Generative AI and Investment Research: Evidence from Analyst Reports

Shawn Huang, Artur Hugon, Tengfei Zhang, Wenting Zheng*

Abstract

We provide large-sample evidence on AI adoption by equity analysts and examine how the use of generative AI affects forecast quality. We develop a supervised machine learning model to detect AI-assisted reports and find such reports are associated with greater earnings forecast accuracy. Leveraging an exogenous shock to analyst workload from overlapping earnings calls, we find that AI usage in reports is likely a causal driver of improved forecast accuracy. Additional tests reveal that AI-assisted reports adopt more valuation models, discuss more topics and stakeholders, and exhibit more analytical reasoning, suggesting that generative AI empowers analysts' research capabilities beyond writing. Consistent with AI-assisted forecasts containing more value relevant information, we find these forecasts trigger stronger capital market reactions.

Keywords: Generative AI; ChatGPT; AI Detection; Analyst Reports; Investment Research; Earnings Forecast; Market Reaction

Shawn Huang: Arizona State University, Email: shawn.huang@asu.edu. Artur Hugon: Arizona State University. Email: Artur.Hugon@asu.edu. Tengfei Zhang: Rutgers University. Email: tengfei.zhang@rutgers.edu. Wenting Zheng: Arizona State University. Email: w.zheng@asu.edu.
We thank GPTZero and Originality.ai for providing API access used for small-sample validation exercises. We thank Alex Cui (GPTZero) for sharing technical documentation. For helpful comments, we thank Leonardo Barcellos, Paul Brockman, Donald Bowen III, Jie Cai, Rebel Cole, Mary Cowx, Naveen Daniel, Chuck Fang, Lucile Faurel, Antoine Giannetti, Steve Hillegeist, Chang Woo (Mark) Hong, Steve Kaplan, Abdul Khizer, Phillip Lamoreaux, Yinghua Li, Matthew Ma, Eldar Maksymov, Michal Matejka, Roni Michaely, Nandu Nayar, Edward Nelling, Greg Nini, Heidi Packard, Ke Shen, Ke Yang, Yuxiang Zheng. We thank seminar participants at Arizona State University, Drexel University, Florida Atlantic University, Lehigh University, Seton Hall University, and Rutgers University-Camden.

"… Financial services companies will be other big beneficiaries of generative AI, accounting for roughly 20% of value creation throughout the rest of the decade. For the financial services industry, generative AI can be integrated into stock market analysis and internal database management systems to improve decision-making …"

— ABI Research, 2025

1. Introduction

With the rapid integration of generative AI into the workplace, a growing finance and accounting literature uses generative AI (GenAI) as a research tool to study topics such as asset management (Sheng et al., 2025), business networks (Breitung and Müller, 2025), corporate culture (Li et al., 2025), and firms' exposure to GenAI (Eisfeldt et al., 2025). However, much less is known about how corporate professionals and market participants themselves adopt GenAI in their daily work (see Brynjolfsson et al., 2025, which examines the adoption of GenAI among customer-service agents). A key barrier to studying real-world AI adoption is measurement. Existing studies that attempt to detect GenAI-generated content, such as in financial reports (Blankespoor et al., 2024), Seeking Alpha articles (Bradshaw et al., 2025), and Reddit's r/WallStreetBets posts (Hirshleifer et al., 2025), typically rely on third-party commercial detection tools (e.g., GPTZero, Originality.ai). Because these tools are proprietary and opaque, their use raises concerns about measurement validity, transparency, and reproducibility. In this paper, we address this methodological gap by developing a supervised machine learning model to detect GenAI usage in analyst reports using linguistically grounded features and validated training data. We then use this measure to examine how AI adoption affects the productivity of equity analysts' investment research.

Sell-side analysts are important information intermediaries in capital markets, providing investors with timely, third-party assessments of firm performance and growth

prospects through their research reports (Asquith et al., 2005; Huang et al., 2014). Producing these reports requires analysts to process and interpret large volumes of information from public filings, earnings calls, and industry data, while also discovering new information and novel insights through activities such as private communications with managers and site visits (Huang et al., 2018; Bradley et al., 2022). At the same time, analysts cover a growing number of companies, and face time pressure and intense competition, all of which limit the attention and resources they can devote to each report. Generative AI offers a potential solution to reduce information-processing costs by rapidly synthesizing large volumes of textual and numerical data. However, investment research also relies heavily on judgment, contextual interpretation, and credibility, and AI-generated content may introduce errors, hallucinations, or overly generic analysis that weakens informational value. As a result, the effect of AI usage on research quality is ambiguous. This raises an important question: Does AI usage improve or impair the quality of investment research in analyst reports?

To investigate how AI affects the production of analyst reports, we develop a novel measure of AI usage by applying a supervised machine-learning detection framework to a large corpus of analyst reports, identifying reports with a higher likelihood of incorporating generative AI. Our analyst report sample consists of approximately 38,000 research reports, covering the period from January 1, 2022, to April 1, 2024. To construct training data for the detection model, we prompt ChatGPT to rewrite analyst reports issued prior to the public release of ChatGPT-3.5 on November 30, 2022, thereby generating AI-assisted versions of originally human-authored text. Using this paired dataset of human-written and AI-assisted analyst reports, we train machine learning models based on 25 linguistic

features documented in the literature (e.g., Guo et al., 2023 and Uzun, 2023) to distinguish AI-assisted content from human-authored content.[1]

We evaluate three supervised classification models, including XGBoost, Random Forest, and Logistic Regression. We find that the XGBoost classifier exhibits the strongest and most stable performance in distinguishing human-authored from AI-assisted analyst reports and is therefore adopted as our primary detection model. The model performs well in settings where portions of human-written reports are substituted with AI-assisted content. We also benchmark our self-trained model against widely used commercial AI-detection services, including GPTZero and Originality.ai. These commercial tools are mostly effective in identifying documents that are largely or fully AI-generated but are less sensitive to more moderate or partial forms of AI assistance. This distinction is important because existing evidence suggests that the share of AI-generated text in practice is relatively small, and analysts are unlikely to rely on AI to produce a full report given the reputational risks in this highly competitive profession. Applying the trained XGBoost classifier, we generate a continuous, document-level probability measure (AI_Prob) for each analyst report, capturing the likelihood that the report incorporates generative AI. We find a clear upward trend in AI usage in analyst reports over the sample period and an uptick of AI usage after the release of GPT-4.

We offer five sets of findings. First, we find that higher AI usage in analyst reports is associated with significantly greater forecast accuracy. On average, an increase in AI usage is linked to a meaningful improvement in relative forecast accuracy, corresponding

[1] We find consistent results when using reports produced in 2021 and using reports produced from 2000 to 2020 as our training data. Also, our results are robust to GPT model choices (ChatGPT-3.5 or ChatGPT-4) and prompting strategies to rewrite human-written reports.

to roughly a 15% gain compared to the sample mean. This relation reflects AI's role in facilitating analysts' information acquisition and insight generation rather than improving writing quality only.

Second, our cross-sectional analyses show that the effectiveness of AI varies systematically with analyst characteristics. Specifically, the accuracy benefits of AI are larger for analysts who cover more firms, follow a broader set of industries, and issue forecasts more frequently, consistent with AI helping analysts manage scale and workload pressures. We further find that the benefits of AI are more pronounced for analysts without CFA certification or All-Star recognition, as well as for less experienced analysts, indicating that AI can partially substitute for differences in human capital and industry experience.

Third, we exploit a plausibly exogenous shock to AI usage in analyst reports to address the endogeneity concern that our AI usage measure may capture underlying report informativeness or analyst skills. We employ reports released on overlapping earnings calls of firms that an analyst covers, which create an exogenous increase in workload and time pressure for analysts to process information and produce reports. We observe higher AI usage in reports released on dates with overlapping earnings calls, compared to reports released on dates without overlapping calls. Using overlapping call dates as an instrument for AI usage in reports in two-stage least squares (2SLS) regressions, we reaffirm the effects of AI usage on forecast accuracy.

Fourth, we find that AI usage in analyst reports elicits significantly stronger reactions from capital markets. Specifically, reports with higher AI usage are associated with larger absolute cumulative abnormal returns (|CAR|) around the forecast date,

indicating that markets respond more strongly to analyst research that is more likely to incorporate generative AI. These results suggest that AI-assisted reports contain greater information content when they deliver investment advice. We conclude that GenAI adoption increases the informativeness of analysts' investment research.

Finally, we study how generative AI empowers analysts' investment research. Using GPT-5.1 as a reasoning agent, we extract detailed research characteristics from analyst reports, including valuation models used, public and private information sources, topics covered, and stakeholders discussed. We find that reports with higher AI usage feature more diverse valuation models, cover broader topics, and address stakeholder impacts more extensively, compared to reports with lower AI usage. We further examine whether AI usage in analyst reports is associated with differences in reasoning patterns. Using LIWC-22, a linguistically based framework widely adopted in prior research (Boyd et al. 2022; Li et al. 2021; Li et al. 2025), we construct measures that capture how analysts process and communicate information rather than general writing style. Specifically, we analyze four dimensions: analytical and authentic language patterns, as well as the use of causal and differentiation terms that reflect explanation-oriented and nuanced reasoning. We find that AI usage is positively associated with all four dimensions, suggesting that generative AI is linked to more structured, explanatory, and nuanced communication, consistent with deeper analytical reasoning rather than purely stylistic changes. These patterns suggest that analysts leverage AI not simply as a writing tool, but as an analytical and reasoning tool that deepens and broadens the scope of investment research.

We make two important contributions. First, we develop a supervised machine learning model to detect AI assistance in analyst reports. Unlike studies that rely on

proprietary commercial detectors (Blankespoor et al., 2024; Bradshaw et al., 2025; Hirshleifer et al., 2025), our approach improves measurement precision and transparency by training a customized model directly on analyst reports. Our measurement helps to uncover the black box of AI detection by bridging linguistic features with investment research metrics. To the best of our knowledge, this study is the first to train and implement an AI detector that is specifically customized for analyst reports, a unique information-production setting. While this approach cannot capture every aspect of individual AI usage, our paper provides the first scalable and reproducible way to quantify analyst-level AI adoption, addressing the growing demand for credible evidence on how AI affects individual behavior in capital markets.

Second, we extend the emerging literature on AI adoption in finance, which has largely focused on institution- or industry-level measures (Blankespoor et al., 2024; Kim et al., 2024), by shifting attention to AI usage at the individual analyst level. This dimension remains largely unexplored but is crucial for understanding how generative AI shapes the production of investment research. Prior to the widespread diffusion of tools such as ChatGPT, AI applications in finance were primarily institution-driven (Coleman et al., 2022; Davenport & Bean, 2017). For example, Shanthikumar and Yoo (2025) show that investment banks' AI investments are positively associated with both the frequency and accuracy of earnings forecasts. More recent studies examine GenAI-related shocks to analyst activity, such as Italy's ChatGPT bans (Bertomeu et al., 2025) or the introduction of GenAI platforms in financial data services (Xue et al., 2025). Our paper complements this literature by measuring AI usage directly at the analyst level using linguistic signals in research reports, which allows us to infer both whether and how individual analysts

incorporate AI into their work, including writing, information acquisition, valuation, topical coverage, and reasoning patterns. We further exploit analyst-level variation in workload shocks from overlapping earnings calls to establish causality and to study the distributional effects of AI adoption across analysts.

The remainder of the paper is organized as follows. Section 2 provides literature review and hypothesis development. Section 3 describes the data and outlines our research design, including the construction of our AI usage measure. Section 4 presents the baseline empirical results. Section 5 discusses our identification strategies. Section 6 examines potential channels. Section 7 examines cross-sectional variation in the impact of AI and the determinants of AI usage. Section 8 concludes.

2. Prior Literature and Hypothesis Development

2.1 AI and Analyst Forecasts

The rapid evolution of artificial intelligence (AI) has transformed how information is produced, disseminated, and consumed. ChatGPT, for instance, achieves human-level performance on complex professional tasks and demonstrates multilingual capabilities across diverse knowledge domains (OpenAI, 2023). Not surprisingly, AI can generate substantial productivity gains (Acemoglu 2025; Eisfeldt et al. 2024; Eloundou et al. 2024; Jia et al. 2024; Noy & Zhang 2023). Noy & Zhang (2023) show that AI assistance can materially improve productivity, for example, reducing the time required for writing tasks by roughly 40 percent while simultaneously improving quality by about 18 percent. Another key benefit of AI lies in its capacity to learn from large amounts of information, enabling it to generate coherent and contextually rich outputs. OpenAI (2025a) notes that

ChatGPT "learn patterns from large amounts of information, including text, images, audio, and video." More recent AI incorporates Retrieval-Augmented Generation (RAG)[2], enabling models to automatically fetch relevant external information at runtime to enrich their responses (OpenAI, 2025b). While such technological breakthroughs have raised concerns about the restructuring of task allocation, with AI complementing high-skill roles that rely on information processing and judgment (Acemoglu & Restrepo, 2018), capital markets are also beginning to grapple with the dual implications of AI for both information producers (e.g., managers, analysts) and information users (e.g., investors, regulators) (Bertomeu et al., 2025; Blankespoor et al., 2024; Bradshaw et al., 2025; Kim et al., 2024; Shanthikumar & Yoo, 2025; Sheng et al., 2025).

Turning to information intermediaries, AI tools are transforming how they process data, draft reports, and deliver insights. While traditional automation technologies have long assisted financial analysts with data collection and modeling, such as Robo-Analysts (Coleman et al., 2022), the rise of LLMs enables data analytics and human-like text generation, which are especially suited for tasks involving communication, synthesis, and interpretation. In the context of financial analysis, these capabilities can enhance analyst productivity both through task automation and task complementarity. Recent research provides early evidence of these effects. Shanthikumar and Yoo (2025) document that the magnitude AI investments by investment banks is positively related to both the frequency and accuracy of earnings forecasts. These effects are most pronounced for firms with high forecasting difficulty, such as those facing greater fundamental uncertainty or information

[2] RAG (Retrieval-Augmented Generation) is an OpenAI-endorsed mechanism. According to the OpenAI Help Center, RAG "improves a model's responses by injecting external context into its prompt at runtime. Instead of relying solely on the model's pre-trained knowledge, RAG retrieves relevant information from connected data sources.

asymmetry. Other studies examine broader implications. Blankespoor et al. (2024) argue that AI-generated text can affect investor interpretation, while Bertomeu et al. (2025) use the temporary ChatGPT ban in Italy to show that restrictions on AI reduce public information processing, leading to decreased news engagement.

2.2 Sell-Side Analyst, Earnings Forecast, and Analyst Reports

Sell-side equity analysts are important information intermediaries who facilitate the efficient functioning of capital markets by synthesizing complex firm-related information into forward-looking insights. Among their various outputs, earnings forecasts are particularly influential, serving not only as direct inputs into investor valuation models but also as the basis for analysts' stock recommendations (Bradshaw, 2004). These forecasts help reduce information asymmetry between firms and investors, thereby enhancing price discovery and capital allocation. Analysts' outputs are closely followed by a wide range of stakeholders, including investors, regulators, the media, and even corporate managers (Bradshaw, 2011). Moreover, recent work highlights that analysts serve dual roles: not only do they discover new information, but they also reframe existing disclosures to reduce investors' information processing costs (Huang et al. 2018).

