Foundation Models for Demand Forecasting via Dual-Strategy Ensembling

Wei Yang
University of Southern California
Los Angeles, United States of America
wyang930@usc.edu

Defu Cao
University of Southern California
Los Angeles, United States of America
defucao@usc.edu

Yan Liu
University of Southern California;
AWS Supply Chain
United States of America
yanliu.cs@usc.edu

ABSTRACT

Accurate demand forecasting is critical for supply chain optimization, yet remains difficult in practice due to hierarchical complexity, domain shifts, and evolving external factors. While recent foundation models offer strong potential for time series forecasting, they often suffer from architectural rigidity and limited robustness under distributional change. In this paper, we propose a unified ensemble framework that enhances the performance of foundation models for sales forecasting in real-world supply chains. Our method combines two complementary strategies: (1) **Hierarchical Ensemble (HE)**, which partitions training and inference by semantic levels (e.g., store, category, department) to capture localized patterns; and (2) **Architectural Ensemble (AE)**, which integrates predictions from diverse model backbones to mitigate bias and improve stability. We conduct extensive experiments on the M5 benchmark and three external sales datasets, covering both in-domain and zero-shot forecasting. Results show that our approach consistently outperforms strong baselines, improves accuracy across hierarchical levels, and provides a simple yet effective mechanism for boosting generalization in complex forecasting environments.

CCS CONCEPTS

• **Computing methodologies** → **Time series analysis**; **Ensemble methods**; **Transfer learning**; • **Applied computing** → **Forecasting**; **Supply chain management**.

KEYWORDS

Supply Chain, Demand Forecasting, Foundation Models, Ensemble Learning

ACM Reference Format:
Wei Yang, Defu Cao, and Yan Liu. 2025. Foundation Models for Demand Forecasting via Dual-Strategy Ensembling. In *Proceedings of the 1st Workshop on "AI for Supply Chain: Today and Future" @ 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

KDD '25, August 3, 2025, Toronto, ON, Canada.

ACM ISBN 979-8-4007-1454-2/25/08.
https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

Accurately forecasting future sales is a fundamental task in modern supply chain management [7, 9, 32, 34]. It drives core decisions in procurement, inventory planning, production scheduling, and logistics. Yet, despite its importance, demand forecasting remains highly challenging in practice [25, 42]. Real-world supply chains are complex, hierarchical, and increasingly sensitive to disruptions such as economic shocks, pandemics, and geopolitical tensions [18, 45, 47, 50]. These disruptions frequently induce regime shifts in demand signals, rendering many forecasting models brittle or unreliable. This situation creates a critical tension. While accurate forecasts are essential for operational resilience, building models that are robust, adaptive, and generalizable across diverse scenarios remains an open research problem [15, 20, 36].

Sales forecasting plays a central role in supply chain optimization [22, 23, 44]. It requires accurate forecasts across product, store and regional levels while accounting for seasonality, promotions and external disruptions. The M5 competition has emerged as a key benchmark in this domain, stimulating diverse modeling strategies [10, 21, 31, 39]. Top solutions fall into three paradigms: tree-based models like LightGBM[19], which rely on rich features and hierarchical ensembling; neural approaches such as DeepAR[38], which leverage autoregressive structures and distribution-aware objectives; and hybrid or statistical methods that remain effective under sparse or intermittent demand. Recently, foundation models like Chronos [2] and TEMPO [3] have gained traction by pretraining on large-scale time series to enable zero-shot generalization. While promising in bypassing task-specific tuning, these models often underperform when faced with domain shifts or hierarchical imbalances [8, 17, 30]. This highlights the need for new methods that can amplify their strengths while mitigating structural brittleness, particularly in complex, real-world supply chain environments.

Despite progress in time series modeling, single-model forecasters remain limited by the trade-off between bias and variance [28, 33]. Models based on trees, neural networks or foundation architectures often rely on fixed assumptions that do not adapt well across forecasting conditions. In supply chains, these limitations are amplified. Demand varies widely across products, stores and time periods, and often shifts due to promotions, seasonality or external events [9, 20, 32]. Even foundation models trained on diverse data can struggle in zero-shot settings when the target domain differs from the pretraining distribution. These failures lead to blind spots across segments and unstable predictions across different hierarchy levels. To address these issues, prior work has explored combining models. Mixture-of-experts methods are one example [1, 40, 46].

They aim to select or weight models dynamically, but often depend on specialized gating functions or fixed architecture designs. As a result, they can be sensitive to noise and fail to generalize under structural shifts. In real-world supply chains, where both data distribution and domain structure vary, such methods often lack robustness [24, 36]. This highlights the need for more flexible ensemble strategies that can integrate diverse model perspectives while remaining stable across different forecasting scenarios.

