# Simulating Analyst Forecast Behavior using LLMs*

Hyuk An[1] Seohyun Lee[2]

June 17, 2026

**Abstract**

We examine whether a commercial large language model (LLM) can match human analysts in numerical earnings per share (EPS) forecasting. Using 121,369 forecasts for 607 Korean listed firms over 2018-2025, we generate paired LLM forecasts utilizing only information publicly available as of the analyst forecast date, with external search disabled. Human analysts remain more accurate on average: median forecast errors are 1.9% of stock price for analysts and 2.3% for the LLM, implying that the LLM underperforms analysts by about 20%. The LLM outperforms consensus forecasts but remains less accurate than analysts' prior forecasts. The analyst advantage is larger for loss-making, highly leveraged, and smaller firms, as well as for firms with greater analyst disagreement. Comparing model generations under identical prompts, GPT-5.4-nano closes 68% of the gap between GPT-4.1-nano and human analysts. Overall, current commercial LLMs do not yet perform on par with human analysts, but their relative performance improves across model generations and is strongest in settings with richer public information and lower uncertainty.

*JEL Classification: G14, G17, M41*
*Keywords: Large language models, Analyst forecasts, EPS forecasting, Generative Artificial Intelligence, Simulated Economic Agents.*

*We thank an anonymous reviewer for constructive comments and suggestions that helped improve the paper.
[1]KDI School of Public Policy and Management, E-mail: hyukan@kdis.ac.kr.
[2]Corresponding author, KDI School of Public Policy and Management, E-mail: slee6@kdis.ac.kr.

## 1 Introduction

Recent research suggests that large language models (LLMs) can match or outperform professional forecasters in predicting macroeconomic variables. Hansen et al. (2026) show that LLM agents perform comparably to human experts in short-term macroeconomic forecasting while outperforming them at medium- to long-term horizons. Beyond expert forecasting, LLMs also appear capable of simulating the economic expectations of the general public. Zarifhonarvar (2025) and Anesti et al. (2025) show that LLMs replicate not only aggregate patterns in survey-based inflation expectations but also biases associated with demographic heterogeneity across households.

In finance, however, the evidence remains more limited. Earlier studies comparing AI and human forecasts focus on traditional machine learning methods, such as random forests and gradient boosting, rather than LLMs. In the context of earnings forecasting, these studies show that ML models can outperform human analysts (Chen et al., 2022; Van Binsbergen et al., 2023). More recently, Kim et al. (2024) extend this comparison to LLMs by studying directional earnings per share (EPS) forecasts and show that GPT performs competitively with analysts on average. Their results also indicate that analysts retain a stronger advantage for smaller firms and firms with more volatile earnings. This pattern is consistent with the broader evidence in Cao et al. (2024), which shows that humans continue to outperform machine learning models when institutional knowledge is critical. Taken together, these findings suggest that the relative performance of LLMs and human analysts depends not only on average forecasting ability but also on the informational environment of the forecasting task.

Despite these advances, it remains unclear whether LLMs can match human analysts in the more demanding task of numerical EPS forecasting. Existing evidence in finance focuses primarily on directional EPS prediction rather than point forecasts at the individual forecast level. We address this gap by constructing 121,369 analyst–LLM forecast pairs from individual EPS estimates for 607 Korean listed firms over 2018–2025, using analyst and consensus data from FnGuide, a major provider of Korean financial market data. For each analyst forecast, we generate a corresponding LLM forecast using only information publicly available as of the forecast date, including

firm financial statements, macroeconomic conditions, historical EPS, the analyst's own prior forecasts, and prevailing market consensus. To isolate intrinsic forecasting performance, rather than the ability to retrieve additional information, we disable external search and use commercially available, off-the-shelf models with default API settings, without fine-tuning or parameter optimization. We use GPT-5.4-nano as the primary model and GPT-4.1-nano to assess improvements across model generations.

We document three main results. First, human analysts outperform the LLM on average, with median forecast errors of 1.9% of stock price for analysts and 2.3% for the LLM, implying an analyst advantage of about 20%. Second, this advantage is highly heterogeneous: it is larger for loss-making firms, firms with greater analyst disagreement, and financially fragile firms, but smaller for large, profitable, low-leverage firms. Third, performance improves substantially across model generations, with the newer model closing 68.1% of the gap relative to human analysts. Additional tests around model training cutoffs provide little evidence that these results are driven primarily by memorization of realized outcomes. Overall, our findings suggest that current commercial LLMs do not yet match human analysts in numerical EPS forecasting, but their relative performance is improving and is strongest in environments with richer public information and lower uncertainty.

Our contribution to the literature is twofold. First, we provide a one-to-one comparison of numerical EPS forecasts between LLMs and individual human analysts in a setting that closely matches the information environment of actual earnings forecasting. In particular, our design places the LLM in an information environment that includes not only standard public information but also the analyst's own forecast history and the prevailing consensus forecast. Both signals are central to EPS forecasting but are typically unavailable in survey-based macroeconomic forecasting settings. By contrast, in the information environment of financial analysts, such information is readily accessible through market data platforms that provide real-time consensus updates and estimate histories.

Second, our design allows us to benchmark LLM forecasts against two reference points: the analyst's own prior forecast and the prevailing consensus forecast. In many forecasting settings,

expectation formation depends on prior beliefs and consensus, even when these inputs are not directly observed by researchers in ex post analyses. Incorporating these benchmark signals into the LLM's information set therefore yields a comparison that more closely matches actual forecasting practice and provides a more interpretable assessment of model performance. We find that the LLM improves on consensus forecasts but remains less accurate than analysts' prior forecasts.

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 outlines the prompt design and empirical framework. Section 4 presents the main results. Section 5 concludes.

## 2 Data

### 2.1 Analyst forecasts

We obtain individual analyst EPS forecasts from FnGuide, a Korean financial data provider. EPS is defined as the net income attributable to controlling shareholders per share under K-IFRS. The sample consists of annual EPS forecasts of Korean listed firms for fiscal years 2018 through 2025. We restrict the sample to annual consolidated forecasts for firms with a December fiscal year-end. We further require the forecast date to fall within the corresponding target fiscal year and retain only observations with available stock price data and consolidated financial statements from DART, Korea's official corporate disclosure system operated by the Financial Supervisory Service. Finally, we require at least three distinct analysts to issue EPS forecasts for the same firm and fiscal year within the same quarter.

Table 1 reports the composition of our sample, comprising 121,369 analyst-firm-date observations across 607 unique firms and 822 analysts. It shows the distribution of observations by target year, market, and industry. KOSPI, the primary exchange in Korea, accounts for 84.7% of the observations, while KOSDAQ, a secondary market oriented toward smaller and growth-oriented firms, comprises 15.3%. By industry, manufacturing constitutes the largest share (54.1%), followed by information & communication (15.7%) and financial & insurance activities (10.5%). The average forecast horizon is 189 days.

### 2.2 Macro and firm-level data

Table 2 reports the variables used as inputs to the LLM forecasts. These variables are chosen to approximate the information set relevant for EPS forecasting, combining firm characteristics, macroeconomic conditions, financial statement information, market indicators, and forecast-based benchmark variables.

Firm characteristics, including firm identity, industry classification, listing market, and market capitalization, are obtained from DART and FnGuide, providing basic information on the firm's business profile and operating environment. Macroeconomic conditions, including GDP growth, inflation, interest rates, exchange rates, and the Bank of Korea's real-time economic outlook, are drawn from the Bank of Korea's Economic Statistics System (BOK ECOS) and the Bank of Korea's Economic Outlook reports. Because earnings expectations are shaped by aggregate demand conditions, monetary policy, and external shocks, macroeconomic information is an important input for EPS forecasting. Its relevance is particularly pronounced in the Korean economy, where firm performance is closely tied to trade exposure, exchange-rate fluctuations, and imported energy costs.

Financial statement information, including the income statement, balance sheet, and cash flow items, is obtained from DART and FnGuide and captures the firm's recent operating performance, financial position, and cash flow dynamics. Market indicators include stock prices from FnGuide, Dubai crude oil prices from Petronet operated by the Korea National Oil Corporation, corporate bond credit spreads from the Korea Financial Investment Association (KOFIA), and global trade and industrial production indices from the CPB Netherlands Bureau for Economic Policy Analysis. These variables reflect forward-looking information about market expectations, financing conditions, and external cost pressures.

Forecast-based benchmark variables, including historical EPS, prevailing consensus, and the analyst's own prior forecasts, are obtained from FnGuide. We include them because forecast formation often depends on both prior beliefs and the benchmark expectations of others.

To preserve the real-time information environment, all variables supplied to the LLM are re-

stricted to information that is publicly observable strictly before the forecast date, $t$. Macroeconomic and external market variables are incorporated on their official release dates, financial statement items on their actual DART filing dates, and analyst consensus, as well as the analyst's own prior forecasts, are limited to those issued before $t$.

