Generative AI and Investment Research: Evidence from Analyst Reports

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

Acknowledgement: The authors have no sources of research support, financial relationships, or other potential conflicts of interest to disclose.

Abstract

This study provides large-sample evidence on generative AI (GenAI) adoption by equity analysts and its impact on investment research quality. Using a supervised machine learning model to detect AI-assisted content, we find that such reports are associated with greater earnings forecast accuracy. This accuracy benefit is more pronounced for analysts facing greater research demands and those with lower baseline skills. To strengthen causality, we leverage an exogenous shock to analyst workload from overlapping earnings calls, and find that AI usage is a significant driver of improved forecast accuracy. We also find that forecasts based on AI-assisted reports trigger stronger capital market reactions, indicating that investors perceive such investment research as more informative. Content analysis of analyst reports reveals that GenAI empowers analysts beyond earnings forecasts: AI-assisted reports employ more diverse valuation models, discuss a broader range of financial topics and stakeholders, and exhibit more structured reasoning patterns. Overall, our findings suggest that generative AI enhances the fundamental capabilities of information intermediaries in capital markets.

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. For helpful comments, we thank Clifton Green (discussant), Leonardo Barcellos, Paul Brockman, Donald Bowen III, Jie Cai, Kimball Chapman, Rebel Cole, Mary Cowx, Naveen Daniel, Chuck Fang, Lucile Faurel, Antoine Giannetti, Steve Hillegeist, Chang Woo (Mark) Hong, Steve Kaplan, Abdul Khizer, Phil Lamoreaux, Yinghua Li, Matthew Ma, Eldar Maksymov, Niccolo Marcarini, Michal Matejka, Roni Michaely, Nandu Nayar, Edward Nelling, Greg Nini, Heidi Packard, Ke Shen, Ke Yang, and Yuxiang Zheng. We thank seminar participants at Arizona State University, Drexel University, Florida Atlantic University, Lehigh University, Seton Hall University, Rutgers University-Camden, and the AI in Finance Conference at University of Maryland. We thank Alex Cui (GPTZero) for sharing technical documentation. Zhang acknowledges data support from the Whitcomb Center for Research in Financial Services at Rutgers University. All errors are our own.

1. Introduction

With the rapid integration of generative AI (GenAI) into the workplace, a growing capital markets literature uses 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 firm value (Eisfeldt et al., 2025). However, much less is known about how investment professionals incorporate GenAI into their daily workflows, likely due to the empirical challenges of measuring such a phenomenon. In this paper, we develop a supervised machine learning model to detect GenAI usage in analyst reports using linguistically grounded features and validated training data.[1] We then utilize this measure to examine how AI adoption affects equity analysts' investment research quality.

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 through activities such as private communications with managers and site visits (Huang et al., 2018; Bradley 2023). At the same time, in recent years, analysts cover a growing number of companies, and face increased time pressure and competition, all of

[1] Our measure cannot and is not intended to identify all forms of GenAI use during analysts' research process; this is especially true for the uses that leave no linguistic footprint in the report (e.g., data synthesis, valuation calculations, or AI-assisted brainstorming). Our objective is more modest: to identify analyst reports whose final written content exhibits linguistic characteristics consistent with AI assistance. Because data on real-world GenAI usage by analysts are not observable, constructing a synthetic ground truth represents a practical approximation. Accordingly, we interpret AI_Prob as a proxy for AI assisted writing in the final report, which, to the extent that AI-assisted research leaves observable traces in written output, serves as a proxy for broader GenAI integration into the analyst's research production process.

which limit the attention and resources they can devote to each report (Rao, Tsekova, and MacDonogh 2025). GenAI offers a potential solution to reducing 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 insights that simply recycle widely known information. As a result, the effect of AI usage on research quality is ambiguous. This raises our central question: Does AI usage improve the quality of investment research?

To investigate how AI affects the production of analyst reports, we develop a novel measure of AI usage by applying a supervised machine-learning framework to a large corpus of analyst reports, identifying reports with a higher likelihood of incorporating GenAI. Over our sample period, two models were feasibly available to analysts: GPT-3.5 and GPT-4. Accordingly, we separately train AI classifiers for each model and match them to reports issued when the corresponding model was predominant. To construct training data for the detection model, we randomly select 2,600 analyst reports from reports issued in 2021 (prior to the public release of ChatGPT-3.5), and prompt ChatGPT to revise them using a controlled rewriting framework, thereby generating AI-assisted versions of originally human-authored text.[2] Using this paired dataset (original and AI-assisted reports), we train machine learning models based on 25 linguistic features documented in

[2] The tools available in 2021 were significantly more primitive and arguably inadequate for the demands of investment research. In particular, the most advanced model was GPT-3 which was not accessible in chatbot form, exhibited high hallucination rates, and had limits on content feeds (Manakul et al., 2023).

the literature that distinguish AI-assisted content from human-authored content (e.g., Guo et al., 2023; Uzun, 2023).[3]

Our testing sample consists of approximately 38,000 analyst research reports, covering the period from January 1, 2022 to April 1, 2024. Applying the trained classifier, we generate a continuous, document-level probability measure (*AI_Prob*) for each analyst report, capturing the likelihood that the report incorporates GenAI. We find a clear upward trend in AI usage in analyst reports over the sample period and a notable uptick of AI usage after the release of GPT-4.

Evaluating the link between AI usage and investment research quality, we find that higher AI usage in analyst reports is associated with significantly greater forecast accuracy, even after controlling for analyst characteristics and including firm-by-year, and analyst-by-year fixed effects. On average, a one standard deviation increase in AI usage (31.6%) is linked to a meaningful improvement in relative forecast accuracy, corresponding to roughly a 4.7% gain compared to the sample mean.

In cross-sectional analyses, we show that the effectiveness of AI varies systematically with analyst workload. 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

[3] In particular, we evaluate three supervised classification models: XGBoost, Random Forest, and Logistic Regression. The XGBoost classifier exhibits the strongest and most consistent performance in distinguishing human-authored from AI-assisted analyst reports, and is therefore adopted as our primary detection model. We also benchmark XGBoost against the following commercial AI-detection services: GPTZero and Originality.ai. We find that the commercial tools are mostly effective in identifying documents that are largely or fully AI-generated but are less effective in detecting moderate or partial forms of AI assistance. This distinction is important because, in practice, analysts are unlikely to rely on AI to generate complete reports, a task requiring nuanced judgment. See Appendix B for details.

certification or All-Star recognition, as well as for less experienced analysts. For instance, the effect for low industry experience analysts is approximately 1.6 times as large as that of high experience analysts. Overall, the results suggest that AI serves at least two beneficial roles: aiding the productivity of analysts facing greater scale demands, and partially substituting for differences in human capital and industry experience among analysts.

A concern with respect to our results thus far is that many of the linguistic features used to detect AI-assisted reports are similar to features that proxy for writing quality, and more experienced analysts may exhibit both higher writing quality and higher research quality. To address this concern, we exploit a plausibly exogenous shock to AI usage in analyst reports. In particular, we examine reports issued on days with overlapping earnings calls among firms covered by the same analyst, creating an exogenous increase in workload and time pressure to process information and produce reports. In the first stage of a two-stage least squares (2SLS) model, we observe higher AI usage in reports released on days with overlapping earnings calls. Critically, if our AI probability measure were capturing writing quality, the association would be the opposite of what we observe empirically. That is, ceteris paribus, we would expect to observe lower writing quality under workload and time pressure. In the second stage, we use overlapping earnings calls as an instrument for AI usage and confirm that AI usage positively impacts forecast accuracy.

Extending our analysis on research quality, we find that AI usage in analyst reports elicits significantly stronger reactions from capital markets. Specifically, forecasts based on reports with higher AI usage are associated with larger absolute abnormal returns around

the forecast date. Beyond mere forecast accuracy, these results suggest that AI-assisted reports increase the informativeness of analysts' investment research.

