A Multimodal Framework for Predicting Surprises in Earnings Metrics of FTSE 100 Companies

Stephanie Keddy
Supervisor: Dr William Marsh
MSc Data Science and Artificial Intelligence, QMUL

Abstract— When year-end earnings disclosures reveal results that differ from consensus forecasts, markets react and valuations change, making it valuable to anticipate these surprises in advance. This paper improves forecasting by predicting the size of surprises in revenue, EBITDA, EBIT and net income. It then recalibrates the forecast by adding the predicted surprise to the prevailing analyst consensus. We introduce a multimodal framework that combines language from earnings calls with changes in the analyst consensus. A rolling origin design fixes transcript features at each disclosure and tracks estimate snapshots at +14, +30, and subsequent +30-day intervals between earnings calls. Using FTSE 100 companies from financial years 2019-2024, we run out-of-sample evaluations with 2024 held out. Language-only models lower mean absolute error across all metrics, with the largest reduction observed for EBIT (24.0% relative to baseline). Consensus-dynamics-only models improve accuracy in a metric dependent way. Hybrid models are most reliable on operating measures, reducing mean absolute error for EBITDA by 9.6% relative to baseline and achieving the highest win rate for EBIT. Improvements can be seen up to 120 days after calls and statistically significant.

Keywords—multimodal, deep ensembles, financial NLP, consensus revision dynamics, surprise magnitude prediction

I. Introduction

Earnings announcements move markets because consensus expectations of earnings metrics are imperfect. The difference between reported results and consensus forecasts is known as an earnings surprise. Anticipating both the direction and magnitude of these surprises before they trigger post-announcement price adjustments is valuable for strategic investment decision-making. While consensus forecasts provide a practical baseline and are widely used by investors, substantial evidence suggests that consensus can underreact to new information and embed predictable errors, particularly around major communications. A significant amount of academic work studies these errors ex post through the lens of price reactions. In contrast, this paper focuses on improving forecasts ex ante by modelling the evolving magnitude of surprises and using those predictions to recalibrate consensus at multiple time horizons before results are released.

Following a study by Ball and Brown (1968), research has largely focused on earnings per share (EPS) surprises. However, in practice, investors typically look to a wider range of earnings metrics, as shown in Table I, including revenue, earnings before interest, tax, depreciation and amortisation (EBITDA), earnings before interest and tax (EBIT), and net income (NI), to get a more comprehensive view of firm performance (Beyer *et al.*, 2010). Crucially, expectations for these metrics evolve over the reporting cycle in response to company-issued guidance. Yet most empirical designs define the "surprise" using a single snapshot of the consensus taken immediately before the results release, effectively ignoring the information embedded in how expectations have evolved to that point. Further still, literature tends to reduce surprise prediction to a simple beat-versus-miss predictive framework, thereby discarding economically meaningful variation in magnitude (Chung and Tanaka-Ishii, 2023; Koval, Andrews and Yan, 2023). To address these limitations, this study models the magnitude of surprises for four earnings metrics, revenue, EBITDA, EBIT and NI, at multiple time horizons leading up to each full year results announcement.

A. Methodological Framework

Accordingly, a multimodal predictive framework is developed that integrates language signals from earnings call transcripts with between-call consensus revision dynamics. This integration captures changes in the target earnings metric and other related earnings metrics at each horizon before the full-year results announcement, enabling us to track how expectations evolve from one call to the next. Transcript features quantify tone, uncertainty, and forward-looking guidance in presentations and Q&As, while revision dynamic measures characterise the level, trend, volatility and persistence of consensus expectations as firm guidance is understood and assimilated. The analysis focuses on FTSE 100 companies from financial year (FY) 2019 to FY 2024, providing multiple reporting cycles and variation in market conditions for differing guidance practices and revision activity. To integrate the transcript and revision measures in a temporally coherent way and to avoid look-ahead bias, a rolling-origin design is used. Each performance-related communication sets an origin date; at that origin, transcript features are fixed, and the evolution of consensus is then tracked at plus-14 days, plus-30 days, and at subsequent 30-day intervals until the next call. These time-stamped snapshots operationalise how guidance is incorporated into the consensus estimates and supply the revision metrics used by the model. For each origin and horizon pair, the model predicts the magnitude of the surprise, and this predicted surprise is added to the contemporaneous consensus to produce a refined point forecast for each earnings metric.

TABLE I. Financial Indicators

Earnings Hierarchy[a]	Analytical Significance for Investors
Revenue	Top-line sales; growth signal
└ EBITDA	Pre-asset, pre-financing, core operating cash proxy
└ EBIT	Post-asset, pre-financing operating profit
└ NI	Bottom-line profit after all expenses
└ EPS	Net income per share; investor earnings

[a] The hierarchy defines valid cross-metric usage: consensus estimates are only used for upstream metrics, while downstream metrics contribute solely via lagged values and estimate change to prevent leakage.

B. Research Objectives

The empirical analysis addresses three questions: (1) whether transcript-based features alone reduce surprises relative to the consensus baseline; (2) whether consensus-dynamic features alone reduce surprises relative to the same baseline; and (3) whether a combined, multimodal specification outperforms both single-source models and the consensus baseline. In doing so, the study assesses whether, and to what extent, a multimodal framework integrating transcript language with consensus-revision dynamics improves year-end performance forecasts beyond the consensus benchmark.

The remainder is organised as follows: Section II reviews related work and positions our contribution; Section III

describes data and preprocessing; Section IV presents textual feature engineering; Section V details dataset construction; Section VI outlines the experimental design and modelling workflow; Section VII reports results and analysis; and Section VIII concludes with findings, conclusions, limitations, and directions for future research.

II. Related Work and Motivation

A. Beyond EPS: Information in Revenue, EBITDA, EBIT, NI

EPS forecasts are a key focus in financial sentiment research and remain a core input in equity valuation models (Loh and Mian, 2006; Ertimur, Sunder and Sunder, 2007). However, as Pope and Wang (2023) argue, forecasting a broader set of earnings metrics beyond EPS can enhance research quality by encouraging deeper analysis of the factors that drive firm performance. Early evidence supporting this broader perspective comes from Bradshaw, Richardson and Sloan (2001) who show that the components of earnings contain information relevant to investors, helping to explain post-announcement drift not captured by EPS alone. Ertimur, Livnat and Martikainen (2003) extend this insight to European firms, demonstrating that revenue surprises help explain abnormal returns even after controlling for EPS, indicating that investors look to what drives earnings, not just the headline figure. This view is further reinforced by recent research showing that surprises in revenue, EBITDA, EBIT, and NI can surpass EPS in explaining announcement returns because they help investors assess the credibility and persistence of the market baseline EPS surprise, which is often a noisy measure of performance. (Hand *et al.*, 2022; Bilinski, 2024). Together, these findings highlight the value of extending research to encompass a broader set of earnings metrics.

B. Leveraging Surprise Magnitudes for Enhanced Forecasting

Extensive research demonstrates that the magnitude of surprises provides valuable forward-looking information beyond their mere direction. For instance, larger surprises trigger proportionally bigger analyst revisions (Kasznik and McNichols, 2002), amplify post earnings announcement drift for months (Livnat and Mendenhall, 2006), steepen the return surprise gradient in the tails of forecast errors (Kinney, Burgstahler and Martin, 2002), and the magnitude of revenue surprises has been shown to predict future earnings growth (Jegadeesh and Livnat, 2006). Collectively, these findings indicate that surprise magnitudes contain systematic, predictive content that could be used to improve ex ante expectations. However, most studies focus on analysing market reactions ex post rather than leveraging these patterns for prospective forecasting. To bridge this gap, our study models historical surprise magnitudes and integrates them into consensus forecasts before earnings announcements, systematically refining baseline expectations. This approach can enhance the precision of valuation models like those of Easton et al. (2002), which rely on unadjusted forecasts, and aligns with evidence from Hou, van Dijk and Zhang (2012) that model-based forecasts outperform analyst estimates in cost of capital estimation. By improving expectations ex ante, this framework enables investors to anticipate market deviations and position themselves proactively.

C. Textual Representation in Financial Forecasting

Using textual data for financial forecasting has attracted growing attention as language carries explicit and latent signals such as sentiment, uncertainty, temporal focus, narrative flow, strategic framing, and thematic emphasis that reveal expectations, perceived risk, and forward-looking intent. These qualities make it a powerful input for predictive modelling, though extracting them remains a core challenge in financial natural language processing (NLP). Early work used lexicon-based methods, such as the Loughran-McDonald dictionary (2011), which improved on general sentiment lexicons by addressing financial-specific misclassifications. While limited by bag-of-words assumptions and weak at capturing contextual nuance, these methods remain valued for transparency and domain relevance, and have been effectively applied in large-scale financial text analysis (Jiang *et al.*, 2019).

As sentiment analysis in NLP has advanced, attention has shifted towards models that represent language in context rather than as isolated terms, most notably FinBERT, a BERT-based transformer fine-tuned on financial language. FinBERT captures sentiment at the phrase and sentence level by applying self-attention mechanisms over financial-domain pretraining, allowing it to model syntactic dependencies and semantic context across tokens. Empirical tests show that FinBERT-derived sentiment variables boost the explanatory and predictive power of financial forecast models relative to both lexicon and generic-BERT benchmarks (Huang, Wang and Yang, 2023; Du *et al.*, 2024). These improvements notwithstanding, FinBERT's performance remains bounded by its parameter scale and 512-token context window, which constrains its ability to capture longer-range semantic and discursive patterns across extended financial narratives.

To address these limitations, recent financial-domain large language models (LLMs) have been designed to better capture long-range dependencies and nuanced sentiment in financial text (Iacovides *et al.*, 2024). Among these, models like FinLLaMA-3-8B generate embeddings that capture not only the polarity and strength of sentiment, but also broader contextual, discursive, and financial-semantic patterns across entire documents, supported by an input capacity of up to 8,192 tokens (Qian *et al.*, 2025). These capabilities are particularly valuable in applications involving long-form financial narratives, where richer textual representations have been shown to enhance the forecasting of market-relevant surprises (Koval, Andrews and Yan, 2023).

D. Analyst Revision Dynamics in Financial Forecasting

The prevailing convention in forecast surprise research benchmarks performance against a single analyst consensus forecast issued immediately prior to earnings announcements (Bernard and Thomas, 1989; Stickel, 1991). However, relying solely on this static snapshot overlooks the informational content embedded in the trajectory of analyst revisions. Han (2025) shows that asymmetric information among forecasters can cause consensus forecasts to underreact to new information, embedding predictable belief errors despite full rationality at the individual level. Similarly, Li, Liu and Li (2024) find that analysts frequently fail to incorporate historical data fully, making the direction, frequency, and magnitude of revisions potentially exploitable.

Additional evidence underscores the predictive value of revision dynamics: Clement (1999) and Kim, Lobo and Song (2011) demonstrate that the timing and responsiveness of analyst revisions influence the informativeness of consensus forecasts. Park and Zach (2025) further link the volatility of these revisions to underlying uncertainty in the information environment. Moreover, Chen et al. (2022) and Hess, Simon and Weibels (2025) show that forecasts based on detailed financial statement data, particularly upstream income statement components such as EBIT and NI, contain complementary information that enhances the predictive accuracy of earnings models.

