Large Language Models in Equity Research

Nicholas Wong
Massachusetts Institute of Technology
nicwjh@mit.edu

Yilu Pan
Massachusetts Institute of Technology
yilup666@mit.edu

Xizhi Fang
Massachusetts Institute of Technology
fangx919@mit.edu

Christelle Saad
Massachusetts Institute of Technology
csaad@mit.edu

Abstract

We study whether large language models can enhance fundamental equity research by analyzing financial statements and predicting future earnings. We prompt three open-source models to perform chain-of-thought analysis of anonymized balance sheets and income statements for Russell 3000 firms from 2010 to 2023, generating structured forecasts and narrative reasoning about each company. We encode the generated text as sentence embeddings and combine them with traditional numeric features in gradient-boosted classifiers to predict earnings changes and earnings surprises. Using these predictions as signals, we construct monthly-rebalanced long-short portfolios and evaluate performance against an equal-weighted benchmark. LLM-derived narrative embeddings improve both classification accuracy and portfolio alpha, with the largest gains appearing when paired with financial statement data rather than precomputed factor characteristics. The best-performing strategy achieves an annualized information ratio above 1.0 and consistently outperforms analyst consensus forecasts. Results hold across prediction lags of one to three months.

1 Introduction

A central task in equity research is forecasting a company's future earnings from its financial statements. Sell-side analysts have long filled this role, but their forecasts are costly to produce, limited in coverage, and subject to well-documented behavioral biases. The recent emergence of large language models capable of processing and reasoning about numerical data opens up the possibility that these models could replicate or improve upon human financial analysis at scale.

Kim, Muhn, and Nikolaev (2024) provided initial evidence that this is feasible. They showed that GPT-4, when prompted with anonymized financial statements and instructed to follow a chain-of-thought reasoning process, can predict the direction of future earnings with accuracy comparable to human analysts. Portfolios constructed from these predictions generated significant alpha in the Fama-French three-factor model. Their results suggest that large language models possess an emergent ability to interpret accounting data and extract economically meaningful signals.

Their analysis, however, relied exclusively on GPT-4, a proprietary model accessed through a paid API, and it remains unclear whether smaller, open-source models can achieve comparable performance. They also evaluated LLM predictions as standalone signals, whereas quantitative investors in practice combine many sources of information, making the more relevant question whether LLM outputs provide incremental value beyond traditional quantitative features. Their portfolio construction used annual rebalancing based on a single model's discrete predictions, and a more realistic evaluation would incorporate continuous signals, higher-frequency rebalancing, and comparison across multiple feature sets.

In this work, we address these gaps. We prompt three open-source models (Meta-Llama-3-8B-Instruct, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-14B) to analyze anonymized balance sheets and income statements for Russell 3000 constituents over the period 2010 to 2023. Following Kim et al. (2024), we use chain-of-thought prompting to elicit structured outputs: trend analysis, ratio analysis, a directional earnings forecast, an estimated magnitude, and a confidence level. Beyond the discrete predictions, each model generates rich narrative text explaining its reasoning.

We extract value from these narratives by encoding the trend analysis, ratio analysis, and rationale sections as dense vector representations using a pretrained sentence transformer. These embeddings, together with the discrete LLM outputs, serve as additional features in XGBoost classifiers trained to predict two targets: the direction of earnings per share changes and the direction of earnings surprises relative to analyst consensus. We evaluate four feature sets that vary the inclusion of 153 monthly asset pricing factors from Jensen, Kelly, and Pedersen (2023), raw financial statement items, and LLM-derived signals. All models are trained using rolling three-year windows and evaluated out of sample across prediction lags of one to three months.

Using the classifier predictions as ranking signals, we construct monthly-rebalanced long-short tercile portfolios and measure alpha relative to the equal-weighted universe return. Our main findings are threefold. LLM-derived narrative embeddings improve classification accuracy and portfolio performance when added to either factor-based or financial-statement-based feature sets, and this improvement is substantially larger when LLM signals are paired with raw financial statement data. We interpret this as evidence that the models provide complementary economic reasoning that raw accounting variables alone do not capture but that is partially redundant with information already summarized in standard factors. The best-performing strategy, combining financial statement features with LLM signals to predict earnings surprises, achieves an annualized long-minus-short information ratio above 1.0 and consistently outperforms portfolios based on analyst consensus forecasts across all configurations and prediction lags of one to three months.

2 Related Work

2.1 Financial Statement Analysis and Earnings Prediction

The use of financial statement data to predict future earnings and stock returns has a long history. Ou and Penman [5] constructed 68 financial ratios and combined them into a probability score using logistic regression, showing that this score predicted future earnings changes and that the market did not fully incorporate the information in these ratios. Chen et al. [1] extended this line of work by applying random forests and gradient boosting to detailed financial data from XBRL filings. Their models outperformed both regression-based methods and professional analyst forecasts in predicting the direction of earnings changes, demonstrating the value of capturing nonlinear interactions among accounting variables. We adopt gradient-boosted classifiers as our primary prediction framework for the same reasons, but augment the feature space with signals derived from large language models.

2.2 Large Language Models in Finance

Recent work has brought large language models into financial analysis. Kim, Muhn, and Nikolaev [4] showed that GPT-4, prompted with anonymized financial statements and chain-of-thought instructions, can predict the direction of future earnings with accuracy comparable to human analysts and specialized machine learning models. Portfolios formed on these predictions generated significant alpha. In a separate study, Kim, Muhn, and Nikolaev [3] demonstrated that LLM-generated summaries of corporate disclosures retain economically relevant information while reducing document length substantially, with summary-based sentiment explaining stock price movements better than full-text sentiment. Yang et al. [7] proposed FinRobot, an open-source platform that integrates LLMs into equity research workflows using retrieval-augmented generation. We depart from these studies in two ways: we use open-source models rather than proprietary APIs, improving reproducibility and scalability, and we encode the models' generated narratives as dense embeddings to test whether they provide incremental predictive power when combined with traditional quantitative features rather than evaluating LLM predictions in isolation.

2.3 LLM Reliability and Calibration

Yoo [6] examined the reliability of self-reported confidence scores from large language models in accounting and finance classification tasks. Chain-of-thought prompting improved failure prediction but worsened calibration and increased overconfidence. Repeated prompting and fine-tuning offered partial remedies, though each involved tradeoffs between accuracy, calibration, and sensitivity to cross-sectional variation. These findings motivate our decision to treat the discrete LLM outputs (direction, magnitude, confidence) as features within a supervised learning framework rather than as direct trading signals, allowing the downstream classifier to learn the appropriate weighting.

