Analysis

Time-Series Foundation Models Enter FMCG Replenishment: Their Value Lies in Cold-Start Priors and Service-Level Cost, Not Zero-Shot Leaderboards

HX Research

HX Research · Industry Modelling Practice · November 4, 2026

Time-series foundation models now forecast unseen series zero-shot, and leaderboard errors improve month by month. Should FMCG companies hand store × SKU × week demand forecasting over to them wholesale? We argue that, in the replenishment chain, their value lies mainly in acting as a cross-series probabilistic prior for cold-start and long-tail series, not in replacing existing methods; models should be selected on newsvendor cost and service level at the critical quantile, after stockout censoring and promotional covariates have been handled. Three lines of evidence support this: on an intermittent-demand benchmark, accuracy rank and service rank are negatively correlated; break-even analysis shows zero-shot pays off on short, seasonal series while classical methods and boosted trees often catch up on long ones; and training on sales systematically underestimates demand. We formalize the equivalence of newsvendor cost and quantile loss, the censored likelihood, hierarchical reconciliation and shrinkage blending; propose a segmentation framework and compare six modelling routes; give a decision-aware selection and reconciliation algorithm with three fixed-seed simulations; and state five falsifiable forecasts.

Keywords: time-series foundation models; demand forecasting; intermittent demand; newsvendor model; quantile loss; stockout censoring; hierarchical reconciliation; FMCG replenishment

Introduction

Over the past three years, time-series forecasting has followed a path much like that of large language models: pretrain one general model on a large, multi-domain corpus of series and, at deployment, skip per-dataset training by feeding the history in as context and forecasting the future directly. Chronos turns values into discrete tokens through scaling and quantization and trains language-model architectures with cross-entropy[1]; TimesFM uses a patched decoder-only architecture[2]; Moirai performs unified training on an open corpus of more than 27 billion observations[3]; and Lag-Llama, TimeGPT, Time-MoE and Sundial push forward through lag features, a commercial API, mixture-of-experts and a flow-matching objective, respectively[4–7]. In 2025–2026, Chronos-2, Moirai 2.0, TiRex and Toto 2.0 brought multivariate inputs, covariates, streaming inference and parameter scale into a single framework[8–11]. According to the paper, Chronos-2 raises skill scores by 5.2 percentage points on average over the strongest baselines across several benchmarks and forecasts 100 series in under one second on a single mid-range GPU[12].

Evaluation has expanded alongside capability. GIFT-Eval covers 23 datasets and 144,000 series[13]; fev-bench contains 100 tasks, 46 of them with covariates, aggregated with bootstrapped confidence intervals[14]; and TIME uses 50 fresh datasets for strict zero-shot evaluation[15]. Leaderboards have made ”zero-shot error” the number most often quoted about foundation models.

Demand forecasting in FMCG, however, is not a problem of error alone. In a store × SKU × week panel, many series have very short histories (new products, new stores, new channels), many are intermittent or lumpy (long-tail SKUs, small stores), promotions, prices and holidays cause large jumps, and stockouts push observed sales below true demand[16,17]. Downstream of the forecast sits replenishment: the order quantity depends on the relative costs of shortage and excess, and the direction of an error matters more than its size. The classical newsvendor model established long ago that the optimal order quantity is the critical quantile of the demand distribution[18], and operations research on integrated prediction and optimization has shown repeatedly that the forecast with the smallest error need not yield the cheapest decision[19,20].

Recent evidence makes this tension concrete. Chin and colleagues compared 38 forecasting methods — classical, intermittent-demand-specific, machine-learning, deep-learning and pretrained foundation models — on data from a contract-logistics spare-parts operator; according to the paper, across 20,330 real multi-line orders, forecast-accuracy rank and order-service rank correlate at −0.555, and service depends more on the direction of cumulative bias, including over-prediction in zero-demand periods[21]. Tan Jerome and colleagues ran a break-even analysis on 30 datasets; according to the paper, foundation

models win at every training fraction on 15 datasets, but on 6 others classical methods beat zero-shot foundation models with as little as 2% of the training data, and one deployment rule needs no training at all: if there are fewer than 700 training samples and seasonality is non-negligible, use zero-shot[22].

Existing discussions leave three gaps. First, model selection mostly reuses leaderboard point or probabilistic errors and never writes the cost asymmetry of replenishment into the evaluation function. Second, leaderboard data usually carry no stockout labels, so "sold little" cannot be told apart from "could not sell because the shelf was empty"; in FMCG this is precisely the source of systematic demand underestimation[23]. Third, foundation models are discussed as replacements and rarely as priors: they are most useful when histories are shortest, their marginal value declines once histories are long enough, and the boundary between the two regimes should be computable.

Our central thesis is that, in the FMCG replenishment chain, foundation models are valuable mainly as cross-series probabilistic priors for cold-start and long-tail series rather than as wholesale replacements; selection and deployment should be governed by newsvendor cost and service level at the critical quantile, with stockout-censoring repair, promotional covariates and hierarchical reconciliation as prerequisites. We make three contributions:

1. A decision-oriented formalization — the equivalence between newsvendor cost and quantile loss at the critical quantile, the stockout-censored likelihood, hierarchical reconciliation, and a prior blend that shrinks with history length — from which we derive a segment-level deployment rule for foundation models.
2. A framework that segments series by history length × demand pattern × demand driver, a systematic comparison of the mechanisms, evidence and limits of six modelling routes, and a review of leakage and benchmark bias in foundation-model evaluation.
3. A decision-aware forecast selection and reconciliation algorithm (DAFSR), three fixed-seed illustrative simulations showing cold-start shrinkage, the divergence between accuracy and service rankings and censoring repair, and five falsifiable forecasts.

Problem formalization and notation

Newsvendor cost and the critical quantile

Let d be the demand for a given SKU in a given store over one replenishment cycle, q the order quantity, c_u the unit shortage cost (lost margin and customer churn) and c_o the unit excess cost (holding, shrinkage and markdowns near expiry). The expected cost and the optimal order quantity are[18]:

$$C(q) = c_u \cdot E[max(d-q,0)] + c_o \cdot E[max(q-d,0)], \quad q^* = F^{-1}(\tau^*), \quad \tau^* = c_u/(c_u+c_o) \tag{1}$$

where F is the cumulative distribution function of demand and τ^* is the critical quantile. In FMCG, c_u and c_o differ by category: short-shelf-life fresh food has high excess costs and hence a low τ^*, whereas high-margin core SKUs with strong substitutes have high shortage costs and hence a high τ^*.

Newsvendor cost is quantile loss at the critical quantile

Rewriting the single-period cost pointwise yields a simple but crucial identity:

$$c_u \cdot max(d-q,0) + c_o \cdot max(q-d,0) = (c_u+c_o)\cdot \ell_{\tau^*}(d,q), \quad \ell_\tau(d,q) = max(\tau(d-q), (\tau-1)(d-q)) \tag{2}$$

ℓ_τ is the quantile (pinball) loss[24]. Equation (2) shows that, for a given τ^*, minimizing newsvendor cost is exactly equivalent to minimizing quantile loss at τ^*; the quantile is the optimal point forecast under this loss, and the corresponding scoring rule is strictly proper[25,26]. Two corollaries follow. First, selecting models by mean absolute error or mean squared error implicitly assumes $\tau^* = 0.5$ or cares only about the mean; when τ^* departs from 0.5, this default systematically picks the wrong model. Second, quantile loss weighted at τ^* can serve directly as the model-selection metric for replenishment, differing from business cost only by a positive constant.

