0.0056***
	(0.0147)	(0.0147)	(0.0150)	(0.0004)	(0.0004)	(0.0009)
$NCSKEW_t$	-0.0061***	-0.0061***	-0.0058***			
	(0.0019)	(0.0019)	(0.0020)			
$Volatility_t$				0.0398***	0.0399***	0.1197***
				(0.0045)	(0.0046)	(0.0392)
InvestorDIS		-0.0288	-0.0103		-0.0031***	-0.0011**
		(0.0277)	(0.0282)		(0.0005)	(0.0005)
AnalystDIS		-0.0002	-0.0001		0.0000*	0.0000*
		(0.0004)	(0.0005)		(0.0000)	(0.0000)
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES
Industry × Year FE	YES	YES	YES	YES	YES	YES
N	292,066	292,000	280,892	294,086	294,002	282,269
R-squared	0.052	0.052	0.052	0.495	0.495	0.510

Note: This table shows the relationship between misinformation and stock crash risk. Regressions are estimated at the firm-month level. We use the negative conditional return skewness (*NCSKEW*) as proxies for stock crash risk following Kim et al. (2011). The control variables include investor disagreement (*InvestorDIS*), analyst disagreement (*AnalystDIS*), Size, Debt, Tangibility, Revenue Growth, and Actions (defined in Table 7). For Panel B, we control for the current period's risk ($NCSKEW_t$ and $Volatility_t$), respectively. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 11 Misinformation, Shareholders Types, and Financial Outcomes

	Panel A: contemporaneous outcomes				
	$Attention_t$	$Trading_t$	$Return_t$	$Volatility_t$	$NCSKEW_t$
	(1)	(2)	(3)	(4)	(5)
MISI	0.6845***	1.1620***	0.0053	0.0177***	0.2120***
	(0.0339)	(0.0497)	(0.0038)	(0.0007)	(0.0164)
HolderType	-0.0350***	0.1278***	-0.0158***	-0.0022***	-0.0235***
	(0.0092)	(0.0154)	(0.0013)	(0.0003)	(0.0065)
$MISI \times HolderType$	0.3623***	0.5914***	0.0730***	0.0120***	-0.0289
	(0.0440)	(0.0753)	(0.0077)	(0.0019)	(0.0282)
Controls	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES
Industry ×Year FE	YES	YES	YES	YES	YES
N	162,702	284,006	283,824	283,719	281,920
R-squared	0.766	0.718	0.297	0.329	0.053
	Panel B: future outcomes				
	$Attention_{t+1}$	$Trading_{t+1}$	$Return_{t+1}$	$Volatility_{t+1}$	$NCSKEW_{t+1}$
	(1)	(2)	(3)	(4)	(5)
MISI	0.5747***	0.8595***	-0.0388***	0.0073***	0.0426**
	(0.0251)	(0.0409)	(0.0028)	(0.0004)	(0.0177)
HolderType	-0.0108	0.1612***	0.0111***	-0.0012***	-0.0264***
	(0.0083)	(0.0143)	(0.0012)	(0.0002)	(0.0065)
$MISI \times HolderType$	0.1949***	0.3517***	-0.0173***	0.0040***	0.0136
	(0.0357)	(0.0644)	(0.0053)	(0.0007)	(0.0281)
Controls	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES
Industry ×Year FE	YES	YES	YES	YES	YES
N	160,075	283,940	283,940	277,807	278,187
R-squared	0.756	0.718	0.289	0.488	0.052

Note: This table shows the interplay between misinformation, shareholder types, and various financial outcomes, including investor attention, trading volume, return, return volatility, and crash risk. Following Leippold et al. (2022), we classify firms each month into two categories based on a threshold of the lowest 30% in average market capitalization per shareholder. Panel A presents current period outcomes, while Panel B presents subsequent period outcomes. The control variables include investor disagreement (*InvestorDIS*), analyst disagreement (*AnalystDIS*), Size, Debt, Tangibility, Revenue Growth, and Actions (defined in Table 7). We are particularly interested in the coefficients of *MISI*× HolderType to analyze how firms with different types of shareholders respond to misinformation. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

APPENDIX

A Additional Technical Details

In this section, we supplement additional technical details of the MMTA framework, detailing how we preprocess text and categorize raw text into comparable subsets.