## 3 Empirical framework

### 3.1 Simulating LLM forecasts

Our objective is to assess whether commercial LLMs can generate analyst-level forecasts in an off-the-shelf setting. Accordingly, we use the API with default parameter values and apply no fine-tuning or parameter optimization (Details of the parameter settings are reported in Appendix Table A1). Forecasts are generated through the OpenAI Responses API using GPT-4.1-nano and GPT-5.4-nano. We focus on these models because they are low-cost, low-latency variants intended for high-volume API use, making them a relevant benchmark for practical deployment in large-scale forecasting exercises. GPT-5.4-nano serves as the primary model for the analyst–LLM accuracy comparison, while GPT-4.1-nano is used as a benchmark in the model advancement comparison.

Appendix A.2 summarizes the prompt structure. At the highest level of instruction, the developer message assigns the LLM the role of an analyst at the same brokerage firm on the same forecast date and explicitly prohibits both external searches and the use of information released after that date. The user prompt then provides six categories of information, each restricted to data available before the forecast date: company overview, macroeconomic indicators, annual financial statements, monthly stock prices, historical EPS and consensus forecasts, and the analyst's own prior forecasts. We highlight three features of the prompt design. First, all inputs are aligned to their actual release dates to preserve the real-time information environment. Second, the prompt incorporates benchmark signals that are central to practical forecasting, especially the prevailing consensus and the analyst's own forecast history. Third, it includes the central bank's real-time projections for key macroeconomic variables, thereby incorporating forward-looking macroeconomic

information available at the time of forecast formation.

We consider potential threats to the validity of the LLM forecasts: external search and memorization from training data of realized EPS. To address these threats, we disable external search and impose a developer message prohibiting reference to events after the forecast date, following Wu et al. (2025). Additionally, we conduct a knowledge cutoff analysis following Cao et al. (2026). If the model were simply retrieving realized EPS, it should systematically outperform human analysts. Instead, as reported in Appendix A.3, analyst superiority remains positive both before and after each model's training cutoff, with small differences across subsamples. For GPT-5.4-nano in particular, analyst superiority declines after the cutoff, which is inconsistent with a retrieval-based explanation.

### 3.2 Evaluating LLM forecasts

We evaluate LLM forecasts by comparing their accuracy with that of human analysts across the forecast horizon. Following Richardson et al. (2004), forecast error (FE) is defined as the absolute difference between realized EPS and the forecasted value, scaled by the stock price at the end of the prior fiscal year, $P_{f,T-1}$:

$$FE^{j}_{i,f,t} = \frac{|EPS_{f,T} - \widehat{EPS}^{j}_{i,f,t}|}{P_{f,T-1}} \tag{1}$$

where $i$ indexes the analyst, $f$ denotes the firm, $t$ the calendar date on which the forecast is issued, and $j \in \{\text{Analyst}, \text{LLM}\}$ indicates whether the forecast is generated by the human analyst or the LLM. Each LLM forecast is matched one-to-one to an analyst forecast instance $i$ of firm $f$ issued on date $t$, by conditioning on the same firm, information date, and forecast target, ensuring that differences in $FE$ reflect forecasting ability rather than differences in the information set or horizon.

Following Bradshaw et al. (2012), we then define analyst superiority as

$$SUP_{i,f,t} = FE^{LLM}_{i,f,t} - FE^{Analyst}_{i,f,t} \tag{2}$$

such that a positive value indicates greater forecast accuracy for the analyst than for the matched LLM forecast. In analyzing the accuracy and superiority metrics, we trim the top and bottom 0.5% of observations to limit the influence of extremes.

As an initial step, we examine how forecast errors vary over the forecast horizon for the LLM and human analysts. We then test whether $E[SUP_{i,f,t}]$ is centered at zero using the Wilcoxon signed-rank test (Wilcoxon, 1992), a nonparametric test for paired data that does not require normality.

To examine whether the accuracy gap between analysts and LLM varies with firm characteristics, we estimate the following specifications using analyst-firm-time panel data:

$$SUP_{i,f,t} = \alpha_i + \delta_f + \gamma_{q(t)} + \beta' X_{f,q(t)-1} + \theta' Z_{i,f,t} + \varepsilon_{i,f,t}, \tag{3}$$

where $SUP_{i,f,t}$ denotes analyst superiority for analyst $i$, firm $f$, and forecast date $t$. $q(t)$ denotes the year-quarter in which the forecast date $t$ falls. $\alpha_i$ denotes analyst fixed effects, $\delta_f$ firm fixed effects, and $\gamma_{q(t)}$ year-quarter time fixed effects. $X_{f,q(t)-1}$ is a vector of lagged firm-level characteristics, and $Z_{i,f,t}$ is a forecast-level control for forecast horizon, measured as $\ln(1 + Horizon_{i,f,t})$ to allow for a potentially nonlinear relationship between horizon and forecast superiority. This specification absorbs time-invariant heterogeneity across analysts and firms, as well as common shocks over time. Using lagged firm characteristics also helps mitigate concerns that contemporaneous firm attributes are jointly determined with the forecasting environment.

Following Kim et al. (2024), the vector $X_{f,q(t)-1}$ includes *Size* (log market capitalization), *BtoM* (book-to-market ratio), *Leverage* (total liabilities divided by total assets), and *Loss* (an indicator equal to one if net income in the preceding quarter is negative). In addition, we add *DU*

(disagreement uncertainty), constructed using the fixed-horizon transformation of Lee (2026).[1] All firm characteristics are measured at the end of the preceding quarter.

We next estimate a specification that replaces separate analyst and firm fixed effects with analyst-firm fixed effects:

$$SUP_{i,f,t} = \alpha_{i,f} + \gamma_{q(t)} + \beta' X_{f,q(t)-1} + \theta' Z_{i,f,t} + \varepsilon_{i,f,t}, \tag{4}$$

where $\alpha_{i,f}$ denotes analyst-firm fixed effects. Relative to Eq. (3), this specification absorbs time-invariant heterogeneity specific to an analyst-firm pair, such as persistent differences in analyst familiarity, coverage intensity, or firm-specific information processing.

We also consider, as a robustness check, a specification that further augments Eq. (4) with industry-time fixed effects:

$$SUP_{i,f,t} = \alpha_{i,f} + \lambda_{k,q(t)} + \beta' X_{f,q(t)-1} + \theta' Z_{i,f,t} + \varepsilon_{i,f,t}, \tag{5}$$

where $\lambda_{k,q(t)}$ denotes industry-time fixed effects for industry $k$ in period $q(t)$. This specification absorbs time-varying shocks common to firms within the same industry, including changes in business conditions, the disclosure environment, and earnings uncertainty. Eq. (5) therefore provides the most restrictive specification, identifying the effect of firm characteristics from within analyst-firm variation over time while accounting for industry-specific shocks in each period.

In all specifications, we cluster standard errors at the firm level. This choice reflects the fact that the regressors of interest vary at the firm-period level, while the dependent variable is observed for multiple analysts covering the same firm. As a result, residuals may be correlated within firms across analysts and over time. As a robustness check, we also report the results based on two-way clustered standard errors by firm and year to allow for common shocks across firms within a given period, and by firm and analyst to account for residual correlation arising from analysts covering multiple firms over time.

[1] See Appendix A.4 for details on the variable construction.

Finally, to assess whether model advancement narrows the analyst-LLM gap, we compare forecasts generated by GPT-4.1-nano and GPT-5.4-nano under identical prompts. Under this design, any difference in accuracy is attributable to differences in model capability. We quantify the fraction of the gap closed by the newer model as

$$RGR = \frac{FE_{4.1}^{LLM} - FE_{5.4}^{LLM}}{FE_{4.1}^{LLM} - FE^{Analyst}}. \tag{6}$$

where $RGR$ denotes the relative reduction in the analyst-LLM accuracy gap achieved by the newer model, and each $FE$ term denotes the median forecast error for the indicated forecaster over the common sample. Subscripts $i$, $f$, and $t$ are suppressed for clarity.

## 4 Results

### 4.1 Forecast accuracy gap

Table 3 reports descriptive statistics. Relative to realized EPS, both analyst and LLM predictions exhibit an optimistic bias. The median realized EPS (in Korean Won, KRW) is 3,294, whereas the median analyst and LLM forecasts are 4,060 and 4,300, respectively. This pattern suggests that LLM forecasts tend to overestimate EPS to a greater extent than analyst forecasts.