We next examine the broader implications of GenAI for the quality and impact of analysts' investment research. More specifically, we extract detailed research characteristics from analyst reports, including valuation models used, topics covered, stakeholders discussed, and public and private information sources. We find that reports with higher AI usage employ a wider range of valuation models, engage with a greater diversity of topics, and consider a broader set of stakeholders. At the same time, AI-assisted reports are less likely to reference private communications with management, suggesting a mechanism whereby AI complements public information processing rather than leveraging private information channels.

Finally, we examine whether AI usage in analyst reports is associated with differences in reasoning patterns. Using a linguistic framework widely adopted in prior research (Boyd et al. 2022; Li et al. 2025; Li et al. 2026), we construct measures that capture how analysts process and communicate information. Specifically, we analyze four dimensions: analytical language, authenticity, causal terms, and differentiation terms, which together reflect structured, explanation-oriented, and nuanced reasoning. We find that AI usage is positively associated with all four dimensions, suggesting that AI-assisted reports are associated with deeper analytical reasoning rather than purely stylistic changes.

We contribute to the emerging literature on AI adoption in financial research and analysis. Recent work in this area includes Bradshaw et al (2026) who study Seeking Alpha contributors and find that GenAI allows them to publish more articles and cover more firms; the authors further find although these reports are less informative, they nevertheless

contribute to liquidity and price discovery. Examining sell-side analysts, Bertomeu et al (2025) rely on a country-level shock (the temporary ChatGPT ban in Italy) to document that the lack of access to ChatGPT results in lower forecast frequency and accuracy by domestic analysts. In addition, Shanthikumar and Yoo (2025) show the magnitude of AI investments by investment banks, as proxied by AI-related job postings, is positively related to both the frequency and accuracy of earnings forecasts.

Relative to these country-level and bank-level findings, we shift the unit of analysis to the analyst-level, which allows us to uncover both which analysts incorporate AI into their research, and under what conditions AI-assisted research improves their forecast performance. Also, in contrast to both analyst studies, we train and customize a machine learning model to detect AI assistance specifically in the domain of sell-side equity reports. This also is a notable difference in approach from other recent work that relies on commercial AI detectors (Blankespoor et al., 2026; Bradshaw et al., 2026; Hirshleifer et al., 2025).[4]

The remainder of the paper is organized as follows. Section 2 provides literature review. Section 3 describes the data and the construction of our AI usage measure. Section 4 presents the baseline empirical results and heterogeneity effects of AI, discusses our identification strategies, and examines capital market reaction and mechanisms. Section 5 examines the robustness of our findings. Section 6 concludes.

[4] In our setting, commercial detection tools such as GPTZero or Originality.ai pose two significant disadvantages. First, they rely on algorithms that are proprietary, which presents challenges in terms of methodological understanding and transparency. Second, perhaps more critically, in the domain of analyst reports, commercial tools appear overly conservative in detecting AI use, which may be an artifact of wishing to avoid false accusations (see our comparative results in Section 5.5).

2. Prior Literature

2.1 AI and Capital Markets

The rapid evolution of artificial intelligence (AI) has transformed the landscape of accounting and finance by changing how information is produced, disseminated, and consumed. For instance, Sheng, Sun, Yang, and Zhang (2025) introduce a novel GenAI reliance measure based on institutional portfolio holdings and document a significant increase in GenAI usage among hedge funds following ChatGPT's launch, resulting in performance gains. Breiting and Muller (2025) use GPT-3 to construct historical business descriptions for US and international firms, and demonstrate the usefulness by obtaining significant alphas, as well as predicting M&A deals.

In the financial reporting area, Anantharaman, Rozario, and Parker (2023) identify public firms that adopt AI using a proprietary dataset. They find that AI adoption is associated with higher financial reporting quality, as reflected in lower discretionary accruals and closer mapping between accruals and cash flows. Blankespoor, DeHaan, and Li (2026) investigate the early adoption of GenAI in corporate financial reporting by applying the detection tool, GPTZero, to financial disclosures. The authors document statistically significant GenAI usage across disclosure types, finding that up to 4.5% of new disclosure text is written by AI, with higher adoption rates concentrated among smaller firms and those experiencing major corporate events.

In financial analysis, Bradshaw et al. (2026) investigate the adoption and consequences of GenAI on the Seeking Alpha platform, where AI-generated content peaked at 13.4% of articles before declining due to a strict anti-AI plagiarism policy. Findings indicate that adopting authors significantly improve their productivity by

publishing 55.1% more articles and expanding coverage to new firms. Although they find that individual AI articles are less informative than human ones, their low production costs enable positive effects on liquidity and price discovery.

2.2 AI and Sell-Side Investment Research

While traditional automation technologies and machine learning have long assisted financial analysis with data collection and modeling (Coleman et al., 2022; van Binsbergen, Han, Lopez-Lira 2022), the rise of AI 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 through both task automation and complementarity. Several recent studies examine the intersection of AI and sell-side investment research.

Li et al. (2025) use GenAI to analyze analyst reports, earnings call transcripts, and employee reviews in an effort to quantify the effects of corporate culture, showing that analysts' culture analyses differ from other stakeholders and are incorporated into their stock recommendations and target prices. Cao et al. (2024) build a supervised machine learning model trained on firm, industry, macro, and textual data to predict 12-month stock returns, finding that it outperforms the majority of human analysts but that humans retain an edge for informationally opaque firms. Interestingly, combining human stock return forecasts with the AI model improves on either alone, outperforming the standalone AI model in 54.8% of forecasts while reducing roughly 91% of analysts' extreme forecast errors.

Of the recent work, the closest in spirit to our study are Bertomeu et al (2025) and Shanthikumar and Yoo (2025). Bertomeu et al (2025) rely on a country-level shock (the

temporary ChatGPT ban in Italy) to document that the lack of access to ChatGPT results in lower forecast frequency and accuracy by domestic analysts. Shanthikumar and Yoo (2025) document that the magnitude of AI investments, as proxied by AI-related job postings, by investment banks is positively related to both the frequency and accuracy of earnings forecasts. Furthermore, these effects are more pronounced for firms with higher forecasting difficulty, such as those facing greater fundamental uncertainty or information asymmetry.

Different from these country and bank level findings, we shift the unit of analysis to the individual analyst level using a large sample of 38,468 reports from 73 brokerage houses. This allows us to examine different economic questions. In particular, which analysts leverage AI, and under what conditions is it beneficial? A second important distinction of our study is our approach to AI identification. We train and customize a machine learning model to detect AI assistance specifically in the domain of sell-side equity reports. This novel approach also sets us apart from recent capital markets research that relies on commercial AI detectors (Hirshleifer et al., 2025; Blankespoor et al., 2026; Bradshaw et al., 2026). In the domain of analyst reports, this is a critical point of differentiation as the commercial tools appear overly conservative at AI detection, which may be an artifact of wishing to avoid false accusations (see our comparative results in Section 5.5)

Prior work documents that human capital and resources impact analyst forecast accuracy. In terms of research quality, factors shown to be important include forecasting experience and employer resources (Clement 1999), industry expertise (Kadan, Madureira, Wang, and Zach 2012), access to in-house economic expertise (Hugon, Kumar, and Li

2016), among others. We therefore expect that AI serves as a similar resource by facilitating rapid synthesis of large volumes of textual and numerical information, resulting in improved research quality.

3 Data and Measurement

3.1 Data

Our empirical analysis focuses on sell-side equity research reports issued between January 1, 2022 and April 1, 2024.[5] 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 publishing institutions (investment banks or brokerage houses), contributing analyst(s), document date, firm names and tickers, report categories, subject tags, and report style classifications. These fields allow us to identify analysts, brokers, and covered firms, and to link textual reports to forecast activity. We restrict the sample to reports with non-missing analyst and firm identifiers.