2.4 Asset Pricing Factors

Jensen, Kelly, and Pedersen [2] organized 153 asset pricing factors into 13 economically meaningful themes and demonstrated an 82.4% replication rate across a global dataset covering 93 countries. Their comprehensive factor set provides a strong baseline for cross-sectional return prediction. We use their monthly stock-level characteristics as one of our primary feature sets, enabling us to test whether LLM-derived signals contain information beyond what is already captured by well-established asset pricing variables.

3 Methods

3.1 Data

Our sample covers Russell 3000 constituents over the period 2010 to 2023 and draws on three data sources. Annual balance sheets and income statements come from WRDS Compustat, filtered for firms with non-missing total assets and stock prices. These financial statements serve as both inputs to the LLM prompts and features in our classification models. We also use the 153 monthly stock-level characteristics constructed by Jensen et al. [2], organized into 13 themes including value, momentum, profitability, and size, downloaded from WRDS Contributed Data Forms. These serve as an alternative feature set representing the current state of cross-sectional asset pricing knowledge. Analyst earnings forecasts from WRDS IBES, including median consensus estimates, actual reported earnings, and announcement dates, provide the earnings surprise target variable and a benchmark signal for portfolio evaluation.

We construct two binary classification targets. The first, `y_epschg_bin`, indicates whether a firm's earnings per share increased relative to the prior year. The second, `y_surprise_bin`, indicates whether reported earnings exceeded the analyst consensus forecast. Both reduce the problem to directional classification, which aligns with practical investment strategies and sidesteps the well-known difficulty of predicting continuous earnings magnitudes [1].

3.2 LLM Prediction Pipeline

We prompt three open-source language models to analyze anonymized financial statements and predict the direction of future earnings. The models are Meta-Llama-3-8B-Instruct, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-14B. The two DeepSeek variants are knowledge-distilled from the larger DeepSeek-R1 model, which has been shown to achieve strong performance on reasoning benchmarks.

Following Kim et al. [4], we use a chain-of-thought prompting strategy. Each prompt instructs the model to assume the role of a financial analyst, then proceed through three steps: (1) identify notable trends in the financial statement items, (2) compute and interpret key financial ratios, and (3) synthesize these findings into a directional earnings prediction with an associated magnitude (small, moderate, or large) and confidence level (low, moderate, or high). Financial statements are anonymized by removing company names and fiscal years to prevent the model from relying on memorized information. The full prompt template is provided in Appendix A.

We evaluate the standalone predictive accuracy of the LLM outputs against both realized EPS changes and realized earnings surprises. When measured against actual EPS changes, all three models achieve accuracy near 50%, roughly equivalent to the naive baseline. Against earnings surprises, accuracy rises above 60% for all models, indicating that the LLMs capture information about the direction of earnings relative to market expectations. There is no significant difference in accuracy across the three models. Composite signals formed by interacting the directional prediction with either magnitude or confidence exhibit positive and statistically significant Spearman rank correlations with realized earnings surprises (0.25 to 0.34 across models), confirming that the LLM outputs contain economically meaningful variation. Cross-model prediction correlations range from 0.42 to 0.54, suggesting that the models capture partially distinct information despite receiving identical inputs.

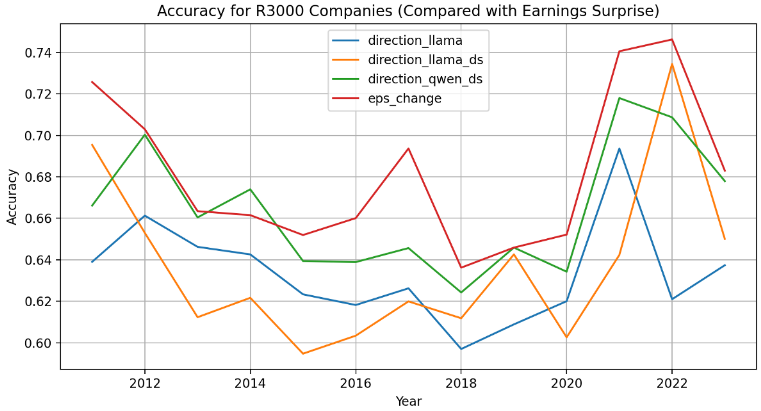

Figure 1: LLM prediction accuracy compared with earnings surprise by year, for each of the three models and the realized EPS change baseline. Accuracy is consistently above 60% across models and years.

3.3 Feature Construction and Embedding

We construct four feature sets to evaluate the incremental contribution of LLM-derived signals. The first two are numeric baselines: (1) the 153 monthly factors from Jensen et al. [2], and (2) the raw balance sheet and income statement items from Compustat. The remaining two augment each baseline with LLM-derived features: (3) monthly factors combined with LLM signals, and (4) balance sheet and income statement items combined with LLM signals.

The LLM features consist of two components. The first is the set of discrete outputs from each model: the predicted direction, magnitude, and confidence level. The second is a set of dense vector representations of the generated text. Specifically, we encode the trend analysis, ratio analysis, and rationale sections from each model's output using the `all-MiniLM-L6-v2` sentence transformer, which maps each text block to a fixed-length vector of dimension 384. Sentence embeddings capture the semantic content of the model's financial reasoning in a form that can be directly combined with numeric features in a supervised learning framework. This approach preserves the nuance of the generated narratives while avoiding the noise inherent in token-level representations.

3.4 XGBoost Classification

We train XGBoost classifiers to predict both binary targets using each of the four feature sets. Models are trained using a rolling window scheme: for each evaluation year t, we train on data from years $t-3$ through $t-1$ and predict on year t. This temporal split ensures strict out-of-sample evaluation and avoids lookahead bias. The rolling window is advanced by one year and the process is repeated, yielding predictions for 2013 through 2022.

To account for the delay between the fiscal year end and the availability of financial data, we evaluate models at prediction lags of 1, 2, and 3 months. At each lag, the classifier uses only information that would have been available to an investor at that point in time.

Figure 2 shows classification accuracy by year for the earnings surprise target at a one-month lag. Adding LLM features improves accuracy for both feature bases, and the improvement is substantially larger for the balance sheet features than for the monthly factors. The balance sheet baseline achieves the lowest accuracy among the four feature sets, but adding LLM signals closes much of the gap with the factor-based models. We believe this reflects the fact that the LLM narratives provide economic reasoning that is new relative to raw accounting data but partially overlaps with the information already encoded in standard asset pricing characteristics. Accuracy is stable across lags of 1 to 3 months (confusion matrices are provided in Appendix C).