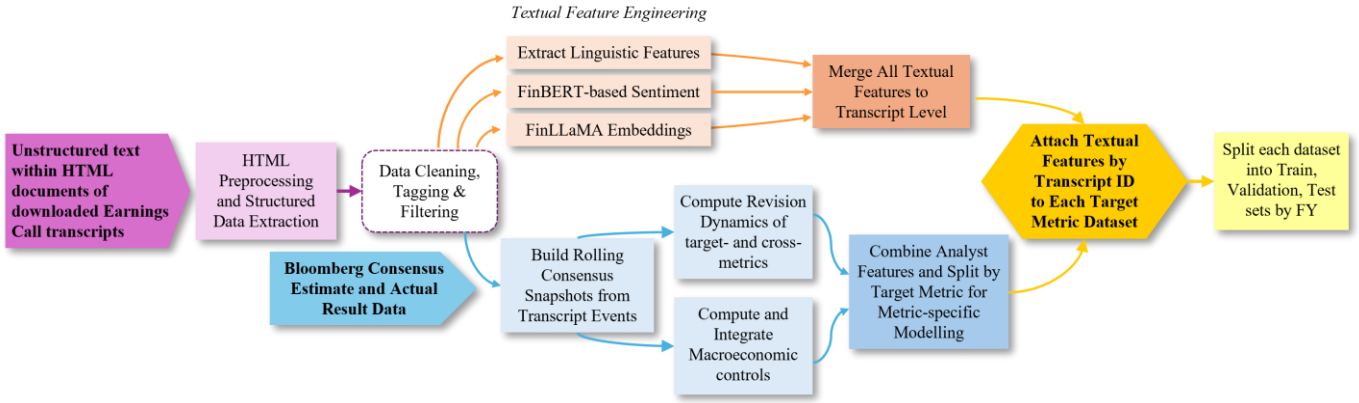

Fig. 1. Call transcripts are preprocessed for feature extraction (linguistic, sentiment, embeddings). Consensus features are computed and combined with macroeconomic data and split into metric-specific datasets; transcript features are then merged into these datasets, which are subsequently temporally split for model training, validation and testing.

Importantly, the persistence of these signals across firms and horizons suggests they are neither transitory nor artefacts of data mining. Accordingly, recent machine learning research reinforces this dynamic view; Van Binsbergen, Han and Lopez-Lira (2023) report that models incorporating revision histories yield unbiased conditional forecasts and reveal horizon-dependent distortions. Likewise, Wang et al. (2025) find that sentiment trajectories drawn from sequences of earnings calls predict firm performance more effectively than isolated disclosures.

Addressing these insights, our study models consensus surprises as continuous outcomes across multiple time windows between performance communications. This enables a contemporaneous and informative characterisation of revision dynamics through features, such as the trailing 12-month revision volatility and count, that reflect the evolving state of expectations at each point.

E. Multimodal Forecasting Frameworks

Recent work in financial forecasting increasingly integrates multiple information channels, particularly textual disclosures and structured analyst data. A growing body of evidence supports the view that combining qualitative and quantitative signals enhances predictive insight. Bradshaw et al. (2021) find that analyst recommendation revisions aligned with broader qualitative cues convey more informative signals than either source alone, reinforcing the value of a multi-modal approach. Ni et al. (2024) further demonstrate that incorporating earnings call transcripts, firm fundamentals, aggregate analyst grades, and market indicators to a transcript-and-fundamentals model improves predictive accuracy, highlighting the value of combining structured and unstructured financial data. Yet even with these advances, few recent machine learning models leverage detailed analyst consensus dynamics, such as revision timing, volatility, and cross-metric estimates, despite the broader evidence that analyst forecast characteristics are informative for market expectations, as previously noted. Collectively, the literature affirms the potential of multimodal modelling while exposing a persistent gap: few studies integrate deep language representations and time-aware analyst forecast structures within a unified framework to predict evolving consensus expectations.

F. Summary and Research Contribution

Prior research highlights the value of looking beyond EPS, modelling expectations as dynamic, and using deep contextual language representations. Yet these strands are rarely unified. We close this gap by introducing a single, multimodal framework that (1) predicts continuous surprise magnitudes for Revenue, EBITDA, EBIT, and NI; (2) fuses long-context transcript signals with time-stamped consensus-revision dynamics in a rolling-origin, multi-horizon design; and (3) recalibrates consensus ex ante by adding model-predicted surprises to contemporaneous estimates. Evaluated out-of-sample on FTSE-100 companies (FY 2019-2024), the approach delivers statistically significant error reductions versus the raw consensus and yields practical, horizon-specific guidance on when language, consensus dynamics, or hybrid signals add most value.

III. Data and Preprocessing

A. Data Sources and Temporal Scope

Earnings call transcripts were retrieved from MarketScreener for all FTSE 100 constituents that issued interim-performance and results announcements, from FY 2019 year-end to FY 2024 year-end. Each year-end disclosure serves as the initial input for forecasting the subsequent FY, with additional announcements within that year providing further forward-looking information. For the corresponding period, sell-side consensus estimates and actuals for Revenue, EBITDA, EBIT, NI, and EPS were obtained from Bloomberg. To account for broader market conditions, two macroeconomic controls were included: the Bank of England (BoE) Bank Rate, sourced from the BoE Database, and realised FTSE 100 volatility over a 252 trading-day window, retrieved from Yahoo Finance using the yfinance package.

B. Target Variables

Four separate target metrics are studied, Revenue, EBIT, EBITDA and NI, resulting in four parallel modelling datasets that share an identical raw data pipeline construction, as shown in Fig. 1 but diverge at the feature selection stage and beyond.

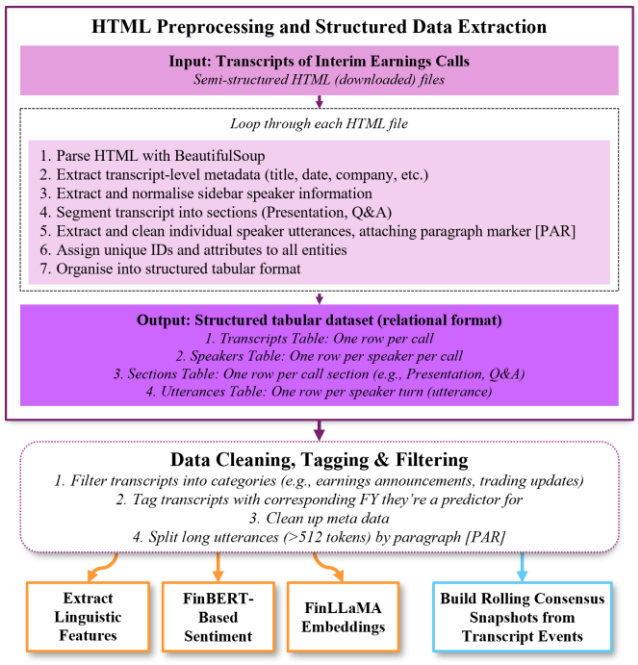

Fig. 2. Preprocessing pipeline converting HTML documents of call transcripts to structured relational tables for feature extraction and financial data integration.

C. HTML Preprocessing and Structured Data Extraction

Raw HTML transcripts were batch-parsed in Python 3.11 using BeautifulSoup 4.13, following the pipeline illustrated in Fig. 2. Each transcript was assigned a UUID and stored across four relational tables: *transcripts* for call-level metadata, *speakers* for unique participant roles, *sections* for presentation and Q&A boundaries, and *utterances* for cleaned, paragraph-delimited speech blocks. Together, these tables provide a consistent, leak-free structure for feature aggregation and model training. A sample of structured tabular datasets can be found in Figs. A.1–A.2 in the Appendix.

IV. Textual Feature Engineering

To convert the unstructured textual data from earnings call transcripts into quantitative predictors for modelling, we extract four complementary families of textual features.

A. Stylistic Signals

We extract stylistic features that shape how easily investors process executive speech, independent of content. Using textstat and preprocessing, we compute four core metrics from executive utterances: Flesch Reading Ease, average sentence length, average word length, and total word count, capturing syntactic complexity, verbosity, and information volume (Li, 2008). To assess narrative stance, we calculate the proportion of first-person plural pronouns ("we") among all pronouns, based on spaCy 3.8 part-of-speech tagging. This reflects the degree of collective, internally aligned framing, which influences perceived competence and credibility (Fladerer *et al.*, 2021). Together, these features capture stylistic cues tied to communication clarity and managerial overconfidence.

B. Lexical Signals

To analyse the semantic content of management language, we use two types of sources: (1) the Loughran and McDonald (2011) dictionary, with example terms shown in Table II; and (2) custom thematic dictionaries based on four lexicons, as summarised in Table III, full breakdown in Table A.XII in the Appendix. For each of these lexicons, term proportions in executive speech are computed. In addition, tone divergence is measured for positive and negative terms by comparing executive term proportions to the unweighted average term proportions of all non-operator participants (Angelo *et al.*, 2025). This process is illustrated in Fig. 3.

TABLE II. Loughran and McDonald (LM) Lexicons

Word Class	Examples
Uncertainty	Cautiously, believe, anticipated, fluctuating, risk, differ
Constraining	Required, obligations, mandatory, limit, prevent, impose
Strong Modal	Always, clearly, definitely, strongly, must, will, never
Weak Modal	May, could, possible, might, depend, appears, somewhat
Positive	Gain, able, advances, improvements, successful, best
Negative	Loss, against, impairment, deficit, adversely, decline

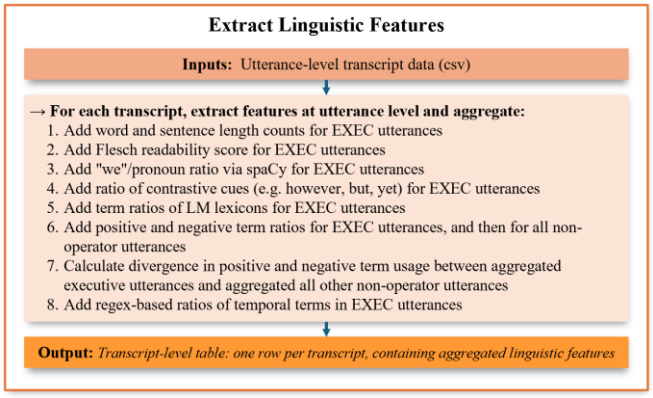

Fig. 3. Overview of the linguistic feature extraction process. The figure illustrates the step-by-step procedure for deriving aggregated transcript-level linguistic features from utterance-level transcript data, highlighting the focus on executive (EXEC) utterances.

TABLE III. Custom Thematic Lexicons

Lexicons	Description and Examples
Contrastive Terms	Rhetorical shift or qualification terms in executive messaging *(e.g., however, although)* (Fraser, 1999; Hyland, 2018)
Strong Forward-Looking	Assertive strategic intent terms around growth and investment *(e.g., expand, build, develop)* (Bozanic, Roulstone and Van Buskirk, 2018)
General Future Focus	General forward-looking orientation or expectation term indicators *(e.g., next quarter, expect, forecast)* (Li, 2010)
Past-Result Focus	Historical outcome framing terms *(e.g., last year, achieved, reported)* (Merkl-Davies, Brennan and McLeay, 2011)

C. Sentiment Metrics (FinBERT)

Raw word counts cannot reveal how sentiment evolves over the course of a call; therefore, we apply FinBERT (Araci, 2019), a BERT-based sentiment classifier trained on finance-specific corpora, to every utterance. From the predicted class probabilities for positive, negative, and neutral sentiment, we derive a continuous sentiment score s for each utterance, defined as: $s = P(positive) - P(negative)$. We aggregate these outputs at the transcript level by computing, for each combination of speaker role (executive or external participant) and call section category, the mean and standard deviation of the sentiment score, the average model confidence in classification, and the proportion of utterances labelled as positive, negative, or neutral. To capture temporal variation, we calculate early-versus-late sentiment averages based on utterance order and estimate a linear sentiment trend using ordinary least squares regression. The full sentiment processing workflow is shown in Fig. 4 below.

