In-Context Forecasting in Supply Chains: Evaluating the Promise and Limits of Tabular Foundation Models

Ole Behre

May 24, 2026

Abstract

Tabular Prior-Data Fitted Networks (TabPFN) have emerged as a promising solution for demand forecasting in data-scarce supply chains, offering zero-shot inference via in-context learning without dataset-specific training. This paper critically analyzes four approaches to their application: univariate temporal adaptation (TabPFN-TS), multivariate extension via table flattening (TabPFN-TS-MV), enterprise deployment of SAP's RPT-1, and architectural correction via ApolloPFN. Across all four, a consistent pattern emerges. TabPFN-TS acts as a strong conditional interpolator but fails to extrapolate trends, creating major issues for cold-start and NPI scenarios. The multivariate extension is severely constrained by a context window that fills linearly with the channel count. The enterprise evaluation highlights a regression gap and deployment hurdles that standard benchmarks miss. Finally, ApolloPFN demonstrates that these failures are architectural rather than fundamental, but fixing them requires adopting temporal inductive biases that blur the line between tabular and dedicated time-series foundation models. We conclude that TFMs occupy a specific but narrow niche, and their practical value in supply chains currently relies on combination with classical methods.

Contents

1 Introduction

Accurate demand forecasting is essential for effective Supply Chain Management (SCM). Forecast accuracy directly impacts the optimization of safety stock, the scheduling of manufacturing runs, and the orchestration of global logistics networks (Fildes et al. 2009). However, in practice, supply chains frequently operate in volatile environments characterized by severe data scarcity (Makridakis et al. 2022). High-value operational scenarios, such as New Product Introductions (NPI), the management of slow-moving capital goods, and highly erratic promotional periods, yield sparse datasets that contradict the stationary conditions assumed by many predictive models.

While advanced machine learning (ML) and deep learning (DL) algorithms have shown great potential in time-series forecasting, their industrial application remains difficult on typical tabular supply chain data. Empirically, small to medium tabular datasets represent the vast majority of traditional machine learning benchmarks (Dua and Karra Taniskidou 2017) and remain everywhere across real-world enterprise applications. In this low-data domain, traditional deep learning architectures are limited; benchmark studies demonstrate that tree-based ensembles, such as Gradient Boosted Decision Trees (GBDTs) (e.g., XGBoost, CatBoost), consistently outperform classical neural networks on medium and small tabular datasets (Grinsztajn et al. 2022). This performance gap persists because neural networks lack proper inductive biases for tabular geometries, struggling to ignore uninformative features and learn irregular target functions when data is scarce (Grinsztajn et al. 2022).

Given these limitations, Tabular Foundation Models (TFMs) have emerged as a compelling alternative. Architectures such as Tabular Prior-Data Fitted Networks (TabPFN) are trained offline once to approximate Bayesian inference on millions of synthetic tabular tasks generated via Structural Causal Models (Hollmann et al. 2023, 2025; Grinsztajn et al. 2026b). During inference, TabPFN utilizes In-Context Learning (ICL) to execute zero-shot predictions in a single forward pass without dataset-specific gradient updates or hyperparameter optimization (Hollmann et al. 2025; Hoo et al. 2026). This makes TFMs an interesting candidate for supply chain scenarios with limited historical data.

However, their application to dynamic, sequential forecasting processes introduces significant operational hurdles that the existing literature does not always fully address. This paper is guided by the following central research question:

> *What are the methodological strengths, limitations, and shared challenges of applying Tabular Foundation Models to demand forecasting in data-scarce supply chains, specifically regarding time-series adaptation, enterprise data integration, and architectural adequacy?*

The remainder of the paper is organized as follows: Section 2 provides foundational background on data scarcity in SCM, the mechanics of TFMs, and common evaluation metrics. Section 3 explains the scope and selection of the analyzed literature. Section 4 systematically surveys four methodological approaches to adapting tabular models for forecasting. Section 5 synthesizes these findings into a comparative discussion. Finally, Section 6 concludes this paper and proposes directions for future research.

2 Background

2.1 Demand Forecasting and the Data Scarcity Challenge

Demand forecasting is a critical first step in SCM, affecting the entire supply network. While high-volume products with stable sales histories are easily modeled, modern industrial catalogs consist largely of intermittent or newly introduced items. Traditionally, intermittent demand has been modeled using specialized heuristics like the Croston method (Croston 1972) and its bias-correcting derivatives, such as the Syntetos-Boylan Approximation (SBA) (Syntetos and Boylan 2005). However, these classical statistical methods provide rigid deterministic point forecasts and often fail to capture complex, nonlinear exogenous factors.

While advanced ML techniques are increasingly used to capture these complexities, their industrial application is frequently bottlenecked by data privacy and severe data scarcity. For New Product Introductions (NPI), the total absence of historical sales creates a "cold-start" problem, rendering standard autoregressive sequence models highly ineffective. Even for existing slow-moving items, the sparsity of non-zero observations prevents gradient-based deep learning algorithms from establishing a stable learning signal.