Next, we link analyst reports to earnings forecasts from the I/B/E/S database using analyst identity, firm ticker, and forecast timing. Following Huang et al. (2014), Lv (2024), and Li et al. (2026), a report is matched to forecasts if it is issued within two calendar days before and after the forecast announcement date. In addition to analyst reports and earnings

[5] We extract raw text from each analyst report PDF using pdfminer.six, a Python library. We remove formatting artifacts introduced by the PDF extraction and retain only analyst-written narrative content. The cleaning process includes correcting broken lines, excluding non-linguistic text and numerical artifacts, and removing regulatory and boilerplate disclosures. The 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.

forecasts, we collect financial data from Compustat, and stock return data 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 hand-collect data from publicly available professional profiles, including LinkedIn, Capital IQ Professional, and similar publicly accessible sources. Following prior work (e.g., Li et al., 2025), we use professional profiles to manually compile information on professional certification 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.

3.2 AI Measurement

Motivated by recent advances in GenAI detection, we construct a supervised machine learning model designed to capture the presence of GenAI involvement in 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).

To construct training data for AI detection, we randomly select 2,600 analyst reports issued in 2021, prior to the public release of ChatGPT 3.5, a period during which GenAI usage in sell-side research was negligible. Figure 1 presents a timeline of key developments in ChatGPT's evolution. 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 or GPT-4—see Appendix B for details, while preserving the original analytical meaning, factual content, and writing structure. The AI-assisted reports are paired with fully human-authored reports as ground-truth data for supervised training and

evaluation of 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.

[Figure 1 here]

We compute 25 document-level linguistic features following Guo et al. (2023) and Uzun (2023)—See Table B1. Using these features, we train three supervised machine learning classifiers, Random Forest, Logistic Regression, and XGBoost, on 80 percent of the training sample 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 for AI detection. Figure 2 reports the relative importance of linguistic features in the GPT-4-trained model using XGBoost.

[Figure 2 here]

The strongest signals distinguishing AI-assisted from human-written analyst reports are grammatical patterns and vocabulary choices (e.g., grammar errors, unique words). This suggests AI-assisted reports tend to be more grammatically consistent and use more distinct vocabulary than their human-written counterparts. Readability and style measures such as Flesch score, spelling errors, and personal pronoun usage also contribute meaningfully, while features like repetition rate and smoothness add relatively little.

To identify potential AI-assisted analyst reports, we apply the trained XGBoost model to reports issued between January 1, 2022 and April 1, 2024. For each report, the

model produces a probability (*AI_Prob*) that the report incorporates GenAI, where a higher probability proxies for greater integration of GenAI into the analyst's research process.

Our approach detects linguistic patterns associated with AI writing, but we use this measure as a proxy for broader AI assistance in analysts' research-production process. Supporting this approach, Section 4 shows that *AI_Prob* is positively associated with richer analytical dimensions, greater valuation model diversity and topic breadth, and more analytical language.

3.3 AI Trend

Figure 3 plots the time-series evolution of the median AI usage probability (*AI_Prob*) during our sample period.[6] Following the introduction of GPT-4 in March 2023, we observe a clear upward trend in AI usage in analyst reports, consistent with broader adoption of GenAI tools in professional settings as these tools became more sophisticated. Notably, the growth in AI usage appears relatively muted in the initial post-release period of GPT 3. This muted pattern coincides with reports that some financial institutions restricted employees' use of ChatGPT during that period because of compliance, confidentiality, and data-leakage concerns.[7]

[Figure 3 here]

4. Empirical Research Design and Results

4.1 Sample Description

[6] In untabulated results, we find that reports issued by brokers reported to have ChatGPT bans exhibit lower AI usage during the restriction period relative to brokers without such restrictions. However, these policies are self-disclosed and media-reported, so it remains unclear whether other, unreported banks also implemented similar bans.

[7] These banks include JPMorgan, Bank of America, Citi, Deutsche Bank, Goldman Sachs, and Wells Fargo.

Table 1 reveals that the sample comprises 55,778 analyst forecast observations linked to 38,468 unique equity research reports, because a single report may correspond to multiple earnings forecasts for different fiscal quarters. These reports are issued by 434 analysts affiliated with 73 brokerage houses, covering 2,016 distinct firms and are distributed relatively evenly across quarterly periods, with coverage ranging from roughly 9% to 13% per quarter.

Table 2 shows the mean probability of AI usage (*AI_Prob*) is 0.241, though the median of 0.062 and wide standard deviation of 0.316 suggest that AI adoption is highly skewed, with many reports showing low AI involvement but a subset exhibiting substantial usage. Forecast accuracy (*Accuracy*) has a mean of 0.174 and median of 0.238. On average, analysts in the sample cover approximately 3 firms across 1.4 industries, and 68% of them are affiliated with top-10 brokerages, suggesting the sample skews toward larger, more established brokerage houses.

4.2 AI Usage and Forecast Accuracy

To examine whether the use of GenAI in analyst reports is associated with forecast accuracy, we follow Ben-Rephael et al. (2025) and Peng et al. (2022) and compute forecast accuracy (*Accuracy*) as the negative of 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.

Following the empirical design in Li et al. (2025), we estimate the following baseline specification:

$$\text{Accuracy}_{i,f,t} = \alpha + \beta_1 \text{AI_Prob}_{i,f,t} + \beta_2 \text{Controls} + \eta_{f,y} + \theta_{i,y} + \varepsilon_{i,f,t}$$

where $Accuracy_{i,f,t}$ denotes forecast accuracy associated with analyst i's report for firm f at time t. The variable *AI_prob* is a continuous measure capturing the probability that the underlying analyst report incorporates GenAI.

Following the analyst literature (Li et al. 2025, Flam et al. 2023, Harford et al. 2019), control variables include analyst and broker characteristics: *FirmsCovered* (the natural logarithm of one plus the number of firms covered by the analyst), *IndCovered* (the natural logarithm of one plus the number of industries covered), *AnalystFirmExperience* (the natural logarithm of the number of years the analyst has covered the firm), *ForecastFrequency* (the natural logarithm of the number of forecasts issued by the analyst during the year), *Horizon* (the natural logarithm of the number of days between the forecast date and the fiscal year-end), and *TOP10* (an indicator equal to one if the analyst is affiliated with a top-10 brokerage house).[8] These controls capture heterogeneity in analyst ability, workload, and information environment. The summary statistics of these control variables in Table 2 are consistent with the literature.

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 effects that are common to each firm-year cell and analyst-year cell. Standard errors are clustered at the analyst–year-quarter level to account for within-analyst correlation in forecast behavior

[8] Forecast frequency in our draft is the natural logarithm of one plus the number of quarterly EPS forecasts the analyst issued during the calendar year. Our raw *ForecastFrequency* has a mean of 298.95 (median 280). Analysts in our sample cover 20.7 firms on average, implying roughly 14 quarterly EPS forecast issuances per firm-year, consistent with an initial forecast plus 2-3 revisions around each quarterly earnings cycle. For comparison, Li et al. (2026) report a mean of 4.57 annual EPS forecasts per analyst-firm-year. Multiplying their benchmark by 20.7 firms per analyst-year and by four quarters per fiscal year yields a magnitude in line with ours.

during the same quarter (Li et al., 2025). All continuous variables are winsorized at the 1st and 99th percentiles.[9]

Table 3 reports the results from regressions examining the association between AI usage in analyst reports and forecast accuracy. Column (1) includes firm–year and analyst–year fixed effects, while Column (2) additionally includes control variables.[10] In both specifications, 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. These results indicate that analyst reports with a higher probability of incorporating GenAI are associated with greater forecast accuracy. In terms of economic significance, a one standard deviation increase in *AI_Prob* (31.6%) is associated with 4.7% improvement in forecast accuracy, relative to the sample mean of *Accuracy*.[11]

[Table 3 here]

4.3 Heterogeneity by Analyst Workload and Busyness

We implement a subsample analysis by partitioning the sample based on analyst workload and estimating the regressions separately across groups. Analysts face limited attention under high workload, which likely increases their reliance on AI for investment research. To test this prediction, we partition the sample based on several analyst workload proxies, including the number of firms covered (*FirmsCovered*), the number of industries covered (*IndCovered*), and the number of forecasts issued during the year

[9] Our results remain robust when we truncate the sample at the 1st–99th, 5th–95th, and 10th–90th percentiles of *AI_Prob*, ruling out the possibility that our findings are driven by outliers.

