

Long-Horizon Forecasting of Complete Financial Statements with FORMA

TRAVIS L. JOHNSON, JIANNAN JIANG, SOUMYABRATA CHAUDHURI,
YIHAO CHEN[†], LAUREN FALVEY[†], and DONAL O’COFAIGH[†]

August 10, 2026

Abstract

Specialist training beats generalist scale when forecasting financial statements. To our knowledge, no prior work jointly forecasts complete financial statements beyond one year, yet in a discounted-cash-flow valuation most firm value sits past that window. We release ProForma-20Q, a reproducible benchmark for forecasting 78 statement line items 1–20 quarters ahead, for anonymized firms, from past statements and an industry code, scored by change-space R^2 . On it, FORMA, a transformer that reads statements as sets of (account, quarter, value) tuples and maximizes a masked-tuple Gaussian likelihood, beats every competitor we field: classical machine learning, chained gradient boosting, a zero-shot time-series foundation model, and frontier large language models. Its lead widens with horizon, where valuation needs accuracy most, and its Gaussian predictive intervals never under-cover. FORMA’s forecasts nearly satisfy accounting identities; exact coherence is recoverable at no statistically significant accuracy cost. Its tuple interface supports scenario analysis without retraining, and we show that pinning future revenue paths sharpens the rest of the statement.

All authors are with the University of Texas at Austin, Austin, TX, USA. [†]These authors contributed equally to this research. We thank Quent Capital, the Langston Family Wealth Management Center, and the AIM Investment Center for financial support. Contact: travis.johnson@mcombs.utexas.edu, jiannanjiang@utexas.edu, schaudhuri@utexas.edu, yihaochen@utexas.edu, lmf2842@my.utexas.edu, do7973@my.utexas.edu.

1. Introduction

A complete *pro forma* financial-statement forecast is the raw material of valuation, credit risk analysis, and financial planning. Yet the academic literature typically supplies only point forecasts of single items, usually earnings, a year ahead. In a discounted cash flow (DCF) valuation, by contrast, most of enterprise value sits in cash flows (not earnings) beyond a year. The largest component is terminal value: the firm’s value at the forecast endpoint (typically 5–10 years), set as a peer or perpetuity multiple of a final-year forecast [39]. A final-year number, however, is only as credible as the statement behind it: margins, reinvestment, and financing must hang together, which is why practitioners build complete pro formas. The economically relevant object is therefore a joint forecast of the whole statement at horizons where value lives. To our knowledge, no prior research provides such forecasts.

We introduce ProForma-20Q, a benchmark for this task. It asks models to predict 78 statement line items at horizons of 1–20 quarters using 12 quarters of lagged values for the same items and an industry code. We release the full protocol, which is reproducible with one command given WRDS access (§3). Because valuation depends on aggregating conditional means across horizons, and we do not want to celebrate propagating persistent levels, models are ranked by out-of-sample R^2 for predicting changes.

Financial statement data resists the standard supervised-learning template. The average firm-quarter in our sample reports only 64 of the 78 line items, individual items are reported in as few as 32% of firm-quarters, and *fewer than 2% of firm-quarters report the complete set*. The data also combines rigid structure (exact cross-account and cross-quarter accounting identities) with soft structure (latent economic states such as competitive moats and financial constraints).

We specify and evaluate FORMA, a transformer-based architecture tailored to this setting. A statement history is encoded as a *set* of (account, quarter, value) tuples, absent items contribute no token, and future items are masked tuples. FORMA outputs a Gaussian predictive mean and variance for each mask and is trained to maximize the likelihood of the value behind the masks.

On ProForma-20Q, a ~ 0.9 M-parameter FORMA beats all competitors (Table 2): classical machine learning (ML) approaches using imputed missing values, such as penalized regressions, random forests (RF) [4, 10], and feed-forward neural networks (FFNNs) [1]; an in-suite re-implementation of its nearest conceptual competitor, the chained gradient boosted machine (GBM) of Geertsema et al. [15]; and two generalists, a zero-shot time-series foundation model (TSFM) and a best-effort panel of three frontier large language models (LLMs). FORMA outperforms every model at every horizon beyond $h=2$ and its edge grows with forecast horizon, reaching at least 3.2 percentage points (pp) of R^2 by $h=20$ (Figure 1). The LLMs’ performance is generally poor: the best frontier model underperforms all but the simplest purpose-trained model, and its R^2 deficit relative to FORMA grows from 5.5pp at $h=1$ to 15.9pp at $h=20$.

Our contributions are:

- **Task and protocol:** We define and release ProForma-20Q, a turnkey protocol for complete financial-statement forecasting, scored in change space on common samples.
- **Architecture-to-problem fit:** We design a tuple-set transformer, FORMA, around the data and task. Unreported items contribute no token (nothing is imputed); identity-aware masking forces the model to learn economics instead of accounting algebra; pinned-future masking trains it to condition on chosen realizations for scenario analysis.
- **Probabilistic statement forecasts:** We produce predictive densities over the statement via a heteroskedastic head and five-seed mixture. These admit exact ex-post reconciliation at statistically insignificant accuracy cost, and Gaussian central intervals never under-cover at any horizon.
- **The specialist-wins-and-widens result:** We document that FORMA outperforms all competitors, including LLMs, and that its advantage widens with horizon.

2. Related Work

Table 1 and the paragraphs below contrast our work with existing research on firm-level financial forecasting across forecast scope, horizon, cross-item modeling, distributional output, and accounting structure. Single-outcome models reach multi-year horizons, whereas multi-output systems cover selected or stylized subsets for at most one year. To our knowledge, FORMA is the first learned system to produce forecasts of a three-statement schema at valuation horizons, and the first to use a tuple-set transformer for this problem.¹

We also discuss three adjacent literatures—LLM forecasting, long-horizon time series, and coherent forecasting—that supply competing models and evaluation methods.

Earnings forecasting. The accounting, finance, and ML literature typically forecasts *earnings* or other individual items. Cross-sectional models in the Hou et al. [24] line [28, 35] and their ML successors [8, 23, 25] predict one target (a point realization or its direction of change) per fitted specification at one- to five-year horizons. Their performance relative to analyst consensus is specification-sensitive [6]. Quantile regressions provide one-year-ahead distributional earnings forecasts [26]. Analyst forecasts and textbook pro formas are practice benchmarks, not learned systems [32, 39].

Statement-level systems. Alberg and Lipton [1] jointly forecast 16 fundamentals one year ahead with multi-task feed-forward and long short-term memory (LSTM) networks; Chauhan et al. [7] extend the design to 17 targets with uncertainty estimates. Divo et al. [12] benchmark 24 diverse models on five items spanning all three statements over four quarters. The closest system, Geertsema et al. [15], forecasts 29 core income-statement and balance-sheet items one year ahead by chaining per-item GBMs and derives 19 more using accounting relations. We re-implement it in-suite and find it underperforms FORMA on its own item footprint.

¹Divo et al. [12] include generic sequence transformers (Transformer, TFT, Chronos) consuming a dense 20-indicator panel to forecast five target items; none attends over sparse tuples, the representation that makes a full 78-item schema tractable.

LLMs and financial statements. Large language models have been proposed as zero-shot forecasters that convert numeric series to token strings and extrapolate them directly [20], though whether the language model itself helps numeric forecasting is disputed [36]. Our LLM panel (§5) gives reasoning-enabled frontier models the same anonymized statements and industry code as every other model, a *stronger* generalist condition than either the token-string prompting of Gruver et al. [20] or the fine-tuned adapters ablated by Tan et al. [36]. We nevertheless show that a small specialist dominates LLMs at every horizon in out-of-sample R^2 .

Long-horizon and foundation-model forecasting. The long-term time-series forecasting (“LTSF”) literature counts “long term” in steps, e.g. hundreds of hourly electric or traffic readings in Zeng et al. [41]. We count “long term” in calendar time—our 20 quarterly steps span five years, far longer than standard LTSF horizons—so the difficulty is signal decay and distribution shift, not sequence length. Zero-shot time-series foundation models [2, 9] supply our second generalist arm (§5.1). Our classical baselines carry Grinsztajn et al. [19]’s trees-versus-deep-learning prior on tabular data: tree ensembles are the strongest competitors here too.

Coherent forecasting. Forecast reconciliation adjusts predictions to satisfy aggregation constraints among hierarchical time series, as in the MinT method of Wickramasuriya et al. [40], with end-to-end and probabilistic extensions [31, 33]. As signed linear relations, accounting identities fall within the general framework of Girolimetto and Di Fonzo [16]. We report unconstrained violations and apply variance-weighted reconciliation in §5.3.

3. ProForma-20Q Task and Protocol

3.1. Task

Each forecasting example is indexed by a firm f and a forecast origin quarter t ; horizon h refers to absolute quarter $t + h$. Let $\mathcal{D}$ denote the 78 accounting-item identifiers and c_f the industry category of firm f . Each value the firm reports is a tuple (h, id, x) , where $\text{id} \in \mathcal{D}$ identifies the accounting item and x is its value at quarter $t + h$ after the standardization of §3.2.

The historical input is a set

$$S_{f,t} = \{(h, \text{id}, x) : h = -11, \dots, 0, \text{id} \in \mathcal{D}, \text{id reported at } t+h\}.$$

For any requested future horizon and accounting item, the forecasting task is to learn

$$\hat{x} = g(S_{f,t}, c_f; h, \text{id}), \quad h = 1, \dots, 20, \quad \text{id} \in \mathcal{D}, \quad (1)$$

where $\hat{x}$ is the forecast of the corresponding standardized value at quarter $t + h$ given the information up to t .

ProForma-20Q thus fixes 12 quarters of history, 78 items, and a 20-quarter horizon. The formulation itself accommodates other item sets, histories, and horizons.

The tuple set is the task’s canonical encoding: it records exactly what the firm reported and lets the reported subset vary freely across firm-quarters. Models need not consume it; the released protocol builds both the canonical *tuple view* and a conventional *tabular view* of the same panel for models requiring rectangular inputs (§5.1).² Forecasts from either view are scored identically.

3.2. Data, Standardization, and Splits

The sample is quarterly U.S. filings (Compustat), excluding financial firms (SIC 6000–6999), whose statements comprise different line items with different economic meaning. Values are deflated by scale (origin-quarter $|\text{total liabilities}| + |\text{shareholders' equity}|$, which nearly always equals total assets), asinh-transformed, and standardized per (item, quarter) with statistics estimated from data available at the origin; standardized inputs are clamped at $|x| \leq 6$.

Splits are temporal — train 1971–2001, validation 2002–2009, test 2010–2024 — with targets purged at each boundary: an example retains only target quarters realized within its own split, so nothing fit or selected before the test sample observes any post-2009 outcome. Every model conditions on the same 12-quarter reported history, origin-quarter scale, and Fama–French 48

²The tabular view’s features include per-item recent levels and year-over-year changes on a fixed grid, with industry dummies and imputed missing values.

industry (FF48) [13]. None sees firm identity. We exclude stock returns, analyst forecasts, and other listed-firm features so that FORMA can be applied to private firms. The panel spans 32,851 firms and 1,173,598 firm-quarters (609,269 train, 211,367 validation, 352,962 test), averaging 758 historical and 1,124 target tuples each.

The underlying data is licensed and cannot be redistributed, so we instead release a repository containing the full pipeline, configurations, evaluation code, and documentation of the complete item list, identities, standardization, and filtering procedures.³ With it, anyone having WRDS access can rebuild the exact data environment in one command, verified by published checksums.

3.3. Evaluation Using R^2 for Changes

We evaluate forecasts of *changes* in standardized values: the truth is $x_{f,t}(h, \text{id}) - x_{f,t}(0, \text{id})$ and a model’s prediction is $\hat{x}_{f,t}^m(h, \text{id}) - x_{f,t}(0, \text{id})$. Because the anchor $x_{f,t}(0, \text{id})$ cancels out, mean absolute error (MAE) and mean squared error (MSE) are numerically identical in levels and changes; only the R^2 metric differs. Every R^2 in this paper reads as skill over assuming sample-average changes; in change space, R^2 of 20–40% at multi-year horizons is strong, and level-space intuitions of 90% or higher do not apply.

A DCF valuation sums discounted *expected* cash flows across horizons. Medians do not add ($\text{med}(A+B) \neq \text{med}(A) + \text{med}(B)$), so median-targeted forecasts cannot feed a valuation. Our primary metric is therefore squared error in standardized space (Panel A of Table 2), which scores conditional-mean performance on the scale where items and firms are comparable. Because the asinh transform is nonlinear, recovering expected dollar cash flows requires the full predictive distribution.⁴ Distributional output is thus the bridge from standardized-space skill to dollar-space valuation, and we evaluate it when available (Panel C of Table 2).

We additionally report an absolute-error track (Panel B of Table 2), which scores conditional medians, for fairness to MAE-native models and to diagnose whether a model’s poor R^2 reflects poorly located forecasts or implicit median-targeting.

³See <https://github.com/forma-lab-mccombs/proforma-20q>.

⁴Models with Gaussian predictive distributions, such as FORMA, admit a closed form: when $X \sim N(\mu, \sigma^2)$ in standardized space, $\mathbb{E}[\sinh(aX + b)] = e^{a^2\sigma^2/2} \sinh(a\mu + b)$.

