Generative AI and Investment Research: Evidence from Analyst Reports

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

Abstract

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

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

Shawn Huang: Associate professor of accounting, School of Accountancy, Arizona State University, Email: shawn.huang@asu.edu. Artur Hugon: Associate professor of accounting, School of Accountancy, Arizona State University. Email: Artur.Hugon@asu.edu. Tengfei Zhang: Assistant professor of finance, Rutgers School of Business-Camden, Rutgers University. Email: tengfei.zhang@rutgers.edu. Wenting Zheng: Ph.D. student, School of Accountancy, Arizona State University. Email: w.zheng@asu.edu. We thank seminar participants at Arizona State University and Rutgers University.

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

— ABI Research, 2025

1. Introduction

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

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

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

To investigate how AI affects the production of analyst reports, we develop a novel measure of AI usage by applying a supervised machine-learning detection framework to a large corpus of analyst reports, identifying reports with a higher likelihood of incorporating generative AI. Our analyst report sample consists of about 51,000 research reports, covering the period from January 2021 to April 2024. To construct training data for the detection model, we prompt ChatGPT to rewrite analyst reports issued prior to the public release of ChatGPT-3.5 on November 30, 2022, thereby generating AI-assisted versions of originally human-authored text. Using this paired dataset of human-written and AI-assisted analyst reports, we train machine learning models based on 25 linguistic features

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

We evaluate three supervised classification models, including XGBoost, Random Forest, and Logistic Regression. We find that the XGBoost classifier exhibits the strongest and most stable performance in distinguishing human-authored from AI-assisted analyst reports and is therefore adopted as our primary detection model. The model performs well in settings where portions of human-written reports are substituted with AI-assisted content. We also benchmark our self-trained model against widely used commercial AI-detection services, including GPTZero and Originality.ai. These commercial tools are mostly effective in identifying documents that are largely or fully AI-generated, but are less sensitive to more moderate or partial forms of AI assistance. This distinction is important because existing evidence suggests that the share of AI-generated text in practice is relatively small, and analysts are unlikely to rely on AI to produce a full report given the reputational risks in this highly competitive profession. Applying the trained XGBoost classifier, we generate a continuous, document-level probability measure (AI_Prob) for each analyst report, capturing the likelihood that the report incorporates generative AI. Using this measure, we document a clear upward trend in AI usage in analyst reports over the sample period. To validate our AI measure, we exploit a ChatGPT ban imposed by three brokerage houses in early 2023. We document that AI usage remains flat in analyst reports issued by these firms, while AI usage continues to increase in reports produced by other brokers.

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

usage is linked to a meaningful improvement in relative forecast accuracy, corresponding to roughly a 15% gain compared to the sample mean. This relation reflects AI's role in facilitating analysts' information acquisition and insight generation rather than improving writing quality only.

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

Third, we study how generative AI empowers analysts' investment research. Using GPT-5.1 as a reasoning agent, we extract detailed research characteristics from analyst reports, including valuation models used, public and private information sources, topics covered, and stakeholders discussed. We find that reports with higher AI usage feature more diverse valuation models, incorporate more public information sources, cover broader topics, address stakeholder impacts more extensively, and rely more heavily on textual analysis relative to quantitative metrics, compared to reports with lower AI usage. We further examine whether AI usage in analyst reports is associated with differences in reasoning patterns. Using LIWC-22, a linguistically based framework widely adopted in prior research (Boyd et al. 2022; Li et al. 2021; Li et al. 2025), we construct measures that

capture how analysts process and communicate information rather than general writing style. Specifically, we analyze four dimensions: analytical and authentic language patterns, as well as the use of causal and differentiation terms that reflect explanation-oriented and nuanced reasoning. We find that AI usage is positively associated with all four dimensions, suggesting that generative AI is linked to more structured, explanatory, and nuanced communication, consistent with deeper analytical reasoning rather than purely stylistic changes. These patterns suggest that analysts leverage AI not simply as a writing tool, but as an analytical and reasoning tool that deepens and broadens the scope of investment research.

Fourth, our cross-sectional analyses show that the effectiveness of AI varies systematically with analyst characteristics. Specifically, the accuracy benefits of AI are larger for analysts who cover more firms and issue forecasts more frequently, consistent with AI helping analysts manage scale and workload pressures. We also find that compared to male analysts, female analysts experience relatively greater accuracy gains from AI usage, despite their lower adoption rates, suggesting that AI may help mitigate pre-existing resource or time constraints. In addition, more experienced analysts realize stronger accuracy improvements when incorporating AI into their work, indicating that AI acts as a complement to, rather than a substitute for, analyst expertise.

Finally, we find that AI usage in analyst reports elicits significantly stronger reactions from capital markets. Specifically, reports with higher AI usage are associated with larger absolute cumulative abnormal returns (|CAR|) around the report date, indicating that markets respond more strongly to analyst research that is more likely to incorporate generative AI. Also, analysts with higher AI usage issue faster reports after earnings calls,

indicating AI increases the timeliness of forecasts. These results suggest that AI-assisted reports contain greater information content when they deliver investment advice. We conclude that GenAI adoption increases the informativeness of analysts' investment research.

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

Second, we extend the emerging literature on AI adoption in finance, which has largely focused on institution- or industry-level measures (Blankespoor et al., 2024; Kim et al., 2024), by shifting attention to AI usage at the individual analyst level. This dimension remains largely unexplored but is crucial for understanding how generative AI shapes the production of investment research. Prior to the widespread diffusion of tools such as ChatGPT, AI applications in finance were primarily institution-driven (Coleman et al., 2022; Davenport & Bean, 2017). For example, Shanthikumar and Yoo (2025) show that

investment banks' AI investments are positively associated with both the frequency and accuracy of earnings forecasts. More recent studies examine GenAI-related shocks to analyst activity, such as Italy's ChatGPT bans (Bertomeu et al., 2025) or the introduction of GenAI platforms in financial data services (Xue et al., 2025). Our paper complements this literature by measuring AI usage directly at the analyst level using linguistic signals in research reports, which allows us to infer both whether and how individual analysts incorporate AI into their work, including writing, information acquisition, valuation, topical coverage, and reasoning patterns. We further exploit analyst-level variation in workload shocks from overlapping earnings calls to establish causality and to study the distributional effects of AI adoption across analysts.

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

2. Prior Literature and Hypothesis Development

2.1 AI and Analyst Forecasts

The rapid evolution of artificial intelligence (AI) has transformed how information is produced, disseminated, and consumed. ChatGPT, for instance, achieves human-level performance on complex professional tasks and demonstrates multilingual capabilities across diverse knowledge domains (OpenAI, 2023). Not surprisingly, AI can generate

substantial productivity gains (Acemoglu 2025; Eisfeldt et al. 2024; Eloundou et al. 2024; Jia et al. 2024; Noy & Zhang 2023). Noy & Zhang (2023) show that AI assistance can materially improve productivity, for example, reducing the time required for writing tasks by roughly 40 percent while simultaneously improving quality by about 18 percent. Another key benefit of AI lies in its capacity to learn from large amounts of information, enabling it to generate coherent and contextually rich outputs. OpenAI (2025a) notes that ChatGPT "learn patterns from large amounts of information, including text, images, audio, and video." More recent AI incorporates Retrieval-Augmented Generation (RAG)[1], enabling models to automatically fetch relevant external information at runtime to enrich their responses (OpenAI, 2025b). While such technological breakthroughs have raised concerns about the restructuring of task allocation, with AI complementing high-skill roles that rely on information processing and judgment (Acemoglu & Restrepo, 2018), capital markets are also beginning to grapple with the dual implications of AI for both information producers (e.g., managers, analysts) and information users (e.g., investors, regulators) (Bertomeu et al., 2025; Blankespoor et al., 2024; Bradshaw et al., 2025; Kim et al., 2024; Shanthikumar & Yoo, 2025; Sheng et al., 2025).

Turning to information intermediaries, AI tools are transforming how they process data, draft reports, and deliver insights. While traditional automation technologies have long assisted financial analysts with data collection and modeling, such as Robo-Analysts (Coleman et al., 2022), the rise of LLMs enables data analytics and human-like text generation, which are especially suited for tasks involving communication, synthesis, and

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

interpretation. In the context of financial analysis, these capabilities can enhance analyst productivity both through task automation and task complementarity. Recent research provides early evidence of these effects. Shanthikumar and Yoo (2025) document that the magnitude AI investments by investment banks is positively related to both the frequency and accuracy of earnings forecasts. These effects are most pronounced for firms with high forecasting difficulty, such as those facing greater fundamental uncertainty or information asymmetry. Other studies examine broader implications. Blankespoor et al. (2024) argue that AI-generated text can affect investor interpretation, while Bertomeu et al. (2025) use the temporary ChatGPT ban in Italy to show that restrictions on AI reduce public information processing, leading to decreased news engagement.

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

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

Despite their central role in capital markets, the process through which sell-side analysts generate earnings forecasts remains largely unobservable. In theory, analysts synthesize vast and heterogeneous information inputs, including corporate disclosures, macroeconomic trends, peer performance, and proprietary data, and exercise substantial discretion in how these inputs are weighted and interpreted. However, this process is rarely observable and not well understood by external stakeholders, creating ambiguity about the origin and objectivity of forecast outputs. For instance, Barniv et al. (2009) note that analysts may strategically adjust their published forecasts in response to reputational concerns or perceived investor expectations, suggesting that forecast outputs may be as much a product of incentive-driven judgment as analytical rigor. The "black box" nature is further compounded by analysts' selective use of peer firms when forming performance benchmarks (De Franco et al., 2015), and the potential influence of cognitive or social biases such as herding (Clement & Tse, 2005). As Chi et al. (2025) emphasize, analysts often lack transparency in their incorporation of "big data" sources. Therefore, it is difficult to know what tools they use in their analysis.