Stockout censoring

Stores observe sales, not demand. Let I_t be the sellable stock at the start of period t; observed sales and the censored likelihood are:

$$s_t = min(d_t, I_t), \quad log\, L(\theta) = \sum_{no\ stockout} log\, f_\theta(s_t) + \sum_{stockout} log\, P_\theta(d_t \geq I_t) \tag{3}$$

Training directly on s_t as if it were d_t replaces the stockout-period likelihood terms with point probabilities and biases the estimator downwards. The bias grows with stockout frequency and with τ^*: the higher the service level one wants to maintain, the more one relies on the right tail of the distribution — exactly the part that has been cut off. According to the paper, FreshRetailNet-50K uses hourly stockout labels to reconstruct latent demand before training forecasting models, cutting systematic underestimation from 7.37% to near zero while improving accuracy by 2.73%[23]; for "double censoring", in which promotions inflate sales while stock caps them, a dedicated quantile-learning framework is also available[27].

Hierarchical reconciliation

FMCG forecasts are naturally hierarchical: store × SKU aggregates to region × category and then to the national level, while weeks aggregate to months. Let $\hat{y}$ be the vector of base forecasts and S the summing matrix; the reconciled forecasts are:

$$\tilde{y} = S \cdot G \cdot \hat{y}, \quad G_{MinT} = (S'W^{-1}S)^{-1}S'W^{-1} \tag{4}$$

where W is the covariance of base-forecast errors. Under unbiasedness, MinT minimizes the total error variance of the reconciled forecasts[28], building on earlier optimal-combination methods[29,30]; probabilistic reconciliation and score optimization have also been treated systematically[31], and a textbook introduction is available[32]. For large retailers, Wang and colleagues scaled optimization-based point

reconciliation beyond four billion forecasted values and proved that, under suitable weighting, least-squares reconciliation is equivalent to share-based allocation[33]; for temporal hierarchies, theory and simulation support the competitiveness of bottom-up aggregation[34]. Reconciliation matters for replenishment because distribution centres order at the aggregate level while stores replenish at the bottom level, and inconsistent numbers between the two are amplified step by step — the bullwhip effect[35].

A prior blend that shrinks with history length

Treating the foundation model as a cross-series prior and the series' own history as the likelihood, the simplest combination is credibility weighting (shrinkage):

$$\hat{\mu}_i(n) = \lambda(n)\cdot\mu_{prior,i} + (1-\lambda(n))\cdot\mu_{local,i}, \quad \lambda(n) = n_0/(n_0+n), \quad n_0 = \sigma^2_\varepsilon / \sigma^2_\eta \tag{5}$$

where n is the number of periods of history for the series, σ^2_ε is the relative variance of single-period observation noise and σ^2_η is the relative variance of true levels across similar series. The meaning of n_0 is intuitive: how many periods of the series' own data the prior is "worth". The noisier the observations and the more alike the peers, the larger n_0 and the longer the prior stays useful. Equation (5) is an approximation at the level of the mean; for quantiles, the same blend can be done at the distribution level, or the few observed periods can be passed as context to a model that supports in-context learning[36].

A segment-level deployment rule

Let q_{base} be the order quantity from a candidate baseline (a statistical method or gradient-boosted trees) and q_{FM} that from a foundation model. Deploying the foundation model on segment g is justified when:

$$\Delta C_g = \sum_{i\in g}[C_i(q_{base}) - C_i(q_{FM})] > K_g \tag{6}$$

where C_i is estimated by equations (1)–(2) on a rolling-origin backtest and K_g is the inference, operations and monitoring cost for the segment. Equation (6) rewrites the break-even logic of Tan Jerome and colleagues[22] from error into cost and lowers the decision granularity from the dataset to the segment.

Notation

Table 1 summarizes the notation.

Table 1 | Main notation

Symbol	Meaning	Where used
d, q, F, τ^*	Demand, order quantity, demand CDF, critical quantile	Eq. (1)
c_u, c_o	Unit shortage cost, unit excess cost	Eqs (1)–(2)
$\ell_\tau(d, q)$	Quantile (pinball) loss at τ	Eq. (2)
s_t, I_t, θ	Observed sales, opening sellable stock, demand-distribution parameters	Eq. (3)
$\hat{y}$, $\tilde{y}$, S, G, W	Base forecasts, reconciled forecasts, summing matrix, mapping matrix, error covariance	Eq. (4)
n, n_0, $\lambda(n)$, σ^2_ε, σ^2_η	History length, prior-equivalent periods, prior weight, observation noise and cross-series dispersion	Eq. (5)
g, ΔC_g, K_g	Segment, cost saving from deploying the foundation model, deployment cost	Eq. (6)

Analytical framework: segmenting by history length, demand pattern and demand driver

Whether foundation models are "good enough" is not a global question but a segmented one. We slice the store × SKU panel along three dimensions. The first is history length n: new products and new stores have a few weeks of data, mature SKUs more than two years. The second is the demand pattern: by average demand interval (ADI) and the squared coefficient of variation of non-zero demand sizes (CV^2), series fall into smooth, erratic, intermittent and lumpy classes[37]. The third is the demand driver: whether demand is shaped mainly by seasonality and trend or by exogenous events such as promotions, prices, displays and holidays.

Each cell in Table 2 corresponds to a mechanism in Section 4. The table also implies directly that neither "replace everything" nor "use nothing" is optimal: the deployment footprint of foundation models should be decided segment by segment through equation (6).

Mechanism analysis and route comparison

Cold start: why a cross-series prior works

The cold-start advantage of foundation models comes from the many "similar shapes" they have seen during pretraining. With only four weeks of data for a new product, its own series cannot reveal annual seasonality, and a local model can only fall back to extrapolating a level or a trend; a foundation model can place those four weeks in context and match them against similar shapes seen in pretraining. Chronos's evaluation on 42 datasets shows zero-shot performance comparable to, and occasionally better than, models trained specifically on the target data[1]; TimesFM reports zero-shot accuracy close to that of the best supervised model for each dataset[2]; and TabPFN-TS even shows that a tabular foundation model with only 11 million parameters, plus lightweight temporal features, can lead on covariate-informed tasks[38].