To overcome the limitations of single-model forecasting in complex supply chain environments, we develop an ensemble framework that combines hierarchical structure awareness with architectural diversity. This design is guided by two key insights. First, supply chain data exhibits inherent hierarchies across stores, categories, and regions. Modeling each semantic level separately allows better alignment with localized demand patterns and improves generalization across structural variations. Second, different forecasting architectures—such as tree-based models, recurrent networks, and transformers distinct inductive biases and capture complementary aspects of temporal dynamics. By integrating these perspectives, our approach incorporates two coordinated strategies. Hierarchical Ensemble (HE) partitions the training and inference process across semantic levels to promote subgroup specialization. Architectural Ensemble (AE) combines predictions from heterogeneous backbones to mitigate model-specific variance and improve robustness. Together, these components enable foundation models to produce more accurate and stable forecasts under both in-domain and zero-shot conditions.

Our contributions are summarized as follows:

- We propose a unified ensemble framework that enhances the performance of foundation models for supply chain forecasting. The framework boosts performance by strategically orchestrating how models are used and their predictions combined, thereby leveraging their collective strengths without requiring modifications to their underlying architecture.
- We introduce two complementary strategies: Hierarchical Ensemble (HE), which models group-specific patterns through semantic partitioning of the data, and Architectural Ensemble (AE), which integrates multiple forecasting backbones to improve robustness under distribution shifts.
- We conduct extensive experiments on the M5 benchmark and external sales datasets, demonstrating that our approach consistently outperforms strong baselines in both in-domain and zero-shot forecasting scenarios.

2 RELATED WORK

2.1 Sales Forecasting Methods

Sales forecasting has long served as a core task in retail supply chains [9, 22, 23, 32, 44], with the M5 competition emerging as a de facto benchmark for evaluating forecasting accuracy under hierarchical and sparse conditions. Tree-based models, particularly LightGBM, have dominated the leaderboard by capturing non-linear interactions and enabling fine-grained control across store, category, and SKU levels [12]. These models typically rely on engineered lag features, recursive training, and Tweedie objectives to address the skewness and intermittency of sales data. In parallel, purely exogenous models eschew autoregressive signals, instead framing the task as probabilistic classification using calendar and pricing features [13]. This shift allows better handling of cold-start or promotion-driven scenarios, especially at the lowest levels of the hierarchy. To reduce redundancy across similar sequences, transfer learning methods share model parameters across related units [48], while hybrid frameworks dynamically combine statistical and learning-based models based on local series characteristics [21]. Neural approaches have also gained traction. Variants of DeepAR incorporate multi-step rolling forecasts, hierarchical embeddings, and Tweedie loss to model zero-heavy counts [14]. Distributional models such as GAMLSS improve interval calibration [54], while simulation-based quantile forecasts align more directly with inventory-level decisions [41].

2.2 Foundation Models for Time Series Forecasting

Transformer-based models have become foundational in time series forecasting due to their scalability and ability to capture long-range dependencies [4, 27, 51, 52]. Early variants such as Autoformer and PatchTST introduce trend decomposition and patch-based attention, respectively, offering architectural inductive biases well-suited for complex temporal dynamics.

Recent work extends this line by exploring large language models (LLMs) for time series, treating numerical sequences as tokenized text [26]. For example, LLMTime and GPT4TS adopt GPT-style models for prompt-based forecasting [16, 53], enabling zero-shot generalization through autoregressive generation. TEMPO improves generality via trend-seasonality decomposition with soft prompts [3], while UniTime aligns forecasts with domain-specific instructions to enhance transferability [29].

Another direction focuses on training foundation models directly on time series, avoiding text pretraining. CHRONOS [2] tokenizes quantized series to build transformer-based models from scratch, and MOIRAI [49] extends this approach to multivariate and irregular sequences. Large-scale variants like TimeGPT, Lag-Llama, and TimesFM aim to generalize across diverse domains [5, 6, 37].

However, these models often overlook external semantics and may struggle under distribution shifts. Our work builds on this foundation by integrating ensemble learning with foundation models to improve robustness and forecasting accuracy across real-world supply chain contexts.

3 THE PROPOSED METHOD

3.1 Foundation Model for Time Series Forecasting

We begin with a powerful backbone: a pre-trained foundation model for time series forecasting. Inspired by recent advances in large language models (LLMs), foundation models for temporal data aim to learn universal representations from large-scale, multi-domain time series [2, 3, 6]. These models possess strong zero-shot and transfer capabilities, enabling cross-domain generalization without task-specific retraining. Specifically, we refer to a "single model" as a unified forecasting architecture that includes both the model backbone and training strategy. This definition does not depend on internal structure, such as the use of mixture-of-experts or parameter sharing, but rather treats each independently trained configuration as a distinct model. In this work, we adopt a transformer-based architecture $\mathcal{F}_\theta$, parameterized by θ, pre-trained on diverse temporal datasets and fine-tuned for hierarchical sales prediction on the M5 dataset.

Given an input sequence $x_{1:T} = [x_1, x_2, \ldots, x_T]$ representing historical sales and associated covariates, the model predicts a future horizon of length H via:

$$\hat{y}_{T+1:T+H} = \mathcal{F}_\theta(x_{1:T}) \tag{1}$$

The foundation model captures both short- and long-term dependencies, and is capable of adapting to novel item-store combinations in zero-shot settings. However, despite its expressive power, it often suffers from empirically observed biases in underrepresented domains, as shown in Table 1 by increased WRMSSE at lower levels (Levels 10–12), motivating the need for robust ensemble strategies.