Common samples. Every comparison is evaluated on a common sample: a cell (firm, origin, item, horizon) contributes only if ground truth is available and every compared model predicts it. Models restrict coverage to different subsets of firms, items, origins, and/or horizons. A single all-model intersection would compound every such restriction and yield a severely limited sample. The protocol instead groups models by restriction and reports one footprint per group, re-scoring the full model suite inside each. Metrics are comparable within a footprint and never across footprints. Table 2 reports three such footprints; §5.1 states which restriction binds each comparator. Finally, a cell at horizon h is scoreable only if the firm reports at $t+h$, so long-horizon evaluation conditions on realized survivors; comparisons remain fair (identical cells for all models), but level interpretations carry that caveat.

Significance. We measure statistical significance by Diebold and Mariano [11] tests, accounting for dependence across cells in the same quarter by collapsing to calendar-quarter means (the effective sample is the 60 test quarters, not 327M cells), with Newey and West [29] standard errors using bandwidth 19 for overlapping origins, and the Harvey et al. [21] (HLN) small-sample correction; positive statistics favor the comparator.

4. FORMA: A Tuple-Set Transformer

Tuple-set representation. Fix a firm-origin example (f, t) . Each observed historical tuple $(h, \text{id}, x) \in S_{f,t}$ is a token, and each requested future pair (h, id) is a query token whose value is hidden from the encoder. The initial representation of a token is

$$z_{h,\text{id}}^{(0)} = E_{\text{acct}}(\text{id}) + E_{\text{horizon}}(h) + E_{\text{value}}(x, m), \quad (2)$$

where m indicates whether the token’s value is hidden and

$$E_{\text{value}}(x, m) = \begin{cases} w_x x, & m = 0, \\ e_{\text{mask}}, & m = 1. \end{cases} \quad (3)$$

Here E_{acct} is a learned account embedding, E_{horizon} a fixed sinusoidal encoding of the relative quarter h , and w_x a learned projection for the standardized value.

An unreported historical item contributes no tuple and hence no token, while masked historical observations and query tokens remain in the set with value e_{mask} . Missingness is therefore native to the representation: incomplete statements can be used without imputation, avoiding complete-case selection and the researcher degrees of freedom afforded by an imputation rule — a choice every tabular model must make (§5.1).

Two context tokens complete the input: the firm’s industry, $z_{\text{ind}}^{(0)} = E_{\text{industry}}(c_f) + E_{\text{horizon}}(0)$, and the origin-quarter scale deflator of §3.2, standardized like any other value (without self-deflation); neither is a prediction target. Tokens are stacked unordered to form $H^{(0)}$, and a Transformer encoder [37] attends over the set as in Lee et al. [27], so every token conditions on the full context. The encoder has 4 layers, $d_{\text{model}} = 128$, and 4 attention heads ($\approx 0.9\text{M}$ parameters).

Training objective. Training uses masked prediction over the tuple set. The loss is evaluated on masked accounting tokens with observed targets, and masking has two designed variations.

(a) *Identity-aware grouped masking.* We randomly mask some reported historical tuples for regularization. However, statement values are tied by exact linear identities, so masking single tuples could teach constraint algebra rather than economics. When masking touches a complete identity group we therefore mask at least two of its members; tuples with no complete identity instance in the tuple set are randomly masked as singletons.

(b) *Pinned-future masking.* Half the training examples mask the entire future (pure forecasting); the other half reveal $\approx 5\%$ of reported future tuple values as inputs. The model thereby learns to condition on partial future information, the capability behind the scenario analysis of §5.4.

The future query grid is constructed ex ante and firm-uniform, so that slot existence never depends on which cells the firm later reports. Otherwise the mere pattern of slots would reveal account additions, removals, or exit from the sample—a survivorship and look-ahead leak whose closure cost ≈ 0.7 pp of test R^2 .

Direct multi-horizon probabilistic forecasts. For each requested pair (h, id) , we map $u = \text{Concat}[z_{h,\text{id}}^{(L)}, E_{\text{horizon}}(h)]$ to a location and a heteroskedastic scale:

$$\mu_{f,t}(h, \text{id}) = f_{\mu}(u), \quad \log \sigma_{f,t}^2(h, \text{id}) = f_{\sigma}(u), \quad (4)$$

with $\mu_{f,t}(h, \text{id})$ serving as the task’s point forecast $\hat{x}$. We implement each mapping as a two-layer multilayer perceptron (MLP).

The primary head is a heteroskedastic Gaussian (μ, σ) , trained with the β -NLL loss ($\beta = 0.5$) [34], under a horizon curriculum; for the MAE track we also train a Laplace head $(\mu, b = \sigma/\sqrt{2})$ with the corresponding β -NLL.

We train five seeds and treat their predictive distributions as an equal-weight mixture. Distributional metrics use the exact mixture rather than averaged parameters: NLL uses the mixture density, CRPS a closed-form expression for the mixture family.⁵

5. Results

5.1. Competing Models

Every comparator receives the information set of §3 and is scored in its native track. The tabular baselines (elastic net, RF, and two FFNNs) consume the benchmark’s tabular view: four recent levels and eight year-over-year changes per item plus industry dummies, with missing features imputed via the local XS estimator of Bryzgalova et al. [5].⁶ Each forecasts every (item,

⁵For the Gaussian mixture, we use Grimit et al. [18]; for the Laplace mixture, we use the analogous expression obtained from $\text{CRPS} = \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'|$.

⁶The imputed feature matrix itself contains lagged levels and changes, so backward-looking information enters the factor model in the spirit of their local B-XS variant.

horizon) pair directly rather than recursively, so comparisons isolate the pipeline, not forecast strategy; the penalized regression and RF fit a separate model per pair, while the FFNNs and FORMA share parameters across all pairs through a vector output head. The FFNNs are five-seed ensembles with the same heteroskedastic head and mixture treatment as FORMA, so they also enter the density track.

Two simple baselines anchor the table. The seasonal random walk repeats each item’s most recent value from the same fiscal quarter, the standard naive expectation for quarterly accounting series [3, 14]. The fade/AR(1) baseline fits one pre-test OLS regression per item and horizon, pooled across firms, on the current level, capturing mean reversion.

The chained GBM of Geertsema et al. [15] is the only prior model we are aware of that forecasts a sizable block of the statement. We re-implement it under our common protocol using the 12-quarter tabular feature view rather than the original paper’s single annual lag. We estimate the original L_1 specification for the absolute-error track and an L_2 variant for the squared-error track. Because 25 of the chained items lie in the ProForma-20Q universe, comparisons use the corresponding item footprint (the “Geert. sample”).

The generalists come in two flavors. We prompt three frontier LLMs (Claude Opus 4.8, GPT-5.5, Claude Sonnet 5; all with extended thinking enabled) with the firm’s industry and the 12-quarter anonymized history as dollar values. They forecast dollar values of the statement primitives at $h=1-20$; we reconstruct the remaining items via the accounting identities, apply the standardization of §3, and score on a calendar-balanced 2,103-origin subsample (the “LLM sample”). Each model runs two prompt arms; Table 2 reports the better arm per model, and our full prompts are available in the release documentation.⁷ The LLMs emit point forecasts without a stated estimand, so we score them in both point tracks.

The zero-shot TSFM Chronos-2 [2] receives the same 12-quarter context in raw dollar values (it applies its own normalization), treating the 78 items as one multivariate group with missing values handled natively. It emits 21 native quantiles, which we map into benchmark space

⁷See <https://github.com/forma-lab-mccombs/forma-release>.

through the origin-frozen transform. Each track scores Chronos-2 at its matched estimand: the quantiles are integrated (trapezoidal, flat tails) to a conditional mean for the squared-error track, read at the median for the absolute-error track, and scored as a predictive distribution in the density track.

5.2. Forecast Performance

Full sample. Panel A of Table 2 presents the headline: FORMA explains 28.9% of the cross-sectional variance of realized changes, ahead of the RF (27.2%), penalized regression (25.8%), and both FFNN baselines (25.3% and 24.7%). Every gap is significant at the 1% level under quarter-clustered DM tests.

Figure 1 shows that the RF slightly outperforms FORMA one quarter ahead (R^2 39.8% vs. 39.0%; DM $t=+12.3$), but the models are at parity at $h=2$, and from $h=3$ FORMA is ahead with significance growing through $h=20$ (0.225 vs. 0.193; t reaching -18.6). RF may therefore be better suited to short-term earnings timing, while FORMA suits medium- and long-term tasks such as valuation or credit risk analysis.

The mechanism is visible in the baseline ladder. Fundamentals mean-revert at conditional, item-specific rates [30]. The seasonal random walk misses reversion entirely ($R^2=-4.1\%$). The pooled fade/AR(1) baseline captures *unconditional* reversion and explains 18.3% of the variation, over half of FORMA’s total. What separates the models is the *conditional* component—reversion speeds that depend on the rest of the statement. That is also where capacity alone fails: the larger FFNN (≈ 4.2 M parameters) *underperforms* its linear sibling (24.7% R^2 vs. 25.3%; ≈ 3.1 M parameters), and both trail the ≈ 0.9 M-parameter transformer by 3.6–4.2 percentage points. Architecture, not capacity, drives FORMA’s edge.

Chained GBM In Panel A, the L_2 GBM variant achieves an R^2 of 18.5%, compared with 24.7% for FORMA on their common sample. On the absolute-error track, the original L_1 specification of Geertsema et al. [15] posts an MAE of 0.400, compared with 0.364 for the MAE-targeting Laplace FORMA, a difference significant at the 1% level. Figure 1(b) shows a similar short-horizon pattern

to the RF comparison: the GBM leads at $h = 1$ (32.6% versus 32.1% R^2 ; DM $t = +6.9$), but FORMA leads from $h = 2$ onward: 25.1% versus 17.2% at $h = 10$ to 19.7% versus 13.3% at $h = 20$.

Generalists: LLMs and TSFMs. On the task metric R^2 , the verdict is one-sided: FORMA achieves 29.9% on the LLM sample, compared to 18.6% for the top-performing LLM. In fact, every purpose-trained model other than fade/AR(1) outperforms every LLM, and every LLM outperforms Chronos-2, which beats only the seasonal random walk. The FORMA–LLM gap widens from 5.5 percentage points at $h=1$ to 15.9 at $h=20$, with the best LLM (Opus 4.8) collapsing from 32.5% to 7.8% while FORMA only decays from 38.0% to 23.7%. Note the comparison is conservatively biased *toward* the generalists as test-period financial statement realizations sit in their pretraining data.

Panel B supports the diagnosis promised in §3.3: the LLMs’ MAEs (0.362–0.368) are only slightly higher than that of the Laplace FORMA (0.348) and lower than those of Chronos-2 and the seasonal random walk. Consistent with conditional medians, these forecasts are competitive in absolute error but underperform under the squared-error criterion essential for valuation.

The LLM approach also has disadvantages in cost and reproducibility. Frontier models cost $\approx \$0.12$ – $\$0.26$ per origin while inference cost is negligible on our ≈ 0.9 M-parameter FORMA. Frontier LLMs with extended thinking produce non-deterministic outputs without the option of a reduced sampling temperature. We quantify this wobble with five repeated forecasts per origin on a 200-origin sample. For Opus 4.8, the most stable model, the cross-call standard deviation grows from 0.03 at $h=1$ to 0.06 at $h=20$ in normalized z -space (cross-sectional s.d. 1). These wobbles are independent across origins and wash out in the full 2,103-origin metrics.

Probabilistic quality. Panel C evaluates the models that produce predictive distributions: the specialist mixtures and Chronos-2’s native quantiles; the LLMs, emitting only a point forecast, drop out. Among the mixtures, the exact five-seed NLL and closed-form CRPS rank FORMA’s Laplace variant clearly first and its Gaussian variant second, both ahead of the FFNN variants.

Beyond the proper scores NLL and CRPS, calibration asks whether conditional quantiles and variances can be taken at face value [17]; we assess it with the mixture probability integral trans-

form (PIT). While individual seeds are mildly overconfident (standardized MSE $\bar{z}^2=1.10$), the five-seed Gaussian mixture is essentially calibrated in total variance ($\bar{z}^2=0.96$, the between-seed spread supplies the missing variance). Its central intervals never under-cover at any horizon—pooled coverage is 72.6/89.6/93.7/95.8% at nominal 50/80/90/95% (per-horizon series in the release).

The PIT is center-heavy, so the intervals are conservative rather than sharp, and by horizon $\bar{z}^2$ drifts from 1.00 to 0.89—the long end errs on the safe side, the right side for terminal-value use. The feed-forward mixtures are the conservative-but-blunt contrast: notably overdispersed (mixture $\bar{z}^2$ of 0.85 and 0.64) with worse CRPS—it is easy to be conservative, hard to be conservative *and* sharp. Chronos-2 posts a competitive CRPS (0.336), with near-nominal 50% central coverage, but still trails both FORMA variants; its NLL is unreported because degenerate zero-width intervals on flat contexts ($\sim 0.6\%$ of cells) admit no usable log score even under a 10^{-4} scale floor.

5.3. *Coherence: Emergent and Recoverable*

A forecast financial statement should add up, and FORMA largely does so without explicit constraints. Across 124.7M enforced identity instances (an identity counts at a firm-origin-horizon when every member account is forecast), the median absolute violation of FORMA’s raw-dollar statements is 3.7% of the identity’s gross scale. The model has substantially learned accounting structure from data. Some violation is expected even from a perfect model because dollar statements plug the standardized-space conditional means into the inverse transform, and means do not commute with the nonlinear map, so part of the violation is a property of the estimand rather than prediction error.