This advantage has clear limits. We introduced the break-even analysis of Tan Jerome and colleagues above; another of its findings is worth adding: according to the paper, LoRA fine-tuning on short series can actually degrade performance[22]. Ebrahim and colleagues built 7,200 series from nine known generating processes that the model cannot have seen; according to the paper, TimesFM-3's error is never more than 1.31 times that of the best method in any scenario, whereas every classical method reaches at least 2.2 times somewhere; the largest classical failures occur

Table 2 | A segmentation framework for FMCG demand series: main risks and the appropriate role of foundation models

Segment	Typical objects	Main risk	Appropriate role for foundation models	Preferred evaluation
Short history × smooth or seasonal	New products, stores, channels	Too little own information; seasonality not estimable	Primary: zero-shot prior, shrunk with n by Eq. (5)	Quantile loss at τ^*, first-month fill rate
Long history × smooth	Mature core SKUs	Structural change, promotional noise	One candidate, backtested alongside statistical methods and boosted trees	Quantile loss at τ^*, bias
Any history × intermittent or lumpy	Long-tail SKUs, small stores	Many zeros; point forecasts drift to zero and service collapses	Supporting: supplies distribution shape, needs bias correction	Fill rate, complete-order rate
Promotion- and price-driven	Featured and campaign items	Jumps driven by exogenous events	Only covariate-capable models, or combined with regression models	Quantile loss in promotion periods
Frequent stockouts	Best-sellers, fresh food	Sales understate demand	Repair censoring before training	Gap between target and achieved service
Aggregate levels	Region × category, distribution centres	Cross-level inconsistency, bullwhip	Base forecasts at bottom or middle levels feed reconciliation	Cross-level coherence, aggregate error

with only two seasonal cycles, when automatic methods fall back to non-seasonal models; yet on seasonal ARIMA data with four or more cycles, automatic ARIMA is 21–24% more accurate than TimesFM-3[39]. Both studies point to the same conclusion: the robustness of foundation models is worth most when histories are short, and local methods catch up as histories lengthen. This is the empirical basis for $\lambda(n)$ declining with n in equation (5).

Intermittent and lumpy demand: point forecasts can "forecast away" the service level

Most weekly demand for long-tail SKUs is zero. For such series, the point forecast that minimizes mean absolute error is the median, and the median is often zero. The illustrative simulation in Fig. 2 shows this mechanism: the "zero forecast" has the lowest MASE and a fill rate of zero. MASE is a scale-free measure designed with intermittent series in mind[40]; the problem lies not in the metric itself but in using it to choose a model for a decision whose τ^* is far above 0.5.

Classical intermittent-demand methods were built for inventory from the outset. Croston's method smooths demand sizes and inter-demand intervals separately[41], and Syntetos and Boylan identified its bias and proposed a correction[42]; deep renewal processes connect these methods to temporal point processes[43]; and the evaluation of count distributions has received dedicated treatment[17]. Recent comparisons show that gradient-boosted trees used alone often underperform specialized methods on intermittent demand, whereas combining a machine-learning model with a specialized method improves accuracy by up to 10%[44]; combinations of probabilistic forecasts beat individual methods, but accuracy trades off against inventory performance[45].

Foundation models have no inherent advantage here. According to the paper, TimesFM-3 shows no advantage over simple history-based benchmarks on intermittent demand[39]. A deeper reason is "forecast collapse": when target predictability is low, the amplitude of well-calibrated point forecasts is compressed and forecasts flatten[46]. Chin and colleagues point to a remedy: service is more closely tied to the direction of cumulative bias than to point accuracy; according to the paper, examining the bias in Chronos-2's instance normalization yields a training-free correction that lifts the per-material fill proxy from 77.5% to 92.0% at a 90% policy target and the complete-order fill rate from 54% to 63%[21]. The implication is that, for intermittent series, foundation models are better suited to supplying the shape of the distribution, which bias correction and quantile choice then "translate" into order quantities.

Promotions and covariates: most foundation models did not originally "see" promotions

Large jumps in FMCG demand are mostly driven by promotions, prices and campaign calendars. Early foundation models were mostly pretrained on univariate series and could not use such information directly[47]. Since 2025, covariate capability has become a main axis of competition: ChronosX injects past and future covariates through pluggable modules, with retail promotions and holidays as its motivating example[48]; COSMIC uses covariates zero-shot via in-context learning[49]; UniCA unifies categorical variables and heterogeneous image and text covariates into series representations[50]; and Chronos-2 shares information across related series, multivariate targets and covariates through group attention — according to the paper, its advantage on covariate tasks is substantial, and the paper includes a retail case study[8].

But "supporting covariates" is not the same as "using them correctly". Berthelier and colleagues ran controlled experiments on simple target–covariate relationships; according to the paper, TabPFN-TS captures these relationships better than Chronos-2, especially at short horizons, and Chronos-2's strong overall benchmark performance does not automatically translate into optimal modelling of simple dependencies[51]. GIFT-Eval-CTX tests whether models can turn verified future information into numerical trajectories; according to the paper, structured effects

reduce relative MASE by 40.7% and 29.6% for two frontier large language models[52]. For FMCG, this means promotion calendars, price plans and display schedules are themselves high-value "future information", and how well they are structured often matters more than which foundation model is chosen.

By contrast, the lessons of the M5 competition still hold: according to the paper, top solutions in the accuracy track relied heavily on gradient-boosted trees such as LightGBM and combined forecasts across the hierarchy[53,54]; the uncertainty track evaluated performance across multiple quantiles[55]; and blends of boosted trees and neural networks are competitive for both point and probabilistic forecasts[56]. For mature SKUs with rich covariates, gradient-boosted trees remain a strong baseline that must be tested head to head.

Censoring and bias: fix the data before choosing the model

Section 2.3 showed that training on sales biases demand estimates downwards. A high τ^* amplifies this bias: the order quantity takes a right-tail quantile, and the right tail is exactly what stockouts cut off. In the illustrative simulation of Fig. 3, about one week in three is a stockout week; training directly on sales underestimates the mean by about 18% on average, and the achieved cycle service level at a 90% target is only 82.4%; after censored maximum-likelihood repair by equation (3), the bias falls to about 1% and achieved service returns to 90.2%. Repair presupposes stockout labels or inferable stock records, which is the value of datasets such as FreshRetailNet-50K with hourly stockout annotations[23].

Censoring repair is just as necessary for foundation models. A zero-shot model that takes sales as demand in its context will likewise learn the "truncated" distribution; in a prior blend, both the prior and the series itself should be based on repaired demand.

Hierarchy, online adaptation and uncertainty

Reconciliation inputs can come from different models. A common combination in practice is foundation models or statistical methods producing probabilistic forecasts at the bottom level, boosted trees with promotional covariates in the middle, and reconciliation by equation (4) at the end. Reconciliation itself does not care which model produced the base forecasts, which lets foundation models be "plugged in" to existing pipelines[28,33].

After deployment, three lightweight adaptation routes are available. The first is online weighting: ELF learns the current distribution with a lightweight forecaster and combines it with the foundation model's forecast; according to the paper, it improved performance in all of its experiments[57]. The second is black-box correction: ECHO corrects foundation-model outputs using historical analogues and past forecast errors; according to the paper, it reduced mean squared error in all 45 combinations of 5 foundation models and 9 real-world datasets, by 9.26% on average[58]. The third is expert editing: DEFT revises a foundation model's candidate trajectories in a trend–seasonal decomposed space under a limited budget of expert scores[59], much as FMCG planners hand-adjust forecasts before a campaign. By contrast, continual fine-tuning causes catastrophic forgetting; according to the paper, smaller models equipped with forgetting mitigation can match foundation models by the end of a continual-learning sequence[60].

Replenishment needs intervals with coverage guarantees. Distribution-free conformal methods can construct prediction intervals for time series[61], and adaptive conformal inference maintains long-run coverage under distribution shift[62]. For the quantiles a foundation model outputs, conformal calibration adds a model-agnostic layer of insurance.