3.2 Hierarchical and Architectural Ensemble Framework

Our method builds on the hierarchical and architectural considerations introduced in the design. We formalize this through an ensemble framework that combines semantic partitioning and model diversity to enhance forecasting performance. The framework is composed of two components: **Hierarchical Ensemble (HE)** and **Architectural Ensemble (AE)**. The Hierarchical Ensemble targets the structural organization of supply chain data. We partition the training data along semantic dimensions such as store, category, and region, and train specialized models for each group. This encourages localized learning and captures subgroup-specific temporal patterns that global models often miss. During inference, predictions are aggregated across levels to preserve consistency while retaining fine-grained accuracy.

The Architectural Ensemble addresses the variability in inductive biases across model types. We instantiate a set of diverse forecasting backbones, including statistical, recurrent and transformer-based models. These models are trained on the same data scope and generate independent forecasts. Their outputs are combined through weighted aggregation to reduce variance and increase robustness under distributional shift. By jointly leveraging hierarchical structure and architectural diversity, the ensemble achieves greater adaptability across forecasting scenarios, including both in-domain and zero-shot settings.

3.2.1 Hierarchical Ensemble Learning (HE). Retail time series data are inherently hierarchical, with different levels exhibiting distinct statistical patterns. We exploit this structure by training independent models $\{\mathcal{M}_\ell^{(i)}\}$ for each group i within a granularity level $\ell \in \{\texttt{store}, \texttt{store+category}, \texttt{store+dept}\}$. For each level, a local model $\mathcal{M}_\ell^{(i)}$ is trained only on the subset of data corresponding to its group i (e.g., a specific store). During inference, the prediction $\hat{y}_\ell^{(i)}$ for a given item is obtained from its corresponding model at level ℓ:

$$\hat{y}_\ell^{(i)} = \mathcal{M}_\ell^{(i)}(x_{1:T}^{(i)}) \tag{2}$$

The final forecast $\hat{y}$ is then computed by aggregating predictions from all levels using a weighted average:

$$\hat{y} = \sum_{\ell=1}^{L} w_\ell \cdot \hat{y}_\ell^{(i)}, \quad \text{with} \quad \sum_{\ell=1}^{L} w_\ell = 1 \tag{3}$$

This multi-resolution modeling strategy aligns with the natural structure of the data and improves generalization across disjoint partitions.

3.2.2 Architectural Ensemble Learning (AE). Different model architectures exhibit distinct inductive biases and error characteristics. For instance, window-based models (e.g., LightGBM) excel at tabular feature interactions, while sequence models (e.g., DeepAR, PatchTST) better capture temporal patterns. We denote the set of K diverse backbones as $\{\mathcal{B}_1, \ldots, \mathcal{B}_K\}$, each producing a prediction $\hat{y}_k$ for the same target.

To combine their strengths and reduce structure-specific bias, we aggregate their outputs via weighted fusion:

$$\hat{y} = \sum_{k=1}^{K} v_k \cdot \hat{y}_k, \quad \text{with} \quad \sum_{k=1}^{K} v_k = 1 \tag{4}$$

This ensemble yields smoother, more stable predictions by integrating architectural diversity, as shown by the consistent WRMSSE reductions in Table 2, and reduces variance from any single backbone. When combined with level-based ensemble, our final forecast benefits from both local specialization and global robustness.

Unified Ensemble Objective. The full prediction pipeline integrates both ensemble layers—hierarchical and architectural—into a unified framework:

$$\hat{y}_{\text{final}} = \sum_{k=1}^{K} v_k \cdot \left(\sum_{\ell=1}^{L} w_\ell \cdot \hat{y}_{\ell,k}^{(i)} \right) \tag{5}$$

where $\hat{y}_{\ell,k}^{(i)}$ denotes the prediction from model backbone k at hierarchy level ℓ for group i. Here, each prediction $\hat{y}_{\ell,k}^{(i)}$ corresponds to a particular forecasting group i, and the summation is applied within each group. Final aggregation is conducted per group instance. For simplicity, we assign equal normalized weights during ensembling.

This hierarchical-heterogeneous ensemble strategy serves as a structural regularizer, harmonizing local specialization with model diversity, and significantly boosts forecast accuracy across all levels.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following research questions:

- **RQ1:** How effective is our ensemble framework in improving forecasting accuracy on supply chain benchmarks?
- **RQ2:** Can our method enhance the zero-shot generalization ability of foundation models when applied to unseen sales datasets with different distributions?
- **RQ3:** Does our ensemble framework consistently improve forecasting accuracy across different datasets under full-shot training conditions?
- **RQ4:** Why does ensemble learning improve performance, and what structural dynamics underlie its effectiveness?

4.1 Experimental Settings

4.1.1 Datasets. Our experiments are conducted on four real-world sales forecasting datasets. **M5 Forecasting**[1] is the primary dataset used for model training and adaptation. It consists of daily sales records for 30,490 Walmart items across multiple hierarchical levels including state, store, category and department. The dataset includes rich covariates such as calendar events and item prices, and defines WRMSSE as the official evaluation metric to emphasize both scale and aggregation consistency.