Exact coherence is recoverable ex post. Following the forecast reconciliation tradition [40], we work in raw dollars, where the identities are linear, and project each forecast statement onto the subspace where every identity holds exactly. We compare two projections: equal-weighted, which minimizes the total squared dollar adjustment, and variance-weighted, which measures each item’s adjustment in units of its predictive standard deviation, mapped into dollars via the delta method, so accounts the model is less certain about absorb more of the residual; cross-account covariances are not modeled.

Table 3 shows that the variance-weighted projection drives violations to numerical zero at no statistically significant squared-error cost (R^2 drops by 3.8 percentage points, quarter-clustered DM $t=-1.4$) while yielding a small but significant MAE *improvement* ($t=+4.8$).⁸ The equal-weight variant is catastrophic (R^2 falls to -5.51 , MAE rises to 0.635): a dollar adjustment spread uniformly across accounts is negligible for total assets but enormous relative to small line items such as minority interest (zero at most firms), and standardized-space errors price exactly that. This contrast illustrates the setting-specific payoff of probabilistic forecasting: the variance head enables affordable reconciliation.

5.4. Scenario Analysis by Conditioning

Scenario analysis is an integral part of the valuation process whereby an analyst makes assumptions or expresses beliefs about line items such as revenues and populates the remainder of the forecast statement based on these assumptions. FORMA natively supports this interface: any future tuple can be revealed as an input, and the remaining query tokens are forecast conditional on it. We train FORMA on this task using the pinned-future masking objective (§4). Arbitrary conditioning patterns can therefore be specified at inference time without retraining.

We evaluate scenario analysis performance with an oracle experiment by pinning the *true realized* revenue path (Q1–Q20) per origin. We then forecast the remaining items, and compare against the unconditional forecast on identical origins. Using the true realized revenue values for revenue assumptions makes our measurement an upper bound on scenario value. Conditional forecasting is a recognized task in the macro literature [38], to our knowledge never posed to a learned firm-level financial statement model.

The results of the oracle experiment, presented in Figure 2, confirm the interface works and that conditioning on realized revenues is economically valuable. Pinning the true revenue path lowers pooled MAE on the remaining items from 0.409 to 0.383 and raises change-space R^2

⁸The reconciled figure should not be compared to Panel A’s entries, none of which add up. A coherent-to-coherent comparison would charge every model its own reconciliation cost. Forecasters without predictive variances would need weights estimated from pre-test residuals, and those cannot be observation-specific. The MAE column scores the Gaussian conditional-mean forecasts, not Panel B’s median-targeting Laplace head, hence 0.408 vs. Panel B’s 0.369 .

from 30.5% to 34.8% (768 origins, 916,846 non-pinned cells; five-seed mixture mean). The gain is small one quarter out (0.9pp) and widens with horizon to 7.4pp by $h=20$. Near-term items are already determined, but at long horizons the revenue path anchors the whole statement. Income-statement and balance-sheet items gain equally in absolute error ($\Delta\text{MAE} -0.032$ each). Balance-sheet items gain most in R^2 (+8.9pp vs. +5.6pp). Cash-flow items barely move ($\Delta R^2 +0.7\text{pp}$), reflecting their weak link to revenue.

6. Conclusion

Forecasting complete financial statements at valuation-relevant horizons is a task that, to our knowledge, no prior work covers. We show that it is tractable: FORMA, a sub-million-parameter transformer with a missingness-native representation and an identity-aware masked objective, beats classical ML, the statement-forecasting system of Geertsema et al. [15], a zero-shot foundation model, and frontier LLMs. FORMA’s lead widens toward the horizons where firm value actually sits, and its Gaussian prediction intervals never fall below nominal coverage.

Some limitations apply. Long-horizon evaluation conditions on realized survivors: comparisons are fair (identical cells for all models), but absolute skill levels describe only the survivors. Our claims are scoped to quarterly U.S. filings, these 78 accounting items, and horizons of 1–20 quarters. We measure forecast quality only; valuation applications require additional inferences about discount rates and terminal values. We do not incorporate stock market or analyst data as additional features, and only consider off-the-shelf generalists rather than fine-tuned ones.

We release ProForma-20Q for adoption as a common task framework for statement forecasting [22]: pipeline, configurations, and evaluation protocol are rebuildable from WRDS with one command (<https://github.com/forma-lab-mccombs/proforma-20q>). We also release code, configurations, and trained weights for FORMA, and seeded regeneration scripts for the learned competitors (<https://github.com/forma-lab-mccombs/forma-release>). The appendix provides additional details.

References

- [1] John Alberg and Zachary C. Lipton. Improving factor-based quantitative investing by forecasting company fundamentals, 2017. NIPS 2017 Time Series Workshop; v2 revised 26 April 2018.
- [2] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. doi: 10.48550/arXiv.2510.15821.
- [3] Victor L. Bernard and Jacob K. Thomas. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics*, 13(4):305–340, 1990. doi: 10.1016/0165-4101(90)90008-R.
- [4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.
- [5] Svetlana Bryzgalova, Sven Lerner, Martin Lettau, and Markus Pelger. Missing financial data. *The Review of Financial Studies*, 38(3):803–882, 2025. doi: 10.1093/rfs/hhae036.
- [6] John L. Campbell, Harrison Ham, Zhongjin (Gene) Lu, and Katherine Wood. Expectations matter: When (not) to use machine learning earnings forecasts. *Management Science*, 2026. doi: 10.1287/mnsc.2024.05808. Articles in Advance, 1–25; published online 12 February 2026.
- [7] Lakshay Chauhan, John Alberg, and Zachary C. Lipton. Uncertainty-aware lookahead factor models for quantitative investing. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1489–1499. PMLR, 2020. URL <https://proceedings.mlr.press/v119/chauhan20a.html>.

- [8] Xi Chen, Yang Ha (Tony) Cho, Yiwei Dou, and Baruch Lev. Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research*, 60(2):467–515, 2022. doi: 10.1111/1475-679X.12429.
- [9] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR, 2024. URL <https://proceedings.mlr.press/v235/das24c.html>.
- [10] Tim de Silva and David Thesmar. Noise in expectations: Evidence from analyst forecasts. *The Review of Financial Studies*, 37(5):1494–1537, 2024. doi: 10.1093/rfs/hhad091.
- [11] Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.
- [12] Felix Divo, Eric Endress, Kevin Endler, Kristian Kersting, and Devendra Singh Dhami. Forecasting company fundamentals. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=haf78jerSt>.
- [13] Eugene F. Fama and Kenneth R. French. Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193, 1997. doi: 10.1016/S0304-405X(96)00896-3.
- [14] George Foster. Quarterly accounting data: Time-series properties and predictive-ability results. *The Accounting Review*, 52(1):1–21, 1977. doi: 10.2308/tar-4493420.
- [15] Paul Geertsema, Helen Lu, and Guang Ma. Projecting financial statements with chained machine learning. *SSRN Working Paper 5039433*, 2026. doi: 10.2139/ssrn.5039433.
- [16] Daniele Girolimetto and Tommaso Di Fonzo. Point and probabilistic forecast reconciliation for general linearly constrained multiple time series. *Statistical Methods & Applications*, 33(2):581–607, 2024. doi: 10.1007/s10260-023-00738-6.

- [17] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.
- [18] Eric P. Grit, Tilmann Gneiting, Veronica J. Berrocal, and Nicholas A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2925–2942, 2006. doi: 10.1256/qj.05.235.
- [19] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*, pages 507–520. Curran Associates, Inc., 2022. doi: 10.52202/068431-0037.
- [20] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pages 19622–19635, 2023. doi: 10.52202/075280-0861.
- [21] David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997. doi: 10.1016/S0169-2070(96)00719-4.
- [22] Oliver Hellum, Theis Ingerslev Jensen, Bryan T. Kelly, and Lasse Heje Pedersen. The power of the common task framework. *SSRN Working Paper 5242901*, 2025. doi: 10.2139/ssrn.5242901.
- [23] Dieter Hess, Frederik Simon, and Sebastian Weibels. Interpretable machine learning for earnings forecasts: Leveraging high-dimensional financial statement data. *SSRN Working Paper 4619313*, 2026. doi: 10.2139/ssrn.4619313. CFR Working Paper 25-06.
- [24] Kewei Hou, Mathijs A. van Dijk, and Yinglei Zhang. The implied cost of capital: A new

- approach. *Journal of Accounting and Economics*, 53(3):504–526, 2012. doi: 10.1016/j.jacceco.2011.12.001.
- [25] Stewart Jones, William J. Moser, and Matthew M. Wieland. Machine learning and the prediction of changes in profitability. *Contemporary Accounting Research*, 40(4):2643–2672, 2023. doi: 10.1111/1911-3846.12888.
- [26] Theodosia Konstantinidi and Peter F. Pope. Forecasting risk in earnings. *Contemporary Accounting Research*, 33(2):487–525, 2016. doi: 10.1111/1911-3846.12158.
- [27] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR, 2019. URL <https://proceedings.mlr.press/v97/lee19d.html>.
- [28] Kevin K. Li and Partha Mohanram. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies*, 19(3):1152–1185, 2014. doi: 10.1007/s11142-014-9282-y.
- [29] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.
- [30] Doron Nissim and Stephen H. Penman. Ratio analysis and equity valuation: From research to practice. *Review of Accounting Studies*, 6(1):109–154, 2001. doi: 10.1023/A:1011338221623.
- [31] Anastasios Panagiotelis, Puwasala Gamakumara, George Athanasopoulos, and Rob J. Hyndman. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, 306(2):693–706, 2023. doi: 10.1016/j.ejor.2022.07.040.

- [32] Min Park and Tzachi Zach. Analysts’ forecasting models and uncertainty about the past. *Review of Accounting Studies*, 30(3):2376–2418, 2025. doi: 10.1007/s11142-025-09898-0.
- [33] Syama Sundar Rangapuram, Lucien D. Werner, Konstantinos Benidis, Pedro Mercado, Jan Gasthaus, and Tim Januschowski. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8832–8843. PMLR, 2021. URL <https://proceedings.mlr.press/v139/rangapuram21a.html>.
- [34] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antić, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=aPOpX1nV1T>.
- [35] Eric C. So. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics*, 108(3):615–640, 2013. doi: 10.1016/j.jfineco.2013.02.002.
- [36] Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, pages 60162–60191, 2024. doi: 10.52202/079017-1922.
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
- [38] Daniel F. Waggoner and Tao Zha. Conditional forecasts in dynamic multivariate models. *The Review of Economics and Statistics*, 81(4):639–651, 1999. doi: 10.1162/003465399558508.

- [39] James M. Wahlen, Stephen P. Baginski, and Mark T. Bradshaw. *Financial Reporting, Financial Statement Analysis and Valuation*. Cengage Learning, 10th edition, 2023. ISBN 9780357722091.
- [40] Shanika L. Wickramasuriya, George Athanasopoulos, and Rob J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019. doi: 10.1080/01621459.2018.1448825.
- [41] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Table 1: Academic research on forecasting firm-level financial statements. No surveyed system jointly forecasts the statement beyond one year, let alone distributionally; FORMA does both.

Research type (representative papers)	Forecast outcome	Outputs	Horizon (Years)	Architecture	Cross-item modeling	Distributional output	Accounting structure
<i>Panel A. Single-item earnings forecasts</i>							
Accounting-based earnings regressions (Hou et al. [24], So [35])	Annual earnings	1	1–5 ^a	OLS	—	—	—
ML earnings models (Campbell et al. [6], Chen et al. [8], Hess et al. [23])	Earnings or earnings direction	1 each	1–5 ^a	Trees, regressions, FFNNs	—	—	—
Quantile earnings-risk model (Konstantinidi and Pope [26])	Distribution of annual earnings	1	1	Quantile regression	—	✓	—
<i>Panel B. Multi-item financial-statement forecasts</i>							
Multitask neural fundamentals (Alberg and Lipton [1], Chauhan et al. [7])	Selected statement items	16–17	1	FFNN, LSTM	Shared model	[7]	—
Multi-model fundamentals benchmark (Divo et al. [12])	Five selected items across all three statements	1 or 5	1	24 models ^d	Varies	Varies	—
Chained statements (Geertsema et al. [15])	Partial income statement and balance sheet	29 + 19 ^b	1	Chained GBMs	Sequential chain	—	Articulated subset
<i>Panel C. Long-horizon complete statement forecasts (this paper)</i>							
FORMA	Standardized income statement, balance sheet, and cash-flow statement	78	5	Tuple-set transformer	Shared model	✓	Optional projection^c

Notes: *Outputs* counts forecast quantities per system. *Cross-item modeling* requires multiple output items modeled jointly or with shared parameters; statement variables used only as predictors do not qualify. A check denotes a continuous distributional forecast (binary class probabilities do not qualify); a dash denotes absent; “Varies” means the cited benchmark spans models with and without the property. ^aRange across the cited studies. ^b29 forecast directly; 19 computed from accounting relations. ^cIdentities can be imposed exactly ex post (§5.3). ^dFrom local statistical baselines to deep sequence models, including generic transformers (Transformer, TFT) and the pretrained Chronos.