Evaluation pitfalls and counter-evidence

Three kinds of evidence about foundation models deserve a discount. The first is information leakage: Meyer and colleagues identify two kinds — train–test sample overlap caused by multi-purpose reuse of datasets, and temporal overlap of correlated series — both of which make results overly optimistic[63]; GIFT-Eval provides a non-leaking pretraining corpus for this reason[13], and TIME switches to entirely new datasets[15]. The second is domain representativeness: IND-Eval finds on industrial data that the overall leader does not lead on every source, so model selection requires domain-representative evaluation[64]. The third is where orchestration gains come from: FairAMO runs controlled comparisons with frozen candidate forecasts; according to the paper, the best agentic orchestrators reduce MASE by less than 6% relative to the best single model, and the gains come from complementarity among candidates rather than their number[65].

The counter-evidence also deserves serious attention. Those who expect foundation models to replace existing methods wholesale can point to Chronos-2's overall lead[12], Toto 2.0's steady scaling from 4 million to 2.5 billion parameters[11] and Moirai 2.0's balance of accuracy, speed and size[9]; inference costs are also falling fast — according to the official blog, Chronos-Bolt is up to 250 times faster and 20 times more memory-efficient than the original Chronos[66], Chronos-2 is a 120-million-parameter encoder-only model[67], and the public TimesFM 2.5 checkpoint has 200 million parameters[68]. We accept that, as in-context learning and covariate capabilities keep improving, the "primary" column of Table 2 may widen. That does not change the core of our thesis: the selection criterion should be cost at τ^* rather than leaderboard error, and censoring repair and covariates are prerequisites — neither of which depends on model size.

No route in Table 3 dominates in every column. The distinctive value of R4 and R5 is concentrated in the "data needs" column: they bring training-free, cross-series priors to cold start. The value of R1 and R2 lies in interpretability and cost: the former has mature intermittent methods, and the latter proved in M5 its ability to exploit promotional and hierarchical features. The gains of R6 are bounded by candidate complementarity. This landscape supports

Table 3 | Mechanism-level comparison of six demand-forecasting routes (evidence mostly self-reported in papers or official materials)

Route	Core mechanism	Data needs	Covariates	Intermittent demand	Probabilistic output	Operating cost	Best-fit segments
R1 Statistical methods (ETS, Theta, Croston/SBA)	Per-series state space or smoothing[69,70]	Enough history per series	Weak	Dedicated methods	Yes, with distributional assumptions	Low	Long smooth histories, intermittent
R2 Gradient-boosted trees (global)	Cross-series feature regression[53,54]	Large panels with features	Strong	Weak when used alone	Needs quantile objectives	Medium	Promotion-driven, mature SKUs
R3 Deep global models (DeepAR, TFT, N-BEATS, etc.)	Cross-series neural networks[71–73]	Large panels, requires training	Medium to strong	Needs dedicated likelihoods	Yes	Medium to high	Large panels, multi-step
R4 Foundation models, zero-shot	Pretrained cross-series prior	No training	Model-dependent	No inherent advantage	Yes	Medium (GPU inference)	Cold start, short histories
R5 Foundation models + covariates / in-context learning	Group attention, adapter modules[8,48]	A few related series plus future information	Strong	Needs bias correction	Yes	Medium	Campaign items, new products with promotions
R6 Orchestration and combination	Select or weight several candidates[65]	Backtests of several candidates	Inherited	Inherited	Inherited	High	When candidates are complementary

our thesis of segmented deployment. The lessons of the M4 competition point the same way: according to the paper, combinations and hybrid methods performed strongly, while pure machine-learning methods did not dominate at the time[74]; a review of forecasting theory and practice likewise systematically collects experience on method selection and combination across settings[75].

From cost parameters to service targets

Many companies configure their replenishment systems with targets such as "95% service level" rather than with c_u and c_o. The two can be converted: by equation (1), a cycle-service target α implies a cost ratio $c_u/c_o = \alpha/(1-\alpha)$. A 95% target means the company believes that one unit short costs 19 times as much as one unit in excess; 90% corresponds to 9 times and 75% to 3 times. This conversion often exposes inconsistent targets: when fresh food with a shelf life of a few days and beverages with a shelf life of a year are given the same service target in one category, the implied cost ratios clearly cannot both hold.

The conversion also clarifies two easily confused metrics. Cycle service level is the probability of no stockout in a period and corresponds to τ^* in equation (1); fill rate is the share of demand that is satisfied, and for the same order quantity it is usually higher than cycle service level because some demand is still met in stockout periods. Fig. 2 and Fig. 3 report these two metrics respectively, and readers should not compare them directly. When choosing models, one should first establish which metric the business is actually held to and then write it into the evaluation function of equation (2); otherwise the same backtest can support entirely different conclusions.

Finally, c_o for fresh and near-expiry products is not constant; it rises as remaining shelf life shrinks. This means τ^* should adapt to each store's turnover and delivery frequency: high-turnover stores can afford a higher τ^*, low-turnover stores should lower it. Turning τ^* into a store × category parameter table, rather than one number for the whole company, often pays off faster than changing the forecasting model.

Method: decision-aware forecast selection and reconciliation

We condense the analysis into an executable procedure. The goal of Algorithm 1 is, for a store × SKU panel, to select or blend forecast sources by segment, repair censoring, backtest on cost at τ^*, reconcile across the hierarchy and output order quantiles, while leaving an auditable selection record and a fallback path.

Algorithm 1 embodies three design choices. First, step 2 precedes step 4: if backtests use censored sales, any model that underestimates demand will look "more accurate", and the selection itself is contaminated. Second, step 4 uses quantile loss at τ^* rather than MASE or CRPS as the primary metric and reports the direction of bias separately — a direct application of Chin and colleagues' finding[21]. Third, step 5 confines the foundation model to segments where equation (6) holds and shrinks it on short histories instead of replacing everything globally; step 8 keeps human review and fallback, consistent with treating forecasts as decision support rather than automated decisions.

Algorithm 1 | Decision-Aware Forecast Selection and Reconciliation (DAFSR)

Input: Store × SKU × week sales, opening stock and stockout flags; promotion, price and campaign calendars; hierarchy S; c_u and c_o per category; candidate pool (statistical methods, gradient-boosted trees, zero-shot foundation models, foundation models with covariates); deployment cost K_g

Output: Forecast source or blend weights per segment; reconciled τ^* quantiles and order suggestions; backtest and selection records; fallback rules

1: Segment: assign series to the segments g of Table 2 by history length n, ADI/CV^2 and promotion intensity
2: Repair censoring: for flagged stockout periods, estimate the demand distribution with the censored likelihood of Eq. (3); where no flags exist but stock records do, use $s_t = I_t$ as a stockout proxy and note this in the record
3: Guard against leakage: use rolling-origin backtests; confirm that the foundation model's pretraining corpus contains no same-source data from the evaluation window and that the window postdates the model's published training cut-off
4: Score backtests: for each candidate, compute quantile loss at τ^* within g by Eq. (2), and report fill rate, the direction of cumulative bias and loss in promotion periods
5: Select and blend: decide by Eq. (6) whether to deploy the foundation model on g; on short-history segments, blend prior and own series by Eq. (5), estimating n_0 from peer-series noise and dispersion; on intermittent segments, apply bias correction to foundation-model outputs
6: Reconcile: reconcile bottom-level and aggregate forecasts by Eq. (4) using MinT or share-based allocation; reconcile probabilistic forecasts at the quantile or sample level
7: Calibrate and order: apply conformal calibration to the reconciled quantiles, take the τ^* quantile as the order quantity, and impose minimum-display and case-pack constraints
8: Monitor and fall back: compare target and achieved service weekly; on sustained deviation beyond a threshold, revert to the previous version or the baseline method; planners review before campaigns