Figure 1 shows how forecast accuracy varies over the forecast horizon for four forecast benchmarks: the analyst, the GPT-5.4-nano model, the consensus forecast, and the prior forecast. The consensus forecast is defined as the median of the most recent forecasts issued by other analysts for the same firm-year, excluding the analyst's own forecast. The prior forecast refers to the analyst's most recent previous forecast. Both the consensus and the prior forecasts are also provided to the LLM as prompt inputs. The upper panel plots their median forecast error (FE) for each benchmark, smoothed using a centered 7-day rolling average. The analyst consistently exhibits the lowest median FE across the forecast horizon, with median FE rising from 0.012 to 0.030 as the horizon increases. The other three forecasts follow a similar pattern at uniformly higher levels. The lower panel presents individual SUP observations for the analyst-LLM comparison (dots), along

with the centered 7-day rolling median (line). The median remains positive throughout the forecast horizon and shows no clear upward or downward pattern. This implies that the relative superiority of human forecasts does not change systematically across the forecast horizon.

To test whether paired differences in forecast errors are statistically distinguishable from zero, Table 4 reports pooled pairwise Wilcoxon signed-rank tests for all pairwise comparisons among the four benchmarks.[2] The results indicate a clear ranking in forecast accuracy. The analyst performs best, with the lowest median and mean FE (0.0198 and 0.0453, respectively), and significantly outperforms the LLM, the consensus forecast, and the prior forecast in all pairwise comparisons (see the first column). Among the remaining three benchmarks, the prior forecast outperforms both the LLM and the consensus forecast, while the LLM in turn outperforms the consensus forecast. All pairwise differences are statistically significant at the 1% level. Although the differences among the non-analyst benchmarks are relatively small in magnitude, the results suggest that the LLM does not merely replicate the prompt inputs it receives. Taken together, these results suggest that the analyst's own information advantage remains difficult for the LLM to match, even when the model is provided with both the analyst's prior forecast and the consensus forecast as inputs.

### 4.2 Firm-level heterogeneity in analyst superiority

While the preceding results establish that analysts outperform the LLM on average, this advantage need not be uniform across firms. We therefore examine firm-level heterogeneity in analyst superiority by relating the analyst's relative forecasting performance to observable firm characteristics. Table 5 reports estimates of Eqs.(3) and (4) for the pooled sample and for each forecast horizon ($h = 1q$ to $h = 4q$). The baseline specification includes analyst, firm, and year-quarter fixed effects, while Eq. (4) replaces the separate analyst and firm fixed effects with analyst$\times$firm fixed effects to absorb analyst-firm specific unobservables, such as an analyst's familiarity with a given firm. The horizon variable is included in the pooled regressions but omitted from the subsamples, where within-quarter variation is limited.

[2]The number of observations in this analysis is smaller than in Table 3 because the sample is restricted to cases in which both the individual analyst's prior forecast and the consensus forecast are available.

In the pooled sample, analyst superiority is greater for loss-making firms, firms with greater analyst disagreement, more highly leveraged firms, and smaller firms, while book-to-market is not significantly related to analyst superiority. This pattern suggests that the analyst's advantage over the LLM is larger when the firm is harder to evaluate. Loss-making firms face greater uncertainty about future performance and viability. Greater analyst disagreement likewise points to a more difficult information environment. Highly leveraged firms are more financially fragile, while smaller firms tend to have less stable fundamentals. By contrast, the gap between analysts and the LLM is narrower for large, profitable, low-leverage firms with greater analyst agreement. These patterns remain after introducing analyst$\times$firm fixed effects in Eq. (4). Although the coefficients attenuate modestly, their signs and statistical significance remain unchanged.

When we split the sample by forecast horizon, the sources of analysts' superiority differ between near-term and more distant forecasts. Loss-making status becomes more important as the forecast approaches earnings realization, suggesting that the analyst's advantage over the LLM is particularly large for loss-making firms when forecasts are made close to the earnings realization. Analyst disagreement is most informative at the shortest and longest horizons. By contrast, the analyst's advantage is larger for smaller and more highly leveraged firms at longer horizons. A similar pattern is also evident when replacing the analyst-firm fixed effects in Eq. (4).

Using the interquartile ranges reported in Panel B of table 3, the magnitudes of the pooled baseline coefficients can be interpreted as follows. Moving from a loss-making firm to a profitable firm reduces the analyst's advantage over the LLM by 0.0091, implying a 0.91%p narrowing in the analyst–LLM gap in absolute forecast errors, scaled by the previous fiscal-year stock price. Lowering analyst disagreement, our proxy for firm-level uncertainty, from the 75th to the 25th percentile narrows the analyst's advantage by 0.30%p. An analogous interquartile decline in leverage reduces the advantage by 0.37%p, while increasing firm size from the 25th to the 75th percentile lowers it by 0.41%p. Each of these changes, considered separately, exceeds the median analyst advantage over the LLM ($SUP$) of 0.0019.

These results are robust to alternative specifications. Table A3 reports the estimation results

of Eq. (5), under both firm-level and two-way clustering by firm and time. Appendix Table A4 re-estimates Eqs.(3) and (4) under two-way clustering by firm and year.[3] Neither the inclusion of industry-by-year-quarter fixed effects nor the adoption of a more conservative clustering structure alters the main findings.

### 4.3 Comparison across LLM generations

The upper panel of Figure 2 plots the median FE for models across the forecast horizon. For both models, FE increases as the forecast horizon lengthens, but GPT-5.4-nano consistently exhibits a lower FE than GPT-4.1-nano throughout. The lower panel shows the share of the GPT-4.1-nano accuracy gap relative to human analysts that is closed by GPT-5.4-nano. This share ranges from 50% to 80%, with a median of 68.1%. These findings suggest that improvements in model capability may further reduce the remaining gap between LLMs and human analysts.

Figure 3 provides additional insight into forecast bias by comparing signed forecast errors for analysts and the two LLMs. GPT-5.4-nano's signed errors track those of analysts closely, as reflected in the high correlation and the large share of same-sign observations. At the same time, its fitted relation lies below the 45-degree line when analyst forecast errors are negative and above it when analyst forecast errors are positive, suggesting that GPT-5.4-nano tends to amplify the direction of the analyst's miss. GPT-4.1-nano exhibits a different pattern. Its fitted relation lies above the 45-degree line over most of the support, implying more positive signed forecast errors than those of analysts even when analysts' own errors are negative. This pattern is consistent with a stronger unconditional optimistic bias in GPT-4.1-nano, whereas GPT-5.4-nano appears less uniformly optimistic and more closely aligned with the direction of analyst forecast errors.

## 5 Conclusions

This study examines whether an LLM can achieve analyst-level accuracy in numerical EPS forecasting. We find that, on average, analysts remain statistically significantly more accurate than

[3]Results under two-way clustering by firm and analyst are reported in Appendix Table A5.

LLMs. The LLM improves on consensus forecasts but remains less accurate than analysts' prior forecasts. The analyst advantage is larger for loss-making, highly leveraged, and smaller firms, as well as for firms with greater analyst disagreement. At the same time, model improvement is substantial: relative to GPT-4.1-nano, GPT-5.4-nano closes about 68% of the analyst-LLM accuracy gap. These findings suggest that current LLMs are better viewed as complements to, rather than substitutes for, human analysts, with their usefulness likely greatest where public information is richer and uncertainty is lower.

Several limitations point to directions for future research. First, our prompts rely exclusively on structured quantitative inputs. Incorporating unstructured information, such as earnings call transcripts, may further reduce the gap between human analysts and LLMs. Second, we focus on nano-class models for cost efficiency, and results may differ for frontier models. Future research could also analyze the textual reasoning generated by reasoning-oriented models to better understand how LLMs process financial information when forming forecast judgments. It could further prompt the LLM to report a self-assessed confidence score alongside each forecast and examine whether higher-confidence forecasts are systematically more accurate.

Finally, this study has implications for AI governance in financial intermediation. As LLMs outperform the analyst consensus and improve rapidly across generations, low-cost models may expand forecast coverage for small and neglected firms. However, because the human–AI gap is largest for harder-to-evaluate firms, broad deployment could lower information quality for the very firms that already face a shortage of high-quality information. These findings have implications for the use and disclosure of AI-generated forecasts, as well as for investor protection as retail investors increasingly rely on such tools.

**Acknowledgements and Disclosures**

The authors received no financial support for this research. The authors have no relevant financial relationships or other potential conflicts of interest to disclose.

## References


[1] Anesti, N., Hill, E., and Joseph, A. (2025). Inflation attitudes of large language models. *arXiv preprint arXiv:2512.14306*.

[2] Bradshaw, M. T., Drake, M. S., Myers, J. N., and Myers, L. A. (2012). A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies*, 17(4):944–968.

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

[4] Cao, S., Jiang, W., and Xu, H. (2026). Seeing the goal, missing the truth: Human accountability for ai bias. *arXiv preprint arXiv:2602.09504*.

[5] Chen, X., Cho, Y. H., Dou, Y., and Lev, B. (2022). Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research*, 60(2):467–515.