It is worth noting that beyond numerical forecasts and recommendations, analysts also provide reports with a rich textual medium that encapsulates analysts' qualitative assessments, strategic insights, and interpretive judgments. These reports are often viewed by institutional investors as the most critical component of sell-side research, surpassing stock picks and earnings estimates in perceived value (Huang et al., 2014). Analysts synthesize firm-specific fundamentals and industry dynamics into report discussions, typically averaging nearly eight pages per report, that both complement and contextualize their quantitative outputs. Empirical evidence suggests that this textual content provides

incremental information beyond contemporaneous earnings forecasts and stock recommendations, aiding investors in interpreting conflicting signals or assessing long-term prospects. For example, Huang et al. (2014) demonstrate that negative textual sentiment in reports is weighted more heavily by markets than positive sentiment and that such text better predicts earnings growth over a five-year horizon. Complementing this view, Asquith et al. (2005) find that the strength of qualitative arguments in analyst reports, especially those written by star analysts, holds explanatory power for abnormal stock returns, even after controlling for forecast revisions. Moreover, the readability and structure of these narratives also matter. De Franco et al. (2015) show that clearer, more readable analyst reports are associated with greater market reaction and improved analyst performance.

2.3 Predictions and Hypothesis Development

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

Prior work documents that forecast accuracy varies significantly across analysts. Clement (1999) shows that accuracy is higher among analysts with greater forecasting experience and employer resources, and lower for those covering broader and more complex portfolios. Malloy (2005) finds that analysts located closer to the firms they cover

are more accurate than distant peers, particularly for small or remote firms, because proximity confers informational advantages. Similarly, Bae et al. (2008) show that local analysts across countries have a consistent accuracy advantage when firms provide less transparent disclosures. Taken together, these findings imply that forecast accuracy improves when analysts can overcome informational frictions through resources or enhanced data. More recently, Chi et al. (2025) provide evidence that analysts who incorporate alternative data into their research produce significantly more accurate earnings forecasts, suggesting that technology-enabled information processing can enhance performance. We therefore expect that AI, by facilitating rapid synthesis of large volumes of textual and numerical information, serves a similar role in reducing forecast errors and boosting stock market reaction.

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

3 Data and Measurement

3.1 Data

Our empirical analysis focuses on sell-side equity research reports that are directly linked to analysts' earnings forecasts. To construct this dataset, we combine analyst reports with forecast-level information and firm fundamentals from multiple sources, with the analyst report serving as the primary unit of analysis. We obtain analyst research reports from Investext, a large-scale research archive maintained by Mergent Online. Investext distributes reports in PDF format and provides rich document-level metadata, including the publishing institution (investment bank or brokerage house), contributing analyst(s),

document date, firm names and tickers discussed, report categories, subject tags, and report style classifications. These metadata fields allow us to identify analysts, brokers, and covered firms, and to link textual reports to forecast activity.

Our sample covers reports issued between January 1, 2021 and April 1, 2024, a period spanning both the pre- and post-ChatGPT environments. Starting from the full universe of reports available in Investext during this window, we restrict the sample to documents for which analyst identities, firm identifiers, and forecast matches can be reliably established.

We link analyst reports to earnings forecasts from the I/B/E/S database using analyst identity, firm ticker, and a narrow timing window around forecast activity. Following Huang et al. (2014) and Lv (2024), a report is matched to forecasts if it is issued within two calendar days before the forecast announcement date and two calendar days after the forecast revision date. The final analysis sample consists of 51,086 equity research reports linked to approximately 74,721 earnings forecasts, issued by 488 analysts affiliated with 75 brokerage houses. Because a single report may correspond to multiple earnings forecasts, the number of forecasts exceeds the number of unique reports. Table 1 summarizes the sample composition and its distribution over time.

Analyst reports are distributed in PDF format, which is unstructured and requires systematic preprocessing prior to textual analysis. We extract raw text from each PDF using pdfminer.six, a Python library for programmatic PDF text extraction, and process the reports using a custom text-cleaning pipeline. We preprocess analyst reports to remove formatting artifacts introduced by PDF extraction and retain only analyst-written narrative content. This includes correcting broken lines, excluding non-linguistic text such as tables

and numerical artifacts, and removing broker-dealer disclosures and regulatory boilerplate in their entirety. The resulting cleaned text forms the basis for all subsequent document-level analyses, including AI-usage detection and prompt-based information extraction. Appendix B provides additional details on the text extraction and cleaning procedures.

In addition to analyst reports and earnings forecasts, we merge several external datasets to construct firm- and analyst-level control variables. Firm financial characteristics are obtained from Compustat, while stock return data are drawn from CRSP. Analyst following and forecast history variables are sourced from the I/B/E/S database. To capture analyst background characteristics that are not available in standard forecast databases, we supplement these data with hand-collected information from publicly available professional profiles. Following prior work (e.g., Li et al., 2025), we manually compile information on analyst gender, educational attainment, and professional certification by reviewing analysts' online biographies and professional profiles, including LinkedIn and similar publicly accessible sources. Analyst gender is identified based on explicit indicators such as pronoun usage and name prefixes. Educational attainment is recorded based on the highest degree listed in the analyst's profile, and professional certification is identified based on whether the analyst holds the Chartered Financial Analyst (CFA) designation. We also obtain information on analysts' professional recognition from the annual Institutional Investor All-Star analyst rankings. Using the most recent All-Star lists available prior to 2022, we construct an indicator variable that equals one if an analyst has ever been recognized as an Institutional Investor All-Star before 2022, and zero otherwise. This measure captures analysts' ex ante reputation and is treated as time-invariant over the sample period.

3.2 AI Measurement

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

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

To construct training data for document-level detection, we draw analyst reports issued prior to the public release of ChatGPT, a period during which generative AI usage in sell-side research was negligible. For each report, we generate AI-assisted document variants by rewriting approximately 50 percent of the original analyst-written content using GPT-3.5-Turbo, while preserving the original analytical meaning, factual content, and professional writing style. The rewritten content is integrated back into the original reports to form partially AI-assisted documents, which are paired with fully human-authored reports as ground-truth data for supervised training and evaluation of document-level AI-

detection models. This 50 percent specification serves as our primary training design, while alternative proportions of AI-assisted content are examined in robustness analyses.

We compute 25 document-level linguistic features following Guo et al. (2023) and Uzun (2023) (Table B1). These features capture multiple dimensions of writing style and structure, including lexical richness, syntactic complexity, spelling and grammar accuracy, sentiment tone, readability, and fluency. Using these features, we train three supervised learning classifiers, Random Forest, Logistic Regression, and XGBoost, on 80 percent of the ground-truth data and evaluate their performance on the remaining 20 percent. As shown in Table B2, XGBoost consistently outperforms the alternative models in distinguishing human-authored versus AI-assisted analyst reports based on the document-level linguistic features. Accordingly, we use XGBoost as our primary model to construct the document-level AI usage measure in the pre- and post-GPT sample. Figure 1 reports the relative importance of linguistic features in the XGBoost model.

[Figure 1 here]

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

We acknowledge that our detection approach may primarily capture AI usage in surface-level editing or drafting of analyst reports rather than in deeper analytical tasks. Any such measurement error would tend to attenuate estimated effects, thereby biasing our

results against finding significance and making our tests conservative. At the same time, given the substantial efficiency gains from using generative AI, analysts who rely on AI for drafting are plausibly more likely to use it in other aspects of the research process. As a result, higher values of AI_Prob likely reflect broader integration of GenAI into analysts' research activities.

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

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

As a second main outcome, we examine the magnitude of the stock market reaction to analyst reports using |CAR|, which is defined as the absolute value of the cumulative two-day abnormal return around the report date (day 0 to +1), where abnormal returns are calculated using the market-adjusted model. This measure captures the strength of investors' responses to the information contained in analyst reports, regardless of whether the reaction is positive or negative.

Examining |CAR| is particularly informative in the context of generative AI. Analyst research reports convey substantially more information than point earnings forecasts, including narrative interpretation, valuation analysis, and the synthesis of firm-

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

We measure forecast timeliness using the forecast horizon in days. Horizon (day) is defined as the number of days between the time stamp of the analyst's forecast and the corresponding earnings announcement. Prior research shows that forecasts issued closer to the announcement typically incorporate more timely information and are associated with higher accuracy (Huang et al., 2022). In our setting, a shorter horizon may also reflect efficiency gains from generative AI, which can lower the time costs of forecast production. At the same time, issuing forecasts closer to the announcement may indicate a reallocation of analyst effort toward faster, more routine forecast generation.

4. Baseline Results

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

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

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

Control variables include analyst characteristics commonly used in the forecast accuracy literature: Male, Education, CFA, AllStar, Forecast Frequency, Forecast Age, Firms Covered, Industries Covered, and Analyst Experience (Li et al. 2025, Flam et al. 2023, Harford et al. 2019). These controls capture heterogeneity in analyst ability, workload, and information environment. Firm-level characteristics are not included separately, as they are absorbed by the firm–year fixed effects.