Despite their central role in capital markets, the process through which sell-side analysts generate earnings forecasts remains largely unobservable. In theory, analysts synthesize vast and heterogeneous information inputs, including corporate disclosures, macroeconomic trends, peer performance, and proprietary data, and exercise substantial discretion in how these inputs are weighted and interpreted. However, this process is rarely observable and not well understood by external stakeholders, creating ambiguity about the origin and objectivity of forecast outputs. For instance, Barniv et al. (2009) note that

analysts may strategically adjust their published forecasts in response to reputational concerns or perceived investor expectations, suggesting that forecast outputs may be as much a product of incentive-driven judgment as analytical rigor. The "black box" nature is further compounded by analysts' selective use of peer firms when forming performance benchmarks (De Franco et al., 2015), and the potential influence of cognitive or social biases such as herding (Clement & Tse, 2005). As Chi et al. (2025) emphasize, analysts often lack transparency in their incorporation of "big data" sources. Therefore, it is difficult to know what tools they use in their analysis.

It is worth noting that beyond numerical forecasts and recommendations, analysts also provide reports with a rich textual medium that encapsulates analysts' qualitative assessments, strategic insights, and interpretive judgments. These reports are often viewed by institutional investors as the most critical component of sell-side research, surpassing stock picks and earnings estimates in perceived value (Huang et al., 2014). Analysts synthesize firm-specific fundamentals and industry dynamics into report discussions, typically averaging nearly eight pages per report, that both complement and contextualize their quantitative outputs. Empirical evidence suggests that this textual content provides incremental information beyond contemporaneous earnings forecasts and stock recommendations, aiding investors in interpreting conflicting signals or assessing long-term prospects. For example, Huang et al. (2014) demonstrate that negative textual sentiment in reports is weighted more heavily by markets than positive sentiment and that such text better predicts earnings growth over a five-year horizon. Complementing this view, Asquith et al. (2005) find that the strength of qualitative arguments in analyst reports, especially those written by star analysts, holds explanatory power for abnormal stock

returns, even after controlling for forecast revisions. Moreover, the readability and structure of these narratives also matter. De Franco et al. (2015) show that clearer, more readable analyst reports are associated with greater market reaction and improved analyst performance.

2.3 Predictions and Hypothesis Development

Building on the preceding discussion, we examine whether the integration of AI into analyst reports influences the quality of their forecasts. Prior literature emphasizes that analysts' work combines information discovery and interpretation, yet the underlying production process remains opaque. To the extent that AI has the potential to enhance this process by improving analysts' efficiency in processing large volumes of information (Acemoglu et al., 2022; Acemoglu, 2025) without being visible to the public, analysts may turn to AI to facilitate their forecasts.

Prior work documents that forecast accuracy varies significantly across analysts. Clement (1999) shows that accuracy is higher among analysts with greater forecasting experience and employer resources, and lower for those covering broader and more complex portfolios. Malloy (2005) finds that analysts located closer to the firms they cover are more accurate than distant peers, particularly for small or remote firms, because proximity confers informational advantages. Similarly, Bae et al. (2008) show that local analysts across countries have a consistent accuracy advantage when firms provide less transparent disclosures. Taken together, these findings imply that forecast accuracy improves when analysts can overcome informational frictions through resources or enhanced data. More recently, Chi et al. (2025) provide evidence that analysts who incorporate alternative data into their research produce significantly more accurate

earnings forecasts, suggesting that technology-enabled information processing can enhance performance. We therefore expect that AI, by facilitating rapid synthesis of large volumes of textual and numerical information, serves a similar role in reducing forecast errors and boosting stock market reaction.

H1: AI usage improves forecast accuracy and leads to stronger stock market reaction.

3 Data and Measurement

3.1 Data

Our empirical analysis focuses on sell-side equity research reports that are directly linked to analysts' earnings forecasts. To construct this dataset, we combine analyst reports with forecast-level information and firm fundamentals from multiple sources, with the analyst report serving as the primary unit of analysis. We obtain analyst research reports from Investext, a large-scale research archive maintained by Mergent Online. Investext distributes reports in PDF format and provides rich document-level metadata, including the publishing institution (investment bank or brokerage house), contributing analyst(s), document date, firm names and tickers discussed, report categories, subject tags, and report style classifications. These metadata fields allow us to identify analysts, brokers, and covered firms, and to link textual reports to forecast activity.

Our sample covers reports issued between January 1, 2022 and April 1, 2024. This window captures approximately 11 months of pre-ChatGPT activity prior to the release of ChatGPT-3.5 (November 2022), followed by the early adoption phase between the release of ChatGPT-3.5 and ChatGPT-4 (March 2023), and the subsequent period after ChatGPT-

4 but before the introduction of more advanced multimodal models (e.g., ChatGPT-4o) and the broader expansion of generative AI tools across multiple providers. During our sample period, ChatGPT-3.5 and ChatGPT-4 represent the dominant generative AI tools in practical applications. While other large language models, including Google Bard (later Gemini), Anthropic Claude, and Meta LLaMA, were introduced during this period, their adoption remained more limited relative to ChatGPT-based tools. As generative AI models evolve, their outputs increasingly resemble human-written text, making detection more challenging. This motivates our focus on the early adoption period, during which generative AI usage is emerging but remains sufficiently distinguishable for measurement.

[Figure 2 here]

Starting from the reports available in Investext between January 1, 2022 and April 1, 2024, we restrict the sample to documents for which analyst identities, firm identifiers, and forecast matches can be reliably established. We link analyst reports to earnings forecasts from the I/B/E/S database using analyst identity, firm ticker, and a narrow timing window around forecast activity. Following Huang et al. (2014) and Lv (2024), a report is matched to forecasts if it is issued within two calendar days before the forecast announcement date and two calendar days after the forecast revision date. The final analysis sample consists of 38,468 equity research reports linked to approximately 55,778 earnings forecasts, issued by 434 analysts affiliated with 73 brokerage houses. Because a single report may correspond to multiple earnings forecasts, the number of forecasts exceeds the number of unique reports. Table 1 summarizes the sample composition and its distribution over time.

Analyst reports are distributed in PDF format, which is unstructured and requires systematic preprocessing prior to textual analysis. We extract raw text from each PDF using pdfminer.six, a Python library for programmatic PDF text extraction, and process the reports using a custom text-cleaning pipeline. We preprocess analyst reports to remove formatting artifacts introduced by PDF extraction and retain only analyst-written narrative content. This includes correcting broken lines, excluding non-linguistic text such as tables and numerical artifacts, and removing broker-dealer disclosures and regulatory boilerplate in their entirety. The resulting cleaned text forms the basis for all subsequent document-level analyses, including AI-usage detection and prompt-based information extraction. Appendix B provides additional details on the text extraction and cleaning procedures.

In addition to analyst reports and earnings forecasts, we merge several external datasets to construct firm- and analyst-level control variables. Firm financial characteristics are obtained from Compustat, while stock return data are drawn from CRSP. Analyst following and forecast history variables are sourced from the I/B/E/S database. To capture analyst background characteristics that are not available in standard forecast databases, we supplement these data with hand-collected information from publicly available professional profiles. Following prior work (e.g., Li et al., 2025), we manually compile information on professional certification by reviewing analysts' online biographies and professional profiles, including LinkedIn and similar publicly accessible sources. Professional certification is identified based on whether the analyst holds the Chartered Financial Analyst (CFA) designation. We also obtain information on analysts' professional recognition from the annual Institutional Investor All-Star analyst rankings. Using the most recent All-Star lists available prior to 2022, we construct an indicator variable that equals

one if an analyst has ever been recognized as an Institutional Investor All-Star before 2022, and zero otherwise. This measure captures analysts' ex ante reputation and is treated as time-invariant over the sample period.

3.2 AI Measurement

Motivated by recent advances in GenAI detection, we construct a supervised machine learning approach designed to capture the presence of GenAI involvement in analyst report writing. Prior studies often rely on commercial detectors such as GPTZero, and Originality.ai to classify entire documents (Blankespoor et al., 2024; Bradshaw et al., 2025). By contrast, we develop our own document-level detection framework using supervised training data derived from analyst reports.

A central premise of our approach is that AI-assisted analyst reports exhibit systematic linguistic patterns that can be detected at the document level (Guo et al., 2023; Mitrović et al., 2023; Uzun, 2023). Our primary measure captures the likelihood that a report incorporates generative AI, rather than attempting to attribute AI involvement to individual textual units. A higher document-level AI probability therefore reflects broader integration of GenAI into the analyst's research process, potentially extending beyond the mechanical drafting of text to other aspects of research production.

To construct training data for document-level detection, we draw analyst reports issued in year 2021, prior to the public release of ChatGPT, a period during which generative AI usage in sell-side research was negligible. For each report, we generate AI-assisted document variants by rewriting approximately 50 percent of the original analyst-written content using GPT-3.5-Turbo and GPT-4, while preserving the original analytical meaning, factual content, and professional writing style. The rewritten content is integrated

back into the original reports to form partially AI-assisted documents, which are paired with fully human-authored reports as ground-truth data for supervised training and evaluation of document-level AI-detection models. This 50 percent specification serves as our primary training design, while alternative proportions of AI-assisted content (e.g., 25 % and 75 %) are examined in robustness analyses.

We compute 25 document-level linguistic features following Guo et al. (2023) and Uzun (2023) (Table B1). These features capture multiple dimensions of writing style and structure, including lexical richness, syntactic complexity, spelling and grammar accuracy, sentiment tone, readability, and fluency. Using these features, we train three supervised learning classifiers, Random Forest, Logistic Regression, and XGBoost, on 80 percent of the ground-truth data and evaluate their performance on the remaining 20 percent. As shown in Table B2, XGBoost consistently outperforms the alternative models in distinguishing human-authored versus AI-assisted analyst reports based on the document-level linguistic features. Accordingly, we use XGBoost as our primary model to construct the document-level AI usage measure in the pre- and post-GPT sample. Figure 2 reports the relative importance of linguistic features in the XGBoost model.

[Figure 2 here]

To identify potential AI-assisted analyst reports, we apply the trained XGBoost model to all analyst reports in our sample, including both the pre- and post-ChatGPT periods. For each report, the model produces a document-level probability (AI_Prob) that the report incorporates generative AI. In the empirical analysis, we use AI_Prob as a continuous measure of AI usage and interact it with an indicator for the post-ChatGPT period to examine how the effects of AI usage differ after the public release of ChatGPT.

We acknowledge that our detection approach may primarily capture AI usage in surface-level editing or drafting of analyst reports rather than in deeper analytical tasks. Any such measurement error would tend to attenuate estimated effects, thereby biasing our results against finding significance and making our tests conservative. At the same time, given the substantial efficiency gains from using generative AI, analysts who rely on AI for drafting are plausibly more likely to use it in other aspects of the research process. As a result, higher values of AI_Prob likely reflect broader integration of GenAI into analysts' research activities.

3.3 AI Trend

Following the public release of ChatGPT-3.5 in November 2022, we observe a clear upward trend in AI usage in analyst reports, as reflected in Figure 3. The increase becomes more pronounced after the introduction of GPT-4 in March 2023, consistent with broader adoption of generative AI tools in professional settings. Notably, the growth in AI usage appears relatively muted in early 2023. This pattern coincides with reported restrictions on ChatGPT usage. Due to compliance risk and data leakage concerns, several banks have been widely reported in the news for restricting the use of ChatGPT starting from late February 2023. These banks include JPMorgan, Bank of America, Citi, Deutsche Bank, Goldman Sachs, and Wells Fargo. For example, JPMorgan was reported to have restricted employees' use of ChatGPT on February 22, 2023, by CNN, Bloomberg, Forbes, Wall Street Journal, among others.

As these restrictions were gradually relaxed or supplemented by internally developed AI systems, AI usage in analyst reports resumed its upward trajectory. For example, JPMorgan stated in its annual report released on April 3, 2023 that "We're

imagining new ways to augment and empower employees with AI through human-centered collaborative tools and workflow, leveraging tools like large language models, including ChatGPT." JPMorgan Chase is giving its over 60,000 employees (including the asset & wealth management division) the LLM Suite, an AI assistant powered by ChatGPT maker OpenAI, reported by CNBC on 8/9/2024. This initial version provided a secure, internal "ChatGPT-like" tool for tasks such as drafting emails, summarizing documents, and generating ideas. We thus take these three brokers as our treatment group. Figure 3 plots the time-series evolution of the mean and median document-level AI usage probability (AI_Prob) for analyst between January 1 2022 and April 1 2024. In untabulated results, we find that reports issued by brokers subject to ChatGPT bans exhibit lower AI usage during the restriction period compared to brokers without such restrictions.

[Figure 3 here]

3.3 Other Measurement: Forecast Accuracy, Market Reactions, and Timeliness

We measure forecast accuracy using Accuracy, defined as the negative of the proportional mean absolute forecast error (PMAFE). Following Ben-Rephael et al. (2023) and Peng et al. (2022), PMAFE is computed as the absolute forecast error for a given analyst, defined as the absolute difference between the forecasted and actual earnings per share, scaled by the average absolute forecast error across all analysts covering the same firm and forecast date. Higher values of Accuracy indicate relatively higher forecast accuracy compared to contemporaneous peers, while lower values indicate poorer relative accuracy.

As a second main outcome, we examine the magnitude of the stock market reaction to analyst reports using |CAR|, which is defined as the absolute value of the cumulative three-day abnormal return around the forecast date (day -1 to +1), where abnormal returns are calculated using the market-adjusted model. This measure captures the strength of investors' responses to the information contained in analyst reports, regardless of whether the reaction is positive or negative.

Examining |CAR| is particularly informative in the context of generative AI. Analyst research reports convey substantially more information than point earnings forecasts, including narrative interpretation, valuation analysis, and the synthesis of firm-specific, industry, and macroeconomic information. Prior research shows that the textual content of analyst reports provides incremental information to investors beyond contemporaneously released quantitative signals and elicits economically meaningful market reactions (Huang et al., 2014). Generative AI may influence |CAR| through channels distinct from numerical forecast precision. By producing more structured and coherent narratives, reducing linguistic noise, and improving the framing of key information, AI-assisted reports may lower investors' information-processing costs and facilitate faster assimilation of complex analyses. Consistent with this view, prior studies show that features related to the presentation and communication of analyst research can significantly affect the magnitude of market reactions (Peng et al., 2022; Birru et al., 2022). Accordingly, |CAR| allows us to assess whether GenAI usage affects not only analysts' forecast precision but also the extent to which analyst reports move capital markets.

4. Baseline Results and Heterogeneity in the Effects of AI

4.1 Main Test

Generative AI may help analysts process information more efficiently, draft forecasts with less effort, and potentially improve the precision of earnings estimates. To examine whether the use of generative AI in analyst reports is associated with forecast accuracy and capital market reactions, we estimate the following baseline specification, following the empirical design in Li et al. (2025):

$$Y_{i,f,t} = \alpha + \beta AI_Prob_{i,f,t} + \lambda' Controls + \eta_{f,y} + \theta_{i,y} + \varepsilon_{i,f,t}$$

where $Y_{i,f,t}$ denotes either Forecast Accuracy associated with analyst i's report for firm f at time t. Improvements in forecast accuracy capture analysts' internal performance and reflect the quality of their information processing and judgment. The variable $AI_Prob_{i,f,t}$ is a continuous, forecast-level measure capturing the probability that the analyst report incorporates generative AI.