[10] These results are robust when replacing firm-year fixed effects with time-varying firm characteristics, such as size, book-to-market, and profitability.

[11] Calculated as the coefficient estimate on *AI_Prob* 0.026 in Column (2) multiplied by standard deviation (0.316), then divided by the sample mean of Accuracy in Table 2 of 0.174.

(*ForecastFrequency*). In Table 4, Columns (1) and (2) 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 improving forecast accuracy in the presence of higher workloads and time constraints.

[Table 4 here]

4.4 Heterogeneity by Analyst Human Capital

Prior research shows that differences in analysts' human capital lead to variation in forecast accuracy (Malloy, 2005; Chi et al., 2025). GenAI can improve the performance of cognitively demanding tasks, such as processing large volumes of information, structuring analysis, detecting patterns, and generating valuation insights, which typically require higher skills. As a result, AI may disproportionately benefit analysts with lower levels of human capital by partially substituting for these capabilities (Acemoglu and Restrepo, 2018). To evaluate this possibility, we partition the sample based on analyst human capital measures, including CFA status (*CFA*), All-Star recognition (*Star*), and analyst industry experience (*IndExpertise*). *CFA* is an indicator that equals one if the analyst is a CFA charterholder. To ensure that All-Star recognition captures ex ante analyst skill, we use the most recent All-Star lists available prior to 2022 to construct *Star*, an indicator variable that equals one if an analyst had been recognized as an Institutional Investor All-Star before 2022, and zero otherwise. This measure is treated as time-invariant over our sample period. *IndExpertise* is measured as the number of years the analyst has covered firms within the same industry.

The results in Table 5 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 the benefit of AI is more pronounced among less-experienced analysts. These findings suggest that GenAI can aid with 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 than other analysts.

[Table 5 here]

4.5 Identification: Overlapping Conference Calls

Because our AI detection measure is based on the linguistic characteristics of analyst reports, a concern is that it may partly reflect unobserved analyst characteristics, such as writing ability or communication skills, that could also be associated with research quality. To mitigate this concern, we implement an instrumental variable strategy that exploits exogenous variation in analysts' workload arising from overlapping earnings conference calls. Prior research shows that overlapping conference calls create workload pressure for analysts covering multiple firms (Driskill et al., 2020; De Franco et al., 2025). Importantly, De Franco et al. (2025) show that the occurrence of overlapping calls is largely driven by factors outside the analyst's control such as firms' fiscal calendars and headquarters' time zones.

A key point is that analyst reports are voluntarily issued; that is, the timing is controlled by the analyst. Hence, if workload simply reduces an analysts' ability to produce accurate research, a natural response would be to delay the publication. In contrast, if

GenAI enables analysts to better manage workloads, overlapping conference calls should primarily affect incentives to rely on such tools to increase the research efficiency.

Following De Franco et al. (2025), we define overlapping calls as multiple conference calls occurring on the same day for firms covered by the same analyst. We construct a report-level overlapping dummy variable, *Overlapping Dummy*, which takes the value of one if a report is released on a day when at least two firms under the analyst's coverage hold earnings calls, and zero otherwise. Our analyst report sample shows that 27% of reports are released on the same day as earnings calls.

[Table 6 here]

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 day as the instrument. Column (1) reports the first-stage results, where the dependent variable is *AI_Prob*, the probability that an analyst report incorporates GenAI. 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. Note here that if our AI probability measure were capturing writing quality, the association would be the opposite of what we observe empirically. In particular, we should observe lower writing quality under workload and time pressure.

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.[12]

4.6 AI-Assisted Forecasts and Capital Market Reaction

Next, to further shed light on the quality of AI-assisted forecasts, we examine the magnitude of the stock market reaction to analyst forecasts (|*CAR*|), defined as the absolute value of the cumulative three-day abnormal return around the forecast day (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 the underlying analyst report, regardless of whether the reaction is positive or negative.

AI may influence |*CAR*| by producing more structured and coherent report narratives, reducing linguistic noise, and improving the framing of key information. Hence, lowering investors' information-processing costs and facilitating 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 an assessment of whether GenAI usage affects not only analysts' forecast precision but also the extent to which analyst research moves capital markets.

To isolate the association between AI usage and market reactions, we estimate specifications that control for the earnings forecast revision, as well as the control variables in our main accuracy regressions (Table 3). 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

[12] In untabulated results, we restrict the sample to reports issued on days when at least one firm in the analyst's coverage portfolio hosts an earnings call, and compare analysts facing multiple concurrent calls with those facing only one. Both stages of the IV regression yield positive and statistically significant coefficients, consistent with our baseline findings.

(t=1.97) in Columns (1) and (2), respectively, suggesting that reports with a higher probability of incorporating generative AI are associated with stronger immediate market reactions. This evidence indicates that AI-assisted analyst research conveys information that investors perceive as more salient or impactful, thereby amplifying the intensity of investor responses to analyst forecasts.

[Table 7 here]

4.7 AI Usage and the Scope of Analytical Content in Analyst Reports

In this section, we delve into analyst reports and examine how GenAI empowers analysts' investment research. Beyond using GenAI as a writing tool, analysts can utilize it as an analytical and reasoning tool to improve their research quality. To explore these mechanisms, we examine several dimensions of research depth and scope documented in the literature.

Because the primary purpose of analyst reports is to guide investors through valuation and forecast implications, we begin by examining the use of valuation models. It is well established that financial analysts rely heavily on multiples-based models and under-utilize present value-based valuation models (Ben-David and Chinco 2025). To examine whether GenAI is associated with more comprehensive valuation analysis, we use GPT-5.1 to read through analyst reports and identify valuation models used in the report (see Appendix D for the prompt).[13] We construct an indicator variable, *Valuation*, which 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.

[13] We use GPT-5.1 because it was the most advanced generative AI model available at the time of implementation. Our results remain robust when replicated with newer models in a smaller sample.

Sheng et al. (2025) identify 14 topics from earnings calls that can be used by large language models to generate forward-looking signals. These topics include the US economy, global economy, financial prospects of the firm, financial prospects of the industry, earnings, revenues, and wages/salaries, among others. Following this framework, we use GPT-5.1 to classify and count the number of distinct topics (up to 14) discussed in each report (*Topics*).

A broader consideration of stakeholders may reflect more comprehensive and forward-looking analysis, as firm performance is shaped not only by shareholders but also by interactions with employees, customers, suppliers, regulators, and other parties. For example, Li et al. (2025) show that there is an increasing trend of environmental and social-related discussions in both earnings calls and analyst reports, highlighting the growing importance of stakeholder-oriented information in financial analysis. Generative AI may facilitate such broader coverage by helping analysts process diverse information sources and incorporate multiple stakeholder perspectives into their reports. We examine the breadth of stakeholder coverage (*Stakeholders*) by considering nine types of stakeholders, including shareholders, debt holders, the management team, employees, customers, suppliers, competitors, regulators, and local communities.

Brown et al. (2015) provide survey evidence that private communication with management is a valuable input to analyst forecasts in addition to their own research and public information sources. Bradley, Jame, and Williams (2023) further document that non-deal roadshows and site visits play an important role in shaping analysts' forecasts and can generate information advantages for institutional investors. However, to the extent that

GenAI enhances analysts' ability to extract insights from public information sources, their continued reliance on private communication as a marginal input may diminish accordingly.

