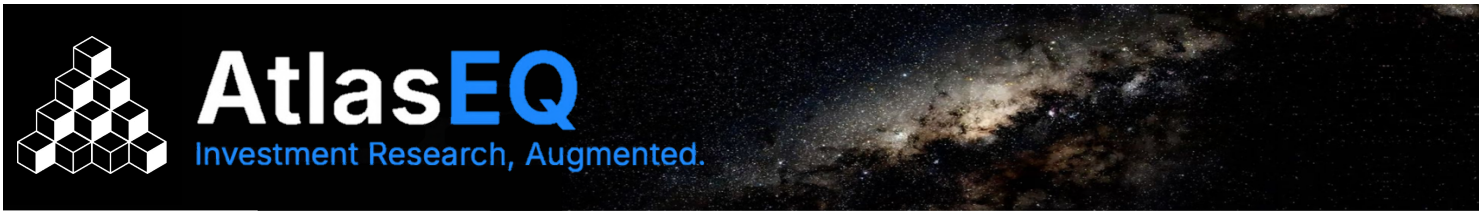

Beyond Consensus

Sequential gradient boosting for full-year EPS, and the returns to analyst disagreement

Mathéo Menges

Abstract

We develop a sequential forecasting model for full-year earnings per share (EPS) that updates as quarterly earnings are reported. It combines an annual forecast, a forecast of the next unreported quarter, and realised earnings using only public information available at each forecast date. We call this sequential updating process a revision cascade. We evaluate the model against the I/B/E/S consensus without using analyst forecasts as model inputs. Across 66,914 firm-year-stage observations, the model's mean absolute error is 0.723 times that of consensus, improving from 0.934 before any quarter is reported to 0.347 after three quarters. The difference between the model and consensus also predicts subsequent earnings surprises and supports a tradable long–short strategy. The results show that the model provides an independent benchmark for analyst consensus and that the disagreement between the two forecasts contains information about future earnings.

IMPORTANT NOTE — RESEARCH AND INFORMATIONAL PURPOSES ONLY
*This research note presents research originally developed by **Matheo Menges** for the **2026 CFA Quant Awards**, an international quantitative-finance research competition organized by participating CFA Societies. The present document is an adapted version prepared for publication by **AtlasEQ** and is not a submission to, publication by, or endorsement of CFA Institute or any participating CFA Society.*
*The research and empirical analysis presented in the original academic work were conducted for research and educational purposes using the data sources and academic research environment described in the paper. Such academic research data were used to develop and test the methodology but were **not incorporated into, connected to, or used as data inputs for the Atlas EQ terminal or its production data infrastructure**. The Atlas EQ terminal therefore does not claim to reproduce the results of this research using those academic data sources.*
*This document is provided **solely for research and informational purposes**. It does not constitute investment advice, a recommendation, solicitation, offer to buy or sell any security or financial instrument, or a representation that any investment strategy or security is suitable for any particular investor. Past research results and historical backtests are not indicative of future performance. Any investment decision should be made independently and with appropriate professional advice where necessary.*
*While reasonable care has been taken in preparing this research, **Atlas EQ and the author make no representation or warranty as to the completeness, accuracy, or continued validity of the information presented**, and neither accepts liability for any loss arising from reliance on this document or its contents.*
Where results, datasets, or analyses originate from academic research conducted using third-party or institutionally licensed data, those materials remain subject to the applicable terms and conditions governing their use. The publication of this adapted research note does not imply that any underlying third-party dataset has been redistributed or made available by Atlas EQ.

1 Introduction

Earnings expectations are central to equity valuation and to the interpretation of reported results. Analyst forecasts have long been used as measures of market expectations, and the literature shows that forecast accuracy varies systematically across analysts and settings (Brown & Rozeff, 1978; Clement, 1999). At the same time, analyst forecasts can contain optimism linked to career incentives and other behavioural forces (Hong & Kubik, 2003; Bordalo et al., 2024). Analyst disagreement is also informative: dispersion can reflect differences in beliefs or uncertainty and is associated with subsequent price reactions and returns (Abarbanell et al., 1995; Barron & Stuerke, 1998; Diether et al., 2002). These findings suggest that consensus is informative but may not fully reflect market expectations. Machine learning provides an alternative source of earnings expectations, but its value relative to analysts is not established. Recent evidence shows that machine-learning forecasts do not systematically outperform analysts, while the strongest gains arise when models exploit predictable analyst biases (Campbell et al., 2026). Other recent evidence finds that nonlinear methods, particularly XGBoost, perform strongly in U.S. earnings forecasting, with gains concentrated in more difficult settings (Chattopadhyay et al., 2025). The question is whether a model using public information can provide information beyond analyst consensus.

We address this question by developing a sequential full-year EPS forecasting model using only public information available at each forecast date. Rather than producing a static annual forecast, the model combines annual and quarterly forecasts and updates the full-year estimate as new earnings information becomes available. Analyst forecasts are excluded from the model inputs, allowing the model to provide an independent benchmark for comparison with consensus. Across 66,914 firm-year-stage observations, the model's mean absolute error is 27.7% lower than consensus, rising to 65.3% lower after three quarters are reported. The advantage also persists in crisis years, with relative error of 0.584 versus 0.746 in other years, and is strongest in several sectors, including Materials (0.500). We make three contributions. First, we develop a sequential forecasting model that combines long-horizon and within-year information and updates the forecast as the fiscal year progresses. Second, we evaluate the model against analyst consensus and a naïve benchmark across forecasting stages and firm characteristics. Third, we examine whether disagreement between the model and consensus predicts subsequent earnings surprises and whether it supports an investable long–short strategy.

2 The revision cascade

2.1 Data

We combine quarterly and annual Compustat fundamentals, I/B/E/S analyst forecasts, macroeconomic indicators from FRED, and daily market data. The Compustat universe contains approximately 5,000 large U.S.-listed firms selected by average market capitalization and includes both active and inactive firms. The forecasting models use accounting, firm, industry, and macroeconomic information. Analyst expectations are taken from the unadjusted I/B/E/S GAAP-basis summary file (MEASURE = 'GPS') to match the GAAP EPS target. The matched dataset contains 166,471 firm-year-stage observations, corresponding to 45,071 distinct firm-years across 4,124 firms for the FY2008–2025 cascade scoring period. After the sample screens and forecast reliability criteria, the evaluation sample contains 66,914 firm-year-stage observations across 2,269 firms over FY2010–2025. Data details are provided in Appendix B.

2.2 Annual and quarterly models

The forecasting problem has two components. The annual model predicts the change in full-year EPS, while the quarterly model predicts the change in the next unreported quarter. For the annual model, prior-year full-year EPS serves as the reference; for the quarterly model, the reference is EPS from the same quarter one year earlier. Both models use expanding walk-forward windows and create out-of-sample forecasts before entering the revision cascade.

Annual model

$$\widehat{EPS_{FY}} = EPS_{FY-1} + \widehat{\Delta_{FY}}$$

Quarterly model

$$\widehat{EPS_{q}} = EPS_{q-4} + \widehat{\Delta_{q}}$$

Each forecasting model is an ensemble of four gradient-boosted tree regressors (XGBoost) with a pseudo-Huber loss. The model inputs, summarized in Appendix C, are constructed from information available at the forecast date. XGBoost was selected after controlled comparisons with alternative boosting methods and is well suited to the nonlinear and sparse structure of the data, while also allowing observation-level weighting within the specialist architecture.

$$L_{\delta}(a) = \delta^2\left[\sqrt{1 + \left(\frac{a}{\delta}\right)^2} - 1\right] \qquad \text{where } a = y - \hat{y}.$$

Here, y is the realized EPS change and $\hat{y}$ the corresponding forecast. The pseudo-Huber loss limits the influence of unusually large EPS errors while retaining all observations in the training sample.

The four specialists share the same inputs, target and model specification, but place greater emphasis on observations within different EPS-magnitude regions. This allows each specialist to focus on a different earnings scale while retaining exposure to the full sample. Their forecasts are then combined using smoothly varying EPS-dependent weights, giving greater influence to the specialist most suited to the firm's earnings scale.

$$\hat{\Delta}_i = \sum_{k=1}^{4} g_{k(a_i)} \hat{\Delta}_i^k, \qquad \sum_{k=1}^{4} g_{k(a_i)} = 1, \qquad \hat{y}_i = a_i + \hat{\Delta}_i$$

Here, a_i denotes the relevant lagged EPS reference, and $g_k(a_i)$ the weight assigned to specialist k. The final forecast is therefore a weighted combination of four predictions of the same EPS change, rather than four different targets.