[6] Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A. (2026). Simulating the survey of professional forecasters. *Available at SSRN 5066286*.

[7] Kim, A., Muhn, M., and Nikolaev, V. (2024). Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

[8] Lee, S. (2026). Forecast disagreement about firm-level profitability and uncertainty. *Pacific-Basin Finance Journal*, page 103092.

[9] Richardson, S., Teoh, S. H., and Wysocki, P. D. (2004). The walk-down to beatable analyst forecasts: The role of equity issuance and insider trading incentives. *Contemporary accounting research*, 21(4):885–924.

[10] Van Binsbergen, J. H., Han, X., and Lopez-Lira, A. (2023). Man versus machine learning: The term structure of earnings expectations and conditional biases. *The Review of financial studies*, 36(6):2361–2396.

[11] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

[12] Wu, J. C., Xi, J., and Xie, S. (2025). Llm survey framework: Coverage, reasoning, dynamics, identification. Technical report, National Bureau of Economic Research.

[13] Zarifhonarvar, A. (2025). Generating inflation expectations with large language models. *Journal of Monetary Economics*, page 103859.

## Tables and Figures

TABLE 1 Analyst forecasts

| | Obs | Firms | Analysts | Average horizon |
|---|---|---|---|---|
| *By Target Year* | | | | |
| 2018 | 14,721 | 344 | 417 | 185 |
| 2019 | 15,833 | 344 | 382 | 188 |
| 2020 | 15,561 | 342 | 376 | 190 |
| 2021 | 13,498 | 326 | 377 | 194 |
| 2022 | 13,281 | 332 | 371 | 185 |
| 2023 | 14,704 | 345 | 393 | 187 |
| 2024 | 16,499 | 372 | 392 | 192 |
| 2025 | 17,272 | 353 | 394 | 188 |
| Total | 121,369 | 607 | 822 | 189 |
| *By Market* | | | | |
| KOSPI | 102,804 | 323 | 773 | 189 |
| KOSDAQ | 18,565 | 284 | 561 | 186 |
| *By Industry (KSIC Major)* | | | | |
| Manufacturing (C) | 65,624 | 363 | 662 | 189 |
| Information & Comm. (J) | 19,079 | 78 | 252 | 189 |
| Financial & Insurance (K) | 12,796 | 56 | 294 | 187 |
| Wholesale & Retail (G) | 9,082 | 34 | 223 | 188 |
| Other | 14,788 | 76 | 346 | 188 |

Notes: Each observation is a unique analyst $\times$ firm $\times$ forecast date $\times$ target year combination. Average horizon is the average of the number of days from the forecast date to the target fiscal year.

TABLE 2 LLM inputs for forecast generation

| Sec. | Variable | Source | Frequency | Reference period |
|---|---|---|---|---|
| *[0] Company Overview* | | | | |
| 0a | Company name and code | FnGuide | — | — |
| 0b | KSIC sector code and name | DART | — | — |
| 0c | Listing market (KOSPI / KOSDAQ) | DART | — | — |
| 0d | Market capitalization (KRW millions) | FnGuide | Daily | latest daily obs. before $t$ |
| *[1a] Domestic Macroeconomic Data* | | | | |
| 1a-1 | GDP growth, YoY (%) | BOK ECOS | Quarterly | 12 quarters preceding $t$ |
| 1a-2 | CPI inflation, YoY (%) | BOK ECOS | Quarterly | 12 quarters preceding $t$ |
| 1a-3 | Unemployment rate (%) | BOK ECOS | Quarterly | 12 quarters preceding $t$ |
| 1a-4 | Employment rate (%) | BOK ECOS | Quarterly | 12 quarters preceding $t$ |
| 1a-5 | Current account balance ($100M) | BOK ECOS | Quarterly | 12 quarters preceding $t$ |
| 1a-6 | BOK macro forecast | BOK | Annual | 2 years ahead of $t$ |
| *[1b] Market Data* | | | | |
| 1b-1 | Dubai crude oil price ($/bbl) | Petronet | Q. avg. + daily | 12Q + latest day |
| 1b-2 | KRW/USD exchange rate | BOK ECOS | Q. avg. + daily | 12Q + latest day |
| 1b-3 | KRW/100JPY exchange rate | BOK ECOS | Q. avg. + daily | 12Q + latest day |
| 1b-4 | KRW/CNY exchange rate | BOK ECOS | Q. avg. + daily | 12Q + latest day |
| 1b-5 | AA− corporate bond credit spread (%p) | KOFIA | Q. avg. + daily | 12Q + latest day |
| 1b-6 | BBB− corporate bond credit spread (%p) | KOFIA | Q. avg. + daily | 12Q + latest day |
| 1b-7 | World merchandise trade volume (index) | CPB | Quarterly | 12 quarters preceding $t$ |
| 1b-8 | Advanced economy trade volume (index) | CPB | Quarterly | 12 quarters preceding $t$ |
| 1b-9 | Emerging economy trade volume (index) | CPB | Quarterly | 12 quarters preceding $t$ |
| 1b-10 | World industrial production (index) | CPB | Quarterly | 12 quarters preceding $t$ |
| 1b-11 | Advanced economy industrial production | CPB | Quarterly | 12 quarters preceding $t$ |
| 1b-12 | Emerging economy industrial production | CPB | Quarterly | 12 quarters preceding $t$ |
| *[2] Summary Financial Statements (Consolidated K-IFRS, KRW Millions)* | | | | |
| 2a | Selected financial statement items | DART | Annual | 3 most recent fiscal years |
| *[3] Stock Price* | | | | |
| 3a | Monthly closing price (KRW) | FnGuide | Monthly | 36 months preceding $t$ |
| 3b | Monthly trading volume (shares) | FnGuide | Monthly | 36 months preceding $t$ |
| *[4] EPS Data* | | | | |
| 4a | Average shares outstanding | FnGuide | Annual | $T-4$ to $T-1$ |
| 4b | Net income (KRW) | FnGuide | Annual | $T-4$ to $T-1$ |
| 4c | Actual EPS (KRW) | FnGuide | Annual | $T-4$ to $T-1$ |
| 4d | Consensus EPS | FnGuide | Annual | $T-4$ to $T$ |
| 4e | Analyst count | FnGuide | Annual | $T-4$ to $T$ |
| *[5] Past Forecasts of the Focal Analyst* | | | | |
| 5a | Analyst's own prior EPS forecasts | FnGuide | Event | all prior forecasts before $t$ |

Notes: Reference period indicates the time span or reference date of the information provided to the LLM relative to the forecast date $t$. "12Q + latest day" denotes twelve quarterly averages supplied together with a single most-recent daily snapshot. BOK macro forecasts, drawn from the Bank of Korea's Economic Outlook reports, include GDP, CPI, unemployment, employment, and current account for the current and next years. Selected financial statement items include revenue, operating income, profit or loss, total assets, total equity, and total liabilities. Net income is reported attributable to owners of the parent. Consensus EPS is the median peer forecast excluding the focal analyst. Past forecasts include forecast date, target fiscal year, and forecasted EPS. *BOK ECOS* refers to the Bank of Korea's Economic Statistics System, *DART* Korea's official corporate disclosure system operated by the Financial Supervisory Service, *Petronet* the platform operated by the Korea National Oil Corporation, *KOFIA* the Korea Financial Investment Association, *CPB* the CPB Netherlands Bureau for Economic Policy Analysis.

TABLE 3 Descriptive statistics

| | $N$ | Mean | Std | P25 | Median | P75 |
|---|---|---|---|---|---|---|
| *Panel A: Forecast metrics* | | | | | | |
| Actual EPS | 120,155 | 6,590 | 12,862 | 1,021 | 3,294 | 8,741 |
| Analyst EPS | 120,155 | 7,953 | 11,462 | 1,604 | 4,060 | 9,596 |
| LLM EPS | 120,155 | 8,191 | 11,149 | 1,800 | 4,300 | 9,885 |
| $P_{T-1}$ | 120,155 | 122,820 | 198,581 | 28,500 | 59,600 | 144,000 |
| $\text{FE}^{\text{Analyst}}$ | 120,155 | 0.0430 | 0.0757 | 0.0070 | 0.0194 | 0.0484 |
| $\text{FE}^{\text{LLM}}$ | 120,155 | 0.0503 | 0.0829 | 0.0085 | 0.0233 | 0.0588 |
| SUP | 120,155 | 0.0074 | 0.0264 | −0.0024 | 0.0019 | 0.0113 |
| *Panel B: Firm characteristics* | | | | | | |
| Size | 112,694 | 15.11 | 1.51 | 14.05 | 14.97 | 16.10 |
| BtoM | 112,671 | 1.18 | 1.16 | 0.43 | 0.83 | 1.57 |
| Leverage | 112,971 | 0.48 | 0.19 | 0.32 | 0.48 | 0.62 |
| Loss | 110,623 | 0.14 | 0.35 | 0.00 | 0.00 | 0.00 |
| DU | 117,433 | 0.023 | 0.020 | 0.010 | 0.018 | 0.030 |

Notes: We trim the top and bottom 0.5% of the observations. EPS values are in KRW. $P_{T-1}$ is the stock price at the end of the prior fiscal year.