A.1 Data preprocessing

We begin with a diverse array of raw data sources, including firm announcements, social media posts, news articles, and analyst reports. Data from social media and news articles are typically in plain text format, while firm announcements and analyst reports are in PDF format. Our preprocessing steps include the following:

Converting all PDF files to text. We utilize PyMuPDF,[34] an open-source PDF toolkit in Python, to extract structured information from PDF documents and transform it into text. PyMuPDF is known for its efficiency and comprehensive functionality in handling PDF files, including text extraction and text property extraction. The converted text documents are formatted into a sentence segmentation module, inclusive of newline characters and other symbols. During this phase, we perform various normalization and cleaning processes, including removing headers and footers, discarding tables, and eliminating inherent text formatting. For the content of analyst reports, we only extract the summary sections. Ultimately, we form 290,048,159 raw articles. After preprocessing, as described in Section 3, we have 254,826,060 articles related to the sample firms.

Chunking segment. Following the methodology outlined in Li et al. (2023), we segment all text files using their inherent paragraph structure and headings to maintain the seman-

[34]https://github.com/pymupdf/PyMuPDF.

tic integrity necessary for subsequent model processing.[35] Specifically, we initially identify headings based on textual conventions and then divide the text into sections accordingly. Each section is subsequently split into paragraphs based on newline characters. For paragraphs exceeding the character limit, we further segment them into semantic units based on punctuation marks, setting a character limit of 200. For these segments, we perform a series of denoising operations. Considering the low signal-to-noise ratio in social media, which contains a high volume of irrelevant or minimal information sentences, we eliminate segments shorter than ten characters. Additionally, for news articles, we retain segments with high relevance to companies by checking if company names, abbreviations, or codes appear in the surrounding text. We discard segments that are excessively templated. This segmentation process results in a total of 141,270,419 distinct segments.

A.2 Categorize raw text into comparable subsets

This section provides additional details regarding the algorithm in Section 2.2.1. As explained in Section 2.2.1, we first perform coarse-grained topic categorization on each text segment, followed by fine-grained entity extraction for texts relevant to our predefined topics. Subsequently, we cluster these entities and assign each piece of information a tuple label (o, e), constructing information sets categorized by these tuples for each firm. To achieve these steps, we employ a machine learning model specifically designed and trained for this task.

We utilize the same foundational pre-trained large language model (LLM) and text processing model for both information classification and extraction, ensuring consistency across our model's logic and processing. Pre-trained LLMs are now the foundation of almost all natural language processing tasks, converting text into fixed-dimension vectors. We use the ERNIE model, which builds on an architecture similar to BERT but is particularly

[35] Pre-trained LLMs, such as BERT and ERNIE, typically have a character limit of 512. Excessively long segments can also negatively affect the accuracy of topic classification and entity extraction.

optimized for Chinese text (Sun et al., 2020). ERNIE is trained on a vast corpus of Chinese knowledge graphs and text, including encyclopedic professional knowledge, news information, and forum dialogues in various language styles. It has shown superiority over BERT in several Chinese text classification tasks, proving more effective in providing contextually relevant embeddings (Sun et al., 2020). We capture representations of each sentence in a 768-dimensional vector space, regardless of varying lengths.

A standard practice is to fine-tune the pre-trained model on a specific task using labeled datasets, adjusting the model parameters (Chen et al., 2022). However, given the context of our research, we adopt a few-shot learning framework to mitigate the need for extensive training datasets. This approach utilizes a smaller number of high-quality training samples, effectively leveraging the pre-trained model's general language understanding capabilities while adapting to task-specific nuances.