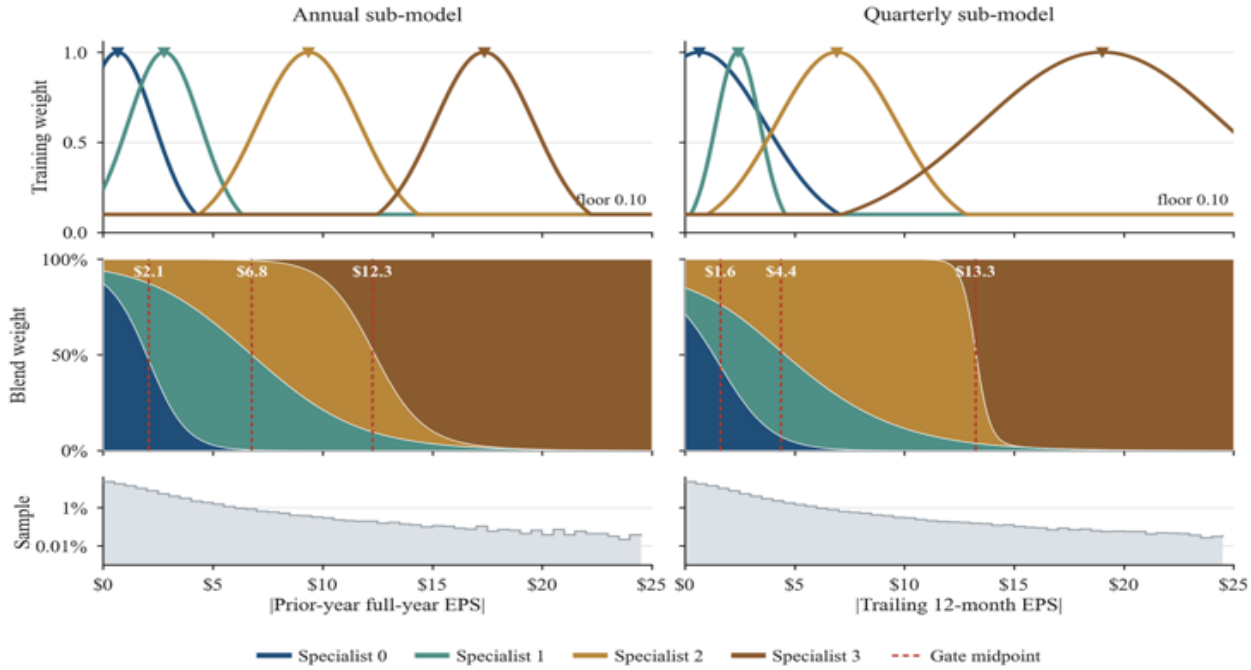

Figure 1. Specialist weighting structure for the annual and quarterly models. *The top panel shows the training weight assigned to each specialist across the EPS reference distribution; the middle panel shows the resulting blend weight used to combine their forecasts; and the bottom panel shows the distribution of observations across EPS scale. Training weights determine which observations each specialist emphasizes during estimation, while blend weights determine its contribution to the final forecast. Dashed vertical lines mark the gate midpoints between adjacent specialists.*

2.3 Revision mechanism

Forecast combination and sequential updating are established approaches (Bates & Granger, 1969). Here, we use them to update the full-year EPS forecast as new information arrives during the fiscal year. At each reporting stage, the quarterly model first provides a forecast for the next unreported quarter, which is combined with the corresponding prior-year EPS to form the stage-specific annual anchor. The revision model then estimates the remaining error in this anchor. **First**, let the quarterly model's forecast for the next unreported quarter be denoted by forecast EPS, and let the corresponding prior-year EPS be the comparison value. We combine the two using their estimated error variances:

$$N_q = w_q \widehat{EPS}_q^Q + (1 - w_q) P_q \qquad w_q = \frac{\sigma^2_P}{\sigma^2_Q + \sigma^2_P}$$

Here, σ^2_P is the historical error variance of the prior-year EPS and σ^2_Q the model-estimated error variance of the quarterly forecast. Thus, the source with the lower estimated error receives the greater weight. Only the next unreported quarter is blended; subsequent quarters retain their prior-year EPS values. Second, the blended quarter is combined with realised quarters and prior-year EPS for the remaining quarters to form the annual anchor. For reporting stage K, the construction is:

$$A_1 = EPS_1 + N_2 + P_3 + P_4$$
$$A_2 = EPS_1 + EPS_2 + N_3 + P_4$$
$$A_3 = EPS_1 + EPS_2 + EPS_3 + N_4$$
$$A_4 = EPS_1 + EPS_2 + EPS_3 + EPS_4$$

Here, EPS_q denotes realised EPS for a reported quarter, P_q the corresponding prior-year EPS, and N_q the blended estimate for the next unreported quarter. The anchor therefore represents the full-year EPS implied by the information available at each stage, before applying the revision model.

Third, the revision model predicts the remaining error in the anchor:

$$\delta_K = EPS_Y^{FY} - A_K$$

The final forecast is:

$$\widehat{EPS}_Y^{FY} = A_K + \widehat{\delta_K}$$

The cascade therefore separates two tasks: first constructing the best available full-year anchor, and then estimating how that anchor should be revised as the fiscal year progresses.

2.4 Feature design and information flow

The feature sets serve different roles within the forecasting system. The annual model uses multi-year fundamentals and context; the quarterly model uses seasonality, earnings history, recent growth, and year-to-date state. The revision cascade combines these model outputs with realised forecast errors, firm track record, forecast uncertainty, the annual model's forecast distribution, and the reliability of the prior-year benchmark.

Figure 2 shows how the contribution of these information groups changes across stages. Prior-model signals account for 44.4% of attribution at stage 0 and 77.7% at stage 3. As more quarters are reported, current-year information becomes available and the cascade relies increasingly on evidence from the current fiscal year.

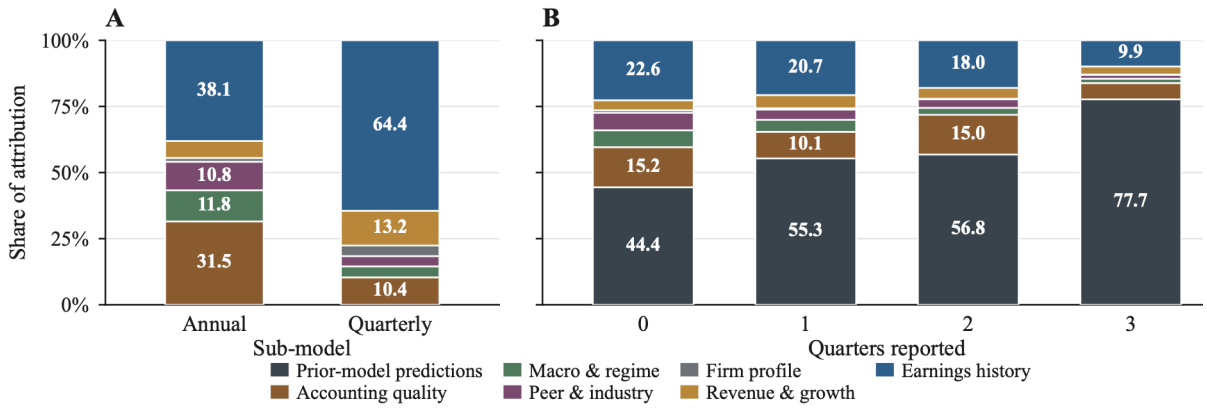

Figure 2. What each level of the cascade learns from. *Panel A shows the share of the revision model's feature attribution assigned to each signal family for the annual and quarterly models. Panel B shows the corresponding attribution across the four reporting stages, including the contribution of prior-model predictions. Percentages are calculated from the fitted model's feature attributions and normalized to sum to 100% within each bar. Prior-model predictions rise from 44% at stage 0 to 78% at stage 3, indicating that the cascade increasingly relies on information produced by earlier forecasting stages.*

2.5 Training and walk-forward validation

All models are trained using expanding walk-forward windows. For target year H, only observations from earlier years enter training; the revision model is estimated separately for stages 0–3 and refit annually. The annual and quarterly models produce out-of-sample forecasts from FY2005; the revision cascade is trained from FY2007 and first scored from FY2008, with annual refits through FY2025 (Appendix B). A separate holdout comparison of XGBoost, CatBoost, and LightGBM found that XGBoost had the lowest test MAE for both annual and quarterly forecasts; the differences were small (0.5% and 1.1%). The point-in-time sensitivity test left the main conclusion unchanged: relative error moved from 0.716 to 0.717 on a paired 65,783-observation sample. Appendix C summarises the expanding walk-forward training, test, tuning, and calibration windows used across stages.

2.6 Point-in-time construction and leakage control

All information is treated as available only if it was known at the forecast date. Each feature has a point-in-time date and enters a stage only if it was available before that date. Compustat observations use the reported availability date where available, with a documented fallback, while I/B/E/S forecasts are matched only when the snapshot date precedes the decision date. Training also excludes each target year from its own training set. The same rule applies to derived and model-generated information: rolling track records, past forecast errors, and stage-level aggregates use only information available before the forecast date. A leakage audit identified two issues: a rolling track-record feature included the evaluated observation, and an aggregate feature used forecasts unavailable at early stages. Both were corrected before the reported results; the first issue produced a correlation of 0.517 before correction and −0.052 after. Details are reported in Appendix D.