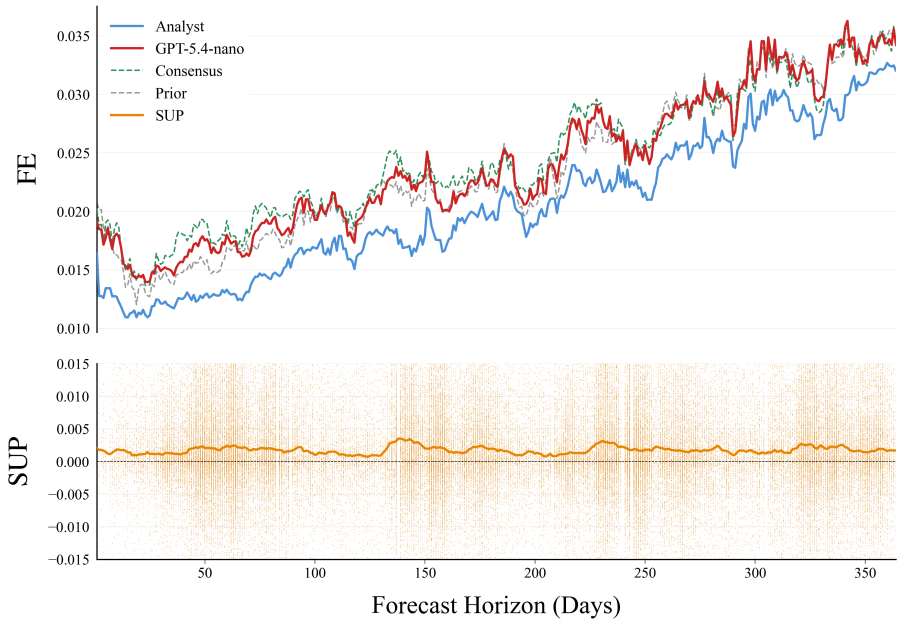



FIGURE 1 Forecast accuracy over the forecast horizon

Notes: The upper panel plots the median FE over the forecast horizon for the analyst, GPT-5.4-nano, the consensus of analysts, and the analyst's prior forecast. The lower panel plots the individual SUP observations (orange dots), along with the 7-day rolling median (orange line).

TABLE 4 Pairwise forecast accuracy comparison: Wilcoxon signed-rank test

| | Analyst | LLM | Consensus | Prior |
|---|---|---|---|---|
| Analyst | - | -99.0*** | -94.4*** | -99.1*** |
| LLM | 99.0*** | - | -7.7*** | 14.7*** |
| Consensus | 94.4*** | 7.7*** | - | 24.5*** |
| Prior | 99.1*** | -14.7*** | -24.5*** | - |
| Median FE | 0.0195 | 0.0233 | 0.0244 | 0.0228 |
| Mean FE | 0.0432 | 0.0505 | 0.0531 | 0.0494 |
| $N$ | | 112,140 | | |

Notes: Each cell reports the Wilcoxon $Z$-statistic. $^{***}p < 0.01$. Positive $Z$ indicates the row forecaster tends to have larger errors than the column forecaster. The number of observations is restricted to cases in which both the individual analyst's prior forecast and the consensus forecast are available.

TABLE 5 Regressions of analyst superiority on firm characteristics

| | Equation (3) | | | | | Equation (4) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Size | $-0.0020^{*}$ | $-0.0013$ | $-0.0006$ | $-0.0022^{*}$ | $-0.0028^{*}$ | $-0.0021^{*}$ | $-0.0008$ | $-0.0011$ | $-0.0022$ | $-0.0026^{*}$ |
| | $(-1.89)$ | $(-0.83)$ | $(-0.39)$ | $(-1.70)$ | $(-1.95)$ | $(-1.94)$ | $(-0.44)$ | $(-0.62)$ | $(-1.49)$ | $(-1.83)$ |
| BtoM | 0.0011 | 0.0024 | 0.0017 | $-0.0001$ | $-0.0005$ | 0.0007 | 0.0017 | 0.0012 | $-0.0001$ | $-0.0009$ |
| | (0.94) | (1.48) | (0.97) | $(-0.14)$ | $(-0.32)$ | (0.59) | (0.88) | (0.65) | $(-0.13)$ | $(-0.65)$ |
| Leverage | $0.0123^{***}$ | 0.0070 | 0.0093 | $0.0130^{**}$ | $0.0127^{**}$ | $0.0102^{**}$ | 0.0029 | 0.0080 | 0.0102 | $0.0130^{*}$ |
| | (2.71) | (0.96) | (1.45) | (2.19) | (2.17) | (1.99) | (0.35) | (1.06) | (1.49) | (1.90) |
| Loss | $0.0091^{***}$ | $0.0195^{***}$ | $0.0178^{***}$ | $0.0039^{***}$ | $0.0041^{**}$ | $0.0082^{***}$ | $0.0193^{***}$ | $0.0181^{***}$ | $0.0033^{***}$ | $0.0035^{*}$ |
| | (7.87) | (7.25) | (8.10) | (3.20) | (2.29) | (7.18) | (6.49) | (7.82) | (2.68) | (1.77) |
| DU | $0.1528^{***}$ | $0.2525^{***}$ | $0.1597^{**}$ | 0.0767 | $0.2074^{***}$ | $0.1220^{***}$ | $0.2105^{***}$ | 0.1128 | 0.0500 | $0.2129^{***}$ |
| | (4.82) | (4.38) | (2.56) | (1.60) | (3.30) | (3.82) | (3.51) | (1.63) | (0.96) | (3.12) |
| Horizon | 0.0000 | – | – | – | – | 0.0001 | – | – | – | – |
| | (0.13) | | | | | (0.45) | | | | |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst×Firm FE | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.148 | 0.244 | 0.188 | 0.136 | 0.161 | 0.196 | 0.274 | 0.201 | 0.168 | 0.217 |
| $N$ | 108,205 | 28,213 | 26,354 | 28,264 | 25,271 | 106,822 | 26,410 | 24,632 | 26,287 | 23,365 |

Notes: $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors are clustered at the firm level. Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns). $h$=$k$q denotes the subsample of forecasts with a target year-end $k$ quarters away, so $h$=1q corresponds to forecasts issued closest to fiscal year-end. SUP is trimmed at the 0.5% tails on the full sample before splitting by quarter.

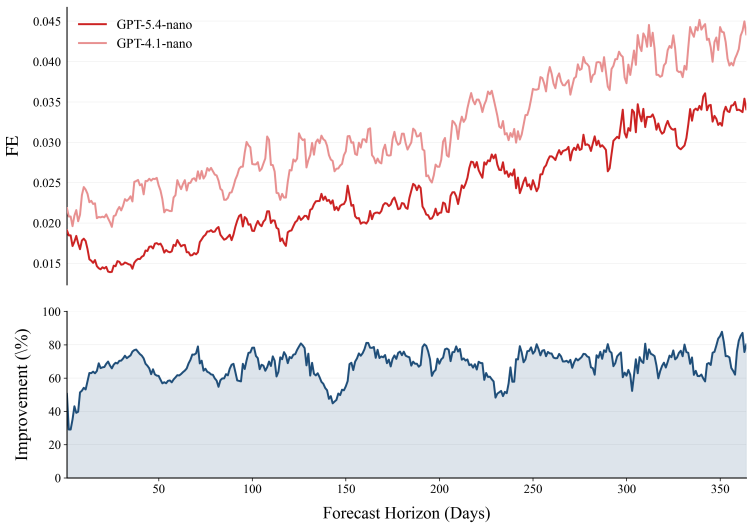



FIGURE 2 Forecast accuracy across LLM generations

Notes: The upper panel plots the median FE over the forecast horizon for GPT-5.4-nano and GPT-4.1-nano. The lower panel plots the median RGR (relative gap reduction, %) = $(FE^{4.1} - FE^{5.4})/(FE^{4.1} - FE^{A}) \times 100$, smoothed with a 7-day centered rolling mean. The sample is trimmed at the 0.5% tails of both $\text{SUP}_{5.4}$ and $\text{SUP}_{4.1}$ ($N = 119{,}400$).