2.2 The Mechanics of Tabular Foundation Models

A key characteristic of recent Tabular Foundation Models (TFMs) is their departure from iterative, dataset-specific training. While traditional ML models use gradient descent to minimize a loss function on a specific dataset, TFMs are pre-trained offline on vast amounts of data and leverage In-Context Learning (ICL) during inference.

Architectures differ significantly in how they achieve this. While some TFMs pre-train on massive repositories of real-world tabular data (Wang and Sun 2022), others utilize the Prior-Data Fitted Network (PFN) framework to approximate Bayesian inference (Müller et al. 2024). TabPFN (Hollmann et al. 2023), a prominent example of the latter, relies entirely on synthetic pre-training. It is exposed to millions of synthetic tabular datasets generated from a prior distribution that incorporates principles of causal learning. TabPFN's prior prefers simple explanations, encoding a form of Occam's razor. By learning universal geometric and causal relationships within these synthetic structures, the Transformer-based architecture can generalize to unseen, real-world tabular data.

During deployment, TFMs like TabPFN utilize ICL by ingesting a support set (the historical training data) alongside a query set (the target features to be predicted) as a single input. The model then outputs predictions in a single forward pass, requiring no hyperparameter tuning. This native probabilistic output is particularly attractive for practitioners, as it directly supports uncertainty-aware safety stock calculations without requiring wrapper techniques.

2.3 Architectural Constraints

While In-Context Learning (ICL) provides extreme sample efficiency, it introduces strict structural bottlenecks. Because transformer attention scales quadratically with input size, these architectures impose hard limits on the

volume of data that can be processed simultaneously. The original TabPFN architecture was strictly constrained to datasets with a context window of up to 1,000 training rows and 100 purely numerical features, showing notably degraded performance on categorical inputs.

Model / Publication	Max Rows	Max Features
Original PFNs (Müller et al. 2024)	∼100	∼60
TabPFN v1 (Hollmann et al. 2023)	1,000	100
TabPFN v2 (Hollmann et al. 2025)	10,000	500
TabPFN-2.5 (Grinsztajn et al. 2026a)	50,000	2,000
TabPFN-3 (Grinsztajn et al. 2026b)	1,000,000	2,000

Table 1: Evolution of context size constraints in PFN-based tabular architectures.

Newer iterations have pushed these boundaries significantly (see Table 1). The second iteration (TabPFN v2) scaled the context window to 10,000 rows and 500 features while introducing categorical support. Most recently, technical reports detailing TabPFN-2.5 and TabPFN-3 have demonstrated context limits reaching up to 1,000,000 rows through advanced context optimization. Despite these advancements, the context window remains a fundamental bottleneck for all transformer-based ICL approaches reviewed here.

2.4 Evaluation Datasets and Metrics

The surveyed literature evaluates TFM architectures across various datasets, ranging from controlled benchmarks to simulated enterprise environments. Establishing these baselines is critical for interpreting the results.

2.4.1 Benchmark Datasets

- **GIFT-Eval (Aksu et al. 2024):** A comprehensive benchmarking suite for zero-shot time-series forecasting comprising 23 datasets and over 144,000 individual time series across seven application domains. Created by researchers from Salesforce AI.
- **FEV-Bench (Shchur et al. 2026):** A benchmark designed to evaluate covariate-informed forecasting. It tests whether models can condition on future-known exogenous variables such as promotional calendars or scheduled events. Created by AWS researchers.
- **M-Competition Forecasting Benchmarks:** A long-running series of forecasting competitions that have served as standard benchmarks in the field for decades. M1 (Makridakis and Hibon 1979), M3 (Makridakis and Hibon 2000), and M4 (Makridakis et al. 2020) cover univariate time series across varying frequencies from diverse domains. M5 (Makridakis et al. 2022) is the most relevant for supply chain applications: drawn from Walmart's retail operations, it covers 30,490 product-level time series with highly intermittent demand and rich exogenous variables including price and calendar effects.

- **Day-Ahead Electricity Price Benchmark (Lago et al. 2021):** A benchmark of day-ahead electricity prices across five major European power markets (Nord Pool, EPEX-BE, EPEX-FR, EPEX-DE, and PJM). The paper itself is a review of state-of-the-art forecasting algorithms accompanied by an open-access dataset spanning multiple years of hourly price data per market, with exogenous variables including grid load and renewable generation forecasts. It is commonly used to evaluate covariate-aware forecasters.
- **Enterprise ERP Data (SAP):** Simulated records mimicking Enterprise Resource Planning (ERP) systems by Lal (2026), focusing on transactional records from Sales and Distribution (SD), Materials Management (MM), Production Planning (PP), and Financial Accounting and Controlling (FI/CO) modules. Unlike the academic benchmarks above, this data mirrors proprietary dimensions and categorical cardinalities.

2.4.2 Performance Metrics

The surveyed papers report a mix of regression, classification, and probabilistic forecasting metrics. We define here only those that actually appear in the results sections of Section 4.