Control variables include analyst and broker characteristics commonly used in the forecast accuracy literature: FirmsCovered, IndCovered, AnalystFirmExperience, ForecastFrequency, Horizon, and TOP10 (Li et al. 2025, Flam et al. 2023, Harford et al. 2019). These controls capture heterogeneity in analyst ability, workload, and information environment. Firm-level characteristics are not included separately, as they are absorbed by the firm–year fixed effects.

The unit of observation is at the analyst–firm–forecast level. The specification includes firm–year fixed effects and analyst–year fixed effects which absorb time-varying firm fundamentals and brokerage-level factors that may jointly affect analysts' research practices and forecast outcomes. This fixed-effects structure allows us to benchmark analysts' forecasts within the same firm-year and analyst-year environments. Standard

errors are clustered at the analyst–year-quarter level to account for serial correlation in analysts' forecast behavior over time (Li et al., 2025). The final estimation sample consists of 55,778 observations for forecast accuracy regressions and market reaction regressions.

Table 2 presents descriptive statistics for the main variables used in our empirical analysis of analyst forecasts.

Table 3 reports the results from regressions examining the association between AI usage in analyst reports and forecast accuracy. Column (1) and (2) shows that the coefficient on AI_Prob is positive and statistically significant with estimates of (0.039, t=4.15) and (0.026, t=2.88), respectively, indicating that analyst reports with a higher probability of incorporating generative AI are associated with greater forecast accuracy. Column (1) includes firm–year and analyst–year fixed effects, while Column (2) additionally includes control variables. These results are obtained after controlling for a rich set of analyst and broker characteristics. Using the sample mean of Accuracy as a benchmark, the point estimate implies a meaningful improvement in relative forecast accuracy as AI usage increases. The coefficient estimates of 0.039 and 0.026 in Columns (1) and (2) of Table 3 correspond to approximately 22% and 15% of the sample mean, respectively.[3]

[Table 3 here]

4.2. Moderators of AI's Effect on Forecast Accuracy

We implement a subsample analysis by partitioning the sample based on analyst workload and characteristics and estimating the regressions separately across groups. Analysts have limited attention when facing concurrent information events such as

[3] The economic magnitude is computed by scaling the estimated coefficient on AI_Prob by the sample mean of Accuracy (0.174; Table2).

overlapping conference calls (De Franco et al., 2025). For Table 4, we partition the sample based on analyst workload (busyness) measures, including FirmsCovered, IndCovered, and ForecastFrequency. The results indicate that the accuracy benefits of AI usage are not uniform across analysts. Columns (1) and (2) of Table 4 show that the effect of AI usage on forecast accuracy is stronger for analysts covering more firms, and Columns (3) and (4) show that the effect is stronger for analysts covering more industries. Columns (5) and (6) further show that the effect is stronger for analysts with higher forecast frequency. These findings are consistent with AI tools helping analysts manage higher workloads and time constraints.

[Table 4 here]

Prior research shows that differences in analysts' information advantages and access to resources lead to variation in forecast accuracy (Malloy, 2005; Chi et al., 2025). Consistent with this view, AI can substitute for tasks traditionally performed by human labor (Acemoglu & Restrepo, 2018). For Table 5, we partition the sample based on human capital measures, including CFA status, All-Star recognition, and AnalystIndExperience.

The results in Columns (1)–(4) indicate that the effect of AI usage varies across analyst credentials, with stronger effects observed for analysts without CFA certification or All-Star recognition. Columns (5) and (6) further show that AI helps compensate for differences in industry experience. These findings suggest that generative AI can substitute for aspects of human capital and industry knowledge, enabling analysts without CFA certification or All-Star recognition, as well as those with less industry experience, to produce more accurate forecasts.

[Table 5 here]

Consistent with these findings, untabulated determinant analysis shows that All-Star analysts and those holding a CFA designation are less likely to rely on generative AI. This negative association is consistent with highly credentialed analysts relying more on established expertise and professional judgment rather than AI assistance. One possible explanation relates to reputational and professional risk considerations. Analysts with stronger reputational capital may face higher expected costs from potential misperceptions about AI-assisted research, including concerns about originality, authorship, or professional judgment. Similarly, CFA charterholders operate under strict professional standards emphasizing independent analysis and accountability, which may lead them to adopt a more conservative approach toward emerging tools whose regulatory and compliance implications are still evolving. As a result, even when access to generative AI is broadly available, analysts with higher reputational or professional stakes may optimally limit their reliance on such tools.

5. Identification and Capital Market Reaction

5.1. Identification: Overlap with Conference Call

A critical concern is that our AI detection measure may conflate true AI usage with underlying report informativeness or analyst ability. It is possible that our detection approach captures reports that are inherently more informative or better written, rather than genuine reliance on generative AI by analysts, even though our model is based solely on linguistic features. If so, this would give rise to a classical endogeneity concern: more informative reports may be both more accurate and more likely to elicit stronger market reactions, independent of AI usage. More skilled analysts may write clearer and more

comprehensive reports, while also being more inclined to adopt AI tools, generating potential confounding bias. Similarly, analysts may deploy AI selectively for reports they expect to be particularly important, valuable, or complex, creating selection concerns. Finally, analysts with higher underlying ability or better performance may be more likely to adopt AI, raising the possibility of reverse causality.

To address these endogeneity concerns, we implement an identification strategy that exploits overlapping earnings calls, which generate exogenous, analyst-specific time pressure that increases reliance on AI tools but is plausibly orthogonal to analyst ability, report informativeness, or forecast outcomes. In addition, we provide supplementary validation in the Appendix using firmwide ChatGPT bans, which serve to assess whether our AI usage measure responds to exogenous restrictions on access to generative AI.

Conference calls have become a common and important disclosure channel for public firms since the passage of Regulation Fair Disclosure (Reg FD) in 2000. Prior studies show that earnings conference calls convey price-related information beyond press releases (Bowen et al., 2002; Bushee et al., 2004; Kimbrough, 2005). Accordingly, earnings conference calls are information hotspots for analysts and investors to acquire information. Driskill et al. (2020) show that analysts are subject to limited attention when multiple firms they cover report earnings on the same day. De Franco et al. (2025) further show that overlapping calls serve as attention shocks to analysts. More importantly, they find no clear evidence that managers strategically adjust call timing to avoid overlaps, and overlap is mainly driven by pre-determined factors such as headquarter time zone and fiscal calendar. Their result suggests that overlapping calls are exogenous shocks to analysts' attention and workload. Since most analysts release reports soon after the earnings calls,

overlapping calls would create an exogenous time pressure for analysts to adopt AI in their analyst reports so that they can release their reports on time.

Overlapping earnings calls generate quasi-random, analyst-specific variation in time pressure that shifts AI usage for reasons unrelated to analyst ability or report quality. When multiple firms covered by the same analyst hold earnings calls on the same day, analysts face a temporary workload shock that reduces available time for report writing and increases reliance on AI tools. Crucially, the scheduling of earnings calls is determined entirely by firms and is orthogonal to analysts' forecast accuracy, writing talent, effort, or preferences. Analysts cannot influence whether two covered firms announce earnings simultaneously, and forecast accuracy cannot cause overlapping calls, eliminating the possibility of reverse causality. This design ensures that variation in AI usage induced by overlap stems from exogenous time constraints rather than analyst skill or selective deployment of AI for complex or high-value reports.

Following De Franco (2022), we define overlapping calls as multiple conference calls occurring on the same date for firms covered by the same analyst. We construct the report-level overlapping dummy variable, Overlapping Dummy, which takes the value of one if a report is released on a date when at least two firms under the analyst's coverage hold earnings calls, and otherwise zero. Our analyst report sample shows that 27% of reports are released on the same date as earnings calls.

[Table 6 here]

We next use overlapping earnings calls as an instrumental variable for AI usage in analyst reports and reexamine the effect of AI usage on forecast accuracy. We follow the same specification as that in Table 3, except that we estimate the model using two-stage

least squares (2SLS). Table 6 presents IV regressions using overlapping earnings calls occurring on the same date as the instrument. Column (1) reports the first-stage results, where the dependent variable is AI_Prob, a document-level measure capturing the probability that an analyst report incorporates generative AI. The overlapping dummy has a positive and statistically significant effect on AI_Prob, indicating that AI usage is higher on report release dates that coincide with overlapping earnings calls for firms covered by the analyst. The results of the F-test, the underidentification test (Kleibergen-Paap rk LM statistic), and the weak identification test (Kleibergen-Paap Wald rk F statistic) confirm the relevance of overlapping calls as a valid instrument for AI usage. Column (2) presents the second-stage results using Accuracy as the dependent variable. Consistent with our baseline findings, instrumented AI usage is positively and significantly associated with forecast accuracy.

5.2 Capital Market Reaction

We next investigate whether the capital market reacts differently to analyst reports that exhibit higher levels of generative AI usage. While forecast accuracy captures analysts' internal performance, it does not fully reflect how investors process and respond to analyst research. We therefore separately examine capital market reactions using cumulative abnormal returns.

We measure investor reactions using the absolute value of cumulative abnormal returns (|CAR|) around the forecast date, which captures the magnitude of the short-window stock price response to analyst research. |CAR| is defined as the three-day market-adjusted abnormal return (−1, +1) around the forecast date. Focusing on |CAR| allows us

to assess whether AI-assisted reports convey information that investors perceive as more salient or impactful, regardless of the direction of the price response. This approach is consistent with prior studies that use short-window abnormal returns to capture the informativeness and salience of analyst disclosures (Birru et al., 2022; De Franco et al., 2015; Huang et al., 2014; Huang et al., 2022).

To isolate the association between AI usage and market reactions, we estimate specifications that control for the same set of control variables as in our main accuracy regressions (Table 3), while absorbing firm-level and analyst-level factors through firm–year and analyst–year fixed effects. This approach extends Huang et al. (2014), who primarily controlled for firm size and book-to-market ratio, and ensures comparability across our empirical settings.

As shown in Table 7, the coefficient on AI_Prob is positive and statistically significant, with estimates of (0.002, t=1.78) and (0.002, t=1.99) in Columns (1) and (2), respectively, suggesting that analyst reports with a higher probability of incorporating generative AI are associated with stronger immediate market reactions. This evidence indicates that AI-assisted reports convey information that investors perceive as more salient or impactful. Consistent with prior evidence that investor reactions depend on the perceived credibility and informativeness of analyst outputs (Peng et al., 2022), generative AI may influence how information is structured, framed, and communicated in analyst reports, thereby amplifying the intensity of investor responses rather than their sign.

[Table 7 here]

6. Mechanisms and Heterogeneity in the Effects of AI

6.1 Mechanism: How AI Empowers Investment Research?

In this section, we delve into analyst reports and examine how generative AI empowers analysts' investment research. Beyond using generative AI as a writing tool, analysts can utilize it as an analytical and reasoning tool to improve their research quality. For example, analysts may use generative AI to process balance sheets and filings to improve their valuation models. At an extreme, analysts may use generative AI to create a whole first draft without any human input, although this unlikely considering the reputation and compliance risk. Through interactions with GenAI, analysts may exhibit more analytical reasoning patterns, including greater use of causal explanations and more nuanced thinking. We examine several dimensions of analysts' research quality well-documented in the literature.

As the goal of analyst report is to guide investors through its forecasts, we first look into valuation models. Ben-David and Chinco (2025) find that most analysts set price targets using multiples, such as multiplying a firm's expected EPS times its trailing PE ratio, in their reports, while few analysts use the present-value based models. Accordingly, we use GPT-5.1 to read through analyst reports and ask it to find out the number of valuation models used in the report. Our prompt is in the Appendix D. Our variable, Valuation, equals zero if a report only employs one type of valuation model, and equals to one if a report uses both multiple-based and present-value based valuation models.

Sheng et al. (2025) use 14 topics that ChatGPT can generate forward-looking information/signal based on earnings calls and they find these signals can predict hedge fund flows. These 14 topics include US economy, global economy, financial prospect of the firm, financial prospect of the industry, earnings, revenue, wages/salaries, etc. We ask

GPT-5.1 to classify the number of distinct topics (up to 14) mentioned in a report accordingly.

Li et al. (2025) show that there is an increasing trend of environmental and social-related discussions in both earnings calls and analyst reports. We then look at the number of distinct stakeholders mentioned in a report. We include nine types of stakeholders, including shareholders, debt holders, the management team, employees, customers, suppliers, competitors, regulators, and communities.

Brown et al. (2015) use a survey to show that private communication with management is a more useful input to analysts' forecasts than their own research and public information, such as filings. Bradley, Jame, and Williams (2022) show that non-deal roadshows play an important role in analysts' biased forecasts and create information advantage for institutional investors. Bradley, Jame, and Williams (2022) further show that hosting analysts issue favorable forecasts proceeding site visits to firms' headquarters and plants. Taking all the documented information sources used by analysts, we ask GPT-5.1 to classify the numbers of private and public meetings with the management separately. Private communications include phone/videos/emails with the management, site visits such as bus tours or field trips, non-deal roadshows, surveys, channels check, fireside chats, and social events such as golf outings. Public communications include earnings calls, company conference presentations, special calls, analyst/investor days, M&A calls, shareholder/analyst calls, annual shareholder meetings, guidance/update calls, and sales/trading statement calls. To avoid counting duplications, we count the number of distinct sources only. For the count-based measures above, we scale them with the total number of words in a report to measure the density of each category in a report.

We examine how AI usage in reports relates to these valuation-related research depth measures. Table 8 presents the results. We use the same specification as previous tables. The results show that AI usage is positively associated with the number of valuation models, the number of distinct topics, and the number of distinct stakeholders, while it is negatively associated with private communication and not significantly associated with public communication discussed in the report. These findings suggest that generative AI is associated with more comprehensive analytical content in analyst research, while potentially reducing reliance on private information sources[4].

[Table 8 here]

We next examine whether AI usage in analyst reports is associated with differences in analysts' reasoning patterns. Table 9 reports these results. To capture reasoning styles, we use LIWC-22 to construct linguistically based measures that have been widely adopted in fields such as politics and psychology (Boyd et al., 2022; Li et al. 2025). These measures are designed to reflect how individuals process and communicate information, rather than general writing style, and are not used in training our machine learning detection model.

Specifically, we consider four dimensions of reasoning. *Analytical* measures the extent of logical and formal reasoning.[5] *Authentic* captures the extent of sincere and straightforward communication. These two measures are based on a proprietary formula that summarizes several categories of function words. *Causation* measures the use of words that explains relationships between events or outcomes (e.g., because, therefore, make, and

[4] Campbell, Zheng, and Zhou (2025) find that the relative ratio between the number of quantitative (numeric numbers) and the number of words correlates with earnings announcement returns as numerical disclosures reduce information risk. We follow their measurement by asking GPT-5.1 to count the number of numbers and words in a report.

[5] LIWC website provides more details on these metrics: https://www.liwc.app/help/liwc.

why), reflecting explanation-oriented reasoning. *Differentiation* captures the fraction of words related to conditions, contrasts, and alternative possibilities (e.g., but, not, if, and or), reflecting more nuanced reasoning. For the count-based measures above, we scale them with the total number of words in a report to measure the density of each category in a report.

The results show that AI usage is positively associated with all four reasoning dimensions. Together, these findings suggest that generative AI is linked to more structured, explanation-oriented, and nuanced communication, consistent with deeper analytical processing rather than merely stylistic changes.

[Table 9 here]

6.2 Robustness to Alternative AI Detection Models

Table C1 assesses the robustness of the relation between AI usage and forecast accuracy to alternative AI detection models and benchmark definitions. We re-estimate the baseline accuracy specification using AI usage probabilities generated from different machine learning classifiers and under alternative definitions of AI involvement. Columns (1) and (2) replace the baseline AI usage measure with AI_ProbLow and AI_ProbHigh, which capture lower- (25%) and higher-intensity (75%) AI involvement based on alternative training benchmarks. Columns (3)–(5) construct AI usage probabilities using a Logistic Regression classifier under different AI-intensity benchmarks, while Columns (6)–(8) present analogous specifications using a Random Forest classifier (model performance metrics are reported in Appendix B). Column (9) reports results using an alternative AI usage measure constructed following the same XGBoost-based procedure

and 50% AI-intensity benchmark as in the baseline specification but trained on analyst reports issued between 2000 and 2020.

Across all specifications, AI usage measures, whether defined using alternative intensity thresholds, classifiers, or training samples, remain positive and statistically significant, with magnitudes comparable to those in the baseline results. Notably, both lower-intensity and higher-intensity AI usage measures are positively associated with forecast accuracy, suggesting that the benefits of generative AI are not limited to heavy reliance on such tools. Even modest levels of AI assistance appear to improve analysts' forecasting performance. These additional analyses reinforce the interpretation that generative AI serves as a productivity-enhancing tool that improves both the efficiency and effectiveness of analysts' forecasting activities, rather than being an artifact of a particular measurement approach.

[Table C1 here]

7. Conclusion

This study examines whether and how analysts' use of generative AI influences forecast quality and capital market outcomes. Using a supervised machine-learning approach to detect AI-assisted content in analyst reports, we show that AI-assisted reports are associated with improved forecast accuracy. In the capital market, we document that investors react more strongly to AI-assisted reports, indicating that markets perceive such reports as incrementally informative.

We establish support for a causal interpretation using exogenous shocks from overlapping earnings calls. We show that overlapping earnings calls increase analysts' reliance on AI tools and that instrumented AI usage is positively associated with forecast accuracy. Our channel analysis suggests that generative AI empowers analysts' investment research through richer valuation analysis and broader informational coverage, including the integration of more valuation models and diverse analytical dimensions. Our cross-sectional analyses show that efficiency gains from AI are more pronounced for less experienced analysts and for analysts facing higher workload demands, while the benefits vary across analyst characteristics. These findings suggest that generative AI functions both as a productivity-enhancing tool and as a partial equalizer of skill differences across analysts.

This study contributes to the growing literature on AI in capital markets by providing large-sample empirical evidence from the analyst setting. More broadly, our findings highlight both the benefits and the risks of AI adoption in information production. Future research could explore how analysts balance efficiency and depth when integrating AI into their workflows, whether AI adoption shapes analysts' career trajectories, and how investors discern between human- versus AI-generated insights. As with any empirical study, our analysis is subject to limitations. Our identification of AI usage relies on linguistic markers of generative AI in analyst reports, which may not capture all forms of AI assistance. Our approach also contributes a transparent and reproducible framework for measuring AI usage in analyst reports, addressing limitations of proprietary detection tools commonly used in prior studies.

References

Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40(121), 13–58.

Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation, and work. In The economics of artificial intelligence: An agenda (pp. 197–236). University of Chicago Press.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40, S293–S340.

Adam, G. A., Cui, A., Tian, E., Thomas, E., Schnell, J., Tian, J. J., Dronavalli, A., & Lee, D. (2024). GPTZero: Scalable and explainable detection of LLM-generated texts. arXiv preprint.

Alam, F., Nakov, P., Habash, N., Gurevych, I., Chowdhury, S., Shelmanov, A., Wang, Y., Artemova, E., Kutlu, M., & Mikros, G. (Eds.). (2025). Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect).

Alamleh, H., AlQahtani, A. A. S., & ElSaid, A. (2023). Distinguishing human-written and ChatGPT-generated text using machine learning. In 2023 Systems and Information Engineering Design Symposium (SIEDS) (pp. 154–158). IEEE.

Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics. Princeton University Press.

Asquith, P., Mikhail, M. B., & Au, A. S. (2005). Information content of equity analyst reports. *Journal of Financial Economics*, 75(2), 245–282.

Bae, K. H., Stulz, R. M., & Tan, H. (2008). Do local analysts know more? *Journal of Financial Economics,* 88(3), 581–606.

Barniv, R., Hope, O. K., Ourring, M. J., & Thomas, W. B. (2009). Do analysts practice what they preach? *The Accounting Review*, 84(4), 1015–1039.

Bastianello, F., Décaire, P. H., & Guenzel, M. (2025). Mental models and financial forecasts. Working paper.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A text-based analysis of corporate innovation. *Management Science*, 67(12), 7604–7627.

Ben-David, I., & Chinco, A. (2025). Capital Budgeting For EPS Maximizers.

Ben-Rephael, A., Carlin, B., Da, Z., & Israelsen, R. (2025). Security analysis and the collection of hard and soft information. Working paper.

Ben-Rephael, A., Carlin, B. I., Da, Z., & Israelsen, R. D. (2025). Uncovering the hidden effort problem. *The Journal of Finance*, 80(2), 1261–1296.

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*.

Birru, J., Gokkaya, S., Liu, X., & Stulz, R. M. (2022). Are analyst short-term trade ideas valuable? *The Journal of Finance*, 77(3), 1829–1876.

Blankespoor, E., Croom, J., & Grant, S. (2025). Generative AI and investor processing of financial information. Working paper.

Blankespoor, E., deHaan, E., & Li, Q. (2025). Generative AI in financial reporting. Working paper.

Bowen, R. M., Davis, A. K., & Matsumoto, D. A. (2002). Do conference calls affect analysts' forecasts? *The Accounting Review*, 77(2), 285-316.

Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW (2022) The development and psychometric properties of LIWC-22. Working paper, University of Texas at Austin.

Bradley, D. (2023). Financial analysts. In Handbook of Financial Decision Making (pp. 356–372).

Bradley, D., Gokkaya, S., & Liu, X. (2017). Before an analyst becomes an analyst: Does industry experience matter? *The Journal of Finance*, 72(2), 751–792.

Bradley, D., Jame, R., & Williams, J. (2023). Hop on! Bus tours and corporate access. Working paper.

Bradshaw, M. T. (2011). Analysts' forecasts: What do we know after decades of work? Working Paper.

Bradshaw, M. T., Ma, C., Yost, B., & Zou, Y. (2025). Generative AI use by capital market information intermediaries: Evidence from Seeking Alpha. Working Paper.

Breitung, C., & Müller, S. (2025). *Global business networks*. Journal of Financial Economics, 166, 104007.

Brown, L. D., Call, A. C., Clement, M. B., & Sharp, N. Y. (2015). Inside the "black box" of sell-side financial analysts. *Journal of Accounting Research*, 53(1), 1–47.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2), 889–942.

Bushee, B. J., Matsumoto, D. A., & Miller, G. S. (2004). Managerial and investor responses to disclosure regulation: The case of Reg FD and conference calls. *The Accounting Review*, 79(3), 617-643.

Campbell, J. L., Zheng, X., & Zhou, D. (2025). Number of numbers: Does a greater proportion of quantitative textual disclosure reduce information risk? *Journal of Corporate Finance*, 94, 102813.

Cao, S., Jiang, W., Wang, J., & Yang, B. (2024). From man vs. machine to man+ machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160, 103910.

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

Chi, F., Hwang, B. H., & Zheng, Y. (2025). The use and usefulness of big data in finance. *Management Science*, 71(6), 4599–4621.

Chugunova, M., Harhoff, D., Hölzle, K., Kaschub, V., Malagimani, S., Morgalla, U., & Rose, R. (2025). Who uses AI in research and for what? *Research Policy*.

Clement, M. B. (1999). Analyst forecast accuracy. *Journal of Accounting and Economics*, 27(3), 285–303.

Clement, M. B., & Tse, S. Y. (2005). Herding behavior in forecasting. The Journal of Finance, 60(1), 307–341.

Coleman, B., Merkley, K., & Pacelli, J. (2022). Human versus machine: Robo-analyst vs. traditional research analyst recommendations. *The Accounting Review*, 97(5), 221–244.

Davenport, T. H., & Bean, R. (2017). How P&G and American Express are approaching AI. *Harvard Business Review*.

De Franco, G., Hope, O. K., & Larocque, S. (2015). Analysts' choice of peer companies. *Review of Accounting Studies*, 20, 82–109.

De Franco, G., Hope, O. K., Vyas, D., & Zhou, Y. (2015). Analyst report readability. *Contemporary Accounting Research*, 32(1), 76–104.

De Franco, G., Li, H., Shen, R., & Yang, S. (2025). Managerial reaction to analysts' limited attention. Working paper.

Driskill, M., Kirk, M. P., & Tucker, J. W. (2020). Concurrent earnings announcements and analysts' information production. *The Accounting Review*, 95(1), 165-189.

Eisfeldt, A. L., Schubert, G., Zhang, M. B., & Taska, B. (2024). Generative AI and firm values. *Journal of Finance*, forthcoming.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs. Science, 384(6702), 1306–1308.

Emett, S., Eulerich, M., Lipinski, E., Prien, N., & Wood, D. A. (2024). Leveraging ChatGPT for the internal audit process. *Accounting Horizons*.

Flam, R. W., Green, J., Lee, J. A., & Sharp, N. Y. (2023). Ethnic minority analysts' participation in public earnings conference calls. *Journal of Accounting Research*, 61(5), 1591-1631.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ... & Wu, Y. (2023). How close is ChatGPT to human experts? arXiv preprint.

Harford, J., Jiang, F., Wang, R., & Xie, F. (2019). Analyst career concerns, effort allocation, and firms' information environment. *The Review of Financial Studies*, 32(6), 2179-2224.

Hirshleifer, D., Peng, L., Wang, Q., Zhang, W., & Zhang, X. (2025). AI, opinion ecosystems, and finance. Working paper.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles. *Management Science*, 64(6), 2833–2855.

Huang, A. H., Lin, A. P., & Zang, A. Y. (2022). Cross-industry information sharing. *Journal of Accounting and Economics*, 74(1).

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A large language model for financial text. *Contemporary Accounting Research*.

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Information content of text in analyst reports. *The Accounting Review*, 89(6), 2151–2180.

Huang, S., Tan, H., Wang, X., & Yu, C. (2021). Valuation uncertainty and analysts' use of DCF models. *Review of Accounting Studies*.

Hwang, B. H. (2017). It pays to write well. *Journal of Financial Economics*, 124(2), 373-394.

IBM. (2024). What is shadow AI?

Irons, C. (2024). Influence of AI-like text on disclosure responses. SSRN Working Paper.

Jabarian, B., & Imas, A. (2025). Artificial writing and automated detection. Working paper.

Jia, N., Li, N., Ma, G., & Xu, D. (2025). Corporate responses to generative AI: Early evidence from conference calls. *Review of Accounting Studies*,1–55.

Ke, S. (2025). Analysts' belief formation in their own words. Working paper.

Kim, A. G., Muhn, M., & Nikolaev, V. V. (2024). Bloated disclosures. Chicago Booth Research Paper.

Kimbrough, M. D. (2005). The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80(1), 189-219.

Kirk, M. (2011). Research for sale. *Journal of Financial Economics*, 100(1), 182–200.

Li, K., Mai, F., Wong, G., Yang, C., & Zhang, T. (2025). Female equity analysts and ESG performance. *Management Science*, forthcoming

Li, Y., Mai, F., Shen, R., Yang, C., & Zhang, T. (2026). Dissecting corporate culture using generative AI. *The Review of Financial Studies*, 39, 253–296.

Loh, R. K., & Stulz, R. M. (2011). When are analyst recommendation changes influential? *The Review of Financial Studies*, 24(2), 593–627.

Lui, D., Markov, S., & Tamayo, A. (2012). Equity analysts and market risk assessment. *Journal of Accounting Research*, 50(5), 1287–1317.

Lv, L. (2024). The value of information from sell-side analysts. arXiv preprint arXiv:2411.13813.

Malloy, C. J. (2005). The geography of equity analysis. *The Journal of Finance*, 60(2), 719–755.

Maidan, L. (2025). AI could wipe out entire Wall Street teams. MarketWatch.

McKinsey. (2024). Extracting value from AI in banking.

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). ChatGPT or human? arXiv preprint.

Noy, S., & Zhang, W. (2023). Productivity effects of generative AI. *Science*, 381(6654), 187–192.

OpenAI. (2023). GPT-4.

OpenAI. (2025a). How ChatGPT and foundation models are developed.

OpenAI. (2025b). Retrieval augmented generation (RAG).

Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., ... & Tworek, J. (2025). GDPval: Evaluating AI model performance. arXiv preprint.

Peng, L., Teoh, S. H., Wang, Y., & Yan, J. (2022). Trait impressions and analyst performance. *Journal of Accounting Research*, 60(2), 653–705.

Rowlands, D. (2024). AI adoption across finance functions. KPMG Report.

Shanthikumar, D. M., & Yoo, I. S. (2025). Artificial intelligence and analyst productivity. SSRN Working Paper.

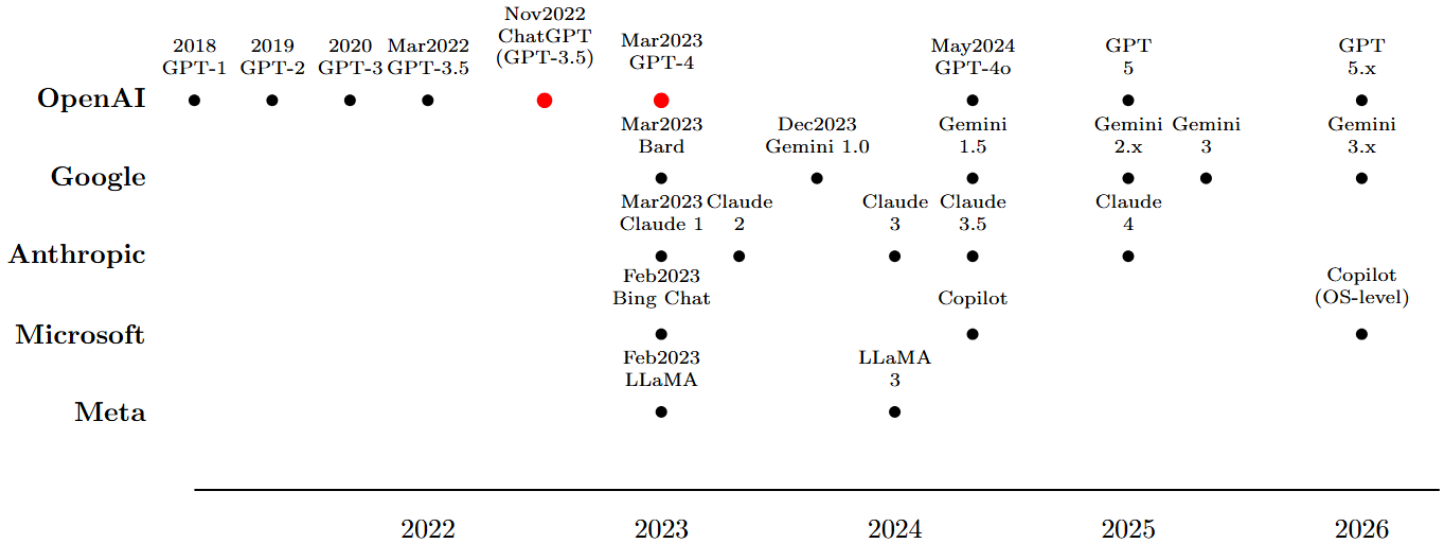
Sheng, J., Sun, Z., Yang, B., & Zhang, A. L. (2025). Generative AI and asset management. *Review of Financial Studies,* Forthcoming.

Uzun, L. (2023). Academic integrity and ChatGPT detection. *Language Education and Technology, 3*(1).

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature, 621*(7980), 672–675.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection. *Computational Linguistics, 51*(1), 275–338.

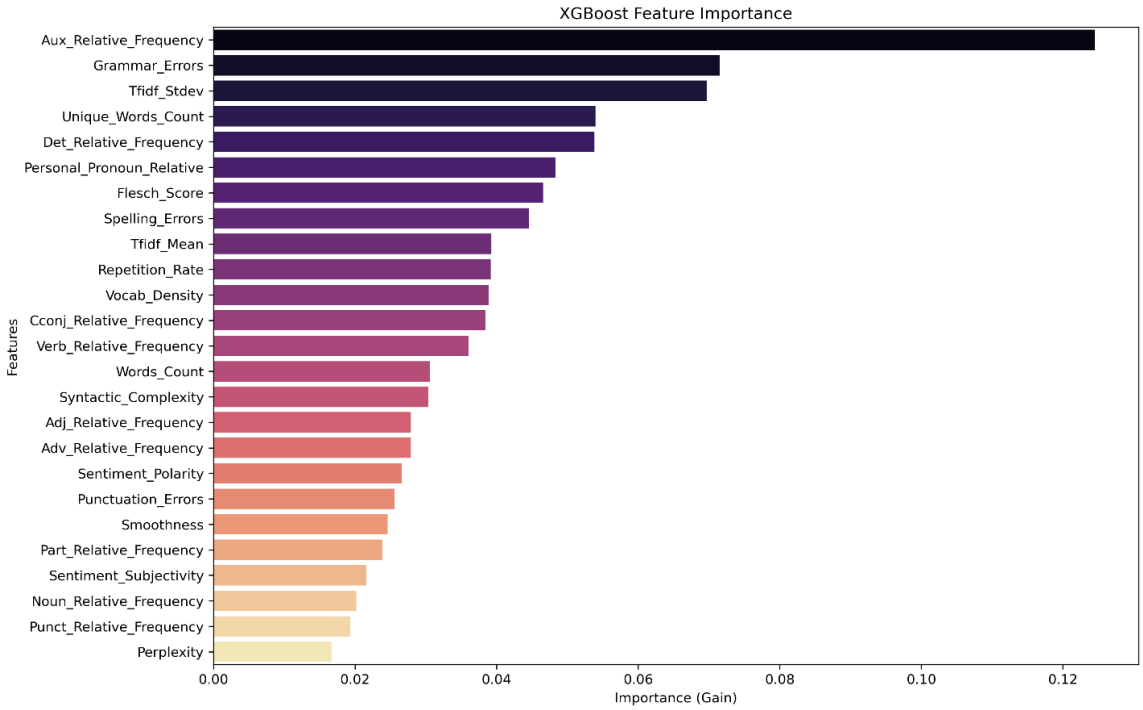
Figure 1: Timeline of Generative AI Model Development



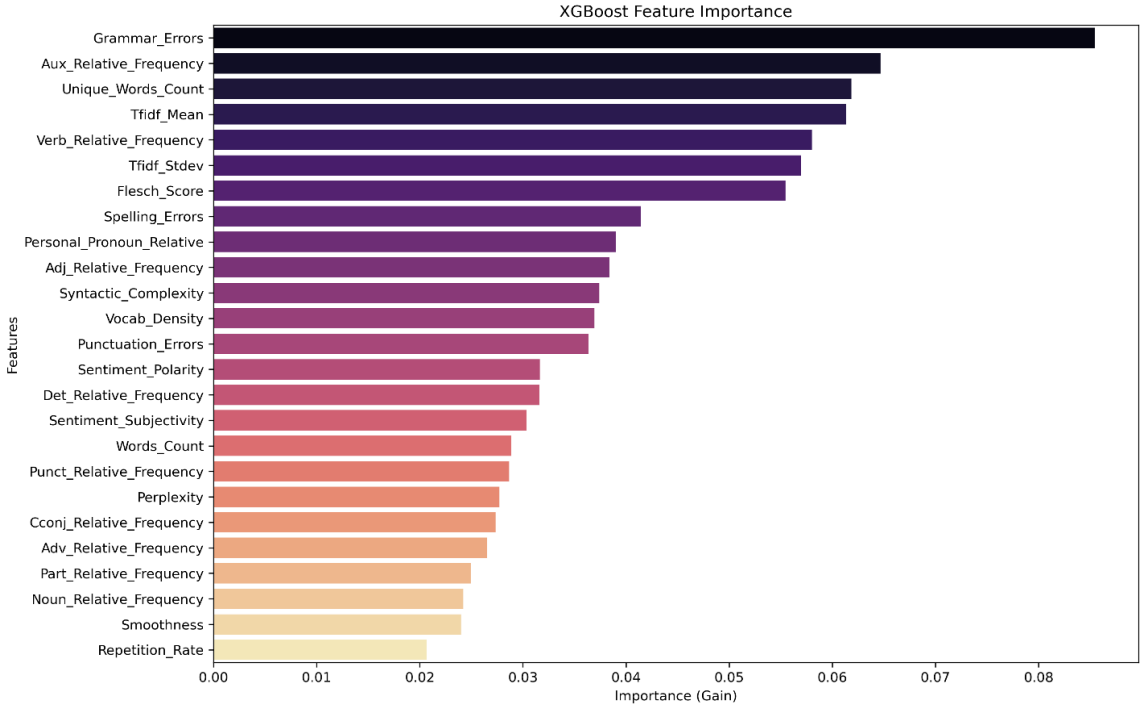
Notes: This figure illustrates the timeline of major generative AI model releases. During our sample period (2022–2024Q1), ChatGPT-3.5 (released November 2022) and GPT-4 (released March 2023) represent the dominant generative AI tools in practical applications. While other large language models, including Google Bard (later Gemini), Anthropic Claude, and Meta LLaMA, were introduced during this period, their adoption remained more limited relative to ChatGPT-based tools. The expansion and competition among generative AI providers accelerated after 2024, with the introduction of more advanced multimodal models (e.g., GPT-4o) and enterprise integration. Accordingly, our model training and empirical design focuses on the ChatGPT-3.5 and GPT-4 periods as the primary phases of early generative AI adoption.

Figure 2. Important Features Considered by XGBoost

Panel A: ChatGPT-3.5–Trained Model

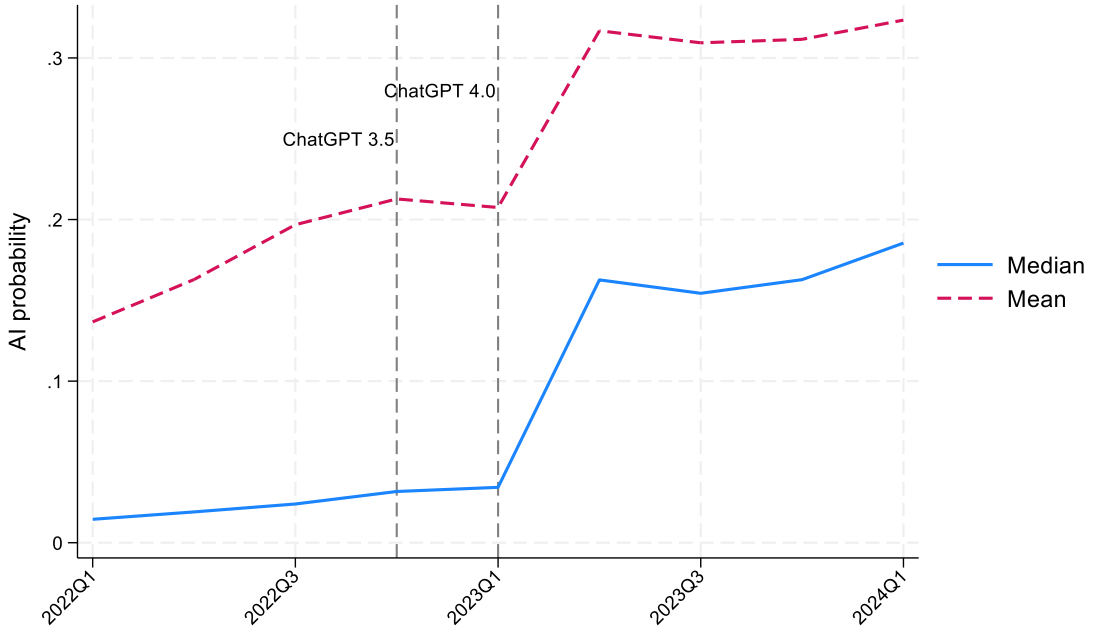


Panel B: ChatGPT-4–Trained Model



Notes: This figure reports feature importance from the XGBoost classifiers used to detect AI-assisted analyst reports. Panel A presents the model trained on GPT-3.5–generated rewrites, while Panel B presents the model trained on GPT-4–generated rewrites. Feature importance is measured using gain, defined as the average reduction in the model's loss function attributable to splits using a given feature across all trees. Higher gain indicates that the feature contributes more to improving classification performance. The two panels show differences in the linguistic signals across model generations. In the GPT-3.5–based model, feature importance is driven more by stylistic patterns, such as auxiliary verb usage. In contrast, the GPT-4–based model places greater weight on grammatical accuracy and lexical features. This shift is consistent with more advanced language models producing text that is closer to human-written content and harder to detec

Figure 3. Trends of AI Probability in Reports



Notes: This figure plots the time-series evolution of the document-level AI usage probability (AI_Prob) in analyst reports, reporting both mean and median values by quarter. The vertical dashed lines mark the introduction of ChatGPT-3.5 (November 2022) and GPT-4 (March 2023). AI usage exhibits a gradual increase following the release of ChatGPT-3.5 and a more pronounced rise after the introduction of GPT-4. The relatively muted growth in early 2023 coincides with reported restrictions on ChatGPT usage by several large financial institutions, including JPMorgan, Bank of America, Citi, Deutsche Bank, Goldman Sachs, and Wells Fargo, due to compliance and data security concerns. For example, JPMorgan restricted employee access to ChatGPT in February 2023. These restrictions were later relaxed or supplemented by internally developed AI systems (e.g., proprietary "ChatGPT-like" tools), consistent with the continued increase in AI usage observed in subsequent periods.

Table 1. Sample Description

Panel A. Analysts, Brokers, Firms, and Equity Reports

	All		Pre-ChatGPT		Post-ChatGPT	
	Total	Unique	Total	Unique	Total	Unique
Broker	55,778	73	23,752	65	32,026	65
Analyst	55,778	434	23,752	351	32,026	399
Firm	55,778	2,016	23,752	1,781	32,026	1,879
Report	55,778	38,468	23,752	15,590	32,026	22,878

Note: 55,778 unique forecasts in this study. One report may contain multiple forecasts.

Panel B. Sample Distribution by Year

Year	Freq.	Percent
2022	25,016	44.85
2023	26,882	48.19
2024	3,880	6.96
Total	55,778	100

Panel C. Sample Distribution by Year-Quarter

Year-Qtr	Freq.	Percent
2022q1	5,237	9.39
2022q2	6,378	11.43
2022q3	6,535	11.72
2022q4	6,866	12.31
2023q1	6,945	12.45
2023q2	7,468	13.39
2023q3	6,565	11.77
2023q4	5,904	10.58
2024q1	3,880	6.96
Total	55,778	100

Panel D. Sample Comparison with IBES

	IBES (2021-2024)						Our Sample					
	N	Mean	P50	SD	P25	P75	N	Mean	P50	SD	P25	P75
BrokerSize	378	1.569	1.386	0.948	0.693	2.303	201	1.992	1.946	0.998	1.099	2.708
FirmsCovered	2,660	2.321	2.565	0.866	1.609	2.996	1,138	2.668	2.833	0.625	2.303	3.135
IndCovered	2,660	1.219	1.099	0.528	0.693	1.609	1,138	1.337	1.386	0.546	0.693	1.792
ForecastFrequency	2,660	4.133	4.394	1.562	2.944	5.476	1,138	4.734	5.075	1.150	3.761	5.680

Note: BrokerSize is measured at the broker-year level. FirmsCovered, IndCovered, and ForecastFrequency are measured at the analyst-year level.

Table 2. Summary Statistics and Correlations

Panel A. Descriptive Statistics

Notes: This table reports summary statistics for the main variables used in the analysis. Panel A presents the mean, median (P50), and standard deviation (SD) for the full sample, as well as separately for the pre-ChatGPT and post-ChatGPT subsamples. AI_Prob is a document-level measure capturing the probability that an analyst report incorporates generative AI. Accuracy is defined as the negative of the proportional mean absolute forecast error (PMAFE), so that higher values indicate greater forecast accuracy. Variables reported include FirmsCovered, IndCovered, AnalystFirmExperience, ForecastFrequency, Horizon, and TOP10. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A.

	All (N=55,778)			Pre-ChatGPT (N=23,752)			Post-ChatGPT (N=32,026)		
	Mean	P50	SD	Mean	P50	SD	Mean	P50	SD
AI_Prob	0.241	0.062	0.316	0.178	0.021	0.288	0.287	0.121	0.328
Accuracy	0.174	0.238	0.572	0.151	0.208	0.582	0.190	0.254	0.564
FirmsCovered	2.990	3.045	0.455	3.008	3.091	0.462	2.976	2.996	0.449
IndCovered	1.443	1.386	0.550	1.465	1.386	0.543	1.427	1.386	0.554
AnalystFirmExperience	1.505	1.479	0.862	1.454	1.433	0.886	1.543	1.510	0.841
ForecastFrequency	5.486	5.638	0.769	5.699	5.813	0.612	5.329	5.568	0.832
Horizon	4.657	4.543	0.753	4.705	4.575	0.778	4.621	4.522	0.732
TOP10	0.684	1.000	0.465	0.679	1.000	0.467	0.687	1.000	0.464

Panel B. Pearson Correlation Matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) AI_Prob	1.00							
(2) Accuracy	0.01***	1.00						
(3) FirmsCovered	0.01**	-0.02***	1.00					
(4) IndCovered	0.00	0.00	0.13***	1.00				
(5) AnalystFirmExperience	0.01	0.04***	0.10***	0.12***	1.00			
(6) ForecastFrequency	-0.03***	-0.02***	0.70***	0.16***	0.10***	1.00		
(7) Horizon	0.01**	-0.17***	0.04***	0.01*	-0.04***	0.11***	1.00	
(8) TOP10	0.10***	0.01***	0.04***	-0.06***	0.06***	0.12***	-0.05***	1.00

Note: ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 3. AI Usage and Forecast Accuracy

Notes: This table reports OLS regression results examining the association between AI usage in analyst reports and forecast accuracy. The dependent variable is Accuracy, defined as the negative of the proportional mean absolute forecast error (PMAFE), so that higher values indicate greater forecast accuracy. The key independent variable is AI_Prob, a document-level measure capturing the probability that an analyst report incorporates generative AI. Control variables include FirmsCovered, IndCovered, AnalystFirmExperience, ForecastFrequency, Horizon, and TOP10. Column (1) includes firm–year and analyst–year fixed effects, while Column (2) additionally includes control variables, as indicated. t-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy
AI_Prob	0.039***	0.026***
	(4.15)	(2.88)
FirmsCovered		0.116**
		(2.01)
IndCovered		-0.072
		(-0.93)
AnalystFirmExperience		0.043***
		(4.54)
ForecastFrequency		-0.061***
		(-5.12)
Horizon		-0.121***
		(-19.66)
TOP10		0.493***
		(3.39)
Firm-Year	Yes	Yes
Analyst-Year	Yes	Yes
Observations	55,778	55,778
R-squared	0.312	0.335
Adj R-squared	0.235	0.260

Table 4. AI Usage and Forecast Accuracy: Heterogeneity by Analyst Workload and Busyness

Notes: This table reports OLS regression results examining cross-sectional variation in the association between AI usage in analyst reports and forecast accuracy using subsample analysis. The dependent variable is Accuracy. The key independent variable is AI_Prob, a document-level measure capturing the probability that an analyst report incorporates generative AI. The sample is partitioned based on analyst workload (busyness) characteristics, including FirmsCovered, IndCovered, and ForecastFrequency. For each characteristic, we split the sample into above-median and below-median groups and estimate the regressions separately. This design allows us to assess whether AI usage is more beneficial in settings with higher analyst workload or busyness. All specifications include the same set of control variables as in Table 3, as well as firm–year and analyst–year fixed effects. t-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
	FirmsCovered		IndCovered		ForecastFrequency	
	Above_Median Accuracy	Below_Median Accuracy	Above_Median Accuracy	Below_Median Accuracy	Above_Median Accuracy	Below_Median Accuracy
AI_Prob	0.043***	0.008	0.026**	0.015	0.031***	0.018
	(3.42)	(0.58)	(2.09)	(1.09)	(2.60)	(1.40)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Year	Yes	Yes	Yes	Yes	Yes	Yes
Analyst-Year	Yes	Yes	Yes	Yes	Yes	Yes
Observations	28,639	26,542	34,346	21,199	27,794	27,362
R-squared	0.376	0.402	0.346	0.364	0.383	0.420
Adj R-squared	0.293	0.310	0.265	0.283	0.303	0.320

Table 5. AI Usage and Forecast Accuracy: Heterogeneity by Analyst Characteristics

Notes: This table reports OLS regression results examining cross-sectional variation in the association between AI usage in analyst reports and forecast accuracy using subsample analysis. The dependent variable is Accuracy. The key independent variable is AI_Prob, a document-level measure capturing the probability that an analyst report incorporates generative AI. The sample is partitioned based on analyst characteristics, including CFA status, All-Star recognition, and AnalystIndExperience. For CFA and All-Star recognition, we split the sample based on indicator values (Yes/No). For AnalystIndExperience, we split the sample into above-median and below-median groups and estimate the regressions separately for each subsample. This design allows us to assess whether the effect of AI usage varies across analysts with different characteristics. All specifications include the same set of control variables as in Table 3, as well as firm–year and analyst–year fixed effects. T-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
	CFA		Star		AnalystIndExperience	
	Yes	No	Yes	No	Above_Median	Below_Median
	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
AI_Prob	0.024	0.024**	0.012	0.029**	0.019*	0.030**
	(1.62)	(2.15)	(0.94)	(2.38)	(1.65)	(2.21)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Year	Yes	Yes	Yes	Yes	Yes	Yes
Analyst-Year	Yes	Yes	Yes	Yes	Yes	Yes
Observations	21,173	34,202	19,241	36,249	27,675	27,689
R-squared	0.403	0.357	0.357	0.375	0.373	0.382
Adj R-squared	0.313	0.276	0.279	0.288	0.295	0.292

Table 6: Exogenous Shocks from Overlapping Earnings Calls: IV Regression

Notes: This table reports instrumental variable (IV) regression results examining the association between AI usage in analyst reports and forecast accuracy using exogenous shocks from overlapping earnings calls. The instrument is an Overlapping Dummy, which equals one if an analyst releases a report on a day with at least two earnings calls for firms the analyst covers, and zero otherwise. Panel A presents IV regressions using overlapping earnings calls on the same date as the instrument. The first-stage dependent variable is AI_Prob. The second-stage dependent variable is Accuracy. All specifications include analyst-level control variables as well as firm–year and analyst–year fixed effects, as indicated. T-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. The table reports first-stage F-tests as well as underidentification and weak IV tests where applicable. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) First-stage AI_Prob	(1) Second-stage Accuracy
Overlapping Dummy	0.020***	
	(3.59)	
AI_Prob		1.924***
		(2.61)
Control	Yes	Yes
Firm-Year	Yes	Yes
Analyst-Year	Yes	Yes
Observations	52,255	52,255
F-test (P-value)	0.0003	
Underidentification test (P-value)	0.0002	
Weak IV test (KP rk Wald F)	12.86	

Table 7. AI Usage and Capital Market Reactions

Notes: This table reports OLS regression results examining the association between AI usage in analyst reports and capital market reactions. The dependent variable is |CAR|, defined as the absolute value of the cumulative three-day abnormal return (−1, +1) around the forecast date, calculated using the market-adjusted model. The key independent variable is AI_Prob, a document-level measure capturing the probability that an analyst report incorporates generative AI. Column (1) includes firm–year and analyst–year fixed effects, while Column (2) additionally includes control variables, as indicated. T-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)
	|CAR|	*|CAR|*
AI_Prob	0.002*	0.002**
	(1.78)	(1.99)
Control	No	Yes
Firm-Year FE	Yes	Yes
Analyst-Year FE	Yes	Yes
Observations	55,778	55,778
R-squared	0.456	0.458
Adj R-squared	0.395	0.397

Table 8: The Relation between AI Usage and Valuation-Related Metrics in Reports

Notes: This table examines AI usage in analyst reports and analysts' analytical research quality. The key independent variable is AI_Prob. This table presents the relation between AI usage and valuation-related metrics in reports. We use GPT-5.1 to classify research metrics in reports. Valuation equals zero if a report only employs one type of valuation model and equals one if a report uses both multiple-based and present-value–based valuation models. Topics is the number of distinct topics mentioned in a report, following Sheng et al. (2025). Stakeholders is the number of distinct types of stakeholders mentioned in a report. Private is the number of distinct types of private communication between an analyst and management mentioned in a report. Public is the number of distinct types of public communication between an analyst and management mentioned in a report. All specifications include the same set of control variables as in Table 3, as well as firm–year and analyst–year fixed effects, as indicated. T-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Valuation	(2) Topics	(3) Stakeholders	(4) Private	(5) Public
AI_Prob	0.089***	0.688***	0.301**	-0.033**	0.020
	(3.12)	(3.95)	(2.42)	(-2.22)	(0.88)
Control	Yes	Yes	Yes	Yes	Yes
Firm-Year FE	Yes	Yes	Yes	Yes	Yes
Analyst-Year FE	Yes	Yes	Yes	Yes	Yes
Observations	55,170	55,170	55,170	55,170	55,170
R-squared	0.777	0.523	0.481	0.310	0.385
Adj R-squared	0.752	0.468	0.422	0.232	0.315

Table 9: AI Usage and Analysts' Analytical Language in Reports

This table reports OLS regression results examining the association between AI usage in analyst reports and analysts' analytical language. The key independent variable is AI_Prob. We use LIWC-22 software to construct language-based measures. Analytical is the fraction of words that reflect logical and formal thinking in a report. Authentic is the fraction of words that reflects perceived honesty and genuineness in a report. Causation is the fraction of causal words (e.g., how, because, make, and why). Differentiation is the fraction of words that reflects cognitive complexity, capturing thinking in conditions, contrasts, and alternatives (e.g., but, not, if, and or). All specifications include the same set of control variables as in Table 3, as well as firm–year and analyst–year fixed effects, as indicated. T-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Analytical	(2) Authentic	(3) Causation	(4) Differentiation
AI_Prob	5.700**	1.471*	0.078**	0.168**
	(2.21)	(1.85)	(2.57)	(2.40)
Control	Yes	Yes	Yes	Yes
Firm-Year FE	Yes	Yes	Yes	Yes
Analyst-Year FE	Yes	Yes	Yes	Yes
Observations	55,168	55,164	55,170	55,170
R-squared	0.128	0.150	0.167	0.159
Adj R-squared	0.0290	0.0539	0.0729	0.0638

Appendix. Variable Definition

Variable	Definition
AI_Prob	The probability that an analyst report is AI-assisted. It is a document-level probability produced by supervised machine learning classifiers trained to distinguish human-written analyst reports from reports that incorporate generative AI text. The classifier applied is aligned with the generative AI technology prevailing at the time of the report.
Accuracy	Accuracy is defined as the negative of the proportional mean absolute forecast error (PMAFE). Following Ben-Rephael et al. (2023), PMAFE is calculated as the absolute difference between the forecasted and actual earnings per share, scaled by the average absolute forecast error across all analysts covering the same firm and forecast date.
|CAR|	|CAR| is defined as the absolute value of the cumulative three-day abnormal return (-1, 1) around on the forecast date.
Male	Male equals one if the analyst is identified as male and zero otherwise.
Edu	Edu is a categorical variable capturing educational attainment, coded from 0 to 3, with higher values indicating higher levels of education.
CFA	CFA is an indicator variable that equals one if the analyst holds the Chartered Financial Analyst (CFA) designation and zero otherwise.
AllStar	AllStar is an indicator variable that equals one if the analyst has ever been recognized as an Institutional Investor All-Star analyst prior to 2022, and zero otherwise.
ForecastFrequency	The logarithm of one plus the number of forecasts made by the corresponding analyst in the corresponding year.
Horizon	The logarithm of one plus the number of calendar days between the forecast date and the corresponding earnings forecast date.
FirmsCovered	The logarithm of one plus the number of firms the corresponding analyst covers in the corresponding year.
IndCovered	The logarithm of one plus the number of two-digit SIC industries the corresponding analyst covers in the corresponding year.
AnalystExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for any firm in the IBES database.
Valuation	Valuation, equals zero if a report only employs one type of valuation models, and equals to one if a report uses both multiple-based and present-value based valuation models.
Private	Private, is the number of distinct types of private communication between an analyst and the management mentioned in a report.
Public	Public, is the number of distinct types of public communication between an analyst and the management mentioned in a report.
Topics	Topics, is the number of distinct topics mentioned in a report, following Sheng et al. (2025).
Stakeholders	Stakeholders, is the number of distinct types of stakeholders mentioned in a report.
Analytical	Analytical is the intensity of language that reflects logical and formal thinking in a report. The measure is constructed using the LIWC-22 software (Boyd et al. 2022).
Authentic	Authentic is the intensity of language that reflects perceived honesty and genuineness in a report. The measure is constructed using the LIWC-22 software (Boyd et al. 2022).
Causation	Causation is the fraction of causal words (e.g., how, because, make, and why). The measure is constructed using the LIWC-22 software (Boyd et al. 2022).

Differentiation	Differentiation is the fraction of words that reflects cognitive complexity, reflecting thinking in conditions, contrasts, and alternatives (e.g., but, not, if, and or). The measure is constructed using the LIWC-22 software (Boyd et al. 2022).
Word_ratio	Word ratio, is the relative ratio of word counts over the sum of word counts and number counts in a report.
AnalystFirmExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for the corresponding firm.
AnalystIndExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for the corresponding industry.
TOP10	TOP10 is an indicator variable that equals one if an analyst works at a top decile brokerage house in year.
SIZE	The value of firm size in a given quarter, measured as the natural logarithm of market value of equity (MVE).
MTB	The market-to-book ratio of equity in a given fiscal quarter.
ROA	Return on assets in a given quarter, calculated as income before extraordinary items divided by total assets.
Loss	Loss is an indicator variable that equals one if the firm reports negative net income in the corresponding fiscal year, and zero otherwise.
Horizon_Day	Days between the forecast date and the earnings announcement date.
AI_ProbLow	A document-level AI usage probability generated by an XGBoost classifier trained on hybrid analyst reports in which approximately 25 percent of sentences are replaced with generative AI rewrites.
AI_ProbHigh	A document-level AI usage probability generated by an XGBoost classifier trained on hybrid analyst reports in which approximately 75 percent of sentences are replaced with generative AI rewrites.
AI_ProbLow_RF	A document-level AI usage probability generated by a Random Forest classifier trained on hybrid analyst reports with approximately 25 percent AI-rewritten content.
AI_Prob_RF	A document-level AI usage probability generated by a Random Forest classifier trained on hybrid analyst reports with approximately 50 percent AI-rewritten content.
AI_ProbHigh_RF	A document-level AI usage probability generated by a Random Forest classifier trained on hybrid analyst reports with approximately 75 percent AI-rewritten content.
AI_ProbLow_LR	A document-level AI usage probability generated by a Logistic Regression classifier trained on hybrid analyst reports with approximately 25 percent AI-rewritten content.
AI_Prob_LR	A document-level AI usage probability generated by a Logistic Regression classifier trained on hybrid analyst reports with approximately 50 percent AI-rewritten content.
AI_ProbHigh_LR	A document-level AI usage probability generated by a Logistic Regression classifier trained on hybrid analyst reports with approximately 75 percent AI-rewritten content.
AI_Prob_2000-2020	A document-level AI usage probability constructed using an alternative training sample consisting of analyst reports issued between 2000 and 2020. The classifier is trained using the same supervised learning framework as the baseline specification but replaces the 2021 training reports with this longer historical sample.

Appendix B. Detecting the Use of Generative AI in Analyst Reports

This appendix describes the construction and validation of the document-level AI usage measure (AI_Prob) used in the main analyses. Because there is no observable ground truth for generative AI usage in analyst reports, our detection framework relies on sentence-level rewriting and feature construction to generate supervised training data. These sentence-level procedures are used solely for model training and validation. AI_Prob captures the likelihood that an analyst report incorporates generative AI and serves as the primary measure of AI usage throughout the paper. The detection framework consists of four steps: 1) construction of labeled training data, 2) text preprocessing and feature extraction, 3) model training and validation (separately for two generative AI models, ChatGPT-3.5 and ChatGPT-4.0), and 4) generation of document-level AI usage probabilities.

B.1 Construction of Training Data

B.1.1 From PDF to Clean Text

We extract raw text from each PDF using pdfminer.six, a Python library for PDF text extraction. To construct ground-truth training data, we randomly select approximately 2,600 analyst research reports (three reports for each analyst) issued in 2021, prior to the public release of ChatGPT, a period during which the use of generative AI in analyst research was negligible.

We preprocess analyst reports to remove formatting artifacts introduced by PDF extraction and retain only analyst-written narrative content. Specifically, we first correct sentence fragmentation introduced by PDF extraction by merging adjacent lines when the preceding line indicates sentence continuation (e.g., abbreviations such as “e.g.” or “vs.”). Second, we remove non-narrative artifacts, including table remnants, symbol-only strings, and lines dominated by numerical or special characters, which primarily reflect formatting noise rather than analyst-written text. Third, we apply a set of heuristic filters to exclude non-linguistic content, such as URLs, excessive punctuation, and numerically dominated strings. These steps ensure that only coherent text is retained for analysis.

Fourth, we identify and remove broker-dealer boilerplate. We manually compile a list of disclaimer opening sentences used by sell-side brokers. For each report, once a disclaimer opening sentence is detected, we remove that sentence and all subsequent content. This approach effectively excludes legally required boilerplate disclosures while preserving analyst-written research text.

After removing noise and boilerplate content, we retain valid textual units and re-establish the narrative sequence within each analyst report. We reorder the filtered sentences according to their original position in the report and construct a continuous index so that numbering begins at one and increases sequentially.

Using the cleaned and re-indexed text, we next construct AI-assisted versions of analyst reports for supervised training. For each selected report, we generate AI-assisted counterparts by rewriting each sentence of the original analyst-written content using generative AI models. Specifically, we use GPT-3.5-Turbo to generate AI-assisted versions for reports issued during the early period between the release of ChatGPT-3.5 (November 2022) and the release of GPT-4 (March 2023), and GPT-4 to generate AI-assisted versions for reports issued during the later period after the release of GPT-4 (March 2023) and before April 2024. These AI-assisted versions are used to train the corresponding detection models. The rewriting process preserves the original analytical meaning and professional tone of the report. Reports rewritten with approximately 50%

AI-assisted content serve as the primary training benchmark. Alternative AI-intensity benchmarks (e.g., 25% AI-assisted content and 75% AI-assisted content) are used in robustness analyses.

Each original report is labeled as human-written, and each AI-assisted version is labeled as AI-assisted. These paired human-written and AI-assisted documents form the labeled dataset used for supervised training.

B.1.2 GPT-Based Sentence Rewriting

We construct AI-generated sentence variants using a controlled rewriting framework. After completing the sentence-level cleaning and re-indexing, we draw a subset of analyst reports and process each remaining sentence through structured rewriting tasks using GPT-3.5-Turbo and GPT-4. Each original sentence is submitted to the model using a predefined prompt configuration. The prompt provides a professional context, instructing the model to act as a sell-side equity research analyst and to reproduce the analytical tone commonly found in broker reports.

'You are an experienced sell-side equity research analyst with deep expertise in financial analysis, industry research, and professional investment writing. Your task is to rewrite the following sentence from a sell-side equity research report while preserving its original meaning and analytical intent.
Follow these instructions carefully:
(1) Professional Tone: Maintain the formal, objective, and analytical style typical of equity research reports.
(2) Length: Keep the rewritten sentence approximately the same length as the original (within ±20% of its word count).
(3) Fidelity: Do not introduce new information, interpretations, opinions, or assumptions beyond what is stated in the original sentence.
(4) Clarity: Improve clarity and readability where appropriate without altering the meaning.
(5) Output Format: Return only the rewritten sentence. Do not include explanations, comments, headings, or additional text.'

Because prompt engineering may influence generative model outputs, we applied a second rewriting prompt that omits the contextual framing used in the primary specification.

'Follow these instructions carefully:
(1) Professional Tone: Maintain the formal, objective, and analytical style typical of equity research reports.
(2) Length: Keep the rewritten sentence approximately the same length as the original (within ±20% of its word count).
(3) Fidelity: Do not introduce new information, interpretations, opinions, or assumptions beyond what is stated in the original sentence.
(4) Clarity: Improve clarity and readability where appropriate without altering the meaning.
(5) Output Format: Return only the rewritten sentence. Do not include explanations, comments, headings, or additional text.'

We compared AI-detection results across the two rewritten versions using a subsample of reports and found no statistically meaningful differences in detection outcomes. Commercial detectors consistently identified the rewritten text as AI-generated under both prompt conditions. This robustness check alleviates concerns that our findings are driven by the specific wording of the rewriting prompt. To study how document-level AI-detection systems perform when analyst reports are only partially generated or edited by generative AI, we construct controlled hybrid

documents in which only a subset of originally human-authored sentences is replaced with AI-rewritten text. For each human-written report in our sample, we preserve the original narrative order and randomly select a fixed proportion of sentences to be replaced with AI-generated rewrites. Replacement is carried out independently for each report, with the random selection process made fully reproducible through seeding. The replacement proportions are set at 25%, 50%, 75%, and 100%.

Figure B1. Rewriting



Each selected sentence is substituted with one of the AI-rewritten variants previously generated under different prompt/model configurations. Because of rounding issue, it could not be exactly 25%, 50%, or 75%

B.2 Text Preprocessing and Feature Engineering

We preprocess the text of each report. The preprocessing steps include converting text to lowercase, removing punctuation and numeric characters, normalizing whitespace, and eliminating English stopwords. Preprocessing is applied in a feature-specific manner. Features that depend on surface-level characteristics, such as grammatical errors, are computed from the original text.

We then extract a comprehensive set of document-level linguistic features designed to distinguish human-written from AI-assisted reports. These features are motivated by prior work on AI text detection (Alamleh et al., 2023; Guo et al., 2023; Mitrović et al., 2023; Uzun, 2023) and fall into the following categories (see Table B1 for detailed definitions):

1) Lexical Features: total word count, vocabulary diversity, and repetition rate.
2) Syntactic Features: clause density, punctuation usage, and the relative frequency of key part-of-speech categories such as nouns, verbs, determiners, and conjunctions.
3) Error Features: the number of spelling mistakes, grammatical issues, and punctuation irregularities in each sentence.
4) Sentiment Features: the polarity (positive vs. negative tone) and subjectivity (objective vs. subjective tone) of the sentence derived using TextBlob.
5) Readability Features: the Flesch Reading Ease Score as well as language model-based fluency metrics such as GPT-2 perplexity and smoothness.

6) TF-IDF (Term Frequency-Inverse Document Frequency) Features: the importance and distribution of terms within the sentence, based on standard term frequency-inverse document frequency scores computed using scikit-learn.

These features serve as the input variables for our machine learning classifier trained to identify AI-generated language (Table B1).

Table B1.2 reports summary statistics for these features from the training sample, separately for human-written and ChatGPT-3.5-rewritten analyst reports. The purpose of this table is to provide a descriptive illustration of how generative AI rewriting alters textual characteristics prior to model training. Because many features are defined as counts, rates, or transformations derived from raw text, their magnitudes can vary substantially across feature types and should not be interpreted on a common scale. In particular, features such as perplexity and smoothness are computed following standard definitions in the language-model literature and therefore appear on scales that differ markedly from frequency-based or error-based measures. These raw feature values are reported for transparency. In model training, all features are standardized so that differences in scale do not affect classification performance. The table is intended to offer intuition on the linguistic signals exploited by the detection models, rather than to support direct economic interpretation of individual feature magnitudes.

Tabel B1.1 Feature Engineering

Feature Name	Description
Lexical Features	
Words_Count	Total number of words in the text.
Unique_Words_Count	Total number of unique words in the text.
Vocab_Density	Vocabulary density, calculated as the percentage of unique words relative to the total word count.
Repetition_Rate	Repetition rate, measured as the percentage of repeated words relative to the total word count.
Syntactic Features	
Syntactic_Complexity	Percentage of clauses per sentence, derived from punctuation and syntactic patterns using spaCy.
Punct_Relative_Frequency	Percentage of punctuation marks relative to the total word count.
Noun_Relative_Frequency	Percentage of nouns in the text relative to the total word count, based on spaCy's part-of-speech tagging.
Verb_Relative_Frequency	Percentage of verbs in the text, calculated similarly using spaCy.
Det_Relative_Frequency	Percentage of determiners (e.g., 'the,' 'a') relative to the total word count.
Adj_Relative_Frequency	Percentage of adjectives in the text.
Aux_Relative_Frequency	Percentage of auxiliary verbs (e.g., 'is,' 'has').
Adv_Relative_Frequency	Percentage of adverbs in the text.
Cconj_Relative_Frequency	Percentage of coordinating conjunctions (e.g., 'and,' 'or').
Part_Relative_Frequency	Percentage of particles (e.g., 'to' in 'to eat').
Personal_pronoun_Relative	Percentage of personal pronouns (e.g., 'he,' 'she,' 'they').
Error Features	
Spelling_Errors	Total number of misspelled words.
Grammar_Errors	Total number of grammar errors.
Punctuation_Errors	Total number of punctuation errors.
Sentiment Features	
Sentiment_Polarity	Sentiment polarity score (range -1 to +1).
Sentiment_Subjectivity	Sentiment subjectivity score (range 0 to 1), where higher values indicate more subjective content.
Readability Features	
Flesch_score	Flesch Reading Ease score = 206.835 - 1.015 * (words_count / sentence_count) - 84.6 * (syllables_count / words_count)
Perplexity	Perplexity score, representing text fluency, measured using the GPT-2 language model from the transformers library, based on pre-processed text. Perplexity is defined as the exponentiation of the average negative log-likelihood per token.
Smoothness	Smoothness of the text, defined as 1 - perplexity, where higher values indicate better fluency, based on pre-processed text.
TFIDF Features	
Tfidf_Mean	Mean of Term Frequency-Inverse Document Frequency (TF-IDF) scores, calculated using scikit-learn, based on pre-processed text.
Tfidf_Stdev	Standard deviation of TF-IDF scores, indicating variability in term importance, based on pre-processed text.

Tabel B1.2 Summary Statistics of Features in the Training Sample (ChatGPT-3.5)

	Feature	Human_Mean	AI_Mean	AI_minus_Human	t_stat	p_value	Significance
1	*Adj_Relative_Frequency*	7.070	7.492	0.422	10.108	0.000	***
2	*Adv_Relative_Frequency*	2.477	2.326	-0.151	-6.109	0.000	***
3	*Aux_Relative_Frequency*	4.609	3.927	-0.682	-20.73	0.000	***
4	*Cconj_Relative_Frequency*	4.131	3.768	-0.362	-9.745	0.000	***
5	*Det_Relative_Frequency*	7.533	7.837	0.303	6.887	0.000	***
6	*Flesch_Score*	23.701	21.321	-2.380	-6.149	0.000	***
7	*Grammar_Errors*	352.22	196.832	-155.388	-6.337	0.000	***
8	*Noun_Relative_Frequency*	23.724	24.355	0.631	6.673	0.000	***
9	*Part_Relative_Frequency*	2.172	2.154	-0.018	-0.894	0.372	
10	*Perplexity*	177269.918	119974.934	-57294.983	-0.588	0.557	
11	*Personal_Pronoun_Relative*	3.110	2.737	-0.373	-11.581	0.000	***
12	*Punct_Relative_Frequency*	15.756	15.686	-0.070	-0.742	0.458	
13	*Punctuation_Errors*	366.267	359.752	-6.515	-0.402	0.687	
14	*Repetition_Rate*	60.575	58.479	-2.096	-6.467	0.000	***
15	*Sentiment_Polarity*	0.068	0.063	-0.005	-4.271	0.000	***
16	*Sentiment_Subjectivity*	0.398	0.400	0.002	1.479	0.139	
17	*Smoothness*	-177268.918	-119973.934	57294.983	0.588	0.557	
18	*Spelling_Errors*	112.21	94.512	-17.698	-5.049	0.000	***
19	*Syntactic_Complexity*	256.107	253.404	-2.703	-1.932	0.053	*
20	*Tfidf_Mean*	0.038	0.037	-0.001	-0.911	0.362	
21	*Tfidf_Stdev*	0.034	0.032	-0.002	-9.881	0.000	***
22	*Unique_Words_Count*	724.42	753.956	29.535	1.751	0.080	*
23	*Verb_Relative_Frequency*	9.461	9.844	0.383	10.617	0.000	***
24	*Vocab_Density*	39.425	41.521	2.096	6.467	0.000	***
25	*Words_Count*	2401.478	2339.875	-61.603	-0.590	0.555	

B.3 Model Training and Validation

B.3.1 Model Training with Three classifiers

Using the labeled document-level dataset, we split the data into training (80%) and testing (20%) samples. We evaluate three widely used supervised learning classifiers: XGBoost, Random Forest, and Logistic Regression. Model performance is assessed using standard classification metrics, including the area under the ROC curve (AUC), accuracy, true positive rate, true negative rate, and F1 score. Across specifications, the XGBoost classifier demonstrates the strongest overall performance in distinguishing human-written from AI-assisted analyst reports, and is therefore adopted as the primary detection model. Detailed performance metrics are reported in Table B2.1 - B2.6.

Table B2.1 Self-Trained document-level XGBoost model (ChatGPT-3.5)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Test set					
Human vs GPT3.5_25%	0.716	0.662	0.629	0.694	0.661
Human vs GPT3.5_50%	**0.947**	**0.877**	**0.860**	**0.894**	**0.877**
Human vs GPT3.5_75%	0.988	0.941	0.940	0.942	0.941
Human vs GPT3.5_100%	0.999	0.983	0.981	0.985	0.983
Test set aligned with GPTZero and Originality.ai evaluation					
Human vs GPT3.5_25%	0.791	0.700	0.720	0.680	0.700
Human vs GPT3.5_50%	0.946	0.830	0.840	0.820	0.830
Human vs GPT3.5_75%	0.974	0.900	0.940	0.860	0.900
Human vs GPT3.5_100%	0.998	0.970	0.980	0.960	0.970

Table B2.2 Self-Trained document-level Random Forest model (ChatGPT-3.5)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Test set					
Human vs GPT3.5_25%	0.662	0.590	0.559	0.621	0.590
Human vs GPT3.5_50%	0.922	0.841	0.836	0.845	0.841
Human vs GPT3.5_75%	0.979	0.916	0.911	0.922	0.916
Human vs GPT3.5_100%	0.998	0.979	0.972	0.985	0.979
Test set aligned with GPTZero and Originality.ai evaluation					
Human vs GPT3.5_25%	0.835	0.750	0.840	0.660	0.748
Human vs GPT3.5_50%	0.934	0.850	0.900	0.800	0.850
Human vs GPT3.5_75%	0.967	0.900	0.940	0.860	0.900
Human vs GPT3.5_100%	0.998	0.970	0.980	0.960	0.970

Table B2.3 Self-Trained document-level Logistical Regression model (ChatGPT-3.5)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Test set					
Human vs GPT3.5_25%	0.745	0.693	0.682	0.705	0.693
Human vs GPT3.5_50%	0.912	0.844	0.834	0.855	0.844
Human vs GPT3.5_75%	0.974	0.935	0.931	0.938	0.935
Human vs GPT3.5_100%	0.994	0.977	0.974	0.980	0.977
Test set aligned with GPTZero and Originality.ai evaluation					
Human vs GPT3.5_25%	0.718	0.690	0.740	0.640	0.689
Human vs GPT3.5_50%	0.887	0.820	0.860	0.780	0.820
Human vs GPT3.5_75%	0.951	0.870	0.920	0.820	0.870
Human vs GPT3.5_100%	0.990	0.950	0.900	1.000	0.950

Table B2.4 Self-Trained document-level XGBoost model (ChatGPT-4.0)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Test set					
Human vs GPT4.0_25%	0.522	0.502	0.488	0.517	0.502
Human vs GPT4.0_50%	**0.862**	**0.778**	**0.749**	**0.808**	**0.778**
Human vs GPT4.0_75%	0.963	0.896	0.875	0.916	0.896
Human vs GPT4.0_100%	0.995	0.962	0.948	0.976	0.962
Test set aligned with GPTZero and Originality.ai evaluation					
Human vs GPT4.0_25%	0.739	0.670	0.780	0.560	0.666
Human vs GPT4.0_50%	0.900	0.800	0.820	0.780	0.780
Human vs GPT4.0_75%	0.962	0.890	0.880	0.900	0.890
Human vs GPT4.0_100%	0.995	0.960	0.980	0.940	0.960

Table B2.5 Self-Trained document-level Random Forest model (ChatGPT-4.0)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Test set					
Human vs GPT4.0_25%	0.486	0.483	0.466	0.500	0.483
Human vs GPT4.0_50%	0.821	0.727	0.706	0.748	0.727
Human vs GPT4.0_75%	0.946	0.860	0.842	0.879	0.860
Human vs GPT4.0_100%	0.992	0.952	0.929	0.974	0.952
Test set aligned with GPTZero and Originality.ai evaluation					
Human vs GPT4.0_25%	0.741	0.660	0.820	0.500	0.651
Human vs GPT4.0_50%	0.870	0.780	0.820	0.740	0.780
Human vs GPT4.0_75%	0.956	0.880	0.860	0.900	0.880
Human vs GPT4.0_100%	0.997	0.950	0.960	0.940	0.950

Table B2.6 Self-Trained document-level Logistical Regression model (ChatGPT-4.0)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Test set					
Human vs GPT4.0_25%	0.651	0.624	0.613	0.636	0.624
Human vs GPT4.0_50%	0.804	0.758	0.749	0.767	0.758
Human vs GPT4.0_75%	0.919	0.868	0.851	0.884	0.868
Human vs GPT4.0_100%	0.983	0.940	0.920	0.961	0.940
Test set aligned with GPTZero and Originality.ai evaluation					
Human vs GPT4.0_25%	0.641	0.590	0.700	0.480	0.585
Human vs GPT4.0_50%	0.786	0.730	0.780	0.680	0.729
Human vs GPT4.0_75%	0.914	0.860	0.840	0.880	0.860
Human vs GPT4.0_100%	0.960	0.930	0.880	0.980	0.930

B.3.2 Validation with GPTZero and Originality.AI

We begin by randomly selecting an additional 50 analyst reports issued in 2021 to benchmark the performance of commercial AI-detection systems (GPTZero and Originality.ai). After constructing both fully human-authored and partially AI-substituted analyst reports, we submit each resulting document to commercial detectors to obtain document-level assessments of AI involvement.

We use GPTZero's 2025-10-17-base model, the most recent release available as of late 2025. GPTZero models follow a date-based versioning scheme, where the version identifier reflects the model's release date rather than the date of inference. We explicitly pin the model version to ensure reproducibility.

We also evaluate the performance of Originality.ai, another commercially deployed AI-detection system widely used in educational and editorial settings. To ensure comparability with prior work and to maintain consistency across all documents in our study, each report, whether fully human-written, fully AI-rewritten, or partially substituted with AI text, is submitted to Originality.ai using the same report-level inputs employed in the GPTZero pipeline. For each document, we transmit the full report text to Originality.ai's model (lite-1.0.0), which corresponds to Bradshaw et al. (2025). The system returns an overall document-level AI probability along with optional block-level predictions.

GPTZero produces a three-way probability distribution over the categories human, AI-generated, and mixed. Following the approach in Blankespoor et al. (2025), and to maintain comparability with prior research, we collapse the three-way prediction into a binary measure by treating the model's estimated probability of human authorship as the complement of AI involvement. Specifically, we construct prob_ai = 1 – prob_human, which incorporates both the AI-generated and mixed components of GPTZero's output. This approach is consistent with the interpretation that any degree of AI contribution, even partial or blended, is relevant for document-level detection.

We classify each document as AI-generated when prob_ai exceeds 50%, and human-authored otherwise. We then evaluate GPTZero's predictions against the true origin of each document, including fully human-written reports, fully AI-rewritten versions, and hybrid

documents formed by selectively replacing 25%, 50%, 75%, or 100% of the original sentences with AI-generated rewrites.

For each comparison group, we compute a comprehensive set of standard performance metrics, including accuracy, the true-positive rate for detecting AI-generated documents, the true-negative rate for identifying human-written documents, precision, the F1-score, and the area under the ROC curve (AUC). These metrics provide a structured assessment of GPTZero's sensitivity and specificity across the full spectrum of AI involvement, from minimally edited hybrid reports to fully AI-generated documents.

We conduct a parallel set of document-level detection tests using Originality.ai, applying the same set of analyst reports evaluated under GPTZero. Originality.ai provides binary probability estimates of AI authorship at the document level, specifically, the model returns the probability that a document is AI-generated and the complementary probability that it is human-written. Because Originality.ai outputs a two-class probability distribution rather than GPTZero's three-class structure, no transformation is required; we directly use the model's AI probability as prob_AI and its human probability as prob_human in our evaluation.

Table B3.1 GPTZero and Originality AI (ChatGPT-3.5)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
GPTZero					
Human vs GPT3.5_25%	0.838	0.500	1.000	0.000	0.000
Human vs GPT3.5_50%	0.937	0.500	1.000	0.000	0.000
Human vs GPT3.5_75%	0.975	0.570	1.000	0.140	0.246
Human vs GPT3.5_100%	0.990	0.670	1.000	0.340	0.508
Originality. AI					
Human vs GPT3.5_25%	0.749	0.540	0.980	0.100	0.179
Human vs GPT3.5_50%	0.916	0.650	0.980	0.320	0.478
Human vs GPT3.5_75%	0.977	0.810	0.980	0.640	0.771
Human vs GPT3.5_100%	0.989	0.990	0.980	1.000	0.990

Table B3.2 GPTZero and Originality AI (ChatGPT-4.0)

Comparison	AUC	Accuracy	TNR	TPR	F1-score
GPTZero					
Human vs GPT4.0_25%	0.793	0.500	1.000	0.00	0.000
Human vs GPT4.0_50%	0.956	0.500	1.000	0.00	0.000
Human vs GPT4.0_75%	0.983	0.510	1.000	0.02	0.039
Human vs GPT4.0_100%	0.997	0.650	1.000	0.30	0.462
Originality. AI					
Human vs GPT4.0_25%	0.677	0.505	0.980	0.020	0.039
Human vs GPT4.0_50%	0.780	0.500	0.980	0.020	0.038
Human vs GPT4.0_75%	0.899	0.550	0.980	0.120	0.211
Human vs GPT4.0_100%	0.982	0.760	0.980	0.540	0.692

B.4 Construction of Document-Level AI Usage Measure

We apply the trained XGBoost classifiers to all analyst reports in our sample, using the model corresponding to each report's period. Specifically, we use the GPT-3.5–trained model for reports issued between November 2022 and March 2023, and the GPT-4–trained model for reports issued between March 2023 and April 2024. We use the 50% AI-assisted benchmark as our primary specification because it delivers the strong detection performance and reflects a realistic degree of partial AI assistance in analyst report writing. For each report, the model produces a continuous document-level probability, denoted AI_Prob, which captures the likelihood that the report incorporates generative AI.

AI_Prob ranges from zero to one, with higher values indicating a greater likelihood of AI-assisted report writing. This document-level probability serves as the primary measure of AI usage in all main analyses. Alternative AI usage measures based on different detection models or higher AI-intensity training benchmarks are used in robustness tests.

Table C1. Robustness: Alternative AI Detection Models

Notes: This table reports OLS regression results examining the robustness of the association between AI usage and forecast accuracy to alternative AI detection models and specifications. The dependent variable is Accuracy. AI_ProbLow and AI_ProbHigh denote alternative AI usage probabilities corresponding to lower and higher degrees of AI involvement in the construction of the training benchmark (e.g., reports with approximately 25% and 75% AI-assisted content, respectively), relative to the baseline 50% specification. AI_Prob_LR and AI_Prob_RF denote AI usage probabilities generated using Logistic Regression and Random Forest classifiers, respectively, with corresponding low- and high-intensity variants (AI_ProbLow_LR, AI_ProbHigh_LR, AI_ProbLow_RF, and AI_ProbHigh_RF). AI_Prob_2000–2020 denotes an alternative AI usage probability constructed using a training sample consisting of analyst reports issued between 2000 and 2020. All specifications include the same set of control variables as in Table 3, as well as firm–year and analyst–year fixed effects, as indicated. T-statistics are reported in parentheses and are based on standard errors clustered at the analyst–year–quarter level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy	(3) Accuracy	(4) Accuracy	(5) Accuracy
AI_ProbLow	0.021**				
	(2.15)				
AI_ProbHigh		0.019*			
		(1.86)			
AI_ProbLow_LR			0.068***		
			(2.95)		
AI_Prob_LR				0.060***	
				(3.88)	
AI_ProbHigh_LR					0.049***
					(3.03)
Control	Yes	Yes	Yes	Yes	Yes
Firm-Year	Yes	Yes	Yes	Yes	Yes
Analyst-Year	Yes	Yes	Yes	Yes	Yes
Observations	55,778	55,778	55,778	55,778	55,778
R-squared	0.335	0.335	0.335	0.336	0.335
Adj R-squared	0.260	0.260	0.260	0.261	0.258

Table C1. Robustness: Alternative AI Detection Models (Continued)

	(6) Accuracy	(7) Accuracy	(8) Accuracy	(9) Accuracy
AI_ProbLow_RF	0.046**			
	(2.25)			
AI_Prob_RF		0.054***		
		(3.07)		
AI_ProbHigh_RF			0.053***	
			(3.09)	
AI_Prob_2001-2020				0.021**
				(2.09)
Control	Yes	Yes	Yes	Yes
Firm-Year	Yes	Yes	Yes	Yes
Analyst-Year	Yes	Yes	Yes	Yes
Observations	55,778	55,778	55,778	55,778
R-squared	0.335	0.335	0.335	0.335
Adj R-squared	0.260	0.260	0.260	0.260

Appendix D. Prompts

PROMPT = """
You are analyzing a single equity analyst report. Ignore the disclaimer or disclosure section at the end of the report; these are boilerplate legal statements from the broker and should NOT be counted toward any classification tasks below. Focus only on the main analytical content of the report.

Read the report text and answer the following eight tasks. You must respond ONLY with a single valid JSON object with the following keys and values, and no extra commentary:

1) "valuation_model": Classify the valuation models mentioned in the report.
- Output 0 if the report does not mention any valuation model.
- Output 1 if it uses only relative multiples such as P/E ratio, EV/EBITDA, EV/sales.
- Output 2 if it uses only present value based models such as discounted cash flow (DCF), dividend discount models, or WACC-based present value models.
- Output 3 if it uses BOTH relative multiples and present value based models.

2) "alt_data_categories_count": Count how many categories of alternative or big data are used as analytical tools in the report. The categories include:
(1) app usage (e.g., number of downloads and active users),
(2) earnings calls and employee-related data (e.g., Glassdoor, job postings),
(3) geospatial data (e.g., store locations, demographic information),
(4) point-of-sale data (e.g., merchant-level transactions, pricing),
(5) satellite imagery (e.g., consumer traffic, inventory levels),
(6) consumer sentiment (e.g., social media data),
(7) web traffic data (e.g., user search and website visits),
(8) other types of alternative or big data not listed above.
Output an integer 0–9, where 0 means no alternative/big data mentioned and up to 9 if the report clearly refers to multiple distinct categories.

3) "macro_micro_topics_count": Determine how many of the following 14 topics are mentioned:
(1) optimism about the US economy,
(2) optimism about the global economy,
(3) optimism about the firm's financial prospects,
(4) optimism about its industry's financial prospects,
(5) the firm's earnings,
(6) the firm's revenue,
(7) the firm's wages and salaries expenses,
(8) demand for its products or services,
(9) production quantity of its products,
(10) prices for its products or services,
(11) prices for its inputs or commodities,
(12) its cost of capital or hurdle rate,
(13) its capital expenditure,

(14) its employment.
A topic counts as mentioned if it is discussed explicitly or implicitly. Output an integer 0–14 for how many of these topics appear.

4) "private_communication_count": Determine how many distinct mentions or categories of private or non-public communication or information-gathering activities by the analyst appear in the report. These include:
- phone or video calls or emails with management (CEO, CFO, IR, etc.), including post-earnings call-backs,
- surveys,
- channel checks,
- site visits or bus tours, or field trips,
- non-deal roadshows,
- fireside chats or roundtables,
- concierge or social events (e.g., golf outings) used to gather information.
Each type of activity or clearly distinct instance should count once, even if repeated in wording. If no such activity is mentioned, the count is 0.
Output a single integer 0 or greater (0, 1, 2, 3, ...).

5) "public_communication_count": Determine how many distinct mentions or categories of public communication events appear in the report. These include:
- earnings conference calls,
- company conference presentations,
- special calls,
- analyst or investor day events,
- M&A calls,
- shareholder or analyst calls,
- annual shareholder meetings,
- guidance or update calls,
- sales or trading statement calls.
Each type of activity or clearly distinct instance should count once, even if repeated in wording. If no such activity is mentioned, the count is 0.
Output a single integer 0 or greater (0, 1, 2, 3, ...).

6) "stakeholder_categories_count": Determine how many of the following nine stakeholder groups are mentioned:
(1) shareholders,
(2) debt holders (bank loans, bonds),
(3) managers or executives (CEO, CFO, top management),
(4) employees or labor unions,
(5) customers,
(6) suppliers,
(7) competitors,
(8) regulators or government policymakers,
(9) local communities.

A stakeholder counts as mentioned if the group, its actions, or its impact on the firm is explicitly or implicitly discussed.
Output a single integer 0–9 for how many of these stakeholder categories appear.

7) "tone": Determine the overall tone of the report toward the firm's prospects, performance, or strategic outlook.
- Output 3 if the tone is positive or optimistic.
- Output 2 if the tone is neutral, mixed, or balanced.
- Output 1 if the tone is negative or pessimistic.

8) "word_to_number": Analyze the report text and:
- Count the total number of word tokens (alphabetic tokens, ignoring punctuation).
- Count the total number of numeric tokens (tokens that are primarily digits, such as 10, 3.5, 2025, 1,000).
- Compute the ratio:
  word_ratio = word_count / (word_count + number_count).
If (word_count + number_count) is 0, set word_ratio = 0.0.
Represent this as a JSON object with fields "word_count", "number_count", and "word_ratio".

Your response MUST be valid JSON of the form:
```
{
 "valuation_model": ...,
 "alt_data_categories_count": ...,
 "macro_micro_topics_count": ...,
 "private_communication_count": ...,
 "public_communication_count": ...,
 "stakeholder_categories_count": ...,
 "tone": ...,
 "word_to_number": {
  "word_count": ...,
  "number_count": ...,
  "word_ratio": ...
 }
}
```

Do not include any explanations, comments, or text outside this JSON object.

REPORT TEXT:
"""