Using GPT-5.1, we classify and count the number of distinct private and public communications referenced in each report. Private communications (*Private*) include direct interactions with management, such as phone calls, video meetings, emails, site visits (e.g., bus tours or field trips), non-deal roadshows, surveys, channel checks, fireside chats, and social events. Public communications (*Public*) include earnings calls, conference presentations, analyst/investor days, M&A calls, shareholder/analyst calls, annual shareholder meetings, guidance, and other publicly accessible disclosures. To avoid double counting, we consider only distinct communication sources within each report. For *Topics*, *Stakeholders*, *Private*, and *Public*, we construct count-based measures and scale them by the total number of sentences in the report (expressed per 100 sentences) to capture the relative density of each dimension.

Table 8 presents the results. We use the same specification as previous tables. The results show that AI usage is positively associated with the use of multiple valuation models, greater topic breadth, and broader stakeholder coverage. In contrast, it is negatively associated with private communication and not significantly associated with public communication discussed in the report. Overall, these findings suggest that GenAI is associated with more comprehensive and structured analytical content in analyst research, while reducing reliance on private information sources.

[Table 8 here]

4.8 AI Usage and Analysts' Analytical Language in Reports

We next examine whether AI usage in analyst reports is associated with differences in analysts' reasoning patterns as proxied by their use of analytical language. If GenAI helps analysts synthesize information and structure arguments, AI-assisted reports may exhibit more analytical language, including greater use of causal explanations and linguistic markers associated with nuanced reasoning. To capture reasoning styles, we use LIWC-22 (Linguistic Inquiry and Word Count), a widely used textual analysis software package, to construct linguistically based measures that have been adopted in fields such as politics and psychology (Boyd et al., 2022; Li et al. 2025).[14]

Specifically, we consider four dimensions of reasoning. *Analytical* measures the extent of logical and formal reasoning, and *Authentic* captures the extent of sincere and straightforward communication.[15] These two measures are based on a proprietary formula that summarizes several categories of function words. *Causation* measures the use of words that explain relationships between events or outcomes (e.g., because, therefore, make, and why), reflecting explanation-oriented reasoning. Finally, *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 these count-based measures, we scale them with the total number of words in a report to measure the density of each category.

The results reported in Table 9 show that AI usage is positively associated with all four reasoning dimensions. Together, these findings suggest that GenAI is linked to more structured, explanation-oriented, and nuanced communication, consistent with deeper analytical processing rather than merely stylistic changes.

[14] These features are designed to reflect how individuals process and communicate information and are not used in training the machine learning detection model.

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

[Table 9 here]

5. Additional Tests

5.1 Placebo Test: Pre-ChatGPT Sample

One concern is that our AI usage measure may just capture general linguistic features of analyst reports rather than actual AI assistance. To further rule out this alternative explanation, we conduct a placebo test using a pre-ChatGPT sample of analyst reports issued between January 2021 and October 2022, prior to the November 2022 public release of ChatGPT-3.5. Because GenAI is unlikely to be present during this period, the AI detection model should not yield a meaningful proxy for AI usage. We re-estimate the baseline regression model on this sample, retaining the same dependent variable, control variables, and fixed-effects as in the main test (Table 3). As reported in Table C1, the coefficient on *AI_Prob* is statistically indistinguishable from zero in both Columns (1) and (2). The absence of a significant relation in the pre-ChatGPT period is consistent with *AI_Prob* capturing GenAI usage rather than linguistic features of analyst writing.

5.2 Robustness: AI Usage Measured by an Indicator Variable

To assess whether our baseline findings are robust to the choice of a dichotomous AI usage measure, we re-estimate the accuracy specification using an indicator variable for AI usage. Specifically, we construct *AI_Indicator*, which equals one if the baseline XGBoost classifier labels the analyst report as AI-assisted (i.e., *AI_Prob* $\geq$ 0.5), and zero otherwise. This binary measure reduces reliance on the exact scaling of the continuous predicted probability and provides a robustness check based on the classifier's discrete AI-assisted classification. As reported in Table C2, the coefficient on *AI_Indicator* is positive and statistically significant at the 1% level without controls (Column 1) and at the 5% level

with controls (Column 2). These results confirm that the positive association between AI usage and forecast accuracy is not an artifact of the continuous *AI_Prob* measure.

5.3 Robustness: Alternative AI Detection Models

We also assess the robustness of the relation between AI usage and our primary dependent variable, forecast accuracy, to alternative assumptions about the degree of AI involvement and to different machine learning classifiers. We first re-estimate the baseline accuracy specification using AI usage measures constructed under alternative definitions of AI involvement. In the baseline setting, we generate ground-truth training data by prompting ChatGPT to rewrite 50% of the content in analyst reports. In Columns (1) and (2) of Table C3, we replace the baseline AI usage measure with *LowAI_Prob* and *HighAI_Prob*, which are based on alternative training data constructed by prompting ChatGPT to rewrite 25% (lower-intensity AI involvement) and 75% (higher-intensity AI involvement) of report content, respectively.

Columns (3)–(5) construct AI usage probabilities using a Logistic Regression classifier under different AI involvement 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 over an even earlier period, 2000–2020.

Across all specifications, AI usage measures, whether defined using alternative involvement 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 GenAI are not limited to heavy reliance on such tools. Even modest levels of AI assistance appear to improve analysts' forecasting performance. Overall, these additional analyses reinforce the interpretation that GenAI serves as a productivity-enhancing tool that improves the effectiveness of analysts' forecasting activities.

5.4 Robustness: Sentence-Level AI Detection

In our main analyses, we measure the probability of AI assistance at the report level. To shed light on the intensity of AI usage, we construct an alternative measure based on sentence-level AI detection. Specifically, we decompose each report in the training sample into sentences and prompt ChatGPT to rewrite each sentence, creating a ground truth sample consisting of originally human-written sentences and their AI-assisted counterparts. After comparing candidate classifiers, we find that the random forest classifier performs best. We then apply this classifier to our sample of analyst reports and re-measure AI usage as the fraction of sentences in a report classified as AI-assisted (*AI_SentRatio*). This alternative measure captures AI involvement at a finer level of granularity and mitigates concerns that the baseline results are sensitive to aggregation choices in the AI detection procedure. As reported in Table C4, the coefficient on *AI_SentRatio* is positive and statistically significant at the 1% level in both Columns (1) and (2). These results provide further evidence that the positive association between AI usage and forecast accuracy continues to hold using an alternative, sentence-level measure of AI assistance.

5.5 Comparison with Commercial Detectors

Finally, we compare model performance to the commercial detectors Originality.ai and GPTZero.[16] Due to cost considerations, we select a random sample of 50 analyst reports issued in 2021, apply the identical procedures to construct the training sample of paired analyst reports, and submit each analyst report to the commercial detectors to obtain document-level assessments of AI involvement.

Originality.ai provides a binary probability estimate of AI authorship at the document level that directly corresponds to our *AI_prob* measure. GPTZero produces a three-way probability distribution over the categories: human, AI-generated, and mixed. Following the approach in Blankespoor et al. (2026), we collapse the three-way prediction into a binary measure by calculating the estimated probability of AI authorship as the complement of human involvement. Specifically, we construct $AI_prob = 1 - prob_human$, which incorporates both the AI-generated and mixed components of GPTZero's output.

Table C5 provides the comparison results. The commercial detectors reveal a gap between discriminative ability and practical classification performance. For instance, GPTZero's AUC of 0.937 is nearly on par with XGBoost's 0.946, indicating it ranks AI-assisted versus human-written reports reasonably well. However, its accuracy is 0.500, with a TNR of 1.000 and TPR of 0.000, which suggests that it essentially never flags a report as AI-assisted at its default operating threshold. Originality.ai shows a similar directional bias. Overall, the results suggest that the commercial tools use an overly conservative threshold, which is likely meant to minimize false accusations of AI use in general-purpose applications. However, this makes them largely ineffective for a task where identifying AI-assisted analyst reports is the entire point.

[16] We use GPTZero's 2025-10-17-base model and originality.ai's model lite-1.0.0.