To evaluate the generalization of foundation and ensemble models, we use three real-world sales forecasting datasets from domains beyond the training data. Specifically, (1) **Sales1 (Walmart Promo)**[2] includes weekly sales data with markdown events and external economic indicators; (2) **Sales2 (Store-Item Benchmark)**[3] is a clean benchmark dataset containing five years of daily sales across various stores and product SKUs; and (3) **Sales3 (Balkan Retail)**[4] is a real-world monthly dataset spanning seven years, covering sales and pricing information for top-selling items across multiple business units in the Balkan region.

4.1.2 Evaluation Metrics. We adopt different evaluation metrics based on the characteristics of the target datasets. For the M5 forecasting task, we follow the official competition protocol and use the **Weighted Root Mean Squared Scaled Error (WRMSSE)** as the primary metric. WRMSSE is a hierarchical-aware error measure that accounts for both the scale and the relative importance of each time series in the hierarchy. It penalizes errors more heavily at aggregate levels and rewards models that maintain consistency across disaggregated units. Formally, WRMSSE extends RMSSE by introducing level-dependent weights based on sales volume and aggregation depth, enabling a unified evaluation across twelve levels of the retail hierarchy.

For the three external sales datasets, we employ standard metrics widely used in forecasting literature: the **Mean Squared Error (MSE)** and the **Mean Absolute Error (MAE)**. These metrics capture complementary aspects of prediction performance: MSE is sensitive to large errors and highlights variance, while MAE reflects median deviation and is more robust to outliers. Together, they provide a comprehensive view of model accuracy in cross-domain generalization settings.

4.1.3 Baselines. We compare our method against the following representative forecasting models:

- **LightGBM** [19]: A strong gradient boosting model with handcrafted temporal and categorical features, widely adopted in M5 competition solutions.
- **DNN** [43]: A feedforward neural network baseline trained on static and lag-based features, representing classical deep learning approaches without temporal modeling.
- **DeepAR** [38]: An RNN-based probabilistic model that forecasts future values by learning autoregressive conditional distributions across multiple time series.
- **PatchTST** [35]: A Transformer model that segments input sequences into patches and processes them in a channel-independent fashion for efficient long-range forecasting.
- **TEMPO** [3]: A foundation model pretrained on time series data using trend-seasonality decomposition and prompt-based adaptation to capture domain-invariant temporal representations.
- **Chronos** [2]: A pre-trained Transformer model that tokenizes time series into quantized sequences and learns probabilistic forecasts via language modeling objectives.

4.1.4 Implementation Details. We follow the unified experimental setup provided by Zhou et al. [53] to ensure consistent and reproducible comparisons across baselines.[5] All models are implemented using PyTorch [11], and trained on a single NVIDIA A100 GPU.

For M5 forecasting, we adopt the standard evaluation split, using the last 28 days of each series as the prediction window. We follow the WRMSSE calculation protocol as defined by the official competition, and apply scale weights across twelve hierarchical levels. For zero-shot evaluation, the foundation models are directly applied to target datasets without any fine-tuning, and we report average MSE and MAE across all prediction windows.

Model-specific configurations (e.g., input sequence length, learning rate, batch size) follow either default values from the Time-Series Library [53] or best-practice settings reported in prior works. We perform minimal hyperparameter tuning to preserve fairness and focus our evaluation on architectural design and ensemble effectiveness.

4.2 Effectiveness of the Ensemble Framework on M5 (RQ1)

4.2.1 Hierarchical Ensemble Performance. Table 1 and Figure 1 reports the forecasting performance of various backbones with and without **Hierarchical Ensemble (HE)** on the M5 dataset. Results are evaluated using the WRMSSE metric across twelve hierarchical levels. Overall, applying HE consistently improves model accuracy across all architectures, demonstrating the effectiveness of hierarchy-aware specialization in capturing structured variation.

The gains are particularly notable for backbones that struggle with heterogeneity in raw form. DeepAR, for example, improves from 0.5556 to 0.5233 in WRMSSE, with especially large reductions at mid-level aggregations where local variation dominates. HE enables DeepAR to specialize across semantic groups, reducing the burden of modeling global variance. Similarly, PatchTST benefits significantly from HE, improving from 0.6997 to 0.6210. As a transformer-based model, PatchTST tends to overfit to coarse-grained signals and suffers from unstable behavior across levels. HE mitigates this by isolating distributional shifts and training within homogeneous partitions.

Even large-scale foundation models like TEMPO and Chronos show measurable gains with HE, despite being pretrained for generalization. This indicates that structural alignment remains a valuable inductive bias, even in the foundation model regime. Overall, HE serves not only as a specialization mechanism but also as a structural regularizer that improves both fine-grained fidelity and cross-level consistency.

[1]https://www.kaggle.com/competitions/m5-forecasting-accuracy/overview
[2]https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/overview
[3]https://www.kaggle.com/competitions/demand-forecasting-kernels-only/overview
[4]https://data.4tu.nl/articles/_/14406134/1

[5]https://github.com/thuml/Time-Series-Library

Table 1: WRMSSE results of different forecasting backbones with and without Hierarchical Ensemble (HE) applied. Results are reported across twelve aggregation levels on the M5 dataset. The Kaggle-Top1 solution corresponds to a LightGBM-based model with a hierarchical ensemble strategy. M1: Window-based, M2: RNN-based, M3: Transformer-based, M4: Foundation Model.