Complexity and scope: Step 2 is $O(T \cdot \Theta)$ per series (Θ is the parameter-grid size); step 4 is $O(M \cdot N \cdot T)$ for M candidates and N series; step 6 is typically a sparse linear solve. Applies to replenishment settings with stock or stockout records and estimable cost parameters; where cost parameters are unavailable, it degrades to reporting quantile-loss curves over several τ

Illustrative simulations

To make the three mechanisms concrete, we ran three illustrative simulations. All numbers are computed on synthetic data and are not measurements of any real company, model or dataset. The "pooled prior" in Simulation 1 stands in for the cross-series prior that a foundation model provides; no real foundation model was run. The scripts fix the random seed at 20261104.

Cold-start shrinkage. We simulate 20 categories with 30 SKUs each; true SKU levels are log-normally distributed around the category level ($\sigma_\eta = 0.3$), weekly demand is negative binomial and $\tau^* = 0.75$. Equation (5) gives $n_0 \approx 7.6$ weeks. Fig. 1 shows the results: using only the SKU's own history, newsvendor cost is 1.56 times the cost with the true mean known after 1 week of history and 1.14 times after 4 weeks; the pooled prior stays between 1.06 and 1.10 times throughout; after 12 weeks the SKU's own history starts to beat the pooled prior. The shrinkage blend is no worse than either at every history length — 1.09 times at 1 week and 1.01 times at 52 weeks.

Divergence between accuracy and service. We simulate 300 intermittent SKUs with 104 training weeks, 26 test weeks and a 90% service target. Of 13 methods, all except the empirical-quantile method — which outputs the τ^* quantile directly — follow common practice and order the 90% quantile of a Poisson distribution whose mean is the point forecast. Fig. 2 shows the results: the zero forecast and the 52-week median have the lowest MASE (about 0.66) but fill rates of 0 and 1.3%; the empirical-quantile method has the highest MASE (2.03), the highest fill rate

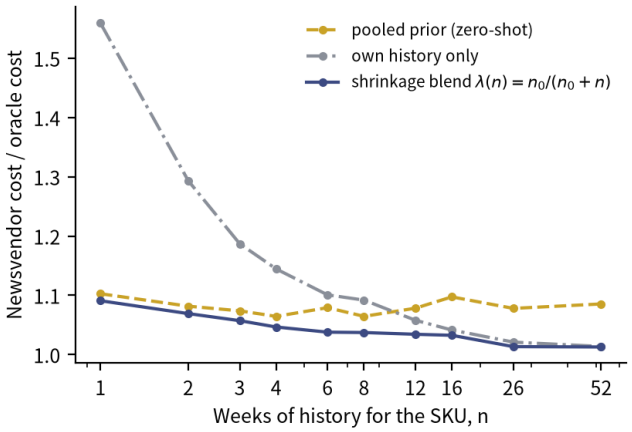


Fig. 1 | The value of the prior declines with history length. The pooled prior (standing in for a foundation model's cross-series prior) far outperforms the SKU's own history when histories are short; the shrinkage blend is no worse than either at any history length. Illustrative data with a fixed random seed, computed by the HX Research Team.

(80.2%) and the lowest unit cost. The Spearman correlation between MASE rank and fill-rate rank is −0.63, in the same direction as Chin and colleagues' finding on real data[21]. Notably, even the best method fails to reach a 90% fill rate, because the Poisson assumption understates the dispersion of lumpy demand and because cycle service level and fill rate are different metrics.

Censoring repair. We simulate 400 SKUs whose opening stock is set at the 70% quantile with perturbation, so

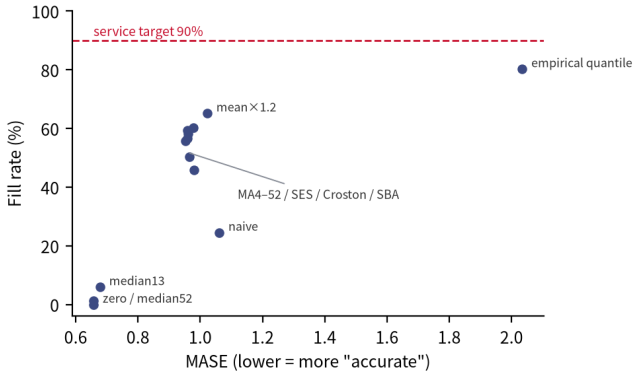


Fig. 2 | On intermittent demand, the "most accurate" forecast can deliver the worst service. The horizontal axis is MASE and the vertical axis fill rate; the zero forecast and median-type methods have the lowest error and the worst service. Illustrative data with a fixed random seed, computed by the HX Research Team.

that about 34% of weeks are stockout weeks. Estimating the mean directly from sales underestimates it by 18.0% on average; censored maximum-likelihood estimation by equation (3) leaves a bias of +1.1%. At targets of 80%, 85%, 90% and 95%, models trained directly on sales achieve cycle service levels of 71.2%, 76.8%, 82.4% and 89.8%, against 80.6%, 85.4%, 90.2% and 95.3% after repair (Fig. 3).

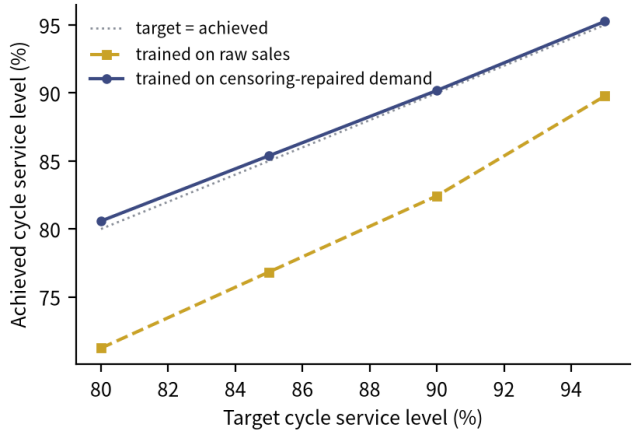


Fig. 3 | Training on sales systematically misses service targets. Stockouts keep sales below demand, so models trained directly on sales fall below the diagonal at every target; after censoring repair they return close to the diagonal. Illustrative data with a fixed random seed, computed by the HX Research Team.

Table 4 makes no claim about any real system. The mechanism it conveys is that, on the same panel, "smallest error" and "lowest cost" can point to different models, and whether the data are repaired matters more than which model is chosen.

Forecasts and research agenda

The following are the views of the HX Research Team, not statements of fact.