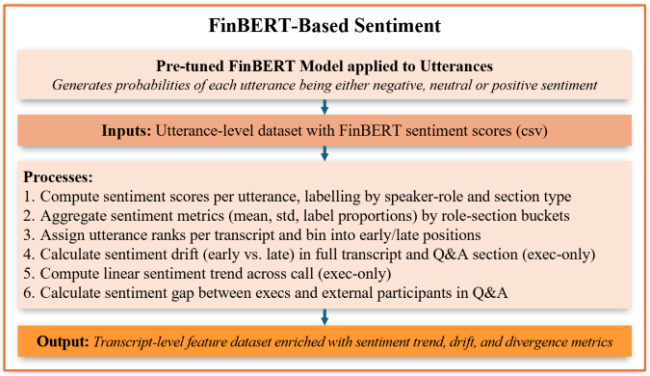

Fig. 4. Overview of sentiment feature derivation with FinBERT. A pre-trained FinBERT model scores utterances for sentiment, which are then aggregated into transcript-level measures of trends, drift, and role-based differences.

D. Semantic Embeddings (FinLLaMA)

To capture nuanced meaning beyond word-level sentiment or lexicon counts, we apply FinLLaMA, a finance-specific large language model based on LLaMA 3 (Qian *et al.*, 2025), to generate high-dimensional embeddings from both executive and external participant utterances, as outlined in Fig. 5. Each transcript is represented by a 4,096-dimensional vector, computed by averaging contextual embeddings across segmented text blocks of up to 8,192 tokens. Prior to embedding, utterances are formatted with speaker and section

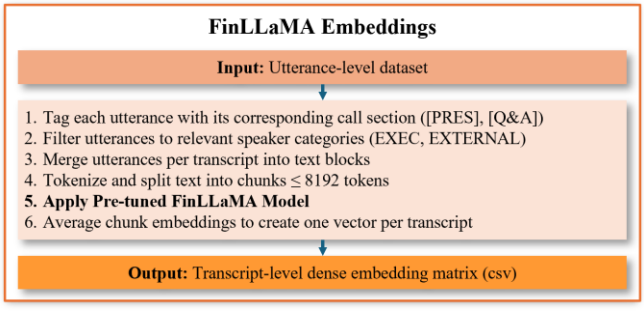

Fig. 5. FinLLaMA high-dimensional embedding workflow. Utterances are filtered, merged, and tokenized, then encoded with a pre-trained FinLLaMA model to generate transcript-level dense embeddings.

tags to preserve structural context. These embeddings encode latent semantics and discourse structure, supporting transcript-level representations of meaning that capture rhetorical framing, narrative emphasis, and thematic orientation. They complement our other textual feature families by providing a dense, context-sensitive representation of narrative content that extends beyond sentiment, lexical choice, or syntax.

V. Dataset Construction and Temporal Splits

A. Building Rolling Consensus Snapshots from Transcripts

We take rolling snapshots of the latest sell-side consensus 14 and 30 days after each event, and subsequently at 30-day intervals until the next communication, denoting each snapshot as window time t, as shown in Fig. 6. As temporal controls, we record two lags: (1) the number of days between the end date of each estimation window and the eventual results announcement; and (2) the cumulative days that have elapsed since the firm's previous communication.

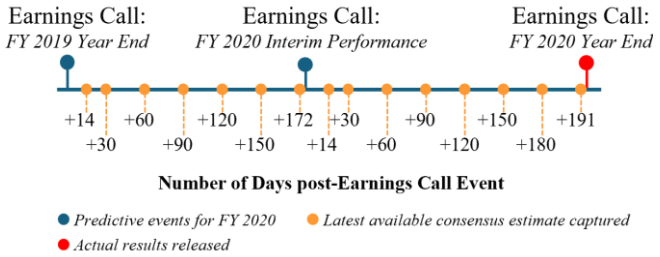

Fig. 6. Illustrative timeline of rolling consensus snapshots within a FY: each window begins at the earnings call and ends at +n days; at the window end, the latest available consensus estimate is recorded.

B. Consensus Revision Dynamics

To incorporate revision dynamics, each snapshot is augmented with variables that describe both the most recent adjustment in consensus and the longer-run behaviour of analyst forecasts. Short-term adjustment is captured by two measures. First, the lag-1 consensus level, defined as the estimate recorded at the end of the immediately preceding estimation window, and second, its first-difference $\Delta^1 C_{i,m,t}$ defined as:

$$\Delta^1 C_{i,m,t} = C^{cons}_{i,m,t} - C^{cons}_{i,m,t-1}\,, \tag{1}$$

where $C^{cons}_{i,m,t}$ is the current window's estimate, and $C^{cons}_{i,m,t-1}$ is the estimate from the previous window for firm i and metric $m \in$ {Revenue, EBITDA, EBIT, NI}. For the first window in a communication cycle, where no earlier window exists, we use the latest estimate dated before the window opens. The long-run component is evaluated over the 365 calendar days preceding the current window end date. Within this interval, the ordered sequence of consensus estimates is converted to successive percentage changes; the change from the final estimate is discarded, and the standard deviation of the remaining changes is taken as the revision-volatility measure σ_{365}. The same sequence yields a revision count, and intervals containing fewer than three changes are flagged as information-sparse. The prior year-end actual is also added for the corresponding target metric. Accordingly, all statistics are derived solely from information available at the snapshot window date to avoid look-ahead bias. Thus, the resulting dataset supports prediction of earnings surprises using a panel of pre-release expectations, incorporating revision dynamics of both target and cross-metrics. Raw consensus estimates for cross-metrics are included only when those metrics precede the target in the earnings hierarchy as shown in Table I.

C. Macroeconomic Controls

For every snapshot we compute the realised FTSE 100 volatility over [−252,−1] trading days, and record the BoE Bank Rate as of t−1. These variables serve as exogenous controls.

D. Metric-specific Dataset Formation and Cleaning

Separate modelling datasets were created for each target metric and restricted to only companies with a complete history of reported actuals for that metric. All estimate-based and derived features were retained, but actuals for non-target metrics were removed to prevent data leakage. Companies were excluded if, in any FY, they recorded no new estimates for the target metric or held no performance-related calls during the 12 months preceding the year-end results announcement, thereby retaining only firms with meaningful analyst coverage and disclosure. If more than five companies lacked cross-metric estimate data, the associated features were removed; otherwise, those companies were excluded. These cleaning steps balanced sample size and feature breadth for subsequent model development. Table IV presents the resulting company-transcript composition for each metric-specific dataset.

TABLE IV. Metric-Specific Dataset Composition

Metric Dataset	FTSE 100 Firms Retained	Transcripts Retained	Number of Target-Metric Estimates
Revenue	53	837	4735
EBITDA	55	876	4979
EBIT	75	1225	6858
NI	85	1387	7857

E. Constructing the Predictive Target: Consensus Surprise

At each window t the consensus surprise for each target metric m is calculated. We define consensus surprise $\varepsilon_{i,m,t}$ as:

$$\varepsilon_{i,m,t} = Y^{actual}_{i,m,t} - C^{cons}_{i,m,t}\,, \tag{2}$$

where $Y^{actual}_{i,m,t}$ is the eventual reported value and $C^{cons}_{i,m,t}$ is the contemporaneous consensus estimate for firm i.

F. Final Dataset Construction and Temporal Splits

For each dataset, the extracted textual features were then attached to the remaining transcript identifiers to provide the full feature set for modelling. Feature-group totals are reported as the denominators or totals-only in Table VI, Section VI-C. Datasets are then temporally partitioned by FY to avoid look-ahead bias: data points categorised as predictive input for FY 2020, 2021 and 2022 are used for training, FY 2023 for validation, and FY 2024 for testing. Transcripts are unique to each fold. Sample counts for each target metric split are shown in Table V.

TABLE V. Sample Counts by Split and Target Metric

Dataset Split	Count of Unique Transcript-Snapshot Observations			
	Revenue	*EBITDA*	*EBIT*	*NI*
Train *(FY 2020-2022)*	2,836	2,982	4,089	4,703
Validation *(FY 2023)*	973	1,028	1,417	1,618
Test *(FY 2024)*	926	969	1,352	1,536

VI. Experimental Design and Modelling Workflow

A. Prediction Task and Forecasting Objective

The regression target is the consensus surprise $\varepsilon_{i,m,t}$, as defined in (2). Accordingly, the objective is to learn a predictive function $f: \boldsymbol{x}_{i,m,t} \mapsto \hat{\varepsilon}_{i,m,t}$, where $\hat{\varepsilon}_{i,m,t}$ denotes the model's estimate of the target. The input vector $\boldsymbol{x}_{i,m,t} \in \mathbb{R}^d$ contains only information observable prior to the window end date and adheres to the earnings hierarchy, allowing only upstream raw cross-metric consensus estimates. The refined forecast $\hat{Y}_{i,m,t}$ is then computed by adding the model's fitted surprise to the raw consensus estimate of the target metric:

$$\hat{Y}_{i,m,t} = C^{cons}_{i,m,t} + \hat{\varepsilon}_{i,m,t} \tag{3}$$

B. Baseline Model

Model performance is benchmarked against a consensus baseline, which uses the raw consensus estimate $C_{i,m,t}^{cons}$ as the forecast for the target metric without any adjustment. This baseline represents the information embedded in analysts' aggregated expectations at the forecast date and serves as a realistic reference point for assessing the incremental value of alternative feature configurations and modelling approaches.

C. Feature Configurations and Screening

We evaluate the incremental value of textual and analyst-revision signals in forecasting surprises across the set of target metrics $m \in$ {Revenue, EBITDA, EBIT, NI}. For each metric independently, three configurations are tested: a "language-only" specification, which uses transcript-derived features without analyst inputs; a "consensus-dynamics-only" specification, which captures evolving changes in the analyst consensus and related signals; and a Hybrid specification, which combines both feature sets. To remove redundancy, we apply a two-stage feature screening process to the training set for each metric-configuration. A composite salience score guides selection, defined as a weighted sum of two min-max scaled terms: the absolute correlation with the target and the "Big Beat-Big Miss" Z-score gap. Stage one prunes any feature pair with pairwise absolute Pearson correlations $|\rho|>0.85$, retaining the feature with higher salience score. Stage two removes features by variance inflation factor (VIF), dropping the highest VIF iteratively until all remaining features have VIF < 6. The full procedure is provided in Algorithm A.1 in the Appendix. Screening is applied to all features except embeddings and controls; Table VI summarises retained/total for screened groups, while embeddings and controls are shown as totals-only.