6. Conclusion

This study examines whether and how analysts' use of GenAI influences forecast quality. Using a supervised machine-learning approach to detect AI-assisted content in analyst reports, we find that AI-assisted reports are associated with improved forecast accuracy. We also document stronger investor reactions to forecasts accompanied by AI-assisted reports, indicating that markets perceive the underlying reports as incrementally informative. Cross-sectional analyses show that the accuracy benefits of AI are more pronounced for analysts facing heavier workload demands, analysts without a CFA designation or prior All-Star recognition, and less experienced analysts. Channel analyses suggest that GenAI improves analysts' investment research through richer valuation analysis and broader informational coverage, including the integration of more valuation models and diverse analytical dimensions.

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 the benefits of AI adoption in information production. 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. 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 GenAI in analyst reports, which may not capture all forms of AI assistance.

References

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.

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.

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

Ben-David, I., & Chinco, A. (2025). Capital budgeting for EPS Maximizers, Working paper.

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,* 101782.

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. (2026). Generative AI in financial reporting. *Journal of Accounting Research* Forthcoming.

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

Bradshaw, M. T. (2004). How do analysts use their earnings forecasts in generating stock recommendations? *The Accounting Review*, 79(1), 25-50.

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. (2026). Generative AI use by capital market information intermediaries: Evidence from Seeking Alpha. *Journal of Accounting Research* Forthcoming.

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.

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.

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

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

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

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. (2025). Generative AI and firm values. *Journal of Finance*, forthcoming.

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

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.

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.

Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9004–9017).

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

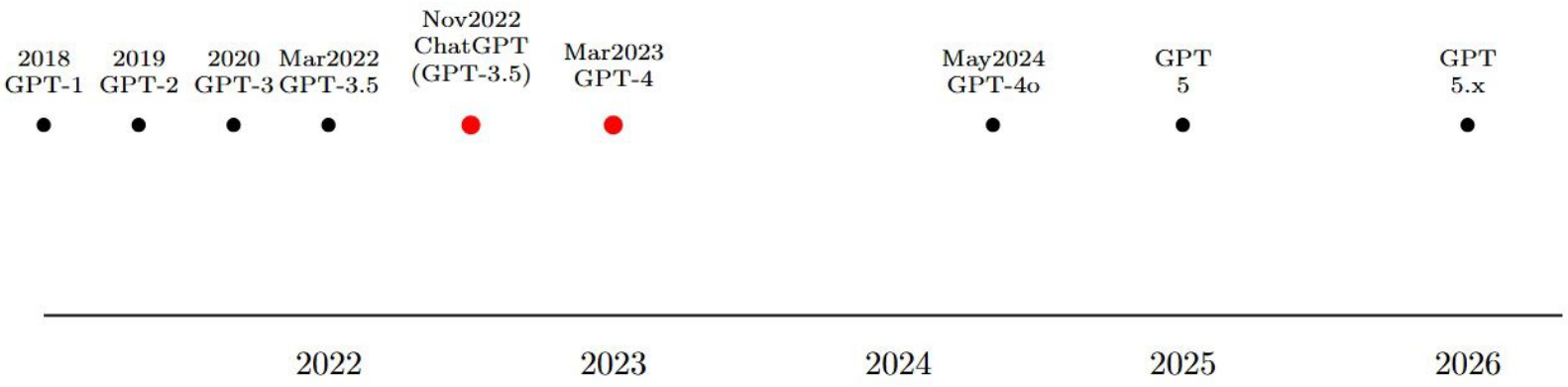
Shanthikumar, D. M., & Yoo, I. S. (2025). Artificial intelligence and analyst productivity. SSRN Working Paper.

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).

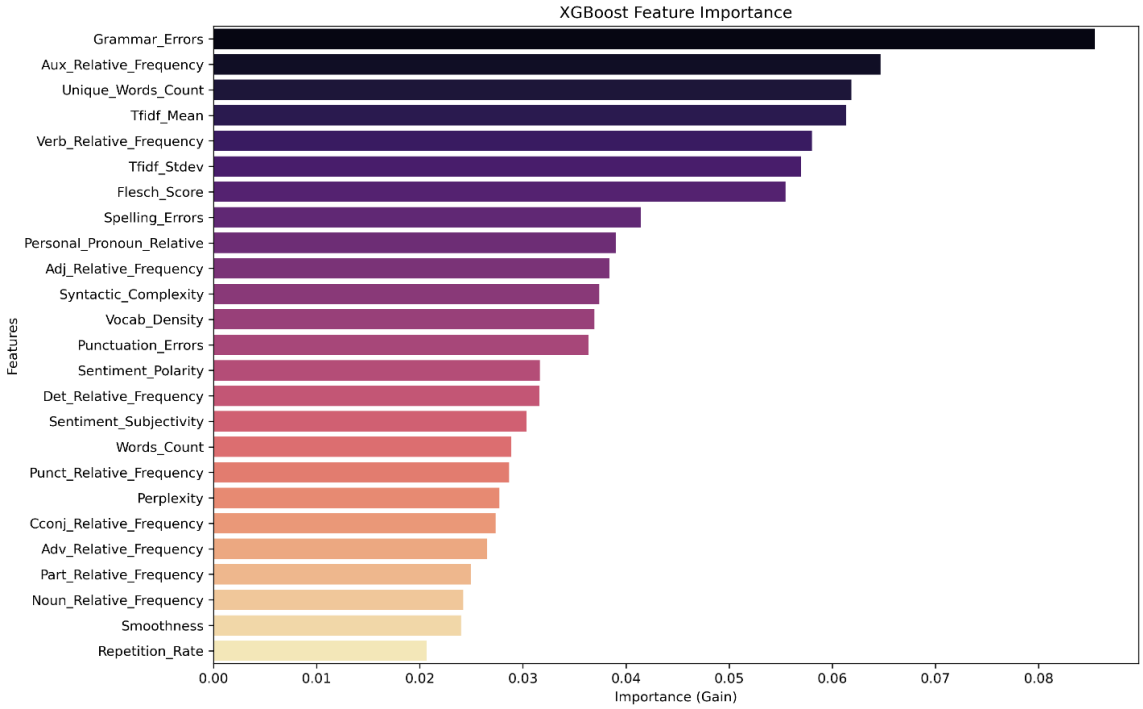
Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection. *Computational Linguistics, 51*(1), 275–338.

Figure 1: Timeline of Generative AI Model Development



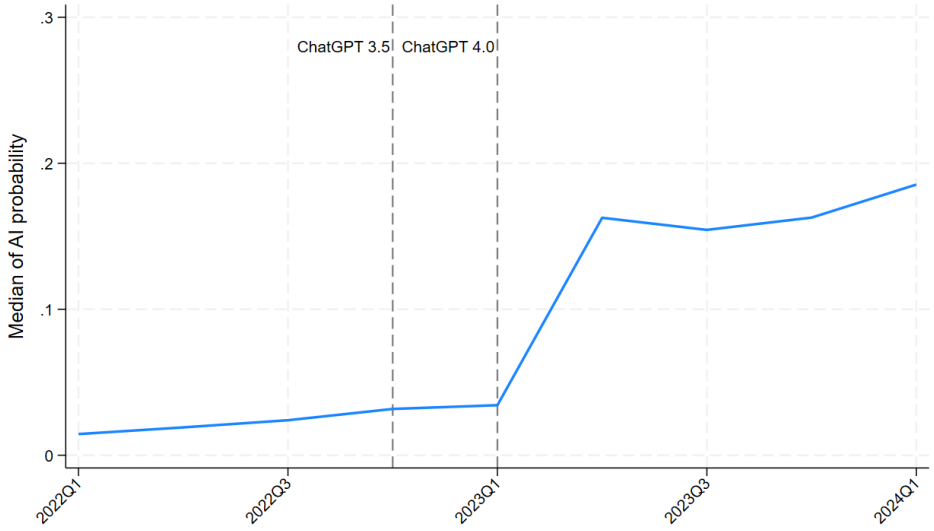
Notes: This figure illustrates the timeline of major generative AI model releases by OpenAI. During our sample period (2022–2024Q1), ChatGPT-3.5 (released November 2022) and GPT-4 (released March 2023) represent the dominant generative AI tools in practical applications, highlighted in red. While other large language models from Google (Bard, later Gemini), Anthropic (Claude), Microsoft (Bing Chat, later Copilot), and Meta (LLaMA) were introduced during this period, their adoption remained limited relative to ChatGPT-based tools. Our training and empirical design focus on the ChatGPT-3.5 and GPT-4 periods.