Mean Absolute Scaled Error (MASE). A scale-invariant point-forecast metric defined as

$$\text{MASE} = \frac{\frac{1}{h}\sum_{t=T+1}^{T+h} |y_t - \hat{y}_t|}{\frac{1}{T-m}\sum_{t=m+1}^{T} |y_t - y_{t-m}|},$$

where the denominator is the mean absolute error of a seasonal-naive forecast with period m. MASE < 1 means the model beats the seasonal-naive baseline. The scaling makes results comparable across series of different magnitudes.

Weighted Quantile Loss (WQL). A probabilistic metric evaluating predicted quantiles $\hat{q}_\tau$ at levels $\tau \in (0, 1)$:

$$\text{QL}_\tau(y, \hat{q}_\tau) = 2 \cdot \big[\tau \cdot (y - \hat{q}_\tau)_+ + (1 - \tau) \cdot (\hat{q}_\tau - y)_+\big],$$

averaged over a set of quantile levels (typically nine levels from 0.1 to 0.9) and normalized by the sum of absolute targets. Lower is better. Unlike MASE, WQL rewards well-calibrated predictive distributions, not just accurate point forecasts.

Coefficient of Determination (R^2). Standard regression metric:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where $\bar{y}$ is the target mean. $R^2 = 1$ is perfect prediction, $R^2 = 0$ matches the mean baseline, and negative values are worse than predicting the mean. Used by Lal (2026) for the demand and payment-timing regression tasks.

Area Under the ROC Curve (AUC-ROC). For binary classification, the probability that a randomly chosen positive example receives a higher predicted score than a randomly chosen negative example. AUC = 0.5 is random, AUC = 1.0 is perfect ranking. Threshold-independent, which is convenient for tasks like stockout or anomaly classification where the operational decision threshold is set separately.

Scaled Continuous Ranked Probability Score (sCRPS). A probabilistic metric generalizing MAE to predictive distributions:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 dz,$$

where F is the predicted CDF. CRPS reduces to MAE for point forecasts, and rewards both sharpness and calibration. The scaled variant normalizes by the magnitude of the target series, making it comparable across datasets.

3 Scope

The application of Tabular Foundation Models to time-series and demand forecasting is a growing research area. This paper restricts its scope to a curated selection of literature addressing the practical operational challenges of deploying TFMs in supply chains. We also briefly compare these to dedicated Time-Series Foundation Models (TSFMs) such as Chronos (Ansari et al. 2024), Moirai (Woo et al. 2024), and TimesFM (Das et al. 2024) where relevant in benchmarks. However, a full review of the TSFM landscape is outside the scope of this paper. The focus here remains on the tabular paradigm. The four selected papers represent a progressive methodological build-up:

1. **Univariate Temporal Adaptation (Hoo et al. 2026):** Shows how a static tabular architecture (TabPFN) can be adapted for time-series forecasting via temporal featurization.
2. **Multivariate Extension (Jayawardhana et al. 2026):** Extends TabPFN-TS to multi-channel supply chain data by flattening multivariate inputs into a single table.
3. **Enterprise Deployment (Lal 2026):** Evaluates SAP's Relational Pre-trained Transformer (RPT-1) against GBDTs on synthetic ERP data, providing a deployment-oriented perspective.
4. **Architectural Correction (Potapczynski et al. 2026):** Characterizes the structural failures of TabPFN-TS and proposes ApolloPFN, a time-aware PFN with native covariate support.

4 Methodological Perspectives

4.1 Univariate Temporal Adaptation: Reframing Time-Series as Tabular Regression

4.1.1 Approach

Hoo et al. (2026) establish the foundational mechanism for adapting static TFMs to time-series forecasting. Their contribution, TabPFN-TS, converts a time-series sequence into a tabular regression dataset through a temporal featurization scheme, enabling TabPFN-v2 (Hollmann et al. 2025) to produce forecasts without time-series-specific pretraining.

The featurization operates in three layers. First, a scalar index $\Phi_{\text{index}}(t) = t$ encodes the global temporal position. Second, eight cyclic calendar components (e.g., hour of day, day of week) are encoded as sin/cos pairs alongside the calendar year, producing a 17-dimensional vector $\Phi_{\text{cal}}(t)$. Third, domain-specific seasonal patterns are captured automatically via spectral analysis, identifying the k most prominent periodic components as $\Phi_{\text{auto}}(t)$. The full feature vector per time step concatenates all three layers with any known covariates $\mathbf{z}_t$. This yields a standard supervised regression dataset that TabPFN-v2 ingests as context to produce a predictive distribution in a single forward pass.

4.1.2 Results

Hoo et al. (2026) evaluate TabPFN-TS on the 97 tasks of the GIFT-Eval benchmark (Aksu et al. 2024). TabPFN-TS achieves a mean WQL rank of 5.39 and a relative MASE of 0.692, placing it fourth among thirteen evaluated models. This is competitive with substantially larger time-series foundation models despite having only 11M parameters. A particularly strong result emerges on FEV-Bench (Shchur et al. 2026), where covariates are known over the forecast horizon. Here, TabPFN-TS achieves the best performance of all evaluated models (relative WQL 0.503, relative MASE 0.666). This advantage is structural: TabPFN-TS treats future-known covariates as ordinary tabular columns, whereas dedicated time-series architectures lack native mechanisms for this kind of conditioning.