TABLE VI. SCREENED FEATURE RETENTION WITH UNSCREENED TOTALS

Features		Metric-Specific Dataset			
		Revenue	***EBITDA***	***EBIT***	***NI***
Language-based	*Language Style*	4 / 5	4 / 5	4 / 5	4 / 5
	Transcript Meta	6 / 6	6 / 6	6 / 6	6 / 6
	Lexical-derived	12 / 13	12 / 13	12 / 13	12 / 13
	FinBERT Sentiment	13 / 26	13 / 26	13 / 26	13 / 26
	FinLLaMA embeddings	4,096	4,096	4,096	4,096
Consensus Dynamics	*Target Revisions*	7 / 7	6 / 7	7 / 7	7 / 7
	Cross-Metric-Related	6 / 8	6 / 12	5 / 8	5 / 9
Controls	*Macroeconomic*	2	2	2	2
	Temporal	2	2	2	2

D. Model Families and Hyperparameter Tuning

To assess the predictive utility of each feature configuration across the four metrics, we conducted systematic model development and benchmarking for each metric-feature pair. Three model classes were developed: gradient boosting machines (XGBoost, LightGBM) and multilayer perceptrons (MLPs).

XGBoost and LightGBM were included to establish strong tree-based baselines due to their widespread use in tabular financial prediction tasks (Padhi *et al.*, 2021). The models were implemented using sklearn pipelines and underwent hyperparameter tuning via grid search over task-specific parameter grids, with time-series cross-validation (TSCV) used to minimise out-of-sample mean squared error (MSE).

MLPs were implemented in PyTorch and embedded within a scikit-learn ColumnTransformer that standardised numeric fundamentals, compressed textual embeddings with PCA to reduce embeddings to 100 principal components, and one-hot encoded categorical controls. Model development followed an empirical hyperparameter strategy with successive waves of experiments exploring deeper and wider layer configurations, residual connections, and regularisation techniques such as dropout, layer normalisation, and weight decay. Training trials combined MSE loss with Adam-family optimisers, a One-Cycle learning-rate schedule, gradient clipping, mixed-precision arithmetic, and Xavier-uniform weight initialisation.

To further enhance predictive robustness and quantify uncertainty in the MLP forecasts, Monte Carlo (MC) dropout inference was implemented, generating predictive samples that were aggregated in two ways: first, a pure ensemble that averaged the draws to form point forecasts and empirical intervals, and second, an adaptive blend that transformed composite signal-to-noise (SNR) features through isotonic regression and a sigmoid function to produce residual weights. The blending coefficients and thresholds were grid-searched with out-of-fold TSCV, giving both ensemble variants calibrated forecasts that remained directly comparable with the tree-based baselines. An overview of the end-to-end MLP modelling pipeline can be understood in Fig. 7.

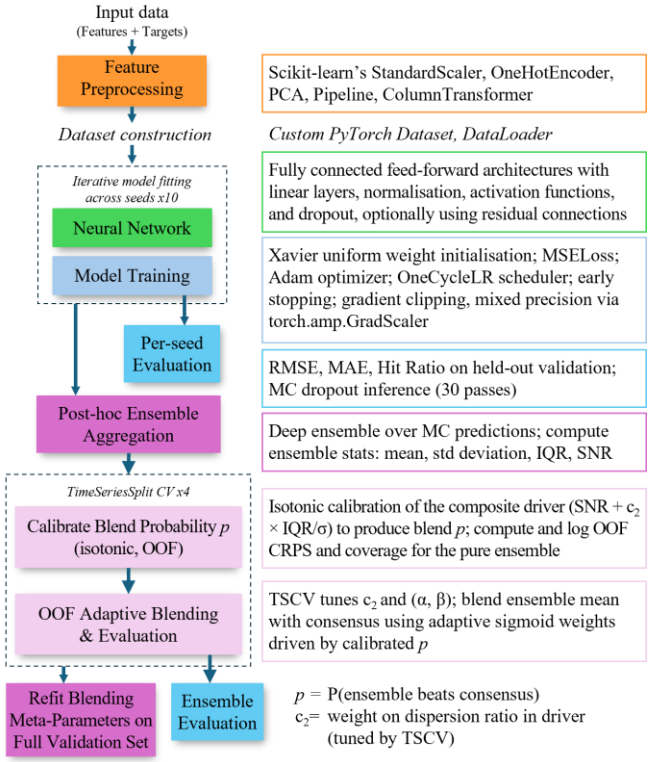

Fig. 7. End-to-end MLP modelling and evaluation pipeline, from preprocessing and multi-seed training to deep-ensemble outputs; pipeline produces two variants: pure ensemble and TSCV-tuned adaptive blend.

E. Evaluation Framework

All models, including the consensus baseline, are evaluated on identical, temporally consistent out-of-sample partitions. Accuracy is assessed primarily through mean absolute error (MAE) and sample-level model wins (Win %), where Win % is defined as the fraction of instances in which the candidate model achieves lower absolute error than the consensus baseline. In this way, MAE serves as a point-forecast accuracy metric, capturing the magnitude of typical forecast errors, while Win % reflects the model's consistency in outperforming the baseline on a case-by-case basis. Additional accuracy metrics, including root mean squared error (RMSE), mean absolute percentage error (MAPE), the coefficient of determination (R^2) and hit ratio, are reported for completeness.

F. Statistical Significance Testing Framework

Statistical differences in forecast accuracy are evaluated using the Diebold–Mariano (DM) test on absolute-error losses, with Newey–West standard errors to account for heteroskedasticity and autocorrelation, and a one-tailed alternative hypothesis that

the candidate model achieves lower MAE than the baseline, with statistical significance assessed at $\alpha = 0.05$. For each configuration, the DM statistic, its p-value (DM p), and the change in MAE (Δ MAE) relative to the baseline are reported. For MLP models, predictive distributions are generated using Monte Carlo (MC) dropout, with thirty samples per seed drawn to compute the continuous ranked probability score (CRPS) and empirical 90% prediction-interval coverage, thereby assessing uncertainty calibration. All metrics are calculated strictly on out-of-sample data to ensure fair, temporally consistent evaluation across individual models and ensemble variants.

G. Final Model Selection and Configuration

Applying the evaluation framework to validation results, MLPs consistently outperformed XGBoost and LightGBM baselines across all target metrics and feature configurations in both surprise-prediction accuracy and downstream actual-forecast accuracy. Consequently, the boosting methods were not advanced to hybrid experimentation or final out-of-sample testing.

As a result, final model selection was conducted exclusively within the MLP family. Selection was performed separately for each target metric, with the same overarching base architecture applied across its three feature configurations. Across the four metrics, two MLP architectures delivered the strongest performance, a residual-style MLP for NI and Revenue, and a more compact feed-forward MLP for EBIT and EBITDA, as shown in Fig. 8. For a given target metric, activation functions and layer structures were fixed across configurations, while training hyperparameters were tuned per configuration to accommodate the distinct statistical and structural characteristics of each feature set. Tuned elements included dropout rates, number of epochs, learning-rate schedules and annealing strategies (One-CycleLR with configuration-specific maximum learning rates), early-stopping criteria, and gradient-clipping thresholds.

Ensemble choice was also optimised per metric–configuration pair using out-of-fold TSCV, with the final selection between a pure ensemble and an adaptive-blend ensemble determined by MAE performance and corroborated by Win %. These final configurations form the basis for the out-of-sample results presented in Section VII. Detailed hyperparameters and optimisation procedures per dataset are provided in Tables A.I–A.III in the Appendix.

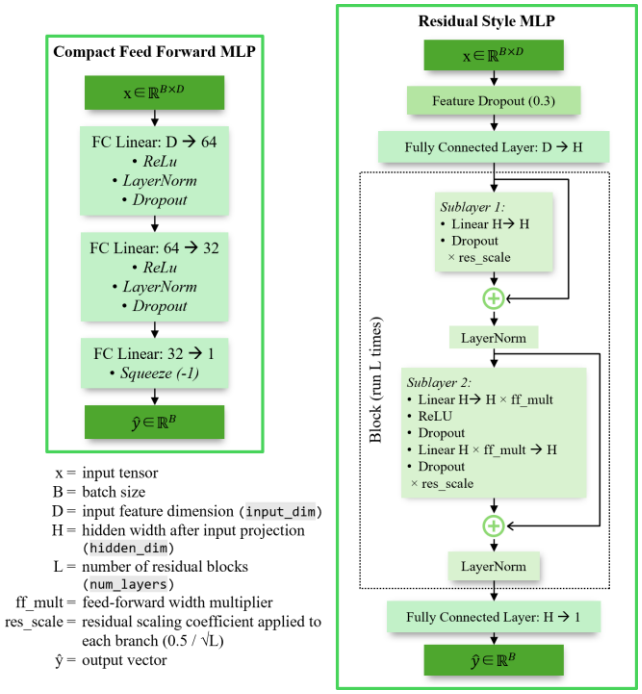

Fig. 8. *Left*: compact two-hidden-layer MLP. *Right*: residual MLP with feature dropout and L residual blocks with skip connections; skips scaled by res_scale, LayerNorm in each block, and widening in sublayer 2 via ff_mult.

VII. Results and Analysis

We evaluate FY 2024 out-of-sample forecasts using MAE, RMSE, and Win %, applying the DM test for significance, as shown in Table VII. A more detailed breakdown of results can be found in Tables A.IV–A.XI in the Appendix. Across the four earnings metrics, the chosen MLP deep ensemble "adaptive-blend" yields the largest forecast error reductions relative to the unadjusted consensus, with gains attributable to both linguistic and revision-dynamic information.

TABLE VII. Out-Of-Sample Test Results

Target and *Arch.*	Model Performance vs. Baseline					
	Model	***RMSE***	***MAE***	***Δ MAE***	***DM p***	***Win %***
Revenue: *Adaptive Blend*	Baseline	5989.72	2321.19	-	-	-
	Language-only	5980.86	2317.54	**-3.65**	0.0000	54.3%
	Cons-Dyn-only	5844.51	2309.75	-11.44	0.1064	43.5%
	Hybrid	5982.76	2318.22	-2.97	0.0000	**55.7%**
EBITDA: *Adaptive Blend*	Baseline	1575.87	522.88	-	-	-
	Language-only	1292.79	489.75	-33.13	0.0053	50.5%
	Cons-Dyn-only	1502.98	506.03	-16.85	0.0001	50.1%
	Hybrid	1224.94	472.94	**-49.95**	0.0011	**56.0%**
EBIT: *Adaptive Blend*	Baseline	2493.44	933.14	-	-	-
	Language-only	1880.91	709.19	**-223.95**	0.0000	56.0%
	Cons-Dyn-only	2286.78	915.75	-17.40	0.0641	37.4%
	Hybrid	2074.54	755.29	-177.85	0.0000	**62.9%**
NI: *Adaptive Blend*	Baseline	1851.68	755.51	-	-	-
	Language-only	1829.17	731.96	-23.55	0.0000	**66.9%**
	Cons-Dyn-only	1822.54	730.14	**-25.37**	0.0000	64.4%
	Hybrid	1828.20	731.57	-23.94	0.0000	66.0%

A. Language-Only Features

Language-only models improve MAE for every target with statistically significant DM tests with all p-values ≤0.0053. The magnitude of gains varies by metric: Revenue exhibits a small but reliable improvement of decreasing MAE by -0.16% compared to the baseline, whereas EBIT shows a larger meaningful reduction of -24.00% vs. the baseline. Collectively, these findings indicate that sentiment trajectories, long-range semantic embeddings, and stylistic cues embedded in earnings call narratives convey forward-looking information that the consensus does not fully capture.