Type	Backbone	Avg.	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11	Level 12
M1	Kaggle-Top1 w/HE	**0.5230**	**0.2006**	**0.3123**	**0.3992**	**0.2843**	**0.3725**	**0.3953**	**0.4792**	**0.4815**	**0.5742**	**0.9649**	**0.9283**	**0.8838**
M1	DNN	0.6984	0.3405	0.4427	0.7106	0.4258	0.5741	0.5723	0.6955	0.8444	0.9187	0.9905	0.9551	0.9104
M1	DNN w/HE	**0.6596**	**0.2954**	**0.4183**	**0.6413**	**0.3784**	**0.5239**	**0.5379**	**0.6560**	**0.7692**	**0.8514**	**0.9862**	**0.9513**	**0.9057**
M2	DeepAR	0.5556	0.2631	0.3469	0.4413	0.3366	0.4306	0.4312	0.5237	0.5222	0.6181	0.9560	0.9203	0.8775
M2	DeepAR w/HE	**0.5233**	**0.2054**	**0.3096**	**0.4071**	**0.2708**	**0.3815**	**0.3909**	**0.4913**	**0.4873**	**0.5920**	**0.9504**	**0.9175**	**0.8761**
M3	PatchTST	0.6997	0.4626	0.5441	0.6421	0.5015	0.6032	0.6098	0.6869	0.6968	0.7652	0.9848	0.9600	0.9397
M3	PatchTST w/HE	**0.6210**	**0.3498**	**0.4552**	**0.5509**	**0.4028**	**0.5129**	**0.5363**	**0.6211**	**0.6238**	**0.7080**	**0.9260**	**0.8995**	**0.8654**
M4	TEMPO	0.9706	**0.8021**	0.9281	1.1122	**0.8197**	0.8746	0.9530	0.9920	1.0953	1.1054	1.0447	0.9884	0.9321
M4	TEMPO w/HE	**0.9137**	0.8106	**0.8429**	**0.9088**	0.8300	**0.8679**	**0.8899**	**0.9354**	**0.9460**	**0.9966**	**1.0373**	**0.9788**	**0.9201**
M4	Chronos	2.4358	3.1928	2.9399	2.8101	3.1571	3.1303	2.8454	2.7765	2.5914	2.4350	1.3178	1.0884	0.9450
M4	Chronos w/HE	**2.2051**	**2.8506**	**2.6301**	**2.5219**	**2.8129**	**2.8066**	**2.5452**	**2.4973**	**2.3275**	**2.2008**	**1.2624**	**1.0651**	**0.9400**

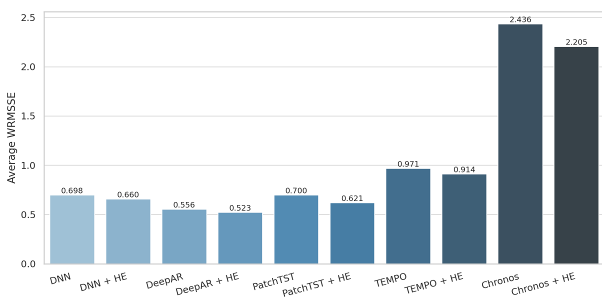

Figure 1: Average WRMSSE for backbone models and their Hierarchical Ensemble (HE) variants on the M5 dataset. HE consistently improves accuracy across model families, including tree-based, neural, and foundation models.

4.2.2 Architectural Ensemble Performance. Table 2 presents the results of **Architectural Ensemble (AE)**, where predictions from structurally diverse models are combined through weighted fusion. This strategy aims to mitigate the limitations of individual architectures by integrating complementary modeling perspectives. Across all configurations, AE consistently improves over single backbones, indicating that architectural diversity contributes to more stable and accurate forecasting.

The combination of LightGBM and PatchTST achieves the best overall performance, reducing the WRMSSE from 0.5230 to 0.4989. This ensemble benefits from the contrasting strengths of its components. LightGBM excels at modeling high-level aggregation with strong performance on sparse tabular data, while PatchTST adapts well to fine-grained temporal dynamics at lower levels. Their error patterns show low correlation, enabling the ensemble to cancel out systematic biases and enhance generalization.

Fusing LightGBM with DeepAR yields similar improvements, further confirming that pairing models with distinct inductive assumptions can reduce variance and improve robustness. While the PatchTST and DeepAR ensemble is slightly less effective, it still outperforms the individual models, reinforcing the value of architectural complementarity. These results support the view that AE serves as an effective bias correction mechanism. By integrating forecasts from models with divergent inductive behaviors, AE suppresses architecture-specific errors and promotes more balanced predictions across hierarchical levels.

4.3 Foundation-Model Ensemble for Zero-Shot Forecasting (RQ2)

We evaluate the zero-shot generalization of foundation models and their ensemble-enhanced variants on real-world sales forecasting datasets outside the M5 domain. In this setting, models are applied without fine-tuning, simulating deployment to unseen markets where no target labels are available. The goal is to assess whether ensemble integration can improve robustness under distribution shift, even when foundation models are pre-trained to generalize.