The research agenda contains four open problems. First, can the foundation model supply n_0 itself? If the width of its output distribution is reliable enough, the prior's equivalent periods could be estimated per series rather than set per category. Second, can censoring repair and foundation models be combined end to end — is passing stockout flags as covariates to a covariate-capable model enough to replace an explicit censored likelihood? Third, multi-level quantile reconciliation still lacks methods tied directly to replenishment cost; how an aggregate-level service target decomposes to the store level is a problem that spans forecasting and operations research. Fourth, the freshness of evaluation corpora: the gap between a foundation model's pretraining cut-off and the evaluation window should be disclosed as mandatorily as a data card.

Where common industry views may be wrong

"Lead the zero-shot leaderboard, and you should replace everything." Leaderboards measure average error, mostly on series without stockout labels; replenishment is measured by cost at τ^*. The two rankings can be reversed[21].

"The longer the history, the more useful the foundation model." The opposite holds: the longer the history, the better local methods can identify seasonality and trend, and the smaller the relative advantage of foundation models[22,39]. Foundation models are worth most for new products with only a few weeks of data.

"A stronger model will fix poor forecasts for stockout-prone items." If the training data are censored sales, any model will learn understated demand. Fix the data first, then change the model.

"Orchestrating several foundation models is always better." Orchestration gains are bounded by the complementarity among candidates, and a redundant candidate can erase them[65].

Limitations

This paper has five limitations. First, the illustrative simulations use synthetic data and simplified demand distributions, and the pooled prior is only a stand-in for a foundation model; it cannot represent the performance of any real model. Second, most foundation-model results cited are self-reported in papers or official materials, some in preprints, and may be subject to evaluation leakage and selective reporting[63]. Third, c_u and c_o are hard to estimate precisely in practice, and a misspecified τ^* affects our selection rule; we recommend sensitivity analysis over a range of τ. Fourth, we mainly discuss the single-period newsvendor framework and do not develop multi-period inventory, lead-time uncertainty or supply constraints. Fifth, data quality varies widely across FMCG companies, and the availability of stockout labels and stock records determines how far Algorithm 1 can be carried out.

Table 4 | Summary of illustrative simulations (fixed random seed 20261104)

Simulation	Key settings	Results	Equation
Cold-start shrinkage	600 SKUs, $\sigma_\eta = 0.3$, $\tau^* = 0.75$	$n_0 \approx 7.6$ weeks; cost ratio at 1 week: own history 1.56, prior 1.10, blend 1.09; own history beats prior after 12 weeks	Eqs (1), (5)
Accuracy versus service	300 intermittent SKUs, 90% target, 13 methods	Correlation between MASE rank and fill-rate rank −0.63; zero forecast has the lowest MASE and zero fill rate	Eq. (2)
Censoring repair	400 SKUs, about 34% stockout weeks	Mean bias −18.0% → +1.1%; achieved service at 90% target 82.4% → 90.2%	Eq. (3)

Table 5 | Five forecasts (views)

No.	Forecast	Time scale	Main basis	Falsification condition	Implication for enterprises
1	Major time-series benchmarks will add decision-oriented metrics (loss at specified quantiles, fill rate or inventory cost) reported alongside point error and CRPS	12–24 months	Accuracy–service divergence[21]; decision-focused learning[19,20]; benchmark expansion[14,15]	By end-2028, major benchmarks such as GIFT-Eval and fev-bench still report only point error and CRPS-type metrics	Use quantile loss at τ^* as the primary acceptance metric
2	Public retail datasets with stockout labels or stock records will become standard in demand-forecasting research, and censoring repair a routine step in retail evaluations of foundation models	12–24 months	FreshRetailNet-50K[23]; double censoring[27]	By end-2028, fewer than a third of newly released retail forecasting benchmarks include stockout information	Complete stockout and stock data before evaluating models
3	Enterprise deployments will mostly use hybrid architectures — foundation models as cold-start priors, statistical methods or boosted trees for mature SKUs — with wholesale replacement in the minority	6–18 months	Break-even analysis[22]; simulation study[39]; ceiling on orchestration gains[65]	Within a year, most public retail and FMCG case studies report wholesale replacement that beats hybrid architectures on service level	Decide the deployment footprint segment by segment by Eq. (6)
4	Structured input of future information — promotions, prices, campaigns — will become the main source of differentiation in retail applications of foundation models, with gains exceeding those from swapping backbones	12–24 months	Covariate models[8,48,49]; contextual-effects benchmark[52]	Controlled comparisons show that adding structured promotion information generally gains less than switching between comparable backbones	Prioritize data interfaces for promotion calendars and price plans
5	The weakness of foundation models on intermittent demand will be offset mainly by bias correction and quantile post-processing, not by larger pretraining	12–24 months	Instance-normalization bias correction[21]; forecast collapse[46]; black-box correction[58]	Within two years, a foundation model appears that, without post-processing, consistently beats Croston/SBA-type methods on service level for intermittent demand	Keep specialized methods and bias monitoring for long-tail SKUs

Conclusion

Time-series foundation models are changing the starting point of demand forecasting: new products, new stores and long-tail series have, for the first time, a training-free prior that yields a full distribution. But in the FMCG replenishment chain, their place is that of a prior, not the end point. Writing newsvendor cost as quantile loss at τ^* by equation (2) brings the selection criterion back from leaderboard error to business cost; repairing censoring by equation (3) lets any model learn true demand; and shrinking by history length with equation (5), while deciding the deployment footprint by segment with equation (6), allows the value of foundation models to be measured accurately.

On the implementation side, according to the company website, the ”new-product and regional demand forecasting” module of HX’s FMCG solution uses hierarchical time-series and machine-learning forecasts, deployed and refreshed on Deep Hope S1, and delivers store × SKU × week forecasts and replenishment suggestions[76]; Deep Hope S1 provides an on-premises environment for inference and training[77], and the Harness governance engine of the Enterprise Agent Platform executes tool calls through a controlled interface with the governance layer retaining the final say[78] — both able to carry the selection records, human

review and fallback of Algorithm 1. For enterprises, the transferable asset is not any single model but a forecasting process that selects by cost, deploys by segment and can fall back.

Key takeaways:

1. Select models with Eq. (2): newsvendor cost equals quantile loss at τ^*; the smallest error is not the lowest cost.
2. Repair data with Eq. (3): fix stockout censoring before backtesting and selection.
3. Manage the hierarchy with Eq. (4): reconcile bottom-level and aggregate forecasts to avoid amplification across levels.
4. Use the prior with Eq. (5): the foundation model's weight shrinks with history length.
5. Decide deployment with Eq. (6): use foundation models only on segments where cost savings exceed deployment cost.

Data availability

No new experimental data were generated. Figures quoted come from the listed references, official materials and the company website; the illustrative simulations were generated by scripts under the assumptions stated in the text.

Code availability

The script for the illustrative simulations and Figs 1–3 is src/figs.py (Python, fixed random seed 20261104), provided with the article package for verification.

Competing interests

The authors are employees of Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. The FMCG solution, Deep Hope S1 and the Enterprise Agent Platform's Harness mentioned in the text are publicly offered products and solutions of the company.