B. Consensus-Dynamics-Only (Cons-Dyn-only) Features

Consensus-dynamics-only models add incremental value for EBITDA and NI, reducing MAE by -3.22% and -3.36% compared to baselines, respectively. For these two metrics, the gains are modest in magnitude but statistically reliable, indicating that the timing and volatility of revisions embed useful information. By contrast, results for EBIT and revenue are weaker and not statistically significant with p-values > 0.05. Together, this indicates that revision-based signals alone are not uniformly valuable across metrics.

C. Hybrid: Multi-Modal Integration

Hybrid configurations generally yield the strongest, and statistically significant, performance on operating fundamentals. Accordingly, for EBITDA, the hybrid model delivers the largest MAE reduction of -9.55%, a substantial RMSE decline of -22.27%, and a 56% sample level Win % compared to the baseline. For EBIT, the hybrid achieves a large reduction in MAE, reducing by -19.06% and with a 62.9% sample level win % compared to the baseline. Gains on Revenue and NI are smaller but statistically reliable with p-values < 0.0001.

D. Evolution of Predictive Performance After Earnings

To assess how accuracy evolves with time since disclosures,

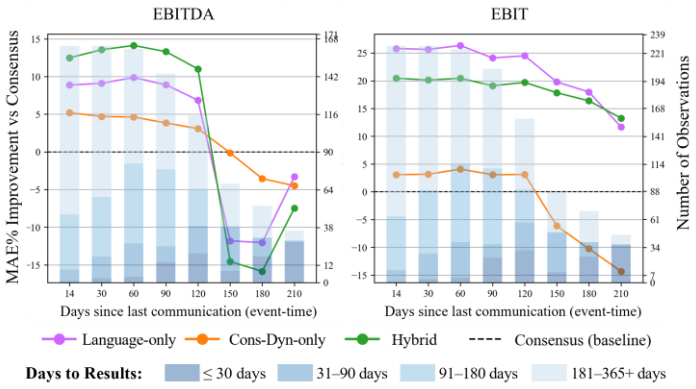

Fig. 9. Out-of-sample FY 2024 results for EBITDA and EBIT across post-call windows. Bar height shows the number of snapshots; within-bar gradient shows proximity of snapshots to year-end results. Overlaid lines report MAE improvement vs. consensus with above the line indicating better performance.

we group forecasts by post-call event buckets and report results for EBITDA and EBIT, as shown in Fig. 9. Across both metrics and all feature configurations, reductions in MAE relative to baselines are evident up to 120 days post-call. Beyond this horizon, predictive signals for EBITDA taper and turn adverse. For EBIT, consensus dynamics also turn adverse after 120 days, though language and hybrid signals retain predictive power beyond this point, albeit with gradually diminishing strength. Similar patterns can be seen for both Revenue and NI as shown in Fig. A.3 in the Appendix. Overall, observation counts fall at longer post-call event windows, so inference there warrants caution. Practically, across all metric trajectories, recalibration of the consensus using the trained models at +14, +30, +60 and even +90 days is favoured, with shrinkage toward consensus beyond ~120+ days.

E. Company-Level Illustrative Example

To contextualise the aggregate results, we use EBITDA as the illustrative metric and present Coca-Cola Europacific Partners (CCEP) as a representative case of the potential firm-level gains. As shown in Fig. 10, all three models outperform consensus across all forecast horizons, with the language model demonstrating the most pronounced advantage. Notably, following the May interim earnings call in the highlighted section, the consensus baseline moves in the wrong direction, while the language and hybrid models correct the trajectory and significantly improve accuracy.

VIII. Conclusion

A. Summary of Findings

In sum, comparing models to the consensus baseline on the FY-2024 test set shows systematic out-of-sample gains from both linguistic and consensus-dynamics information, with the hybrid configuration delivering the most consistent case-level dominance (higher Win%) but not uniformly the largest MAE reductions. Language-only achieves the strongest average error reductions for EBIT (−24.00% MAE vs. baseline) and a small but reliable gain for Revenue; the hybrid is best on EBITDA (−9.55% MAE), and consensus-dynamics-only narrowly leads NI. These patterns indicate complementarity rather than strict dominance: transcripts embed durable, forward-looking signals that consensus underweights, while revision dynamics add context whose value depends on the target metric.

B. Conclusion

By capturing how the analyst consensus evolves between earnings calls and linking these changes to the language of the latest transcript, this study provides a reproducible way to examine the ongoing market impact of corporate disclosures. Preserving the exact information available at each point in time removes look-ahead bias and creates a foundation for studying how unstructured communication interacts with shifting

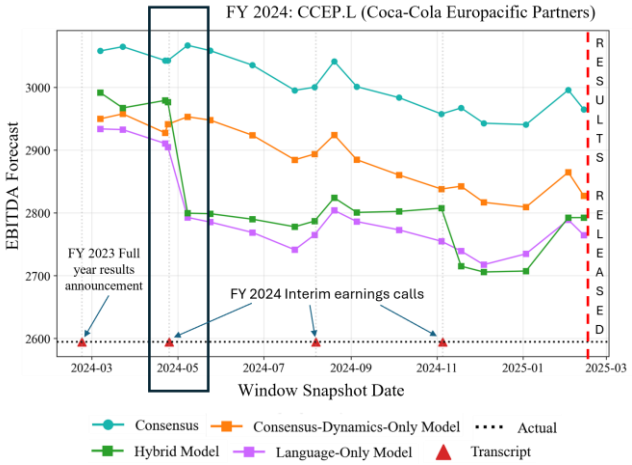

Fig. 10. EBITDA FY2024 out-of-sample forecasts for CCEP, comparing model forecasts against the consensus baseline, with key disclosure events marked.

expectations across multiple horizons. In doing so, it addresses a key barrier in financial NLP and expands the scope for robust, time-aware predictive modelling.

Building on this foundation, we find that combining interpretable lexicon indicators, utterance-level sentiment from FinBERT and domain embeddings from FinLLaMA is an effective representation strategy for corporate transcripts. These features alone often outperform both consensus-dynamics and hybrid inputs in terms of MAE, underscoring the predictive value of textual information. Even so, adding consensus revision dynamics improves robustness, with the hybrid model achieving higher Model Win % overall. For EBITDA in particular, combining language and consensus variables delivers MAE and accuracy gains that neither source achieves on its own, directly supporting the project's research objective and demonstrating the value of a multi-modal forecasting framework. At the architecture level, the compact MLP delivered larger MAE reductions on EBIT and EBITDA, consistent with a smaller model reducing overfitting. By contrast, the residual MLP showed muted MAE gains, consistent with greater sensitivity to noise from its extra flexibility and skip connections; adaptive blending steadied its outputs and increased Win%, even when MAE gains were modest, showing the value of advanced deep learning methods.

Taken together, the pipeline provides a practical path from static surprise snapshots to multi-horizon, uncertainty-aware inference on real disclosures and consensus revisions. The results show that aligning unstructured language with time-varying analyst consensus changes supports richer, horizon-aware predictive models and offers a scalable, reproducible framework for multimodal financial NLP.

C. Limitations and Future Work

Several technical limits apply. The study covers FTSE 100 companies from FY 2019-2024 and uses a single test year; event-time bins thin out beyond about 120 to 150 days, lowering statistical power. The target distribution is heavy-tailed; we did not apply reweighting or transformation to preserve realism, which limits performance in the extremes. Future work within this framework includes multi-year validation and final testing under different conditions, horizon-aware weighting across event time, a staged hit/miss-then-regression design per metric, and using predicted fundamentals to forecast EPS and assess links to post-announcement returns.

Acknowledgement

Thanks are due to Dr. William Marsh for his support, advice and guidance in writing this paper. Thank you also to Alexandra Bull (OneIM) and Gregory Keddy, CFA (LGIM) for their invaluable financial industry insight.

References

Angelo, B. *et al.* (2025) "Tone Distance: Managerial Tone Divergence and Market Reaction to Earnings Announcements," *Financial Review* [Preprint]. Available at: https://doi.org/10.1111/fire.70002.

Araci, D. (2019) "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv preprint* [Preprint]. Available at: http://arxiv.org/abs/1908.10063.

Ball, R. and Brown, P. (1968) "An Empirical Evaluation of Accounting Income Numbers," *Journal of Accounting Research*, 6(2), pp. 159–178. Available at: https://doi.org/10.2307/2490232.

Bernard, V.L. and Thomas, J.K. (1989) "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium?," *Journal of Accounting Research*, 27, pp. 1–36. Available at: https://doi.org/10.2307/2491062.

Beyer, A. *et al.* (2010) "The financial reporting environment: Review of the recent literature," *Journal of Accounting and Economics*, 50(2–3), pp. 296–343. Available at: https://doi.org/10.1016/j.jacceco.2010.10.003.

Bilinski, P. (2024) "Beyond the street EPS surprise–when 'other surprises' matter in explaining earnings announcement returns," *Accounting and Business Research* [Preprint]. Available at: https://doi.org/10.1080/00014788.2024.2400875.

Van Binsbergen, J.H., Han, X. and Lopez-Lira, A. (2023) "Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases," *Review of Financial Studies*, 36(6), pp. 2361–2396. Available at: https://doi.org/10.1093/rfs/hhac085.

Bozanic, Z., Roulstone, D.T. and Van Buskirk, A. (2018) "Management earnings forecasts and other forward-looking statements," *Journal of Accounting and Economics*, 65(1), pp. 1–20. Available at: https://doi.org/10.1016/J.JACCECO.2017.11.008.

Bradshaw, M.T. *et al.* (2021) "Soft Information in the Financial Press and Analyst Revisions," *The Accounting Review*, 96(5), pp. 107–132. Available at: https://doi.org/10.2308/TAR-2018-0264.

Bradshaw, M.T., Richardson, S.A. and Sloan, R.G. (2001) "Do analysts and auditors use information in accruals?," *Journal of Accounting Research*, 39(1), pp. 45–74. Available at: https://doi.org/10.1111/1475-679X.00003.

Chen, X. *et al.* (2022) "Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data," *Journal of Accounting Research*, 60(2), pp. 467–515. Available at: https://doi.org/10.1111/1475-679X.12429.

Chung, A. and Tanaka-Ishii, K. (2023) "Predictability of Post-Earnings Announcement Drift with Textual and Contextual Factors of Earnings Calls," in *ICAIF 2023 – 4th ACM International Conference on AI in Finance*. Association for Computing Machinery, Inc, pp. 401–408. Available at: https://doi.org/10.1145/3604237.3626861.

Clement, M.B. (1999) "Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter?," *Journal of Accounting and Economics*, 27(3), pp. 285–303. Available at: https://doi.org/10.1016/S0165-4101(99)00013-0.

Du, K. *et al.* (2024) "Financial Sentiment Analysis: Techniques and Applications," *ACM Computing Surveys*, 56(9). Available at: https://doi.org/10.1145/3649451.

Easton, P. *et al.* (2002) "Using forecasts of earnings to simultaneously estimate growth and the rate of return on equity investment," *Journal of Accounting Research*, 40(3), pp. 657–676. Available at: https://doi.org/10.1111/1475-679X.00066.