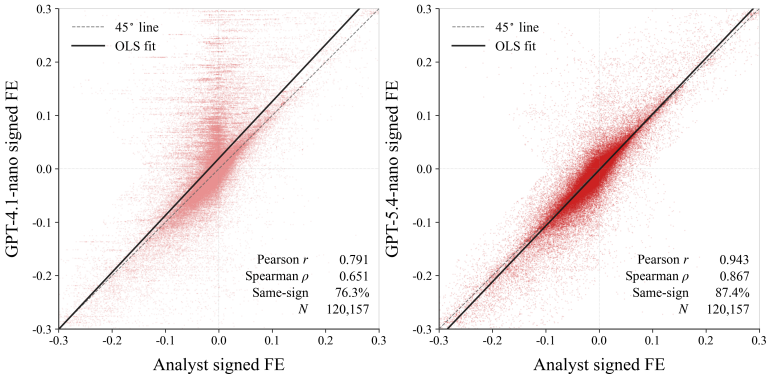



FIGURE 3 Forecast bias: analyst vs. LLM

Notes: Signed forecast errors for the analysts (x-axis) plotted against GPT-4.1-nano (left) and GPT-5.4-nano (right). SUP is trimmed at the 0.5% tails separately for each model ($N = 120{,}157$ for each scatter).

# A Appendix

## A.1 Model Configuration

Table A1 summarizes the API settings for both models, accessed through the OpenAI Responses API. All other parameters available in the API configuration are set to their default values.

TABLE A1 Model Configuration

| Parameter | GPT-4.1-nano | GPT-5.4-nano | Description |
|---|---|---|---|
| API | OpenAI Responses API | OpenAI Responses API | API endpoint |
| temperature | 1.00 | — | Sampling randomness |
| top_p | 1.00 | — | Sampling probability cutoff |
| reasoning.effort | — | medium | Internal reasoning depth |
| verbosity | — | medium | Response verbosity |
| store | false | false | API-side logging of the call |
| JSON schema | enforced | enforced | Structured numeric output |

## A.2 Prompt example

Each LLM prediction is generated from a structured prompt consisting of a developer message (system prompt) and a prompt message (user prompt). All data inputs are filtered to exclude information published on or after the forecast date. Placeholders shown in {blue braces} denote observation-specific values that are filled in for each forecast; the remaining text is fixed across all observations.

**DEVELOPER MESSAGE (SYSTEM PROMPT)**

```
You are an equity analyst at {brokerage firm}, producing a forecast in
{forecast date}.
Your most covered companies: {top covered companies}
Your most covered sectors: {top covered sectors}
Do not reference or rely on any events or developments that occurred
after this date.
Do not search for real data or base your answers on actual economic
figures or official statistics.
```

**PROMPT MESSAGE (USER PROMPT)**

```
You will be provided with financial and macroeconomic data below.
[0] COMPANY OVERVIEW
Company: {company name} ({firm code})
Sector: {sector name} (KSIC {KSIC code})
Market: {market}
```

Market Cap ({last trading day < forecast date}): {market cap} (KRW Millions)

[1a] DOMESTIC MACROECONOMIC DATA

{quarterly values of GDP YoY, CPI YoY, Unemployment, Employment, Current Account --- 12 most recent quarters published before forecast date}

--- BOK Forecast (published {latest publication date < forecast date}) ---

{GDP, CPI, Unemployment, Employment, CA forecasts for current and next fiscal year}

[1b] MARKET DATA

{quarterly values of Dubai oil, World Trade, World Industrial Production, KRW/USD-JPY-CNY exchange rates, AA-/BBB- credit spreads --- 12 most recent quarters before forecast date}

As of {last trading day < forecast date}: {daily values of Dubai oil, KRW/USD, AA- and BBB- credit spreads}

[2] FINANCIAL STATEMENTS (KRW Millions)

{Revenue, Operating income, Profit (loss), Total assets, Total equity, Total liabilities, and other IFRS items --- 3 most recent fiscal years with publication date < forecast date}

[3] STOCK PRICE (Monthly Close, KRW)

{monthly closing price and trading volume --- 36 most recent months before forecast date}

[4] EPS DATA (as of {last trading day < forecast date})

{Target Year, Avg Shares, Net Income, Actual EPS, Consensus EPS, # Analysts --- 5 fiscal years (t−4 through target year)}

[5] YOUR PAST FORECASTS FOR THIS COMPANY

{Forecast Date, Target Year, EPS (KRW) --- analyst's own prior forecasts for same firm-target year before forecast date}

Predict the FY{target year} annual EPS for the following company. Respond with ONLY a single number in KRW.

### A.3 Knowledge cutoff analysis

Table A2 reports the median and mean analyst superiority for four subgroups based on whether the target year falls before or after each model's training cutoff.

TABLE A2 Knowledge cutoff analysis

| | $N$ | Median SUP | Mean SUP | SUP $> 0$ (%) |
|---|---|---|---|---|
| *GPT-4.1-nano (training cutoff: 2024-06-30)* | | | | |
| Pre-cutoff target years (2018–2023) | 86,804 | 0.0045 | 0.0209 | 66.3 |
| Post-cutoff target years (2024–2025) | 33,353 | 0.0047 | 0.0227 | 64.3 |
| *GPT-5.4-nano (training cutoff: 2025-08-31)* | | | | |
| Pre-cutoff target years (2018–2024) | 103,171 | 0.0019 | 0.0073 | 62.3 |
| Post-cutoff target years (2025) | 16,986 | 0.0017 | 0.0077 | 57.8 |

Notes: SUP is trimmed at the 0.5% tails within each model separately. Pre-cutoff and post-cutoff target years are defined according to whether the corresponding realized EPS predates the model's training cutoff. SUP $> 0$ (%) denotes the share of paired analyst–LLM forecasts for which analyst superiority is positive.

### A.4 Construction of firm-level disagreement uncertainty

We construct $DU$ as in (8). For each firm $f$, let $y^C_{t,i}$ denote analyst $i$'s within-quarter mean EPS forecast for the current fiscal year, issued at time $t$, and $y^N_{t,i}$ the corresponding forecast for the next fiscal year. The transformed one-year-ahead forecast is constructed as a weighted average of the two:

$$\hat{y}_{t,i} = w_1\, y^C_{t,i} + w_2\, y^N_{t,i}, \tag{7}$$

where the weights depend on the timing of the forecast within the current calendar year and satisfy $w_1 + w_2 = 1$. For quarterly data,

$$w_1 = \frac{5 - q_t}{4}, \qquad w_2 = 1 - w_1, \tag{8}$$

where $q_t \in \{1, 2, 3, 4\}$ indicates the calendar quarter during which the forecast is made. The weights correspond to how much of the next twelve-month period lies within the current calendar year at the time the forecast is issued.

$DU$ for firm $f$ at time $t$ is the cross-analyst sample standard deviation of $\hat{y}$, scaled by the firm's quarter-end closing price $P_{f,q}$:

$$DU_{f,q} = \frac{1}{P_{f,q}} \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \big(\hat{y}_{t,i} - \bar{\hat{y}}_t\big)^2}, \qquad \bar{\hat{y}}_t = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_{t,i}. \tag{9}$$

## A.5 Robustness checks

TABLE A3 Regressions of analyst superiority on firm characteristic: Equation (5)

| | Firm clustered s.e. | | | | | Firm and year clustered s.e. | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Size | −0.0012 | −0.0004 | 0.0010 | −0.0004 | −0.0042** | −0.0012 | −0.0004 | 0.0010 | −0.0004 | −0.0042* |
| | (−0.85) | (−0.17) | (0.47) | (−0.22) | (−2.12) | (−0.80) | (−0.18) | (0.39) | (−0.18) | (−2.25) |
| BtoM | 0.0011 | 0.0013 | 0.0019 | 0.0011 | −0.0009 | 0.0011 | 0.0013 | 0.0019 | 0.0011 | −0.0009 |
| | (0.69) | (0.52) | (0.86) | (0.90) | (−0.55) | (0.69) | (0.54) | (0.81) | (1.04) | (−0.41) |
| Leverage | 0.0117** | 0.0111 | 0.0167* | 0.0092 | 0.0091 | 0.0117** | 0.0111 | 0.0167* | 0.0092** | 0.0091 |
| | (2.18) | (1.30) | (1.79) | (1.47) | (1.27) | (3.39) | (1.20) | (2.22) | (2.51) | (1.84) |
| Loss | 0.0067*** | 0.0156*** | 0.0163*** | 0.0024** | 0.0019 | 0.0067*** | 0.0156*** | 0.0163*** | 0.0024* | 0.0019 |
| | (6.92) | (5.59) | (7.84) | (2.26) | (1.15) | (6.15) | (6.19) | (8.83) | (2.11) | (1.03) |
| DU | 0.1475*** | 0.2995*** | 0.1448** | 0.0579 | 0.2508*** | 0.1475*** | 0.2995** | 0.1448 | 0.0579 | 0.2508** |
| | (4.38) | (4.02) | (2.27) | (1.09) | (2.92) | (4.45) | (2.68) | (1.86) | (1.17) | (3.21) |
| Horizon | 0.0003 | – | – | – | – | 0.0003 | – | – | – | – |
| | (1.05) | | | | | (0.77) | | | | |
| Analyst×Firm FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry×Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.269 | 0.332 | 0.274 | 0.250 | 0.281 | 0.269 | 0.332 | 0.274 | 0.250 | 0.281 |
| $N$ | 106,821 | 26,408 | 24,630 | 26,286 | 23,365 | 106,821 | 26,408 | 24,630 | 26,286 | 23,365 |

Notes: $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Columns (1)–(5) cluster standard errors at the firm level; columns (6)–(10) use two-way (firm × year) clustering. Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns).