Table 2: ProForma-20Q test-set results by loss geometry. Panel A scores squared error (change-space R^2) on conditional-mean forecasts, Panel B absolute error (MAE) on median forecasts, Panel C proper scores on predictive densities. Columns are exact common samples—Full, Geert. [15], and LLM; compare within columns, never across. Dashes: no forecast on that sample (for Chronos-2’s NLL, no usable log score; see text). Significance: Diebold–Mariano vs. FORMA within column (quarter-clustered, Newey–West, HLN-corrected); */**/** FORMA better at 10/5/1%, † comparator better. Model specifications: §5.1.

<i>Panel A: squared-error track — change-space $R^2 \uparrow$ (conditional-mean forecasts)</i>				<i>Panel B: absolute error — MAE↓ (median)</i>			
Model	Full sample 327.2M	Geert. sample 109.1M	LLM sample 2.15M	Model	Full	Geert.	LLM
FORMA (Gaussian, 5-seed)	0.289	0.247	0.299	FORMA (Laplace, 5-seed)	0.369	0.364	0.348
Random Forest	0.272***	0.231***	0.279***	Chronos-2 (median)	0.431***	0.418***	0.407***
Elastic Net	0.258***	0.217***	0.269***	Seasonal random walk	0.445***	0.426***	0.425***
FFNN (linear, 5-seed)	0.253***	0.207***	0.264***	Chained GBM (L1) [15]	—	0.400***	—
FFNN (large, 5-seed)	0.247***	0.200***	0.253***	Claude Opus 4.8	—	—	0.362**
Fade / AR(1)	0.183***	0.170***	0.179***	GPT-5.5	—	—	0.363**
Chronos-2 (mean)	0.155***	0.108***	0.152***	Claude Sonnet 5	—	—	0.368***
Seasonal random walk	−0.041***	−0.068***	−0.041***				
Chained GBM (MSE) [15]	—	0.185***	—	<i>Panel C: density track (full sample)</i>			
Claude Opus 4.8	—	—	0.186***	Model	NLL↓	CRPS↓	Cov ₅₀ (0.50)
GPT-5.5	—	—	0.174***				Cov ₉₀ (0.90)
Claude Sonnet 5	—	—	0.158***	FORMA (Laplace)	0.160	0.293	0.637
				FORMA (Gaussian)	0.660	0.318	0.726
				FFNN (large)	0.842	0.349	0.700
				FFNN (linear)	0.992	0.369	0.781
				Chronos-2	—	0.336	0.509
							0.875

Table 3: Coherence and its cost. For each enforced identity instance (firm $\times$ origin $\times$ horizon $\times$ identity where all member accounts are forecast), we calculate the absolute violation $|\sum_j s_j \hat{v}_j|$ in raw dollars as a share of the identity’s gross scale $\sum_j |\hat{v}_j|$; we report the median in the identity-violations column. Reconciliation is an ex-post projection in raw-dollar space; untouched cells pass through unchanged, so all rows share the Panel A common sample. Significance: Diebold–Mariano vs. the raw transformer within column (quarter-clustered, Newey–West, HLN-corrected): $^*/^{**}/^{***}$ raw better at 10/5/1%, † variant better at 1%.

Variant	$R^2 \uparrow$	MAE $\downarrow$	Identity viol. $\downarrow$
Transformer-only (raw)	0.289	0.408	3.7%
+ variance-weighted reconciliation	0.251	0.407†	≈ 0
+ equal-weighted reconciliation	−5.51	0.635 ***	≈ 0

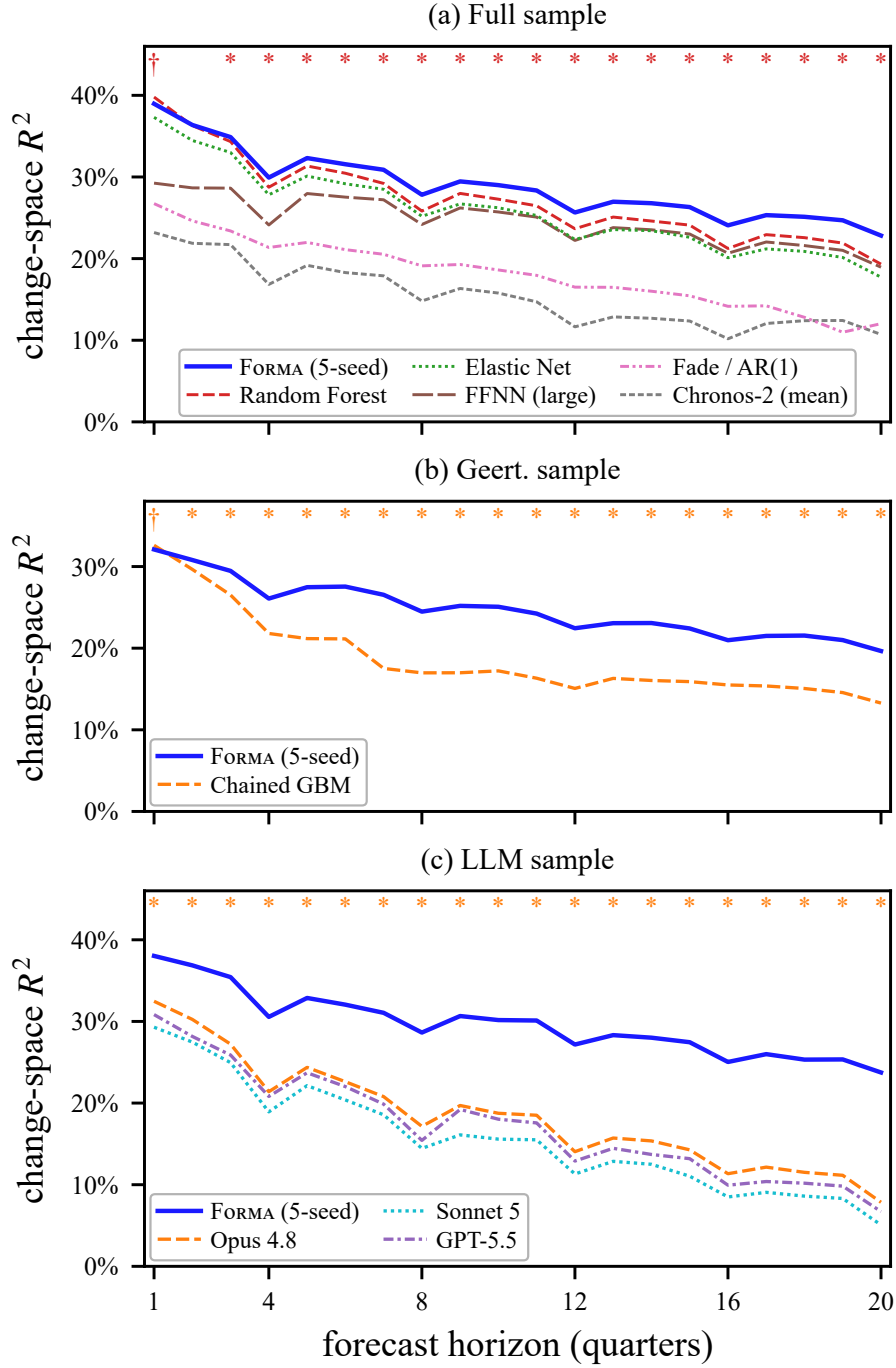


Figure 1: Change-space R^2 by forecast horizon on (a) the full sample (FORMA vs. five baselines), (b) the Geert. common sample (FORMA vs. the chained GBM), and (c) the LLM common sample (FORMA vs. three frontier LLMs). Markers atop each panel show quarter-clustered DM significance at the 5% level against RF, the GBM, and Opus 4.8 (the strongest LLM), respectively ($\dagger$ comparator better, $*$ FORMA better).

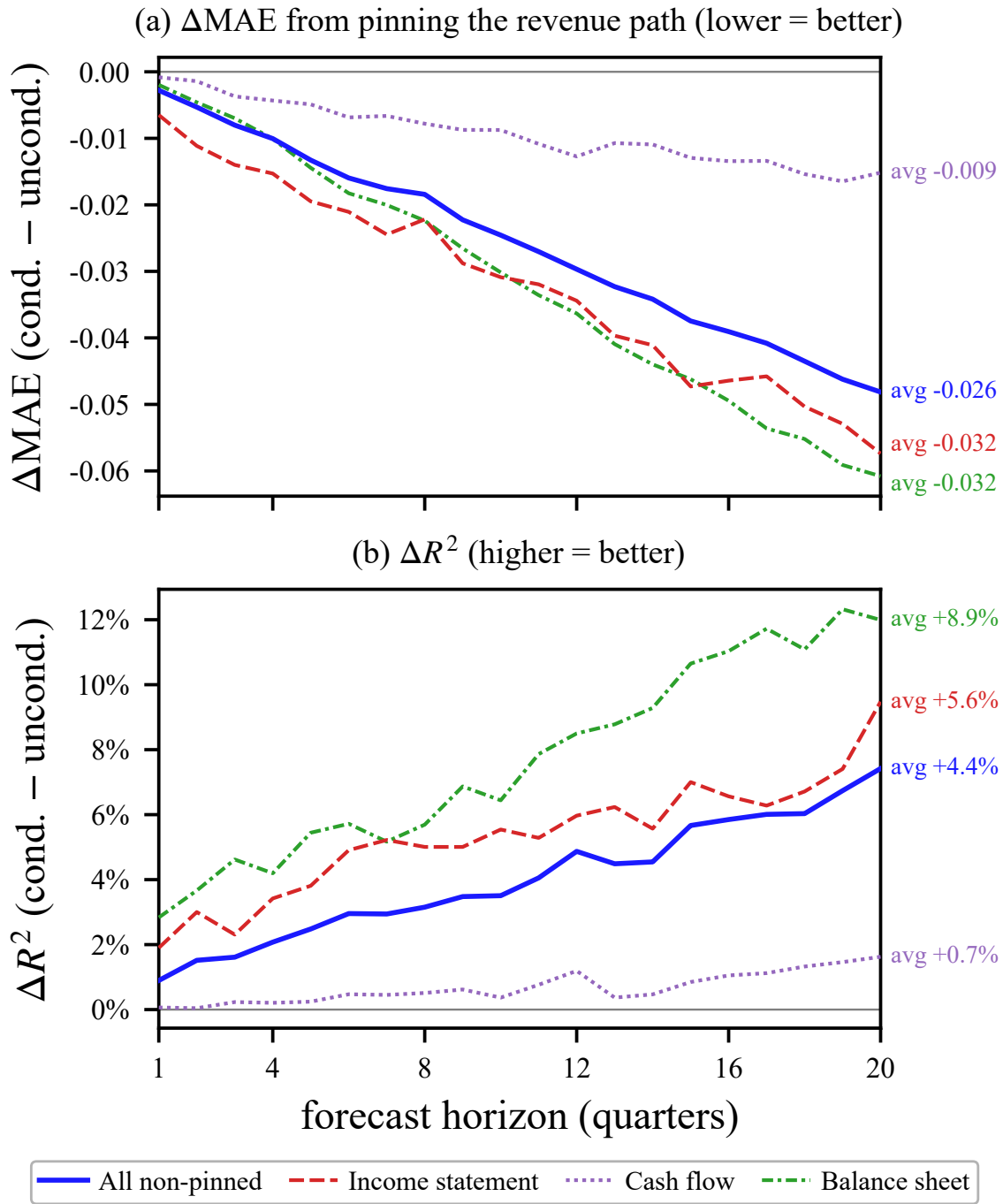


Figure 2: Scenario conditioning. Within-model change from pinning the true realized revenue path and forecasting the remaining statement, vs. the unconditional forecast on the same origins: (a) ΔMAE and (b) ΔR^2 by horizon, overall and by statement class.

Appendix A. Data

This appendix documents the data underlying the benchmark: the pipeline from the raw Compustat pull to the modeled firm-quarter origins (§A.1), the standardization that maps reported accounting values into model space (§A.2), the 78-item target universe (§A.3), its reporting availability (§A.4), and the accounting identities that link the items (§A.5).

Throughout, a *forecast origin* is a firm-quarter (f, t) from which a model forecasts quarters $t+1, \dots, t+20$ using the 12 quarters through the origin $(t-11, \dots, t)$. The benchmark provides the same panel in two views: the *tuple view*, in which each firm-origin is a variable-length set of (relative quarter, item, value) tuples, consumed by FORMA and the LLM protocol, and the *tabular view*, a fixed-width matrix of recent levels and year-over-year changes per item, consumed by the tabular baselines (§B.2).

A.1. Sample formation and splits

Table 4 gives the sample waterfall; the paragraphs below document each stage in order. The scale deflator that defines the final filter is specified in §A.2.

Table 4: Sample formation waterfall. Row 1 is the raw comp.fundq pull and each subsequent row applies one filter or transform to the row above. Row 1 is a WRDS re-query at a marginally later vintage than the downloaded panel, so the implied link-window trim is approximate (see text). The last waterfall row is the deflator-valid firm-quarter panel whose counts the main paper quotes (dropping the 203 deflator-valid firm-quarters dated 1970, which serve only as history); the rows below the rule partition it by split. Origin *eligibility* additionally requires four quarters of history through the origin (see text).