Ertimur, Y., Livnat, J. and Martikainen, M. (2003) *Differential Market Reactions to Revenue and Expense Surprises*, *Review of Accounting Studies*.

Ertimur, Y., Sunder, J. and Sunder, S. V. (2007) "Measure for measure: The relation between forecast accuracy and recommendation profitability of analysts," *Journal of Accounting Research*, 45(3), pp. 567–606. Available at: https://doi.org/10.1111/j.1475-679X.2007.00244.x.

Fladerer, M.P. *et al.* (2021) "The Value of Speaking for 'Us,'" *Journal of Business and Psychology*, 36(2), pp. 299–313. Available at: https://doi.org/10.2307/48763107.

Fraser, B. (1999) "What are discourse markers?," *Journal of Pragmatics*, 31(7), pp. 931–952. Available at: https://doi.org/10.1016/S0378-2166(98)00101-5.

Han, L.J. (2025) "Announcements, expectations, and stock returns with asymmetric information," *Journal of Monetary Economics*, 151. Available at: https://doi.org/10.1016/j.jmoneco.2025.103751.

Hand, J.R.M. *et al.* (2022) "Explaining firms' earnings announcement stock returns using FactSet and I/B/E/S data feeds," *Review of Accounting Studies*, 27(4), pp. 1389–1420. Available at: https://doi.org/10.1007/s11142-021-09597-6.

Hess, D., Simon, F. and Weibels, S. (2025) "Interpretable Machine Learning for Earnings Forecasts: Leveraging High-Dimensional Financial Statement Data *," *SSRN Working Paper*, pp. 1–70. Available at: https://doi.org/https://dx.doi.org/10.2139/ssrn.4619313.

Hou, K., van Dijk, M.A. and Zhang, Y. (2012) "The implied cost of capital: A new approach," *Journal of Accounting and Economics*, 53(3), pp. 504–526. Available at: https://doi.org/10.1016/j.jacceco.2011.12.001.

Huang, A.H., Wang, H. and Yang, Y. (2023) "FinBERT: A Large Language Model for Extracting Information from Financial Text*," *Contemporary Accounting Research*, 40(2), pp. 806–841. Available at: https://doi.org/10.1111/1911-3846.12832.

Hyland, K. (2018) *Metadiscourse : Exploring Interaction in Writing*. 2018 reissued. London, UNITED KINGDOM: Bloomsbury Publishing Plc. Available at: http://ebookcentral.proquest.com/lib/gmul-ebooks/detail.action?docID=5560196.

Iacovides, G. *et al.* (2024) "FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading," in *ICAIF 2024 – 5th ACM International Conference on AI in Finance*. Association for Computing Machinery, Inc, pp. 134–141. Available at: https://doi.org/10.1145/3677052.3698696.

Jegadeesh, N. and Livnat, J. (2006) "Revenue surprises and stock returns," *Journal of Accounting and Economics*, 41(1–2), pp. 147–171. Available at: https://doi.org/10.1016/j.jacceco.2005.10.003.

Jiang, F. *et al.* (2019) “Manager sentiment and stock returns,” *Journal of Financial Economics*, 132(1), pp. 126–149. Available at: https://doi.org/10.1016/j.jfineco.2018.10.001.

Kim, Y., Lobo, G.J. and Song, M. (2011) “Analyst characteristics, timing of forecast revisions, and analyst forecasting ability,” *Journal of Banking and Finance*, 35(8), pp. 2158–2168. Available at: https://doi.org/10.1016/j.jbankfin.2011.01.006.

Kinney, W., Burgstahler, D. and Martin, R. (2002) “Earnings surprise ‘materiality’ as measured by stock returns,” *Journal of Accounting Research*. Blackwell Publishing Inc., pp. 1297–1329. Available at: https://doi.org/10.1111/1475-679X.t01-1-00055.

Koval, R., Andrews, N. and Yan, X. (2023) “Forecasting Earnings Surprises from Conference Call Transcripts,” in A. Rogers, J. Boyd-Graber, and N. Okazaki (eds) *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 8197–8209. Available at: https://doi.org/10.18653/v1/2023.findings-acl.520.

Li, F. (2008) “Annual report readability, current earnings, and earnings persistence,” *Journal of Accounting and Economics*, 45(2–3), pp. 221–247. Available at: https://doi.org/10.1016/j.jacceco.2008.02.003.

Li, F. (2010) “The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach,” *Journal of Accounting Research*, 48(5), pp. 1049–1102. Available at: https://doi.org/10.1111/j.1475-679X.2010.00382.x.

Li, L., Liu, K. and Li, G. (2024) “What can analyst forecasts tell us about imperfect information?,” *International Review of Economics and Finance*, 92, pp. 1059–1073. Available at: https://doi.org/10.1016/j.iref.2024.02.071.

Livnat, J. and Mendenhall, R.R. (2006) “Comparing the post-earnings announcement drift for surprises calculated from analyst and time series forecasts,” *Journal of Accounting Research*, 44(1), pp. 177–205. Available at: https://doi.org/10.1111/j.1475-679X.2006.00196.x.

Loh, R.K. and Mian, G.M. (2006) “Do accurate earnings forecasts facilitate superior investment recommendations?,” *Journal of Financial Economics*, 80(2), pp. 455–483. Available at: https://doi.org/10.1016/j.jfineco.2005.03.009.

Loughran, T. and McDonald, B. (2011) “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks,” *Journal of Finance*, 66(1), pp. 35–65. Available at: https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Merkl-Davies, D.M., Brennan, N.M. and McLeay, S.J. (2011) “Impression management and retrospective sense-making in corporate narratives: A social psychology perspective,” *Accounting, Auditing and Accountability Journal*, 24(3), pp. 315–344. Available at: https://doi.org/10.1108/09513571111124036.

Ni, H. *et al.* (2024) “Harnessing Earnings Reports for Stock Predictions: A QloRA-Enhanced LLM Approach,” in *2024 6th International Conference on Data-Driven Optimization of Complex Systems, DOCS 2024*. Institute of Electrical and Electronics Engineers Inc., pp. 909–915. Available at: https://doi.org/10.1109/DOCS63458.2024.10704454.

Padhi, D.K. *et al.* (2021) “A fusion framework for forecasting financial market direction using enhanced ensemble models and technical indicators,” *Mathematics*, 9(21). Available at: https://doi.org/10.3390/math9212646.

Park, M. and Zach, T. (2025) “Analysts’ forecasting models and uncertainty about the past,” *Review of Accounting Studies* [Preprint]. Available at: https://doi.org/10.1007/s11142-025-09898-0.

Pope, P.F. and Wang, T. (2023) “Analyst ability and research effort: non-EPS forecast provision as a research quality signal,” *Review of Accounting Studies*, 28(3), pp. 1263–1315. Available at: https://doi.org/10.1007/s11142-023-09791-8.

Qian, L. *et al.* (2025) “Fino1: On the Transferability of Reasoning-Enhanced LLMs and Reinforcement Learning to Finance,” *arXiv preprint*, arXiv:2502.08127. Available at: https://doi.org/https://doi.org/10.48550/arXiv.2502.08127.

Stickel, S.E. (1991) *Common Stock Returns Surrounding Earnings Forecast Revisions: More Puzzling Evidence*, *Source: The Accounting Review*.

Wang, Z. *et al.* (2025) “Temporal Evolution of Sentiment in Earnings Calls and Its Relationship with Financial Performance,” *Applied and Computational Engineering*, 141, pp. 195–206. Available at: https://doi.org/10.54254/2755-2721/2025.21983.

APPENDIX A

A. Figures

	transcript_id	company_name	transcript_name	published_date	category
0	T_d634c664	Airtel Africa plc	Airtel Africa Plc - Shareholder/Analyst Call	2024-07-03	Shareholder/Analyst Call
1	T_ea266cd0	Airtel Africa plc	Airtel Africa Plc, 2020 Earnings Call, May 13,...	2020-05-13	2020 Earnings Call
2	T_2d036c36	Airtel Africa plc	Airtel Africa Plc, 2021 Earnings Call, May 12,...	2021-05-12	2021 Earnings Call
3	T_55bbaea9	Airtel Africa plc	Airtel Africa Plc, 2022 Earnings Call, May 11,...	2022-05-11	2022 Earnings Call
4	T_bfcf523c	Airtel Africa plc	Airtel Africa Plc, 2023 Earnings Call, May 11,...	2023-05-11	2023 Earnings Call
...	...	...	...	...	...
2312	T_d8657e17	Whitbread PLC	Whitbread plc, Q3 2022 Sales/ Trading Statemen...	2022-01-12	Q3 2022 Sales/ Trading Statement Call
2313	T_2e1627db	Whitbread PLC	Whitbread plc, Q3 2023 Sales/ Trading Statemen...	2023-01-12	Q3 2023 Sales/ Trading Statement Call
2314	T_056851ba	Whitbread PLC	Whitbread plc, Q3 2024 Sales/ Trading Statemen...	2024-01-11	Q3 2024 Sales/ Trading Statement Call
2315	T_22e7fb26	Whitbread PLC	Whitbread plc, Q3 2025 Sales/ Trading Statemen...	2025-01-16	Q3 2025 Sales/ Trading Statement Call
2316	T_18b63abc	Whitbread PLC	Whitbread plc, Q4 2025 Earnings Call, May 01, ...	2025-05-01	Q4 2025 Earnings Call

2317 rows × 5 columns

Fig. A 1. Screenshot of transcript relational table after HTML document preprocessing Screenshot of the relational table derived from transcript data following HTML preprocessing. Each row corresponds to a transcript segment, with metadata and content fields extracted into normalised columns, enabling systematic querying and feature engineering.

	utterance_id	transcript_id	company_name	call_category	published_date	section_id	section_type	section_sequence	turn_order
0	U_50213c13	T_d634c664	Airtel Africa plc	Shareholder/Analyst Call	2024-07-03	SEC_d23858e3	Presentation	1	1
1	U_fc67bf13	T_d634c664	Airtel Africa plc	Shareholder/Analyst Call	2024-07-03	SEC_d23858e3	Presentation	1	2
2	U_87ed796d	T_d634c664	Airtel Africa plc	Shareholder/Analyst Call	2024-07-03	SEC_d23858e3	Presentation	1	3
3	U_d84fb99d	T_d634c664	Airtel Africa plc	Shareholder/Analyst Call	2024-07-03	SEC_d23858e3	Presentation	1	4
4	U_c3ce4499	T_d634c664	Airtel Africa plc	Shareholder/Analyst Call	2024-07-03	SEC_4a8ccf68	Q&A	2	1
...	...	...	...	...	...	...	...	...	...
115379	U_cd7f335d	T_dc3d8755	NatWest Group plc	Shareholder/Analyst Call	2024-04-23	SEC_e0a13119	Q&A	2	13
115380	U_003972d1	T_dc3d8755	NatWest Group plc	Shareholder/Analyst Call	2024-04-23	SEC_e0a13119	Q&A	2	14
115381	U_75751f1c	T_dc3d8755	NatWest Group plc	Shareholder/Analyst Call	2024-04-23	SEC_e0a13119	Q&A	2	15
115382	U_2df186a1	T_dc3d8755	NatWest Group plc	Shareholder/Analyst Call	2024-04-23	SEC_e0a13119	Q&A	2	16
115383	U_86d5c3c1	T_dc3d8755	NatWest Group plc	Shareholder/Analyst Call	2024-04-23	SEC_e0a13119	Q&A	2	17