Figure 2. Important Features in the Classification Model



Notes: This figure reports feature importance from the XGBoost classifiers used to detect AI-assisted analyst reports trained on GPT-4–generated rewrites (Note: similar feature importance with GPT-3.5-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.

Figure 3. Trends of AI Probability in Reports



Notes: This figure plots the time-series evolution of the median document-level AI usage probability (*AI_Prob*) in analyst reports by quarter. The vertical dashed lines mark the introduction of ChatGPT-3.5 (November 2022) and GPT-4 (March 2023). The median AI probability remains relatively flat during the GPT-3.5 period, consistent with limited early adoption due to reputational concerns, nascent technology, and 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. AI usage exhibits a more pronounced rise after the introduction of GPT-4, as these restrictions were gradually relaxed or supplemented by internally developed AI systems.

Table 1. Sample Description

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

	Total	Unique
Broker	55,778	73
Analyst	55,778	434
Firm	55,778	2,016
Report	55,778	38,468

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

Panel B. 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

Table 2. Summary Statistics

Notes: This table reports summary statistics for the main variables used in the analysis. It presents the mean, median (P50), and standard deviation (SD) for the full sample. *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*. For logged variables, the corresponding raw values are also reported with the suffix *_Raw*. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A.

	All (N=55,778)				
	Mean	P50	SD	P25	P75
AI_Prob	0.241	0.062	0.316	0.006	0.414
Accuracy	0.174	0.238	0.572	-0.137	0.609
FirmsCovered	2.990	3.045	0.455	2.773	3.296
FirmsCovered_Raw	20.729	20.000	8.389	15.000	26.000
IndCovered	1.443	1.386	0.550	1.099	1.946
IndCovered_Raw	3.908	3.000	2.642	2.000	6.000
AnalystFirmExperience	1.505	1.479	0.862	0.809	2.176
AnalystFirmExperience_Raw	5.491	3.389	5.748	1.247	7.811
ForecastFrequency	5.486	5.638	0.769	5.247	5.948
ForecastFrequency_Raw	298.954	280.000	168.288	189.000	382.000
Horizon	4.657	4.543	0.753	4.454	5.204
Horizon_Raw	130.749	93.000	84.755	85.000	181.000
TOP10	0.684	1.000	0.465	0.000	1.000

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. 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
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. 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)
Controls	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
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 *IndExpertise*. For CFA and All-Star recognition, we split the sample based on indicator values (Yes/No). For *IndExpertise*, 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. 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		IndExpertise	
	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)
Controls	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
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. 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	(2) Second-stage Accuracy
Overlapping Dummy	0.020***	
	(3.59)	
AI_Prob		1.924***
		(2.61)
Controls	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. Control variables include those used in Table 3 (*FirmsCovered*, *IndCovered*, *AnalystFirmExperience*, *ForecastFrequency*, *Horizon*, and *TOP10*), as well as *Revision*, defined as the absolute difference between the analyst's current forecast and the trailing 30-day consensus forecast. 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. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) |CAR|	(2) |CAR|
AI_Prob	0.002*	0.002**
	(1.78)	(1.97)
Revision		0.013***
		(2.66)
FirmsCovered		0.001
		(0.20)
IndCovered		-0.000
		(-0.05)
AnalystFirmExperience		0.002*
		(1.72)
ForecastFrequency		-0.003***
		(-2.98)
Horizon		0.005***
		(9.33)
TOP10		0.003
		(0.19)
Firm-Year	Yes	Yes
Analyst-Year	Yes	Yes
Observations	55,778	55,778
Adj R-squared	0.395	0.397

Table 8. AI Usage and the Scope of Analytical Content in Analyst 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. *Topics*, *Stakeholders*, *Private*, and *Public* are scaled per 100 sentences of report text. 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. 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)
Controls	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
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. 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)
Controls	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
Adj R-squared	0.0290	0.0539	0.0729	0.0638

Appendix A. Variable Definitions

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. Specifically, reports issued between the release of ChatGPT-3.5 (November 2022) and GPT-4 (March 2023) are classified using a model trained on GPT-3.5 rewritten text, while reports issued after GPT-4's release are classified using a model trained on GPT-4 rewritten text.
Accuracy	*Accuracy* is defined as the negative of the proportional mean absolute forecast error (*PMAFE*). Following Ben-Rephael et al. (2025), *PMAFE* is calculated as the analyst's absolute forecast error minus the average absolute forecast error across all analysts covering the same firm and forecast period, scaled by the average absolute forecast error across all analysts covering the same firm and forecast period.
|*CAR*|	*/CAR/* is defined as the absolute value of the cumulative three-day abnormal return [-1, 1] around the forecast date.
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.
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.
AnalystFirmExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for the corresponding firm.
ForecastFrequency	The logarithm of one plus the number of quarterly EPS forecasts the analyst issued during the calendar year
Horizon	The logarithm of one plus the number of calendar days between the forecast date and the corresponding earnings forecast date.
TOP10	*TOP10* is an indicator variable that equals one if an analyst works at a top 10 brokerage house (ranked by the number of equity analysts) in year.
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, scaled per 100 sentences.
Public	*Public*, is the number of distinct types of public communication between an analyst and the management mentioned in a report, scaled per 100 sentences.

Topics	*Topics*, is the number of distinct topics mentioned in a report, following Sheng et al. (2025), scaled per 100 sentences.
Stakeholders	*Stakeholders*, is the number of distinct types of stakeholders mentioned in a report, scaled per 100 sentences.
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.
IndExpertise	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for the corresponding industry.
SIZE	The value of firm size in a given quarter, measured as the natural logarithm of market value of equity.
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.
Revision	The absolute value of the analyst's forecast revision, defined as the absolute difference between the analyst's current earnings forecast and the trailing consensus forecast, where the consensus is measured as the mean forecast across all analysts covering the same firm over the preceding 30 calendar days.
AI_Indicator	An indicator variable equal to one if the document-level classification label produced by the baseline XGBoost classifier identifies the analyst report as AI-assisted (i.e., *AI_Prob* ≥ 0.5), and zero otherwise.
LowAI_Prob	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.
HighAI_Prob	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.
LowAI_Prob_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.
HighAI_Prob_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.
LowAI_Prob_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.
HighAI_Prob_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.
AI_SentRatio	A sentence-level measure of AI usage in an analyst report. It is computed as the fraction of sentences in a report classified as AI-assisted by a sentence-level Random Forest classifier trained on paired human-written analyst sentences and their GPT-rephrased counterparts. Random Forest is adopted at the sentence level because it outperforms XGBoost and Logistic Regression in distinguishing human- from AI-generated sentences based on sentence-level linguistic features. The measure ranges from zero to one, with higher values indicating greater AI involvement.

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

This appendix describes the construction and validation of the analyst report AI usage measure (*AI_Prob*). *AI_Prob* captures the probability that an analyst report incorporates a specific level of GenAI—50% in the main tests and 25% or 75% in the robustness tests—and serves as the primary measure of AI usage throughout the paper. The detection framework consists of four steps: 1) construction of training data, 2) feature extraction, 3) model training, and 4) generation of the *AI_Prob* measure.