Specifically, we use the Unified Semantic Matching (USM) model proposed by Lou et al. (2023), a widely employed and highly effective algorithm designed for diverse text-based semantic classification and information extraction tasks. As of February 2023, this model ranks first in ZeroCLUE and FewCLUE competition tasks. This model vectorizes both user-defined labels and text, then constructs a shared semantic space for labels and text through Directed-Token-Linking (DTL) and performs semantic matching. This process involves linking tokens within the text (Token-Token Linking, TTL), linking labels to text tokens (Label-Token Linking, LTL), and linking tokens to labels (Token-Label Linking, TLL). Each linking score is computed as follows, ultimately determining the semantic matching relationship:

$$s_{\text{TTL}}(t_i, t_j) = \text{FFNN}^{\text{left}}_{\text{TTL}}(h_{t_i})^T \cdot R_{j-i} \cdot \text{FFNN}^{\text{right}}_{\text{TTL}}(h_{t_j}), \tag{A.1}$$

$$s_{\text{LTL}}(l_i, t_j) = \text{FFNN}^{\text{label}}_{\text{LTL}}(h_{l_i})^T \cdot R_{j-i} \cdot \text{FFNN}^{\text{text}}_{\text{LTL}}(h_{t_j}), \tag{A.2}$$

$$s_{\text{TLL}}(t_i, l_j) = \text{FFNN}^{\text{text}}_{\text{TLL}}(h_{t_i})^T \cdot R_{j-i} \cdot \text{FFNN}^{\text{label}}_{\text{TLL}}(h_{l_j}). \tag{A.3}$$

Here, h_{t_i}, h_{t_j}, h_{l_i}, and h_{l_j} represent the embeddings of tokens and labels, while R_{j-i} is a

relative position encoding matrix. FFNNs (Feed-forward neural networks) process these embeddings to calculate the linking scores. For this combined pre-trained large language model and unified semantic matching model, we construct training sets for both classification and extraction tasks separately.

For the information classification task, we handle a multi-label assignment involving eight core topics plus an "others" category. The goal of the extraction task is to identify and describe relevant entities within the text. We randomly extract 600 samples from each data source and use the same batch for annotations. Initially, we label the training sets for both tasks using ChatGPT-4. We employ a chain-of-thought prompting strategy, as detailed in Li et al. (2023). The prompt used is as follows:

You are a financial expert specializing in analyzing and understanding various types of information within the financial market. The text sources include company announcements, research reports, news articles, and social media. Your task is to analyze the provided Chinese text content and complete the following steps, thank you:
1) Determine whether the text forms a complete sentence or paragraph, i.e., whether it is semantically coherent. If so, proceed to the next question; if not, please respond in the following format and do not continue with the task: Topic: Other Entity: Semantically incoherent
2) Classify the text into one or more of the most relevant topics related to specific company behaviors. The selectable topics are "Financial Performance," "Operational Business," "Corporate Governance," "Human Capital," "Social Responsibility," "Environmental Responsibility," "Risk Management," "R&D Innovation." If there is no suitable topic or if the text is particularly incoherent, classify it as "other." Please note the following:
a. The chosen topics must reflect the company's behavior or status and not be generic. If the text is too general, classify it as "other."
b. Ensure that the topic terms strictly follow the provided words without creating new ones; this is very important.

c. The text should be classified by its core topic.

3) Extract the core specific described entities (such as products, etc.) from the text, as well as descriptive content and stance (positive, negative, or neutral). The extracted entities and contents must be directly taken from the original text without creating new phrases. Try to avoid vague entities such as "the company." If the text mentions multiple described entities, separate them with semicolons. If a part is missing, denote it as "not mentioned."

Please respond in the following format:

"Topic: [Your Answer]"

"Entity: [Your Answer]"

Subsequently, we ensure the accuracy and comprehensiveness of labels through manual annotation. We then divide the dataset into training, validation, and testing sets in an 80/10/10 ratio and optimize the model based on the Micro-F1 metric, computed as follows:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FP}_i}, \tag{A.4}$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FN}_i}, \tag{A.5}$$

$$F1_{\text{micro}} = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}. \tag{A.6}$$

The trained classification model achieves good performance, with a Micro-F1 score of 92.04 on the test set. For the information extraction task, we similarly use the F1 score as the optimization goal, ultimately achieving a score of 78.36.

We perform clustering on the extracted entities to facilitate canonicalization. For this purpose, we employ the hierarchical clustering technique, a widely used method well-documented in Ward Jr (1963); Hastie et al. (2009). For monthly text data slices from each firm, we use the scikit-learn library (https://github.com/scikit-learn/scikit-learn) in

Python to assign the same cluster ID to synonymous entities. This approach ensures that entities with similar meanings are grouped together, enhancing the analysis' consistency and reliability.