The unit of observation is at the analyst–firm–forecast level. The specification includes firm–year fixed effects and broker–year fixed effects which absorb time-varying firm fundamentals and brokerage-level factors that may jointly affect analysts' research practices and forecast outcomes. This fixed-effects structure closely follows Li et al. (2025) and allows us to benchmark analysts' forecasts within the same firm-year and brokerage-year environments. Standard errors are clustered at the analyst–year level to account for

serial correlation in analysts' forecast behavior over time (Li et al., 2025). The final estimation sample consists of 74,634 observations for forecast accuracy regressions and 74,721 observations for market reaction regressions.

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

Table 3 reports the results from regressions examining the association between AI usage in analyst reports and forecast accuracy. Column (1) shows that the coefficient on AI_Prob is positive and statistically significant, indicating that analyst reports with a higher probability of incorporating generative AI are associated with greater forecast accuracy. Using the sample mean of Accuracy as a benchmark, the point estimate implies a meaningful improvement in relative forecast accuracy as AI usage increases.

These results are obtained after controlling for a rich set of analyst characteristics, including Male, Education, CFA, AllStar, ForecastFrequency, ForecastAge, FirmsCovered, IndCovered, and AnalystExperience. We also control the firm–year and broker–year fixed effects to account for time-varying firm and broker characteristics, such as brokers' report templates variations. Standard errors are clustered at the analyst–year level. Overall, the evidence in Table 3 suggests a robust positive association between AI usage in analyst reports and forecast accuracy. In Appendix Table C3, we find that our results stand when controlling analyst fixed effects. Though the statistical significance is lower, this result suggests that the relation between AI usage in reports and forecast accuracy is not driven by analysts' time-invariant characteristics (such as analysts' risk preferences).

[Table 3 here]

5. Identification

A critical concern is that our AI detection measure may conflate true AI usage with underlying report informativeness or analyst ability. It is possible that our detection approach captures reports that are inherently more informative or better written, rather than genuine reliance on generative AI by analysts, even though our model is based solely on linguistic features. If so, this would give rise to a classical endogeneity concern: more informative reports may be both more accurate and more likely to elicit stronger market reactions, independent of AI usage. More skilled analysts may write clearer and more comprehensive reports, while also being more inclined to adopt AI tools, generating potential confounding bias. Similarly, analysts may deploy AI selectively for reports they expect to be particularly important, valuable, or complex, creating selection concerns. Finally, analysts with higher underlying ability or better performance may be more likely to adopt AI, raising the possibility of reverse causality.

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

Conference calls have become a common and important disclosure channel for public firms since the passage of Regulation Fair Disclosure (Reg FD) in 2000. Prior studies show that earnings conference calls convey price-related information beyond press releases (Bowen et al., 2002; Bushee et al., 2004; Kimbrough, 2005). Accordingly,

earnings conference calls are information hotspots for analysts and investors to acquire information. Driskill et al. (2020) show that analysts are subject to limited attention when multiple firms they cover report earnings on the same day. De Franco et al. (2025) further show that overlapping calls serve as attention shocks to analysts. More importantly, they find no clear evidence that managers strategically adjust call timing to avoid overlaps, and overlap is mainly driven by pre-determined factors such as headquarter time zone and fiscal calendar. Their result suggests that overlapping calls are exogenous shocks to analysts' attention and workload. Since most analysts release reports soon after the earnings calls, overlapping calls would create an exogenous time pressure for analysts to adopt AI in their analyst reports so that they can release their reports on time.

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

Following De Franco (2022), we define overlapping calls as multiple conference calls occurring on the same date for firms covered by the same analyst. We construct the

report-level overlapping dummy variable, *Overlapping Dummy*, which takes the value of one if a report is released on a date when at least two firms under the analyst's coverage hold earnings calls, and otherwise zero. Using reports released on the next day following multiple earnings calls, we create an additional variable, *Overlapping Dummy ([-1])*, which takes the value of one if the report is released one day after overlapping calls of firms that the analyst covers, and zero otherwise. We will use this as an additional instrumental variable if necessary. Our analyst report sample shows that 27% of reports are released on the same date as earnings calls, and 24% of reports are released on the following date of earnings calls, and the distribution of report release dates is very diverse on other subsequent dates.

[Table 4 here]

We next use overlapping earnings calls as an instrumental variable for AI usage in analyst reports and reexamine the effect of AI usage on forecast accuracy. We follow the same specification as that in Table 3, except that we estimate the model using two-stage least squares (2SLS). Panel A of Table 4 presents IV regressions using overlapping earnings calls occurring on the same date as the instrument. Column (1) reports the first-stage results, where the dependent variable is AI_Prob, a document-level measure capturing the probability that an analyst report incorporates generative AI. The overlapping dummy has a positive and statistically significant effect on AI_Prob, indicating that AI usage is higher on report release dates that coincide with overlapping earnings calls for firms covered by the analyst. The results of the F-test (of excluded instruments), underidentification test (cluster-robust Kleibergen-Paap rk LM statistic), and 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.

Panel B extends the analysis by using overlapping earnings calls on the same date and the previous date (a report is released one day after the earnings call date) as instruments, which allows us to conduct an overidentification test of the exclusion restriction. The estimated coefficients remain positive and statistically significant. We fail to reject the Hansen J overidentification test ($p = 0.136$), suggesting that the instruments are jointly valid and uncorrelated with the second-stage error term. This evidence supports the validity of the exclusion restriction. The 2SLS estimates remain consistent with those reported in Panel A.

6. Mechanisms and Heterogeneity in the Effects of AI

6.1 Mechanism: How AI Empowers Investment Research?

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

nuanced thinking. We examine several dimensions of analysts' research quality well-documented in the literature.

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

We then examine tone in analyst report. Extensive literature has shown that analysts tend to write favorable reports to generate trading commissions from their clients and business opportunities with the firms they cover. Huang, Zang, and Zheng (2014) find that investors react more strongly to negative text than to positive text in analyst reports. Generative AI may help analysts tone down their optimism, as it stands for an average investor and tends to offer more balanced views. We ask GPT-5.1 to classify tone in an analyst report. *Tone*, takes the value of 1 if a report is negative or pessimistic, 2 if a report is neutral, mixed, or balanced, and 3 if a report is positive or optimistic.

Brown et al. (2015) use a survey to show that private communication with management is a more useful input to analysts' forecasts than their own research and public information, such as filings. Bradley, Jame, and Williams (2022) show that non-deal roadshows play an important role in analysts' biased forecasts and create information advantage for institutional investors. Bradley, Jame, and Williams (2022) further show that

hosting analysts issue favorable forecasts proceeding site visits to firms' headquarters and plants. Taking all the documented information sources used by analysts, we ask GPT-5.1 to classify the numbers of private and public meetings with the management separately. Private communications include phone/videos/emails with the management, site visits such as bus tours or field trips, non-deal roadshows, surveys, channels check, fireside chats, and social events such as golf outings. Public communications include earnings calls, company conference presentations, special calls, analyst/investor days, M&A calls, shareholder/analyst calls, annual shareholder meetings, guidance/update calls, and sales/trading statement calls. To avoid counting duplications, we count the number of distinct sources only.

Chi, Hwang, and Zheng (2025) classify the alternative data used in analyst reports into eight categories and show that reports with higher big data usage have more accurate forecasts. These eight categories include app usage, earnings calls/employee-related data, geospatial data, point-of-sale data, satellite images, consumer sentiment, web traffic, and others. We ask GPT-5.1 to classify the numbers of distinct categories of big data used in a report accordingly.

Sheng et al. (2025) use 14 topics that ChatGPT can generate forward-looking information/signal based on earnings calls and they find these signals can predict hedge fund flows. These 14 topics include US economy, global economy, financial prospect of the firm, financial prospect of the industry, earnings, revenue, wages/salaries, etc. We ask GPT-5.1 to classify the number of distinct topics (up to 14) mentioned in a report accordingly.

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

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

For the count-based measures above, we scale them with the total number of words in a report to measure the density of each category in a report. Alternatively, we use the total number of sentences or pages in a report to scale these measures and get consistent results.

We examine how AI usage in reports relates to these valuation-related research depth measures. Panel A of Table 5 presents the results. We use the same specification as previous tables, except that we include one more control, the sentence count in a report as it is used as the scaler. The results show that AI usage in a report is positively associated with the number of valuation models, the number of distinct topics, the number of distinct stakeholders, the word ratio, and the extent of public communication discussed in the report. These suggest that generative AI is associated with deeper and more comprehensive analytical content in analyst research.

[Table 5 here]

We next examine whether AI usage in analyst reports is associated with differences in analysts' reasoning patterns. Panel B of Table 5 reports these results. To capture reasoning styles, we use LIWC-22 to construct linguistically based measures that have been widely adopted in fields such as politics and psychology (Boyd et al., 2022; Li et al. 2025). These measures are designed to reflect how individuals process and communicate information, rather than general writing style, and are not used in training our machine learning detection model.

Specifically, we consider four dimensions of reasoning. *Analytical* measures the extent of logical and formal reasoning.[2] *Authentic* captures the extent of sincere and straightforward communication. These two measures are based on a proprietary formula that summarizes several categories of function words. *Causation* measures the use of words that explains relationships between events or outcomes (e.g., because, therefore, make, and why), reflecting explanation-oriented reasoning. *Differentiation* captures the fraction of words related to conditions, contrasts, and alternative possibilities (e.g., but, not, if, and or), reflecting more nuanced reasoning.