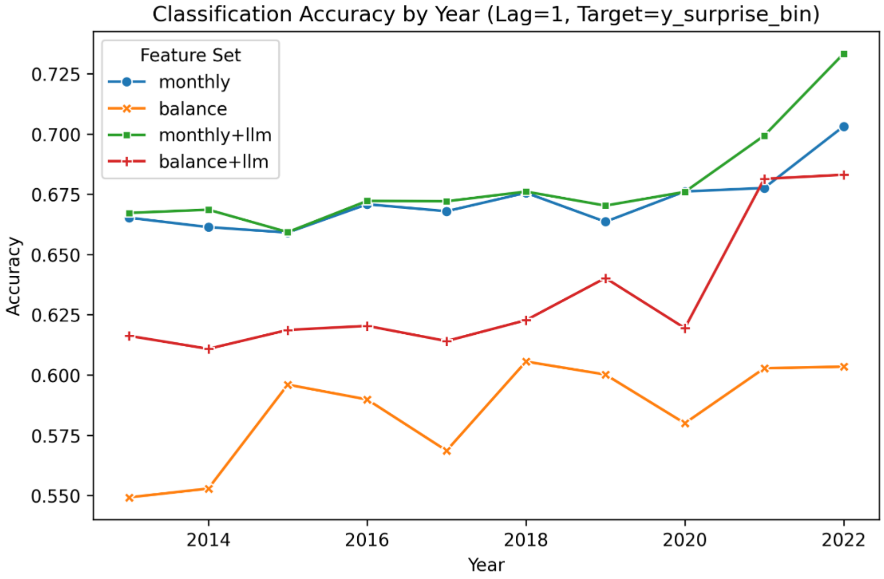

Figure 2: Classification accuracy by year for the earnings surprise target (lag = 1 month) across four feature sets. Adding LLM features improves accuracy for both bases, with a larger gain for balance sheet features.

3.5 Portfolio Construction

We translate the XGBoost predictions into portfolio signals and evaluate their economic value. For each of the eight model configurations (two targets × four feature sets), we rank stocks monthly by their predicted probability of a positive outcome. Stocks in the top tercile form the long portfolio and stocks in the bottom tercile form the short portfolio. All portfolios are equal-weighted and rebalanced monthly. We also construct a ninth signal based on analyst consensus forecasts from IBES as a benchmark.

We define alpha as the difference between the portfolio return and the equal-weighted universe return:

$$\alpha_t = R_t^{\text{portfolio}} - R_t^{\text{universe}}. \tag{1}$$

This measures the cross-sectional return to stock selection, isolating the value of the signal from broad market exposure. We report annualized average alpha and annualized information ratios (mean alpha divided by the standard deviation of alpha) for the long, short, and long-minus-short (HML) portfolios over the evaluation period 2013 to 2022.

4 Results

4.1 Portfolio Performance

Table 1 reports average alpha and information ratios for all nine signals. Long portfolios generate positive alpha across nearly all signals, while short portfolios perform poorly, with several producing slightly positive or negative alpha. This asymmetry suggests that the classification models are better at identifying stocks that will outperform than stocks that will underperform.

Among the long portfolios, the LLM-enhanced signals consistently outperform their non-LLM counterparts. The strongest long-side performance comes from the Surprise (Financial Statements + LLM) signal, which achieves an average monthly alpha of 0.17% and an annualized information ratio of 1.45. The analyst consensus benchmark produces the weakest long-side alpha (0.05%) and is statistically insignificant.

For the long-minus-short portfolios, the two signals combining financial statements with LLM features achieve the highest alphas (0.20% per month) and information ratios above 1.0. Signals based on factors alone or financial statements alone produce HML alphas below 0.10% with information ratios below 0.50.

Table 1: Average monthly alpha (%) and annualized information ratio for long, short, and HML portfolios across all signals. Evaluation period: 2013–2022.

		Average Alpha (%)			Information Ratio		
	Signal	Long	Short	HML	Long	Short	HML
4*EPS Chg	153 Factors	0.12	0.03	0.09	1.09	0.23	0.49
	Financial Statements	0.09	0.00	0.09	0.76	−0.01	0.42
	153 Factors + LLM	0.10	−0.08	0.18	0.91	−0.57	1.08
	Fin. Statements + LLM	0.16	−0.04	0.20	1.28	−0.30	1.08
4*Surprise	153 Factors	0.09	0.04	0.05	0.67	0.22	0.27
	Financial Statements	0.06	0.06	0.00	0.50	0.36	0.01
	153 Factors + LLM	0.14	0.01	0.13	1.01	0.08	0.61
	Fin. Statements + LLM	0.17	−0.02	0.20	1.45	−0.18	1.07
	Analyst Prediction	0.05	0.11	−0.07	0.43	0.73	−0.34

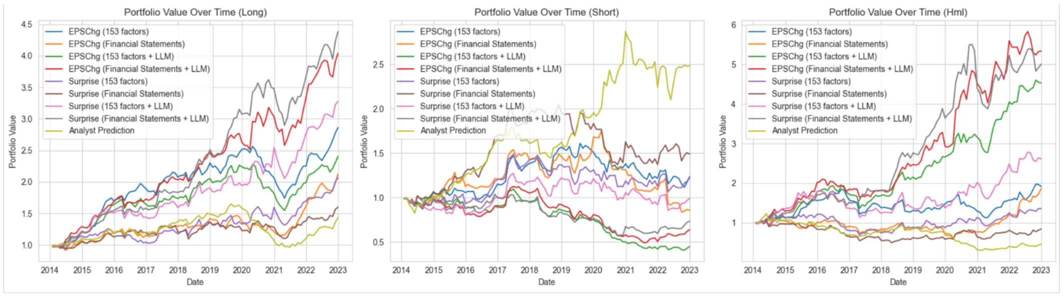

Figure 3: Portfolio value over time (2013–2022) for long, short, and HML portfolios across all signals. LLM-enhanced strategies (particularly those combining financial statements with LLM signals) dominate on the long side. Analyst predictions consistently underperform.

4.2 Statistical Significance

Figure 4 reports t-statistics for the null hypothesis that alpha equals zero. On the long side, LLM-enhanced signals produce the highest t-statistics, with Surprise (Financial Statements + LLM) reaching 4.32 and EPSChg (Financial Statements + LLM) reaching 3.83. Most non-LLM signals are also significant on the long side, though with smaller t-statistics. The analyst benchmark is not statistically significant on the long side ($t = 1.30$).