4.1.3 Discussion

TabPFN-TS demonstrates that a tabular foundation model can produce a competitive time-series forecaster without gradient updates on real temporal data. Two limitations require attention.

The first concerns how the authors interpret "zero-shot". In NLP, zero-shot typically means the model receives no examples of the target task. TabPFN operates differently: at inference time it receives the full historical training set as a support context, from which it conditions its predictions in a single forward pass. This is more accurately described as in-context learning, and the historical data essentially acts as a "shot". While the absence of gradient-based optimization on the target dataset is a meaningful distinction from standard supervised learning, the model still requires exposure to the target domain's history. This might seem like a terminological technicality, but precise wording

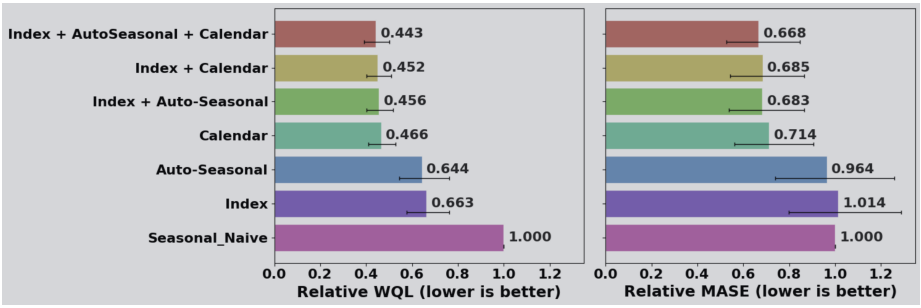

Figure 1: Ablation of temporal featurization components in TabPFN-TS. Relative WQL (left) and MASE (right) show that all components contribute meaningfully. Scores are relative to the `Seasonal_Naive` baseline. Image taken directly from Hoo et al. (2026).

matters here because "zero-shot" claims drive a lot of the current excitement around TFMs and obscure what they actually require at inference time.

Furthermore, the approach requires substantive preprocessing decisions: handling missing values, selecting k seasonal components, and truncating long contexts. As shown in Figure 1, removing calendar or automatic seasonal features visibly degrades performance.

The second limitation is more critical for supply chain applications. Hoo et al. (2026) show that TabPFN-TS struggles to extrapolate trends when targets fall outside the range of the conditioning set. For both linear and exponential growth trajectories, predictions tend to collapse toward the boundary of observed values rather than continuing the trend. The root cause is structural: during pretraining, context and query sets are sampled from the same generating distribution, so query targets almost always fall within the range already observed. The model is optimized purely for interpolation, and there is no pretraining signal for extrapolation.

For NPI and cold-start scenarios, this is a significant flaw. A newly introduced product naturally grows beyond historically observed levels. A model unable to extrapolate will systematically underestimate demand during growth phases and overestimate it during decline, producing directionally persistent errors in safety stock calculations. Even when informative trend covariates are provided, the interpolation barrier sits upstream of the covariate integration mechanism entirely.

4.2 Multivariate Extension: Unrolling Cross-Channel Dependencies

4.2.1 Approach

Jayawardhana et al. (2026) present a straightforward approach to extend TabPFN-TS to multivariate forecasting through the TabPFN-TS-MV framework. Rather than modifying the architecture, they serialize the multivariate structure into a tabular input format.

Given a multivariate time series with d channels, each time step is expanded into d rows. Each row carries the temporal features, an integer channel in-

dicator $\eta \in \{1, \ldots, d\}$, and the observed value. The context window therefore expands by a factor of d. Before flattening, each channel is independently z-score normalized to handle different scales. An ablation confirms that first-order differencing as an alternative strategy degrades performance (Jayawardhana et al. 2026). The central trade-off is obvious: context length scales linearly with the number of channels, quickly running into TabPFN-v2's limits.

4.2.2 Results

Across the multivariate subset of GIFT-Eval (Aksu et al. 2024), TabPFN-TS-MV achieves an average MASE of 1.2032 against 1.2139 for the original TabPFN-TS, a marginal improvement of roughly 0.9%, and outperforms the channel-independent baseline on 60% of the evaluated datasets. On the probabilistic WQL metric, it performs slightly worse (0.1558 vs. 0.1514). Standard deviations across random seeds are omitted, which prevents a proper statistical significance assessment.

In a comparison against state-of-the-art models, Chronos-2 (Ansari et al. 2024) achieves an average MASE of 6.81, but this figure is driven almost entirely by the Jena Weather dataset. Because Jena Weather has 21 variates, Chronos-2 collapses under the channel count. The authors report both the full average and an average excluding Jena, which is methodologically transparent. With Jena excluded, Chronos-2's MASE drops to 1.04, outperforming both TabPFN-TS-MV (1.24) and TabPFN-TS (1.27), indicating that Chronos-2 remains highly competitive when extreme high-variate edge cases are excluded.