2.7 Reliability tiering and uncertainty

Forecast generation and forecast reliability are treated as separate steps. At each stage, a LightGBM classifier estimates whether the forecast is likely to fall within a stage-specific error threshold, using the forecast itself, conformal interval widths, and available panel information. The thresholds tighten from \$0.75 at stage 0 to \$0.25 at stage 3, reflecting the greater precision expected later in the fiscal year. The classifier assigns each forecast to one of three reliability tiers:

$$\textit{HIGH}: p \geq 0.65 \quad \textit{MED}: 0.35 \leq p < 0.65 \quad \textit{LOW}: p < 0.35$$

Figure 3 shows the tiers separate materially different error distributions: HIGH forecasts account for 51.9% of the screened sample with median absolute error of \$0.151, versus \$0.550 for MED and \$1.304 for LOW, while mean absolute error rises from \$0.263 to \$0.831 and \$2.433. Headline results therefore use HIGH and MED forecasts, which together comprise 66,914 observations. The remaining 12.2% carry no classifier score, mostly during the FY2008–2012 warm-up, and are excluded alongside LOW.

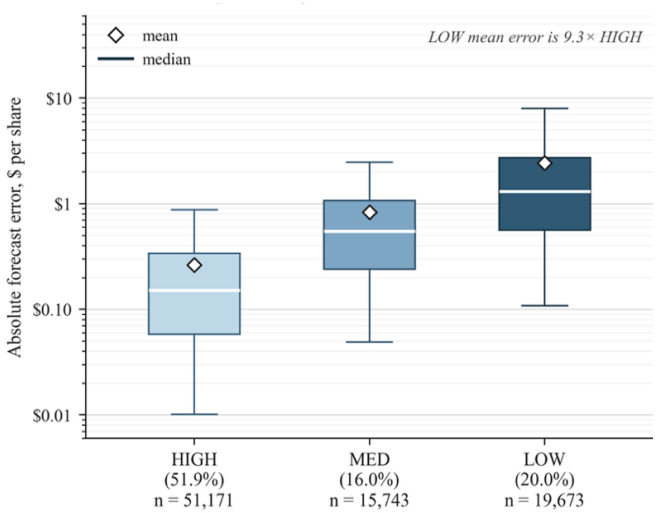

Figure 3. Forecast error by reliability tier. *Boxplots show the distribution of absolute full-year forecast errors for HIGH, MED, and LOW forecasts pooled across stages; diamonds denote the mean and horizontal lines the median. Tier shares and observation counts are shown beneath each category.*

3 Results

3.1 Forecast accuracy and bias

We measure forecast accuracy using scale-free absolute error. Each forecast error is divided by prior-year absolute EPS, with the denominator floored at \$0.50. Relative error is the ratio of our mean absolute error to consensus mean absolute error, so values below one indicate greater accuracy. Inference uses two-way firm × year cluster-robust standard errors (Cameron, Gelbach & Miller, 2011) and the Diebold–Mariano test (Diebold & Mariano, 1995) on absolute-error differentials; definitions are provided in Appendix F.

Across 66,914 firm-year-stage observations from FY2010–2025, the model has a relative error of 0.723 and outperforms consensus at every stage. Its advantage grows as information accumulates, with relative error falling from 0.934 at stage 0 to 0.347 at stage 3. Consensus is systematically optimistic, with mean signed error of +0.131, while the model is slightly conservative at −0.046, consistent with documented evidence of optimism in analyst forecasts (Hong & Kubik, 2003; Bordalo et al., 2024).

Figure 4 shows how accuracy and bias change through the year. Before any quarter is reported, the model is only modestly more accurate than consensus and wins on 48.2% of observations. As realised earnings accumulate, its advantage grows: relative error falls to 0.656 at stage 2 and 0.347 at stage 3, while the win rate rises to 69.1%. Consensus optimism declines through the year but remains positive at every stage, from +0.187 at stage 0 to +0.083 at stage 3.

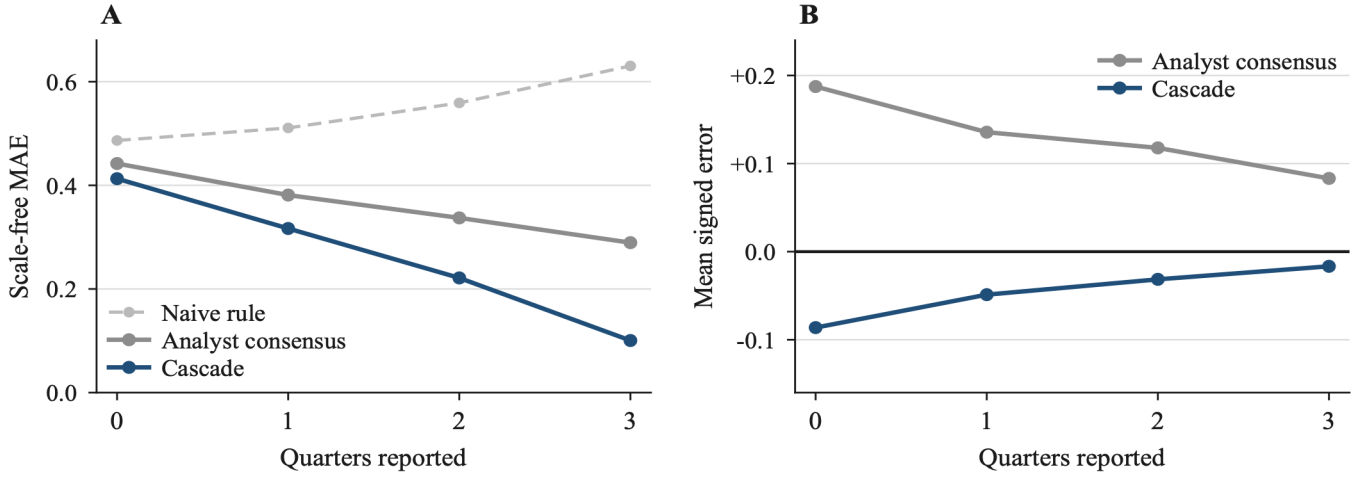

Figure 4. Forecast error and bias by stage of the fiscal year, FY2010–2025. *Panel A: scale-free mean absolute error for the naive rule, the consensus, and the cascade. Panel B: mean signed error, positive above the zero line.*

t	n	Firms	Naive	Cons.	Model	Rel. err	Win	DM (HLN)	M.bias	C.bias
0	16,485	2,105	0.487	0.442	0.413	0.934	48.2%	2.36	-0.086	+0.187
1	16,960	2,154	0.511	0.381	0.317	0.831	51.6%	3.25	-0.049	+0.136
2	16,315	2,098	0.559	0.337	0.221	0.656	56.4%	4.77	-0.031	+0.118
3	17,154	2,140	0.630	0.289	0.100	0.347	69.1%	6.32	-0.017	+0.083
All	**66,914**	**2,269**	**0.547**	**0.362**	**0.262**	**0.723**	**56.4%**	**5.49**	**-0.046**	**+0.131**

Table 1. Forecast accuracy and bias by stage. *Stage is the number of target-year quarters reported at the forecast date; n is the number of forecasts and Firms the number of distinct firms. Naive, Cons., and Model denote the no-change, consensus, and cascade forecasts, respectively. Rel. err is model MAE relative to consensus MAE; Win is the share of forecasts with lower absolute error than consensus; DM (HLN) is the Diebold–Mariano statistic with the Harvey–Leybourne–Newbold correction; M.bias and C.bias are mean signed model and consensus errors. Errors are scaled by prior-year | EPS |, floored at $0.50; DM is clustered by firm and year.*

3.2 Performance through time and in crises

Figure 5 shows that the forecasting advantage persists throughout the out-of-sample period and is also present during crisis years. Relative error declines from 1.195 in FY2010 to 0.665 in FY2025, with a trend of 0.027 per year; the first half of the sample averages 0.831 versus 0.668 in FY2018–2025. The improvement is not limited to normal conditions: relative error falls to 0.598 in FY2020 and 0.565 in FY2022, compared with 0.746 in other years. Across the two crisis years, relative error is 0.584 with a DM t-statistic of 13.0, versus 0.746 and 7.2 in other years.

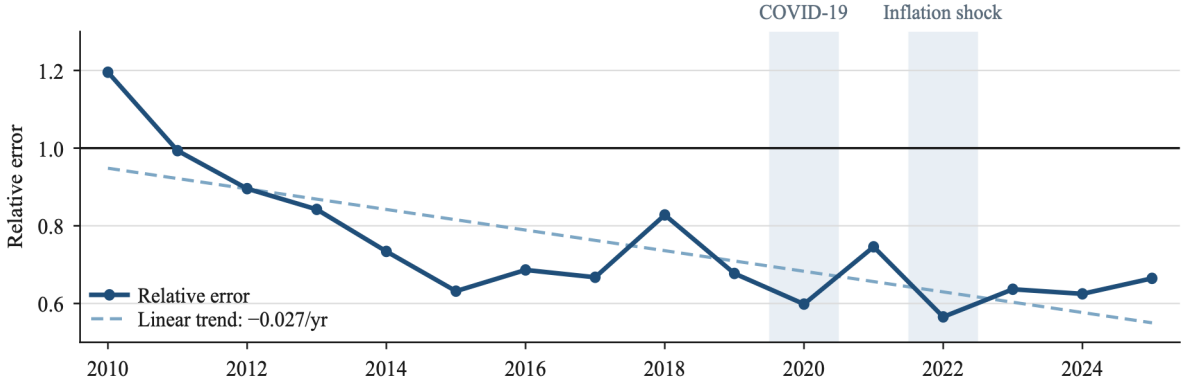

Figure 5. Relative error by fiscal year, *FY2010–2025, with fitted trend. Values below 1.00 mean the cascade is closer than the consensus. Shaded bands mark the FY2020 and FY2022 crisis years.*

3.3 Where the advantage is strongest

The advantage is strongest where analyst information is limited or uncertain (Table 2). Relative error is 0.658 with one to three analysts versus 0.931 with 17 or more, and 0.638 among the most dispersed forecasts versus 0.904 among closely agreeing analysts. It is also strongest for smaller EPS firms, at 0.584 below $0.72 versus 0.793 above $3.86. The pattern holds across operating sectors, while financial firms are a statistical tie (1.045, t = 1.2); full sector results are reported in Appendix G.