On the short side, most signals produce insignificant or negative t-statistics. The only significant short-side result is the analyst prediction ($t = 2.18$), which is notable given that analysts produce the weakest long-side performance.

For the HML portfolios, the LLM-enhanced signals are the only ones that achieve consistent significance, with t-statistics above 3.0 for all four LLM-augmented configurations. Non-LLM signals produce HML t-statistics below 1.5, none of which reach significance at conventional levels.

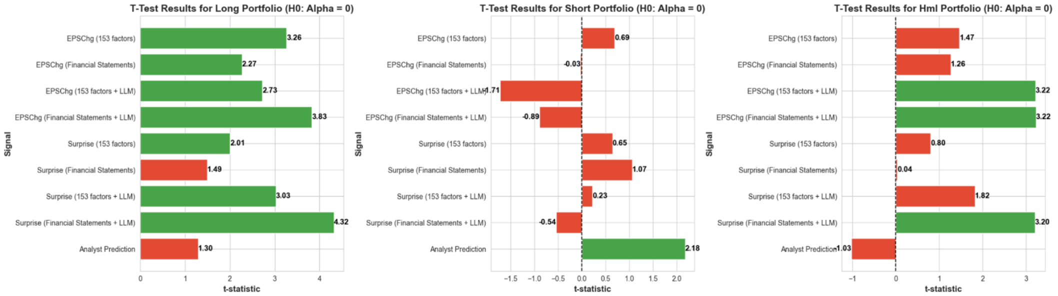

Figure 4: T-statistics for the null hypothesis that alpha equals zero, for long, short, and HML portfolios. Green bars indicate $t > 0$ and red bars indicate $t < 0$. LLM-enhanced signals produce the most significant results on the long side and HML.

4.3 Incremental Value of LLM Signals

To isolate the contribution of LLM features, we compare each signal directly with its LLM-enhanced counterpart. Figure 5 shows the percentage increase in long-side alpha and the absolute increase in information ratio from adding LLM signals. The gains differ sharply across feature bases. For the EPS change target, adding LLM features to factors increases alpha by 16%, while adding them to financial statements increases alpha by 80%. For the surprise target, the gap widens further: a 59% increase for factors versus a 190% increase for financial statements. The information ratio improvements follow the same ordering.

We think the explanation is straightforward. The 153 monthly factors from Jensen et al. [2] already encode processed cross-sectional information, including profitability, accruals, and other characteristics that overlap with what the LLM extracts from financial statements. Raw accounting data contains the same underlying information but in an unprocessed form. The LLM narratives effectively perform a layer of financial analysis that transforms raw accounting numbers into economic interpretations, and this layer adds the most value precisely where such interpretation is otherwise absent.

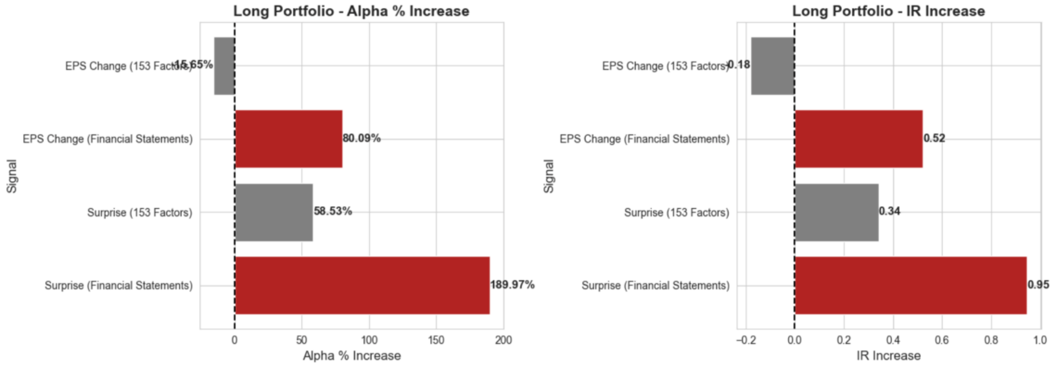

Figure 5: Percentage increase in long-side alpha (left) and absolute increase in information ratio (right) from adding LLM features. Gains are substantially larger when LLM signals are paired with financial statement data (red) than with precomputed factor characteristics (gray).

4.4 Robustness Across Prediction Lags

We evaluate robustness by varying the prediction lag from 1 to 12 months. Figure 6 reports annualized average alpha for all signals across lags of 1, 2, 3, and 12 months. Alpha declines with longer lags for all signals, as expected given that older information is less predictive. LLM-enhanced signals, however, degrade more gracefully than their non-LLM counterparts. At a 3-month lag, the LLM-enhanced financial statement signals still produce higher alpha than the non-LLM factor signals at a

1-month lag in several configurations. Even at a 12-month lag, most LLM-enhanced signals retain positive alpha, while some non-LLM signals approach zero. The persistence of these signals at longer horizons suggests that the information the LLMs extract from financial statements is not quickly incorporated into prices.

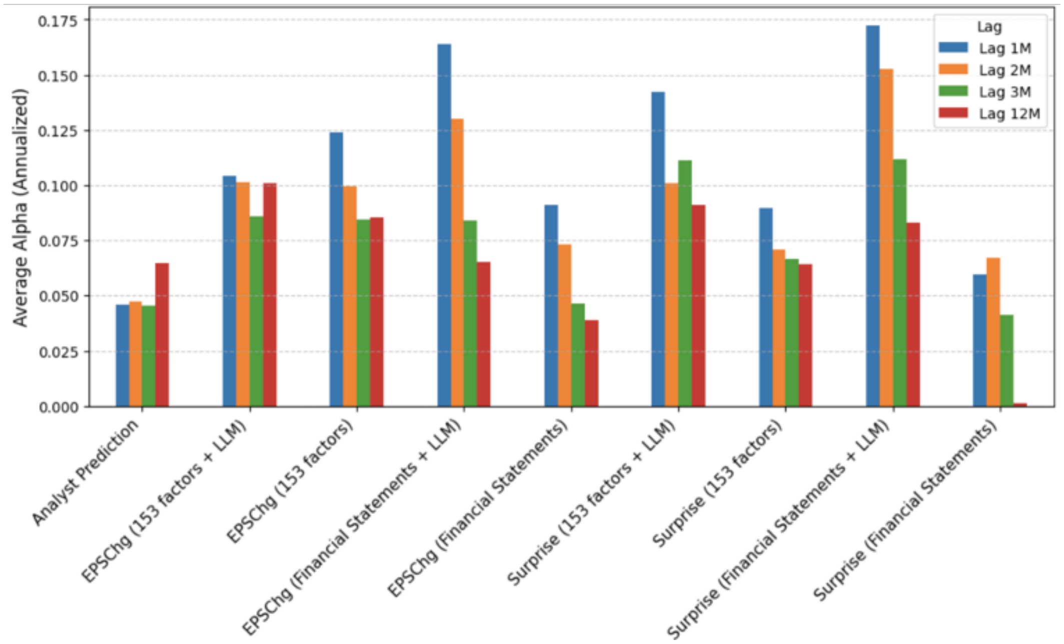

Figure 6: Annualized average long-side alpha across prediction lags of 1, 2, 3, and 12 months for all signals. LLM-enhanced signals (those with "+ LLM") consistently produce higher alpha and degrade more slowly with increasing lag.