4.2.3 Discussion

The most critical issue here is the baseline comparison. TabPFN-TS-MV combines data from all d channels into a single flattened context of d rows. Since this multiplies the row count by d, TabPFN-TS-MV can only look back $1/d$ as far for the same context budget. The authors compensate by artificially restricting the TabPFN-TS baseline to the same reduced lookback window. While this makes the comparison controlled, the MASE delta of 0.011 is best interpreted merely as an upper bound on the benefit of cross-channel modeling.

A more fundamental concern is scalability. A supply chain scenario with 50 Stock Keeping Units (SKUs) and 52 weeks of history produces a flattened table of 2,600 rows, straining the model's limits. The authors note that newer models like TabPFN 3 (Grinsztajn et al. 2026b) support much larger context windows, but they do not test them. Thus, the scalability argument relies entirely on future capabilities. This contribution is best understood as a proof of concept rather than a production-ready framework.

4.3 Enterprise Deployment: In-Context Learning on Synthetic ERP Data

4.3.1 Approach

Lal (2026) present an evaluation of SAP's Relational Pretrained Transformer (RPT-1) on enterprise business process prediction tasks. RPT-1 is built on the ConTextTab architecture Spinaci et al. (2025), pretrained on a large corpus of real-world tables. It processes tabular data through semantic embeddings for

text, date, and numeric columns, using alternating cross-column and cross-row attention.

Three scenarios are evaluated: demand forecasting from Sales and Distribution (SD) and Materials Management (MM) data, anomaly classification from master data, and payment risk scoring from simulated Financial Accounting and Controlling (FI/CO) data. All datasets are synthetically generated to mirror SAP table schemas and statistical distributions. RPT-1-OSS is benchmarked against tuned XGBoost, LightGBM, and CatBoost baselines under 5-fold cross-validation.

4.3.2 Results

On classification tasks, RPT-1-OSS achieves an AUC-ROC of 0.912 on the integrity task, trailing the best GBDT by 3.6 percentage points. The gap is larger on regression: RPT-1-OSS achieves an R^2 of 0.781 on demand forecasting versus XGBoost's 0.892. A context size ablation reveals that at 50 context rows, RPT-1-OSS actually outperforms all GBDTs by 3.7 percentage points. GBDTs overtake it at approximately 75–100 rows as they gain sufficient training signal, and the gap then widens monotonically: from $\Delta = -0.015$ AUC at 100 rows to $\Delta = 0.036$ at 2,048 rows.

4.3.3 Discussion

The main limitation of this evaluation is its exclusive reliance on synthetic data. Lal (2026) acknowledge that production SAP data may differ due to complex feature interactions and data quality issues. The synthetic datasets mirror SAP schemas and statistical distributions, but they necessarily lack the irregularities of live ERP environments, such as schema variation across installations or missing value patterns tied to operational events. The results should therefore be read as an upper bound on deployment performance.

Before drawing comparisons to the TabPFN-based approaches above, it is worth noting how RPT-1 and TabPFN-v2 relate. The two share the core paradigm of in-context learning over tabular data, but differ sharply in their pretraining strategy. TabPFN-v2 (Hollmann et al. 2025) is pretrained on synthetic datasets generated from Structural Causal Models, learning a prior over abstract data-generating processes with no awareness of column semantics. RPT-1, built on ConTextTab (Spinaci et al. 2025), is pretrained on a real-world table corpus and embeds column names and categorical values through a sentence transformer, processing labels such as "Material Group" or "Payment Terms" as semantic signals. RPT-1 thus serves as an enterprise-oriented data point for the same paradigm, with the semantic embedding layer representing a design choice particularly suited to ERP data where column names carry domain meaning.

The crossover finding is the most useful practical result in the paper. RPT-1-OSS outperforms tuned XGBoost at very low sample sizes but is overtaken around 100 rows, with the gap widening monotonically afterward. This supports the observation that TFMs are most valuable in extreme data-scarce regimes where GBDT hyperparameter optimization lacks the signal to function. The specific crossover point is dataset-dependent and should not be read as a universal threshold, but the directional pattern is informative for practitioners deciding when ICL-based screening is preferable to a trained GBDT pipeline.

The regression gap on the demand forecasting task is also noteworthy. An 11 percentage point R^2 deficit is substantial and perfectly aligns with the broader difficulty in-context learners exhibit when predicting targets outside the provided context range, as seen in TabPFN-TS. Lal (2026) do not investigate the architectural drivers of this gap, leaving open whether the issue stems from RPT-1's regression head, the semantic embedding strategy, or the same i.i.d. pretraining structure that limits TabPFN-TS.