Overall, the model adds the most value when analyst coverage is limited, disagreement is high, or earnings are small.

	Bucket	n	Cons. MAE	Model MAE	Rel. err	DM t
Analyst coverage	1-3 analysts	19,077	0.458	0.302	0.658	8.77
	4-8	27,693	0.360	0.263	0.731	7.27
	9-16	15,196	0.288	0.226	0.782	4.81
	17+	4,948	0.228	0.212	0.931	1.00
Analyst agreement	agree closely	15,231	0.143	0.130	0.904	2.08
	mostly agree	15,232	0.208	0.180	0.867	3.42
	some spread	15,224	0.353	0.280	0.793	4.97
	disagree	15,229	0.686	0.438	0.638	8.70
Earnings scale	under $0.72	13,436	0.557	0.325	0.584	8.58
	$0.72–1.47	13,441	0.411	0.309	0.753	5.34
	$1.47–2.37	13,313	0.317	0.251	0.792	3.70
	$2.37–3.86	13,360	0.269	0.220	0.820	3.90
	over $3.86	13,364	0.254	0.202	0.793	3.65
Regime	crisis	8,049	0.432	0.252	0.584	12.97
	calm years	58,865	0.352	0.263	0.746	7.24

Table 2. Rel. err by analyst coverage, agreement, earnings scale, and market regime. *Agreement is quartiles of consensus dispersion and requires at least two analysts (5,998 single-analyst observations excluded), scale is quintiles of realised |EPS|, crisis is 2020 and 2022. Relative error is model MAE ÷ consensus MAE; DM t clustered by firm and year*

3.4 Surprise detection

The forecast also contains information about subsequent earnings surprises. Because consensus is the benchmark, it cannot provide an independent signed surprise forecast; we therefore compare the model with analyst dispersion and forecast revisions. Detection is similar at stage 0 (AUC 0.680 vs. 0.690) but improves thereafter, reaching 0.869 at stage

3 versus 0.724 for dispersion. Figure 6 shows the corresponding improvement in detection AUC and direction correctness as quarters are reported. The model also predicts surprise direction: it correctly classifies the sign in 78.5% of surprising observations overall, rising to 89.5% at stage 3 versus 52.7% for the naïve rule. Accuracy is higher for negative surprises (86.0%) than positive ones (66.1%), consistent with the optimism documented in consensus.

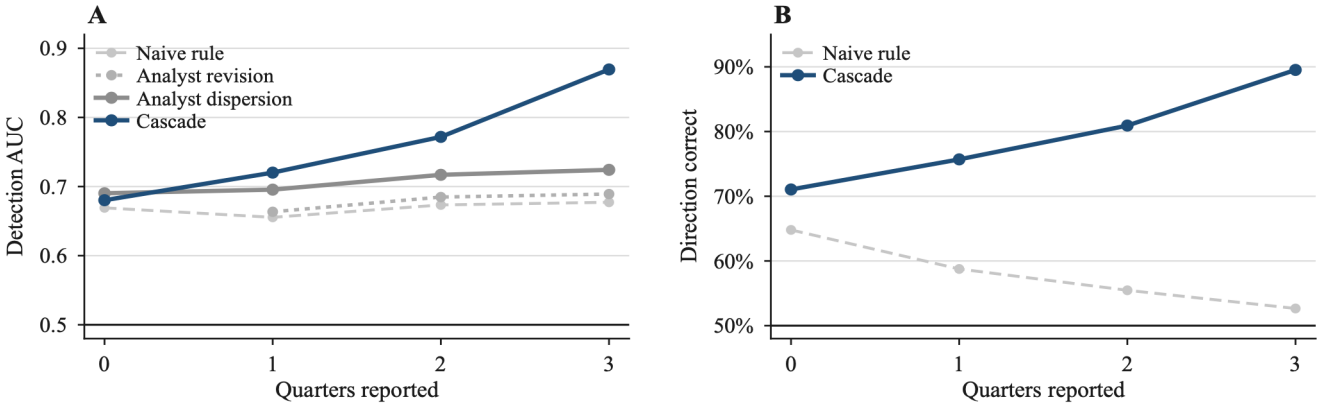

Figure 6: Surprise detection and direction by stage. *A surprise is defined as reported EPS differing from consensus by more than 10% of prior-year EPS. Panel A reports detection AUC for the cascade, analyst dispersion, analyst revisions, and the naïve rule; 0.5 represents no predictive ability. Panel B reports the percentage of surprising observations for which the cascade correctly predicts the sign of the earnings surprise, compared with the naïve rule.*

3.5 Does the model know when it is wrong

The forecast intervals also provide information about uncertainty. Sorting observations by predicted interval width, the narrowest quintile has a relative error of 0.450 and a 31.9% large-surprise rate, compared with 0.788 and 85.4% in the widest quintile. The relationship is monotonic: wider intervals are associated with both larger forecast errors and greater surprise risk. After walk-forward conformal calibration (Romano et al., 2019), the nominal 80% intervals achieve 82.2% coverage; stage-level calibration results are reported in Appendix E.

4 The investment strategy

4.1 Signal construction

Forecast accuracy alone does not imply investment value. We therefore use the point-in-time disagreement between the model forecast and analyst consensus, scaled by prior-year EPS, as the investment signal. The intuition is that a large disagreement identifies firms where the model and the market's prevailing expectation differ most, creating the greatest potential for information to be incorporated into prices. The signal is formed at each rebalance using only information available at that date. Because the model is slightly conservative on average, we rank firms by disagreement rather than trade its raw sign. The signal can then be implemented either through quarterly rebalancing or around annual earnings announcements.

4.2 Portfolio construction

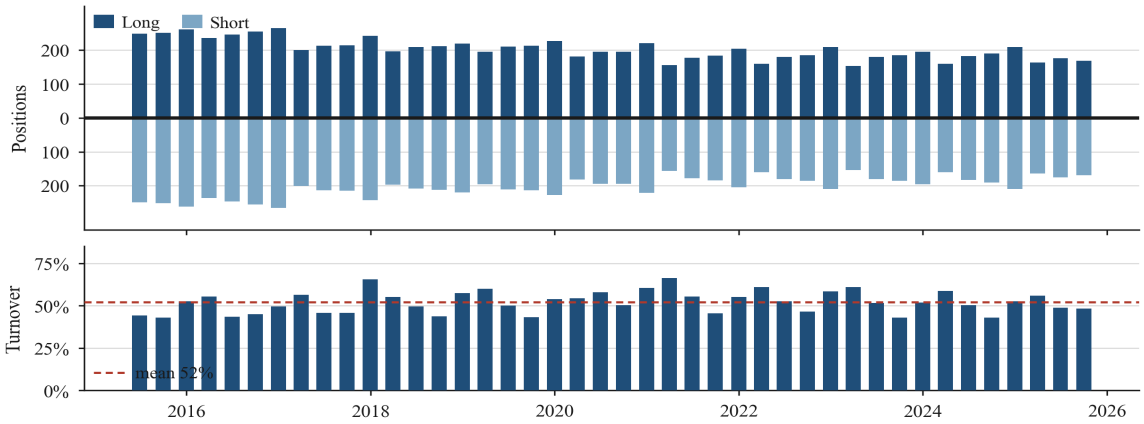

Figure 7. Long and short position counts and portfolio turnover by quarter. *Longs and shorts are the top and bottom 20% of the model-consensus disagreement signal; turnover is the share of the book replaced at each quarterly rebalance.*

We implement the signal in two settings. The quarterly book ranks firms at fixed quarter-ends, goes long the top 20% and short the bottom 20%, and holds the positions for three months; Figure 7 shows the resulting position counts and turnover over time. This is the primary long-horizon test. The event strategy applies the same long–short ranking around each firm's annual earnings announcement to test shorter-term repricing. The quarterly book runs from March 2010 to September 2025 across 63 quarterly rebalances, 2,313 firms, and 26,198 positions and is survivorship-free. The event strategy runs from February 2014 to December 2025 with 3,308 positions across 904 firms, but uses a daily price panel with incomplete historical coverage and is therefore subject to survivorship bias. We treat it as supporting evidence rather than the headline strategy.