115384 rows × 17 columns

speaker_id	speaker_name	speaker_role	speaker_company	raw_text	text	token_len	speaker_cat
S_62912057	Sunil Mittal	Non-Executive Chair	Airtel Africa plc	Ladies and gentlemen, I would like to welcome ...	Ladies and gentlemen, I would like to welcome ...	374	EXEC
S_54df2678	Unknown Executive	Manager	Airtel Africa plc	Thank you, Chairman. Voting on all resolutions...	Thank you, Chairman. Voting on all resolutions...	457	EXEC
S_62912057	Sunil Mittal	Non-Executive Chair	Airtel Africa plc	Thank you, Simon. In a challenging year, marke...	Thank you, Simon. In a challenging year, marke...	667	EXEC
S_54df2678	Unknown Executive	Manager	Airtel Africa plc	[indiscernible] has a question.	[indiscernible] has a question.	4	EXEC
S_4e7012dd	Unknown Attendee	Attendees	Unknown	I'm [Sam Holland]. I've got a question in te...	I'm [Sam Holland]. I've got a question in te...	67	EXTERNAL
...	...	...	...	...	...	...	...
S_7f76df30	Unknown Attendee	Attendee	Unknown	I'm hoping this is working? Is it?	I'm hoping this is working? Is it?	7	EXTERNAL
S_a30b6c50	Richard Haythornthwaite	Executive Board Member	NatWest Group plc	Yes, it's working.	Yes, it's working.	3	EXEC
S_7f76df30	Unknown Attendee	Attendee	Unknown	Just a shareholder, since 2008, the biggest sh...	Just a shareholder, since 2008, the biggest sh...	49	EXTERNAL
S_a30b6c50	Richard Haythornthwaite	Executive Board Member	NatWest Group plc	Yes. I think -- so this is a matter of percept...	Yes. I think -- so this is a matter of percept...	221	EXEC
S_a30b6c50	Richard Haythornthwaite	Executive Board Member	NatWest Group plc	I now declare the voting closed. All right. Th...	I now declare the voting closed. [PAR] All rig...	87	EXEC

Fig. A 2. Screenshot of the relational table of utterances obtained after preprocessing the HTML transcript document. Each row corresponds to a single speaker utterance, with text content and associated metadata extracted into normalised columns to enable structured querying and downstream feature construction. First image shows columns 1-9, second image shows columns 10-17.

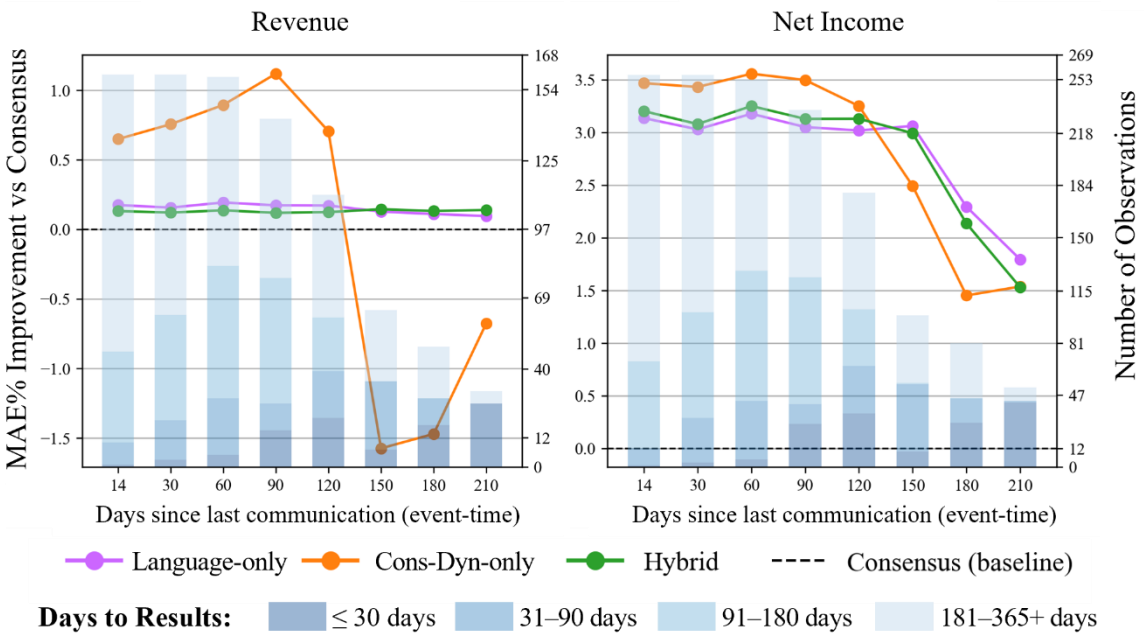

Fig. A 3. Out-of-sample FY 2024 results for Revenue and Net Income across post-call windows. Bar height shows the number of snapshots; within-bar gradient shows proximity of snapshots to year-end results. Overlaid lines report MAE improvement vs. consensus with above the line indicating better performance.

B. Algorithms

Algorithm A.1. Iterative VIF–Correlation Feature Screening

Require: Feature matrix $\mathbf{X}$; absolute correlations ρ; gap score[b] g; weights (w_1, w_2); VIF threshold τ; pairwise threshold κ, mandatory set $\mathcal{M}$

Ensure: Pruned feature set $\mathcal{F}$ and final VIFs

$\mathcal{F}$ = all columns in $\mathbf{X}$

Compute priority score $\pi_j = w_1 \cdot \text{scale}(|\rho_j|) + w_2 \cdot \text{scale}(g_j)$

while TRUE **do**

Compute $|\text{corr}(x_i, x_j)|$ for all $(I, j) \in \mathcal{F}$

if any $> \kappa$ **then**

Drop lower π non-mandatory feature (resolve one pair per-iteration)

continue

end if

Compute VIF_j for all $x_j \in \mathcal{F}$

if all $\text{VIF}_j \leq \tau$ **then**

break

else

Drop non-mandatory feature with highest VIF (tie-break on lowest π)

end if

end while

Recompute final VIFs on cleaned $\mathcal{F}$

return pruned feature set $\mathcal{F}$ and corresponding VIFs

b. Absolute mean Z-score gap between Big Beat and Big Miss (>±10%) surprise groups per feature

C. Tables

TABLE A.I Model Class Configurations by Dataset

Dataset	Model	Model Class Configurations
EBITDA	LangMLP	input_dim → [64, ReLU, LN, Dropout(0.2)] → [32, ReLU, LN, Dropout(0.1)] → [1]
	AnalystMLP	input_dim → [64, ReLU, LN, Dropout(0.3)] → [32, ReLU, LN, Dropout(0.2)] → [1]
	HybridMLP	input_dim → [64, ReLU, LN, Dropout(0.4)] → [32, ReLU, LN, Dropout(0.3)] → [1]
EBIT	LangMLP	input_dim → [64, ReLU, LN, Dropout(0.2)] → [32, ReLU, LN, Dropout(0.1)] → [1]
	AnalystMLP	input_dim → [64, ReLU, LN, Dropout(0.3)] → [32, ReLU, LN, Dropout(0.2)] → [1]
	HybridMLP	input_dim → [64, ReLU, LN, Dropout(0.4)] → [32, ReLU, LN, Dropout(0.3)] → [1]
Revenue	LangResMLP AnalystResMLP HybridResMLP	input_dim → Linear(128) → [(Linear(128→128), residual scale = 0.5/√L, LayerNorm) + (FeedForward 128→256→128, ReLU, residual scale = 0.5/√L, LayerNorm)] ×2 → LayerNorm → Linear(128→1) with feature dropout = 0.3, and dropout = 0.4 applied after the attention-like linear layer, between the two feedforward layers, and after the feedforward output. Num_layers = 2 = L
Net Income	LangResMLP AnalystResMLP	input_dim → Linear(128) → [(Linear(128→128), residual scale = 0.5/√L, LayerNorm) + (FeedForward 128→256→128, ReLU, residual scale = 0.5/√L, LayerNorm)] ×2 → LayerNorm → Linear(128→1) with feature dropout = 0.3, and dropout = 0.4 applied after the attention-like linear layer, between the two feedforward layers, and after the feedforward output. Num_layers = 2 = L
	HybridResMLP	input_dim → Linear(128) → [(Linear(128→128), residual scale = 0.5/√L, LayerNorm) + (FeedForward 128→256→128, ReLU, residual scale = 0.5/√L, LayerNorm)] ×2 → LayerNorm → Linear(128→1) with feature dropout = 0.3, and dropout = 0.3 applied after the attention-like linear layer, between the two feedforward layers, and after the feedforward output. Num_layers = 2 = L

TABLE A.II Hyperparameters for Each Model by Dataset

Dataset	Model	Batch size	Epochs	Optimizer	LR	Weight decay	Grad clip	Early stop	Scheduler	Scheduler args
EBITDA	LangMLP	28	100	Adam	5e-4	1e-6	0.5	15	OneCycleLR	max_lr=5e-3, total_steps=100×steps_per_epoch
EBITDA	AnalystMLP	64	150	Adam	5e-4	1e-3	0.5	25	OneCycleLR	max_lr=2e-3, pct_start=0.1, anneal=cos, total_steps=150×steps_per_epoch
EBITDA	HybridMLP	128	150	Adam	5e-4	1e-5	0.5	20	OneCycleLR	max_lr=4.5e-3, pct_start=0.1, anneal=cos, total_steps=150×steps_per_epoch
EBIT	LangMLP	128	100	Adam	5e-4	1e-6	1	15	OneCycleLR	max_lr=5e-3, total_steps=100×steps/epoch
EBIT	AnalystMLP	64	200	Adam	5e-4	1e-3	1	25	OneCycleLR	max_lr=1.5e-3, pct_start=0.1, anneal=cos, total_steps=200×steps/epoch
EBIT	HybridMLP	64	100	Adam	5e-4	1e-6	1	15	OneCycleLR	max_lr=2e-3, pct_start=0.1, anneal=cos, total_steps=100×steps/epoch
Revenue	LangResMLP	64	100	Adam	5e-6	1e-4	0.5	15	OneCycleLR	max_lr=1e-4, pct_start=0.1, anneal=linear, total_steps=100×steps/epoch
Revenue	AnalystResMLP	64	100	Adam	5e-4	1e-4	1	15	OneCycleLR	max_lr=1.5e-3, pct_start=0.1, anneal=linear, total_steps=100×steps/epoch
Revenue	HybridResMLP	128	100	Adam	5e-5	1e-5	0.5	20	OneCycleLR	max_lr=1e-4, pct_start=0.1, anneal=linear, total_steps=100×steps/epoch
Net Income	LangResMLP	64	100	Adam	5e-6	1e-4	0.5	15	OneCycleLR	max_lr=1e-4, pct_start=0.1, anneal=linear, total_steps=100×steps/epoch
Net Income	AnalystResMLP	64	100	Adam	5e-6	1e-4	1	15	OneCycleLR	max_lr=1e-4, pct_start=0.1, anneal=linear, total_steps=100×steps/epoch
Net Income	HybridResMLP	64	100	Adam	5e-6	1e-4	1	15	OneCycleLR	max_lr=1e-4, pct_start=0.1, anneal=linear, total_steps=100×steps/epoch