TABLE A4 Regressions of analyst superiority on firm characteristic: Eqs. (3) and (4) with two-way clustered standard errors (firm $\times$ time)

| | Equation (3) | | | | | Equation (4) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Size | $-0.0020^{*}$ | $-0.0013$ | $-0.0006$ | $-0.0022$ | $-0.0028^{*}$ | $-0.0021^{*}$ | $-0.0008$ | $-0.0011$ | $-0.0022$ | $-0.0026$ |
| | $(-1.95)$ | $(-0.81)$ | $(-0.31)$ | $(-1.62)$ | $(-2.19)$ | $(-2.29)$ | $(-0.42)$ | $(-0.57)$ | $(-1.69)$ | $(-1.88)$ |
| BtoM | 0.0011 | 0.0024 | 0.0017 | $-0.0001$ | $-0.0005$ | 0.0007 | 0.0017 | 0.0012 | $-0.0001$ | $-0.0009$ |
| | (0.94) | (1.43) | (1.09) | $(-0.23)$ | $(-0.33)$ | (0.59) | (1.00) | (0.67) | $(-0.15)$ | $(-0.49)$ |
| Leverage | $0.0123^{***}$ | 0.0070 | $0.0093^{*}$ | $0.0130^{***}$ | $0.0127^{**}$ | $0.0102^{***}$ | 0.0029 | 0.0080 | $0.0102^{**}$ | $0.0130^{*}$ |
| | (8.33) | (1.11) | (2.04) | (7.15) | (2.84) | (4.18) | (0.33) | (1.22) | (2.50) | (2.34) |
| Loss | $0.0091^{***}$ | $0.0195^{***}$ | $0.0178^{***}$ | $0.0039^{***}$ | $0.0041^{***}$ | $0.0082^{***}$ | $0.0193^{***}$ | $0.0181^{***}$ | $0.0033^{**}$ | $0.0035^{*}$ |
| | (8.53) | (9.12) | (9.86) | (3.57) | (3.76) | (7.70) | (7.76) | (9.37) | (3.01) | (2.04) |
| DU | $0.1528^{**}$ | $0.2525^{**}$ | $0.1597^{*}$ | 0.0767 | $0.2074^{**}$ | $0.1220^{**}$ | $0.2105^{**}$ | 0.1128 | 0.0500 | $0.2129^{**}$ |
| | (3.49) | (3.40) | (1.99) | (1.38) | (2.79) | (3.35) | (2.82) | (1.37) | (0.99) | (2.91) |
| Horizon | 0.0000 | – | – | – | – | 0.0001 | – | – | – | – |
| | (0.11) | | | | | (0.37) | | | | |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst$\times$Firm FE | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.148 | 0.244 | 0.188 | 0.136 | 0.161 | 0.196 | 0.274 | 0.201 | 0.168 | 0.217 |
| $N$ | 108,205 | 28,213 | 26,354 | 28,264 | 25,271 | 106,822 | 26,410 | 24,632 | 26,287 | 23,365 |

Notes: $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors are clustered at the firm and year level (two-way). Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns).

TABLE A5 Regressions of analyst superiority on firm characteristic: Eqs (3), (4), and (5) with two-way clustered standard errors (firm $\times$ analyst)

| | Equation (3) | | | | | Equation (4) | | | | | Equation (5) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) | (14) | (15) |
| Size | $-0.0020^{*}$ | $-0.0013$ | $-0.0006$ | $-0.0022^{*}$ | $-0.0028^{*}$ | $-0.0021^{*}$ | $-0.0008$ | $-0.0011$ | $-0.0022$ | $-0.0026^{*}$ | $-0.0012$ | $-0.0004$ | 0.0010 | $-0.0004$ | $-0.0042^{**}$ |
| | $(-1.91)$ | $(-0.85)$ | $(-0.38)$ | $(-1.70)$ | $(-1.94)$ | $(-1.93)$ | $(-0.44)$ | $(-0.61)$ | $(-1.50)$ | $(-1.84)$ | $(-0.87)$ | $(-0.17)$ | (0.47) | $(-0.22)$ | $(-2.14)$ |
| BtoM | 0.0011 | 0.0024 | 0.0017 | $-0.0001$ | $-0.0005$ | 0.0007 | 0.0017 | 0.0012 | $-0.0001$ | $-0.0009$ | 0.0011 | 0.0013 | 0.0019 | 0.0011 | $-0.0009$ |
| | (0.94) | (1.47) | (0.96) | $(-0.14)$ | $(-0.32)$ | (0.59) | (0.85) | (0.63) | $(-0.13)$ | $(-0.62)$ | (0.69) | (0.52) | (0.84) | (0.96) | $(-0.54)$ |
| Leverage | $0.0123^{***}$ | 0.0070 | 0.0093 | $0.0130^{**}$ | $0.0127^{**}$ | $0.0102^{*}$ | 0.0029 | 0.0080 | 0.0102 | $0.0130^{*}$ | $0.0117^{**}$ | 0.0111 | $0.0167^{*}$ | 0.0092 | 0.0091 |
| | (2.67) | (0.97) | (1.46) | (2.15) | (2.09) | (1.95) | (0.35) | (1.08) | (1.49) | (1.80) | (2.12) | (1.29) | (1.77) | (1.47) | (1.25) |
| Loss | $0.0091^{***}$ | $0.0195^{***}$ | $0.0178^{***}$ | $0.0039^{***}$ | $0.0041^{**}$ | $0.0082^{***}$ | $0.0193^{***}$ | $0.0181^{***}$ | $0.0033^{***}$ | $0.0035^{*}$ | $0.0067^{***}$ | $0.0156^{***}$ | $0.0163^{***}$ | $0.0024^{**}$ | 0.0019 |
| | (7.85) | (7.35) | (8.21) | (3.20) | (2.35) | (7.12) | (6.50) | (7.85) | (2.67) | (1.80) | (6.83) | (5.57) | (7.80) | (2.26) | (1.15) |
| DU | $0.1528^{***}$ | $0.2525^{***}$ | $0.1597^{**}$ | 0.0767 | $0.2074^{***}$ | $0.1220^{***}$ | $0.2105^{***}$ | 0.1128 | 0.0500 | $0.2129^{***}$ | $0.1475^{***}$ | $0.2995^{***}$ | $0.1448^{**}$ | 0.0579 | $0.2508^{***}$ |
| | (4.62) | (4.39) | (2.57) | (1.47) | (3.22) | (3.65) | (3.47) | (1.63) | (0.87) | (3.06) | (4.13) | (3.87) | (2.24) | (1.00) | (2.85) |
| Horizon | 0.0001 | – | – | – | – | 0.0002 | – | – | – | – | 0.0003 | – | – | – | – |
| | (0.16) | | | | | (0.49) | | | | | (1.11) | | | | |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Firm FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No | No | No | No | No | No |
| Analyst FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No | No | No | No | No | No |
| Analyst$\times$Firm FE | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry$\times$Time FE | No | No | No | No | No | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.148 | 0.244 | 0.188 | 0.136 | 0.161 | 0.196 | 0.274 | 0.201 | 0.168 | 0.217 | 0.269 | 0.332 | 0.274 | 0.250 | 0.281 |
| $N$ | 108,205 | 28,213 | 26,354 | 28,264 | 25,271 | 106,822 | 26,410 | 24,632 | 26,287 | 23,365 | 106,821 | 26,408 | 24,630 | 26,286 | 23,365 |

Notes: $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors are clustered at the firm and analyst level (two-way). Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns).