B Validating the FT corpus

This section investigates the validity of the FT corpus. Theoretically, the diversity and independence of data sources are critical. As discussed in Surowiecki (2005) and Allen et al. (2021), the quality of an information set hinges not merely on volume but on the variety and autonomy of the sources. For instance, while each analyst might represent an individual data source, they are not entirely independent, owing to possible herding behaviors within the analyst community that could undermine collective wisdom (Kremer et al., 2014; Prelec et al., 2017). Additionally, reducing information redundancy is essential. Excessive or biased redundancy can obscure meaningful signals, leading to noise amplification rather than enhanced understanding (Ash, 1990; Schmidt, 2020; Xu et al., 2021a).

Thus, we quantitatively assess the effectiveness of our data sources. Fedyk and Hodson (2023) define old news as news that has some overlap with historical news and discuss the informational value of "old news."We apply a similar principle to evaluate the contributions of each data source, known as Information Gain Analysis (IGA). This analysis helps to quantify the incremental information each new data brings to our corpus. Information gain diminishes as the dataset approaches an "information saturation point," beyond which additional data fails to significantly enhance content value.

Specifically, IGA is a quantitative approach derived from information theory, widely used in machine learning for feature selection (Li et al., 2017), and it involves several key steps:

Preprocessing. Prior to calculating information gain, textual data must undergo preprocessing, including the removal of stop words and segmentation of the text into words, to

facilitate effective analysis of words.

Entropy calculation. Entropy, denoted as $H(D)$, quantifies the uncertainty within the dataset D and is calculated using the equation:

$$H(D) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i), \tag{B.1}$$

where $p(x_i)$ is the frequency of occurrence of the i-th word in dataset D. We then introduce an additional dataset D_{new} and compute the entropy of the combined datasets $H(D \cup D_{\text{new}})$.

Information gain calculation. Given the textual nature of our corpus, we adopt a simplified version of standard calculation. Information gain here is calculated as the difference between the original entropy and the combined entropy:

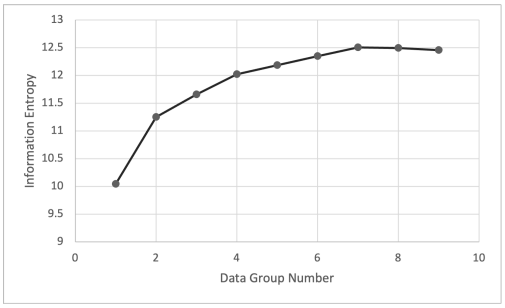
$$\Delta H = H(D \cup D_{\text{new}}) - H(D). \tag{B.2}$$

This metric assesses the reduction in uncertainty and thus the value added by the new data to D. A decrease in information gain below a specified threshold (here, 0.001) suggests a stable contribution level from the data source.

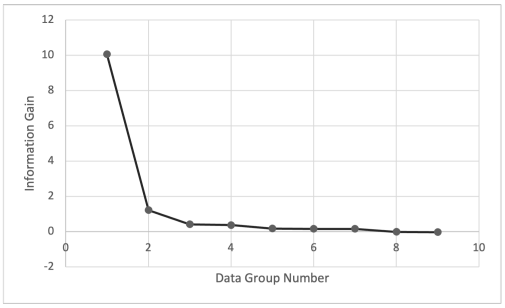
We apply IGA based on principles of source diversity, independence, and minimal redundancy to ensure the validity of our corpus. First, we select data groups based on the interactions of three primary financial market participants—firms, investors, and financial intermediaries to ensure diversity. Our selection includes nine distinct groups: firm announcements, investor Q&A sessions, analyst reports, social media posts, governmental newspapers, mainstream economic and financial newspapers, online financial platforms, and financial information websites. To validate stability, we randomly initialize the data groups and iteratively add new groups in a randomized order. Across 50 experiments, we calculate the average information gain and entropy at each iteration as shown in Panels (a) and (b) of

Figure 4. The results show that the corpus's stability is achieved with the inclusion of the nine data groups.