4.3 Strategy performance

The quarterly strategy earns a lower absolute return than the S&P 500 but delivers a higher Sharpe ratio and lower maximum drawdown. Capped volatility scaling increases returns but leaves the Sharpe ratio broadly unchanged while materially increasing drawdown. The event strategy shows a similar pattern but is treated as supporting evidence because of its weaker historical price coverage. The quarterly strategy therefore provides the clearest evidence that forecast disagreement contains investable information. Full strategy and test details are reported in Appendix H.

	Quarterly book	Quarterly, capped vol-scaled	Event trade	Event, capped vol-scaled	Equal-weight univ.	S&P 500 (div.)
Annual return	9.51%	22.10%	9.34%	18.02%	9.58%	13.72%
Volatility	5.61%	15.04%	6.42%	15.12%	18.89%	15.60%
Excess Sharpe	**1.40**	**1.41**	**1.16**	**1.12**	**0.50**	**0.83**
Max drawdown	-8.20%	-25.46%	-0.61%	-3.72%	-30.81%	-23.93%
t-statistic	5.61	4.66	4.65	3.84	1.90	2.97
Growth of 1$	$2.60	$8.14	$2.55	$5.70	$2.61	$3.86

Table 3. Strategy performance against benchmarks. *All series run 2015-06-30 to 2025-09-30, the common window in which every strategy and benchmark is live, and are rebased to $1 at the start; Growth of $1 is the terminal value. Strategy returns are net of 15bps per side and 50bps a year borrow on the short leg; the two benchmarks are gross. Vol-scaled columns apply $L = \min(3, 0.15/\hat{\sigma})$ from the four preceding quarters, lagged one, with borrowing charged at the 13-week bill plus 100bps. Sharpe is on excess return over the bill, with idle cash held in bills. The event trade uses a survivorship-limited panel.*

4.4 Sources of return

The portfolio return comes from both sides of the disagreement signal. Relative to the equal-weight universe, the long leg earns 1.38% per quarter ($t = 5.78$, $p < 0.001$), while the short leg contributes a further 0.67% per quarter ($t = 2.32$, $p = 0.025$). Together, the two legs generate 2.06% per quarter of excess return, with 67.2% attributable to the long leg and 32.8% to the short leg. Both sides of the signal therefore contribute statistically significant excess returns.

5 Conclusion and limitations

We develop a public-data forecasting system that updates a full-year EPS forecast as information accumulates through the fiscal year, without using analyst expectations. The resulting forecast is more accurate than analyst consensus at every reporting stage and avoids the consensus's persistent optimism. The difference between the two forecasts is also economically meaningful: forecast disagreement predicts the direction of subsequent earnings surprises and supports a tradable long–short strategy. Taken together, the results show that a sequential forecasting process can be built from observable accounting, market, and macroeconomic information while providing an independent benchmark for consensus expectations.

The analysis has several limitations. The forecast comparison is restricted to firms with sufficient analyst coverage and therefore does not describe the full listed universe, and the model does not outperform consensus for financial firms. The primary quarterly portfolio is not subject to survivorship bias, whereas the event strategy relies on a more limited historical daily price panel and is treated as supporting evidence. The analysis is also limited to U.S.-listed firms, and the short announcement window limits capacity and liquidity. Finally, strategy returns are not uniform through time; the quarterly book went through a multi-year trough around 2020–2022 before recovering, and the persistence of the investment edge remains a question for future research.

References

Abarbanell, J. S., Lanen, W. N., & Verrecchia, R. E. (1995). Analysts' forecasts as proxies for investor beliefs in empirical research. Journal of Accounting and Economics, 20(1), 31–60. https://doi.org/10.1016/0165-4101(94)00392-I

Barron, O. E., & Stuerke, P. S. (1998). Dispersion in analysts' earnings forecasts as a measure of uncertainty. Journal of Accounting, Auditing & Finance, 13(3), 245–270. https://doi.org/10.1177/0148558X9801300305

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. Journal of the Operational Research Society, 20(4), 451–468. https://doi.org/10.1057/jors.1969.103

Bordalo, P., Gennaioli, N., La Porta, R., O'Brien, M., & Shleifer, A. (2024). Long-term expectations and aggregate fluctuations. NBER Macroeconomics Annual, 38(1), 311–347. https://doi.org/10.1086/729198

Brown, L. D., & Rozeff, M. S. (1978). The superiority of analyst forecasts as measures of expectations: Evidence from earnings. The Journal of Finance, 33(1), 1–16. https://doi.org/10.1111/j.1540-6261.1978.tb03385.x

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. Journal of Business & Economic Statistics, 29(2), 238–249. https://doi.org/10.1198/jbes.2010.07136

Campbell, J. L., Ham, H., Lu, Z., & Wood, K. (2026). Expectations matter: When (not) to use machine learning earnings forecasts. Management Science. https://doi.org/10.1287/mnsc.2024.05808

Chattopadhyay, A., Fang, B., & Mohanram, P. S. (2025). Machine learning for earnings forecasting: US and international evidence. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5941658

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

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. The Journal of Finance, 57(5), 2113–2141. https://doi.org/10.1111/0022-1082.00490

Hong, H., & Kubik, J. D. (2003). Analyzing the analyst: Career concerns and biased earnings forecasts. The Journal of Finance, 58(1), 313–351. https://doi.org/10.1111/1540-6261.00526

Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. In Advances in Neural Information Processing Systems 32 (pp. 3543–3553).

Data sources

Federal Reserve Bank of St. Louis. (2026). Federal Reserve Economic Data (FRED).

LSEG. (2026). I/B/E/S historical estimates and summary data. Data accessed via Wharton Research Data Services (WRDS).

S&P Global Market Intelligence. (2026). Compustat North America. Data accessed via Wharton Research Data Services (WRDS).

Yahoo Finance. (2026). Historical market data.

Appendix A Cascade architecture and information flow

This appendix sets out the full cascade. Figure A1 shows the annual and quarterly feature panels, the Vega and Polaris specialist layers, and the Sirius stage-aware revision model, together with how their outputs enter the cascade at each stage. Table A1 lists each component, its role, its estimator, and how it is validated.

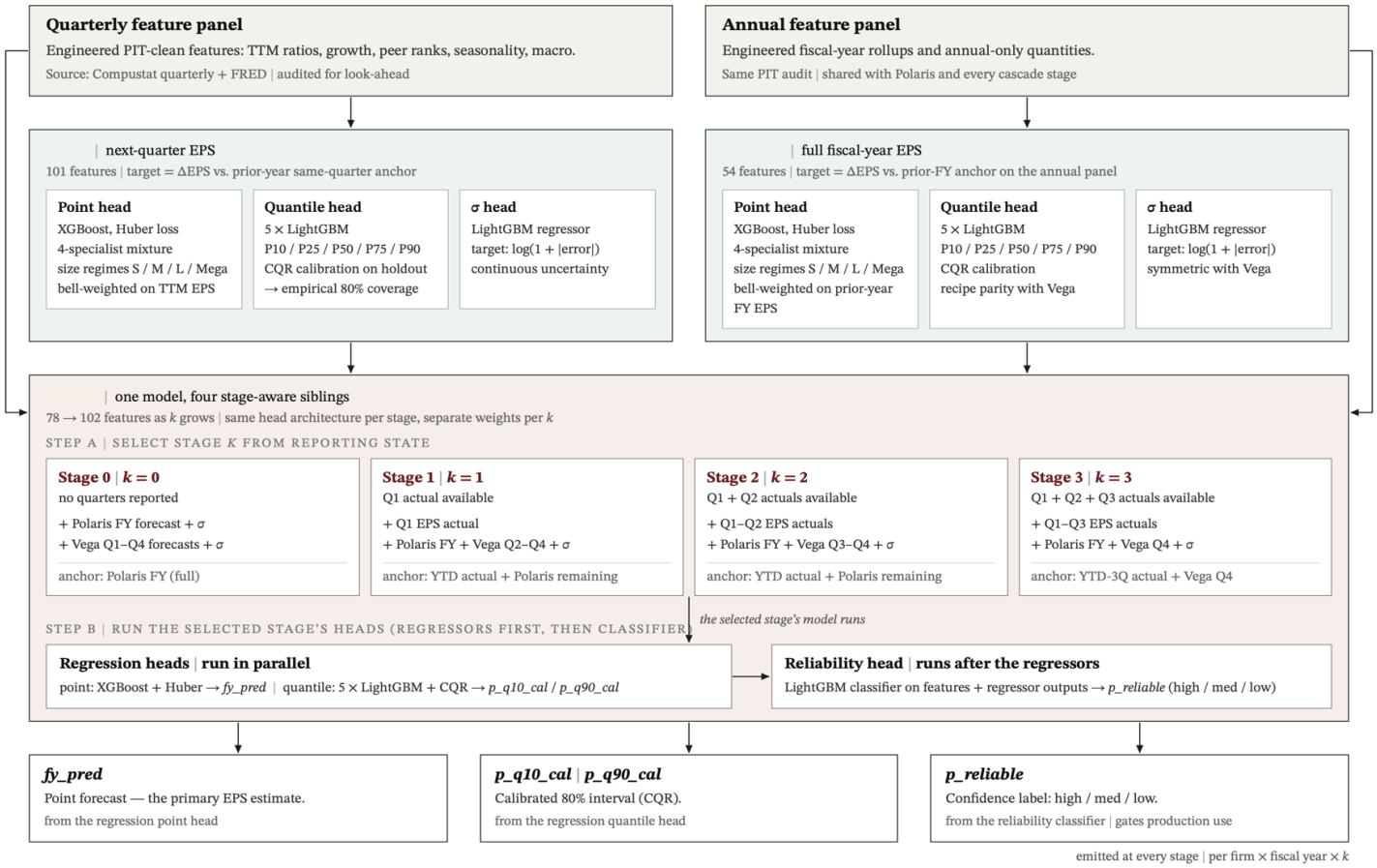

Figure A1: Cascade architecture: the feature panels, the Vega and Polaris specialists, and the Sirius stage-aware cascade.