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

6.2 Moderators of AI's Effect on Forecast Accuracy

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

Table 6 examines whether the association between AI usage in analyst reports and forecast accuracy varies systematically across analyst characteristics. We interact our primary AI usage measure, AI_Prob, with several analyst attributes that capture differences in workload and experience, including FirmsCovered, ForecastFrequency, Male, and AnalystExperience.

The results indicate that the accuracy benefits of AI usage are not uniform across analysts. Column (1) shows that the interaction between AI_Prob and FirmsCovered is positive and statistically significant, suggesting that analysts who cover a larger number of firms experience greater accuracy improvements from AI usage. Column (2) reports a positive and significant interaction between AI_Prob and ForecastFrequency, indicating that AI assistance is particularly beneficial for analysts who issue forecasts more frequently. These findings are consistent with AI tools helping analysts manage higher workloads and time constraints.

Columns (3) and (4) explore heterogeneity along analyst demographic and experience dimensions. Column (3) shows that the interaction between AI_Prob and Male is negative and statistically significant, suggesting that the association between AI usage and forecast accuracy differs across analyst gender. One interpretation is that, holding AI access constant, female analysts may benefit more from AI assistance, potentially because generative AI helps mitigate constraints related to time pressure, information processing, or communication style. Column (4) reports a positive and significant interaction between AI_Prob and AnalystExperience, indicating that more experienced analysts are better able to translate AI usage into forecast accuracy improvements. This pattern is consistent with the view that while generative AI tools are broadly accessible, their effectiveness depends

on users' accumulated human capital, with more experienced analysts better positioned to integrate AI outputs into high-quality forecasts.

[Table 6 here]

7. Additional Tests

7.1 Capital Market Reaction

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

Following Huang et al. (2014), we measure investor reactions using the absolute value of cumulative abnormal returns (|CAR|) around the report date, which captures the magnitude of the short-window stock price response to analyst research. |CAR| is defined as the two-day market-adjusted abnormal return beginning on the report date. Focusing on |CAR| allows us to assess whether AI-assisted reports convey information that investors perceive as more salient or impactful, regardless of the direction of the price response.

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

As shown in Table 7, the coefficient on AI_Prob is positive and statistically significant, suggesting that analyst reports with a higher probability of incorporating generative AI are associated with stronger immediate market reactions. This evidence indicates that AI-assisted reports convey information that investors perceive as more salient or impactful. Consistent with our interpretation, generative AI may influence how information is structured, framed, and communicated in analyst reports, thereby amplifying the intensity of investor responses rather than their sign.

[Table 7 here]

7.2 Determinant Model

To address potential selection concerns regarding the use of generative AI in analyst reports, we estimate a determinant model to examine which analyst- and firm-level characteristics are associated with a higher likelihood of AI-assisted report writing. The dependent variable is AI_Prob, a continuous, document-level measure capturing the probability that an analyst report incorporates generative AI.

The specifications in determinant analysis (Table C1) are constructed in a stepwise manner following the determinant frameworks in prior studies. Column (1) includes a baseline set of analyst characteristics consistent with Table 3, and follows the fixed effects and clustering approach of Flam et al. (2023). Column (2) augments the baseline specification by incorporating additional analyst and brokerage characteristics motivated by Harford et al. (2019), while Columns (3) and (4) further expand the model by adding firm-level and industry-related determinants following Peng et al. (2022) and Huang et al. (2022), respectively. This stepwise design allows us to assess the incremental explanatory

power of different classes of determinants while maintaining a consistent estimation framework across specifications.

Across the specifications reported in Table C1, several analyst characteristics consistently explain variation in AI usage. In particular, male analysts and analysts with longer overall experience exhibit a higher likelihood of using AI, while All-Star analysts and those holding a CFA designation are less likely to rely on generative AI. The positive association between analyst experience and AI usage suggests that more experienced analysts may be better positioned to integrate AI tools into their research workflow. In contrast, the negative association with All-Star status and CFA certification is consistent with highly credentialed analysts relying more on established expertise and professional judgment rather than AI assistance. One possible explanation for the lower propensity of All-Star analysts and CFA charterholders to use generative AI relates to reputational and professional risk considerations. Analysts with stronger reputational capital may face higher expected costs from potential misperceptions about AI-assisted research, including concerns about originality, authorship, or professional judgment. Similarly, CFA charterholders operate under a strict professional code emphasizing independent analysis and accountability, which may lead them to adopt a more conservative approach toward emerging tools whose regulatory and compliance implications are still evolving. As a result, even when access to generative AI is broadly available, analysts with higher reputational or professional stakes may optimally limit their reliance on such tools. Firm- and brokerage-level characteristics also contribute to explaining variation in AI usage. Analysts affiliated with top brokerage houses are significantly more likely to incorporate AI into

their reports, potentially reflecting greater access to technological resources and institutional support.

7.3 Forecast Timeliness

In this section, we present additional analyses that further evaluate the implications of generative AI usage in analyst reports and assess the robustness of our main findings. Table 8 examines the association between AI usage and forecast timeliness. The dependent variable, Horizon_Day, measures the number of days between the issuance of an analyst forecast and the corresponding earnings announcement, with smaller values indicating timelier forecasts. Column (1) reports a negative and statistically significant coefficient on AI_Prob, suggesting that analyst reports with higher AI usage are associated with forecasts issued closer to earnings announcements. This finding is consistent with generative AI facilitating more efficient information processing and report preparation, allowing analysts to update forecasts in a timelier manner. Among the control variables, forecast timeliness varies systematically with analyst characteristics. More experienced analysts and those covering a larger number of firms tend to issue forecasts closer to earnings announcements, while analysts with higher forecast frequency issue forecasts further in advance. These patterns are consistent with prior evidence on analyst workload and information acquisition constraints and indicate that the relation between AI usage and timeliness is incremental to well-established determinants of forecast horizons.

[Table 8 here]

7.4 Validation: Bank ChatGPT Bans

Due to compliance risk and data leakage concerns, several banks have been widely reported in the news for restricting the use of ChatGPT starting from late February 2023.

These banks include JPMorgan, Bank of America, Citi, Deutsche Bank, Goldman Sachs, and Wells Fargo. For example, JPMorgan was reported to have restricted employees' use of ChatGPT on February 22, 2023, by CNN, Bloomberg, Forbes, Wall Street Journal, among others. We conduct a comprehensive search of all brokerage houses in our sample online and find that three big banks implement ChatGPT bans, including JPMorgan, Deutsche Bank, and Wells Fargo. It is unclear when these restrictions are lifted, but *Fortune* provides evidence that these three banks still implemented the policy as of May 19, 2023. For example, JPMorgan stated in its annual report released on April 3, 2023 that "We're imagining new ways to augment and empower employees with AI through human-centered collaborative tools and workflow, leveraging tools like large language models, including ChatGPT." JPMorgan Chase is giving its over 60,000 employees (including the asset & wealth management division) the LLM Suite, an AI assistant powered by ChatGPT maker OpenAI, reported by CNBC on 8/9/2024. This initial version provided a secure, internal "ChatGPT-like" tool for tasks such as drafting emails, summarizing documents, and generating ideas. We thus take these three brokers as our treatment group.

Bank-level ChatGPT bans offer a powerful source of exogenous variation in analysts' access to generative AI that directly addresses the major endogeneity concerns. These bans were imposed for firm-wide IT, cybersecurity, and compliance reasons rather than analyst performance, writing skill, or forecasting ability. As a result, they break the link between an analyst's unobserved ability and their likelihood of using AI: highly skilled analysts did not "self-select" into AI usage, nor were the bans triggered by changes in report quality or forecast accuracy. Furthermore, because analysts cannot influence whether or when their employer restricts AI access, reverse causality, whereby more

accurate or more informative analysts choose to use AI, cannot explain our findings. The ban events operate as plausibly exogenous supply-side shocks that remove or restore AI access independently of analysts' reporting behavior, allowing us to isolate the effect of AI usage on forecast quality rather than the reverse.

Figure 2 plots the time-series evolution of the mean document-level AI usage probability (AI_Prob) for analyst reports issued by brokerage houses with and without internal ChatGPT bans. Prior to the public release of ChatGPT in late 2022, AI usage follows similar trends across the two groups. After ChatGPT becomes publicly available, reports issued by non-banned brokerage houses exhibit a sharp increase in AI usage, while reports issued by banned brokerage houses show a substantially slower adoption pattern.

The divergence in AI usage between the two groups is most pronounced in the months immediately following the introduction of ChatGPT, consistent with the timing of internal restrictions on generative AI tools at several large banks. Over time, the gap narrows, potentially reflecting the introduction of alternative generative AI tools, changes in enforcement intensity, or the rollout of internal AI systems by brokerage houses.

[Figure 2 here]

We then examine whether reports provided by our treatment group during their ChatGPT ban period exhibit lower AI usage, compared to brokers who do not explicitly mention restrictions on ChatGPT. Our implicit assumption is that the three treatment brokers are large banks and thus have strong enforcement mechanisms to restrict or block employees' access to ChatGPT from company networks or devices. In contrast, brokers in the control group lack comparable enforcement to prevent employees from using ChatGPT, despite having general policies regarding AI usage.