Stage	Firm-quarters
Raw Compustat quarterly rows (1970–2024 pull)	2,002,987
After CRSP link-window trim (see text)	1,708,447
Distinct firm-quarters (after deduplication)	1,706,140
After financial-sector exclusion (SIC 6000–6999)	1,531,223
With a valid scale deflator	1,173,801
Deflator-valid firm-quarters (1971–2024 window)	1,173,598
Train (1971–2001)	609,269
Validation (2002–2009)	211,367
Test (2010–2024)	352,962

Download and universe. The raw panel is Compustat Fundamentals Quarterly (comp.fundq) via WRDS, restricted to indfmt=INDL, datafmt=STD, consol=C, popsrc=D, with datadate in 1970–2024. Historical SIC codes (sich) come from comp.co_industry, attached to each firm-quarter as the most recent industry record at or before the statement date. The financial-sector filter drops firm-quarters with $sich \in [6000, 6999]$; firm-quarters with *missing* SIC are retained and map to the Unknown industry bucket.

The download also intersects the panel with the CRSP–Compustat link table (crsp.ccmxpf_1lnkhist, link types LU/LC, primary flags P/C): firms that appear in the link table retain only firm-quarters

whose statement date falls within a link window (dropping quarters before listing and after delisting), while the 16,418 firms with no link-table entry are retained in full (495,440 firm-quarters, 29% of the retained panel). The trim removes 294,540 firm-quarters, 15% of the raw pull, and leaves 26,917 link-matched firms alongside the 16,418 unmatched ones. Because WRDS updates comp.fundq and the link table in place, the raw count in Table 4 is a marginally later vintage than the downloaded panel; replaying the trim entirely on the re-queried vintage drops 296,589 firm-quarters, about 0.1% of the raw pull — 0.7% of the trim itself — away from the 294,540 implied here.

Deduplication and quarter alignment. Firm-quarters are indexed by *calendar* quarter: each row’s quarter is the calendar quarter-end of its datadate. When multiple rows share a (firm, calendar-quarter) pair (e.g. around fiscal-year changes), rows are sorted by fiscal year and quarter and the last record is kept. Year-to-date accumulation, by contrast, follows the *fiscal* calendar (next paragraph), so flow conversion is unaffected by fiscal/calendar offsets.

Year-to-date conversion. Compustat reports the 20 cash-flow items year-to-date. Each is converted to a quarterly flow within the firm–fiscal-year group: fiscal Q1 values pass through unchanged, and fiscal Q2–Q4 values are differenced against the immediately preceding fiscal quarter’s year-to-date value. If that preceding fiscal quarter is not in the file, the quarterly flow is left missing rather than imputed; no differencing ever crosses a fiscal-year boundary.

Derived items. The six derived items are computed after year-to-date conversion, exactly as defined in Table 5: gross profit ($gpq = revtq - cogsq$), free cash flow ($fcfq = oancfq - capxq$), working capital ($wcapq = actq - lctq$), and the three carve-outs aoq_ex_intanq , loq_ex_dr , $xsgaq_ex_rd$, in which the subtracted component ($intanq$, $dr1tq$, $xrdq$) is treated as zero when unreported.

Origin eligibility and splits. A firm-quarter is an eligible forecast origin if it has a valid deflator and at least four quarters of history through the origin. The history requirement is enforced per firm at load time, so a firm’s first three panel quarters cannot serve as origins; the tabular build instead drops the panel’s first three calendar quarters globally. The two rules differ only at firm entry, and scored comparisons are unaffected because every comparison is made on the common sample of cells all compared models predict. The counts in Table 4 are the deflator-valid firm-quarter panel *before* this final gate; eligible-origin counts are smaller and depend on a consumer’s additional requirements (e.g. the LLM protocol’s test pool of 321,892 origins, §C.1). Every input window spans the trailing 12 quarters, with unreported (item, quarter) cells contributing no tuple. Splits are by the calendar year of the origin quarter: 1971–2001 train, 2002–2009 validation, 2010–2024 test. Split boundaries also purge targets: a train (validation) example retains only target quarters dated on or before 2001Q4 (2009Q4) — later targets are set to missing in the tabular view and are absent from the split’s tuple file — so origins near a boundary contribute only their realized short horizons and no target crosses a split boundary. Hyperparameter and model-selection decisions for every model use the validation split only, whose targets end in 2009Q4; final models are refit on train+validation with all choices frozen (§B.1, §B.2), so nothing trained or selected before test conditions on an outcome after 2009Q4.

A.2. Standardization

Reported values span orders of magnitude across firms and are heavy-tailed within firm. The pipeline therefore maps every accounting value into a standardized space before it reaches any model, in four steps: deflation by firm size, an asinh transform, cross-sectional standardization, and clipping. We call the full mapping from an original value v to a model value x *standardization*, and its inverse *de-standardization*. All standardization parameters are fixed at the forecast origin. The same procedure serves the tuple view and the tabular view, and forecasts are evaluated in the standardized space.

Scale deflator. Let $v_{f,q,\text{id}}$ denote the reported value of item id for firm f at quarter q . The deflator of firm-quarter (f, q) is

$$z_{f,q} \equiv \begin{cases} |v_{f,q,\text{ltq}}| + |v_{f,q,\text{seqq}}| + 10^{-3} & \text{when valid,} \\ |v_{f,q,\text{atq}}| + 10^{-3} & \text{otherwise,} \end{cases}$$

i.e. total liabilities plus total stockholders' equity, which by the balance identity (§A.5) equals total assets less noncontrolling interests, up to sign conventions. The primary definition is valid when both components are reported and the sum is finite and positive; the total-assets fallback rescues 6,350 firm-quarters, and firm-quarters with no valid deflator under either definition are dropped from the panel (357,422 firm-quarters; Table 4). We call the surviving firm-quarters *scale-valid*. For an origin (f, t) , every model input and target — at every lead and lag in the window — is deflated by the origin deflator $z_{f,t}$; deflators at other quarters enter only the normalization statistics below.

Scaling and transform. For a firm f with forecast origin t , the scaled value of an accounting tuple (h, id, v) — item id reporting value v at relative quarter h , i.e. absolute quarter $t + h$ — is

$$v_{h,\text{id}}^{\text{scaled}}(f, t) \equiv \text{asinh}\left(k_{\text{id}} \frac{v}{z_{f,t}}\right),$$

so the entire example, leads and lags alike, is expressed on one origin-fixed scale. The account-specific constant k_{id} is estimated once on training-period observations (quarters through 2001Q4) of the lagged-ratio pool defined for σ below, and is then held fixed: it is the value on a 250-point logarithmic grid over $[10^{-2}, 10^3]$ that brings the excess kurtosis of the pooled transformed values as close as possible to 3 (total kurtosis 6, a mildly heavier-tailed target than the Gaussian).

Cross-sectional standardization. Scaled values are standardized per (item, quarter) with statistics built from trailing cross-sections. Let $\mathcal{F}$ denote the cross-sectional pool of firms. For each calendar quarter q , define the raw per-quarter statistics

$$\mu_{\text{id},q}^{\text{raw}} \equiv \text{mean}_{f' \in \mathcal{F}} \text{asinh}\left(k_{\text{id}} \frac{v_{f',q,\text{id}}}{z_{f',q}}\right), \quad \sigma_{\text{id},q}^{\text{raw}} \equiv \text{std}_{f' \in \mathcal{F}, r \in \{0,-8\}} \text{asinh}\left(k_{\text{id}} \frac{v_{f',q+r,\text{id}}}{z_{f',q-4}}\right).$$

The mean is the contemporaneous cross-section of item id at quarter q , each observation deflated by its own quarter’s scale. The standard deviation is computed over the pooled values at quarters q and $q - 8$, both deflated by the scale of the intermediate quarter $q - 4$ (the lagged-ratio pool). In both constructions a lag indexes positions in the firm’s reported series for the item, so reporting gaps collapse. The origin-quarter parameters average the four most recent raw statistics (fewer at the panel start),

$$\mu_{\text{id},t} \equiv \frac{1}{4} \sum_{q=t-3}^t \mu_{\text{id},q}^{\text{raw}}, \quad \sigma_{\text{id},t} \equiv \frac{1}{4} \sum_{q=t-3}^t \sigma_{\text{id},q}^{\text{raw}},$$

and the normalized value is

$$v_{h,\text{id}}^{\text{normalized}}(f,t) \equiv \frac{v_{h,\text{id}}^{\text{scaled}}(f,t) - \mu_{\text{id},t}}{\sigma_{\text{id},t} + 10^{-8}},$$

with 10^{-8} a numerical floor. All means and standard deviations use available reported observations only. Every lag in the construction is non-negative and the four-quarter averages are trailing, so $(\mu_{\text{id},t}, \sigma_{\text{id},t})$ depend only on data through quarter t and are updated each quarter; the per-item constants k_{id} are the only standardization parameters frozen at the 2001Q4 estimation cutoff.

Clipping. With $\text{clip}(u, a, b) \equiv \min\{b, \max\{a, u\}\}$, the model value of a tuple (h, id, v) is

$$x_{f,t}(h, \text{id}) \equiv \text{standardize}_{f,t}(h, \text{id}, v) \equiv \text{clip}\left(v_{h,\text{id}}^{\text{normalized}}(f,t), -6, 6\right).$$

The same clipped value serves as a model input when the tuple is visible and as the supervised target when it is masked. In the tuple view the clip is applied at load time; in the tabular view it is applied at build time, to levels and targets at 6 and to year-over-year difference features, which are computed from the unclipped standardized levels, at $6\sqrt{2}$ (the ± 6 bound scaled by $\sqrt{2}$, the standard deviation of the difference of two independent unit-variance values). Because clipping is not invertible, an original value whose standardized value lies at either boundary cannot generally be recovered uniquely.

The scale token. In the tuple view the deflator itself enters as one additional input tuple per origin, under a dedicated scale account at $h = 0$, so models observe firm size. Its value is $z_{f,t}$, standardized like any other account through an asinh transform with its own constant k_{scale} and per-quarter statistics $(\mu_{\text{scale},t}, \sigma_{\text{scale},t})$ estimated on the cross-section of contemporaneous deflator values; the deflator is not divided by itself. The scale account is never queried at future horizons and receives no future-grid slots (§B.1); like any historical tuple, it can be masked for reconstruction during training. In the tabular view the deflator enters as one feature on the same standardized basis.

De-standardization. Given a predicted model value $\hat{x}_{f,t}(h, \text{id})$, the prediction in original accounting units is

$$\hat{v}_{f,t+h,\text{id}} \equiv z_{f,t} \frac{\sinh(\hat{x}_{f,t}(h, \text{id}) \sigma_{\text{id},t} + \mu_{\text{id},t})}{k_{\text{id}}}$$

(numerical floors of order 10^{-8} omitted). De-standardization uses the deflator and normalization parameters fixed at the forecast origin t , never the realization quarter $t + h$. The map is strictly increasing in $\hat{x}$, so it carries quantiles of a predictive distribution in model space to quantiles in accounting units, a property the Chronos-2 comparison relies on (§B.2).

A.3. The ProForma-20Q variable universe

ProForma-20Q targets the 78 quarterly Compustat line items in Table 5: 34 balance-sheet, 18 income-statement, 20 cash-flow, and 6 derived items, jointly at horizons $h = 1, \dots, 20$. The same items serve as both conditioning inputs (reported history) and prediction targets (future quarters); ProForma-20Q additionally conditions on a Fama–French-48 industry category. Cash-flow items are converted from year-to-date to quarterly flows (§A.1); derived items are exact functions of raw items (definitions in the table). Mnemonics are Compustat quarterly field names.

Table 5: The 78-item ProForma-20Q target universe, ordered alphabetically by Compustat mnemonic: description, statement (St: BS balance sheet, IS income statement, CF cash flow, Der derived), and reporting availability. % raw is the fraction of all 1,706,140 Compustat quarterly firm-quarters (all sectors and years, no scale-validity requirement) that report the item after year-to-date→quarterly conversion, i.e. the reporting rate before any sample filter. The split columns are that same fraction restricted to the scale-valid firm-quarters of each split: train 1971–2001 ($N=609,269$), validation 2002–2009 ($N=211,367$), test 2010–2024 ($N=352,962$). Items near-zero in the training era (e.g. stkcoq, txbcfq, drcq, acomincq) are modern disclosure fields absent from early filings; FORMA represents them as missing rather than imputing them.