Furthermore, Lal (2026) highlight severe deployment constraints: inference with 2,048 context rows requires approximately 28 seconds on a single GPU (12–80 GB VRAM), challenging real-time processing. More importantly, because RPT-1 requires context examples at inference time, sensitive business data must be constantly transmitted to the model endpoint, introducing major data governance hurdles that do not apply to locally deployed GBDTs. While these figures are specific to RPT-1-OSS, the underlying constraints (GPU memory for attention over context and the need to ship context data at inference time) are structural to the ICL paradigm and apply in varying degrees to all tabular foundation models reviewed in this paper.

4.4 Architectural Correction: Fixing the Inductive Bias of Tabular PFNs

4.4.1 Approach

Potapczynski et al. (2026) introduce ApolloPFN, a PFN that extends the TabPFN paradigm with time series-specific inductive biases while retaining native covariate support. The authors trace the failures of TabPFN-TS to a single architectural root cause: TabPFN's sample attention mechanism is permutation invariant with respect to time. This is a deliberate design choice for the i.i.d. tabular setting, where the order of rows is meaningless. In a time series context, however, order defines the structure of the prediction problem. Temporal features are imposed post hoc on top of this order-invariant architecture, but they do not inherently compensate for the absence of temporal inductive biases at the attention level.

Section 3 of Potapczynski et al. (2026) describes five concrete failure modes that follow from this root cause:

- **Reliance on manual frequency features:** Without explicit frequency inputs, TabPFN-TS simply predicts the mean of past values. Even when frequency features are provided, it only captures patterns that match them and misses the rest, such as the largest spikes.
- **Inability to represent ordered patterns:** Patterns that play out over several steps, like a slow ramp-up to a promotion and a sharp drop after, cannot be learned. The model treats earlier and later rows as interchangeable, so it has no way to connect them.
- **Weak trend extrapolation:** The same problem already raised in Section 4.1. Without a sense of order, the model cannot project a trend forward.

- **Lack of recency bias:** All past observations receive equal weight, causing the model to adapt slowly to shifts in the data and rely on outdated context.
- **Generic confidence intervals:** The predicted intervals tend to cover the entire historical variance rather than reflecting the localized uncertainty of the current trend.

ApolloPFN addresses these failures through two complementary interventions. On the data generation side, the synthetic data generator is replaced with a Single Root Node Random Growing Network (SRNGN), which produces graphs with longer causal paths and a single root node. Crucially, root node values are sampled from a stochastic combination of sine and cosine functions with randomly drawn frequencies and amplitudes, rather than independently for each observation. This introduces genuine temporal autocorrelation into the synthetic pretraining distribution. On the architectural side, Rotary Position Embeddings (RoPE) (Su et al. 2023) are incorporated into the sample attention mechanism, encoding relative positional distance directly into the attention computation. The attention mask is additionally relaxed to allow future test points to attend to each other, enabling the model to exploit future-known covariate information across the full forecast horizon simultaneously.

4.4.2 Results

On the day-ahead electricity price benchmark (Lago et al. 2021), ApolloPFN achieves a 12% improvement in sCRPS over TabPFN-TS. On M5, ApolloPFN matches or beats TabPFN-TS on most aggregation levels. On the classical M-competition benchmarks without covariates (M1, M3, M4, and Tourism (Hyndman et al. 2008)), ApolloPFN outperforms TabPFN-TS by 10% on average while retaining the same 11M parameter count, remaining competitive with much larger models like Chronos-Large.

4.4.3 Discussion

ApolloPFN successfully retains the structural advantage of TabPFN-TS (handling future-known covariates seamlessly) while adding the temporal awareness it lacked. The failure modes documented in Section 3 of Potapczynski et al. (2026) confirm that the trend extrapolation and ordered pattern challenges share a common architectural root, and ApolloPFN demonstrates that these failures are tied to specific choices made in TabPFN's pretraining and architecture rather than to the in-context learning paradigm itself. Replacing the i.i.d. pretraining prior with a temporally structured one, and adding positional encodings, recovers a meaningful fraction of performance.

However, limitations remain. On M5, the authors do not compare against the LightGBM-based solutions that actually dominated the original competition; the baselines reported are other foundation models (TabPFN-TS, Moirai-Large, Chronos-Large) rather than the tree-based methods that set the bar in the competition itself. Furthermore, Potapczynski et al. (2026) acknowledge that the standard quadratic attention bottleneck still limits applicability to very long time series.

The main takeaway is that fixing TabPFN requires a departure from the pure tabular formulation. By pretraining on synthetically generated time series rather than generic tabular data, ApolloPFN blurs the boundary between tabular and time-series foundation models.

5 Discussion

5.1 Evaluation and Comparability

While all four papers benchmark against established baselines, their results are not directly comparable. Hoo et al. (2026), Jayawardhana et al. (2026), and Potapczynski et al. (2026) use overlapping but non-identical subsets of public benchmarks, while Lal (2026) assesses whether an off-the-shelf TFM is competitive with tuned GBDTs on simulated ERP data. Direct cross-paper comparison of absolute numbers is therefore not particularly meaningful.