Component	Target / role	Estimator	Key design	Validation
Annual specialist	Full-year EPS change vs. prior-year EPS	XGBoost	54 features; 4 specialist models; pseudo-Huber loss; quantile heads	Expanding walk-forward; annual refit
Quarterly specialist	Next-quarter EPS YoY change	XGBoost	101 features; 4 specialist models; pseudo-Huber loss; soft error weighting	Expanding walk-forward; annual refit
Revision cascade	Revision to stage-specific annual EPS anchor	XGBoost	4 separate stage models; precision-weighted anchor; stage-specific tuning	Expanding walk-forward; annual refit
Forecast uncertainty	Error scale / prediction intervals	LightGBM	Point error-scale model + 10/25/50/75/90% quantiles	Annual walk-forward refit
Conformal calibration	Prediction-interval coverage	CQR	80% and 50% intervals	Walk-forward calibration
Reliability classifier	Probability forecast meets stage error threshold	LightGBM	Forecast, interval widths and stage information	Expanding walk-forward; annual refit
Walk-forward rule	Point-in-time validation	—	Train only on years strictly earlier than evaluation year	No target-year observations in training

Table A1: Cascade components, roles, estimators, and validation.

Appendix B Data and sample construction

This appendix documents the data and the construction of the estimation sample.

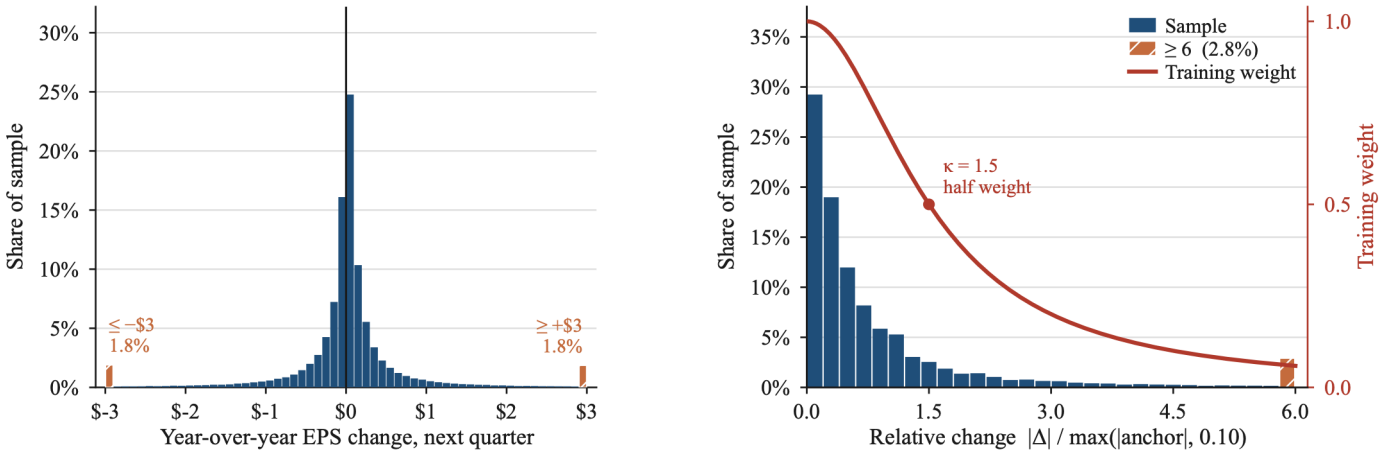

Figure B1. Sample construction and coverage across the historical panel.

The forecasting inputs combine quarterly and annual Compustat fundamentals, I/B/E/S analyst forecasts, FRED macroeconomic indicators, and daily market data. The feature panel begins in FY1996, providing the long historical training panel for the annual and quarterly models. Both models are trained with expanding walk-forward windows and produce out-of-sample forecasts from FY2005; the revision cascade is trained only after these forecasts and their rolling track records have accumulated. The cascade first trains in FY2007 and produces scored forecasts from FY2008; the final evaluation sample begins in FY2010 after the reliability classifier's walk-forward warm-up.

Appendix C Training protocol and walk-forward validation

This appendix documents how the specialists are combined and how the models are trained and validated out of sample. Figure C1 shows the expanding walk-forward scheme, with the training, test, tuning, and calibration windows used at each stage. Table C1 shows a representation of inputs for the annual and quarterly base ensemble models.

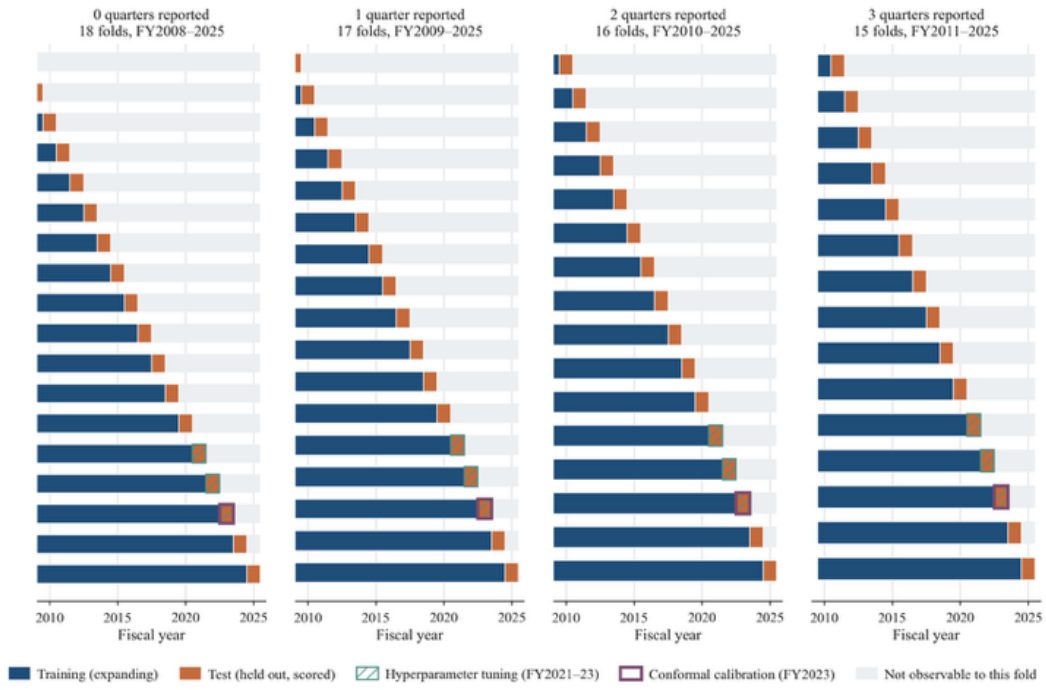

Figure C1. Walk-forward training, test, tuning, and calibration windows by stage.

Feature family	Representative annual inputs	Representative quarterly inputs
EPS / earnings history	eps_fy_lag1/2/3; eps_yoy_growth_lag1; eps_3y_trend_slope	eps_2q_momentum; eps_qoq_growth/delta; eps_growth_q_yoy
Revenue / operating trends	rev_yoy_growth_lag1; rev_3y_trend_slope; rev_5y_growth_vol	revenue_growth_q_yoy; rev_qoq_growth/delta
Profitability / margins	roe; gross/operating/net margin; roa/gpm/opm/npm 3y trends	operating_margin_q; net_margin_q_accel/qoq_chg; roa_q_ttm
Earnings quality	accruals_ratio; accruals_5y_avg; cfo_to_ni_3y_avg	accruals_ratio; cfo_to_ni_q_roll4_std; effective_tax_rate
Leverage / capital allocation	debt_to_equity; net_debt_to_ebitda; dividend_payout; buyback_ratio	leverage_change; net_debt_to_ebitda; total_payout_ratio
Industry / peer position	revenue_rank_in_industry; rev_share_trend_industry; net_margin_rank_industry	roa/roe/net_margin_rank_industry; earnings_gap_industry; mkt_share_yoy_chg
Macro / regime	macro_growth_score; macro_credit_score; gdp_accel; yield_spread; consumer_sentiment	fed_funds_rate; treasury_10y; cpi_yoy; unemployment; macro_growth/credit/demand_score
Firm characteristics	years_since_ipo; log_revenue	log_assets; size_decile; years_since_ipo
Quarterly specific information	—	fqtr; quarters_known_in_fy; seasonal_dev_eps/ni; fy_eps_to_date_yoy; beat_rate_8q

Table C1. Representative model inputs, *the table reports representative inputs rather than the complete feature set. The annual and quarterly specifications use different feature sets, reflecting their respective forecasting horizons. All features are constructed point-in-time using information available before the forecast date.*

Appendix D Point-in-time and leakage audit

This appendix reports the leakage audit behind the point-in-time construction in Section 2.6. It documents the two leakage issues found in early testing and shows model behaviour before and after each correction.