Mnemonic	Description	St	% raw	% Tr	% Val	% Te
acomincq	Accumulated other comprehensive income	BS	38.7	1.5	92.3	98.0
acoq	Current assets, other	BS	72.2	92.1	96.6	99.7
actq	Current assets, total	BS	65.5	89.6	94.0	97.5
ancq	Non-current assets, total	BS	62.6	85.0	90.3	93.8
aoq	Assets, other	BS	78.1	98.4	99.8	99.9
aoq_ex_intanq	Other assets excl. intangibles	Der	78.1	98.4	99.8	99.9
apq	Accounts payable	BS	76.9	96.8	98.9	99.3
aqcq	Acquisitions	CF	61.9	64.8	90.2	91.2
atq	Assets, total	BS	78.8	100.0	100.0	100.0
capsq	Capital surplus / APIC	BS	75.8	93.6	95.1	94.9
capxq	Capital expenditures	CF	62.7	65.1	91.9	93.7
ceqq	Common equity, total	BS	79.8	98.8	99.5	99.8
cheq	Cash & short-term investments	BS	77.8	97.8	99.8	99.9
cogsq	Cost of goods sold	IS	79.7	97.5	98.9	99.5
cstkq	Common stock	BS	76.7	95.0	95.3	96.2
dlcchq	Change in current debt	CF	34.0	32.6	51.7	54.9
dlcq	Debt in current liabilities	BS	75.2	93.8	97.4	97.5
dltisq	Long-term debt issuance	CF	60.6	62.7	87.4	91.8
dltrq	Long-term debt reduction	CF	61.1	62.9	88.9	92.3

continued on next page

Table 5, continued

Mnemonic	Description	St	% raw	% Tr	% Val	% Te
dlttq	Long-term debt	BS	79.5	98.6	99.2	99.3
dpactq	Accumulated depreciation	BS	51.1	73.2	57.9	69.6
dpq	Depreciation & amortization	IS	69.5	81.9	92.8	96.2
drcq	Deferred revenue, current	BS	29.5	0.1	67.4	89.3
drltq	Deferred revenue, long-term	BS	31.3	0.1	69.9	92.6
dvq	Cash dividends	CF	63.1	66.1	91.4	93.4
exreq	Exchange rate effect on cash	CF	57.0	49.1	92.3	94.1
fcfq	Free cash flow = oancfq – capxq	Der	56.3	48.0	91.9	93.7
fiaoq	Financing activities, other	CF	56.9	49.1	92.1	93.8
fincfq	Financing cash flow	CF	57.1	49.3	92.5	94.3
fopoq	Funds from operations, other	CF	60.1	59.7	89.7	93.0
gdwlq	Goodwill	BS	39.3	4.2	90.9	97.1
gpq	Gross profit = revtq – cogsq	Der	74.0	92.0	95.7	99.2
ibq	Income before extraordinary items	IS	82.6	99.6	99.5	99.7
intanoq	Intangibles, other	BS	37.8	4.3	87.4	92.6
intanq	Intangibles, total	BS	43.0	8.9	99.0	99.5
invtq	Inventory	BS	75.9	95.0	97.7	98.1
ivacoq	Investing activities, other	CF	56.9	49.1	92.2	93.9
ivchq	Increase in investments	CF	60.7	63.3	88.2	90.6
ivncfq	Investing cash flow	CF	57.1	49.3	92.5	94.3
ivstchq	Short-term investments, change	CF	45.7	45.0	77.1	73.5
lcoq	Current liabilities, other	BS	72.0	91.5	96.4	99.6
lctq	Current liabilities, total	BS	66.0	90.2	94.1	97.6
loq	Liabilities, other	BS	78.1	98.3	99.8	99.9
loq_ex_dr	Other liab. excl. deferred rev.	Der	78.1	98.3	99.8	99.9
ltq	Liabilities, total	BS	78.4	99.1	99.9	99.9
mibtq	Noncontrolling interests (BS)	BS	74.5	91.5	93.7	98.1
miiq	Noncontrolling interest (income)	IS	76.2	88.9	91.8	96.9
niq	Net income	IS	82.6	99.6	99.4	99.7
nopiq	Non-operating income	IS	79.9	98.2	98.9	99.5
oancfq	Operating cash flow	CF	57.1	49.3	92.5	94.3
oiadpq	Operating income after D&A (EBIT)	IS	79.0	96.9	98.4	99.2
oibdpq	Operating income before D&A (EBITDA)	IS	70.6	83.3	93.2	96.1
piq	Pretax income	IS	81.8	98.5	99.3	99.7
ppegqtq	Gross PP&E	BS	51.0	73.2	57.9	69.6
ppentq	Net PP&E	BS	76.6	99.2	99.5	99.6
prstkcoq	Purchase of common & preferred stock	CF	59.9	62.4	85.2	89.5
pstkq	Preferred stock	BS	79.8	98.8	99.4	99.7
rectq	Receivables, total	BS	75.4	94.3	97.6	97.4
req	Retained earnings	BS	76.3	94.5	95.7	95.3
revtq	Revenue, total	IS	76.8	93.8	96.1	99.4
seqq	Stockholders' equity, total	BS	80.3	99.9	99.9	99.9
sivq	Sale of investments	CF	60.8	63.2	88.6	91.0
spiq	Special items	IS	77.1	89.8	98.8	98.3
sppeq	Sale of PP&E	CF	52.0	52.3	76.0	79.8
sstkq	Sale of common & preferred stock	CF	62.3	64.9	90.7	92.4
stkcoq	Stock compensation expense	IS	30.1	0.1	62.5	83.9
tstkq	Treasury stock	BS	74.5	86.5	98.9	99.1
txbcofq	Excess tax benefit, stock options	CF	30.8	0.0	49.0	93.5
txditcq	Deferred taxes & investment tax credit	BS	68.5	87.8	91.3	93.4

continued on next page

Table 5, continued

Mnemonic	Description	St	% raw	% Tr	% Val	% Te
txpq	Income taxes payable	BS	66.5	84.1	88.2	92.7
txtq	Income taxes, total	IS	81.9	98.6	99.3	99.7
wcapq	Working capital = actq – lctq	Der	65.4	89.4	93.9	97.5
xidoq	Extraordinary items & disc. ops	IS	82.6	99.5	99.4	99.7
xintq	Interest expense	IS	65.8	79.0	81.4	88.6
xoprq	Operating expenses, total	IS	79.5	97.2	98.8	99.3
xrdq	R&D expense	IS	24.2	22.3	41.4	43.6
xsgaq	SG&A expense	IS	62.6	77.9	83.4	85.2
xsgaq_ex_rd	SG&A excl. R&D	Der	62.6	77.9	83.4	85.2

A.4. Availability and sparsity

Reporting is pervasively incomplete. Table 6 quantifies the joint availability of the 78-item vector across scale-valid firm-quarters. On the test panel the item×firm-quarter matrix is 93.6% dense and the average firm-quarter reports 73 of 78 items, yet only 4.7% report all 78, so complete-case (listwise) deletion, the implicit requirement of a dense tabular design, would discard 95.3% of the test panel. FORMA’s tuple-set representation accommodates this incompleteness natively; the tabular baselines require an explicit imputation model.

Table 6: Joint availability of the 78-item vector across scale-valid firm-quarters. *Matrix density* is the fraction of item×firm-quarter cells reported; *items per firm-quarter* summarizes how many of the 78 items a firm reports in a given quarter; *complete-case* firm-quarters report all 78.

	Full sample 1971–2024	Test 2010–2024
Scale-valid firm-quarters	1,173,598	352,962
Distinct firms	32,851	13,288
Matrix density (%)	82.2	93.6
Items per firm-quarter, mean	64.1	73.0
median [p10, p90]	68 [46, 76]	75 [68, 77]
With ≥ 40 items (%)	95.1	99.7
With ≥ 70 items (%)	41.4	87.1
Complete-case (all 78, %)	1.73	4.67
<i>Cell density by statement (%), full / test</i>		
Balance sheet	85.4	95.8
Income statement	87.8	93.6
Cash flow	69.7	89.3
Derived	89.4	95.9

A.5. Accounting identities

The 26 exact linear identities below tie the target items. They define the identity groups used by FORMA’s grouped masking objective and the coherence residuals analyzed in the main paper. Each holds in levels at every quarter; contemporaneous subscripts are suppressed. Signs are normalized so all terms are additive (negative-coefficient terms moved to the right-hand side).

Balance sheet (13).

$$\begin{aligned}
&\text{cheq} + \text{invtq} + \text{rectq} + \text{acoq} + \text{ppentq} + \text{aoq} = \text{atq} \\
&\quad \text{actq} + \text{ancq} = \text{atq} \\
&\text{cheq} + \text{invtq} + \text{rectq} + \text{acoq} = \text{actq} \\
&\quad \text{wcapq} + \text{lctq} = \text{actq} \\
&\text{intanq} + \text{aoq_ex_intanq} = \text{aoq} \\
&\quad \text{gdwlq} + \text{intanoq} = \text{intanq} \\
&\quad \text{ppentq} + \text{dpactq} = \text{ppegqtq} \\
&\text{apq} + \text{lcoq} + \text{dlcq} + \text{txpq} + \text{dlttq} + \text{txditcq} + \text{loq} = \text{ltq} \\
&\quad \text{apq} + \text{dlcq} + \text{txpq} + \text{lcoq} = \text{lctq} \\
&\quad \text{drltq} + \text{loq_ex_dr} = \text{loq} \\
&\quad \text{cstkq} + \text{capsq} + \text{req} = \text{tstkq} + \text{ceqq} \\
&\quad \text{seqq} = \text{pstkq} + \text{ceqq} \\
&\quad \text{atq} = \text{ltq} + \text{mibtq} + \text{seqq}
\end{aligned}$$

Income statement (10).

$$\begin{aligned}
&\text{gpq} + \text{cogsq} = \text{revtq} \\
&\text{oibdpq} + \text{xsgaq} = \text{gpq} \\
&\text{xrdq} + \text{xsgaq_ex_rd} = \text{xsgaq} \\
&\quad \text{xoprq} = \text{cogsq} + \text{xsgaq} \\
&\text{oiadpq} + \text{dpq} = \text{oibdpq} \\
&\quad \text{piq} + \text{xintq} = \text{oiadpq} + \text{nopiq} + \text{spiq} \\
&\text{ibq} + \text{txtq} + \text{miiq} = \text{piq} \\
&\quad \text{niq} = \text{ibq} + \text{xidoq} \\
&\text{niq} + \text{txtq} + \text{miiq} = \text{piq} + \text{xidoq} \\
&\text{revtq} + \text{nopiq} + \text{spiq} + \text{xidoq} = \text{cogsq} + \text{xsgaq} + \text{dpq} + \text{xintq} + \text{txtq} + \text{miiq} + \text{niq}
\end{aligned}$$

Cash flow (3).

$$\begin{aligned}
&\text{sivq} + \text{sppeq} + \text{ivstchq} + \text{ivacoq} = \text{capxq} + \text{ivchq} + \text{aqcq} + \text{ivncfq} \\
&\text{sstkq} + \text{dltisq} + \text{dlcchq} + \text{fiaoq} + \text{txbcofq} = \text{prstkq} + \text{dltrq} + \text{dvq} + \text{fincfq} \\
&\quad \text{fcfq} + \text{capxq} = \text{oancfq}
\end{aligned}$$

Appendix B. Models

The main paper specifies FORMA’s representation and objective. This appendix records the implementation detail needed to retrain it from scratch (§B.1) and the exact specification of every competitor (§B.2). The released configuration files are the authoritative source for all values quoted here. Throughout, the *headline table* is the main paper’s headline accuracy table, which scores three tracks on standardized values: a squared-error track summarized by change-space R^2 (the R^2 of predicted against realized changes $x_{t+h} - x_t$, so zero corresponds to predicting the sample-average change), an absolute-error track (MAE), and a density track (log score, CRPS, and coverage) for models that produce predictive distributions.

B.1. FORMA: training and implementation

Table 7 collects the canonical hyperparameters.

Table 7: FORMA canonical hyperparameters (identical across the five mixture seeds; only the seed differs).

<i>Architecture</i>	
Encoder layers / d_{model} / heads	4 / 128 / 4
Feed-forward width / dropout	512 / 0.2
Trainable parameters (incl. variance head)	942,210
<i>Optimization</i>	
Optimizer	AdamW (default β , ϵ)
Learning rate	10^{-4} , constant
Weight decay	0.1 (all parameters)
Gradient-norm clip	1.0
Batch size	32 firm-origin sets
Epochs	12 (final-epoch weights used)
Precision	fp32 (medium matmul precision)
Seeds	60–64 (equal-weight mixture)
<i>Loss</i>	
Objective	Gaussian β -NLL, $\beta = 0.5$
$\log \sigma^2$ clamp	$[-10, 10]$
Absolute-error-track variant	Laplace β -NLL, $b = \sigma/\sqrt{2}$
<i>Horizon curriculum</i>	
Initial $\rightarrow$ max horizon	4 $\rightarrow$ 20 quarters
Step	+4 per epoch after epoch 2
Lookback	12 quarters, fixed
<i>Masking probabilities (per training example)</i>	
Mask entire future block	0.5
else mask each future tuple	0.95
Select historical identity instance	0.1 (mask $k \sim \text{U}\{2, \dots, N\}$ of its N members)
Mask ungrouped historical tuple	0.05

Embeddings. All embedding components share $d_{\text{model}} = 128$: a learned account-id embedding over the build’s account vocabulary, a learned industry embedding over the 48 Fama–French industries plus an Unknown bucket, a fixed (parameter-free) sinusoidal encoding of the quarter offset relative to the origin, a learned value direction w_x that scales with the standardized value, and a learned mask vector substituted for the value component of hidden tuples. The industry token is never masked and carries no prediction target.

Output heads. The final representation of each queried tuple is concatenated with its sinusoidal horizon encoding (width $2d_{\text{model}} = 256$) and passed through two parallel two-layer MLPs ($256 \rightarrow 256 \rightarrow 1$, GELU): one for the mean μ and one for the log-variance $\log \sigma^2$. The log-variance is clamped to $[-10, 10]$ before exponentiation; there is no other floor on σ .