Reporting choices also vary significantly. Jayawardhana et al. (2026) omit standard deviations, which prevents a statistical significance assessment for the small MASE improvements they report. Potapczynski et al. (2026) exclude the LightGBM-based solutions that dominated the original M5 competition from their baseline set. Lal (2026) evaluate strictly on synthetic data and frame their results as an upper bound on deployment performance. None of these are disqualifying, but together they reinforce that conclusions about TFM performance in supply chain settings should be interpreted directionally rather than precisely.

5.2 TFMs vs. Time-Series Foundation Models

The benchmark results suggest that dedicated TSFMs (like Chronos-2) are consistently competitive with TFM-based approaches on standard time-series tasks. However, TFMs retain a clear structural advantage in their native handling of future-known covariates as tabular columns. This advantage is clearly demonstrated by TabPFN-TS on FEV-Bench and by ApolloPFN on the M5 and electricity price benchmarks. ApolloPFN complicates this picture further: by introducing temporal synthetic data and positional encodings, it absorbs the core inductive biases of TSFMs into the PFN framework, making the distinction between the two paradigms increasingly fluid.

5.3 Preprocessing Constraints and Inductive Bias

Despite representing different approaches, a recurring observation is that TFMs still require substantial preprocessing. Temporal featurization, Fourier transforms, z-score normalization, channel indicator variables, and multivariate unrolling all constitute feature engineering; the work is simply shifted from model training to data preprocessing. ApolloPFN's approach of absorbing seasonal structure into the pretraining prior highlights this: TabPFN-TS's dependence on manually injected frequency features is precisely the symptom ApolloPFN tries to eliminate. The reframing does carry genuine advantages, as preprocessing logic is explicit and deterministic, but TFMs shift the engineering burden rather than eliminating it.

5.4 Industrial Feasibility and Deployment Constraints

Unlike deterministic point estimates, the native probabilistic output of TabPFN-based models is a real operational advantage for practitioners calculating safety stock. This advantage is partially offset by the comparative weakness of TFMs on explainability. GBDTs come with mature post-hoc interpretability tooling such as SHAP values and split-based feature importances, which are widely used in enterprise forecasting workflows to justify ordering decisions to operations and finance stakeholders. ICL-based TFMs offer no equivalent: predictions emerge from attention over a context set rather than from a fixed learned function, and none of the four reviewed papers report on attribution methods, counterfactual explanations, or any other interpretability layer. For regulated industries and audit-heavy procurement processes, this is a non-trivial deployment gap.

Beyond interpretability, Lal (2026) highlight severe deployment constraints: high GPU requirements, slow inference times for large contexts, and the necessity of transmitting sensitive historical data to the model endpoint on every prediction call. The last of these is structural rather than incidental. Because ICL requires the support set at inference, sensitive historical transactions must be shipped to the model endpoint on every prediction, which is materially different from the GBDT deployment pattern where a trained model can be served locally with no further exposure of training data. Together, these factors introduce data governance and infrastructure considerations that apply even before accuracy is evaluated.

5.5 Hierarchical Consistency in Supply Chains

A shared limitation across these architectures is the handling of hierarchical reasoning. The reviewed models operate on flattened sequences of observations and produce an independent forecast for each series. Real supply chains require item-level forecasts to aggregate coherently to department, store, and regional totals. A forecast that is individually accurate but does not sum correctly up the hierarchy is difficult to operationalize.

The M5 evaluation in Potapczynski et al. (2026) implicitly acknowledges this, evaluating at state and store aggregation levels rather than at the full SKU level. The broader supply chain forecasting literature has developed two families of responses. Hierarchical reconciliation methods, of which the trace-minimization estimator MinT (Wickramasuriya et al. 2019) is the most widely cited, offer a post-processing path to consistency by adjusting independently produced base forecasts so that they sum coherently across the hierarchy. Applying such methods to a tabular foundation model would treat the model as a black-box base forecaster, leaving its inductive biases unchanged. An alternative is to train or query separate models at each aggregation level and combine them, as is common in the M5-winning solutions (Makridakis et al. 2022), though this multiplies inference cost and inherits the same context window constraints discussed in Section 4.2.

A recent position paper by Klein and Hoffart (2026) generalizes this critique. They argue that current tabular foundation models are constrained by their isolation from operational context: the declarative business rules and procedural code that govern how enterprise data is generated. Their proposed direction, Foundation Models for Semantically Linked Tables (FMSLT), would integrate

relational data with executable business logic, learned via a two-stage scheme that pretrains on open-source code-data pairs and synthetic system-table pairs before transferring to organizational data. By learning to apply business logic to synthetic data and only retrieving private artifacts at inference, architectures like FMSLT aim to navigate the privacy constraints observed in Section 4.3.

6 Conclusion

This paper analyzed four methodological perspectives on applying Tabular Foundation Models to demand forecasting in data-scarce supply chains. Our analysis reveals a clear progression: each approach identifies limitations, proposes targeted solutions, and in turn reveals deeper architectural constraints.