We create two indicator variables. Banned_Bank is an indicator that takes the value of one if a report is released by one of the three treatment banks, and zero otherwise. Banned_Period is an indicator that takes the value of one if a report is released between December 1, 2022, and March 31, 2023, and zero otherwise. For this test, we restrict the sample to reports issued after December 1, 2022, the public release date of ChatGPT-3.5. We use the interaction between these two indicators to proxy for the period in which banks' ChatGPT bans are in effect.

We then implement a difference-in-differences regression as a supplementary validation exercise, following the same specification as in Table 3, with AI_Prob as the dependent variable. Table 9 presents the results. Column (1) reports the baseline difference-in-differences estimate, showing that the interaction term has a significantly negative effect on AI_Prob, while Banned_Period has a significantly positive effect on AI_Prob. Column (2) adds analyst-level control variables, and Column (3) further includes broker fixed effects; the estimated effects remain negative and statistically significant across specifications.

[Table 9 here]

Taken together, these results show that firmwide ChatGPT bans lead to a significant reduction in AI usage in analyst reports, supporting the interpretation that AI usage responds to exogenous restrictions on access to generative AI tools.

7.5 Robustness to Alternative AI Detection Models

Table C2 assesses the robustness of the relation between AI usage and forecast accuracy to alternative AI detection models and benchmark definitions. We re-estimate the baseline accuracy specification using AI usage probabilities generated from different

machine learning classifiers and under more stringent definitions of AI involvement. Column (1) replaces the baseline AI usage measure with AI_ProbHigh, which captures higher-intensity AI involvement based on a stricter training benchmark. The coefficient on AI_ProbHigh remains positive and statistically significant. Columns (2) and (3) further construct AI usage probabilities using a Random Forest classifier, while Columns (4) and (5) rely on a Logistic Regression classifier. Across all specifications, both continuous and high-involvement AI usage measures load positively and significantly, with magnitudes comparable to those in the baseline results. These additional analyses reinforce the interpretation that generative AI serves as a productivity-enhancing tool that improves both the efficiency and effectiveness of analysts' forecasting activities rather than being an artifact of a particular measurement approach.

8. Conclusion

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

We establish support for a causal interpretation using exogenous shocks from overlapping earnings calls. We find that the effects of AI usage in reports on forecast accuracy are concentrated in dates with overlapping earnings calls from firms under an analyst's coverage. Our channel analysis suggests that GenAI empowers analysts'

investment research partially through richer valuation analysis and broader informational coverage, including the integration of more valuation models and diverse analytical dimensions. Our cross-sectional analyses show that efficiency gains from AI are more pronounced when reports are produced by more experienced analysts or by analysts facing higher workload demands, while the benefits vary across analyst characteristics.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bertomeu, J., Lin, Y., Liu, Y., & Ni, Z. (2025). The impact of generative AI on information processing: Evidence from the ban of ChatGPT in Italy. *Journal of Accounting and Economics*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

IBM. (2024). What is shadow AI?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

OpenAI. (2023). GPT-4.

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

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

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

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

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

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

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

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

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

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

Figure 1. Important Features Considered by XGBoost

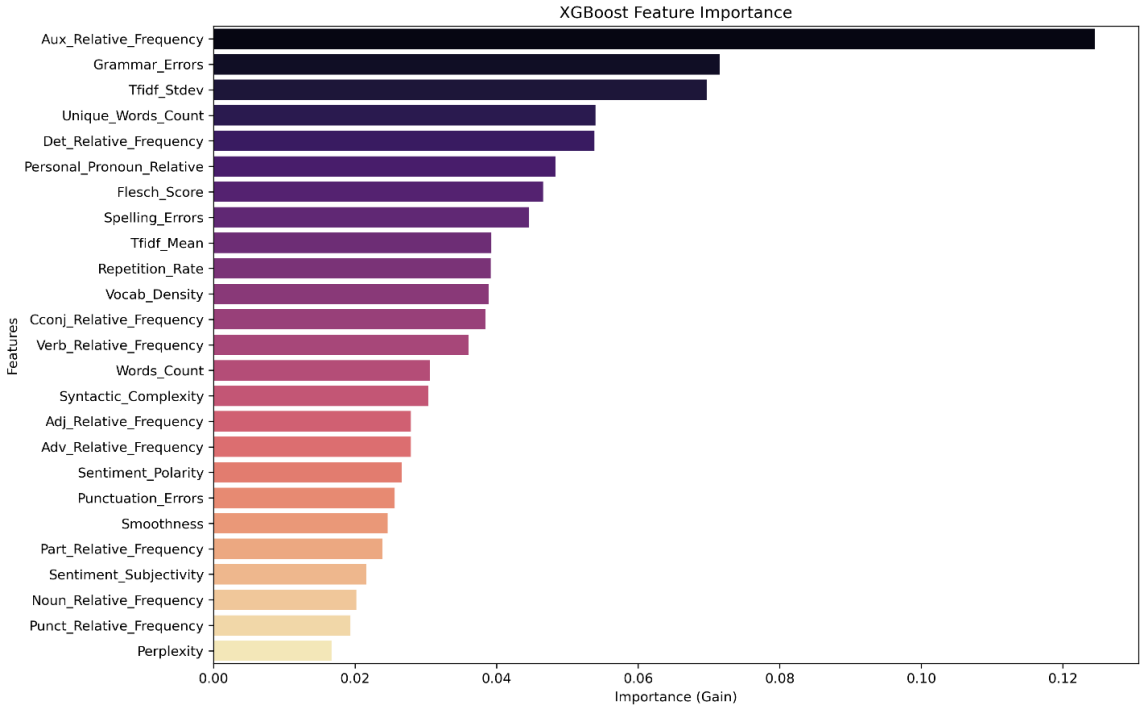
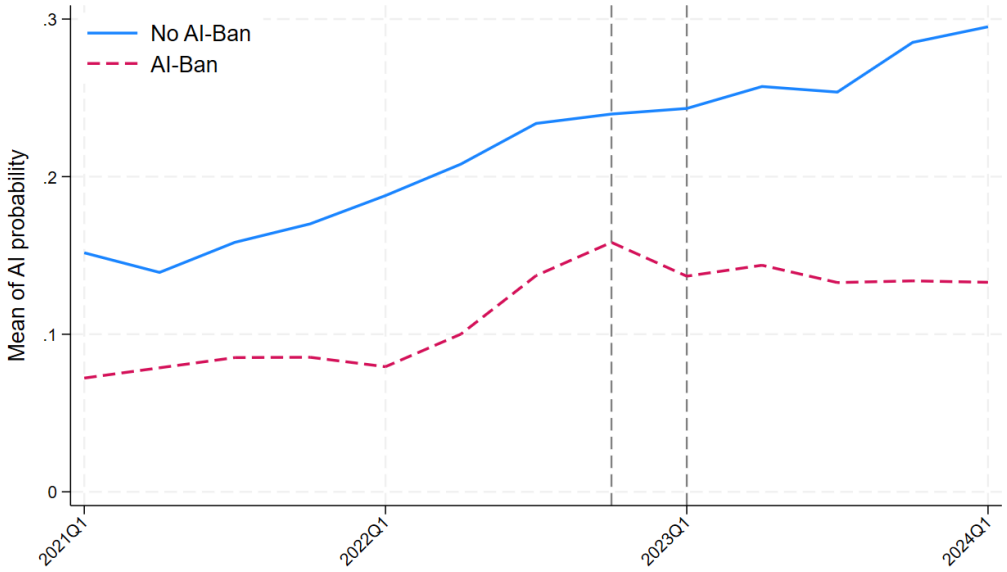


Figure 2. Trends of AI Usage in Reports



Notes: This figure plots the time-series of the mean document-level AI usage probability (AI_Prob) in analyst reports over time. The solid line represents reports issued by brokerage houses without internal bans on generative AI usage ("No AI-Ban"), while the dashed line represents reports issued by brokerage houses that implemented internal bans on ChatGPT usage ("AI-Ban"). The first vertical dashed lines mark key dates surrounding the public release of ChatGPT on November 30, 2022. The second vertical dashed lines mark the AI ban period. The figure illustrates differences in the evolution of AI usage across brokerage houses with and without AI usage bans.

Table 1. Sample Description

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

	Total	Unique
Analyst	74,721	488
Broker	74,721	75
Firm	74,721	2,123
Report	74,721	51,086

Note: 74,721 unique forecasts in this study. One report may contain multiple forecasts.

Panel B. Sample Distribution by Year

Year	Freq.	Percent
2021	18,418	24.65
2022	25,424	34.03
2023	26,941	36.06
2024	3,938	5.27
Total	74,721	100.00

Panel C. Sample Comparison with IBES

	IBES (2021-2024)						Our Sample					
	N	Mean	P50	SD	P25	P75	N	Mean	P50	SD	P25	P75
BrokerSize	378	1.569	1.386	0.948	0.693	2.303	225	1.943	1.792	0.980	1.099	2.708
FirmsCovered	2,660	2.321	2.565	0.866	1.609	2.996	1,347	2.603	2.773	0.698	2.197	3.135
IndCovered	2,660	1.219	1.099	0.528	0.693	1.609	1,347	1.314	1.386	0.539	0.693	1.792
ForecastFrequency	2,660	4.133	4.394	1.562	2.944	5.476	1,347	4.698	5.050	1.184	3.689	5.673

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

Table 2. Summary Statistics and Correlations

Panel A. Descriptive Statistics

Notes: This table reports summary statistics for the main variables used in the analysis. Panel A presents the mean, median (P50), standard deviation (SD), and interquartile range (P25, P75) 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. Controls include Male, Education, CFA, AllStar, ForecastFrequency, ForecastAge, FirmsCovered, IndCovered, and AnalystExperience. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A.