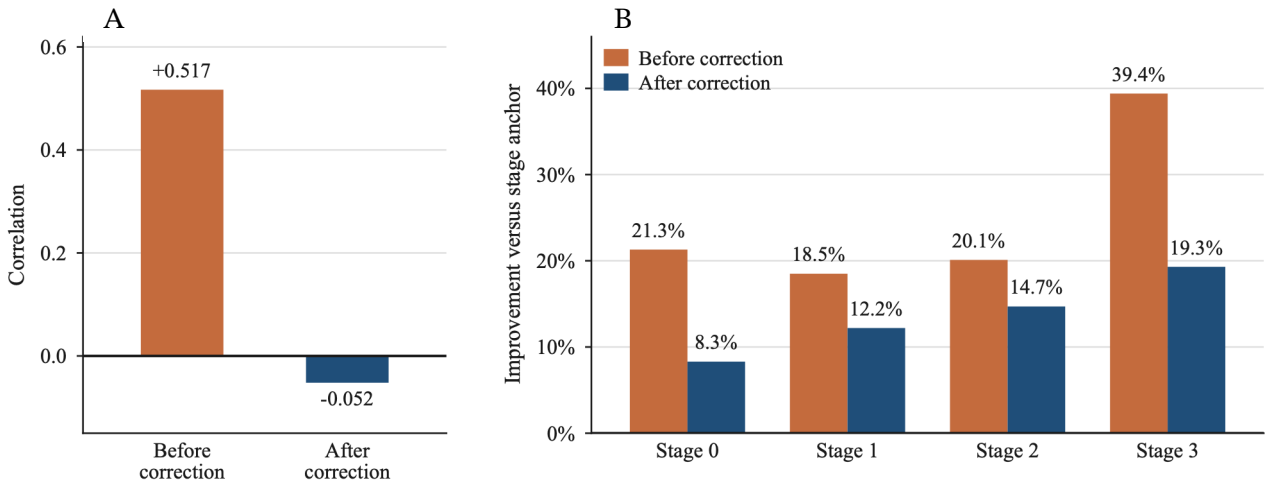

Figure D1. Leakage audit. Panel A shows the correlation associated with the rolling track-record feature before and after correction. Panel B shows financial-model improvement relative to the stage anchor before and after correction. Both leakage issues were identified and corrected before the reported evaluation.

Appendix E Uncertainty quantification and interval calibration

This appendix reports how forecast uncertainty is quantified and how well the prediction intervals are calibrated. Figure E1 shows empirical coverage of the nominal 80% intervals before and after walk-forward conformal calibration, and relates interval width to realised error and to the frequency of large earnings surprises.

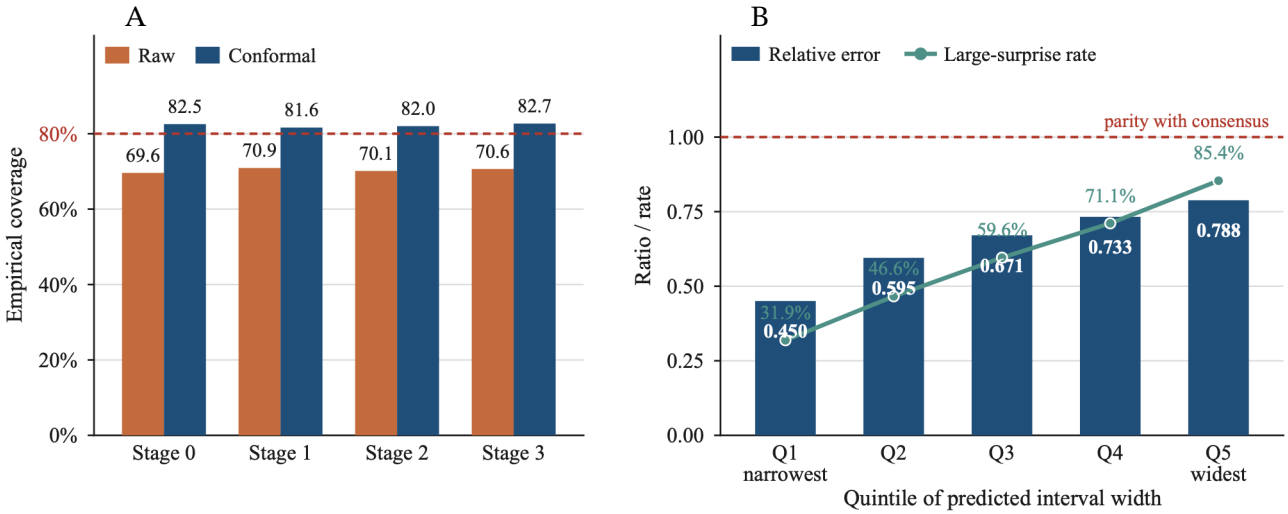

Figure E1. Forecast uncertainty and calibration. Panel A reports empirical coverage of nominal 80% prediction intervals before and after walk-forward conformal calibration. Panel B shows relative forecast error and the frequency of large earnings surprises across quintiles of predicted interval width, from narrowest to widest.

Appendix F Statistical inference and evaluation

This appendix records the inference and screening procedures used in the results. Table F1 reports the forecast-error inference, with Diebold-Mariano statistics and cluster-robust standard errors by stage. Table F2 lists the symmetric screen that holds the model and the consensus to the same set of observations.

Stage	N	Firms	Mean d	HAC SE	DM (HLN)	p-value
Stage 0	16,485	2,105	0.0314	0.0129	2.3596	0.0323
Stage 1	16,960	2,154	0.0576	0.0171	3.2539	0.0053
Stage 2	16,315	2,098	0.0990	0.0201	4.7673	0.0002
Stage 3	17,154	2,140	0.1573	0.0241	6.3170	0.0000
All stages	66,914	2,269	0.0930	0.0164	5.4918	0.0001

Table F1. Forecast-error inference.

Step	% of pre-screen
Consensus forecast missing	37.1%
Model forecast missing	0.0%
Compustat GAAP actual missing	0.0%
I/B/E/S GAAP actual missing	2.8%
Consensus implausible	0.2%
Model implausible	0.1%
Actuals disagree	0.6%

Table F2. Symmetric evaluation screen.

Appendix G Full results by sector, scale, and robustness

This appendix reports the breakdown. Table G1 gives performance by sector, Table G2 by EPS scale and stage, Table G3a the robustness checks, and Table G3b persistence by prior track record.

Sector	N	Consensus MAE	Model MAE	Relative error	DM t
Banks/Insurance	16,150	0.3191	0.3334	1.0450	-1.2147
Industrials	10,930	0.3083	0.2054	0.6663	5.8869
Info Tech	9,974	0.3848	0.2578	0.6699	7.7064
Cons. Disc.	7,106	0.3391	0.2367	0.6979	3.8940
Health Care	6,417	0.3704	0.2401	0.6482	5.2941
Materials	3,763	0.5704	0.2852	0.5001	6.0838
Cons. Staples	3,524	0.2988	0.1733	0.5800	5.2332
Energy	3,326	0.5378	0.3180	0.5913	5.1863
Utilities	3,242	0.3013	0.2028	0.6730	3.7612
Comm. Services	2,482	0.4450	0.2775	0.6235	3.8133

Table G1. Sector performance.

EPS bracket	Stage 0	Stage 1	Stage 2	Stage 3
< $0.72	0.6890	0.6648	0.5489	0.3083
$0.72–1.47	0.9324	0.8501	0.6835	0.3874
$1.47–2.37	1.0810	0.9151	0.7144	0.3530
$2.37–3.86	1.1765	0.9425	0.7143	0.3473
> $3.86	1.1828	0.9463	0.7000	0.3542

Table G2. Relative error by EPS scale and stage.

Specification	N	Relative error	DM t (clustered)	DM t (iid)
Baseline	66,914	0.7231	7.5696	40.0928
$0.10 like-for-like	55,505	0.7841	5.9615	32.7592
Unscreened	67,512	0.4563	2.6845	9.3630
Street-basis	69,170	0.5755	1.2707	4.9727

Table G3a. Robustness.

Track record	N	Firms	Consensus MAE	Model MAE	Relative error	DM t
Proven	14,953	1,049	0.1778	0.1284	0.7220	5.8246
Mixed	28,879	1,698	0.3745	0.2544	0.6794	8.3573
Unproven	6,752	814	0.5793	0.4055	0.7000	6.8170

Table G3b. Persistence by track record.

Appendix H Trading strategy construction and results

This appendix provides the results of the trading strategy in Section 4. Table H1 reports the strategy and test detail, including turnover and leverage. Figure H1 shows performance over time for the quarterly and event implementations against the benchmarks.