References

1. Ansari, A. F. et al. Chronos: Learning the Language of Time Series. *Trans. Mach. Learn. Res.* (2024); https://arxiv.org/abs/2403.07815.
2. Das, A., Kong, W., Sen, R. & Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Proc. 41st International Conference on Machine Learning* (PMLR, 2024); https://arxiv.org/abs/2310.10688.
3. Woo, G. et al. Unified Training of Universal Time Series Forecasting Transformers. In *Proc. 41st International Conference on Machine Learning* (PMLR, 2024); https://arxiv.org/abs/2402.02592.
4. Rasul, K. et al. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. Preprint at https://arxiv.org/abs/2310.08278 (2023).
5. Garza, A., Challu, C. & Mergenthaler-Canseco, M. TimeGPT-1. Preprint at https://arxiv.org/abs/2310.03589 (2023).
6. Shi, X. et al. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In *Proc. International Conference on Learning Representations* (2025); https://arxiv.org/abs/2409.16040.
7. Liu, Y. et al. Sundial: A Family of Highly Capable Time Series Foundation Models. In *Proc. 42nd International Conference on Machine Learning* (PMLR, 2025); https://arxiv.org/abs/2502.00816.
8. Ansari, A. F. et al. Chronos-2: From Univariate to Universal Forecasting. Preprint at https://arxiv.org/abs/2510.15821 (2025).
9. Liu, C. et al. Moirai 2.0: When Less Is More for Time Series Forecasting. Preprint at https://arxiv.org/abs/2511.11698 (2025).
10. Auer, A. et al. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. Preprint at https://arxiv.org/abs/2505.23719 (2025).
11. Khwaja, E. et al. Toto 2.0: Time Series Forecasting Enters the Scaling Era. Preprint at https://arxiv.org/abs/2605.20119 (2026).
12. Ansari, A. F. et al. Chronos-2: A foundation model for multivariate time series forecasting. Preprint at https://arxiv.org/abs/2610.41956 (2026).
13. Aksu, T. et al. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation. Preprint at https://arxiv.org/abs/2410.10393 (2024).
14. Shchur, O. et al. fev-bench: A Realistic Benchmark for Time Series Forecasting. Preprint at https://arxiv.org/abs/2509.26468 (2025).
15. Qiao, Z. et al. It's TIME: Towards the Next Generation of Time Series Forecasting Benchmarks. Preprint at https://arxiv.org/abs/2602.12147 (2026).
16. Fildes, R., Ma, S. & Kolassa, S. Retail forecasting: Research and practice. *International Journal of Forecasting* **38**, 1283–1318 (2022); https://doi.org/10.1016/j.ijforecast.2019.06.004.
17. Kolassa, S. Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting* **32**, 788–803 (2016); https://doi.org/10.1016/j.ijforecast.2015.12.004.
18. Arrow, K. J., Harris, T. & Marschak, J. Optimal Inventory Policy. *Econometrica* **19**, 250 (1951); https://doi.org/10.2307/1906813.
19. Bertsimas, D. & Kallus, N. From Predictive to Prescriptive Analytics. *Management Science* **66**, 1025–1044 (2020); https://doi.org/10.1287/mnsc.2018.3253.
20. Elmachtoub, A. N. & Grigas, P. Smart "Predict, then Optimize". *Management Science* **68**, 9–26 (2022); https://doi.org/10.1287/mnsc.2020.3922.
21. Chin, J. E., Cheng, S.-F. & Gunawan, A. Accuracy Is Not Service: A Decision-Aware Benchmark for Intermittent-Demand Forecasting. Preprint at https://arxiv.org/abs/2609.13840 (2026).
22. Jerome, N. T. & Simon, F. When Do Foundation Models Pay Off? A Break-Even Analysis of Pretrained Time Series Forecasters. Preprint at https://arxiv.org/abs/2607.04919 (2026).
23. Wang, Y. et al. FreshRetailNet-50K: A Stockout-Annotated Censored Demand Dataset for Latent Demand Recovery and Forecasting in Fresh Retail. Preprint at https://arxiv.org/abs/2505.16319 (2025).
24. Koenker, R. & Bassett, G. Regression Quantiles. *Econometrica* **46**, 33 (1978); https://doi.org/10.2307/1913643.
25. Gneiting, T. Quantiles as optimal point forecasts. *International Journal of Forecasting* **27**, 197–207 (2011); https://doi.org/10.1016/j.ijforecast.2009.12.015.
26. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007); https://doi.org/10.1198/016214506000001437.
27. Ye, Z., Wei, J., Song, T., Chen, C. & Ma, G. A Deep Learning Framework for Quantile Demand Forecasting in E-commerce with Doubly Censored Data. Preprint at https://arxiv.org/abs/2610.40138 (2026).
28. Wickramasuriya, S. L., Athanasopoulos, G. & Hyndman, R. J. Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association* **114**, 804–819 (2019); https://doi.org/10.1080/01621459.2018.1448825.
29. Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G. & Shang, H. L. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* **55**, 2579–2589 (2011);