Loss. With $\beta = 0.5$ the per-tuple Gaussian loss is

$$\ell = \left[\frac{(y-\mu)^2}{2\sigma^2} + \frac{1}{2} \log \sigma^2 \right] \cdot [\sigma^2]_{\text{sg}}^\beta,$$

where y is the standardized target value of §A.2, $[\cdot]_{\text{sg}}$ denotes stop-gradient [34], and $\beta = 0$ recovers the standard NLL. The loss averages over tuples that are both masked and observed: synthetic query tokens without realized values (below) and the industry token never enter the loss. The absolute-error-track variant is a separately trained five-seed family, identical except that the head is read as a Laplace scale $b = \sigma/\sqrt{2}$ and trained with the Laplace β -NLL.

Masking procedure. For each training example: (i) with probability 0.5 the entire future block is masked (pure forecasting); otherwise each future tuple is masked independently with probability 0.95, revealing $\approx 5\%$ of realized future values as inputs (the mechanism that lets a user pin assumed future values and condition the remaining forecasts on them; see the main paper’s scenario analysis); (ii) each complete historical identity instance (all member accounts reported in a quarter) is selected with probability 0.1, and a selected instance masks k members with k drawn uniformly from $\{2, \dots, N\}$; k is never exactly one, so no masked value is recoverable from its own identity; (iii) historical tuples of *accounts* that participate in no complete historical identity instance in the window are masked independently with probability 0.05. An account appearing in a complete instance in any historical quarter is exempt from this singleton masking at all its quarters.

Curriculum. Training runs 12 epochs with the maximum queried horizon following the schedule 4, 4, 8, 12, 16, 20, ..., 20 (the first expansion comes after the second epoch); the training loader is rebuilt each epoch at the current horizon. The canonical run trains on the merged train and validation panel (1971–2009), like the refit tabular competitors of §B.2: all hyperparameters were fixed beforehand from validation-split experiments, no validation pass runs during training, and final-epoch weights are used. The 12-quarter lookback is fixed for every example; the 4-quarter minimum history only gates which firm-quarters are eligible origins.

Batching. One example is the full tuple set of one firm-origin. No maximum sequence length is imposed and no example is truncated; sets are batched 32 at a time and padded to the longest

set in the batch with a boolean key-padding attention mask. The average origin carries ≈ 758 observed historical tuples and $\approx 1,124$ realized future targets; the model’s token set additionally includes the synthetic query tokens for unrealized (account, horizon) cells (below), the scale token of §A.2, and one industry token.

Ex-ante future grid. Future queries are built ex ante: for each origin, a synthetic query tuple is created for every (account, horizon) pair with the account drawn from the set the firm reports *at the origin quarter* and the horizon running $1, \dots, 20$, truncated only where the horizon extends past the split’s last calendar quarter. That cut is identical for every firm and does not depend on the firm’s own survival. Queries whose realization is later observed are loss-eligible targets; queries never realized (e.g. the firm exits) remain in the set as masked tokens but are excluded from the loss and masked at scoring time.

Inference. At test time all future tuples are masked and a single forward pass yields (μ, σ) for every queried (item, horizon) pair jointly, with no iterated roll-forward. Forecasts are produced in fp32; $\sigma = \exp(\frac{1}{2} \log \sigma^2)$ is saved alongside μ . No reconciliation is applied at inference; the ex-post projections studied in the main paper are computed downstream from the saved forecasts.

Mixture. The five seeds are trained independently and combined as an equal-weight mixture. Distributional metrics use the exact mixture: log-scores from the mixture density and CRPS from closed-form mixture expressions.

Compute. Each seed trains in ≈ 16.5 hours on a single NVIDIA A100-SXM GPU; the five-seed run cost $\approx \$120$ of on-demand cloud compute. A trained checkpoint is ≈ 11 MB.

B.2. Competitor specifications

All learned tabular competitors share one protocol. They consume the tabular view: for each of the 78 items, four standardized recent levels and eight standardized year-over-year changes, plus the standardized deflator (§A.2) and 48 industry dummies, with the Unknown bucket as the omitted reference (985 features in all). Missing features are imputed with the cross-sectional latent-factor model of Bryzgalova et al. [5] with 10 factors, fit one calendar quarter at a time and filling only unobserved cells. Hyperparameters are selected on the 2002–2009 validation split (never test) and final models are refit on train+validation.

Elastic net. One sklearn elastic net per (item, horizon) pair (1,560 fits; rows with any missing feature or target are dropped per fit). Grid: $\alpha \in \{0.01, 0.1, 1, 10, 100, 1000\} \times \ell_1\text{-ratio} \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$, selected by validation MSE per horizon on a reference item (net income) and shared across items at that horizon; coordinate descent, max 10,000 iterations.

Random forest. One GPU (cuML) regression forest per (item, horizon) pair. Fixed parameters: 25 minimum samples per leaf, 50% bootstrap per tree. Grid (same selection protocol as the elastic net): $\text{trees} \in \{50, 100, 200\}$, $\text{max_depth} \in \{5, 10, 20\}$, $\text{min_samples_split} \in \{100, 200\}$. The full 1,560-model fit runs on the order of 40 hours on one GPU.

Feed-forward networks. Each FFNN is a single multi-output network mapping the 985 features to two 1,560-wide heads (a mean and a log-variance per (item, horizon)), trained with the same masked Gaussian β -NLL ($\beta = 0.5$, $\log \sigma^2$ clamped to $[-10, 10]$) as FORMA; missing targets contribute zero loss rather than dropping the row, and rows are dropped only for residual missing features (rare after imputation). The *linear* variant has no hidden layer (3,076,320 parameters); the *large* variant has GELU hidden layers of widths 1024/1024/512 with dropout 0.1 (4,184,624 parameters), 3.3 and 4.4 times FORMA’s 942,210 respectively. AdamW, learning rate 10^{-4} , no weight decay, batch 1024; trained for a fixed 6 epochs (the best-validation epoch from a held-out selection run) on train+validation. Like FORMA, each variant is a five-seed (60–64) equal-weight mixture with saved σ , so both enter the density track.

Chained GBM. Our re-implementation of the chained gradient-boosting approach of Geertsema et al. [15] under the common protocol. LightGBM regressors are organized in a 14-step chain over 28 of the original specification’s 29 core items (iva is dropped as scarcely populated in the quarterly file); the chain runs revenue first, then costs and working capital, through the capital-structure and tax blocks. At each step the feature matrix is augmented with the chain’s upstream items: realized values during training (teacher forcing), predicted values at inference (cascade), with unreported upstream values entering as zero. Per item: up to 500 trees with early stopping (patience 10) on the validation split, leaves $\in \{15, 31\}$ selected on validation, refit on train+validation at the selected size. Two arms are trained, an L_1 objective (the original specification, absolute-error track) and an L_2 objective (squared-error track). Of the 28 chained items, 25 lie in the 78-item ProForma-20Q universe; the headline table’s Geert. column is scored on this 25-item footprint.

Pooled fade/AR(1). For each (item, horizon), one pooled OLS regression of the standardized future value on the standardized origin value, $\hat{x}_{t+h} = \hat{\alpha}_{id,h} + \hat{\rho}_{id,h} x_t$, fit on train+validation (direct per-horizon regressions, never iterated), requiring at least 100 observations.

Seasonal random walk. $\hat{x}_{t+h} = x_{t-((4-h) \bmod 4)}$: the forecast repeats the most recent observation falling in the same fiscal quarter, so the base sits 4, 8, 12, 16, or 20 quarters before the target, and at $h \in \{4, 8, 12, 16, 20\}$ it coincides with the no-change forecast $\hat{x}_{t+h} = x_t$. Quarterly accounting series are seasonally differenced [14], and the seasonal random walk is the standard naive expectation against which the quarterly-earnings literature measures skill [3].

Chronos-2. The public amazon/chronos-2 checkpoint, zero-shot (no fine-tuning; it applies only its own inference-time normalization). For each origin the model receives the firm’s raw (un-deflated) reported values on a contiguous quarter grid, up to 12 observations through the origin with gaps entered as missing, which Chronos-2 handles natively, and forecasts all items jointly via its multivariate group attention. Its 21 native quantile levels $\{0.01, 0.05, 0.10, \dots, 0.90, 0.95, 0.99\}$ are mapped into benchmark space through the origin-frozen monotone transform of §A.2. The squared-error track uses the conditional mean obtained by trapezoidal integration over the mapped quantile grid (flat tails beyond the 0.01/0.99 levels); the absolute-error track uses the mapped median; the density track scores Chronos-2 directly from the same mapped quantile grid, CRPS by the exact piecewise-linear quantile expression and interval coverage from the matching quantile

pairs. No log score is reported: on flat contexts ($\approx 0.6\%$ of cells) the predicted quantiles collapse to zero-width intervals that admit no usable density even under a scale floor.

LLM panel. The frontier-LLM protocol and detailed prompts are documented in Appendix C.

Appendix C. LLM Benchmark

The LLM panel is evaluated on the same information set as FORMA and scored on the same grid, so differences across models reflect the forecaster rather than the information set. §C.1 documents the protocol; §C.2 reproduces the elicitation prompts.

C.1. Protocol

Inputs (tuple-sourced). History and origins are read directly from FORMA’s stored test tuples, with no re-derivation of any input FORMA receives. Each prompt carries a 12-quarter reported history in original units, millions of dollars ($Q_{-11} \dots Q_0$, FORMA’s `max_lookback`), with Compustat mnemonic labels and *relative* quarter indices, plus the firm’s Fama–French-48 industry rendered as a plain-English sector name. The industry label gives the LLMs the same conditioning as FORMA’s industry token and the baselines’ dummies, and each of the 48 buckets is shared by hundreds of firms. The mnemonics are labels, not assumed knowledge: the system prompt carries a glossary giving every item’s plain-English description and sign convention, plus the formulas of the derived subtotals shown as context (§C.2). No firm identifier and no absolute dates enter the prompt. Returned forecasts are standardized with the origin-frozen parameters of §A.2 before scoring.

Origins and sampler. An eligible origin (firm, q_0) has q_0 in the test window 2010Q1–2024Q4, a valid scale at q_0 , at least four quarters of history, and at least one future quarter to score; the eligible pool is 321,892 origins across 12,230 firms (the deflator-valid test firm-quarters of §A.1 that also pass these gates). For each sampled firm the sampler draws one variable-length block of *consecutive* eligible origins (length 4–20), calendar-balanced across 2010–2024; consecutive origins produce the overlapping-target-quarter panels that a companion analysis of forecast revisions requires. One shared, seeded origins file is reused by every model to maximize pairwise overlap. The scored run is 2,103 pinned origins across 133 firms, run in three nested stages (the first 20, then 220, then all 2,103 origins of the shared file); the two prompt arms are paired on the identical origin set.

Targets. The model forecasts the 55 primitive line items (only the firm’s reported subset; structurally absent items are excluded from its list and treated as zero when reconstructing derived items). The remaining 23 of the 78 targets are reconstructed mechanically from the primitives through the accounting identities of §A.5: statement subtotals such as `atq`, `ltq`, `seqq`, and `niq`. Eighteen of the 23 appear in the prompt’s glossary as labeled derived context; the remaining five (`ancq`, `wcapq`, and the three carve-outs of §A.1) are reconstructed downstream and never shown to the model. Every model is therefore scored on the same $78\text{-target} \times Q_1\text{--}Q_{20}$ grid as FORMA. Every origin is asked for the full $Q_1\text{--}Q_{20}$ horizon regardless of how many future quarters the firm actually has in file; unavailable horizons are masked only at scoring time, so the prompt never leaks the firm’s data-end or delisting.

Two elicitation arms. Both arms share the same inputs, the same 55-primitive schema, and the same strict primitives-only output format; they differ only in one system-prompt block. The *unstructured* arm pairs the line-item glossary and accounting-consistency checks with a free-form steering paragraph that tells the model to reason about trajectory, growth, seasonality,

mean-reversion, the identities, and the drivers, without prescribing a method. The *structured* arm replaces that paragraph with an explicit driver-hierarchy / roll-forward method (revenue year-over-year off the same quarter a year earlier, cost and expense lines as ratios to revenue, working capital via turnover ratios, PP&E by roll-forward, then cash and retained earnings *set* by the cash-flow and retained-earnings rolls), using the identities as the generative construction order. The comparison therefore carries no parse-format risk. The headline table reports each model’s better arm, the unstructured arm for all three models. Unstructured/structured change-space R^2 is 0.186/0.175 for Opus 4.8, 0.174/0.171 for GPT-5.5, and 0.158/0.139 for Sonnet 5; MAE is 0.362/0.367, 0.363/0.368, and 0.368/0.376.

Models and settings. The panel comprises three frontier models, Claude Opus 4.8, GPT-5.5, and Claude Sonnet 5, all with extended reasoning enabled. The Anthropic models run at their adaptive default (bounded, measured at $\approx 4\text{--}6\text{k}$ thinking tokens for Opus and $\approx 12\text{--}17\text{k}$ for Sonnet 5); GPT-5.5 runs at medium reasoning effort. Reasoning is therefore not token-matched across vendors (the batch gateway exposes no common reasoning-effort setting), a run property to bear in mind when comparing models. All calls use default temperature and `max_tokens=64,000` (headroom for heavy reasoning on the structured prompt; billed only when consumed) and route through the Portkey batch gateway: the Anthropic models via the Messages batch endpoint, and GPT-5.5’s scored run via the OpenAI Responses endpoint to capture reasoning summaries ($\approx 82\%$ of origins carry one). The main run is a single call per origin.