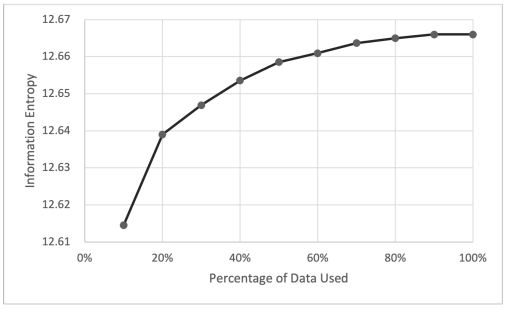
Second, for internal validation within each group, we start with a random 10% of the data, incrementally adding 10% in subsequent rounds. As shown in Panels (c) and (d) of Figure 4, this incremental approach indicates that about 80% of the current data volume is sufficient to reach stability. These results validate the effectiveness of our corpus, reflecting a thoughtful balance between comprehensive coverage and efficient data utilization.



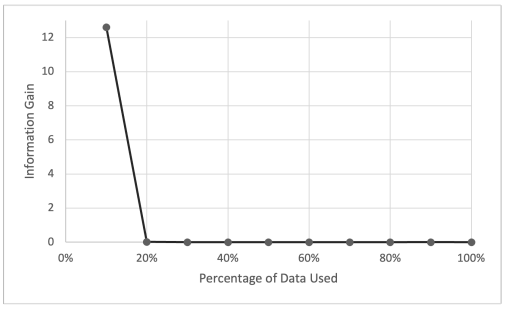
(a) Information Entropy for Data Groups



(b) Information Gain for Data Groups



(c) Information Entropy for Data Volume



(d) Information Gain for Data Volume

Figure 4 Information Gain Analysis for Data Selection

C Additional Tables and Figures

Table C.1 Variable Definitions

Variable	Definition and Construction
Size	Natural logarithm of total assets.
Tangibility	Net fixed assets divided by total assets
Debt	Current debt divided by total debt.
RevGrowth	The growth in total operating revenue, calculated as (Total operating Revenue for the current period - Total operating revenue for the previous period)/(Total operating revenue for the previous period).
FinRisk	The volatility of earnings, calculated as the standard deviation of earnings before interest and taxes over total assets for the past three years, multiplied by one hundred.
FinConstraint	Whited-Wu Index, representing financial constraint, calculated following Chen and Wang (2012) as $-0.091 \times X_1 + 0.06 \times X_2 + 0.01 \times X_3 + 0.044 \times X_4 + 0.10 \times X_5 - 0.03 \times X_6$, where X_1 is cash flow to total assets ratio, X_2 is a dummy variable for cash dividend distribution, X_3 is the ratio of long-term debt to assets, X_4 is the natural logarithm of total assets, X_5 is industry sales growth, and X_6 is the growth rate of sales revenue. A firm with a high WW index is considered more financially constrained.
IndBoard	The proportion of independent directors to the board size.
SupBoard	The number of supervisory board members.
Institution%	The percentage of institutional ownership.
ShareCon	The ratio of the holdings of the largest shareholder to the total holdings of the second to the fifth largest shareholders.

Table C.2 Misinformation and Firm Disclosure Quality

	Dep. Variable: *MISI*								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IDQ^{KV}	0.0158***	0.0137***	0.0136***	0.0133***	0.0161***	0.0137***	0.0157***	0.0134***	0.0135***
	(0.0031)	(0.0022)	(0.0022)	(0.0021)	(0.0058)	(0.0048)	(0.0029)	(0.0021)	(0.0048)
Constant	0.1333***	0.1341***	0.1342***	0.1245***	0.2714***	0.2724***	0.1235***	0.1245***	0.2725***
	(0.0012)	(0.0009)	(0.0009)	(0.0008)	(0.0023)	(0.0019)	(0.0012)	(0.0008)	(0.0019)
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Time FE	NO	YES	YES	NO	YES	YES	NO	YES	YES
Industry FE	NO	NO	YES	NO	NO	YES	NO	NO	YES
N	23,294	23,294	23,294	23,294	23,294	23,294	23,294	23,294	23,294
R-squared	0.260	0.581	0.582	0.592	0.211	0.403	0.275	0.591	0.404

Note: This table presents an analysis of misinformation and firm disclosure quality. In these regressions, the dependent variable is the misinformation measure (*MISI*), and the key independent variable is a market-based measure of disclosure quality (IDQ^{KV}) proposed by Kim and Verrecchia (2001), *with higher values indicating lower disclosure quality*. We collect these data from the MAKE database (www.macrodatas.cn), one of China's largest social science data-sharing platforms. Columns (1) to (3) use the average monthly misinformation measure in a given year as the dependent variable, whereas Columns (4) to (6) use the median value. Columns (7) to (9) use the maximum monthly misinformation measure in a given year as the dependent variable. The results in Columns (1) to (9) are based on annual frequency. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table C.3 Misinformation and Investor Attention

	$Attention_t$			$Attention_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)
MISI	2.5930***	2.6138***	2.4874***	1.9420***	1.9544***	1.8432***
	(0.0924)	(0.0931)	(0.0908)	(0.0645)	(0.0646)	(0.0642)
Constant	10.9404***	10.3783***	8.5351***	11.0426***	10.7672***	8.7512***
	(0.0149)	(0.0439)	(0.2833)	(0.0102)	(0.0426)	(0.2876)
Disagreement	NO	YES	YES	NO	YES	YES
Controls	NO	NO	YES	NO	NO	YES
Firm FE	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES
Industry × Year FE	YES	YES	YES	YES	YES	YES
N	200,321	200,231	193,227	199,446	199,368	192,413
R-squared	0.582	0.584	0.608	0.581	0.581	0.596

Note: This table presents the relationship between misinformation and investor attention. Regressions are estimated at the firm-month level. Investor attention is measured as the logarithm of each company's search volume of investors on 13 major stock trading and stock broker apps. We obtain this data from HMM Technology, a start-up specializing in alternative data. Control variables for disagreement include investor disagreement (*InvestorDIS*) and analyst disagreement (*AnalystDIS*). Additional control variables include Size, Debt, Tangibility, Revenue Growth, and Actions (as defined in Table 7). Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table C.4 Misinformation, Shareholders Types, and Financial Outcomes

	Panel A: contemporaneous outcomes				
	$Attention_t$	$Trading_t$	$Return_t$	$Volatility_t$	$NCSKEW_t$
	(1)	(2)	(3)	(4)	(5)
MISI	0.5994***	0.9438***	0.0219***	0.0164***	0.1729***
	(0.0462)	(0.0694)	(0.0059)	(0.0015)	(0.0214)
HolderType	0.0207	0.1220***	-0.0011	-0.0018***	-0.0092
	(0.0134)	(0.0210)	(0.0014)	(0.0003)	(0.0077)
$MISI \times HolderType$	0.3163***	0.5546***	0.0087	0.0070***	0.0454*
	(0.0518)	(0.0812)	(0.0069)	(0.0017)	(0.0265)
Controls	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES
Industry ×Year FE	YES	YES	YES	YES	YES
N	162,702	288,578	288,396	288,270	286,319
R-squared	0.766	0.713	0.290	0.256	0.052
	Panel B: future outcomes				
	$Attention_{t+1}$	$Trading_{t+1}$	$Return_{t+1}$	$Volatility_{t+1}$	$NCSKEW_{t+1}$
	(1)	(2)	(3)	(4)	(5)
MISI	0.5241***	0.6696***	-0.0314***	0.0068***	0.0362
	(0.0347)	(0.0571)	(0.0036)	(0.0006)	(0.0265)
HolderType	0.0409***	0.1269***	0.0044***	-0.0008***	-0.0018
	(0.0127)	(0.0202)	(0.0013)	(0.0002)	(0.0079)
$MISI \times HolderType$	0.1740***	0.4199***	-0.0160***	0.0022***	0.0164
	(0.0412)	(0.0687)	(0.0044)	(0.0007)	(0.0301)
Controls	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES
Industry ×Year FE	YES	YES	YES	YES	YES
N	160,075	288,511	288,511	282,246	282,635
R-squared	0.756	0.714	0.298	0.487	0.052

Note: This table presents robustness checks for Table 11. Each month, we classify firms into two groups based on the threshold of the lowest 30% in the proportion of institutional investors. Standard errors, clustered at the firm level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.