5 Discussion

Our results demonstrate that open-source language models, despite being substantially smaller than GPT-4, generate economically valuable signals when applied to financial statement analysis. The standalone accuracy of the LLM predictions is modest when measured against realized EPS changes, but meaningful when evaluated against earnings surprises. This distinction matters: the models appear to capture something about firm fundamentals that the market, as reflected in analyst consensus, does not fully anticipate.

The more important finding, in our view, is that the value of LLM outputs is best realized not as a standalone signal but as an input to a supervised learning pipeline. When LLM-derived embeddings are combined with traditional features in an XGBoost classifier, both classification accuracy and portfolio performance improve substantially. The improvement is largest when LLM features are added to raw financial statement data and smallest when added to precomputed factor characteristics. This pattern holds across both prediction targets, across portfolio evaluation metrics, and across prediction lags. The 153 factors from Jensen et al. [2] already summarize much of the economic content that the LLMs extract from financial statements, so when investors have access only to raw accounting data, the LLM narratives fill a gap that would otherwise require either human analysis or engineered features.

The persistent weakness of the short-side portfolios deserves attention. Across nearly all signals, long portfolios produce significant alpha while short portfolios do not. We suspect this reflects two reinforcing forces. The LLM training data likely overrepresents positive or neutral language about companies, making the models less attuned to deterioration. At the same time, standard financial statements are structured around reporting current positions rather than signaling future decline, which may limit what any model can extract about downside risk from this data alone.

Our findings build on those of Kim et al. [4] in a specific way. While they demonstrated that GPT-4 can match analyst accuracy as a standalone predictor, we show that the greater opportunity lies in combining LLM reasoning with quantitative models. The narrative text generated during chain-of-thought prompting encodes economic logic that, when embedded and fed to a classifier, provides incremental predictive power beyond what either source achieves alone.

6 Limitations and Future Work

Despite anonymizing the financial statements by removing company names and fiscal years, we cannot fully rule out the possibility that the LLMs rely on memorized information from their pretraining corpora. Recent work by Yoo [6] and others has raised concerns about lookahead bias in LLM-based financial analysis. Additional debiasing techniques, such as randomizing the least significant digits of financial statement items, could help isolate the models' analytical reasoning from memorized knowledge.

We evaluate only three open-source models ranging from 8 to 14 billion parameters. Larger models, including GPT-4 and more recent open-source alternatives, may produce higher-quality analyses, and our results should be interpreted as a lower bound on the potential of LLM-augmented equity research. Relatedly, we use a single prompting strategy throughout. The sensitivity of LLM outputs to prompt design is well documented, and alternative approaches such as repeated top-3 guesses combined with chain-of-thought prompting [6] may yield more calibrated or diverse predictions.

The financial statement data is available at annual frequency, which limits the freshness of the LLM-derived signals. Incorporating quarterly filings or earnings call transcripts would provide more timely inputs and could meaningfully strengthen the approach.

Our downstream classifier is also limited to XGBoost. Deep learning models that can jointly process numeric features and text representations may capture richer interactions between the two modalities, and we view this as a promising direction for subsequent work.

References

[1] Xi Chen, Yang Ha Cho, Yiwei Dou, and Baruch Itamar Lev. Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research*, 60(2): 467–515, 2022.

[2] Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.

[3] Alex G Kim, Maximilian Muhn, and Valeri V Nikolaev. Bloated disclosures: Can ChatGPT help investors process information? *Chicago Booth Research Paper No. 23-07*, 2024.

[4] Alex G Kim, Maximilian Muhn, and Valeri V Nikolaev. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*, 2024.

[5] Jane A Ou and Stephen H Penman. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*, 11(4):295–329, 1989.

[6] Minji Yoo. How much should we trust large language model-based measures for accounting and finance research? *The Wharton School Research Paper*, 2024.

[7] Tianyu Zhou et al. FinRobot: An open-source AI agent platform for financial applications using large language models. *arXiv preprint arXiv:2411.08804*, 2024.

A Prompt Template

The following is the chain-of-thought prompt used across all three LLM models. The prompt consists of four components: a scenario description establishing the model's role, a specification of the input format, detailed step-by-step instructions, and a structured output template.

> *You are a financial analyst specialized in analyzing accounting data. You will receive a query containing the balance sheet and income statement of an unspecified company. Your current goal is to analyze these two tables and form a prediction about the direction of the company's future EPS.*
>
> *The format of the two tables is as follows. Each entry in the two tables is separated by a tab token. The first column contains the accounting item name. The remaining columns contain the numerical values of each item over various years. The balance sheet table contains data for the present year (t) and the past year (t-1), while the income statement further includes the year before the past (t-2).*
>
> *In order to analyze these tables, follow these steps:*
>
> ***Step 1** – Begin by performing a "trend analysis" where you identify notable changes in certain financial statement items. Focus on items that are relevant for your goal of predicting the direction of EPS. Use short paragraphs to summarize these findings.*
>
> ***Step 2** – Perform a "ratio analysis" where you compute and analyze key financial ratios. Do not limit the set of ratios that need to be computed and consider those that may be relevant. When calculating the ratios, state the formulae first and then perform simple computations. Provide an economic interpretation of the ratio and what it implies for future performance. Use short paragraphs to summarize these findings.*
>
> ***Step 3** – Finally, using the derived quantitative information and your associated insights from it, predict whether the EPS is likely to increase or decrease over the next fiscal year.*
>
> *Structure your output as follows:*
> *Trend analysis: [TREND ANALYSIS HERE]*
> *Ratio analysis: [RATIO ANALYSIS HERE]*
> *Direction: [increase/decrease]*
> *Magnitude: [small/moderate/large]*
> *Confidence: [low/moderate/high]*
> *Rationale: [RATIONALE HERE]*