	All (N=74,721)				
	Mean	SD	P25	P50	P75
AI_Prob	0.176	0.284	0.003	0.023	0.212
Accuracy	0.158	0.568	-0.147	0.209	0.587
Male	0.943	0.232	1.000	1.000	1.000
Edu	1.316	0.963	0.000	2.000	2.000
CFA	0.381	0.486	0.000	0.000	1.000
AllStar	0.358	0.479	0.000	0.000	1.000
ForecastFrequency	5.525	0.761	5.268	5.684	5.986
ForecastAge	4.662	0.761	4.454	4.543	5.204
FirmsCovered	2.992	0.476	2.773	3.045	3.296
IndCovered	1.443	0.543	1.099	1.386	1.946
AnalystExperience	2.556	0.740	2.023	2.783	3.120

Panel B. Pearson Correlation Matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) AI_Prob	1.00										
(2) Accuracy	0.01*	1.00									
(3) Male	0.06***	0.02***	1.00								
(4) Edu	0.00	0.00	0.11***	1.00							
(5) CFA	-0.03***	0.02***	0.04***	-0.08***	1.00						
(6) AllStar	-0.03***	0.03***	0.12***	0.01***	0.16***	1.00					
(7) ForecastFrequency	-0.01***	-0.02***	0.08***	0.14***	0.08***	0.28***	1.00				
(8) ForecastAge	0.02*	-0.17*	0.01*	0.05*	-0.02*	-0.04*	0.09*	1.00			
(9) FirmsCovered	0.00	-0.02***	0.12***	0.23***	0.09***	0.26***	0.73***	0.04***	1.00		
(10) IndCovered	-0.01***	0.00	0.08***	-0.01***	0.05***	-0.12***	0.16***	0.00	0.14***	1.00	
(11) AnalystExperience	0.01***	0.02***	0.13***	0.30***	0.02***	0.28***	0.19***	0.00	0.23***	0.20***	1.00

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

Table 3. AI Usage and Forecast Accuracy

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

	(1) Accuracy	(2) Accuracy
AI_Prob	0.031***	0.023**
	(0.010)	(0.010)
Male	0.020	-0.022
	(0.027)	(0.022)
Edu	0.002	0.000
	(0.006)	(0.006)
CFA	0.042***	0.035***
	(0.011)	(0.011)
AllStar	0.041***	0.059***
	(0.012)	(0.016)
ForecastFrequency	-0.039***	-0.049***
	(0.011)	(0.011)
ForecastAge	-0.126***	-0.125***
	(0.007)	(0.007)
FirmsCovered	-0.015	-0.018
	(0.022)	(0.023)
IndCovered	0.037**	0.044**
	(0.016)	(0.017)
AnalystExperience	0.002	0.006
	(0.007)	(0.008)
Firm-Year FE	Yes	Yes
Broker-Year FE	No	Yes
Observations	74,634	74,634
R-squared	0.280	0.286
Adj R-squared	0.215	0.220

Table 4: Exogenous Shocks from Overlapping Earnings Calls

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. Panel B presents IV regressions using overlapping earnings calls on the same date and the previous date as instruments. The specification and interpretation parallel those in Panel A, with multiple instruments used to capture exogenous variation in analysts' reliance on AI induced by overlapping earnings-call schedules. All specifications include analyst-level control variables as well as firm–year fixed effects and broker–year fixed effects, as indicated. Standard errors are clustered at the the analyst–year level. All continuous variables are winsorized at the 1st and 99th percentiles. The table reports first-stage F-tests as well as underidentification, weak IV, and overidentification tests where applicable. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Panel A: IV regression using overlapping calls on the same date as the IV

	(1) First-stage AI_Prob	(2) Second-stage Accuracy
Overlapping Dummy	0.018***	
	(0.006)	
AI_Prob		3.048***
		(0.004)
Analyst controls	Yes	Yes
Firm-Year FE	Yes	Yes
Broker-Year FE	Yes	Yes
Observations	74,634	74,634
F-test (P-value)	0.002	
Underidentification test (P-value)	0.001	
Weak IV test (P-value)	0.000	

Panel B: IV regression using overlapping calls on the same date and the previous date as the IVs

	(1) First-stage AI_Prob	(2) Second-stage Accuracy
Overlapping Dummy([0])	0.024***	
	(0.006)	
Overlapping Dummy ([-1])	-0.008**	
	(0.004)	
AI_Prob		1.980***
		(0.736)
Analyst/Broker controls	Yes	Yes
Firm-Year FE	Yes	Yes
Broker-Year FE	Yes	Yes
Observations	74,634	74,634
F-test (P-value)	0.001	
Underidentification test (P-value)	0.000	
Weak IV test (P-value)	0.000	
Overidentification test (P-value)	0.000	

Table 5: AI usage and analysts' analytical research quality

Notes: This table examines AI usage in analyst reports and analysts' analytical research quality. The key independent variable is AI_Prob. Panel A 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. Word_ratio is the relative ratio of word counts over the sum of word and numeric counts 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 mentioned in a report. Big_data is the number of distinct sources of big or alternative data mentioned in a report. Panel B presents the relation between AI usage and analytical language metrics in reports. We use LIWC-22 software to measure analytical metrics. Analytical is the fraction of words that reflects 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, reflecting 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 fixed effects and broker–year fixed effects, as indicated. Standard errors are clustered at the analyst–year level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Panel A: the relation between AI usage and valuation-related metrics in reports

	(1) Valuation	(2) Topics	(3) Stakeholders	(4) Word_ratio	(6) Private	(7) Public	(8) Big_data
AI_Prob	0.136***	0.910***	0.530***	0.012***	-0.014	0.079***	-0.011
	(0.045)	(0.149)	(0.127)	(0.002)	(0.017)	(0.025)	(0.018)
Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Broker-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	73,890	73,890	73,890	74,721	73,890	73,890	73,890
R-squared	0.702	0.563	0.525	0.510	0.268	0.340	0.265
Adj R-squared	0.674	0.522	0.480	0.464	0.199	0.278	0.196

Panel B: AI usage and analysts' analytical language in reports

	(1) Analytical	(2) Authentic	(3) Causation	(4) Differentiation
AI_Prob	1.053***	1.259***	0.020**	0.112***
	(0.058)	(0.295)	(0.010)	(0.038)
Control	Yes	Yes	Yes	Yes
Firm-Year FE	Yes	Yes	Yes	Yes
Broker-Year FE	Yes	Yes	Yes	Yes
Observations	74,719	74,715	74,721	74,721

R-squared	0.702	0.563	0.525	0.510
Adj R-squared	0.574	0.718	0.725	0.687

Table 6 Cross-Sectional Variation in the Effect of AI Usage on Forecast Accuracy

Notes: This table reports OLS regression results examining cross-sectional variation in the association between AI usage and forecast accuracy. The dependent variable is Accuracy. The key independent variables are interaction terms between AI_Prob analyst characteristics, including FirmsCovered, ForecastFrequency, Male, and AnalystExperience. These interaction terms capture how the association between AI usage and forecast accuracy varies across different analyst characteristics. All specifications include the same set of analyst-level control variables as in Table 3, as well as firm–year fixed effects and broker–year fixed effects, as indicated. Standard errors are clustered at the analyst–year level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy	(3) Accuracy	(4) Accuracy
AI_Prob_FirmsCovered	0.070***			
	(0.020)			
AI_Prob_ForecastFrequency		0.039***		
		(0.012)		
AI_Prob_Male			-0.083*	
			(0.042)	
AI_Prob_AnalystExperience				0.024*
				(0.013)
AI_Prob	-0.187***	-0.194***	0.103**	-0.038
	(0.062)	(0.065)	(0.042)	(0.032)
FirmsCovered	-0.030	-0.016	-0.017	-0.018
	(0.022)	(0.023)	(0.023)	(0.023)
Male	-0.021	-0.022	-0.011	-0.020
	(0.022)	(0.022)	(0.023)	(0.022)
Edu	0.000	0.001	-0.000	-0.000
	(0.006)	(0.006)	(0.006)	(0.006)
CFA	0.035***	0.035***	0.035***	0.035***
	(0.011)	(0.011)	(0.011)	(0.011)
AllStar	0.058***	0.059***	0.058***	0.060***
	(0.016)	(0.016)	(0.016)	(0.016)
ForecastFrequency	-0.049***	-0.058***	-0.049***	-0.049***
	(0.011)	(0.011)	(0.011)	(0.011)
ForecastAge	-0.125***	-0.125***	-0.125***	-0.125***
	(0.007)	(0.007)	(0.007)	(0.007)
IndCovered	0.046***	0.045***	0.044**	0.044**
	(0.017)	(0.017)	(0.017)	(0.017)
AnalystExperience	0.006	0.006	0.006	0.001
	(0.008)	(0.008)	(0.008)	(0.009)
Firm-Year FE	Yes	Yes	Yes	Yes
Broker-Year FE	Yes	Yes	Yes	Yes
Observations	74,634	74,634	74,634	74,634
R-squared	0.286	0.286	0.286	0.286
Adj R-squared	0.220	0.220	0.220	0.220