Architectural Ensemble consistently improves the performance of all tested backbones. This effect is especially pronounced for PatchTST, a transformer-based model trained without cross-domain pretraining. PatchTST alone shows limited transferability and unstable error patterns across datasets. When combined with structurally distinct models through AE, its predictions become more stable and better aligned with local temporal structures, suggesting that ensemble fusion can compensate for narrow inductive biases.

Foundation models like Chronos and TEMPO exhibit stronger baseline performance, yet also benefit from AE. In particular, Chronos gains from the integration of models that specialize in coarse-grained seasonality or localized trends. For TEMPO, which already includes architectural adaptations such as decomposition and modulation, AE acts as a form of bias regularization. It smooths over residual overfitting and introduces modeling perspectives that help balance representation across varying domains. Pretraining provides broad generalization capacity, but it does not eliminate structural blind spots or ensure robustness under extreme domain shifts. AE complements this by integrating heterogeneous inductive signals, enabling more resilient zero-shot forecasting without requiring retraining or domain-specific adaptation.

Table 2: Architectural Ensemble (AE) performance on the M5 dataset. Forecasts from structurally diverse backbone models, including LightGBM, DeepAR, and PatchTST, are aggregated using weighted fusion. This ensemble leverages complementary inductive biases to reduce model-specific error patterns and improve robustness across all hierarchical levels. For reference, the top-ranked Kaggle solution using only LightGBM is also included.

Backbone	avg.	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11	Level 12
Kaggle-Top1	**0.5230**	**0.2006**	**0.3123**	**0.3992**	**0.2843**	**0.3725**	**0.3953**	**0.4792**	**0.4815**	**0.5742**	**0.9649**	**0.9283**	**0.8838**
AE (LightGBM, PatchTST)	**0.4989**	**0.1455**	0.2907	0.3830	**0.2244**	**0.3392**	0.3688	0.4631	0.4625	0.5623	**0.9528**	**0.9188**	**0.8758**
AE (LightGBM, DeepAR)	0.5003	0.1537	**0.2903**	**0.3792**	0.2330	0.3427	**0.3677**	**0.4627**	**0.4591**	**0.5607**	0.9556	0.9212	0.8777
AE (PatchTST, DeepAR)	0.5216	0.2002	0.3119	0.4163	0.2601	0.3800	0.3946	0.4948	0.4954	0.5955	0.9366	0.9064	0.8674

Table 3: Zero-shot forecasting performance (MSE / MAE) of foundation models and their ensemble-enhanced variants across three external sales datasets. HE denotes Hierarchical Ensemble. Foundation models are pre-trained on the M5 dataset without access to target domains.

Model	Sales1	Sales2	Sales3
	MSE / MAE	*MSE / MAE*	*MSE / MAE*
PatchTST	7.3854 / 1.3052	11.7808 / 3.1509	2.7013 / 1.1475
PatchTST w/HE	**5.4312** / **1.0393**	**8.5738** / **2.3874**	**2.4085** / **1.0712**
TEMPO	1.1690 / 0.8357	1.1814 / 0.8792	1.7293 / 0.9556
TEMPO w/HE	**1.1173 / 0.8038**	**1.1503 / 0.8674**	**1.6725 / 0.9487**
Chronos	2.5253 / 1.1935	2.6585 / 1.1721	3.1250 / 1.4787
Chronos w/HE	**2.3375** / **1.1372**	**2.3751** / **1.0965**	**2.8533** / **1.3514**

Table 4: Full-shot forecasting performance (MSE / MAE) of foundation models and their ensemble-enhanced variants across three external sales datasets. HE denotes Hierarchical Ensemble. Foundation models are trained on each specific dataset.

Model	Sales1	Sales2	Sales3
	MSE / MAE	*MSE / MAE*	*MSE / MAE*
PatchTST	0.0420 / 0.0949	0.0896 / 0.2305	0.7161 / 0.5177
PatchTST w/HE	**0.0403** / **0.0938**	**0.0881** / **0.2282**	**0.6673** / **0.4938**
TEMPO	0.0408 / 0.0939	0.0878 / 0.2284	0.5869 / 0.4751
TEMPO w/HE	**0.0391 / 0.0921**	**0.0869 / 0.2263**	**0.5791 / 0.4711**
Chronos	0.0411 / 0.0933	0.0887 / 0.2296	0.6019 / 0.4838
Chronos w/HE	**0.0398** / **0.0925**	**0.0875** / **0.2284**	**0.5827** / **0.4787**

4.4 Cross-Dataset Generalization of Hierarchical Ensemble in Full-Shot Forecasting (RQ3)

While RQ1 establishes the effectiveness of Hierarchical Ensemble (HE) on the M5 dataset in a full-shot training setting, it remains unclear whether the same benefits generalize to other domains. In this section, we evaluate the robustness of HE across three external sales datasets by retraining foundation models from scratch on each dataset and comparing performance with and without HE.