TABLE A.III Key settings and Hyperparameters for Deep Ensemble Construction

Step	Description	Key settings / Hyperparameters
Inputs	Collect residual samples from all ensemble members (MC dropout × seeds)	mc_passes = 30, deep_ensemble = True
Pure Ensemble	Consensus + mean residual; evaluate directly.	Report: MAE, RMSE, HitRatio, CRPS, Coverage
Adaptive Blend	Out-of-fold (4-fold) adaptive correction using SNR, IQR, variance, p_win	Grids: c2 ∈ {0,0.25,0.5,0.75,1.0}, α ∈ {-8..8}, β ∈ {-4..4}; fallback = consensus if NaN
Meta-param refit	Refit on all data: isotonic composite + adaptive sigmoid correction	Store c2, isotonic thresholds (x,y), threshold, α, β; driver = isotonic_composite
Final output	Ensemble predictions + calibrated meta-parameters stored for test-time use	adaptive = sigmoid-mae

TABLE A.IV Revenue Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.1

Target Metric	Model Type	Method	RMSE	MAE	MAPE	R2
Revenue	Baseline	Consensus	5989.72	2321.19	0.1268	0.80
Revenue	Task A: Language-only	Ensemble - AdaptiveBlend-MAE	5980.86	2317.54	0.1271	0.80
Revenue	Task B: Analyst-only	Ensemble - AdaptiveBlend-MAE	5844.51	2309.75	0.1313	0.81
Revenue	Task C: Hybrid	Ensemble - AdaptiveBlend-MAE	5982.76	2318.22	0.1272	0.80
Revenue	Task A: Language-only	Pure Ensemble	5980.78	2317.52	0.1271	0.80
Revenue	Task B: Analyst-only	Pure Ensemble	5844.36	2309.99	0.1314	0.81
Revenue	Task C: Hybrid	Pure Ensemble	5982.76	2318.22	0.1272	0.80

TABLE A.V Revenue Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.2

Target Metric	Model Type	Method	HitRatio	DM score	P value	Δ MAE	Model wins %
Revenue	Task A: Language-only	Ensemble - AdaptiveBlend-MAE	0.5551	-4.7030	0.0000	-3.65	54.30%
Revenue	Task B: Analyst-only	Ensemble - AdaptiveBlend-MAE	0.5475	-1.2460	0.1064	-11.44	43.50%
Revenue	Task C: Hybrid	Ensemble - AdaptiveBlend-MAE	0.5680	-5.3350	0.0000	-2.97	55.70%
Revenue	Task A: Language-only	Pure Ensemble	0.5551	-4.6960	0.0000	-3.68	54.30%
Revenue	Task B: Analyst-only	Pure Ensemble	0.5475	-1.2170	0.1118	-11.20	43.10%
Revenue	Task C: Hybrid	Pure Ensemble	0.5680	-5.3350	0.0000	-2.97	55.70%

TABLE A.VI EBITDA Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.1

Target Metric	Model Type	Method	RMSE	MAE	MAPE	R^2
EBITDA	Baseline	Consensus	1575.87	522.88	0.24	0.96
EBITDA	Language-only	Ensemble - AdaptiveBlend-MAE	1292.79	489.75	0.26	0.97
EBITDA	Analyst-only	Ensemble - AdaptiveBlend-MAE	1502.98	506.03	0.24	0.96
EBITDA	Hybrid	Ensemble - AdaptiveBlend-MAE	1224.94	472.94	0.26	0.97
EBITDA	Language-only	Pure Ensemble	1288.39	492.97	0.27	0.97
EBITDA	Analyst-only	Pure Ensemble	1502.26	506.64	0.24	0.96
EBITDA	Hybrid	Pure Ensemble	1220.50	487.05	0.28	0.97

TABLE A.VII EBITDA Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.2

Target Metric	Model Type	Method	HitRatio	DM score	P value	Δ MAE	Model wins %
EBITDA	Task A: Language-only	Ensemble - AdaptiveBlend-MAE	0.8215	-2.553	0.0053	-33.1295	50.50%
EBITDA	Task B: Analyst-only	Ensemble - AdaptiveBlend-MAE	0.8019	-3.657	0.0001	-16.8458	50.10%
EBITDA	Task C: Hybrid	Ensemble - AdaptiveBlend-MAE	0.8122	-3.061	0.0011	-49.945	56.00%
EBITDA	Task A: Language-only	Pure Ensemble	0.8215	-2.244	0.0124	-29.9108	49.30%
EBITDA	Task B: Analyst-only	Pure Ensemble	0.8019	-3.489	0.0002	-16.2394	49.90%
EBITDA	Task C: Hybrid	Pure Ensemble	0.8122	-2.105	0.0176	-35.8267	47.20%

TABLE A.VIII EBIT Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.1

Target Metric	Model Type	Method	RMSE	MAE	MAPE	R^2
EBIT	Baseline	Consensus	2493.4423	933.1441	0.7718	0.5597
EBIT	Language-only	Ensemble - AdaptiveBlend-MAE	1880.91	709.19	0.5936	0.75
EBIT	Analyst-only	Ensemble - AdaptiveBlend-MAE	2286.78	915.75	0.8245	0.63
EBIT	Hybrid	Ensemble - AdaptiveBlend-MAE	2074.54	755.29	0.6135	0.70
EBIT	Language-only	Pure Ensemble	1880.58	709.31	0.5936	0.75
EBIT	Analyst-only	Pure Ensemble	2283.96	918.37	0.8293	0.63
EBIT	Hybrid	Pure Ensemble	2074.22	755.28	0.6134	0.70

TABLE A.IX EBIT Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.2

Target Metric	Model Type	Method	HitRatio	DM score	P value	Δ MAE	Model wins %
EBIT	Task A: Language-only	Ensemble - AdaptiveBlend-MAE	0.8735	-10.473	0.00	-223.95	56.00%
EBIT	Task B: Analyst-only	Ensemble - AdaptiveBlend-MAE	0.8646	-1.521	0.0641	-17.40	37.40%
EBIT	Task C: Hybrid	Ensemble - AdaptiveBlend-MAE	0.8772	-12.574	0.00	-177.85	62.90%
EBIT	Task A: Language-only	Pure Ensemble	0.8735	-10.46	0.00	-223.83	55.80%
EBIT	Task B: Analyst-only	Pure Ensemble	0.8646	-1.27	0.1021	-14.78	37.10%
EBIT	Task C: Hybrid	Pure Ensemble	0.8772	-12.561	0.00	-177.87	62.60%

TABLE A.X Net Income Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.1

Target Metric	Model Type	Method	RMSE	MAE	MAPE	R2
NI	Baseline	Consensus	1851.68	755.51	1.0165	0.5800
NI	Task A: Language-only	Ensemble - AdaptiveBlend-MAE	1829.17	731.96	0.9449	0.5901
NI	Task B: Analyst-only	Ensemble - AdaptiveBlend-MAE	1822.54	730.14	0.9452	0.5931
NI	Task C: Hybrid	Ensemble - AdaptiveBlend-MAE	1828.20	731.57	0.9455	0.5905
NI	Task A: Language-only	Pure Ensemble	1829.11	731.93	0.9449	0.5901
NI	Task B: Analyst-only	Pure Ensemble	1822.54	730.14	0.9452	0.5931
NI	Task C: Hybrid	Pure Ensemble	1828.17	731.56	0.9455	0.5906

TABLE A.XI Net Income Full Breakdown of Performance Results for Pure Ensemble and Adaptive Blend – Pt.1

Target Metric	Model Type	Method	HitRatio	DM score	P value	Δ MAE	Model wins %
NI	Task A: Language-only	Ensemble - AdaptiveBlend-MAE	0.7559	-16.0150	0.0000	-23.55	66.90%
NI	Task B: Analyst-only	Ensemble - AdaptiveBlend-MAE	0.7715	-13.6500	0.0000	-25.37	64.40%
NI	Task C: Hybrid	Ensemble - AdaptiveBlend-MAE	0.7526	-15.5410	0.0000	-23.94	66.00%
NI	Task A: Language-only	Pure Ensemble	0.7559	-15.9890	0.0000	-23.58	66.90%
NI	Task B: Analyst-only	Pure Ensemble	0.7715	-13.6470	0.0000	-25.37	64.40%
NI	Task C: Hybrid	Pure Ensemble	0.7526	-15.5290	0.0000	-23.95	65.90%

TABLE A.XII Full List of Terms Included in Custom Lexicons

Self-constructed category	*Thematic groupings*	**Terms included**
Contrastive terms		albeit, although, but, contrary to, conversely, despite, even though, even so, even if, however, in spite of, in contrast, instead, much as, nevertheless, notwithstanding, nonetheless, on the contrary, on the other hand, that being said, then again, though, whereas, while, whilst, yet
Modal_invest_expand_ phrase (strong_future)	*Investment-oriented*	plan to invest, plans to invest, planning to invest, will invest, intend to invest, expects to invest, is expected to invest, anticipate investing, investment will be, investments will be
	Growth-oriented	expect growth, expects growth, expected growth, project growth, projects growth, projected growth, forecast growth, forecasts growth, growth will be, growth is expected to
	Expansion-oriented	plan to expand, plans to expand, planning to expand, will expand, expects expansion, is expected to expand, forecast expansion, project expansion, expansion will be, expansion is expected to
	Development-oriented	plan to develop, plans to develop, planning to develop, will develop, expects development, is expected to develop, anticipate development, development will be, development is expected to, development is projected to.
Future focus terms	*Temporal-markers*	will, would, shall, going to, next, later, future, upcoming, forthcoming, soon, sooner, subsequent, eventually, imminent, imminently, impending, looking-forward, looking forward, forward-looking, forward looking, moving forward, as we look ahead, as we move forward, over the next, coming quarter, coming year, next quarter, next month, next year, long-term, short-term, near-term, over time, sometime, henceforth, hence after, heading into, heading toward, heading towards, head into, head toward, head towards.
	Guidance-oriented	outlook, guidance, guidance range
	Planning/ opportunity-oriented	plan to, plans to, planned, planning to, planning on, strategy, strategic, roadmap, pipeline, intend to, intends to, intent, target, targets, aim, aims, goal, goals, aspire, aspiration, prepare, preparing, approach, approaching, opportunity, opportunities, potential, ramp-up, ramp up, scale, monetize, monetise, launch, launching, rollout, innovation, develop, development, expand, expanding, expansion, growth.
	Expectations / projections-oriented	expect, expects, is expected, expecting, anticipate, anticipated, forecast, forecasts, forecasted, project, projects, is projected, projection, projecting, estimate, estimates, to increase, be increasing, to result in, foresee, foreseeable, predict, predicts, predicting.
	Conditional / stance cues	likely, unlikely, we believe, we see, we will, we are going to, possibility, possible, probable, probability, hopeful, hopefully, optimistic
Past result focus terms	*Past-time markers*	last quarter, prior quarter, previous quarter, last year, prior year, previous year, year-ago, year-to-date, ytd, in the quarter, in the prior, during the quarter, as of, ended, as we ended.
	Past-tense performance verbs	delivered, achieved, recorded, generated, realized, realised, posted, grew, increased, declined, decreased, improved, deteriorated, expanded, contracted, beat, exceeded, missed, fell short, rose, up, down, higher, lower, flat, sequentially, year-over-year, yoy, quarter-over-quarter, qoq.