TabPFN-TS establishes that a tabular model can produce competitive forecasts, but its optimization for interpolation makes it struggle with trend extrapolation, which is a major issue for NPI and cold-start scenarios. The multivariate extension provides a framework for cross-channel integration but is severely bottlenecked by context window limits. The enterprise evaluation confirms that TFMs can outperform tuned GBDTs at very low sample sizes, but introduces significant data governance and infrastructure hurdles. Finally, ApolloPFN shows that temporal limitations can be fixed by modifying the pretraining distribution, though doing so blurs the line between tabular and time-series models.

Ultimately, tabular foundation models currently occupy a specific but narrow niche: data-scarce, interpolation-heavy forecasting tasks where GBDT hyperparameter tuning lacks sufficient training signal. Outside that niche, tree-based methods remain a strong and practical baseline. Future success in supply chains will likely depend on combining in-context inference with hierarchical reconciliation and the deeper integration of relational and operational context proposed by recent position work (Klein and Hoffart 2026).

With SAP SE's recent announcement to acquire Prior Labs (SAP SE 2026), with commiting to invest more than 1 billion euros, aiming to scale TabPFN into an AI research lab for structured business data and integrate the technology into SAP S/4HANA, the industrial relevance of this research is clear. As research converges with enterprise deployment, resolving issues around trend extrapolation, explainability, and data governance will determine the operational success of these models.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation, 2024. URL https://arxiv.org/abs/2410.10393.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

J. D. Croston. Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23(3):289–303, 1972. doi: 10.2307/3007885.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Dheeru Dua and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Robert Fildes, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25:3–23, 2009. doi: 10.1016/j.ijforecast.2008.11.010.

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 (NeurIPS)*, volume 35, pages 507–520, 2022.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-2.5: Advancing the state of the art in tabular foundation models, 2026a. URL https://arxiv.org/abs/2511.08667.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin, Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026b. URL https://arxiv.org/abs/2605.13986.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second, September 2023. URL http://arxiv.org/abs/2207.01848. arXiv:2207.01848 [cs.LG].

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045): 319–326, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From Tables to Time: Extending TabPFN-v2 to Time Series Forecasting, January 2026. URL http://arxiv.org/abs/2501.02945. arXiv:2501.02945 [cs.LG].

Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing: the State Space Approach*. Springer, 2008.

Mayuka Jayawardhana, Nihal Sharma, Kazem Meidani, Bayan Bruss, Tom Goldstein, and Doron Bergman. Zero-shot Multivariate Time Series Forecasting Using Tabular Prior Fitted Networks, April 2026. URL http://arxiv.org/abs/2604.08400. arXiv:2604.08400 [cs.LG] version: 1.

Tassilo Klein and Johannes Hoffart. Position: Foundation Models for Tabular Data within Systemic Contexts Need Grounding, January 2026. URL http://arxiv.org/abs/2505.19825. arXiv:2505.19825 [cs.LG] version: 2.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. doi: 10.1016/j.apenergy.2021.116983.

Amit Lal. Evaluating SAP RPT-1 for Enterprise Business Process Prediction: In-Context Learning vs. Traditional Machine Learning on Structured SAP Data, February 2026. URL http://arxiv.org/abs/2602.19237. arXiv:2602.19237 [cs.LG].

Spyros Makridakis and Michele Hibon. Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society*, 1979.

Spyros Makridakis and Michele Hibon. The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, 2000.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 2020.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336, 2022. doi: 10.1016/j.ijforecast.2021.07.007.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers Can Do Bayesian Inference, August 2024. URL http://arxiv.org/abs/2112.10510. arXiv:2112.10510 [cs.LG].

Andres Potapczynski, Ravi Kiran Selvam, Tatiana Konstantinova, Shankar Ramasubramanian, Malcolm Wolff, Kin G. Olivares, Ruijun Ma, Mengfei Cao, Michael W. Mahoney, Andrew Gordon Wilson, Boris N. Oreshkin, and Dmitry Efimov. Time-Aware Prior Fitted Networks for Zero-Shot Forecasting with Exogenous Variables, March 2026. URL http://arxiv.org/abs/2603.15802. arXiv:2603.15802 [cs.LG].

SAP SE. SAP to acquire Prior Labs to establish a globally leading frontier AI lab in Europe. SAP News Center, May 2026. URL https://news.sap.com/2026/05/sap-to-acquire-prior-labs-establish-frontier-ai-lab-europe/. Press release, accessed May 24, 2026.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting, 2026. URL https://arxiv.org/abs/2509.26468.

Marco Spinaci, Marek Polewczyk, Maximilian Schambach, and Sam Thelin. Contexttab: A semantics-aware tabular in-context learner, 2025. URL https://arxiv.org/abs/2506.10707.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

A. A. Syntetos and J. E. Boylan. The accuracy of intermittent demand estimates. *International Journal of Forecasting*, 21(2):303–314, 2005. doi: 10.1016/j.ijforecast.2004.10.001.

Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables, 2022. URL https://arxiv.org/abs/2205.09328.

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.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.