Table 4 presents the results for PatchTST, TEMPO, and Chronos when trained directly on each target dataset. Across all models and datasets, we observe consistent improvements in both MSE and MAE after applying HE. For example, PatchTST w/HE outperforms its non-ensemble variant on all datasets, with particularly notable gains on Sales3 (MSE reduced from 0.7161 to 0.6673). Similar trends are observed for TEMPO and Chronos, confirming the general effectiveness of HE even in domains outside the M5 benchmark.

These results suggest that the benefits of HE extend beyond a single dataset and are not specific to any particular domain structure. Even when retraining on diverse datasets with different temporal resolutions and demand patterns, HE consistently enhances performance. This supports the claim that HE functions as a general-purpose structural inductive bias, encouraging subgroup-level specialization and reducing variance due to global overfitting.

In contrast to RQ2, which focused on zero-shot transfer from a single source dataset (M5), this experiment demonstrates that HE retains its effectiveness in realistic full-shot deployment scenarios where training on the target domain is feasible. Together with the results from RQ1 and RQ2, this cross-dataset analysis highlights the broad applicability of the HE strategy and motivates its adoption in both transfer and direct-learning forecasting workflows.

4.5 Why Does Ensemble Work? A Layer-Wise View (RQ4)

To better understand the efficacy of ensemble learning, we examine the performance of PatchTST trained independently at three semantic levels: Store, Store+Department, and Store+Category and compare them with their ensemble combination. Table 5 reports the WRMSSE across all 12 evaluation levels defined in the M5 competition, which span from total sales aggregated across all stores (Level 1) to individual product-store combinations (Level 12).

A key insight emerges when aligning performance patterns with the semantic meaning of each level. Levels 1–5 correspond to coarse-grained aggregations such as total sales by state or category. At these levels, the *Store+Department* model performs best, as it captures relatively stable and high-volume sales patterns associated with department-level trends. However, at mid-level aggregations (Levels 6–9), which include cross-structured dimensions such as Store+Category or State+Department, the performance advantage shifts slightly toward *Store+Category*, likely due to its finer specialization along consumer preference lines. At the lowest levels (Levels 10–12), which evaluate item-level forecasts by store or region, the

Table 5: WRMSSE of PatchTST under different hierarchical modeling strategies on the M5 dataset. "Store-Dept", "Store-Cate", and "Store" denote models trained on different semantic levels. "Ensemble" corresponds to the Hierarchical Ensemble (HE) of the three levels.

Model	Avg.	Level1	Level2	Level3	Level4	Level5	Level6	Level7	Level8	Level9	Level10	Level11	Level12
PatchTST (Store-Dept)	0.6349	0.3374	0.4547	0.5650	0.3982	0.5069	0.5412	0.6390	0.6538	0.7571	0.9245	0.9224	0.9187
PatchTST (Store-Cate)	0.6387	0.3559	0.4542	0.5713	0.4118	0.5289	0.5499	0.6357	0.6563	0.7347	0.9276	0.9220	0.9159
PatchTST (Store)	0.6339	0.3615	0.4716	0.5625	0.4173	0.5344	0.5528	0.6368	0.6330	0.7126	0.9170	0.9080	0.8994
PatchTST (Ensemble)	**0.6210**	**0.3498**	**0.4552**	**0.5509**	**0.4028**	**0.5129**	**0.5363**	**0.6211**	**0.6238**	**0.7080**	**0.9260**	**0.8995**	**0.8654**

Store-only model shows superior adaptability, as it has learned from more homogeneous, localized patterns.

These observations highlight that no single hierarchical partition captures all relevant signal components across levels. Each level-specific model overfits or undergeneralizes in different structural regimes. Importantly, the HE fusion of these three perspectives consistently improves or matches the best component model across all levels, resulting in an overall WRMSSE reduction from 0.6349 to 0.6210.

From a theoretical standpoint, this reinforces the notion that structural diversity among models introduces orthogonal inductive biases, which when aggregated, cancel out local biases and yield a more balanced predictor. HE not only reduces forecast variance but acts as a structure-aware alignment mechanism, reconciling inconsistencies between different views of the data hierarchy. The ensemble is not simply averaging—it is synthesizing semantically complementary forecasts that individually dominate in specific regimes but are collectively incomplete.

4.6 Discussion and Limitations

Our findings reveal that effective demand forecasting in real-world supply chains increasingly hinges not on refining a single model architecture, but on orchestrating diverse inductive perspectives across structural and architectural dimensions. The dual ensemble strategies proposed in this work address complementary sources of forecasting difficulty: HE mitigates distributional fragmentation caused by semantic heterogeneity (e.g., store, category, department), while AE balances architectural biases by fusing models with distinct representational priors.

Notably, our analysis demonstrates that no single semantic partition dominates performance across all levels of the hierarchy. Instead, different granularities specialize in different structural regimes, and their ensemble synthesis yields performance superior to any individual specialization. Similarly, the combination of tree-based models and transformer-based foundation models leverages non-overlapping error characteristics, leading to more robust and stable forecasts, particularly under distribution shifts.