B.1 Construction of Training Data

We randomly select 2,600 analyst research reports issued in 2021 (the selection criterion is at least three reports by each available analyst), prior to the public release of ChatGPT. The text is extracted from the PDF-based analyst reports using pdfminer.six, a Python library for PDF text extraction. We clean the text by removing formatting artifacts, non-linguistic content (e.g., URLs), and boilerplate disclaimers.

We next construct AI-assisted versions of analyst reports for supervised training. For each report, we generate AI-assisted counterparts by rewriting 50% of the original analyst-written content using a GenAI model. Specifically, we use GPT-3.5 to generate AI-assisted versions for reports issued 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 after the release of GPT-4 (March 2023). Each original report is submitted to the model for sentence-by-sentence rewriting using the following prompt:[17]

'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.'[18]

[17] Sentence-level rewriting serves two purposes. First, it allows us to calibrate the intensity of AI usage, which we set at 50% in our main tests and at 25% and 75% in our robustness tests. Second, as a practical matter, our cleaning procedures fragment a report's original paragraph structure, making sentence-level rewriting a more feasible unit of analysis relative to paragraph-rewriting.

[18] As a robustness test on our prompt engineering, we applied a second rewriting prompt that omits the contextual framing; that is, it only includes the instruction steps 1-5. We compared AI-detection results across the two approaches using a subsample of reports and found no significant differences in detection outcomes.

The paired analyst reports, i.e., original and AI-assisted, form the labeled dataset used for supervised training. As an alternative to our 50% AI-intensity benchmark, we also construct 25% and 75% AI-assisted benchmarks and find consistent results

B.2 Feature Engineering

We extract a comprehensive set of document-level linguistic features designed to distinguish human-written from AI-assisted reports. These features are based on 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):[19]

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.

Table B.2 reports summary statistics for these features from the training sample, separately for human-written and AI-assisted analyst reports.[20] 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 model training, all features are standardized so that differences in scale do not affect classification performance.

[19] Features that depend on surface-level characteristics, such as grammatical errors, are computed from the pre-processed text.

[20] For conciseness, only GPT-3.5 statistics are reported.

Table B1. 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.

Table B2. Summary Statistics of Features in the Training Sample

	Feature	Human_Mean	AI_Mean	Diff.	Tstat	Pvalue	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

Using the 2,600 labeled analyst reports, 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 (TPR), true negative rate (TNR), 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. For brevity, we only report detailed performance metrics for the GPT-3.5 model in Table B3. (The GPT 4 model results are qualitatively similar.)

Table B3. Classification Performance Results

Comparison	AUC	Accuracy	TNR	TPR	F1-score
XGBoost	**0.947**	**0.877**	**0.860**	**0.894**	**0.877**
Random Forest	0.922	0.841	0.836	0.845	0.841
Logistic Regression	0.912	0.844	0.834	0.855	0.844

B.4 Construction of Document-Level AI Usage Measure

We apply the trained XGBoost classifiers to the 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 after March 2023. We use the 50% AI-assisted benchmark as our primary specification because it 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 GenAI. *AI_Prob* ranges from zero to one, with higher values indicating a greater likelihood of AI assistance.

Appendix C: Additional Tests

Table C1. Placebo Test: Pre-ChatGPT Sample

Notes: This table reports OLS regression results of placebo tests that re-estimate the baseline accuracy specification on a pre-ChatGPT sample of analyst reports issued between January 2021 and October 2022, prior to the November 2022 public release of ChatGPT-3.5. Because generative AI usage in sell-side research was rare or absent during this period, the baseline AI detection model should not yield a meaningful association between *AI_Prob* and forecast accuracy. The dependent variable is *Accuracy*. The key independent variable is *AI_Prob*, the document-level AI usage probability produced by the baseline XGBoost classifier. 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. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy
AI_Prob	0.009	0.006
	(0.67)	(0.49)
FirmsCovered		-0.150
		(-1.39)
IndCovered		0.092
		(1.20)
AnalystFirmExperience		0.028**
		(2.52)
ForecastFrequency		0.097
		(1.54)
Horizon		-0.099***
		(-14.46)
TOP10		-0.133*
		(-1.95)
Firm-Year	Yes	Yes
Analyst-Year	Yes	Yes
Observations	35,778	35,778
Adj R-squared	0.274	0.288

Table C2. Robustness: AI Usage Measured by an Indicator Variable

Notes: This table reports OLS regression results examining the robustness of the baseline result to an alternative, discrete measure of AI usage. The dependent variable is Accuracy. The key independent variable is *AI_Indicator*, an indicator equal to one if the baseline XGBoost classifier labels the analyst report as AI-assisted (i.e., *AI_Prob* $\geq$ 0.5), and zero otherwise. 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. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy
AI_Indicator	0.019***	0.013**
	(2.99)	(2.14)
FirmsCovered		0.117**
		(2.02)
IndCovered		-0.072
		(-0.93)
AnalystFirmExperience		0.043***
		(4.54)
ForecastFrequency		-0.061***
		(-5.14)
Horizon		-0.121***
		(-19.68)
TOP10		0.491***
		(3.36)
Firm-Year	Yes	Yes
Analyst-Year	Yes	Yes
Observations	55,778	55,778
Adj R-squared	0.235	0.260

Table C3. 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*. *LowAI_Prob* and *HighAI_Prob* 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 (*LowAI_Prob_LR*, *HighAI_Prob_LR*, *LowAI_Prob_RF*, and *HighAI_Prob_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. 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
LowAI_Prob	0.021**				
	(2.15)				
HighAI_Prob		0.019*			
		(1.86)			
LowAI_Prob_LR			0.068***		
			(2.95)		
AI_Prob_LR				0.060***	
				(3.88)	
HighAI_Prob_LR					0.049***
					(3.03)
Controls	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
Adj R-squared	0.260	0.260	0.260	0.261	0.258

Table C3. Robustness: Alternative AI Detection Models (Continued)

	(6) Accuracy	(7) Accuracy	(8) Accuracy	(9) Accuracy
LowAI_Prob_RF	0.046**			
	(2.25)			
AI_Prob_RF		0.054***		
		(3.07)		
HighAI_Prob_RF			0.053***	
			(3.09)	
AI_Prob_2000-2020				0.021**
				(2.09)
Controls	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
Adj R-squared	0.260	0.260	0.260	0.260

Table C4. Robustness: Sentence-Level AI Ratio

Notes: This table reports OLS regression results examining the robustness of the baseline result to a sentence-level measure of AI usage. The dependent variable is *Accuracy*. The key independent variable is *AI_SentRatio*, the fraction of sentences in an analyst report identified as AI-assisted, with higher values indicating a greater share of AI-assisted content within a report. 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. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy
AI_SentRatio	0.178***	0.168***
	(4.38)	(4.25)
FirmsCovered		0.123**
		(2.14)
IndCovered		-0.079
		(-1.03)
AnalystFirmExperience		0.043***
		(4.46)
ForecastFrequency		-0.061***
		(-5.17)
Horizon		-0.119***
		(-19.54)
TOP10		0.491***
		(3.34)
Firm-Year	Yes	Yes
Analyst-Year	Yes	Yes
Observations	55,170	55,170
Adj R-squared	0.235	0.260

Table C5. Comparison with Commercial Detectors: Originality.ai and GPTZero

Note: This table reports classification performance on the 50-report sample described in Section 5.5, evaluated at the 50% AI intensity benchmark. Model performance is assessed using standard classification metrics, including the area under the ROC curve (AUC), accuracy, true positive rate (TPR), true negative rate (TNR), and F1 score.

Comparison	AUC	Accuracy	TNR	TPR	F1-score
Originality.ai	0.916	0.650	0.980	0.320	0.478
GPTZero	0.937	0.500	1.000	0.000	0.000
XGBoost	0.946	0.830	0.840	0.820	0.830
Random Forest	0.934	0.850	0.900	0.800	0.850
Logistic Regression	0.887	0.820	0.860	0.780	0.820

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:
"""