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 two-day cumulative abnormal return around the report date, calculated using the market-adjusted model. All specifications include the same set of control variables as in Table 3. Column (1) includes firm–year fixed effects, while Column (2) additionally includes broker–year fixed effects, as indicated. Standard errors are clustered at the analyst–year level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) |CAR|	(2) |CAR|
AI_Prob	0.004***	0.003***
	(0.001)	(0.001)
Male	0.007***	0.005***
	(0.002)	(0.002)
Edu	0.000	-0.000
	(0.000)	(0.000)
CFA	0.000	0.000
	(0.001)	(0.001)
AllStar	-0.001	-0.001
	(0.001)	(0.001)
ForecastFrequency	-0.001	-0.002*
	(0.001)	(0.001)
ForecastAge	0.003***	0.003***
	(0.000)	(0.000)
FirmsCovered	0.001	0.002
	(0.002)	(0.002)
IndCovered	-0.000	0.000
	(0.001)	(0.001)
AnalystExperience	0.002***	0.003***
	(0.001)	(0.001)
Firm-Year FE	Yes	Yes
Broker-Year FE	No	Yes
Observations	74,721	74,721
R-squared	0.426	0.431
Adj R-squared	0.375	0.378

Table 8. Horizon Test: AI Usage and Forecast Timeliness

Notes: This table reports OLS regression results examining the association between AI usage and forecast timeliness. The dependent variable is Horizon_Day, defined as the number of days between the time stamp of an analyst's forecast and the corresponding earnings announcement. The key independent variable is AI_Prob. Control variables include the same set of analyst characteristics as in Table 3. All specifications include firm–year fixed effects and broker–year fixed effects, as indicated. Standard errors are clustered at the analyst–year level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Horizon_Day	(2) Horizon_Day
AI_Prob	-5.611***	-3.308**
	(1.657)	(1.640)
Male	8.181*	13.348***
	(4.596)	(4.740)
Edu	2.522**	1.882*
	(0.981)	(1.106)
CFA	-2.490	-3.720**
	(1.813)	(1.865)
AllStar	-7.999***	-1.157
	(2.246)	(2.611)
ForecastFrequency	55.984***	58.251***
	(2.156)	(2.196)
FirmsCovered	-59.005***	-60.049***
	(4.385)	(4.934)
IndCovered	-7.770**	-7.747**
	(3.247)	(3.664)
AnalystExperience	-6.674***	-11.159***
	(1.355)	(1.571)
Firm-Year FE	Yes	Yes
Broker-Year FE	No	Yes
Observations	74,721	74,721
R-squared	0.283	0.292
Adj R-squared	0.219	0.226

Table 9. Validation: Exogenous shock using banks' ChatGPT ban

Notes: This table reports OLS regression results examining the effect of banks' ChatGPT usage bans on AI usage in analyst reports. The dependent variable is AI_Prob. The key independent variable is Banned_Bank_PostGPT, an indicator that equals one for analyst reports issued by banks that implemented internal bans on ChatGPT usage during the post-ChatGPT period, and zero otherwise. Banned_Bank is an indicator for banks that imposed such bans, and GPT_Period is an indicator for the post-ChatGPT period. Column (1) includes firm–year fixed effects only. Column (2) adds analyst-level control variables. Column (3) further includes broker fixed effects to control for time-invariant differences across brokerage houses. Across all specifications, firm–year fixed effects are included to absorb time-varying firm characteristics. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)
	AI_Prob	AI_Prob	AI_Prob
Banned_Bank _PostGPT	-0.039**	-0.034*	-0.032***
	(0.019)	(0.018)	(0.012)
Banned_Bank	-0.093***	-0.117***	
	(0.015)	(0.015)	
GPT_Period	0.031***	0.033***	0.212**
	(0.010)	(0.010)	(0.008)
Analyst controls	No	Yes	Yes
Firm-year fixed effects	Yes	Yes	Yes
Broker fixed effects	No	No	Yes
Observations	74,721	74,721	74,721
R2	0.274	0.284	0.375

Appendix. Variable Definition

Variable	Definition
AI_Prob	The probability that an analyst report is AI-assisted. It is a document-level probability, produced by a supervised machine learning classifier trained to distinguish human-written analyst reports from reports that incorporate generative AI text.
Accuracy	Accuracy is defined as the negative of the proportional mean absolute forecast error (PMAFE). Following Ben-Rephael et al. (2023), PMAFE is calculated as the absolute difference between the forecasted and actual earnings per share, scaled by the average absolute forecast error across all analysts covering the same firm and forecast date.
|CAR|	|CAR| is defined as the absolute value of the cumulative two-day abnormal return beginning on the report date.
Male	Male equals one if the analyst is identified as male and zero otherwise.
Edu	Edu is a categorical variable capturing educational attainment, coded from 0 to 3, with higher values indicating higher levels of education.
CFA	CFA is an indicator variable that equals one if the analyst holds the Chartered Financial Analyst (CFA) designation and zero otherwise.
AllStar	AllStar is an indicator variable that equals one if the analyst has ever been recognized as an Institutional Investor All-Star analyst prior to 2022, and zero otherwise.
ForecastFrequency	The logarithm of one plus the number of forecasts made by the corresponding analyst in the corresponding year.
ForecastAge	The logarithm of one plus the number of calendar days between the forecast date and the corresponding earnings report date.
FirmsCovered	The logarithm of one plus the number of firms the corresponding analyst covers in the corresponding year.
IndCovered	The logarithm of one plus the number of two-digit SIC industries the corresponding analyst covers in the corresponding year.
AnalystExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for any firm in the IBES database.
Valuation	Valuation, equals zero if a report only employs one type of valuation models, and equals to one if a report uses both multiple-based and present-value based valuation models.
Private	Private, is the number of distinct types of private communication between an analyst and the management mentioned in a report.
Public	Public, is the number of distinct types of public communication between an analyst and the management mentioned in a report.
Tone	Tone, takes the value of 1 if a report is negative or pessimistic, and 2 if a report is neutral, mixed, or balanced, and 3 if a report is positive or optimistic.
Big_data	Big data, is the number of distinct sources of big or alternative data mentioned in a report.
Topics	Topics, is the number of distinct topics mentioned in a report, following Sheng et al. (2025).
Stakeholders	Stakeholders, is the number of distinct types of stakeholders mentioned in a report.
Analytical	Analytical is the intensity of language that reflects logical and formal thinking in a report. The measure is constructed using the LIWC-22 software (Boyd et al. 2022).
Authentic	Authentic is the intensity of language that reflects perceived honesty and genuineness in a report. The measure is constructed using the LIWC-22 software (Boyd et al. 2022).

Causation	Causation is the fraction of causal words (e.g., how, because, make, and why). The measure is constructed using the LIWC-22 software (Boyd et al. 2022).
Differentiation	Differentiation is the fraction of words that reflects cognitive complexity, reflecting thinking in conditions, contrasts, and alternatives (e.g., but, not, if, and or). The measure is constructed using the LIWC-22 software (Boyd et al. 2022).
Word_ratio	Word ratio, is the relative ratio of word counts over the sum of word counts and number counts in a report.
AnalystFirmExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for the corresponding firm.
AnalystIndExperience	The logarithm of one plus number of years since the corresponding analyst first issued a forecast for the corresponding industry.
TOP10	TOP10 is an indicator variable that equals one if an analyst works at a top decile brokerage house in year.
SIZE	The value of firm size in a given quarter, measured as the natural logarithm of market value of equity (MVE).
MTB	The market-to-book ratio of equity in a given fiscal quarter.
ROA	Return on assets in a given quarter, calculated as income before extraordinary items divided by total assets.
Loss	Loss is an indicator variable that equals one if the firm reports negative net income in the corresponding fiscal year, and zero otherwise.
Horizon_Day	Days between the forecast date and the earnings announcement date.
AI_ProbHigh	A document-level AI usage probability generated by an alternative XGBoost classifier trained on hybrid analyst reports in which approximately 75 percent of sentences are replaced with ChatGPT-3.5–generated rewrites.
AI_Prob_RF	A document-level AI usage probability generated by a Random Forest classifier trained on ChatGPT-3.5–rewritten analyst reports and hybrid reports with approximately 50 percent AI-rewritten content.
AI_ProbHigh_RF	A document-level AI usage probability generated by a Random Forest classifier trained on ChatGPT-3.5–rewritten analyst reports and hybrid reports with approximately 75 percent AI-rewritten content.
AI_Prob_LR	A document-level AI usage probability generated by a Logistic Regression classifier trained on ChatGPT-3.5–rewritten reports and 50 percent hybrid reports.
AI_ProbHigh_LR	A document-level AI usage probability generated by a Logistic Regression classifier trained on ChatGPT-3.5–rewritten reports and 75 percent hybrid reports.

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

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

B.1 Construction of Training Data

B.1.1 From PDF to Clean Text

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

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

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

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

Using the cleaned and re-indexed text, we next construct AI-assisted versions of analyst reports for supervised training. For each selected report, we generate AI-assisted counterparts by rewriting each sentence of the original analyst-written content using OpenAI’s GPT-3.5-Turbo. The rewriting process preserves the original analytical meaning and professional tone of the report. Reports rewritten with approximately 50% AI-assisted content serve as the primary training benchmark. Alternative AI-intensity benchmarks (e.g., 75% AI-assisted content) are used in robustness analyses.