B Standalone LLM Prediction Analysis

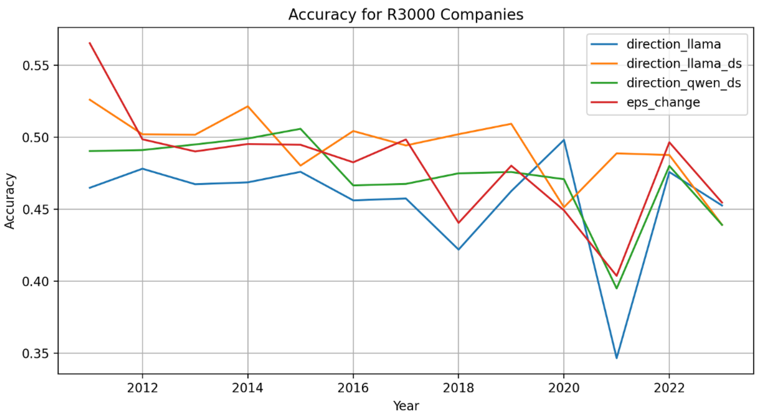

Figure 7: LLM prediction accuracy compared with actual EPS change by year. Accuracy hovers near 50% for all models, indicating that standalone LLM predictions do not substantially outperform the naive baseline for this target.

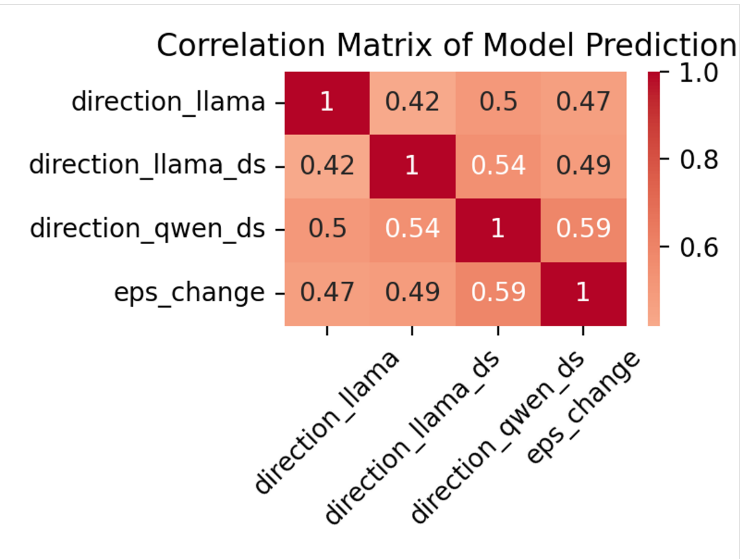

Figure 8: Correlation matrix of directional predictions across the three LLM models and the realized EPS change. Cross-model correlations range from 0.42 to 0.54, indicating partially distinct predictions across models.

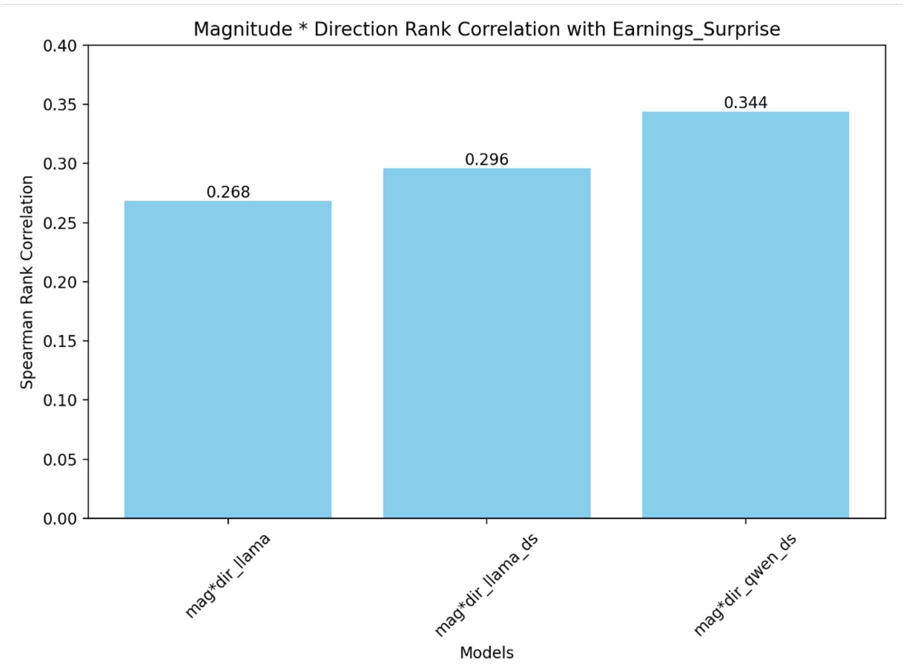

Figure 9: Spearman rank correlation between magnitude × direction and realized earnings surprise. All correlations are positive and significant.

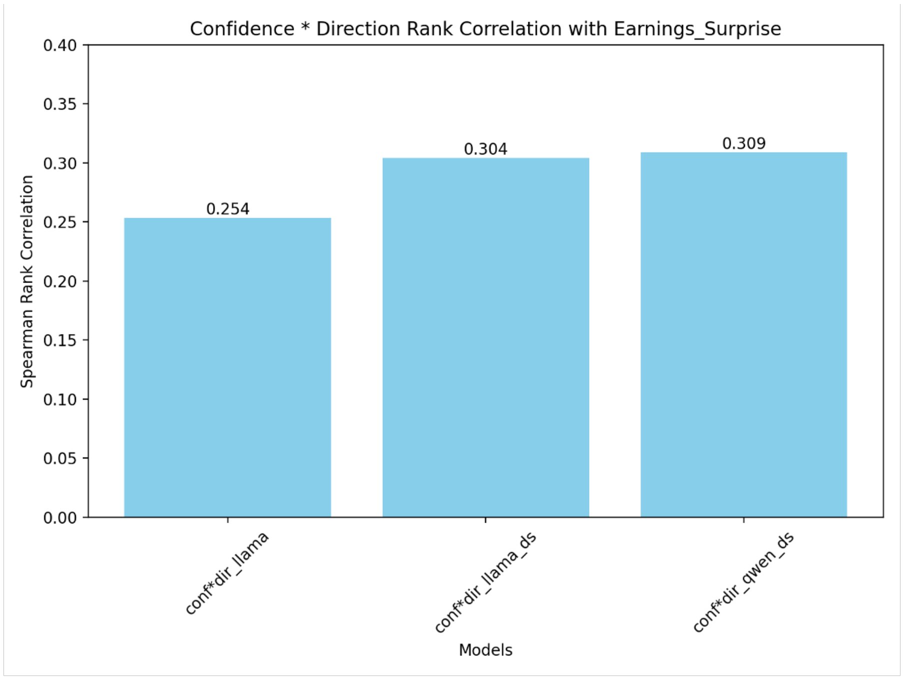

Figure 10: Spearman rank correlation between confidence × direction and realized earnings surprise. All correlations are positive and significant.