## A.6 Replication of main results using GPT-4.1-nano

TABLE A6 Regressions of analyst superiority on firm characteristic: GPT-4.1-nano, Eqs. (3) and (4)

| | Equation (3) | | | | | Equation (4) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Size | −0.0001 | 0.0034 | 0.0027 | −0.0006 | −0.0063** | −0.0034 | 0.0008 | 0.0007 | −0.0031 | −0.0106*** |
| | (−0.05) | (0.91) | (0.78) | (−0.23) | (−2.23) | (−1.16) | (0.18) | (0.18) | (−0.97) | (−3.55) |
| BtoM | −0.0008 | −0.0019 | −0.0022 | −0.0005 | −0.0021 | −0.0003 | −0.0022 | −0.0015 | 0.0000 | −0.0033 |
| | (−0.32) | (−0.52) | (−0.78) | (−0.26) | (−0.75) | (−0.13) | (−0.49) | (−0.45) | (0.02) | (−1.16) |
| Leverage | 0.0214** | 0.0083 | 0.0009 | 0.0408*** | 0.0217 | 0.0253** | 0.0074 | 0.0064 | 0.0418*** | 0.0333** |
| | (2.02) | (0.55) | (0.06) | (3.15) | (1.57) | (1.97) | (0.35) | (0.31) | (3.24) | (2.10) |
| Loss | 0.0063*** | 0.0165*** | 0.0132*** | 0.0001 | 0.0035 | 0.0061*** | 0.0157*** | 0.0151*** | 0.0011 | 0.0031 |
| | (3.98) | (5.48) | (4.17) | (0.03) | (1.15) | (4.10) | (4.52) | (5.15) | (0.63) | (1.05) |
| DU | 0.4108*** | 0.7770*** | 0.5922*** | 0.2761*** | 0.3095*** | 0.3429*** | 0.7341*** | 0.5235** | 0.2493** | 0.2539** |
| | (5.39) | (5.24) | (3.09) | (3.08) | (3.05) | (4.26) | (4.71) | (2.50) | (2.53) | (2.18) |
| Horizon | −0.0010* | – | – | – | – | −0.0013** | – | – | – | – |
| | (−1.69) | | | | | (−2.06) | | | | |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst×Firm FE | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.167 | 0.244 | 0.198 | 0.158 | 0.159 | 0.236 | 0.272 | 0.259 | 0.217 | 0.230 |
| $N$ | 108,272 | 28,291 | 26,400 | 28,258 | 25,217 | 106,889 | 26,478 | 24,670 | 26,295 | 23,311 |

Notes: Dependent variable is $\text{SUP}^{41} = FE^{LLM\text{-}4.1} - FE^{Analyst}$. $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors are clustered at the firm level. Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns).

TABLE A7 Regressions of analyst superiority on firm characteristic: GPT-4.1-nano, Eqs. (3) and (4) with two-way clustered standard errors

| | Equation (3) | | | | | Equation (4) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Size | $-0.0001$ | 0.0034 | 0.0027 | $-0.0006$ | $-0.0063^{*}$ | $-0.0034$ | 0.0008 | 0.0007 | $-0.0031$ | $-0.0106^{***}$ |
| | $(-0.04)$ | (0.79) | (0.65) | $(-0.22)$ | $(-2.00)$ | $(-1.09)$ | (0.16) | (0.20) | $(-1.17)$ | $(-3.53)$ |
| BtoM | $-0.0008$ | $-0.0019$ | $-0.0022$ | $-0.0005$ | $-0.0021$ | $-0.0003$ | $-0.0022$ | $-0.0015$ | 0.0000 | $-0.0033$ |
| | $(-0.31)$ | $(-0.50)$ | $(-0.74)$ | $(-0.30)$ | $(-0.63)$ | $(-0.13)$ | $(-0.45)$ | $(-0.53)$ | (0.03) | $(-1.15)$ |
| Leverage | $0.0214^{**}$ | 0.0083 | 0.0009 | $0.0408^{**}$ | 0.0217 | $0.0253^{*}$ | 0.0074 | 0.0064 | $0.0418^{***}$ | $0.0333^{*}$ |
| | (2.39) | (0.66) | (0.06) | (3.33) | (1.57) | (2.17) | (0.38) | (0.30) | (3.92) | (2.22) |
| Loss | $0.0063^{***}$ | $0.0165^{***}$ | $0.0132^{***}$ | 0.0001 | 0.0035 | $0.0061^{***}$ | $0.0157^{***}$ | $0.0151^{***}$ | 0.0011 | 0.0031 |
| | (3.63) | (5.82) | (4.31) | (0.04) | (1.00) | (3.56) | (5.22) | (5.37) | (0.67) | (0.86) |
| DU | $0.4108^{***}$ | $0.7770^{***}$ | $0.5922^{**}$ | $0.2761^{**}$ | $0.3095^{***}$ | $0.3429^{***}$ | $0.7341^{***}$ | $0.5235^{**}$ | $0.2493^{**}$ | $0.2539^{**}$ |
| | (4.97) | (4.06) | (2.64) | (3.19) | (4.08) | (3.83) | (3.64) | (2.42) | (2.57) | (2.57) |
| Horizon | $-0.0010^{**}$ | – | – | – | – | $-0.0013^{**}$ | – | – | – | – |
| | $(-2.69)$ | | | | | $(-2.71)$ | | | | |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst FE | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No |
| Analyst×Firm FE | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.167 | 0.244 | 0.198 | 0.158 | 0.159 | 0.236 | 0.272 | 0.259 | 0.217 | 0.230 |
| $N$ | 108,272 | 28,291 | 26,400 | 28,258 | 25,217 | 106,889 | 26,478 | 24,670 | 26,295 | 23,311 |

Notes: Dependent variable is $\text{SUP}^{41} = FE^{LLM\text{-}4.1} - FE^{Analyst}$. $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors are clustered at the firm and year level (two-way). Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns).

TABLE A8 Regressions of analyst superiority on firm characteristic: GPT-4.1-nano: Equation (5)

| | Firm clustered s.e. | | | | | Firm and year clustered s.e. | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q | All | $h$=1q | $h$=2q | $h$=3q | $h$=4q |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Size | $-0.0032$ | 0.0001 | 0.0016 | $-0.0026$ | $-0.0126^{***}$ | $-0.0032$ | 0.0001 | 0.0016 | $-0.0026$ | $-0.0126^{**}$ |
| | $(-1.14)$ | (0.03) | (0.48) | $(-0.92)$ | $(-3.50)$ | $(-0.81)$ | (0.02) | (0.37) | $(-0.78)$ | $(-3.35)$ |
| BtoM | 0.0004 | $-0.0004$ | $-0.0002$ | 0.0008 | $-0.0039$ | 0.0004 | $-0.0004$ | $-0.0002$ | 0.0008 | $-0.0039$ |
| | (0.13) | $(-0.09)$ | $(-0.06)$ | (0.43) | $(-1.27)$ | (0.14) | $(-0.09)$ | $(-0.08)$ | (0.38) | $(-1.38)$ |
| Leverage | $0.0248^{**}$ | 0.0090 | 0.0200 | $0.0332^{**}$ | $0.0356^{**}$ | 0.0248 | 0.0090 | 0.0200 | $0.0332^{**}$ | $0.0356^{*}$ |
| | (1.97) | (0.47) | (1.09) | (2.25) | (2.12) | (1.86) | (0.43) | (0.98) | (2.48) | (2.18) |
| Loss | $0.0056^{***}$ | $0.0135^{***}$ | $0.0142^{***}$ | 0.0010 | 0.0015 | $0.0056^{**}$ | $0.0135^{***}$ | $0.0142^{***}$ | 0.0010 | 0.0015 |
| | (4.02) | (3.64) | (4.95) | (0.61) | (0.62) | (3.16) | (3.95) | (5.86) | (0.68) | (0.38) |
| DU | $0.3454^{***}$ | $0.7299^{***}$ | $0.4907^{***}$ | $0.2814^{**}$ | $0.2908^{**}$ | $0.3454^{***}$ | $0.7299^{***}$ | $0.4907^{**}$ | $0.2814^{**}$ | $0.2908^{**}$ |
| | (5.44) | (5.63) | (3.57) | (2.49) | (2.36) | (5.15) | (4.90) | (3.12) | (2.78) | (2.85) |
| Horizon | $-0.0013^{**}$ | – | – | – | – | $-0.0013^{**}$ | – | – | – | – |
| | $(-2.10)$ | | | | | $(-2.42)$ | | | | |
| Analyst×Firm FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry×Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.301 | 0.348 | 0.324 | 0.286 | 0.274 | 0.301 | 0.348 | 0.324 | 0.286 | 0.274 |
| $N$ | 106,889 | 26,476 | 24,670 | 26,294 | 23,311 | 106,889 | 26,476 | 24,670 | 26,294 | 23,311 |

Notes: Dependent variable is $\text{SUP}^{41} = FE^{LLM\text{-}4.1} - FE^{Analyst}$. $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Columns (1)–(5) cluster standard errors at the firm level; columns (6)–(10) use two-way (firm × year) clustering. Time fixed effect is year-quarter fixed effect. Horizon is $\ln(1 + \text{Days to year-end})$ and is included only in pooled regressions (All columns).