Strategy	Window	N quarters	Annual return	Volatility	Excess Sharpe	Max drawdown	t-stat	Turnover	Leverage
Quarterly book	2015-06-30 to 2025-09-30	42	9.51%	5.61%	1.40	-8.20%	5.61	52.08%	—
Quarterly vol-scaled	2015-06-30 to 2025-09-30	42	22.10%	15.04%	1.41	-25.46%	4.66	52.08%	2.78
Event trade	2015-06-30 to 2025-09-30	42	9.34%	6.42%	1.16	-0.61%	4.65	—	—
Event vol-scaled	2015-06-30 to 2025-09-30	42	18.02%	15.12%	1.12	-3.72%	3.84	—	2.35
Equal-weight universe	2015-06-30 to 2025-09-30	42	9.58%	18.89%	0.50	-30.81%	1.90	—	—
S&P 500	2015-06-30 to 2025-09-30	42	13.72%	15.60%	0.83	-23.93%	2.97	—	—

Table H1. Strategy and test detail, including turnover and leverage.

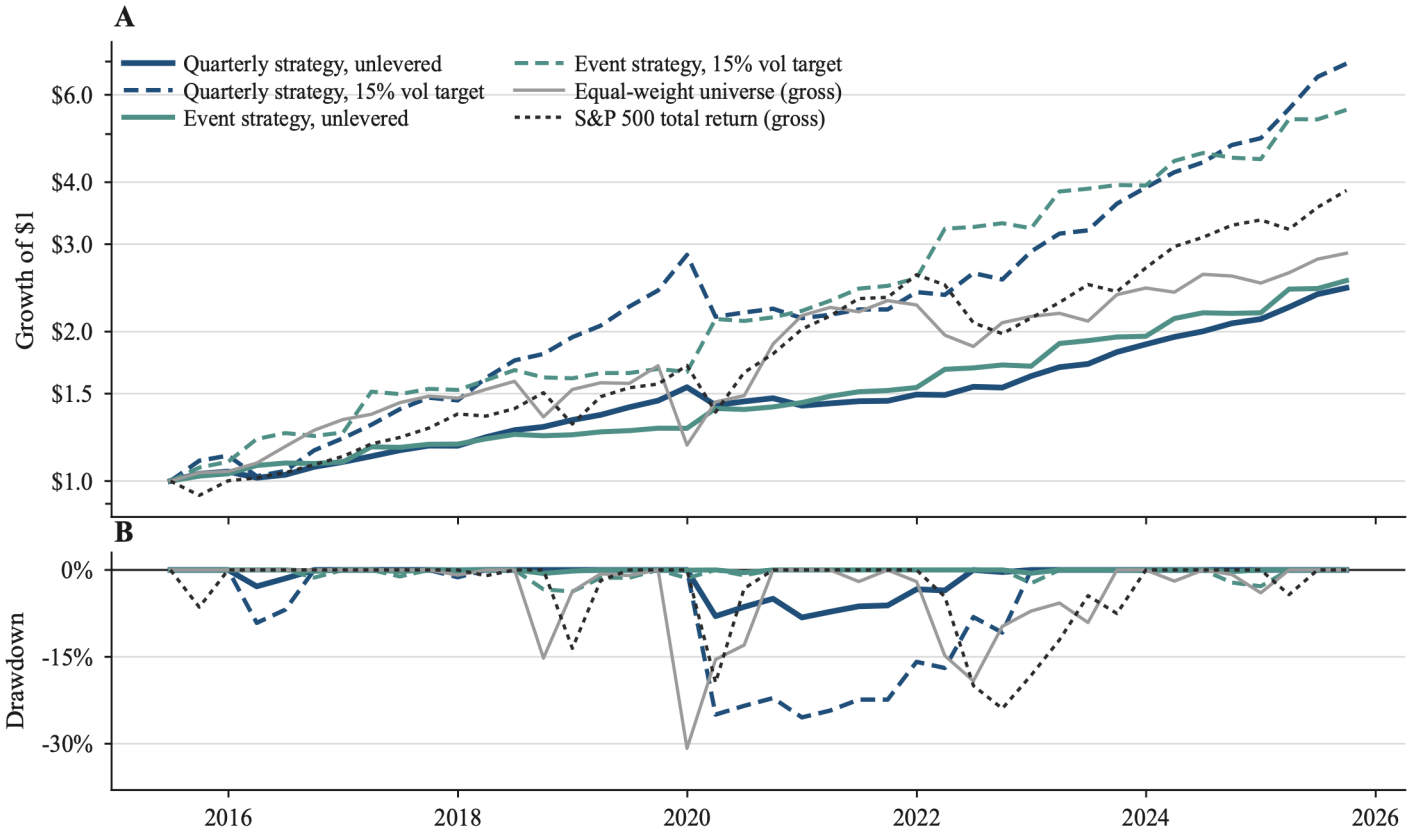

Figure H1: Strategy performance over time. *Panel A: growth of $1, log scale, rebased at 2015-06-30. Panel B: drawdown from running peak. Strategies are net of costs, benchmarks gross; vol-targeted lines gear to 15% at a 3× cap, financed at the bill plus 100 bps.*

Appendix I Worked firm-year examples

This appendix illustrates the cascade on individual firms. Figure I1 traces the model forecast, the calibrated 80% interval, the analyst consensus, and the reported actual for a set of large firms as quarters are reported through the fiscal year.

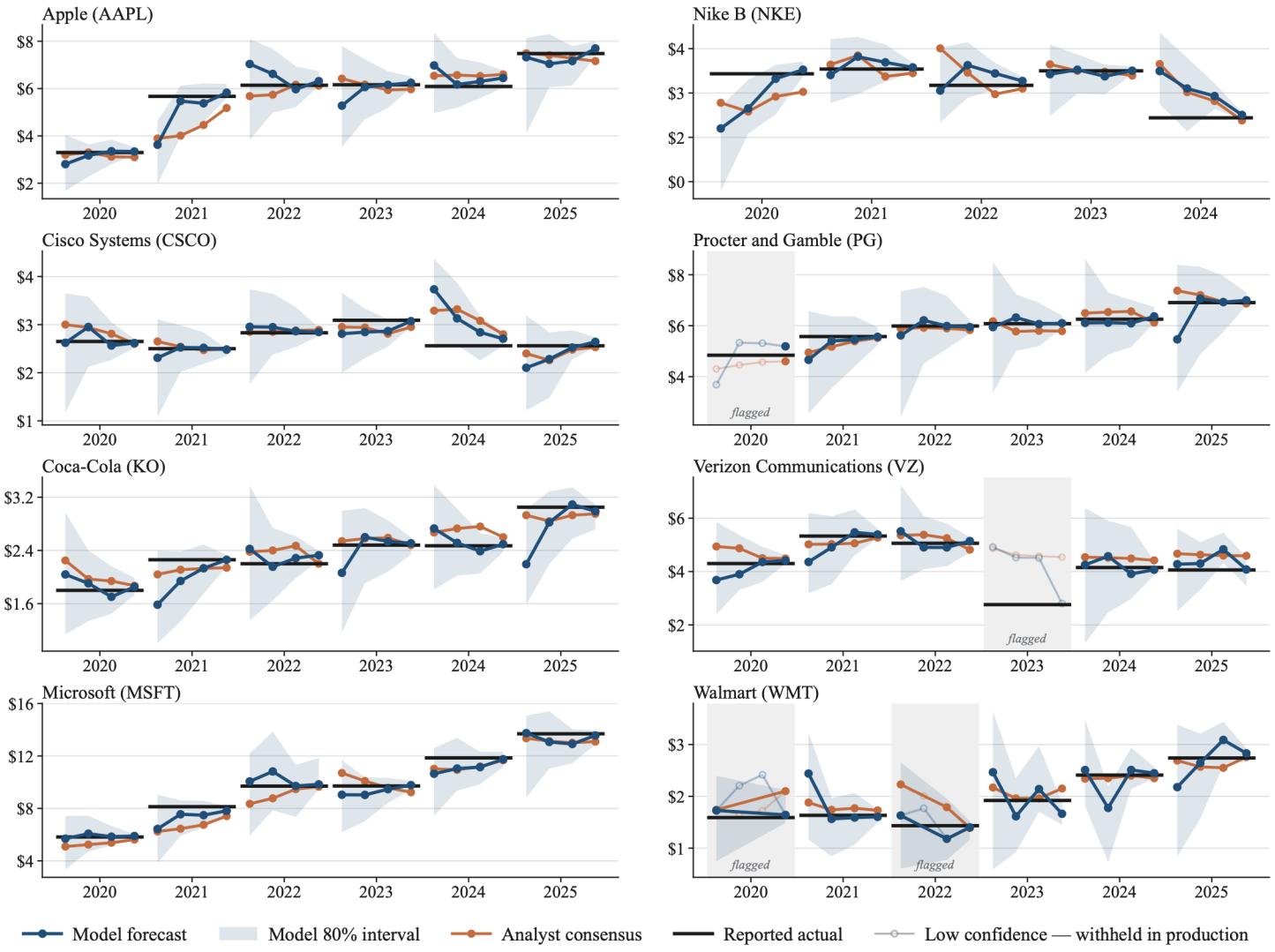

Figure I1. Worked firm-year examples: model forecast, calibrated 80% interval, analyst consensus, and reported actual; flagged years correspond to predictions made with the classifier flag of *LOW*.