C Confusion Matrices Across Prediction Lags

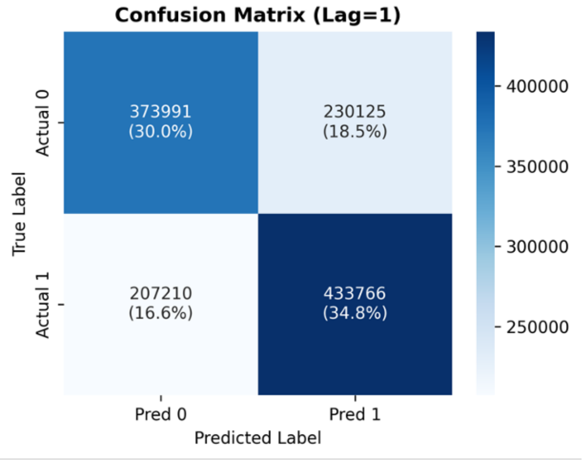

Figure 11: Confusion matrix aggregated across all feature sets and rolling windows at lag = 1 month. Overall accuracy is 64.8%.

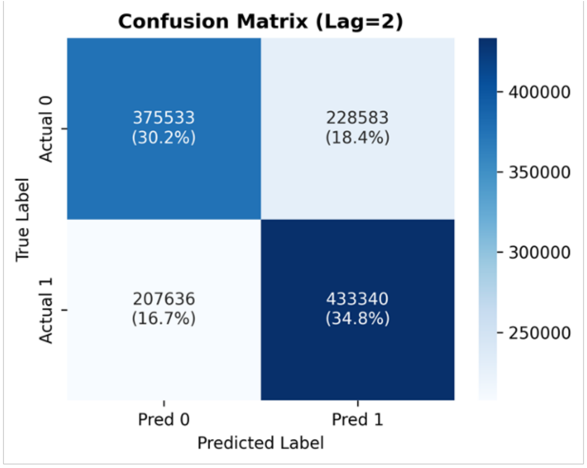

Figure 12: Confusion matrix at lag = 2 months. Overall accuracy is 65.0%, nearly identical to the 1-month lag.

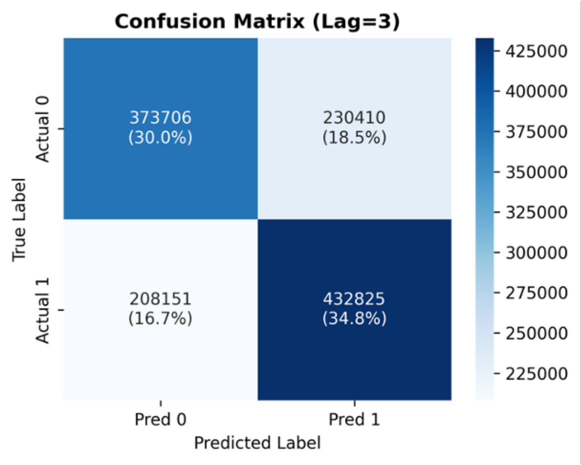

Figure 13: Confusion matrix at lag = 3 months. Overall accuracy is 64.8%, confirming stability across prediction lags.

D Cumulative Alpha Comparison

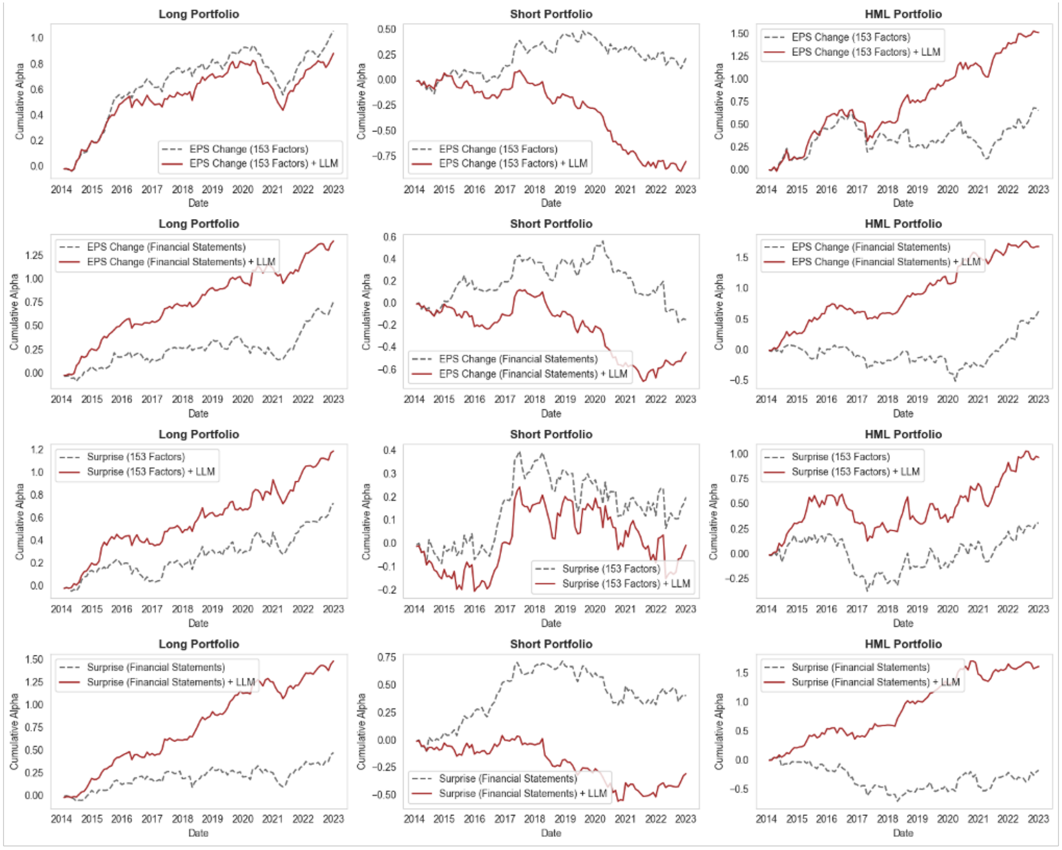

Figure 14: Cumulative alpha over time for each signal (dashed gray) and its LLM-enhanced counterpart (solid red), shown for long, short, and HML portfolios. The LLM-enhanced version consistently outperforms on the long side and HML.