Decoding-noise subset. A 200-origin simple-random subset (drawn from the main sample) is queried $K=5$ times on every model and both arms at default temperature to measure across-call dispersion; the main run’s single pass serves as draw 1, so the extra cost is four replicate runs per cell. The subset supports the companion analysis of forecast revisions and does not enter the headline results.

Common sample. The LLM column of the headline table is an exact common sample: a (firm, origin, item, horizon) cell counts only where all seven forecasters (FORMA and the six LLM runs, two prompt arms $\times$ three models) predict it finitely and a realized value is available to score against. That intersection is 2,152,314 cells, of a possible $78 \times 20 \times 2,103 = 3,280,680$ (targets $\times$ horizons $\times$ origins), a 66% subset; the full model suite is then re-scored inside that footprint. Parse failures are essentially absent (every model returned forecasts on $\approx 100\%$ of requested cells), so exclusions come from cells with no realized value to score or outside a model’s covered item set at that origin.

C.2. Exact prompts

Prompts are reproduced from the released prompt files with Unicode box-drawing and math glyphs transliterated to ASCII for typesetting; the byte-exact originals are included in the release. The `{lookback_qs}` / `{lookback_first}` / `{max_horizon}` tokens are substituted from config at load time and render as 12 / 11 / 20 for the production run.

C.2.1. System prompt, unstructured arm, in full

You are a financial forecasting model. Given `{lookback_qs}` quarters of historical financial data (Q-`{lookback_first}` through Q0) for one company, forecast forward quarters Q1 through Q`{max_horizon}`.

Forecasting these line items accurately is a hard, multi-step reasoning problem. Before committing to any numbers, reason carefully and thoroughly: study the firm's recent trajectory and growth rates, seasonality, mean-reversion, the accounting identities that link the statements, and the likely path of each driver. Think deeply and check your work before answering. Do all of this reasoning internally -- the visible output must still be ONLY the Q-lines specified in the OUTPUT FORMAT section below.

Each user message states the firm's industry using the Fama-French 48 sector classification. Treat it as a prior -- sector-typical margins, seasonality, asset intensity, capital structure, and growth differ markedly across industries -- but always defer to the firm's own reported history where the two disagree.

All values are in millions USD. Negative = losses / outflows / contra-equity. Variable names are Compustat quarterly codes. All cash-flow items are QUARTERLY (pipeline has already converted YTD reporting to single-quarter flows). Signs follow Compustat conventions -- SEE THE SIGN COLUMN in each glossary row.

=====

GLOSSARY -- primitives you MUST forecast are marked [P]; derived items are computed mechanically from primitives and are shown here as context only.

=====

-- INCOME STATEMENT -----

[P] revtq	Revenue, total	(+)
[P] cogsq	Cost of goods sold	(+)
[P] xsgaq	Selling, general & admin expense -- INCLUDES xrdq	(+)
[P] xrdq	Research & development expense (informational sub of xsgaq)	(+)
[P] dpq	Depreciation & amortization expense	(+)
[P] stkcoq	Stock-based compensation expense (informational)	(+)
[P] xintq	Interest expense	(+)
[P] nopiq	Non-operating income/(expense), net	(signed)
[P] spiq	Special items (impairments, restructuring, etc)	(signed; usually <= 0)
[P] txtq	Income tax expense	(signed)
[P] miiq	Noncontrolling interest (income attrib. to minority)	(signed)
[P] xidoq	Extraordinary items & discontinued operations	(signed)
gpq	Gross profit = revtq - cogsq	(derived)
oibdpq	EBITDA = gpq - xsgaq	(derived)
oiadpq	EBIT = oibdpq - dpq	(derived)
piq	Pretax income = oiadpq - xintq + nopiq + spiq	(derived)
ibq	Income before extra = piq - txtq - miiq	(derived)
niq	Net income = ibq + xidoq	(derived)
xoprq	Total operating exp. = cogsq + xsgaq	(derived)

-- BALANCE SHEET -- ASSETS -----

[P] cheq	Cash & short-term investments	(+)
[P] rectq	Accounts receivable, trade	(+)
[P] invtq	Inventories	(+)
[P] acoq	Other current assets	(+)
[P] ppentq	Property, plant & equipment, NET	(+)
[P] dpactq	Accumulated depreciation	(+)
[P] gdlq	Goodwill	(+)
[P] intanoq	Other intangibles (excl. goodwill)	(+)
[P] aoq	Other assets -- AGGREGATE that already contains intanq	(+)
intanq	Intangibles total = gdlq + intanoq	(derived)
ppegqtq	PP&E gross = ppentq + dpactq	(derived)
actq	Current assets total = cheq + rectq + invtq + acoq	(derived)
atq	Total assets = actq + ppentq + aoq	(derived)

-- BALANCE SHEET -- LIABILITIES -----

[P] apq	Accounts payable, trade	(+)
[P] dlcq	Debt in current liabilities (short-term debt)	(+)
[P] txpq	Income taxes payable	(+)
[P] drcq	Deferred revenue, current	(+)
[P] lcoq	Other current liabilities	(+)
[P] drltq	Deferred revenue, long-term -- subset of loq	(+)
[P] dlrtq	Long-term debt	(+)
[P] txditcq	Deferred taxes & investment tax credit	(signed; usually +)
[P] loq	Other liabilities -- AGGREGATE that already contains drltq	(+)
lctq	Current liab. total = apq + dlcq + txpq + lcoq	(derived)
	(drcq is typically inside lcoq;	
	do not add drcq again)	

```

ltq      Total liabilities      = lctq + dlttq + txditcq + loq      (derived)

-- BALANCE SHEET -- EQUITY -----
[P] cstkq  Common stock, par value              (+)
[P] capsq  Additional paid-in capital / capital surplus      (+)
[P] req    Retained earnings -- already includes AOCI        (signed)
[P] acomincq AOCI component inside req (disclosure breakout) (signed)
[P] tstkq  Treasury stock at cost -- POSITIVE magnitude, subtracted (+)
[P] pstkq  Preferred stock                          (+)
[P] mibtq  Noncontrolling interest, total              (+)
ceqq      Common equity          = cstkq + capsq + req - tstkq    (derived)
seqq      Total stockholders eq. = ceqq + pstkq                (derived)

-- CASH FLOW (all quarterly; YTD already converted) -----
Operating:
[P] oancfq  Cash from operating activities              (signed)
[P] fopopq  Other operating CF adjustments              (signed)
Investing (individual components; invncfq is DERIVED from them):
[P] capxq  Capital expenditures -- positive OUTFLOW        (+)
[P] ivchq  Increase in LT investments -- positive OUTFLOW   (+)
[P] aqcq   Acquisitions -- positive OUTFLOW                (+)
[P] sivq   Sale of investments -- positive INFLOW           (+)
[P] sppeq  Sale of property -- positive INFLOW             (+)
[P] ivstchq Change in ST investments (separate from ivchq) (signed)
[P] ivacoq  Other investing activities                    (signed)
invncfq    Cash from investing      = -capxq - ivchq - aqcq
          + sivq + sppeq + ivstchq + ivacoq (derived)
Financing (individual components; fincfq is DERIVED from them):
[P] sstkq  Stock issuance -- positive INFLOW               (+)
[P] prstkq  Stock repurchases -- positive OUTFLOW           (+)
[P] dltsq  Long-term debt issuance -- positive INFLOW       (+)
[P] dltrq  Long-term debt repayment -- positive OUTFLOW     (+)
[P] dlchq  Changes in short-term debt                      (signed)
[P] dvq    Dividends paid -- positive OUTFLOW               (+)
[P] txbcfq  Excess tax benefit of stock options             (signed; often 0 post-2016)
[P] fiaq   Other financing activities                      (signed)
fincfq     Cash from financing      = sstkq - prstkq + dltsq - dltrq
          + dlchq - dvq + txbcfq + fiaq (derived)
FX & free cash flow:
[P] exreq  Effect of exchange rate on cash                  (signed)
fcfq      Free cash flow      = oancfq - capxq              (derived)

=====
ACCOUNTING CONSISTENCY CHECKS (your primitives must satisfy these)
=====

(1) Accounting equation:  atq = ltq + mibtq + seqq
(2) Cash rollforward:    cheq[t] = cheq[t-1] + oancfq + invncfq + fincfq + exreq
(3) Retained earnings:   req[t] ~ req[t-1] + niq - dvq + Delta acomincq
                        (approximate; non-cash AOCI + reclassifications add noise)

=====
OUTPUT FORMAT
=====

Each user message lists the EXACT subset of [P] items this firm reports
(items it never discloses are excluded -- they are treated as 0 by us). You
must:
* Forecast EVERY item in that per-firm list.
* Do NOT output any item not in that list (no hallucinated zeros for
  structurally-absent items).
* Do NOT output any derived item -- they are recomputed mechanically.

One line per horizon, values comma-separated. Output Q1 through Q{max_horizon}
(forecast EVERY horizon in that range, even if you are uncertain a later
quarter will materialize -- do not stop early):

Q1: revtq=X.XX, cogsq=X.XX, xsgaq=X.XX, ...
Q2: revtq=X.XX, cogsq=X.XX, xsgaq=X.XX, ...
...
Q{max_horizon}: revtq=X.XX, cogsq=X.XX, xsgaq=X.XX, ...

Output ONLY the Q-line forecasts. No markdown, no commentary, no code fences.

```

C.2.2. Structured arm: the block that replaces the steering paragraph

The remainder of the prompt (industry prior, units and sign conventions, glossary, consistency checks, output format) is unchanged from the unstructured arm.

PROJECTION METHOD -- forecast a few DRIVERS, then DERIVE the rest. Reason through this sequence internally (in growth rates and ratios), then convert back to dollar levels before emitting:

1. Revenue (revtq) first. Anchor on the SAME fiscal quarter a year earlier: $\text{revtq}[t] \sim \text{revtq}[t-4] \times (1 + g)$, where g is a year-over-year growth rate read from the recent trailing-4-quarter trend and FADED toward a modest long-run rate across the 20-quarter horizon -- do not let a recent spike persist to Q20. Anchoring on $t-4$ preserves seasonality.
2. Operating lines as ratios to revenue. Project cogsq, xsgaq (which includes xrdq), dpq, and stkcoq each as a share of revtq, held near its trailing-4-quarter median and drifting only with a clear operating-leverage reason; convert back to a level = ratio $\times$ revtq[t]. Income tax (txtq) $\sim$ tax rate $\times$ pretax income, and tracks the sign of pretax income.
3. Working capital via turnover. Project receivables (rectq) off revtq (days-sales-outstanding) and inventory (invttq) and payables (apq) off cogsq (days-inventory / days-payable); hold each turnover near its trailing level, then convert back to a level using projected revtq/cogsq.
4. PP&E by roll-forward. $\text{ppentq}[t] \sim \text{ppentq}[t-1] + \text{capxq} - \text{dpq}$, with capxq projected as a share of revtq.
5. Cash flow first, then SET cash and retained earnings by roll-forward -- do NOT forecast cheq or req independently:
 - * Operating cash flow must be consistent with earnings and working capital: $\text{oancfq} \sim \text{niq} + \text{dpq} - \Delta \text{rectq} - \Delta \text{invttq} + \Delta \text{apq}$. A RISE in receivables or inventory is a USE of cash (subtract); a rise in payables is a SOURCE (add); dpq is a positive add-back. capxq is a cash OUTFLOW (in invncfq); debt issuance is a source while repayments and dividends (dvq) are uses (in fincfq).
 - * Cash: $\text{cheq}[t] = \text{cheq}[t-1] + \text{oancfq} + \text{invncfq} + \text{fincfq} + \text{exreq}$. Set cheq to this rollforward result, not to an independent guess.
 - * Retained earnings: $\text{req}[t] \sim \text{req}[t-1] + \text{niq} - \text{dvq} + \Delta \text{acomincq}$. Set req to this roll.

Before emitting each quarter, silently confirm the three ACCOUNTING CONSISTENCY identities below hold (accounting equation, cash rollforward, retained-earnings roll); if one fails, adjust the dependent line (cheq, req, or an equity line) so it holds.

Do all of this reasoning internally -- the visible output must still be ONLY the Q-lines specified in the OUTPUT FORMAT section below.

C.2.3. Per-origin user message (both arms)

Assembled per origin: the fixed horizon instruction, the firm's Fama-French-48 industry, the firm's active-primitive list, then the 12-quarter history as CSV-like blocks (one per statement; rows null across all history are suppressed). Schematic, with illustrative values and the history truncated to 3 of 12 quarters for space:

Forecast horizons: Q1 through Q20.

Firm industry (Fama-French 48): Retail.

This firm reports 47 primitive line items. Forecast ONLY these items -- do not output any other primitives:

revtq, cogsq, xsgaq, dpq, xintq, ..., oancfq, capxq, dvq

Items not in this list are structurally absent for this firm (consistently null in history); they will be treated as 0.

Historical data:

```
=== INCOME STATEMENT ===
Account,Q-11,...,Q-1,Q0
revtq,1203.40,...,1456.10,1502.77
cogsq,742.10,...,889.30,910.55
...
=== BALANCE SHEET - ASSETS ===
Account,Q-11,...,Q-1,Q0
cheq,310.22,...,402.10,419.88
...
```

The model returns one comma-separated line per horizon (Q1: revtq=..., cogsq=..., ...→Q20: ...), primitives only; the 23 derived and subtotal items are reconstructed mechanically downstream.