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

B.1.2 GPT-Based Sentence Rewriting

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

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

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

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

We compared AI-detection results across the two rewritten versions using a subsample of reports and found no statistically meaningful differences in detection outcomes. Commercial detectors consistently identified the rewritten text as AI-generated under both prompt conditions. This robustness check alleviates concerns that our findings are driven by the specific wording of the rewriting prompt. To study how document-level AI-detection systems perform when analyst reports are only partially generated or edited by generative AI, we construct controlled hybrid documents in which only a subset of originally human-authored sentences is replaced with AI-rewritten text. For each human-written report in our sample, we preserve the original narrative order and randomly select a fixed proportion of sentences to be replaced with AI-generated rewrites. Replacement is carried out independently for each report, with the random selection

process made fully reproducible through seeding. The replacement proportions are set at 25%, 50%, 75%, and 100%.

Figure B1. Rewriting



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

B.2 Text Preprocessing and Feature Engineering

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

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

1) Lexical Features: total word count, vocabulary diversity, and repetition rate.
2) Syntactic Features: clause density, punctuation usage, and the relative frequency of key part-of-speech categories such as nouns, verbs, determiners, and conjunctions.
3) Error Features: the number of spelling mistakes, grammatical issues, and punctuation irregularities in each sentence.
4) Sentiment Features: the polarity (positive vs. negative tone) and subjectivity (objective vs. subjective tone) of the sentence derived using TextBlob.
5) Readability Features: the Flesch Reading Ease Score as well as language model-based fluency metrics such as GPT-2 perplexity and smoothness.
6) TF-IDF (Term Frequency-Inverse Document Frequency) Features: the importance and distribution of terms within the sentence, based on standard term frequency-inverse document frequency scores computed using scikit-learn.

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

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

Tabel B1.1 Feature Engineering

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

Tabel B1.2 Summary Statistics of Features in the Training Sample

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

B.3 Model Training and Validation

B.3.1 Model Training with Three classifiers

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

Table B2.1 Self-Trained document-level XGBoost model

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

Table B2.2 Self-Trained document-level Random Forest model

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

Table B2.3 Self-Trained document-level Logistical Regression model

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

B.3.2 Validation with GPTZero and Originality.AI

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

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

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

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

Using this binary probability of AI authorship, we classify each document as AI-generated when prob_ai exceeds 50%, and human-authored otherwise. We then evaluate GPTZero's predictions against the true origin of each document, including fully human-written reports, fully AI-rewritten versions, and hybrid documents formed by selectively replacing 25%, 50%, 75%, or 100% of the original sentences with AI-generated rewrites.

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

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

Table B3 GPTZero and Originality AI

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

B.4 Construction of Document-Level AI Usage Measure

We apply the trained XGBoost classifier to all analyst reports in our sample. We use the 50% AI-assisted benchmark as our primary specification because it delivers the strong detection performance and reflects a realistic degree of partial AI assistance in analyst report writing. For each report, the model produces a continuous document-level probability, denoted AI_Prob, which captures the likelihood that the report incorporates generative AI.

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

Table C1. Determinants of AI Usage

Notes: This table reports determinant analysis. Column (1) includes a baseline set of analyst characteristics that mirrors the main test. The specification in Column (1) follows Flam et al. (2023): no fixed effects are included, and standard errors are clustered at the firm and year–quarter levels. Column (2) extends Column (1) by adding additional characteristics following Harford et al. (2019). Column (3) incorporates additional determinants following Peng et al. (2022). Column (4) incorporates additional determinants following Huang et al. (2022). Across Columns (1) – (4), the fixed effects structure and clustering approach remain consistent. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) AI_Prob	(2) AI_Prob	(3) AI_Prob	(4) AI_Prob
Male	0.083***	0.091***	0.091***	0.092***
	(0.016)	(0.016)	(0.016)	(0.017)
Edu	-0.004	0.000	0.001	0.002
	(0.005)	(0.005)	(0.005)	(0.005)
CFA	-0.015	-0.030***	-0.030***	-0.031***
	(0.009)	(0.009)	(0.009)	(0.009)
AllStar	-0.023*	-0.071***	-0.073***	-0.072***
	(0.012)	(0.011)	(0.011)	(0.011)
ForecastFrequency	-0.004	-0.011	-0.012	-0.013
	(0.015)	(0.011)	(0.011)	(0.010)
ForecastAge	0.006**	0.009***	0.009***	0.009***
	(0.003)	(0.002)	(0.002)	(0.002)
FirmsCovered	0.008	0.016	0.019	0.021
	(0.021)	(0.017)	(0.016)	(0.017)
IndCovered	-0.012	-0.013	-0.013	-0.015*
	(0.008)	(0.008)	(0.008)	(0.008)
AnalystExperience	0.008	0.020***	0.019**	0.020**
	(0.006)	(0.006)	(0.006)	(0.008)
AnalystFirmExperience		-0.001	-0.003	-0.003
		(0.005)	(0.005)	(0.006)
TOP10		0.103***	0.098***	0.097***
		(0.012)	(0.012)	(0.012)
SIZE			0.004**	0.003
			(0.002)	(0.002)
MTB			-0.000	-0.000
			(0.000)	(0.000)
AnalystIndExperience				-0.002
				(0.008)
ROA				-0.036
				(0.059)
Loss				-0.016*
				(0.008)
Observations	74,721	74,721	74,436	74,415
R-squared	0.007	0.027	0.027	0.028
Adj R-squared	0.00665	0.0267	0.0273	0.0277

Table C2. Robustness: Alternative AI Detection Models

Notes: This table reports OLS regression results examining the robustness of the association between AI usage and forecast accuracy to alternative AI detection models and specifications. The dependent variable is Accuracy. AI_ProbHigh denotes an alternative AI usage probability corresponding to a higher degree of AI involvement in the construction of the training benchmark (e.g., reports with approximately 75% AI-assisted content, relative to the 50% baseline) using the primary XGBoost-based model. AI_Prob_RF and AI_Prob_LR denote AI usage probabilities generated using Random Forest and Logistic Regression classifiers, respectively. AI_ProbHigh_RF and AI_ProbHigh_LR denote corresponding alternative AI usage probabilities constructed under the same higher AI-involvement benchmark using these models. All specifications include the same set of control variables, firm–year fixed effects, and broker–year fixed effects as in Table 3. Standard errors are clustered at the analyst–year level. All continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) Accuracy	(3) Accuracy	(4) Accuracy	(5) Accuracy
AI_ProbHigh	0.023**				
	(0.010)				
AI_Prob_RF		0.037**			
		(0.017)			
AI_ProbHigh_RF			0.040**		
			(0.016)		
AI_Prob_LR				0.037***	
				(0.014)	
AI_ProbHigh_LR					0.030**
					(0.013)
Male	-0.022	-0.022	-0.022	-0.022	-0.022
	(0.023)	(0.022)	(0.022)	(0.022)	(0.022)
Edu	-0.000	0.000	-0.000	-0.000	-0.000
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
CFA	0.035***	0.035***	0.035***	0.034***	0.034***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
AllStar	0.059***	0.060***	0.059***	0.060***	0.059***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
ForecastFrequency	-0.049***	-0.049***	-0.049***	-0.049***	-0.049***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
ForecastAge	-0.125***	-0.125***	-0.125***	-0.125***	-0.125***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
FirmsCovered	-0.017	-0.018	-0.017	-0.019	-0.018
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
IndCovered	0.044**	0.045**	0.044**	0.046***	0.045***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
AnalystExperience	0.006	0.006	0.006	0.006	0.006
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Firm-Year FE	Yes	Yes	Yes	Yes	Yes
Broker-Year FE	Yes	Yes	Yes	Yes	Yes
Observations	74,634	74,634	74,634	74,634	74,634
R-squared	0.286	0.286	0.286	0.286	0.286
Adj R-squared	0.220	0.220	0.220	0.220	0.220

Table C3. Robustness: Incorporating Analyst Fixed Effects

This table reports OLS regression results using analyst fixed effects. The design mirrors the baseline specification in Table 3 but replaces the analyst-level control variables with analyst fixed effects, which absorb all time-invariant analyst characteristics. We exclude the analyst-level controls used in the baseline regressions (e.g., gender, education, CFA designation, All-Star status). Additionally, we include firm-level fundamentals and brokerage characteristics (e.g., TOP10, SIZE, MTB, ROA, Loss). The dependent variables are Accuracy and |CAR|. All regressions include analyst fixed effects and year fixed effects. Standard errors are clustered at the year level. Variable definitions are provided in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1) Accuracy	(2) |CAR|
AI_Prob	0.015**	0.003***
	(0.006)	(0.001)
ForecastFrequency	-0.037**	-0.000
	(0.011)	(0.002)
ForecastAge	-0.125***	0.003**
	(0.013)	(0.001)
IndCovered	0.063	-0.002
	(0.047)	(0.002)
AnalystIndExperience	0.019**	0.001
	(0.006)	(0.001)
TOP10	0.042	0.005
	(0.059)	(0.005)
SIZE	0.005**	-0.004***
	(0.002)	(0.000)
MTB	-0.001	0.000
	(0.000)	(0.000)
ROA	-0.146*	-0.049*
	(0.054)	(0.022)
Loss	0.024	0.016***
	(0.016)	(0.001)
Analyst FE	Yes	Yes
Year FE	Yes	Yes
Observations	74,307	74,385
R-squared	0.072	0.177
Adj R-squared	0.0660	0.172

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:
"""