E Information Ratio Across Prediction Lags

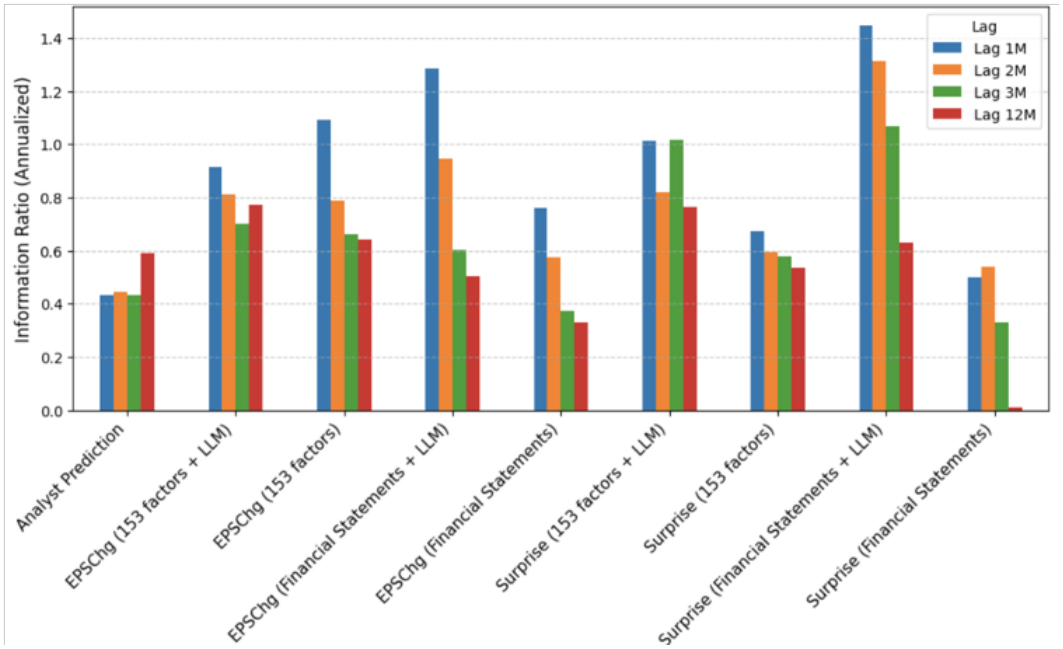

Figure 15: Annualized information ratio across prediction lags of 1, 2, 3, and 12 months for all signals. The pattern mirrors that of alpha: LLM-enhanced signals maintain higher information ratios at longer lags.

F Portfolio Turnover

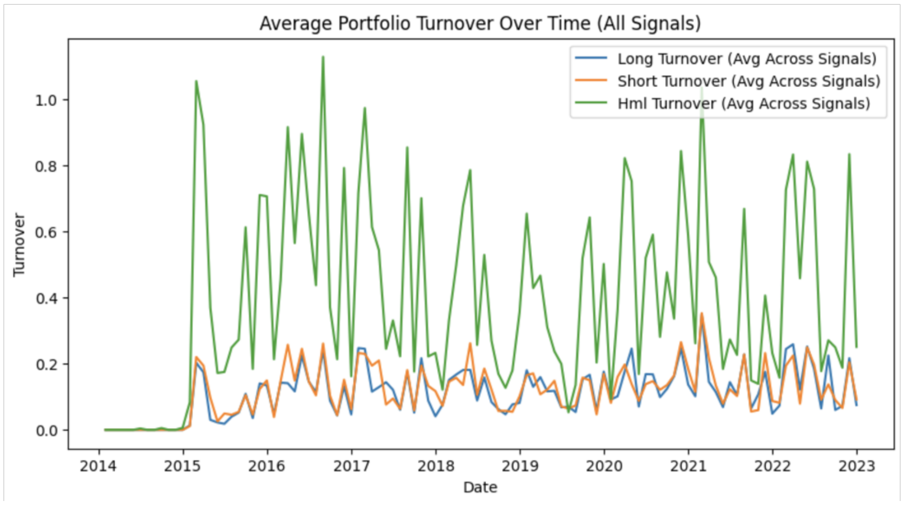

Figure 16: Average monthly portfolio turnover over time, averaged across all signals. Long and short portfolio turnover remains low (10–20%), while HML turnover is substantially higher due to the combined rebalancing of both sides.

G Earnings Announcement Coverage

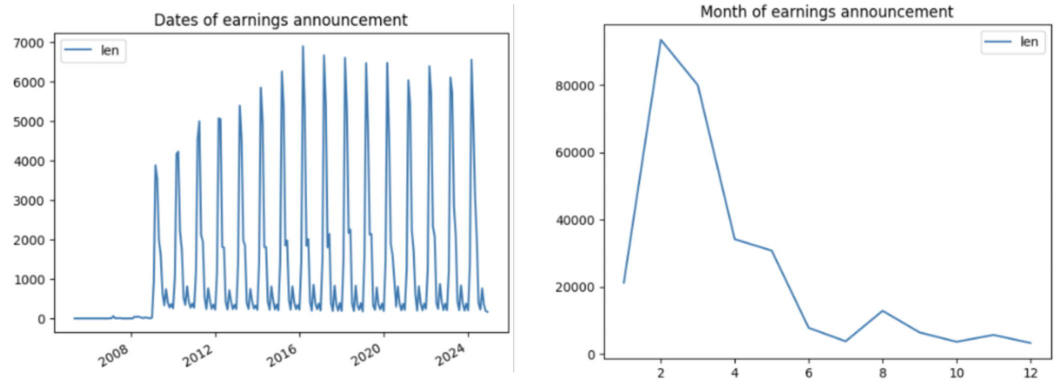

Figure 17: Left: frequency of earnings announcements over time, showing sample coverage beginning around 2009 and stabilizing thereafter. Right: distribution of earnings announcements by calendar month, with a concentration in the first quarter corresponding to fiscal year-end reporting.