https://doi.org/10.1016/j.csda.2011.03.006.
30. Athanasopoulos, G., Ahmed, R. A. & Hyndman, R. J. Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* **25**, 146–166 (2009); https://doi.org/10.1016/j.ijforecast.2008.07.004.
31. Panagiotelis, A., Gamakumara, P., Athanasopoulos, G. & Hyndman, R. J. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research* **306**, 693–706 (2023); https://doi.org/10.1016/j.ejor.2022.07.040.
32. Hyndman, R. J. & Athanasopoulos, G. *Forecasting: Principles and Practice* 3rd edn, Ch. 11 (OTexts, 2021); https://otexts.com/fpp3/hierarchical.html.
33. Wang, T., Johnson, M. C., Klee, S. & Malloy, M. L. Billions-Scale Forecast Reconciliation. Preprint at https://arxiv.org/abs/2602.05030 (2026).
34. Neubauer, L. & Filzmoser, P. Rediscovering Bottom-Up: Effective Forecasting in Temporal Hierarchies. Preprint at https://arxiv.org/abs/2407.02367 (2024).
35. Lee, H. L., Padmanabhan, V. & Whang, S. Information Distortion in a Supply Chain: The Bullwhip Effect. *Management Science* **43**, 546–558 (1997); https://doi.org/10.1287/mnsc.43.4.546.
36. Das, A., Faw, M., Sen, R. & Zhou, Y. In-Context Fine-Tuning for Time-Series Foundation Models. In *Proc. 42nd International Conference on Machine Learning* (PMLR, 2025); https://arxiv.org/abs/2410.24087.
37. Syntetos, A. A., Boylan, J. E. & Croston, J. D. On the categorization of demand patterns. *Journal of the Operational Research Society* **56**, 495–503 (2005); https://doi.org/10.1057/palgrave.jors.2601841.
38. Hoo, S. B., Müller, S., Salinas, D. & Hutter, F. From Tables to Time: Extending TabPFN-v2 to Time Series Forecasting. Preprint at https://arxiv.org/abs/2501.02945 (2025).
39. Ebrahim, E. K., Ali, S. M. & El-Kotory, A. Foundation or Formula? A Simulation-Based Comparison of Google's TimesFM-3 and Classical Time-Series Models. Preprint at https://arxiv.org/abs/2610.00589 (2026).
40. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**, 679–688 (2006); https://doi.org/10.1016/j.ijforecast.2006.03.001.
41. Croston, J. D. Forecasting and Stock Control for Intermittent Demands. *Journal of the Operational Research Society* **23**, 289–303 (1972); https://doi.org/10.1057/jors.1972.50.
42. Syntetos, A. A. & Boylan, J. E. The accuracy of intermittent demand estimates. *International Journal of Forecasting* **21**, 303–314 (2005); https://doi.org/10.1016/j.ijforecast.2004.10.001.
43. Turkmen, A. C., Wang, Y. & Januschowski, T. Intermittent Demand Forecasting with Deep Renewal Processes. Preprint at https://arxiv.org/abs/1911.10416 (2019).
44. Kislinskii, V. & Hameed, M. Are Gradient Boosting Models Suitable for Intermittent Demand Forecasting? Preprint at https://arxiv.org/abs/2609.14718 (2026).
45. Wang, S., Kang, Y. & Petropoulos, F. Combining Probabilistic Forecasts of Intermittent Demand. Preprint at https://arxiv.org/abs/2304.03092 (2023).
46. Wan, S. et al. Forecast Collapse in Time-Series Foundation Models. Preprint at https://arxiv.org/abs/2608.14106 (2026).
47. Qin, G. et al. CoRA: Covariate-Aware Adaptation of Time Series Foundation Models. Preprint at https://arxiv.org/abs/2510.12681 (2025).
48. Arango, S. P. et al. ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables. In *Proc. 28th International Conference on Artificial Intelligence and Statistics* (PMLR, 2025); https://arxiv.org/abs/2503.12107.
49. Auer, A. et al. Zero-Shot Time Series Forecasting with Covariates via In-Context Learning. Preprint at https://arxiv.org/abs/2506.03128 (2025).
50. Han, L. et al. UniCA: Unified Covariate Adaptation for Time Series Foundation Model. Preprint at https://arxiv.org/abs/2506.22039 (2025).
51. Berthelier, G. et al. Investigating simple target-covariate relationships for Chronos-2 and TabPFN-TS. Preprint at https://arxiv.org/abs/2605.12200 (2026).
52. Liu, X. et al. GIFT-Eval-CTX: A Benchmark For Contextual Forecasting Model Evaluation. Preprint at https://arxiv.org/abs/2610.17794 (2026).
53. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting* **38**, 1346–1364 (2022); https://doi.org/10.1016/j.ijforecast.2021.11.013.
54. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* **30** (2017); https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.
55. Makridakis, S. et al. The M5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting* **38**, 1365–1385 (2022); https://doi.org/10.1016/j.ijforecast.2021.10.009.
56. Nasios, I. & Vogklis, K. Blending gradient boosted trees and neural networks for point and probabilistic forecasting of hierarchical time series. *International Journal of Forecasting* **38**, 1448–1459 (2022); https://doi.org/10.1016/j.ijforecast.2022.01.001.
57. Lee, T. L., Toner, W., Singh, R., Joosen, A. & Asenov, M. Lightweight Online Adaption for Time Series Foundation Model Forecasts. Preprint at https://arxiv.org/abs/2502.12920 (2025).
58. Zhang, B. et al. ECHO: Correcting Time Series Foundation Model Forecasts with Historical Analogues and Errors. Preprint at https://arxiv.org/abs/2610.16433 (2026).
59. Le, H., Nguyen, M. H., Nguyen, M., Nguyen, H. H. & Do, D. Expert-Guided Forecast Editing for Time-Series Foundation Models. Preprint at https://arxiv.org/abs/2607.19659 (2026).
60. Karaouli, N., Coquenet, D., Fromont, E., Mermillod, M. & Reyboz, M. Foundation vs. Specialized Models: Evaluating Catastrophic Forgetting in Continual Time Series Forecasting. Preprint at https://arxiv.org/abs/2510.00809 (2025).
61. Xu, C. & Xie, Y. Conformal prediction for time series. In *Proc. 38th International Conference on Machine Learning* (PMLR, 2021); https://arxiv.org/abs/2010.09107.
62. Gibbs, I. & Candès, E. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems* **34** (2021); https://arxiv.org/abs/2106.00170.
63. Meyer, M., Kaltenpoth, S., Zalipski, K. & Müller, O. Rethinking Evaluation in the Era of Time Series Foundation Models: (Un)known Information Leakage Challenges. Preprint at https://arxiv.org/abs/2510.13654 (2025).
64. Haugland, V. et al. IND-Eval: Aligning Evaluation and Simulation for Industrial Time Series Foundation Models. Preprint at https://arxiv.org/abs/2610.16652 (2026).
65. Guan, T., Zhan, T., Pan, S. & Jin, M. Rethinking Agentic Model Orchestration for Time Series Forecasting. Preprint at https://arxiv.org/abs/2610.29504 (2026).
66. Ansari, A. F., Turkmen, C., Shchur, O. & Stella, L. Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon. *AWS Machine Learning Blog* https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/ (2024).
67. Amazon. amazon/chronos-2 (model card). *Hugging Face* https://huggingface.co/amazon/chronos-2 (2025).
68. Google Research. timesfm-2.5-200m-pytorch (model card). *Hugging Face* https://huggingface.co/google/timesfm-2.5-200m-pytorch (2025).
69. Hyndman, R. J., Koehler, A. B., Snyder, R. D. & Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18**, 439–454 (2002); https://doi.org/10.1016/S0169-2070(01)00110-8.
70. Assimakopoulos, V. & Nikolopoulos, K. The theta model: a decomposition approach to forecasting. *International Journal of Forecasting* **16**, 521–530 (2000);

https://doi.org/10.1016/S0169-2070(00)00066-2.
71. Salinas, D., Flunkert, V., Gasthaus, J. & Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* **36**, 1181–1191 (2020); https://doi.org/10.1016/j.ijforecast.2019.07.001.
72. Lim, B., Arık, S. Ö., Loeff, N. & Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* **37**, 1748–1764 (2021); https://doi.org/10.1016/j.ijforecast.2021.03.012.
73. Oreshkin, B. N., Carpov, D., Chapados, N. & Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *Proc. International Conference on Learning Representations* (2020); https://arxiv.org/abs/1905.10437.
74. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **36**, 54–74 (2020); https://doi.org/10.1016/j.ijforecast.2019.04.014.
75. Petropoulos, F. et al. Forecasting: theory and practice. *International Journal of Forecasting* **38**, 705–871 (2022); https://doi.org/10.1016/j.ijforecast.2021.11.001.
76. 弘犀智能（Great Unicorn Intelligence）. 快消与营销解决方案（FMCG & marketing solutions）. https://www.hxallai.com/solutions/fmcg/ (2026).
77. 弘犀智能（Great Unicorn Intelligence）. Deep Hope S1 · 推训一体大模型平台（Deep Hope S1 · unified inference–training LLM platform）. https://www.hxallai.com/platform/deep-hope/ (2026).
78. 弘犀智能（Great Unicorn Intelligence）. 企业智能体平台 · Harness（Enterprise Agent Platform · Harness）. https://www.hxallai.com/platform/agent/ (2026).

Copyright Notice: © 2026 Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. All rights reserved. Any reposting, quotation in commentary, or adaptation must credit the authors (Great Unicorn Intelligence Research Team) and cite the source (Great Unicorn Intelligence, https://www.hxallai.com).