From a broader perspective, these results suggest that forecasting accuracy in complex domains emerges not from deeper architectures, but from structured diversity—a principle aligned with emerging trends in foundation model research. While pretrained models such as TEMPO and CHRONOS offer strong zero-shot capabilities, they remain susceptible to inductive blind spots introduced by coarse pretraining objectives or flattened hierarchies. Our ensemble framework effectively regularizes these foundation models by anchoring them to localized patterns and complementary views.

This insight carries broader implications. As supply chains become more volatile and fragmented, future forecasting systems must be modular, adaptive, and capable of reconciling signals across levels and representations. Rather than seeking universal predictors, we advocate for systems that learn to coordinate partial experts, each aligned with a semantically coherent subspace or architectural strength. Such systems not only improve accuracy but offer greater resilience, transparency, and extensibility for downstream decision-making.

Despite these advances, several limitations and open questions remain. While our ensemble strategy is effective in fusing structural and architectural diversity, it currently operates with fixed combination schemes and does not explicitly adapt to context or input uncertainty. Moreover, although foundation models pretrained on large-scale sales data demonstrate strong generalization, their robustness across domains with distinct temporal patterns or operational semantics (e.g., manufacturing or healthcare) is still not well understood. Finally, while accuracy improves significantly, the interpretability and explainability of ensemble outputs remain an open challenge, especially in decision-making environments.

5 CONCLUSION

In this paper, we present a unified ensemble framework for sales forecasting in supply chain settings, combining hierarchical structure awareness with architectural diversity to enhance the performance of pretrained foundation models. Our method integrates Hierarchical Ensemble (HE) and Architectural Ensemble (AE) to mitigate model-specific biases, capture fine-grained local patterns, and enable robust generalization across both in-domain and zero-shot scenarios. Empirical results on the M5 benchmark and three external datasets demonstrate that our approach consistently improves forecasting accuracy and stability, particularly when applied to foundation models such as TEMPO and Chronos.

Looking forward, we envision several promising research directions. First, the development of domain-specialized foundation models tailored for supply chain data could significantly enhance representational alignment, especially when incorporating structured metadata. Second, integrating LLM-based temporal reasoning offers a powerful avenue for uncovering latent causal structures and interpreting forecasting decisions. Third, Retrieval-Augmented Generation (RAG) can help forecasting models adapt to changing environments by retrieving relevant signals such as promotions, macroeconomic trends and calendar events. Finally, combining generative modeling with discriminative objectives may improve both forecast accuracy and interpretability, which is essential for decision support in supply chain operations.

REFERENCES

[1] Khaled Alkilane, Yihang He, and Der-Horng Lee. 2024. MixMamba: Time series modeling with adaptive expertise. *Information Fusion* 112 (2024), 102589.

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815* (2024).

[3] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948* (2023).

[4] Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. 2024. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. *arXiv preprint arXiv:2402.05956* (2024).

[5] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*.

[6] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv preprint arXiv:2310.03589* (2023).

[7] Iman Ghalehkhondabi, Ehsan Ardjmand, Gary R Weckman, and William A Young. 2017. An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems* 8 (2017), 411–447.

[8] Peiliang Gong, Emadeldeen Eldele, Min Wu, Zhenghua Chen, Xiaoli Li, and Daoqiang Zhang. [n.d.]. Towards Adaptive Time Series Foundation Models Against Distribution Shift. ([n. d.]).

[9] MD Rokibul Hasan, Md Raisul Islam, and Md Anisur Rahman. 2025. Developing and implementing AI-driven models for demand forecasting in US supply chains: A comprehensive approach to enhancing predictive accuracy. *Edelweiss applied science and technology* 9, 1 (2025), 1045–1068.

[10] Hansika Hewamalage, Pablo Montero-Manso, Christoph Bergmeir, and Rob J Hyndman. 2021. A look at the evaluation setup of the m5 forecasting competition. *arXiv preprint arXiv:2108.03588* (2021).

[11] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. 2021. PyTorch. *Programming with TensorFlow: solution for edge computing applications* (2021), 87–104.

[12] Yeonjun IN. 2020. *1st Place Solution: M5 Forecasting - Accuracy.* https://github.com/YeonJun-IN/data.science.competition/blob/master/4.%20%5Bkaggle%5DM5-Accuracy/ Accessed May 2025.

[13] Yeonjun IN. 2020. *1st Place Solution: M5 Forecasting - Accuracy.* https://github.com/YeonJun-IN/data.science.competition/blob/master/4.%20%5Bkaggle%5DM5-Accuracy/ Accessed May 2025.

[14] Yeonjun IN. 2020. *1st Place Solution: M5 Forecasting - Accuracy.* https://github.com/YeonJun-IN/data.science.competition/blob/master/4.%20%5Bkaggle%5DM5-Accuracy/ Accessed May 2025.

[15] Md Abrar Jahin, Asef Shahriar, and Md Al Amin. 2024. Mcdfn: Supply chain demand forecasting via an explainable multi-channel data fusion network model integrating cnn, lstm, and gru. *arXiv e-prints* (2024), arXiv–2405.

[16] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[17] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position Paper: What Can Large Language Models Tell Us about Time Series Analysis. *arXiv preprint arXiv:2402.02713* (2024).

[18] Hatim Kagalwala, GV Radhakrishnan, Irshadullah Asim Mohammed, