

Generative AI, Investor Composition, and Information Barriers*

Encarna Guillamón Saorín
Universidad Carlos III de Madrid
eguillam@emp.uc3m.es

María Gutiérrez Urtiaga
Universidad Carlos III de Madrid
mgurtiag@emp.uc3m.es

Rongyi Yao
Universidad Carlos III de Madrid
ryao@emp.uc3m.es

December 20, 2025

*Funded by TED2021-129861B-I00 Unión Europea NextGenerationEU/PRTR. We acknowledge the support of the Google Cloud Research Credits program with the award GCP313990497 and the funding provided by NextGeneration-EU, PRTR, The Spanish Institute of Cybersecurity (INCIBE-CYBERTHREAT C115/23), and financial support from the Ministry of Science and Education (PID2019-111143GB-C33, TED2021-129861B-I00). We thank Edoardo Acabbi, David García Méndez, Juan Manuel García Lara, Beatriz García Osma, Marco Giometti, Bing Guo, Wayne Landsman, Li Liu, David Martínez Miera, Eduardo Melero Martín, Huiping Ren, Anna Toldrà Simats, participants at the PRICIT 2024 workshop, XVIII International Accounting Research Symposium 2024, and FMARC 2025 doctoral symposium for their invaluable comments.

Abstract

Which type of investor benefits more in the era of Large Language Models (LLMs) remains unclear. We study how managers from different investor bases adjust qualitative disclosure differentially after the introduction of large language models (LLMs). Focusing on 10-K MD&A sections, we exploit the release of ChatGPT 3.5 as an exogenous shock to investors' information-processing technologies and decompose linguistic complexity into an information component (reflecting business-related complexity) and an obfuscation component (reflecting residual complexity unexplained by fundamentals). We show that, on average, the information component of linguistic complexity decreases after the release of ChatGPT 3.5, consistent with managers providing less information-rich narrative content. This reduction in the information component is more pronounced for firms whose investor base is tilted toward less sophisticated investors and for firms with more concentrated institutional ownership. By contrast, we find no systematic relation between the obfuscation component and the overall investor base, although obfuscation increases in firms with higher ownership by active (non-index) mutual funds. Overall, our evidence is consistent with managers adjusting narrative disclosure in response to LLMs in a manner that tends to preserve, and in some cases enhance, the relative informational advantages of sophisticated investors over less sophisticated investors.

For more than forty years, I've studied the documents that public companies file. Too often, I've been unable to decipher just what is being said or, worse yet, had to conclude that nothing was being said. [...] Maybe we simply don't have the technical knowledge to grasp what the writer wishes to convey. Or perhaps the writer doesn't understand what he or she is talking about. In some cases, moreover, I suspect that a less-than scrupulous issuer doesn't want us to understand a subject it feels legally obligated to touch upon.

-Warren Buffett, 1998¹

1 Introduction

Narrative disclosure is the main channel through which firms communicate with capital markets, yet these disclosures are often linguistically complex (Blankespoor, 2019; Miller, 2010; You & Zhang, 2009). This complexity can increase with informativeness, reflecting the need to convey technical details (R. Bloomfield, 2008; Bushee et al., 2018). However, it can also be used to obfuscate and impede public information processing, sustaining private-information advantages (R. Bloomfield, 2002; Maffett, 2012; You & Zhang, 2009). Compared to sophisticated investors such as institutions, less sophisticated investors suffer more from linguistic complexity and are less likely to read financial reports due to limited attention and high disclosure processing costs (Blankespoor, 2019; Blankespoor et al., 2020; Hirshleifer & Teoh, 2003).

Although technological advancements may shrink the gap of information processing capacity between naive and savvy investors, most previous empirical findings indicate the opposite (Lawrence, 2013; Miller, 2010). For instance, retail trading activities are reduced when filings are longer and more linguistically complex, suggesting that technologies at that time were not sufficiently powerful to significantly advance the gap (Miller, 2010). However, the recent advent of large language models (LLMs) is revolutionizing disclosure processing and trading behaviors (Bertomeu et al., 2025; Cheng et al., 2025), as they lower the cost of extracting and analyzing

¹From *A plain English Handbook: How to create clear SEC disclosure documents* (SEC, 1998).

information from text (Bilinski, 2024; A. Kim et al., 2023a, 2023b). In addition to their analytical capabilities, LLMs have other strengths that help them move from niche technology circles into mainstream public use.² From turning on the computer to obtaining a fully customized analysis, LLMs help less sophisticated investors analyze financial reports in minutes, providing a rare opportunity for them to narrow the information and skill gaps against more sophisticated investors.

Investors are adopting LLMs swiftly, but it is unclear whether the information gap between sophisticated and less sophisticated investors is narrowed or widened after LLM adoption (Blankespoor et al., 2024a; Cheng et al., 2025; A. Kim et al., 2024a). Recent papers that study this issue treat the release of ChatGPT as an exogenous shock to the information environment that is taken as given, but do not consider potential *strategic* changes to disclosure policy followed by the shock (Bertomeu et al., 2025; Blankespoor et al., 2024b). As the theory of investor sophistication affecting disclosure strategy is well documented (Aghamolla & Smith, 2023; Boone & White, 2015; Maffett, 2012), we address this information gap question from the angle of managers. *How do managers adjust linguistic complexity based on investor base after the release of ChatGPT?*

Our argument is that managers will make disclosure more opaque (i.e., less transparent) after the release of LLMs, and the effect will be more pronounced for firms with a greater percentage of less sophisticated investors. Previous literature demonstrates channels such as broker-hosted investor conferences for institutions in the post-regulation FD world to gain an information advantage against the rest of the investors (Green et al., 2014). Such an advantage brings private benefits from informed trading. To maximize the benefits, sophisticated investors prefer firms with

²Firstly, not only does ChatGPT’s dictionary-level multi-language support reduce users’ language barriers, it also creates, detects, and analyzes human-like tones and narratives (Ahmad, 2023). Secondly, the response is not only swift but the most user-customized ever. It saves users’ time in creating algorithms in reading and analyzing financial reports more efficiently, reducing the financial literacy threshold and users’ costs of learning and information collection (Helter, 2023). Thirdly, while investment software and prior AI tools are well known inside the investment circle, GPT’s easy accessibility and friendly user interface overstep such a circle to even the most naive investors, gaining worldwide popularity (Hu & Hu, 2023). In other words, domestic markets can increase participation from foreign new-coming investors at every knowledge level.

an opaque information environment to maintain the information gap against the rest of investors (Maffett, 2012). Given the adoption level and analysis capacity of LLMs, less sophisticated investors can extract more information from the firm. If a firm does not increase opacity after the boom of generative AI, sophisticated investors can switch, through "voting with their feet," to another firm with higher opacity to maintain the information gap against their counterparts. In this way, they can obtain constant private benefits from informed trading. To keep sophisticated investors from switching firms, all managers are going to signal with increased opacity in equilibrium. If all firms hold the same amount of information gaps, the higher the percentage of less sophisticated investors in a firm, the more information gaps can be potentially narrowed with LLMs, the fewer opportunities sophisticated investors foresee to reap benefits from private information in this firm, and the greater effort managers have to put to widen the information gaps narrowed by LLMs. Therefore, firms with a greater percentage of less sophisticated investors are incentivized to raise opacity more³. Methodologically, a more opaque information environment means that managers need to provide less information (more obfuscation) in the financial report.

We use a comprehensive dataset of 10-K and 10-Q filings for U.S. firms from the period 2018 to 2023. The sample contains 31,225 firm-quarter-filing-level observations. Our focus is on MD&A disclosure because, among all qualitative sections in the financial report, MD&A is a template-free section where managers have the largest discretion under safe harbor protection to disclose. Thus, if managers react strategically to the advent of LLMs by changing qualitative disclosure in financial reports, the MD&A section should be the one showing the greatest changes⁴. We perform textual analysis to decompose linguistic complexity into informative and obfuscating components (Bushee et al., 2018). Then, we exploit the plausibly exoge-

³Although managers can use LLMs for producing financial reports (Blankespoor et al., 2024b), based on the previous literature on readers' behaviors and strategic disclosure (Aghamolla & Smith, 2023; Blankespoor et al., 2020; Boone & White, 2015; Cao et al., 2023a; Cohen et al., 2020b; Maffett, 2012; Miller, 2010; Zhang, 2001), we hold the belief that a crucial reason for their post-shock report modification, no matter using LLMs or not, is out of strategic concerns.

⁴Our result is consistent after switching the sample to Risk Factor section. Results upon request.

nous shock of ChatGPT 3.5’s release to isolate the potential impact of LLMs on the information environment. This shock is effective magnitude-wise due to ChatGPT 3.5’s quick growth in public attention upon its release (Bertomeu et al., 2025; Hu & Hu, 2023), strong processing potential (Bilinski, 2024; A. Kim et al., 2023a, 2023b), and easy accessibility and interaction method (OpenAI, 2022). We measure linguistic complexity using the Bog and Fog Index and decompose these measures into information and obfuscation components (Bonsall et al., 2017; Bushee et al., 2018; Li, 2008). Our proxies for investor sophistication are the percentage of institutional ownership and stock internalization.

Our results show that the release of ChatGPT is associated with less information component of linguistic complexity. This indicates that managers believe investors can extract more information from their releases after the introduction of ChatGPT 3.5, and they reduce the information in these releases to maintain the level of information in the market constant at some optimal level. More importantly, the effect of our results is more pronounced for firms with less sophisticated investor bases. This is consistent with our argument that managers with a greater amount of less sophisticated investors are more concerned about the potentially narrowed gaps between more and LLM-equipped less sophisticated investors.

Additionally, we find that firms reduce the information component of linguistic complexity more when their institutional investors are more concentrated (i.e., the firm has a higher Herfindahl-Hirschman Index of institutional investor ownership stakes). This should allow the institutions with large stakes, greater influence, and easier access to inside information to maintain their informational advantage over institutions with smaller stakes. Compared to the previous analysis, here we narrow down our scope to institutional investors only. Even so, the implication that managers and sophisticated investors are aligned and opposed to the interests of less sophisticated investors still holds. We conclude that managers’ reaction, making disclosures less informative after the boom of LLMs, only protects the informational advantages of their most important sophisticated institutional investors.

We document a different pattern for the obfuscation component. On average, there is a statistically significant post-ChatGPT increase in obfuscation, but there

is no significance after we add investor sophistication. However, we find a positive association in the obfuscation component when firms have a large share of their equity in the hands of active institutional investors. Such investors make money by generating proprietary information and trading with uninformed, less sophisticated investors. They are likely to see through obfuscation and, therefore, higher obfuscation may serve as a signal that managers are willing to increase their information advantage over less sophisticated investors, even to a level surpassing the effect of LLM capability.

Our contributions are threefold. First, we contribute to the literature on how to retain optimal financing resources through balancing public and private information (Bertomeu et al., 2024; Bushman, 1991; Diamond & Verrecchia, 1991; Easley & O’hara, 2004; Zhang, 2001). When investors have a public tool to increase the level of informativeness, the advantage to the better-informed investors declines (Zhang, 2001). Consistent with this literature, we find that managers potentially preserve information advantages for influential, sophisticated investors despite widespread public availability of AI tools. This result empirically contributes to the literature by highlighting indigenous investor characteristics as another important factor for managers to balance transparency and financing costs (Aghamolla & Smith, 2023; Zhang, 2001). This balance influences information dissemination, investor behavior, and ultimately affects corporate resource retention and allocation.

Second, we connect to work on firms’ overall attitudes towards artificial intelligence reflected in corporate disclosure policies (Bertomeu et al., 2024; Blankespoor et al., 2024b; Cao et al., 2023a, 2023b; Miller, 2010). Recent studies already show the capital market consequences of LLMs and how managers change their voluntary disclosure and linguistic complexity (Bertomeu et al., 2024, 2025; Blankespoor et al., 2024b). Building on this, our paper digs deeper by looking into the cross-sectional variations in the investor base. Rather than inferring a single motive, we test predictions implied by convergence versus hysteresis in qualitative information. We show that managers’ responses vary systematically with the investor base, fund characteristics, and concentration. Furthermore, following Bushee et al. (2018), we transplant their proposed mechanisms to the financial reporting context and sepa-

rate information and obfuscation components from linguistic complexity. Our results extend the literature by showing that managers’ responses to the boom of LLMs are conservatively focused on protecting the interests of their core financiers.

Third, our paper contributes to the expanding literature on how the technological revolution in information acquisition, processing, and dissemination methods benefits less sophisticated investors (Blankespoor et al., 2019, 2020, 2024a, 2024b; A. Kim et al., 2023a, 2023b, 2024b; A. G. Kim et al., 2024). Prior work suggests that LLM technologies can be used to reduce information asymmetries (A. Kim et al., 2023a). We show that this effect is not unconditional. Firms adjust narrative transparency in ways that can offset potential convergence for less sophisticated investors. Our result extends the literature by showing that managers’ protection of sophisticated investors sacrifices less sophisticated investors’ benefits to take advantage of the leveled playing field, given their increased ability to process and analyze financial reports efficiently.

2 Literature review

2.1 Linguistic complexity and investor base

2.1.1 Managers’ motivations behind disclosure complexity

Over the past decades, the Securities Exchange Commission (SEC) has issued regulations to enhance investors’ comprehension of financial information. (SEC, 2017, 2020a, 2020b). It is important for regulators to understand the reasons behind firms presenting linguistically complex financial reports. Literature identifies two alternative reasons why managers can intentionally increase linguistic complexity in financial reports: to obfuscate readers or to provide additional, informative but complex details (R. Bloomfield, 2002, 2008; Bushee et al., 2018; Gosselin et al., 2021).

Obfuscation refers to managers’ attempts to reduce or postpone the market impact of bad news. It is backed by the incomplete revelation hypothesis (R. Bloomfield, 2002). Intentional obfuscation aims to increase disclosure processing costs, avert litigation, or attribute bad news to external factors to cloak poor management

(Blankespoor et al., 2020; R. Bloomfield, 2002, 2008). Since investors have limited attention and processing capability (Hirshleifer & Teoh, 2003), deliberate obfuscation can confuse market participants and postpone their trading activities (You & Zhang, 2009), while maintaining inflated impressions of the firm (García Osma & Guillamón-Saorín, 2011).

Empirical evidence below echoes this result. Bad performance or managed earnings is associated with less readable disclosure for investors, especially when there is a motive to present the firm as favorably as possible (Asay et al., 2018; Li, 2008, 2010; Lo et al., 2017). Managers also obfuscate investors from bad news by purposefully discussing good news with relatively lower linguistic complexity (Aghamolla & Smith, 2023; Bliss et al., 2018; R. Bloomfield, 2008). To strengthen the effect of obfuscation, when filing bad events, managers could issue unrelated press releases with noisy content or select a time slot where investors’ attention is relatively low (Rawson et al., 2023; Segal & Segal, 2016).

Nevertheless, linguistic complexity may not necessarily reflect intentional attempts to obfuscate (R. Bloomfield, 2008). As Bushee et al. (2018) argue, complex language can reflect the provision of intricate firm characteristics rather than obfuscation. They support the argument with the empirical results that information content is negatively associated with information asymmetry, yet obfuscation content is positively associated with information asymmetry (Bushee et al., 2018). Moreover, Mauritz et al. (2023) document that companies with the same audit partner exhibit similarities in language, content, and structure in narrative disclosure. Li (2010) discovers that firm size and history can also predict linguistic complexity.

2.1.2 Disclosure processing costs and investors’ sophistication level

According to Blankespoor et al. (2020), disclosure processing costs encompass three subdimensions: (i) awareness costs, to increase one’s probability of knowing the existence of a specific disclosure; (ii) acquisition costs, to gain and quantify a disclosure signal for valuation; and (iii) integration costs, to combine and refine acquired signals into an estimate and, finally, a decision. Differences across investor types re-

garding each subdimension are reflected in differences in investors' decision-making procedures. Thus, the consequences financial reporting complexity has on investors' decision making depend on how costly it is for them to process disclosures (Tan et al., 2014).

Information processing costs are higher for less sophisticated investors. Lee (2012) investigates three specific practical skills that can reduce the costs of information processing, but which less sophisticated investors generally lack. Firstly, a sense of stock value precision, which makes investors less likely to be misled and overprice the stock. Secondly, a capability to deal with large amounts of information and filter out the less salient pieces. Thirdly, a direct and efficient searching pattern. Given these skill gaps, the higher disclosure processing costs of less sophisticated investors reduce their trading activities and make their interpretations more diverse than their counterparts, which can lead to higher volatility of stocks with a less sophisticated investors base (Loughran & McDonald, 2011; Miller, 2010; You & Zhang, 2009). Furthermore, Rennekamp (2012) and Tan et al. (2014) document that not only do the disclosure processing costs limit retail investors' abilities, but the existence of such costs per se may influence investors' subconscious feelings about how easy it is to process the information, and their subsequent decision to rely on it. Lawrence (2013) considers low readability and excessive length "serious barriers" for less sophisticated investors when extracting relevant information. Consistently, Lawrence (2013) and Zhang (2001) present evidence showing that less sophisticated investors benefit more from higher disclosure quality.

Empirical research demonstrates consistent results. You and Zhang (2009) document investors' sluggish reactions after the release of 10-K, and such effect is stronger when firms have complex 10-K reports. Miller (2010) finds that complexity has a negative effect on investors' overall trading, and small investors interpret the information more diversely when reports are more complex. Lee (2012) discovers that stock prices reflect less information related to earnings for those firm-quarters that present unexpectedly longer and more complex quarterly reports. Rennekamp (2012) finds that individuals tend to have more favorable judgments when the news is good and more negative judgments when the news is bad. Lawrence (2013) finds positive

associations between concise disclosure and individuals' investment amount and return, but the former effect is less pronounced for more intense high-frequency trading activities and more financially literate individuals. Tan et al. (2014) document that less sophisticated investors' earnings judgments are susceptible to positive framing effects, even if the positive language may be inconsistent with managers' negative implications. Nevertheless, sophisticated investors consider positive language less credible. Asay et al. (2017) show that when investors encounter complex reports, their sensitivity to incorporate outside information increases (especially for analysts' reports), such that they feel more comfortable evaluating the firm, but they might over-rely on more readable disclosure. Ben-Raphael et al. (2022) demonstrate that retail attentions on 8-K spike only if there is abnormal traditional media coverage, while sophisticated investors trade during the pre-filing period of 8-K.

2.1.3 Optimal disclosure policy depending on the firms' investors base

Overall, the previous discussion indicates that (i) managers choose the level of linguistic complexity and its informativeness and obfuscation components, and (ii) sophisticated investors can benefit more from more complex disclosures because they can extract more information while being less affected by obfuscation.

Most literature that examines the implications of disclosures for specific groups of investors, such as individual investors (Lawrence, 2013) or institutional investors (Bushee & Noe, 2000; Healy & Palepu, 2001) only focuses on one investor group at a time. For example, seminal work by Bushee & Noe (2000) indicates that sophisticated investors invest more in firms with better public disclosures.⁵ However, this result does not tell us whether institutional investors invest more in firms with more public disclosure either because they prefer a higher level of disclosure or because public disclosures increase their information advantage and trading profits vis-à-vis individual investors. Some recent papers (Anchev et al., 2022; Kalay, 2015) have studied this issue. They found that when firms increase public disclosure, they change their investor base towards more sophisticated investors, consistent with the hypothesis

⁵Although the effect is more pronounced for active investors and quasi-indexers than for dedicated investors, since the latter probably do not need to rely on public disclosures

that firm disclosures increase the information asymmetries between investors who differ in their ability to bear the costs of processing such disclosures.

The remaining question is why firms prefer to cater to more/less sophisticated investors. Interestingly, an investor base with more retail investors may increase liquidity (Balakrishnan et al., 2014; Diamond & Verrecchia, 1991). Empirical evidence consistent with this idea is provided by Blankespoor et al. (2014) who show that firms’ liquidity improves and firms experience more retail trading after using social media such as Twitter for information dissemination. On the other hand, risk-averse managers want to minimize pricing errors (Bertomeu et al., 2023) and can view retail investors as noise traders whose activity will increase stock price volatility and even increase firms’ cost of capital (De Bondt, 1998; De Long et al., 1990; Eaton et al., 2025). In fact, prior research shows that more trading by retail investors increases return volatility (Barber & Odean, 2000; Barber et al., 2009, 2022; De Long et al., 1990). Moreover, even if managers do not have specific preferences regarding the investors’ base, the pressure of sophisticated investors will align the firms’ disclosure policies with the preferences of more sophisticated investors for greater disclosure through information channels that are more accessible to them than to retail investors. For instance, empirical evidence from Zheng et al. (2023) shows that managers counteract the increase in retail investors’ trading interest caused by stay-at-home policies during the COVID pandemic by reducing the information they provided via social media (which is easily accessed by non-sophisticated investors) but not changing information channels that are more appealing to sophisticated investors, such as 8-K filings or earnings guidance.

2.2 Artificial intelligence and disclosure

The emergence of large language models (LLMs), such as ChatGPT, marks a significant shift in how investors process financial disclosures. LLMs are advanced computational systems capable of analyzing and generating human-like text with efficiency and accuracy. For example, ChatGPT 3.5, launched by OpenAI in November 2022, reached 100 million users in two months, surpassing the adoption rates of social me-

dia platforms like TikTok and Instagram (Hu & Hu, 2023; Singh, 2024). ChatGPT is trained with diverse and extensive text from the internet, including books, articles, websites, and other written content (OpenAI, 2022). This helps the model learn a wide range of language patterns, facts, and contexts, including knowledge from finance and accounting (Talerico, 2024).

Early research demonstrates the potential of ChatGPT 3.5 to reduce information asymmetry, allowing less sophisticated investors to analyze financial reports more effectively (Bilinski, 2024; Hansen & Kazinnik, 2024; A. Kim et al., 2023a). Moreover, it provides accessible and accurate interpretations and predictions of financial data (AIContentfy, 2023; Bilinski, 2024; A. Kim et al., 2024b). It can also improve the screening of misinformation, as market reactions to misstated earnings and to COVID-19 fake-news shocks are muted when AI-adopting institutions hold more of the stock (J. M. Kim & Leone, 2025). These functions have the potential to narrow the skill gaps of financial statement analysis between retail and institutional investors (Cheng et al., 2025; Guo et al., 2022; A. Kim et al., 2023a). Our aim in this paper is to investigate whether and how the potential democratization brought about by the introduction of advanced textual analysis tools through LLMs influences managerial disclosure strategies and how managers' reactions to the new tools depend on the level of sophistication of the firm's investor base. With this study, we contribute to a growing academic literature that studies the relationship between artificial intelligence and firm disclosures that can be classified into four different research lines.

2.2.1 Artificial intelligence and measures of disclosure quality

The first line of research into the use of LLMs in the analysis of financial reports uses artificial intelligence to build measures of language complexity. Bilinski (2024) documents that sentiment and linguistic complexity both contain value-relevant information. Especially, the study creates measurements of complexity and sentiment scores from ChatGPT 4. They are associated with price reactions to report publication, profitability, and investor beliefs in a way that is consistent with a variety of

previous literature (Bilinski, 2024). Kim et al. (2023a) define "bloated disclosure" as the relative amount of a document's length that can be reduced by ChatGPT 3.5, measuring the degree of redundant or irrelevant information. Consistent with the obfuscation theory (R. Bloomfield, 2008; Li, 2008; Lo et al., 2017), their findings reveal that "bloated disclosure" is higher when a firm "reports losses, has negative sentiment, and experiences negative stock market reactions", and is associated with "adverse capital market consequences". Emerging research also proposes measures of linguistic or business complexity with ChatGPT and compares them with previous methodologies (Bernard et al., 2023; Zang et al., 2024).

2.2.2 Artificial intelligence adoption by investors

A second research stream studies whether investors make use of LLMs for the analysis of financial reports. Recent evidence shows that AI tools, especially LLM applications, were taken up quickly by investors and are already used to process qualitative disclosures. In a 2023 global survey of 130 systematic investors managing 22.5 trillion dollars, about half reported using AI in the investment process, with many deploying NLP for sentiment analysis and for translating earnings calls and reports. Respondents also describe using ChatGPT internally to summarize research and tailor communications to different audiences⁶. Complementing this survey evidence, finance appears to be an early AI adopter, as institutional "AI ownership" (ownership by institutions recruiting AI-skilled staff) rose from roughly 5% in 2012 to 20% by 2022 (J. M. Kim & Leone, 2025).

Academic literature also demonstrates evidence that retail investors are using ChatGPT. For instance, Cheng et al. (2025) find that after ChatGPT outages, retail trading volume significantly declines. Interestingly, the magnitude of the drop is smaller than for non-retail trading volume. This seems to imply that both sophisticated and retail investors use ChatGPT to analyze financial information to varying extents. Guo et al. (2022) investigate a recent Chinese retail-targeted application with LLM functions. They find that the use of LLMs positively influences investment

⁶See 2023 Invesco Global Systematic Investing Study (<https://www.invesco.com/content/dam/invesco/apac/en/pdf/insights/2023/october/IGSIS-FullStudy-APAC.pdf>)

decisions, redemption activities, and overall returns. The effect is more pronounced for more experienced retail investors. Most relevant to our argument, Blankespoor et al. (2024b) survey 2,175 retail investors. They find that half of them use LLMs, while those more sophisticated ones "understand and leverage LLMs' strengths to a greater extent." They know how to use it to gather and process complex information from broad sources and handle difficult tasks quickly and easily.

2.2.3 Which type of investors benefits more from artificial intelligence?

A key question here is which type of investor benefits more from the new technologies that reduce the costs of analyzing financial information. This is important because sophisticated investors' ability to offer above-market returns critically depends on their information advantage over their counterparts. Bertomeu et al. (2025) use Italy's temporary ChatGPT ban as a natural shock to examine the economic impact of LLMs on the capital market. For information users, they find that firms with fewer sophisticated investors and lower analyst coverage experience a more significant increase in bid-ask spread during the ban period. This implies that ChatGPT offers "different investors, especially small local investors, a conversational and interactive platform to gather insights and navigate the complexities of financial markets," and it is being used by retail investors. Cheng et al. (2025) find that when ChatGPT encounters technical issues, the trading volumes of both retail and institutional investors plummet.

If this effect is more important for less sophisticated investors, one would expect their trading behavior to converge to the trading of sophisticated investors, reducing information asymmetry and the informational advantage gap. For example, the implementation of the EDGAR system disseminates corporate disclosure through the Internet. While this may not be relevant for sophisticated investors that can access corporate disclosures through data providers, it can be very relevant for individual, less sophisticated investors, who can now get more information about future stock returns (M. Gao & Huang, 2020). Likewise, the XBRL adoption from the SEC can level the playing field by reducing investors' data aggregation costs for retail investors.

Smaller institutions benefit more than the larger ones from technical and financial competencies (Bhattacharya et al., 2018), supporting the convergence argument.

Furthermore, ChatGPT can be important for convergence because less sophisticated investors can also benefit from the fast-evolving online resources that help process financial reports with ChatGPT. For example, social media users are sharing prompts for analyzing quantitative and qualitative data in financial reports and trading strategies (BH.Falcon27, 2023; Starks-Technology, 2024). Also, a bunch of emerging websites provide LLM users with specific prompt instructions to use ChatGPT for efficiently reading financial reports (Ahmad, 2023; Arshad, 2023; Talerico, 2024). Custom GPTs lower the cost of running qualitative analyses, and domain-tuned models now achieve strong performance—for example, Kim et al. (A. Kim et al., 2024b) on earnings-change prediction and trading performance, and evidence of an AI analyst outperforming most human analysts in return prediction. Furthermore, their tailored GPT model provides trading strategies with higher Jensen’s Alphas and Sharpe ratios than other machine-learning strategies (A. Kim et al., 2024b). Collectively speaking, researchers and social media users are developing and sharing online prompts of ChatGPT for financial report analysis, accelerating the progress of proper LLM adoptions for general investors.

Among all the previous technologies that can help this convergence, such as the introduction of Internet, Google Search, XBRL, or deep learning tools, LLMs such as ChatGPT have some unique characteristics that can promote convergence. Firstly, not only does ChatGPT’s dictionary-level multi-language support reduce users’ language barriers, it also creates, detects, and analyzes human-like tones and narratives (Ahmad, 2023). Secondly, its swift, code-attached response saves users’ time in creating algorithms, reading and analyzing financial reports more efficiently, meanwhile reducing the financial literacy threshold and thus users’ learning costs (Helter, 2023). Thirdly, while investment software and AI tools are well-known inside the investment circle, GPT’s easy accessibility oversteps such a circle and gains worldwide popularity. In other words, domestic markets can increase participation from foreign new-coming investors at every knowledge level. For instance, a foreign investor with no knowledge of the country’s language can now easily analyze a domestic language

financial report. A person without financial knowledge can improve financial literacy and invest merely with ChatGPT’s advice. Recent literature also documents ChatGPT’s usefulness in providing better portfolio recommendations for retail investors. (Oehler & Horn, 2024),

2.2.4 Artificial intelligence and disclosure strategy

The third line of research is directly related to our work and studies how artificial intelligence has changed managers’ disclosure strategy. Cao et al. (2023a) investigate disclosure tonality and find that managers seem to be aware of the ease with which readers’ automated programs process and parse firms’ disclosures and analyze tone. Therefore, firms avoid using negative words perceived by algorithms, but such an effect is moderated by imminent external financing needs and/or higher litigation risk (Cao et al., 2023a). They also argue that tonality will quickly lose relevance if firms can use positive language without costs to impress readers. A more recent study from Cao et al. (2023b) concludes that the existence of machine readers reduces financial misreporting occurrences. This effect is more pronounced when a report is complex or when alternative data is available.

Our work builds on all these literature streams by investigating how managers’ reactions to technology that reduces disclosure processing costs may depend on their main investor base.

3 Hypotheses development

From the literature review, we can conclude that the new technologies that help investors analyze financial reports have been adopted by all types of investors, and managers are aware of that. Moreover, less sophisticated investors seem to have benefited more from the reduction in information processing costs brought about by these new technologies (Lawrence, 2013; Zhang, 2001). How do managers perceive and react to this new informational environment where all investors are better and more similarly informed? How do they manage linguistic complexity?

Previous literature shows that when mortgage disclosures are simplified, it becomes a less attractive investment opportunity to banks because (i) consumers' improved understanding of mortgage info reduces the interest and fees that banks are able to charge on mortgages, and (ii) banks may face challenges in selling mortgages on the secondary market (Nicoletti & Zhu, 2023). Similarly, we argue managers are concerned that average investors' improved efficiency of processing disclosure reduces the profits that sophisticated investors are able to obtain from private information, as it becomes harder to monetize information asymmetry in the equity market (Maffett, 2012). To maintain trading benefits, sophisticated investors are likely to switch to similar firms where the information barriers hold. Managers may be concerned about this potential change to their investor base either because (i) sophisticated investors add value to the firm through their monitoring activities, (ii) sophisticated investors are active investors that vote and can discipline the manager, or (iii) simply because managers do not like the higher stock volatility suffered by firms with more unsophisticated investors in their investors base. Whatever the cause, managers will try to maintain the information barrier to prevent the firm's sophisticated investors from leaving for another firm with higher information opacity. Eventually, all managers have to engage in more opaque disclosure practices, attempting to roll the information environment back to the situation where sophisticated investors can benefit from trading with less informed investors. To increase the information opacity, managers need to decrease the release of information.

Interestingly, this effect should be more pronounced for firms with a greater amount of less-sophisticated investors. When ownership is very concentrated, liquidity is limited and trading is not an important source of benefits for sophisticated investors. In Appendix C we develop a simple theory model that illustrates our argument, showing that it can be optimal to limit disclosure when the presence of an active shareholder can increase firm value, even if this means that uninformed investors will lose money when trading with the better informed active investor.

Additionally, managers may also increase the level of obfuscation. It is assumed in the literature that sophisticated investors can see through firms' attempts to obfuscate them with linguistic complexity (Blankespoor et al., 2020; M. Bloomfield,

2021; Hirshleifer & Teoh, 2003; You & Zhang, 2009). So obfuscation is targeted to non-sophisticated investors and can be used to keep the information barrier between sophisticated and non-sophisticated investors (Maffett, 2012). Nevertheless, after the release of ChatGPT, attempts to obfuscate retail investors lose effectiveness. As modifying financial reports is costly (Cohen et al., 2020a), we do not believe obfuscation will increase. Therefore, we list our hypotheses as follows:

H1: *Firms with a higher percentage of less sophisticated investors in their investor base decrease the information component of linguistic complexity more.*

H2: *Firms with a higher percentage of less sophisticated investors in their investor base do not change the obfuscation component of linguistic complexity more after the release of ChatGPT 3.5.*

So far, we have considered differences in the use of ChatGPT between more and less sophisticated investors, but not all sophisticated investors are equal in terms of their stake in the firm. As explained above, there may be several reasons why managers may change information disclosure in line with the preferences of more sophisticated investors. Nevertheless, this effect is likely to be more important when their stake is more concentrated. Large institutional investors will have direct contact with the managers, are likely to sit on the board of directors, and they will not trade often. Moreover, different from less sophisticated investors, but also from sophisticated investors with small stakes, they will not obtain their information from public reports. Large institutional investors are dedicated, long-term investors that usually maintain large stakes in a few firms and exert voice to bring about changes in strategy to align managers actions to their preferences (Bushee & Noe, 2000). Therefore, if managers adapt their disclosure policies to the preferences of their more influential investors, to maintain the informational advantage of these large investors over the rest of the investors, firms where institutional ownership is more concentrated should release less information after the introduction of ChatGPT.

H3: *Firms with a higher concentration of institutional investors in their investor*

base will increase information opacity after the release of ChatGPT 3.5.

4 Empirical design

4.1 Disclosure channel selection

We study the effect of investor base sophistication on the level of linguistic complexity used by the managers in the firm’s EDGAR Form 10-K and 10-Q filings. We focus on these specific disclosure channels because both forms are essential for investors, as they provide regular updates on a company’s financial health and operations, helping them make informed investment decisions. Moreover, both forms offer a degree of flexibility that managers can use to make these disclosures more or less informative. Our analyses focus specifically on the management discussion and analysis (MD&A) section in these financial reports. It is important to acknowledge that the way qualitative texts are written is revised by the SEC (e.g., PWC, 2024), and does not easily change from period to period (Cohen et al., 2020b). However, compared with other qualitative sections, managers have the highest degree of freedom in the MD&A section to tweak linguistic complexity because (i) MD&A is defined by the SEC to present results ”through the eyes of management” without a clear template; and (ii) MD&A allows forward-looking discussion under safe-harbor protection, giving leeway in tone and content.

4.2 Data and sample construction

We retrieve 10-K and 10-Q financial reports and tags (Hoitash & Hoitash, 2018) from SEC EDGAR, accounting quarterly fundamentals data from Compustat, institutional ownership data from Refinitiv, daily stock returns data from CRSP, and analyst summary data from I/B/E/S. All data spans from 2018 to 2023.

We merge the Compustat Quarterly fundamentals with the textually analyzed dataset. After that, we left-joined the dataset from Hoitash and Hoitash (2018),

Compustat Annual Fundamentals, CRSP stock daily, FINRA OTC trading volume, and Analyst forecast dataset. Since the dataset of analyst forecasts is monthly, we merge by the first monthly observation with the analysts’ next-period forecast after the public earnings announcement release of the last period.⁷ Then, to obtain a balanced panel dataset, we delete all observations with missing or NA variables and winsorize the dataset at a level between 1 and 99 percent. The final sample includes 3,629 firms with 31,225 filing-year-quarter level observations. Appendix A shows the sample construction process in detail.

4.3 Investor sophistication measurement

We use two proxies for investor sophistication measurements: institutional ownership and internalization⁸ (García-Méndez, 2023). Institutional ownership serves as a critical indicator because it reflects the level of professionalism and expertise within the investor base. Obviously, a higher proportion of institutional investors correlates with more sophisticated decision-making processes for both investors and managers.

Internalization is an over-the-counter activity. It happens when a brokerage firm trades securities with clients from its own inventory rather than routing clients’ orders to an external exchange⁹. A high level of internalized trades can indicate significant retail activities, as these trades are more common among retail investors compared to sophisticated investors, who typically trade in larger volumes on public exchanges. Internalization of retail investors currently represents about 16% of the total trading volume in the consolidated US equity market (Comerton-Forde et al., 2018), which

⁷Although this might underestimate the number of analysts, we argue that the next-period forecasts on such observations attribute the most to the latest public earnings release.

⁸We do not include corporate investors due to data limitations. Corporate investors typically constitute approximately 10-15% of ownership stakes. They may exhibit qualitatively distinct behaviors due to strategic rather than purely informational or financial motives.

⁹Brokerage firms in the US must report their trading to the Financial Industry Regulatory Agency (FINRA). The association makes its trading records publicly available through its official website. In their dataset, each observation shows the volume executed by brokers of a specific firm on a specific date. We clean up the dataset by summing up the trading volume in quarters. Then, we merge it with the stock volume retrieved from CRSP. The internalization level is calculated as the quarterly internalized trading volume of a firm divided by the quarterly volume of CRSP stock trading firms.

indicates its materiality in overall market dynamics.

4.4 Linguistic complexity measurements and decomposition

For the textual analysis, we choose the Fog Index in the bag-of-words method as the main measurement of linguistic complexity because of its ease of replicability ¹⁰. Fog index was used to measure firm obfuscations when first introduced in the accounting literature (Li, 2008). It estimates the number of years of education required for a person to understand the text. The calculation is as follows:

$$FogIndex = 0.4[(\frac{\#Words}{\#Sentences}) + 100(\frac{\#ComplexWords}{\#Words})] \quad (1)$$

Where $\#Words$, $\#Sentences$, and $\#ComplexWords$ mean the count number of *Words*, *Sentences*, and *ComplexWords*. Moreover, *ComplexWords* means words with more than three syllables.

We decompose the linguistic complexity of the report into information and obfuscation components (Bushee & Huang, 2024; Bushee et al., 2018). This method is widely applied in the recent accounting literature (Bushee & Huang, 2024; Mekhaimer et al., 2024). The information component refers to texts that provide additional information that readers can use to make decisions. The obfuscation component refers to texts that are included at the discretion of managers and do not provide additional information for report readers to make decisions. Bushee et al. (2018) provide comprehensive steps on decomposing linguistic complexity. We follow their structure while accommodating the specifications into our settings. The reason is that we use financial reporting instead of a conference call as the setting. In particular, we regress linguistic complexity measurements on variables that are exogenous to the firm’s reporting decisions. The decomposition model is specified as follows:

¹⁰We also include other popular measurements of bag-of-words method in the robustness check.

$$Fog_{MD\&A} = \theta_0 + \theta_1 BUSSEG + \theta_2 GEOSEG + \theta_4 FirmAge + IndustryFE + \delta \quad (2)$$

Where the $Fog_{MD\&A}$ is the Fog Index of the MD&A section. $BUSSEG$ is the number of business segments. $GEOSEG$ is the number of geographic segments. $FirmAge$ is the age of the firm. These business complexity variables are not identical to Loughran and McDonald (2024) or Hoitash and Hoitash to measure business complexity (2018) because we restrict the variables to those whose measurements cannot be manipulated by managers. Following Bushee et al. (2018), we define the fitted value estimated from the regression as the information component, and the residual from the regression as the obfuscation component. Next, we estimate equation (3) using these components as dependent variables. Table 2 provides the detailed decomposition results.

4.5 Main Model

Our empirical strategy exploits the common timing of the late-2022 ChatGPT release and cross-sectional variation in investor base composition. We estimate a heterogeneous response model with firm and time fixed effects, comparing each firm’s post-shock outcomes to its own pre-shock history, and allowing the post-shock change to vary with the investor base (e.g., institutional ownership, concentration, active share). The model is specified as follows:

$$Fog_{Component} = \beta_0 + \beta_1 PostGPT_{2022Q3} + \beta_2 InvBase + \beta_3 PostGPT_{2022Q3} * InvBase + \sum_{n=4}^n \beta_n Controls + FirmFE + QuarterFE + \epsilon \quad (3)$$

The dependent variable $Fog_{Component}$ measures, in different specifications, the information component and the obfuscation component of linguistic complexity.

Among the independent variables, *PostGPT* is a calendar time dummy variable equal to 1 if a report is published 1 period after the release of ChatGPT 3.5 (November 2022), and 0 otherwise. *InvBase* is either a dummy or continuous variable that proxies for the firm’s investor base characteristics. We use alternative definitions of investor sophistication in different estimations to capture different characteristics of the firm’s investor base as measured at the time of the release of ChatGPT. We first use *InstOwn*, the institutional ownership of the firm. We also use the dummy *InstHHI*, the Herfindahl-Hirschman index (i.e., concentration) of institutional investors.

We follow Bernard et al. (2023) to construct control variables at the firm-quarter level. *is10K* is a dummy variable equal to one if the filing is a 10-K and zero otherwise. *Size* is the total assets of the firm. *Stdev* is the standard deviation of analyst estimates. *AnalystFollow* is the natural log of one plus the number of analysts for a specific quarter. *Leverage* is the leverage ratio of the firm. *BM* is the book-to-market ratio. *Capex* is the capital expenditures divided by sales. *Loss* is a dummy variable equal to one if the firm’s earnings before extraordinary items are negative, and zero otherwise. *R&D* is the R&D expense over sales. *RecentM&A* is a dummy variable equal to one if the firm reported acquisition charges in the current or prior two fiscal quarters, and zero otherwise. *ROA* is the return on assets. *LargeAccFiler* is a dummy variable equal to one if the firm is a large accelerated filer at that time and zero otherwise. *NonAccFiler* is a dummy variable equal to one if the firm is not an accelerated filer at that time and zero otherwise.¹¹ To address time and cross-sectionally dependent residual correlations, we apply firm and year-quarter double-clustering model on our results (Petersen, 2009).

¹¹Appendix B contains all variables described above.

5 Results

5.1 Summary statistics

Table 1 shows the descriptive statistics. From 2018 to 2023, firms in our sample have a sophisticated investors base, with an 80% median value of institutional investors' ownership. Regarding other variables, firm use on average 276 distinct XBRL tags in their financial reports. They are followed by approximately 2 analysts. Retail trading volumes of these firms are about 20 percent of the total trades. Firms have an average ROA of -0.01, an average book-to-market ratio of 0.54, an average leverage ratio of 0.29, an average Capex-over-sales ratio of 0.20, and an average R&D-over-sales ratio of 0.14. Approximately 33% of observations report a loss in the previous period. An average firm has approximately 2 business segments and 2 geographic segments. We find that 15% of the sample firms are non-accelerated filers and 65% of the sample firms are large accelerated filers. Only 5% of the sample firms have recently performed mergers or acquisitions. Tables 1a and 1b demonstrate the correlation matrix of our variables. Our model is free from multicollinearity problems.

Table 2 shows the result of the decomposition of the linguistic complexity. Panel A in Table 2 demonstrates the impact of the number of business segments, the age of the firm, the geographic segment, and the size on linguistic complexity. Panel B in Table 2 shows the descriptive statistics of the decomposed variables.

5.2 Post-GPT changes on information components of linguistic complexity

Table 3 shows the OLS panel regression results of changes to the information component of linguistic complexity after the release of ChatGPT 3.5. *PostGPT* equals one for observations at least one quarter after the release of ChatGPT, and zero otherwise. Column (1) to (3) presents the result of the information component change post-ChatGPT 3.5. The coefficients of *PostGPT* are overall significantly negative, suggesting that the information component of linguistic complexity decreases after the release of ChatGPT 3.5. This result is consistent with the conclusions from

the recent literature, reflecting the idea that managers want to maintain an optimal level of information disclosure (Bertomeu et al., 2024; Blankespoor et al., 2024b). In doing so, they balance the advantages of offering more information (higher liquidity and lower cost of capital) with the cost of giving away more information to competitors. The release of ChatGPT increases the information that can be extracted from the firms' reports and induces managers to decrease the information component of linguistic complexity so as to return to the optimal disclosure level.

We are particularly interested in differences depending on the sophistication of the investor base, which is captured by the $InstOwn \times PostGPT$ interaction variable. Columns (3) and (4) in Table 3 test the effect of institutional ownership on the information component of linguistic complexity before and after the release of ChatGPT. The coefficients of the interaction term are significantly positive. The magnitude of the coefficient in column (4) is approximately 1.77% of the standard deviation of the information component of the Bog Index. This indicates that, while on average companies reduce the information component of linguistic complexity after the introduction of ChatGPT, the decrease is more (less) pronounced in firms with a (lower) higher percentage of less sophisticated investors. Therefore, we confirm Hypothesis 1 that managers believe the release of ChatGPT leads to convergence between more and less sophisticated investors. This could lead to a reduction in the trading gains that sophisticated investors can obtain from uninformed investors. And these gains are more important when the percentage of ownership in the hands of less sophisticated investors is higher and there is more market liquidity. Thus, managers of firms with a larger percentage of less sophisticated investors react more, lowering the information component to bring back total information to its previous level.

We repeat the analysis in Table 4 for the obfuscation component. The coefficients of the time dummy ($PostGPT$) and the interaction term ($InstOwn \times PostGPT$) are insignificant. This is consistent with our hypothesis that the obfuscation component of linguistic complexity does not differ among investor compositions. It implies that managers' obfuscation signal to keep sophisticated investors is irrespective of the percentage of less sophisticated investors in their base. Sophisticated investors are able to see through obfuscation, and, after the release of ChatGPT, this is also likely

to be true for less sophisticated investors. Firms do not gain by modifying obfuscation, irrespective of their investor base, with the cost of text stickiness. Therefore, this result can only imply that managers’ obfuscation signal to keep sophisticated investors is irrespective of the percentage of less sophisticated investors in their base.

5.3 Post-shock changes in linguistic complexity as a function of investor concentration

We now want to study the differences between different types of sophisticated investors. To do this, we first look at the concentration of institutional investors. This is interesting because the importance of reports as a guide for investment and trading decisions is not the same for all institutional investors. Institutional investors that keep a large stake in the firm are likely to be long-term investors who do not trade often and have direct access to managers, particularly through positions in the board of directors. This means that, if managers favor the informational preferences of their large investors, they will decrease the information component to a greater extent the higher the concentration of institutional owners, as stated in Hypothesis 3.

Table 5 presents the analysis of how institutional investor concentration, measured by the Herfindahl–Hirschman Index (*InstHHI*), influences firms’ changes in the linguistic complexity of their MD&A disclosures following the release of ChatGPT. The dependent variables separately measure the information and component of linguistic complexity using the Fog Index measures. The primary explanatory variable is the interaction term *InstOwn* × *PostGPT*, capturing the differential effect of institutional investor concentration after the introduction of ChatGPT.

The findings indicate a statistically significant negative relationship between institutional investor concentration and the information component of linguistic complexity. In other words, after ChatGPT’s introduction, firms with higher institutional investor concentration significantly reduce the informational clarity of their disclosures compared to firms with less concentrated institutional ownership. The effect is around 15% of the information component standard deviation. This result supports

Hypothesis 3, which argues that firms with more concentrated institutional investors prefer to limit the public information content of disclosures, thereby preserving the informational advantage enjoyed by these large investors who likely have privileged access to private information channels.

Conversely, when studying the obfuscation component of linguistic complexity, the coefficient for $InstHHI \times PostGPT$ is significantly positive. This suggests that firms with concentrated institutional investors significantly increase their level of obfuscation after ChatGPT’s release. Given that LLMs make complex disclosures easier to interpret, managers at these firms prioritize sending signals to leading institutional investor(s) to keeping sticky obfuscation content. Overall, this table indicates that institutional investor concentration influences managerial disclosure choices through keeping an opaque information environment. Managers appear strategically motivated to protect the informational benefits for large institutional investors following the introduction of AI-based analytical tools such as ChatGPT.

6 Robustness checks and additional analyses

6.1 Alternative Proxies

6.1.1 Bog Index and file size as alternative proxies for linguistic complexity

Loughran and McDonald (2014, 2016) criticize the use of the Fog Index to measure reports because its algorithm defines multi-syllabic words frequently used in business as linguistically complex. Therefore, among different bag-of-words measurements for linguistic complexity, we use the most advanced one, Bog Index, for the robustness analysis of linguistic complexity measurement (Bonsall et al., 2017). The Bog Index is a state-of-the-art proxy that captures most aspects mentioned in the SEC Plain English Handbook, such as sentence length, passive voice, weak verbs, overused words, complex words, and jargon (SEC, 1998). Born from the Fog Index, it penalizes multi-syllabic words less severely and addresses more concerns from the SEC Plain English Handbook (Bonsall et al., 2017). The Fog Index is still widely used in the

studies of conference call settings since in spoken English multi-syllabic words are less frequent (Bushee & Huang, 2024; Bushee et al., 2018; Mekhaimer et al., 2024). The Bog Index is calculated as follows:

$$BogIndex = SentenceBog + WordBog - Pep \quad (4)$$

There are three components on the right-hand side of the equation. The first component, *Sentence Bog*, captures the linguistic complexity originating from sentence length. Longer sentences signify higher linguistic complexity. The second component, *Word Bog*, comprises two subcomponents. The first refers to the style problems highlighted in the SEC’s Plain English Handbook¹². The second is the word difficulty for general vocabulary, abbreviations, and specialist terms. The last component in (1), *Pep*, captures the attributes that facilitate the reader’s understanding of the texts¹³. Since the lexicon of the Bog Index is not completely disclosed to the public, we purchased the 2024 professional version of StyleWriter 4. We calculate the Bog Index by loading the MD&A part of the financial reports in our sample into the software.

Our second measure is the file size (Loughran & McDonald, 2014). We use it due to two reasons: (i) conceptually, file size captures the ”effective communication of valuation-relevant information” (Loughran & McDonald, 2014). If managers intend to obscure mandated information, they would likely “bury the awkward revelation in an overwhelming amount of uninformative text and data” rather than merely inflate syllable counts (Loughran & McDonald, 2014). File size can therefore reflect whether management is adding excessive bulk and complexity, which in turn complicates the investor’s or analyst’s task of assessing the firm’s value. (ii) Empirically, file size in bytes of the SEC filings bypasses parsing, which creates a straightforward way for

¹²It includes passive verbs, hidden verbs, overwriting, legal terms, clichés, abstract words, and wordy phrases (Bonsall et al., 2017).

¹³Pep is calculated as follows: $Pep = \frac{\sum c_i \times 25}{W + \frac{10\sigma_\ell}{\bar{\ell}}}$. Where c_i is the count of i th interesting word,

W is the total number of words, σ_ℓ is the standard deviation of sentence lengths, and $\bar{\ell}$ is the average sentence length (Bonsall et al., 2017)

researchers and regulators to avoid manual or algorithmic errors arising from large chunks of text.

Table 6 presents the result. It shows that the time dummy *PostGPT* is significantly negative, meaning there is a decrease in information component after the release of ChatGPT, but there is no significant change on the obfuscation component. As for the interaction term, one standard deviation of institutional ownership increases the information component of linguistic complexity more after the release of ChatGPT than before. Overall, the result is consistent with our previous analysis using information components of the Fog Index as the main dependent variable.

6.1.2 Non-decomposition proxy

One might question the validity of our result due to the recent criticisms about using decompositions in regressions (Chen et al., 2018, 2022). To address the concern, we include the decomposition variables as controls and rerun the main analysis using firms with high and low business complexity. We define high and low business complexity as the above-/below-mean fitted value from our decomposition result. As the decomposition variables are included, we proxy the information component as the Fog and Bog index of the MD&A section.

Table 7 shows the overall result. The table indicates that after the release of ChatGPT 3.5, the overall value of linguistic complexity increases on subsamples with high rather than low business complexity (BC). As obfuscation is useless, the only possibility to change the linguistic complexity is changes in information. Combined together, these results imply that managers' level of concerns on non-sophisticated investors differ on their BC, and the ones with higher BC change the amount of information released regarding their investor base after the release of ChatGPT 3.5. Such implications are consistent with those drawn from our main analysis.

6.1.3 Internalization as an alternative proxy for investors' sophistication

Throughout the analysis, we have used the percentage ownership of institutional investors as our measure of investors' sophistication. However, this is not a perfect

proxy because it leaves out non-institutional large owners (such as families or firms) which are also considered to be sophisticated investors. To ensure our measurement of investors' sophistication is robust, we repeat the analysis considering internalization, as defined in Section 4.3, as an alternative to institutional ownership. Internalization is trying to capture the importance of less sophisticated investors in the investor base. Therefore, we expect the effect of internalization on linguistic complexity to be opposite to the effect of institutional ownership, as internalization proxies retail investors' interests in the firm.

Table 8 shows the results. The dependent variable measures the information component of the *BogIndex(Bog-Info)* or *FogIndex(Fog-Info)*, and the independent variable is the estimator. The independent variable *Internalization* equals the percentage of internalization level of the stock. Overall, the magnitude and significance of the robustness check are consistent with the results in Table 4. This confirms that after the release of ChatGPT, the information component of linguistic complexity goes down, as indicated by the negative and significant coefficient on the *PostGPT* variable, and this effect is stronger for firms with a higher interest from less sophisticated investors.

6.2 Identification

We claim our result from Hypothesis 3 causal. In our model, identification rests on several assumptions. Firstly, we argue that the release of ChatGPT 3.5 is unlikely to be anticipated by the general public. This implies that the release of ChatGPT 3.5 was an exogenous and unanticipated shock for the investors and managers of the firms in our sample. Secondly, the Stable Unit Treatment Value Assumption (SUTVA) indicates that the institutional investor concentration level does not change in the quarters following the release of ChatGPT 3.5. We classify firms with an institutional HHI value higher than the sample mean as having a high institutional concentration. We contend that it is unlikely for the institutional investor concentration level the firm to change drastically. In our sample, 97 percent of the sample satisfies the SUTVA assumption. Our result is robust whether we include or exclude the sample

in violation.

If the first and the second assumption are satisfied, the parallel trend assumption will also hold (Cunningham, 2021). We provide an event study to prove that the changes in information depending on the concentration level of the investors are due to the release of ChatGPT 3.5 instead of the continuation of previous trends. Figure 1 presents the parallel trends of the coefficient for the information components in Hypotheses 3. The x-axis is the calendar year-quarter. The y-axis refers to the coefficients of the regressions. The graphs show that, compared with the baseline period, firms with high and low levels of concentration of institutional ownership are not significantly different from each other before the shock in 2022Q4, but are different afterward. We cannot reject the hypothesis that the two groups are equal in linguistic complexity before the treatment, but we can reject the hypothesis for several consecutive periods after the treatment. This implies that even though the information levels observed before the release of ChatGPT were probably different across firms depending on their investor base, the difference is only changing at the time of the ChatGPT 3.5 release.

The result of the difference-in-difference analysis is shown in Table 9. The coefficient for information component shows consistently and significantly negative result as in Table 5. Although the obfuscation component shows insignificant result, the coefficient is still positive. Overall, the results imply that managers increase information opacity after the release of ChatGPT 3.5 despite the cost of altering sticky text in financial report, consistent with Hypothesis 3.

6.3 Placebo test

If the increase in linguistic complexity among different sophistication of investor is due to the release of ChatGPT 3.5, there shouldn't be any difference in information components among different sophistication of investor before the release, and there could be an even larger difference when large language models are updated.

To address this concern, we divide the institutional ownership into a higher and lower-than-mean group at the point of the ChatGPT 3.5 release and redo a difference-

in-difference-style parallel trend analysis ¹⁴ If the difference between the two subsamples is not different from zero before the release of ChatGPT, but significantly different from zero afterwards, it increases the credibility of our belief that managers change the information disclosure pattern after the release of ChatGPT 3.5.

Figure 2 presents the parallel trend result. The x-axis is the calendar year-quarter. The y-axis refers to the coefficients of the information component. The figure shows that, compared with the baseline period, firms with high and low levels of institutional ownership are not significantly different from each other before the shock in 2022Q4, but are different afterward. We cannot reject the hypothesis that the two groups are equal in linguistic complexity before the treatment, but we can reject the hypothesis for all consecutive periods after the treatment. Furthermore, the information component coefficient becomes higher at the last period, contemporaneous to the release of ChatGPT 4.5. This implies that even though the information levels observed before the release of ChatGPT were probably different across firms depending on their investor base, the difference is only changing at the time of the release of different ChatGPT versions.

Then, we perform the analysis using the high and low institutional ownership groups considering the ownership structure before the release of ChatGPT 3.5. Table 10 presents the result. It is significantly positive for the information component and not significant for the obfuscation component, consistent with our main analysis that after the release of managers reduce the information component based on the composition of their investor base before the release of ChatGPT, while leaving the obfuscation component unchanged.

6.4 Active versus passive mutual funds and changes to linguistic complexity after the release of ChatGPT

Besides our main hypotheses, we argue that the informational content of public disclosures will be more important for active relative to passive index funds. Active

¹⁴We do not claim causality because institutional ownership can change after the exogenous shock of ChatGPT release.

funds are usually short-term investors that do not maintain large positions in the firms where they invest, and they do not have direct contact with the managers, so as not to compromise the liquidity of their stake. They gain from trading with the less sophisticated investors. Therefore, one would expect active investors to favor firms that provide more information in their disclosures. However, this effect may be more pronounced after the release of ChatGPT. According to the convergence hypothesis, the level of information of less sophisticated investors converges to the level of information of more sophisticated investors.

Prior research documents that active institutional investors benefit from more opacity and asymmetric information because they can produce their own private proprietary information, which offers them an advantage when trading with uninformed investors (Bushee, 2012; P. Gao & Liang, 2013; Maffett, 2012). Thus, active institutional investors will gain more from trading with less sophisticated investors if it is more difficult for the latter to extract information from the public reports. Therefore, when active investors are an important constituency in the investor base, managers may generally favor more obfuscation after the release of ChatGPT to maintain the trading advantage of active investors. Nevertheless, if managers believe obfuscation is not effective after the release of ChatGPT, obfuscation may drop more in firms with more active investors. Therefore, managers of firms with a higher proportion of active institutional investors will increase the obfuscation of their disclosures to favor the informational advantage of active institutional investors.

We follow Appel et al. (2016) to retrieve active mutual fund shares for each firm. The active (i.e., non-index) mutual fund percentage is calculated using the number of active mutual fund shares divided by the total mutual fund shares. We specify the model as equation (4) in part 4.5, in which *Investor Base* refers to *HighNonIndex*. This measure takes the value one (zero) if the active mutual fund percentage over total institutional ownership is higher (lower) than the mean at the time of the release of ChatGPT.

Table 11 investigates whether the presence of active mutual funds influences how firms adjust their linguistic complexity in MD&A disclosures after the release of ChatGPT. The dependent variable captures the information component of the Bog

Index (columns 1–4) and the obfuscation component of the Bog Index (columns 5–8). The primary independent variable is the interaction term, where *HighNonIndex* equals 1 if firms have a higher-than-average share of active mutual funds among their institutional investors at the ChatGPT release date, and 0 otherwise.

The results show that the coefficients for the information component (columns 1–2) are not statistically significant. One possible explanation for this non-significant result is that, although active institutional investors generally favor more informative disclosures, their advantage from detailed disclosures diminishes after the introduction of ChatGPT. Therefore, managers have limited incentives to further adjust the information component to cater specifically to active investors following ChatGPT’s release.

However, results for the obfuscation component (columns 3–4) are significantly positive. Specifically, the presence of active mutual funds is associated with an increase in the obfuscation component of about 7% of its standard deviation after ChatGPT’s introduction. This suggests that firms with greater active mutual fund ownership respond to the introduction of LLMs by increasing the obfuscation of their MD&A disclosures. Managers appear motivated to maintain or reinforce the informational advantage enjoyed by active institutional investors, who benefit from trading with less informed investors. Nevertheless, the ultimate success of this obfuscation strategy remains uncertain, given the enhanced capacity of less sophisticated investors to detect obfuscation using tools like ChatGPT.

7 Conclusions

The emergence of ChatGPT has the potential to boost LLM-equipped readers’ tolerance to financial reporting complexity. How do managers from different firms consider it? In this paper, we hypothesize that managers anticipate investors’ higher tolerance to linguistic complexity and react by modifying the information and obfuscation components of linguistic complexity in financial reports towards protecting the interests of sophisticated investors. We find that after the release of ChatGPT, managers decrease the information component of linguistic complexity, which is more

pronounced in firms with a higher percentage of less sophisticated investors. Moreover, they increase the obfuscation component of linguistic complexity, irrespective of the investor base. Firms with high institutional investor concentration reduce more of the information component, and firms with a high percentage of active mutual funds increase the obfuscation component of linguistic complexity. Our results imply that managers are aware of the revolution of LLMs and its benefits to market investors, and tweak their disclosure strategy to maintain the status quo. Most importantly, managers maintain the informational advantage of influential sophisticated investors at the cost of less sophisticated investors. Our results provide insights into the new approach to financial reports by both managers and investors. These insights offer valuable implications for managers looking to optimize their communication strategies in the era of LLM and for investors looking to interpret these evolving practices.

References

- Aghamolla, C., & Smith, K. (2023). Strategic complexity in disclosure. *Journal of Accounting and Economics*, 101635. <https://doi.org/10.1016/j.jacceco.2023.101635>
- Ahmad, U. (2023, April). ChatGPT – a powerful tool for financial analysts. Retrieved October 17, 2023, from <https://www.acuitykp.com/blog/chatgpt-a-powerful-tool-for-financial-analysts/>
- AIContentfy. (2023, January). ChatGPT in Finance: Automating Reports and Analyses. Retrieved May 19, 2024, from <https://aicontentfy.com/en/blog/chatgpt-in-finance-automating-reports-and-analyses>
- Anchev, S., Hellström, J., & Olsson, R. (2022, October). Firm Disclosure and Individual and Institutional Investors. <https://doi.org/10.2139/ssrn.3557246>
- Appel, I. R., Gormley, T. A., & Keim, D. B. (2016). Passive investors, not passive owners. *Journal of Financial Economics*, 121(1), 111–141. <https://doi.org/10.1016/j.jfineco.2016.03.003>
- Arshad, U. (2023, December). Revolutionizing Financial Analysis with ChatGPT Prompts for. Retrieved June 11, 2024, from <https://medium.com/@umararshad2/revolutionizing-financial-analysis-with-chatgpt-prompts-for-a74feaf60f2>
- Asay, H. S., Elliott, W. B., & Rennekamp, K. (2017). Disclosure Readability and the Sensitivity of Investors' Valuation Judgments to Outside Information [Publisher: American Accounting Association]. *The Accounting Review*, 92(4), 1–25. Retrieved November 23, 2023, from <https://www.jstor.org/stable/26550672>
- Asay, H. S., Libby, R., & Rennekamp, K. (2018). Firm performance, reporting goals, and language choices in narrative disclosures. *Journal of Accounting and Economics*, 65(2), 380–398. <https://doi.org/10.1016/j.jacceco.2018.02.002>
- Balakrishnan, K., Billings, M. B., Kelly, B., & Ljungqvist, A. (2014). Shaping Liquidity: On the Causal Effects of Voluntary Disclosure [eprint: <https://onlinelibrary.wiley.com/doi/p>]. *The Journal of Finance*, 69(5), 2237–2278. <https://doi.org/10.1111/jofi.12180>

- Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2022). Attention-Induced Trading and Returns: Evidence from Robinhood Users [eprint: <https://onlinelibrary.wiley.com/doi/p>]. *The Journal of Finance*, 77(6), 3141–3190. <https://doi.org/10.1111/jofi.13183>
- Barber, B. M., & Odean, T. (2000). Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors [eprint: <https://onlinelibrary.wiley.co>]. *The Journal of Finance*, 55(2), 773–806. <https://doi.org/10.1111/0022-1082.00226>
- Barber, B. M., Odean, T., & Zhu, N. (2009). Do Retail Trades Move Markets? *The Review of Financial Studies*, 22(1), 151–186. <https://doi.org/10.1093/rfs/hhn035>
- Ben-Rephael, A., Da, Z., Easton, P. D., & Israelsen, R. D. (2022). Who Pays Attention to SEC Form 8-K? *The Accounting Review*, 97(5), 59–88. <https://doi.org/10.2308/TAR-2019-0425>
- Bernard, D., Blankespoor, E., de Kok, T., & Toynbee, S. (2023, December). A Modular Measure of Business Complexity. <https://doi.org/10.2139/ssrn.4480309>
- Bertomeu, J., Hu, P. (, & Liu, Y. (2023). Disclosure and Investor Inattention: Theory and Evidence. *The Accounting Review*, 98(6), 1–36. <https://doi.org/10.2308/TAR-2022-0122>
- Bertomeu, J., Lin, Y., Liu, Y., & Ni, Z. (2024, November). Voluntary Disclosure, Misinformation, and AI Information Processing: Theory and Evidence. <https://doi.org/10.2139/ssrn.5026360>
- Bertomeu, J., Lin, Y., Liu, Y., & Ni, Z. (2025). The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy. *Journal of Accounting and Economics*, 101782. <https://doi.org/10.1016/j.jacceco.2025.101782>
- Bhattacharya, N., Cho, Y. J., & Kim, J. B. (2018). Leveling the Playing Field between Large and Small Institutions: Evidence from the SEC’s XBRL Mandate. *The Accounting Review*, 93(5), 51–71. <https://doi.org/10.2308/accr-52000>
- BH_Falcon27. (2023, May). ChatGPT in finance. Retrieved June 11, 2024, from www.reddit.com/r/ChatGPTPro/comments/13ajlla/chatgpt_in_finance/

- Bilinski, P. (2024, February). The Usefulness of ChatGPT for Textual Analysis of Annual Reports. <https://doi.org/10.2139/ssrn.4723503>
- Blankespoor, E. (2019). The Impact of Information Processing Costs on Firm Disclosure Choice: Evidence from the XBRL Mandate. *Journal of Accounting Research*, 57(4), 919–967. <https://doi.org/10.1111/1475-679X.12268>
- Blankespoor, E., Croom, J., & Grant, S. M. (2024a, December). Generative AI and Investor Processing of Financial Information. <https://doi.org/10.2139/ssrn.5053905>
- Blankespoor, E., deHaan, E., & Li, Q. (2024b, October). Generative AI in Financial Reporting. <https://doi.org/10.2139/ssrn.4986017>
- Blankespoor, E., deHaan, E., & Marinovic, I. (2020). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2-3), 101344. <https://doi.org/10.1016/j.jacceco.2020.101344>
- Blankespoor, E., Dehaan, E., Wertz, J., & Zhu, C. (2019). Why Do Individual Investors Disregard Accounting Information? The Roles of Information Awareness and Acquisition Costs [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12248>]. *Journal of Accounting Research*, 57(1), 53–84. <https://doi.org/10.1111/1475-679X.12248>
- Blankespoor, E., Miller, B. P., & White, H. D. (2014). Initial evidence on the market impact of the XBRL mandate. *Review of Accounting Studies*, 19(4), 1468–1503. <https://doi.org/10.1007/s11142-013-9273-4>
- Bliss, B. A., Partnoy, F., & Furchtgott, M. (2018). Information bundling and securities litigation [Publisher: Elsevier]. *Journal of Accounting and Economics*, 65(1), 61–84. Retrieved August 24, 2025, from <https://www.sciencedirect.com/science/article/pii/S0165410117300812>
- Bloomfield, M. (2021). The Asymmetric Effect of Reporting Flexibility on Priced Risk. *Journal of Accounting Research*, 59(3), 867–910. <https://doi.org/10.1111/1475-679X.12346>
- Bloomfield, R. (2002). The "incomplete revelation hypothesis" and financial reporting [Num Pages: 11 Place: Sarasota, United States Publisher: American Ac-

- counting Association]. *Accounting Horizons*, 16(3), 233–243. Retrieved December 1, 2023, from <https://www.proquest.com/docview/208897359/abstract/BC0B4E304A114863PQ/1>
- Bloomfield, R. (2008). Discussion of “Annual report readability, current earnings, and earnings persistence”. *Journal of Accounting and Economics*, 45(2), 248–252. <https://doi.org/10.1016/j.jacceco.2008.04.002>
- Bonsall, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2), 329–357. <https://doi.org/10.1016/j.jacceco.2017.03.002>
- Boone, A. L., & White, J. T. (2015). The effect of institutional ownership on firm transparency and information production. *Journal of Financial Economics*, 117(3), 508–533. <https://doi.org/10.1016/j.jfineco.2015.05.008>
- Bushee, B. J. (2012). Discussion of “Financial reporting opacity and informed trading by international institutional investors”. *Journal of Accounting and Economics*, 54(2), 221–228. <https://doi.org/10.1016/j.jacceco.2012.09.001>
- Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic Complexity in Firm Disclosures: Obfuscation or Information? [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/679X.12179>]. *Journal of Accounting Research*, 56(1), 85–121. <https://doi.org/10.1111/1475-679X.12179>
- Bushee, B. J., & Huang, Y. (2024). Do Analysts and Investors Efficiently Respond to Managerial Linguistic Complexity during Conference Calls? *The Accounting Review*, 1–26. <https://doi.org/10.2308/TAR-2019-0358>
- Bushee, B. J., & Noe, C. F. (2000). Corporate Disclosure Practices, Institutional Investors, and Stock Return Volatility [Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley]]. *Journal of Accounting Research*, 38, 171–202. <https://doi.org/10.2307/2672914>
- Bushman, R. M. (1991). Public Disclosure and the Structure of Private Information Markets [Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley]]. *Journal of Accounting Research*, 29(2), 261–276. <https://doi.org/10.2307/2491049>

- Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023a). How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. *The Review of financial studies*, 36(9), 3603–3642. <https://doi.org/10.1093/rfs/hhad021>
- Cao, S., Liang, Y., & Moon, Y. (2023b). Machine Readership and Financial Reporting Decisions. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4650113>
- Chen, W., Hribar, P., & Melessa, S. (2018). Incorrect Inferences When Using Residuals as Dependent Variables [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12195>]. *Journal of Accounting Research*, 56(3), 751–796. <https://doi.org/10.1111/1475-679X.12195>
- Chen, W., Hribar, P., & Melessa, S. J. (2022). On the Use of Residuals as Dependent Variables. *Journal of Financial Reporting*, 7(1), 69–83. <https://doi.org/10.2308/JFR-2021-008>
- Cheng, Q., Lin, P., & Zhao, Y. (2025). Does Generative AI Facilitate Investor Trading? Early Evidence from ChatGPT Outages. *Journal of Accounting and Economics*, 101821. <https://doi.org/10.1016/j.jacceco.2025.101821>
- Cohen, L., Lou, D., & Malloy, C. J. (2020a). Casting Conference Calls [Publisher: INFORMS]. *Management Science*, 66(11), 5015–5039. <https://doi.org/10.1287/mnsc.2019.3423>
- Cohen, L., Malloy, C., & Nguyen, Q. (2020b). Lazy Prices [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12885>]. *The Journal of Finance*, 75(3), 1371–1415. <https://doi.org/10.1111/jofi.12885>
- Comerton-Forde, C., Malinova, K., & Park, A. (2018). Regulating dark trading: Order flow segmentation and market quality. *Journal of Financial Economics*, 130(2), 347–366. <https://doi.org/10.1016/j.jfineco.2018.07.002>
- Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.
- De Bondt, W. F. M. (1998). A portrait of the individual investor. *European Economic Review*, 42(3), 831–844. [https://doi.org/10.1016/S0014-2921\(98\)00009-9](https://doi.org/10.1016/S0014-2921(98)00009-9)
- De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise Trader Risk in Financial Markets [Publisher: The University of Chicago Press]. *Journal of Political Economy*, 98(4), 703–738. <https://doi.org/10.1086/261703>
- Diamond, D. W., & Verrecchia, R. E. (1991). Disclosure, Liquidity, and the Cost of Capital [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540->

- 6261.1991.tb04620.x]. *The Journal of Finance*, 46(4), 1325–1359. <https://doi.org/10.1111/j.1540-6261.1991.tb04620.x>
- Easley, D., & O'hara, M. (2004). Information and the Cost of Capital [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2004.00672.x>]. *The Journal of Finance*, 59(4), 1553–1583. <https://doi.org/10.1111/j.1540-6261.2004.00672.x>
- Eaton, G. W., Guo, F., Liu, T., & Tu, D. (2025). The Cost of Equity: Evidence from Investment Banking Valuations. *Journal of Financial and Quantitative Analysis*, 1–39. <https://doi.org/10.1017/S0022109024000851>
- Gao, M., & Huang, J. (2020). Informing the Market: The Effect of Modern Information Technologies on Information Production. *The Review of Financial Studies*, 33(4), 1367–1411. <https://doi.org/10.1093/rfs/hhz100>
- Gao, P., & Liang, P. J. (2013). Informational Feedback, Adverse Selection, and Optimal Disclosure Policy [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12019>]. *Journal of Accounting Research*, 51(5), 1133–1158. <https://doi.org/10.1111/1475-679X.12019>
- García Osma, B., & Guillamón-Saorín, E. (2011). Corporate governance and impression management in annual results press releases. *Accounting, Organizations and Society*, 36(4-5), 187–208. <https://doi.org/10.1016/j.aos.2011.03.005>
- García-Méndez, D. (2023). Does Internalization Increase Information Acquisition? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4629902>
- Goldman, E., & Wang, W. (2021). Weak Governance by Informed Active Shareholders. *The Review of Financial Studies*, 34(2), 661–699. <https://doi.org/10.1093/rfs/hhaa052>
- Gosselin, A.-M., Le Maux, J., & Smaili, N. (2021). Readability of Accounting Disclosures: A Comprehensive Review and Research Agenda* [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1911-3838.12275>]. *Accounting Perspectives*, 20(4), 543–581. <https://doi.org/10.1111/1911-3838.12275>
- Green, T. C., Jame, R., Markov, S., & Subasi, M. (2014). Access to management and the informativeness of analyst research. *Journal of Financial Economics*, 114(2), 239–255. <https://doi.org/10.1016/j.jfineco.2014.07.003>

- Guo, Y., Tong, S., Chen, T., Kumar, S., & Ouyang, S. (2022, November). The Impact of Generative Artificial Intelligence on Individual Manual Investment Decisions: Empirical Evidence from Mutual Funds. Retrieved January 13, 2025, from <https://papers.ssrn.com/abstract=4644157>
- Hansen, A. L., & Kazinnik, S. (2024, April). Can ChatGPT Decipher Fedspeak? <https://doi.org/10.2139/ssrn.4399406>
- Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1), 405–440. [https://doi.org/10.1016/S0165-4101\(01\)00018-0](https://doi.org/10.1016/S0165-4101(01)00018-0)
- Helter, A. (2023). What Is ChatGPT? Everything You Need to Know. Retrieved October 18, 2023, from <https://www.techtarget.com/whatis/definition/ChatGPT>
- Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1-3), 337–386. <https://doi.org/10.1016/j.jacceco.2003.10.002>
- Hoitash, R., & Hoitash, U. (2018). Measuring Accounting Reporting Complexity with XBRL. *The Accounting Review*, 93(1), 259–287. <https://doi.org/10.2308/accr-51762>
- Hu, K., & Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Retrieved June 17, 2024, from <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- Kalay, A. (2015). Investor sophistication and disclosure clienteles. *Review of Accounting Studies*, 20(2), 976–1011. <https://doi.org/10.1007/s11142-015-9317-z>
- Kim, A., Kim, D. S., Muhn, M., Nikolaev, V. V., & So, E. C. (2024a). AI, Investment Decisions, and Inequality. *Chicago Booth Accounting Research Center Research Paper, Fama-Miller Working Paper, MIT Sloan Research Paper*. Retrieved May 19, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5075727

- Kim, A., Muhn, M., & Nikolaev, V. (2023a, October). Bloated Disclosures: Can ChatGPT Help Investors Process Financial Information? [arXiv:2306.10224 [cs, econ, q-fin]]. Retrieved December 15, 2023, from <http://arxiv.org/abs/2306.10224>
- Kim, A., Muhn, M., & Nikolaev, V. V. (2023b). From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4593660>
- Kim, A., Muhn, M., & Nikolaev, V. V. (2024b, May). Financial Statement Analysis with Large Language Models. <https://doi.org/10.2139/ssrn.4835311>
- Kim, A. G., Muhn, M., Nikolaev, V. V., & Zhang, Y. (2024, November). Learning Fundamentals from Text. <https://doi.org/10.2139/ssrn.5047947>
- Kim, J. M., & Leone, A. J. (2025, June). AI Investing and Misinformation. <https://doi.org/10.2139/ssrn.5283695>
- Kyle, A. S. (1985). Continuous Auctions and Insider Trading [Publisher: [Wiley, Econometric Society]]. *Econometrica*, 53(6), 1315–1335. <https://doi.org/10.2307/1913210>
- Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, 56(1), 130–147. <https://doi.org/10.1016/j.jacceco.2013.05.001>
- Lee, Y.-J. (2012). The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices* [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1911-3846.2011.01152.x>]. *Contemporary Accounting Research*, 29(4), 1137–1170. <https://doi.org/10.1111/j.1911-3846.2011.01152.x>
- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2), 221–247. <https://doi.org/10.1016/j.jacceco.2008.02.003>
- Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach [Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley]]. *Journal of Accounting Research*, 48(5), 1049–1102. Retrieved October 16, 2023, from <https://www.jstor.org/stable/40929537>

- Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, 63(1), 1–25. <https://doi.org/10.1016/j.jacceco.2016.09.002>
- Loughran, T., & McDonald, B. (2024). Measuring Firm Complexity. *Journal of Financial and Quantitative Analysis*, 59(6), 2487–2514. <https://doi.org/10.1017/S0022109023000716>
- Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01625.x>]. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12162>]. *The Journal of Finance*, 69(4), 1643–1671. <https://doi.org/10.1111/jofi.12162>
- Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12123>]. *Journal of Accounting Research*, 54(4), 1187–1230. <https://doi.org/10.1111/1475-679X.12123>
- Maffett, M. (2012). Financial reporting opacity and informed trading by international institutional investors. *Journal of Accounting and Economics*, 54(2), 201–220. <https://doi.org/10.1016/j.jacceco.2012.09.002>
- Mauritz, C., Nienhaus, M., & Oehler, C. (2023). The role of individual audit partners for narrative disclosures. *Review of Accounting Studies*, 28(1), 1–44. <https://doi.org/10.1007/s11142-021-09634-4>
- Mekhaimer, M., Soliman, M., & Zhang, W. (2024). Does Political Uncertainty Obfuscate Narrative Disclosure? *The Accounting Review*, 99(4), 367–394. <https://doi.org/10.2308/TAR-2021-0884>
- Miller, B. P. (2010). The Effects of Reporting Complexity on Small and Large Investor Trading [Publisher: American Accounting Association]. *The Accounting Review*, 85(6), 2107–2143. Retrieved October 30, 2023, from <https://www.jstor.org/stable/27895913>

- Nicoletti, A., & Zhu, C. (2023). Economic Consequences of Transparency Regulation: Evidence from Bank Mortgage Lending [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/679X.12498>]. *Journal of Accounting Research*, 61(5), 1827–1871. <https://doi.org/10.1111/1475-679X.12498>
- Oehler, A., & Horn, M. (2024). Does ChatGPT provide better advice than robo-advisors? *Finance Research Letters*, 60, 104898. <https://doi.org/10.1016/j.frl.2023.104898>
- OpenAI. (2022). Introducing ChatGPT. Retrieved June 20, 2024, from <https://openai.com/index/chatgpt/>
- Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *The Review of Financial Studies*, 22(1), 435–480. <https://doi.org/10.1093/rfs/hhn053>
- PWC. (2024, December). 2024 SEC comment letter trends: MD&A. Retrieved December 23, 2024, from https://viewpoint.pwc.com/dt/us/en/pwc/podcasts/podcasts_US/2024seccommentlettertrendsmda.html
- Rawson, C., Twedt, B. J., & Watkins, J. C. (2023). Managers’ Strategic Use of Concurrent Disclosure: Evidence from 8-K Filings and Press Releases. *The Accounting Review*, 98(4), 345–371. <https://doi.org/10.2308/TAR-2021-0088>
- Rennekamp, K. (2012). Processing Fluency and Investors’ Reactions to Disclosure Readability [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-679X.2012.00460.x>]. *Journal of Accounting Research*, 50(5), 1319–1354. <https://doi.org/10.1111/j.1475-679X.2012.00460.x>
- SEC. (1998, August). A Plain English Handbook: How to create clear SEC disclosure documents. Retrieved January 17, 2024, from <https://www.sec.gov/pdf/handbook.pdf>
- SEC. (2017). Annual Report to Congress on the Dodd-Frank Whistleblower Program. Retrieved November 28, 2023, from <https://www.sec.gov/whistleblower/reports-and-publications/annual-reports/sec-2017-annual-report-whistleblower-program>

- SEC. (2020a). SEC Adopts Amendments to Modernize and Enhance Management’s Discussion and Analysis and other Financial Disclosures. Retrieved November 29, 2023, from <https://www.sec.gov/news/press-release/2020-290>
- SEC. (2020b). SEC Adopts Rule Amendments to Modernize Disclosures of Business, Legal Proceedings, and Risk Factors Under Regulation S-K. Retrieved November 29, 2023, from <https://www.sec.gov/news/press-release/2020-192>
- Segal, B., & Segal, D. (2016). Are managers strategic in reporting non-earnings news? Evidence on timing and news bundling. *Review of Accounting Studies*, 21(4), 1203–1244. <https://doi.org/10.1007/s11142-016-9366-y>
- Singh, S. (2024, May). ChatGPT Statistics (June 2024) — Users, Growth & Revenue. Retrieved June 17, 2024, from <https://www.demandsage.com/chatgpt-statistics/>
- Starks-Technology. (2024, January). Here’s how I’m using ChatGPT to analyze financial statements. First stop: Cathie Wood’s ARKK. Retrieved June 11, 2024, from www.reddit.com/r/ChatGPT/comments/1ab6i6v/heres_how_im_using_chatgpt_to_analyze_financial/
- Talerico, A. (2024). How Finance and Banking Professionals Can Use ChatGPT. Retrieved June 11, 2024, from <https://corporatefinanceinstitute.com/resources/elearning/finance-chatgpt/>
- Tan, H.-T., Ying Wang, E., & Zhou, B. (2014). When the Use of Positive Language Backfires: The Joint Effect of Tone, Readability, and Investor Sophistication on Earnings Judgments [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12039>]. *Journal of Accounting Research*, 52(1), 273–302. <https://doi.org/10.1111/1475-679X.12039>
- You, H., & Zhang, X.-j. (2009). Financial reporting complexity and investor under-reaction to 10-K information. *Review of Accounting Studies*, 14(4), 559–586. <https://doi.org/10.1007/s11142-008-9083-2>
- Zang, A., Zheng, J., & Zheng, R. (2024, March). Measuring Readability with Language Predictability: A Large Language Model Approach. <https://doi.org/10.2139/ssrn.4764707>

- Zhang, G. (2001). Private Information Production, Public Disclosure, and the Cost of Capital: Theory and Implications [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1506/N6G3-RWX7-Y15L-BWPV>]. *Contemporary Accounting Research*, 18(2), 363–384. <https://doi.org/10.1506/N6G3-RWX7-Y15L-BWPV>
- Zheng, K. (2023, January). Voluntary Disclosure and Retail Trading. <https://doi.org/10.2139/ssrn.5378006>

Appendix A: Sample construction

Steps	Details	N
1. Inner join Compustat and Institutional ownership dataset.	Merging with eight-digit CUSIP, calendar year and quarter	318,259
2. Inner join the Fog index dataset.	Merging with CIK, calendar year and quarter with reporting year and quarter.	186,401
3. Inner join Hoitash and Hoitash (2018) dataset.	Merging with SEC filing ID	82,485
4. Left join 94-year Compustat Annual fundamentals to calculate the inception year of a firm.	Merging with eight-digit CUSIP and calendar with the earliest fiscal year available in the dataset.	82,485
5. Left join CRSP dataset.	Merging with ticker, calendar year and quarter with reporting year and quarter.	82,485
6. Left join FINRA dataset.	Merging with ticker, calendar year and quarter with reporting year and quarter.	82,485
7. Left join Analyst forecast dataset.	We keep the earliest analyst report observation released after each firm's report announcement ¹⁵ . Merging with ticker, calendar year and quarter with the earliest next-period forecast end year and quarter.	82,485
8. Construct a balanced panel.	Remove missing data from variables. Winsorize at a level between 1 and 99 percent.	31,225

Appendix B: Variable definitions

Name	Description	Source
Complexity variables		
<i>Bog_MD&A</i>	Bog Index for the MD&A item in txt file, measured by Sentence Bog + Word Bog - Pep (Bonsall et al., 2017).	SEC filings
<i>Bog_Info</i>	Information component of <i>Bog_MD&A</i> .	
<i>Bog_Obfu</i>	Obfuscation component of <i>Bog_MD&A</i> .	
<i>Fog_MD&A</i>	Fog Index for the MD&A item in txt file, measured by $0.4 \times (\text{Average sentence length} + (100 \times \text{Percentage of complex words}))$ (Li, 2008).	SEC filings
<i>Fog_Info</i>	Information component of <i>Fog_MD&A</i> .	
<i>Fog_Obfu</i>	Obfuscation component of <i>Fog_MD&A</i> .	
Investor base variables		
<i>HighInstOwn</i>	Dummy variable of institutional ownership. An observation has high institutional ownership if institutional ownership percentage is greater than the mean sample in the ChatGPT release period.	Derived from variable <i>InstOwn</i>
<i>InstOwn</i>	The institutional ownership of a firm in percentage.	Refinitiv
<i>Internalization</i>	The percentage of retail trading volume over the total trading volume.	FINRA and CRSP
<i>InstHHI</i>	Institutional ownership concentration level.	Refinitiv and CRSP
<i>NonIndexOwn</i>	Non-index fund percentage.	Appel et al. (2016)
Other variables		
<i>AnalystsFollow</i>	The logarithm of one plus the number of analysts providing a quarterly EPS forecast for quarter q.	I/B/E/S
<i>ARC</i>	Accounting reporting complexity, measured by the number of XBRL tags in the financial statement footnotes (Hoitash & Hoitash, 2018).	SEC filings

Name	Description	Source
<i>BM</i>	Book-to-market ratio ($CEQQ / CSHOQ * PRCCQ$).	Compustat
<i>Capex</i>	Capital expenditures divided by sales in quarter q. Set to zero if missing.	Compustat
<i>FirmAge</i>	The year in the panel observations minus the inception year retrieved from the earliest data on Compustat.	Compustat
<i>is10K</i>	Dummy variable equal to one if the filing is a 10-K (i.e., 4th fiscal quarter), and zero otherwise.	SEC 10-K and 10-Q filings
<i>LargeAccFiler</i>	Indicator variable equal to one if the firm is a large accelerated filer, and zero otherwise.	Audit Analytics
<i>Leverage</i>	Long-term plus short-term debt divided by total assets at the end of quarter q ($DLCQ + DLTTQ / ATQ$).	Compustat
<i>Loss</i>	Indicator variable equal to one if the firm's earnings before extraordinary items are negative and zero otherwise.	Compustat
<i>NonAccFiler</i>	Indicator variable equal to one if the firm is a non-accelerated filer, and zero otherwise	Audit Analytics
<i>NonIndexFund</i>	The percentage of non-index fund compared to index fund.	Refinitiv and CRSP
<i>R&D</i>	R&D expense for the quarter divided by sales in quarter q. Set to zero if missing.	Compustat
<i>RecentM&A</i>	Indicator variable equal to one if the firm reported acquisition charges in the current or prior two fiscal quarters, and zero otherwise.	Compustat
<i>ROA</i>	Earnings before extraordinary items in quarter q (IBQ) divided by total assets at the beginning of the quarter (ATQ).	Compustat
<i>Size</i>	Natural log of total assets at the end of quarter q.	Compustat
<i>Stdev</i>	The standard deviation of analyst estimates.	I/B/E/S

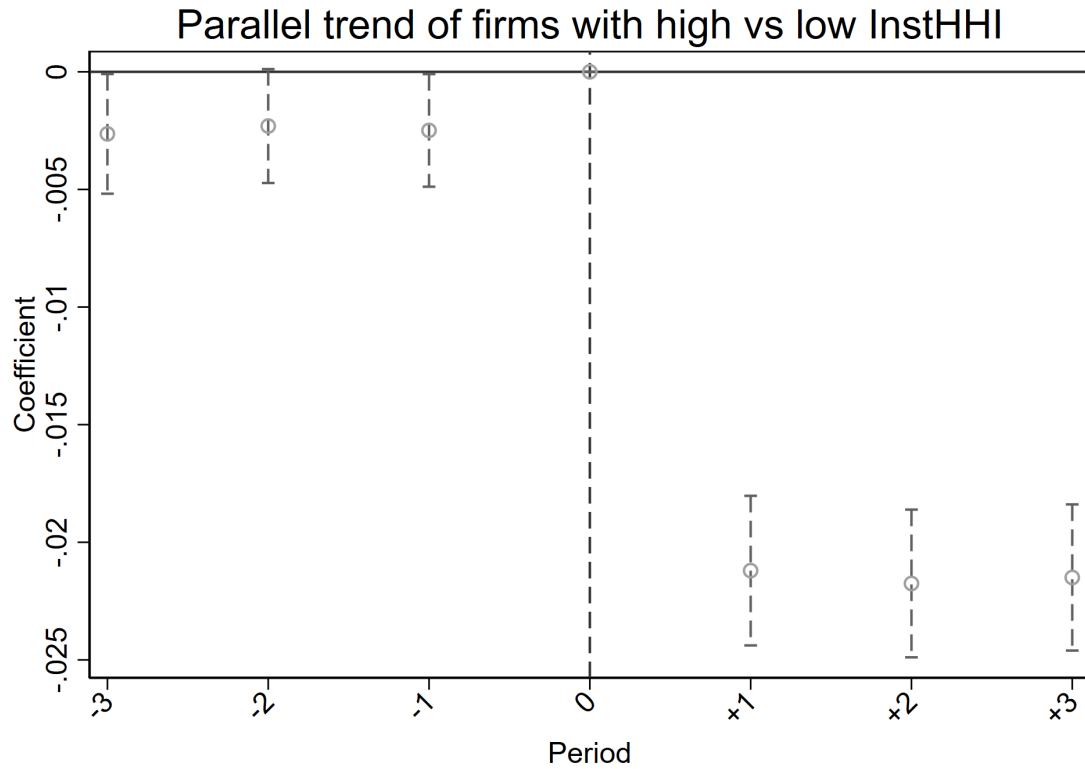


Figure 1: Parallel trend of firms with high vs low institutional concentration

Note: This figure shows the parallel trend for two groups above and below the mean of institutional concentration percentage when ChatGPT 3.5 is released. The Y axis is the information component of the Fog Index from the decomposition. The X axis shows the year-quarter. Period 2022Q4 is served as the benchmark and is thus omitted from the X axis.

Table 1: Summary Statistics

<i>Variables</i>	N	mean	sd	p10	p25	p50	p75	p90
<i>Bog_MD&A</i>	31,225	85.05	12.25	69.00	77.00	85.00	93.00	100.00
<i>Fog_MD&A</i>	31,225	22.58	1.71	20.44	21.43	22.50	23.65	24.81
<i>ARC</i>	31,225	275.69	135.17	142.00	183.00	241.00	336.00	456.00
<i>AnalystsFollow</i>	31,225	1.98	0.67	1.10	1.39	1.95	2.48	2.89
<i>BM</i>	31,202	0.54	0.52	0.07	0.20	0.42	0.78	1.15
<i>BUSSEG</i>	31,225	2.04	1.68	1.00	1.00	1.00	3.00	4.00
<i>Capex</i>	27,409	0.20	0.30	0.00	0.02	0.07	0.19	0.81
<i>FirmAge</i>	31,225	2.89	0.82	1.61	2.20	3.04	3.47	4.04
<i>GEOSEG</i>	31,225	2.16	2.55	0.00	1.00	1.00	3.00	5.00
<i>InstHHI</i>	31,155	0.08	0.09	0.04	0.05	0.06	0.08	0.14
<i>InstOwn</i>	31,225	0.74	0.24	0.36	0.61	0.80	0.92	0.99
<i>Internalization</i>	31,225	0.20	0.10	0.11	0.13	0.17	0.24	0.34
<i>is10K</i>	31,225	0.22	0.42	0.00	0.00	0.00	0.00	1.00
<i>LargeAccFiler</i>	31,225	0.65	0.48	0.00	0.00	1.00	1.00	1.00
<i>Leverage</i>	29,082	0.29	0.25	0.02	0.08	0.26	0.44	0.62
<i>Loss</i>	31,225	0.33	0.47	0.00	0.00	0.00	1.00	1.00
<i>NonAccFiler</i>	31,225	0.15	0.35	0.00	0.00	0.00	0.00	1.00
<i>NonIndexFund</i>	31,155	0.04	0.04	0.01	0.02	0.04	0.06	0.08
<i>R&D</i>	31,225	0.14	0.30	0.00	0.00	0.00	0.10	0.70
<i>RecentM&A</i>	31,225	0.05	0.22	0.00	0.00	0.00	0.00	0.00
<i>ROA</i>	31,221	-0.01	0.05	-0.07	-0.01	0.00	0.01	0.03
<i>Size</i>	31,224	7.82	1.97	5.25	6.52	7.81	9.07	10.40
<i>Stdev</i>	29,087	0.10	0.14	0.01	0.03	0.05	0.11	0.23

Table 1a: Correlation Coefficient Matrix (Columns 1-5)

	<i>Bog_MD&A</i>	<i>Fog_MD&A</i>	<i>InstHHI</i>	<i>InstOwn</i>	<i>Internalization</i>
<i>Bog_MD&A</i>	1.00				
<i>Fog_MD&A</i>	0.38***	1.00			
<i>InstHHI</i>	0.02***	-0.00	1.00		
<i>InstOwn</i>	0.05***	0.13***	-0.32***	1.00	
<i>Internalization</i>	0.06***	-0.03***	0.36***	-0.58***	1.00
<i>AnalystsFollow</i>	0.08***	0.12***	-0.30***	0.42***	-0.31***
<i>ARC</i>	0.06***	0.02***	-0.12***	0.07***	-0.20***
<i>BM</i>	-0.06***	-0.01	0.06***	-0.14***	0.12***
<i>Capex</i>	0.10***	0.10***	0.05***	-0.10***	0.13***
<i>FirmAge</i>	-0.06***	-0.13***	-0.26***	0.22***	-0.24***
<i>is10K</i>	0.11***	-0.08***	-0.02***	0.00	-0.01
<i>LargeAccFiler</i>	0.06***	0.07***	-0.33***	0.52***	-0.46***
<i>Leverage</i>	0.11***	0.08***	0.05***	0.15***	0.01
<i>Loss</i>	0.15***	0.10***	0.19***	-0.14***	0.34***
<i>NonAccFiler</i>	0.05***	0.03***	0.33***	-0.42***	0.45***
<i>R&D</i>	0.22***	0.17***	0.12***	-0.14***	0.30***
<i>RecentM&A</i>	0.03***	0.02***	-0.02***	0.03***	-0.02***
<i>ROA</i>	-0.18***	-0.12***	-0.21***	0.24***	-0.38***
<i>Size</i>	0.01*	0.11***	-0.27***	0.33***	-0.44***
<i>Stdev</i>	0.07***	0.04***	-0.04***	0.08***	-0.00

Note: Table 1a and 1b is the Pearson correlation matrix of our variables. Our model does not suffer multicollinearity problem. ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 1b: Correlation Coefficient Matrix (Columns 6-20)

	<i>AnalystsFollow</i>	<i>ARC</i>	<i>BM</i>	<i>Capex</i>	<i>FirmAge</i>
<i>AnalystsFollow</i>	1.00				
<i>ARC</i>	0.16***	1.00			
<i>BM</i>	-0.21***	0.20***	1.00		
<i>Capex</i>	0.01	-0.03***	0.04***	1.00	
<i>FirmAge</i>	0.20***	0.27***	0.03***	-0.15***	1.00
<i>is10K</i>	-0.01	0.51***	-0.01*	0.11***	0.01
<i>LargeAccFiler</i>	0.54***	0.24***	-0.14***	-0.12***	0.42***
<i>Leverage</i>	0.14***	0.03***	-0.26***	-0.09***	0.06***
<i>Loss</i>	-0.14***	-0.26***	-0.01**	0.23***	-0.38***
<i>NonAccFiler</i>	-0.31***	-0.25***	0.01	0.20***	-0.41***
<i>R&D</i>	-0.09***	-0.33***	-0.14***	0.46***	-0.39***
<i>RecentM&A</i>	0.08***	0.02***	-0.03***	-0.03***	0.02***
<i>ROA</i>	0.17***	0.22***	0.03***	-0.36***	0.36***
<i>Size</i>	0.56***	0.54***	0.14***	-0.11***	0.48***
<i>Stdev</i>	0.15***	0.05***	0.04***	0.10***	0.08***
	<i>is10K</i>	<i>LargeAccFiler</i>	<i>Leverage</i>	<i>Loss</i>	<i>NonAccFiler</i>
<i>is10K</i>	1.00				
<i>LargeAccFiler</i>	0.04***	1.00			
<i>Leverage</i>	0.01*	0.15***	1.00		
<i>Loss</i>	-0.02***	-0.29***	-0.01	1.00	
<i>NonAccFiler</i>	-0.05***	-0.57***	-0.16***	0.32***	1.00
<i>R&D</i>	-0.01**	-0.26***	-0.20***	0.53***	0.40***
<i>RecentM&A</i>	0.01	0.05***	0.02***	-0.00	-0.02***
<i>ROA</i>	0.02***	0.32***	0.07***	-0.66***	-0.39***
<i>Size</i>	0.02***	0.60***	0.12***	-0.45***	-0.48***
<i>Stdev</i>	0.04***	0.11***	0.08***	0.04***	-0.06***
	<i>R&D</i>	<i>RecentM&A</i>	<i>ROA</i>	<i>Size</i>	<i>Stdev</i>
<i>R&D</i>	1.00				
<i>RecentM&A</i>	-0.03***	1.00			
<i>ROA</i>	-0.71***	0.03***	1.00		
<i>Size</i>	-0.48***	0.06***	0.46***	1.00	
<i>Stdev</i>	0.03***	0.01	0.00	0.17***	1.00

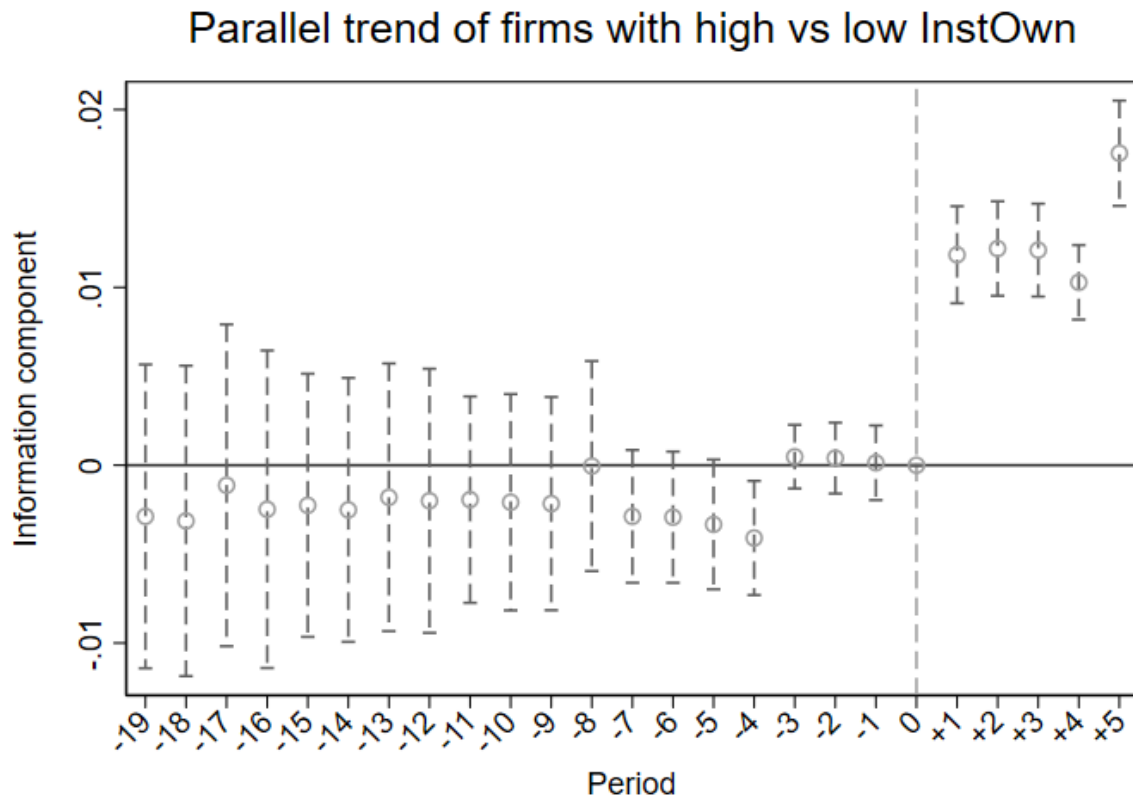


Figure 2: Parallel trend of firms with high vs low institutional ownership

Note: This figure shows the parallel trend for two groups above and below the mean of institutional ownership when ChatGPT 3.5 is released. The Y axis is the information component of the Fog Index from the decomposition. The X axis shows the year-quarter. Period 2022Q4 is served as the benchmark and is thus omitted from the X axis.

Table 2: Linguistic Complexity Decomposition

Panel A		
	(1)	(2)
	<i>Bog_MD&A</i>	<i>Fog_MD&A</i>
<i>BUSSEG</i>	0.324*** (7.05)	0.009 (1.47)
<i>GEOSEG</i>	-0.091*** (-2.69)	-0.008* (-1.82)
<i>FirmAge</i>	-0.544*** (-5.21)	-0.362*** (-25.69)
<i>Size</i>	1.060*** (21.88)	0.244*** (37.33)
N	31,155	31,155
Adj R-squared	0.224	0.276
Industry FE	Yes	Yes

Panel B

<i>Variables</i>	N	mean	sd	p10	p25	p50	p75	p90
<i>Bog_Info</i>	31,155	85.04	2.09	82.34	83.66	85.00	86.38	87.75
<i>Fog_Info</i>	31,155	22.58	0.43	22.03	22.31	22.59	22.86	23.10
<i>Bog_Obfu</i>	31,155	-0.00	10.73	-14.43	-6.84	0.55	7.09	13.05
<i>Fog_Obfu</i>	31,155	0.00	1.45	-1.78	-0.95	-0.05	0.86	1.81

Note: This table reports the complexity components. Panel A shows the results of decomposition of linguistic complexity. The dependent variable shows the linguistic complexity measurement of the MD&A text. We take out the fitted value as the information component and the residuals as the obfuscation component of the linguistic complexity. Variables are winsorized at an 1% and 99% levels. Industry FE is the industry fixed effect. ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test. Panel B shows the descriptive statistics of the decomposition. The information component is the fitted value of the regression. The obfuscation component is the residual of the regression.

Table 3: Changes to information component after the release of ChatGPT 3.5

	Information component of Fog Index			
	(1)	(2)	(3)	(4)
<i>PostGPT</i>	-0.019*** (-5.26)	-0.048*** (-11.07)	-0.071*** (-12.76)	
<i>InstOwn</i> × <i>PostGPT</i>			0.030*** (6.09)	0.037*** (7.59)
<i>InstOwn</i>		0.025* (1.76)	0.022 (1.44)	-0.030*** (-2.96)
<i>AnalystsFollow</i>		-0.004 (-0.94)	-0.004 (-0.89)	-0.017*** (-5.60)
<i>ARC</i>		0.000* (1.78)	0.000* (1.90)	-0.000 (-1.13)
<i>BM</i>		-0.008* (-1.84)	-0.008* (-1.78)	-0.010*** (-4.23)
<i>Capex</i>		0.025*** (3.30)	0.025*** (3.32)	0.015*** (3.92)
<i>is10K</i>		-0.003 (-0.32)	-0.004 (-0.36)	0.004 (0.79)
<i>LargeAccFiler</i>		-0.020*** (-4.77)	-0.020*** (-4.76)	-0.014*** (-5.14)
<i>Leverage</i>		-0.066*** (-7.09)	-0.066*** (-7.05)	-0.040*** (-5.58)
<i>Loss</i>		0.003** (2.18)	0.003** (2.23)	0.002** (2.30)
<i>NonAccFiler</i>		0.011* (2.02)	0.011* (2.00)	0.018*** (4.07)
<i>RD</i>		0.017* (1.85)	0.017* (1.83)	0.008 (1.20)
<i>RecentMA</i>		0.000 (0.14)	0.000 (0.13)	-0.000 (-0.06)
<i>ROA</i>		0.067*** (2.99)	0.069*** (3.15)	0.017 (1.17)
<i>Size</i>		0.196*** (33.84)	0.196*** (34.35)	0.240*** (84.95)
<i>stdev</i>		-0.016** (-2.52)	-0.016** (-2.55)	0.007 (1.52)
N	31107	23595	23595	23595
Adj R-squared	0.980	0.994	0.994	0.996
Firm FE	Yes	Yes	Yes	Yes
Year-Quarter FE	No	No	No	Yes
Double Clustering	Yes	Yes	Yes	Yes

Note: This table shows the effect of institutional ownership on the information component of the Fog Index. The dependent variable (*Fog_Info*) proxies the information component of the linguistic complexity of the report text. The main independent variable is the interaction between the institutional ownership (*InstOwn*) and a dummy variable indicating the period before or after ChatGPT 3.5 release date (i.e. *PostGPT*). Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the year-quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 4: Changes to obfuscation component after the release of ChatGPT 3.5

	Obfuscation component of Fog Index			
	(1)	(2)	(3)	(4)
<i>PostGPT</i>	0.245*	0.116	0.026	
	(1.94)	(0.95)	(0.20)	
<i>InstOwn</i> × <i>PostGPT</i>			0.117	0.044
			(1.12)	(0.44)
<i>InstOwn</i>		-0.358**	-0.374**	-0.091
		(-2.24)	(-2.36)	(-0.67)
<i>AnalystsFollow</i>		-0.116*	-0.115*	-0.002
		(-1.97)	(-1.96)	(-0.05)
<i>ARC</i>		-0.000	-0.000	-0.000
		(-1.01)	(-0.97)	(-0.45)
<i>BM</i>		-0.011	-0.010	-0.012
		(-0.20)	(-0.19)	(-0.33)
<i>CapeX</i>		0.116	0.117	-0.022
		(1.30)	(1.31)	(-0.35)
<i>is10K</i>		-0.337***	-0.338***	-0.452***
		(-3.88)	(-3.91)	(-3.63)
<i>LargeAccFiler</i>		0.010	0.010	-0.024
		(0.20)	(0.20)	(-0.55)
<i>Leverage</i>		0.316**	0.316**	0.114
		(2.56)	(2.56)	(1.02)
<i>Loss</i>		0.006	0.007	0.032
		(0.23)	(0.24)	(1.37)
<i>NonAccFiler</i>		0.007	0.006	-0.051
		(0.13)	(0.12)	(-1.22)
<i>RD</i>		0.053	0.052	0.210
		(0.38)	(0.37)	(1.49)
<i>RecentMA</i>		0.025	0.025	0.023
		(0.52)	(0.52)	(0.52)
<i>ROA</i>		-0.241	-0.234	-0.036
		(-0.93)	(-0.89)	(-0.16)
<i>Size</i>		0.151**	0.149**	-0.153***
		(2.16)	(2.12)	(-3.51)
<i>stdev</i>		0.383**	0.382**	0.175*
		(2.37)	(2.37)	(1.96)
N	31107	23595	23595	23595
Adj R-squared	0.701	0.713	0.713	0.735
Firm FE	Yes	Yes	Yes	Yes
Year-Quarter FE	No	No	No	Yes
Double Clustering	Yes	Yes	Yes	Yes

Note: This table shows the effect of institutional ownership on the obfuscation component of the Fog Index. The dependent variable (*Fog-Obfu*) proxies the obfuscation component of the linguistic complexity of the report text. The main independent variable is the interaction between the institutional ownership (*InstOwn*) and a dummy variable indicating the period before or after ChatGPT 3.5 release date (i.e. *PostGPT*). Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the year-quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 5: Robustness check: changes to information component after ChatGPT release depending on internalization

	Decomposed Fog Index			
	Information component		Obfuscation component	
	(1)	(2)	(3)	(4)
<i>InstHHI</i>	-0.130** (-2.65)	0.090*** (4.39)	0.532** (2.09)	0.315 (1.05)
<i>InstHHI</i> $\times$ <i>PostGPT</i>	-0.206*** (-5.74)	-0.139*** (-5.59)	0.423** (2.33)	0.796*** (2.82)
N	31107	23595	31107	23595
Adj R-squared	0.980	0.997	0.732	0.735
Controls	No	Yes	No	Yes
Firm FE	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes
Double Clustering	Yes	Yes	Yes	Yes

Note: This table shows the effect of institutional ownership concentration on linguistic complexity. The dependent variable (i.e. *Fog_Info/Fog_Obfu*) proxies the information/obfuscation component of the report text. The main independent variable is the interaction between institutional ownership concentration (*InstHHI*) and ChatGPT 3.5 release date (*PostGPT*). Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the Year-Quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 6: Robustness Check: Changes to linguistic complexity components after ChatGPT 3.5 release using Bog Index and File Size

	Information component				Obfuscation component			
	Bog Index		File Size		Bog Index		File Size	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>PostGPT</i>	-0.067*** (-7.31)		-0.328*** (-6.14)		0.584 (0.89)		-1.214 (-0.51)	
<i>InstOwn</i> × <i>PostGPT</i>		0.053** (2.68)		0.243* (1.96)		0.708 (0.90)		-1.282 (-0.97)
<i>InstOwn</i>	0.010 (0.23)	-0.092** (-2.35)	0.046 (0.18)	-0.503** (-2.21)	-5.732*** (-3.38)	-0.988 (-0.76)	-0.874 (-0.27)	-0.910 (-0.48)
N	23595	23595	23595	23595	23595	23595	23595	23595
Adj R-squared	0.994	0.994	0.993	0.993	0.494	0.528	0.688	0.725
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	No	Yes	No	Yes	No	Yes	No	Yes
Double Clustering	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table shows the effect of institutional ownership on the information component of the Bog Index. The dependent variables proxy the information component of the linguistic complexity of the report text. The main independent variable is the interaction between the institutional ownership (*InstOwn*) and a dummy variable indicating the period before or after ChatGPT 3.5 release date (i.e. *PostGPT*). Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the year-quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 7: Robustness Check: Non-decomposition analysis

	Bog Index			
	(1)	(2)	(3)	(4)
<i>InstOwn</i> × <i>PostGPT</i>	2.618** (2.03)	2.618* (1.86)	0.716 (0.91)	0.716 (0.75)
<i>InstOwn</i>	0.613 (0.42)	0.613 (0.26)	-0.212 (-0.22)	-0.212 (-0.15)
<i>AnalystsFollow</i>	0.302 (0.68)	0.302 (0.41)	-0.416 (-1.17)	-0.416 (-0.68)
<i>ARC</i>	0.004*** (3.08)	0.004 (1.56)	0.013*** (6.33)	0.013*** (3.86)
<i>BM</i>	-0.033 (-0.10)	-0.033 (-0.06)	-0.264 (-0.89)	-0.264 (-0.69)
<i>CapeX</i>	-1.591** (-2.41)	-1.591 (-1.51)	-0.516 (-1.18)	-0.516 (-0.77)
<i>is10K</i>	4.238*** (5.63)	4.238*** (4.80)	1.596** (2.24)	1.596* (2.04)
<i>LargeAccFiler</i>	-1.569*** (-2.79)	-1.569 (-1.67)	-0.565* (-1.90)	-0.565 (-1.14)
<i>Leverage</i>	1.517 (1.05)	1.517 (0.59)	0.068 (0.10)	0.068 (0.06)
<i>Loss</i>	0.155 (0.57)	0.155 (0.57)	0.255 (1.03)	0.255 (0.93)
<i>NonAccFiler</i>	-2.189*** (-2.94)	-2.189* (-1.89)	-0.617** (-2.02)	-0.617 (-1.24)
<i>RD</i>	3.083 (1.47)	3.083* (1.77)	-0.359 (-0.40)	-0.359 (-0.27)
<i>RecentMA</i>	-0.336 (-0.77)	-0.336 (-0.67)	-0.464 (-1.17)	-0.464 (-0.83)
<i>ROA</i>	-8.773** (-1.96)	-8.773* (-2.04)	-9.793*** (-4.14)	-9.793*** (-3.96)
<i>Size</i>	3.215*** (7.20)	3.215*** (4.11)	-0.570* (-1.88)	-0.570 (-1.16)
<i>stdev</i>	1.245* (1.77)	1.245 (1.44)	4.416*** (5.41)	4.416*** (3.73)
N	10917	10917	12641	12641
Adj R-squared	0.601	0.601	0.655	0.655
Firm FE	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes
Double Clustering	No	Yes	No	Yes

Note: This table shows the effect of institutional ownership on linguistic complexity based on two subsamples with different levels of business complexity (BC). The dependent variable is the Bog Index of the MD&A section. We choose the Bog Index because it is the most advanced overall linguistic complexity measurement among all bag-of-words methods. High (Low) BC means the business complexity of the subsample is higher (lower) than the sample mean. The main independent variable is the interaction between the institutional ownership (*InstOwn*) and a dummy variable indicating the period before or after ChatGPT 3.5 release date (i.e. *PostGPT*). Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the year-quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 8: Robustness check: changes to information component after ChatGPT release depending on internalization

	Decomposed Fog Index			
	Information component		Obfuscation component	
	(1)	(2)	(3)	(4)
<i>Internalization</i> × <i>PostGPT</i>	-0.109*** (-8.38)	-0.120*** (-8.72)	-0.121 (-0.54)	-0.200 (-0.99)
<i>PostGPT</i>	-0.025*** (-5.45)		0.141 (1.18)	
N	23595	23595	23595	23595
Adj R-squared	0.994	0.997	0.713	0.735
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Year-Quarter FE	No	Yes	No	Yes
Double Clustering	Yes	Yes	Yes	Yes

Note: This table shows the robustness check results from the main model. The dependent variable (*Fog_Info*/*Fog_Obfu*) proxies the information or obfuscation component of the linguistic complexity of the report text. The main independent variable is the interaction between internalization (*Internalization*) and the ChatGPT 3.5 release date (*PostGPT*). *HighInt* is a dummy that takes on the value 1 (0) if the value of the continuous variable *Internalization* is higher (lower) than the mean of the sample at the time of ChatGPT 3.5 release. *PostGPT* is a time dummy indicating the date before or after the release of ChatGPT 3.5. Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the year-quarter fixed effect. SE Clustering shows the standard error is clustered at the firm level. ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 9: Changes to linguistic complexity components after ChatGPT depending on concentration

	Decomposed Fog Index	
	Information component	Obfuscation component
	(1)	(2)
<i>HighInstHHI</i> × <i>PostGPT</i>	-0.029*** (-6.28)	0.030 (0.45)
N	23595	23595
Adj R-squared	0.997	0.735
Controls	Yes	Yes
Firm FE	Yes	Yes
Year-Quarter FE	Yes	Yes
Double Clustering	Yes	Yes

Note: This table shows the results from the model. The dependent variable *Fog_Info*/*Fog_Obfu* proxies the information/obfuscation component of the report text. The main independent variable is the interaction between (*HighInstHHI*) and ChatGPT 3.5 release date (*PostGPT*). *HighInstHHI* takes the value 1 (0) for firms that have higher (lower) than the median values of institutional investor concentration (i.e. Herfindahl–Hirschman index) at the release period of ChatGPT. Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the Year-Quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 10: Robustness Check: Changes to linguistic complexity components after ChatGPT 3.5 release, depending on the investors' base composition before the release

	Information component			Obfuscation component		
	Fog Index (1)	Bog Index (2)	File Size (3)	Fog Index (4)	Bog Index (5)	File Size (6)
<i>HighInstOwn</i> × <i>PostGPT</i>	0.016*** (5.91)	0.030** (2.76)	0.140** (2.10)	0.085 (1.63)	0.484 (1.16)	-0.611 (-0.85)
<i>InstOwn</i>	-0.025** (-2.62)	-0.087** (-2.23)	-0.478** (-2.12)	-0.089 (-0.66)	-0.918 (-0.71)	-1.050 (-0.56)
<i>AnalystsFollow</i>	-0.017*** (-5.57)	-0.062*** (-4.70)	-0.358*** (-4.37)	-0.001 (-0.01)	-0.058 (-0.12)	-0.919 (-1.21)
N	23595	23595	23595	23595	23595	23595
Adj R-squared	0.996	0.994	0.993	0.735	0.528	0.725
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Double Clustering	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table shows the effect of institutional ownership on the information and obfuscation component of the linguistic complexity. We use three proxies for linguistic complexity as dependent variable: The Fog Index, the Bog Index, and File Size (Bonsall et al., 2017; Li, 2008; Loughran and Mcdonald, 2014). Each of them is decomposed using the method from Bushee et al. (2018). The main independent variable is the interaction between a dummy variable indicating whether a firm had high institutional ownership (*HighInstOwn*) during the period previous to the release of ChatGPT 3.5 and a dummy variable indicating the period before or after ChatGPT 3.5 release date (i.e. *PostGPT*). We set high institutional ownership as 1 if the institutional ownership of the firm is higher than the sample mean during the period before the release of ChatGPT 3.5, and 0 otherwise. Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the year-quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Table 11: Robustness check: changes to information component after ChatGPT release depending on non-index fund percentage

	Decomposed Bog Index			
	Information component		Obfuscation component	
	(1)	(2)	(3)	(4)
<i>NonIndexOwn</i> $\times$ <i>PostGPT</i>	-0.029 (-0.24)	0.035 (0.29)	9.680** (2.26)	7.152* (1.78)
<i>PostGPT</i>	-0.065*** (-6.27)		0.137 (0.21)	
<i>NonIndexOwn</i>	0.148 (1.26)	0.045 (0.44)	-6.607 (-1.70)	-2.818 (-0.69)
N	23595	23595	23595	23595
Adj R-squared	0.994	0.996	0.713	0.735
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Year-Quarter FE	No	Yes	No	Yes
Double Clustering	Yes	Yes	Yes	Yes

Note: This table shows the effect of non-index mutual fund percentage on different components of linguistic complexity. The dependent variable *Bog_Info/Bog_Obfu* proxies the information/obfuscation component of the report text. The main independent variable is the interaction between non-index institutional ownership *NonIndexOwn* and *PostGPT*. Variables are winsorized at 1% and 99% levels. Firm FE is the firm fixed effect. Year-Quarter FE is the Year-Quarter fixed effect. We apply double clustering at the firm and year-quarter levels to account for error correlation within firms over time and within periods across firms (Petersen, 2009). The standard error is clustered at the firm or industry level. ***, ** and * indicate significance at 1%, 5% and 10% levels in a two-tailed test.

Appendix C: Theory model

This simple theory model shows that it can be optimal to limit disclosure when the presence of an active shareholder can increase firm value, even if this means that uninformed investors will lose money when trading with the better informed active investor. It is based on Goldman & Wang (2021) and their interpretation of seminal paper Kyle (1985)

C.1 Model Setup

C.1.1 The firm

We model an all-equity financed firm with an initial value V_0 at $t = 0$ and a final value V_1 at time $t = 1$. There are two types of shareholders and a market maker. The discount rate is normalized to zero and all shareholders and market maker are assumed to be risk neutral.

The manager runs the firm and the final value V_1 will take value V_H with probability p and V_L with probability $(1 - p)$, with $V_H > V_0 > V_L$.

Value V_1 is determined at an interim period $t = i$. Depending on the firm's disclosure policy, V_1 can be learned by market participants before termination with probability d or remains unknown until time $t = 1$ with probability $(1 - d)$. Nevertheless, all agents understand the distribution of value and, if they do not learn the true value, they assign the expected intrinsic value determined by nature.

C.1.2 Shareholders

There are two types of potential shareholders in our model.

The first type of shareholder can be thought of as an active informed institutional shareholder. This active shareholder, if present in the firm at time $t = 0$ with a stake X_0 , can influence managerial decisions and increase the probability of a high value to $(p + a)$ at a cost $C(a)$. We assume active influence is cost efficient (i.e. $a(V_H - V_L) > C(a)$). Moreover, we assume the required initial stake, X_0 , is observable and exogenously given and equal to the minimum stake required to exert

influence. We normalize the total number of shares outstanding to one and assume $X_0 < 1$.

Other shareholders hold the remaining $(1 - X_0)$ shares. These shareholders are either passive or inactive and cannot influence firm actions. However, they can trade with the market maker for liquidity reasons and the aggregate net order flow from these investors, U , is expected to be zero ($E(U) = 0$).

We assume that the active shareholder is informed because, by virtue of its active role, he collects private information about firm value and always learns V_1 at the interim, irrespectively of the firm's disclosure policy, d . Therefore, in our model, with probability $(1 - d)$ the active shareholder has an information advantage relative to the rest of shareholders in the market.

C.1.3 Trading

If the active shareholder is present, he can trade the firm's stocks at the interim period after receiving his private information regarding value. The active shareholder's trading activity is characterized by the standard Kyle (1985) model.

Specifically, if the informed, active shareholder chooses to trade X shares, he submits an order to the market maker who sets a transaction price P'_i based on the aggregate order flow as follows:

$$P'_i = P_i + \lambda[X + U - E(X)] \quad (5)$$

where

- P_i represents the market maker's assessed stock price at the interim period after disclosure and before the trading starts;
- X represents the demand of the active shareholder; U represents the demand of liquidity traders with an expected value of $E(U) = 0$; and
- $E(X)$ represents the market maker's expectation of the demand by the active shareholder.

- λ captures market illiquidity.

The market maker only observes the combined order flow $X + U$ and must set the price to satisfy the equilibrium condition that $E(P'_i) = P_i$, implying the market is efficient. This is achieved as long as the market maker correctly anticipates the expected demand by the active shareholder. The parameter λ captures the price impact of trades and is assumed exogenous in our setting, without loss of generality. A high value of λ represents low stock liquidity.

The active shareholder has an information advantage, because, when disclosure fails, the true value of the firm may be still unknown to the market when he trades. In the next section, we present the detailed time line of the model, which demonstrates the information sets and the actions taken by the different agents.

C.1.4 Model time line

At $t = 0$ disclosure level d is chosen and the active shareholder determines whether to invest in the firm. At the interim period, $t = i$, there are three consecutive actions.

- First, nature determines the value of the firm.
- Second, after disclosure by the firm, the active shareholder learns the true value and the rest of market participants learns it with probability d .
- Third, trade happens.

At $t=1$, if disclosure failed at the interim, terminal value V_1 is fully revealed to the market at this time. Shareholders are remunerated according to their stake.

C.2 Equilibrium

To solve the model, we need to characterize:

1. The optimal choice of d at $t = 0$.
2. The optimal investment choice of the active shareholder at time $t = 0$.

3. The optimal trading of the active shareholder at the interim $t = i$ when the market maker sets the price by taking into account the active shareholder's expected trading.

We derive the solution to the model by backward induction, starting by deriving the optimal trading quantity.

C.3 Optimal trading by the active shareholder

If the active shareholder is present in the firm, at the interim period we can have two different situations. First, with probability d , the final value is known by all market participants at the interim so that $P_i = P_1 = V_1$. Trading is not profitable and trades of the uninformed investors happen at price P_i for liquidity reasons with net value U and $E(U) = 0$. Second, with probability $(1 - d)$ market participants do not learn the final firm value until termination so $P_i = P_0 = E(V_1)$. In this second case, (i) the active shareholder chooses the amount of shares to trade, X , so as to maximize his expected gains from trading; (ii) nature determines the realization of net liquidity trades, U ; (iii) the market maker sets the trading price, P'_i , according to (1) after observing the joint total order flow $(X + U)$ and comparing that to its expected value $E(x)$.

The active shareholder solves the following maximization problem:

$$\begin{aligned} \max_X \quad & X(V_1 - E[P'_i|X]) \\ \text{s.t.} \quad & P'_i = P_i + \lambda[X + U - E(X)] \end{aligned} \tag{6}$$

where V_1 is either V_H or V_L . Given X and knowing $E[U] = 0$ we have:

$$E[P'_i|X] = P_i + \lambda[X - E(X)] \tag{7}$$

Substituting (3) into (2) we have:

$$\max_X \quad X(V_1 - P_i - \lambda[X - E(X)]) \tag{8}$$

And, we can express the F.O.C. as:

$$V_1 - P_i - 2\lambda X + \lambda E[X] = 0 \quad (9)$$

The S.O.C. is satisfied, since $-2\lambda < 0$. And the optimal trade of the active shareholder is:

$$X = \frac{1}{2\lambda}(V_1 - P_i) + \frac{1}{2}E[X] \quad (10)$$

This is correctly anticipated by the market maker who computes:

$$E[X] = \frac{1}{2\lambda}(E[V_1] - P_i) + \frac{1}{2}E[X] \quad (11)$$

Since the market anticipates correctly $P_i = E[V_1]$, the previous expression becomes $E[X] = \frac{1}{2}E[X]$ which requires $E[X] = 0$. So that the equilibrium condition $P_i = E[P_i]$ is also satisfied. Therefore, the active shareholder optimal trade is:

$$X = \frac{1}{2\lambda}(V_1 - P_i) \quad (12)$$

Therefore, optimal trading of the active shareholder is as follows.

C.3.1 If disclosure is successful

If the market learns final value at the interim time (with probability d), the active shareholder does not trade.

C.3.2 If disclosure fails

If disclosure fails (with probability $(1 - d)$) the active shareholder trades depending on the final value.

With probability $(p + a)$, $V_1 = V_H$ he buys a fraction of shares:

$$X = \frac{1}{2\lambda}(V_H - ((p + a)V_H + (1 - p - a)V_L)) \quad (13)$$

and his trading profit is

$$X(V_H - P'i) = \frac{1}{2\lambda}(1 - p - a)^2(V_H - V_L)^2 \quad (14)$$

With probability $(1 - p - a)$, $V_1 = V_L$ he sells shares:

$$X = \frac{1}{2\lambda}(V_L - ((p + a)V_H + (1 - p - a)V_L)) \quad (15)$$

and his trading profit is

$$X(V_L - P'i) = \frac{1}{2\lambda}(p + a)^2(V_L - V_H)^2 \quad (16)$$

C.4 Trading profit

The anticipated trading profit of the active shareholder is:

$$\begin{aligned} E[\Pi_T] &= (1 - d)(p + a)\frac{1}{2\lambda}[(1 - p - a)^2(V_H - V_L)^2] + \\ &\quad (1 - d)(1 - p - a)\frac{1}{2\lambda}[(p + a)^2(V_L - V_H)^2] = \\ &= \frac{1}{2\lambda}(1 - d)(1 - p - a)(p + a)(V_H - V_L)^2 \end{aligned} \quad (17)$$

C.5 Optimal investment choice of the active shareholder

At time $t = 0$ the active shareholder decides whether to invest in the firm. At this point, the active shareholder has to buy X_0 and his trade is disclosed, so that he does not have an informational advantage over the market maker and buys at the expected value:

$$P_0 = E[V_1] = (p + a)V_H + (1 - p - a)V_L \quad (18)$$

The final net payoff of the active shareholder is:

$$\Pi = (X_0 + X)V_1 - X_0P_0 - XP'_i - C(a) \quad (19)$$

At time $t=0$ his expected profit is:

$$E[\Pi] = E[\Pi_T] - C(a) = \frac{1}{2\lambda}(1-d)(1-p-a)(p+a)(V_H - V_L)^2 - C(a) \quad (20)$$

Therefore, the active investor will invest if:

$$d \text{ is lower than } \bar{d} = 1 - \frac{C(a)}{(1-p-a)(p+a)\frac{1}{2\lambda}(V_H - V_L)^2} \quad (21)$$

C.6 Optimal disclosure policy d^*

If the active shareholder does not enter the firm, the expected value is

$$E[V_1] = pV_H + (1-p)V_L \quad (22)$$

so the value of d does not affect firm value.

If the active shareholder enters the firm, total value increases to:

$$E[V_1] = (p+a)V_H + (1-p-a)V_L - C(a) \quad (23)$$

Notice that the trading profit of the active shareholder does not enter this equation because it is a gain for the active shareholder and a loss for the rest of the shareholders, so it does not affect total value, but only its division among shareholders. Interestingly, the value for market participants is given by:

$$E[V_{1M}] = (p+a)V_H + (1-p-a)V_L - (1-d)(1-p-a)(p+a)\frac{1}{2\lambda}(V_H - V_L)^2 \quad (24)$$

which is increasing in d .

Therefore, the optimal level of disclosure is the maximum level that ensures the participation of the active shareholder. i.e. $d^* = \bar{d}$.

C.7 Optimal disclosure and ChatGPT

The release of Chat GPT is an information shock that, for a given disclosure policy d , increases the probability of the information about the firm value reaching market investors to $d + S$. Therefore, it is optimal to reduce disclosure back to the level that satisfies active shareholder participation, so that after this shock disclosure is set at $d^{*'} = d^* - S$.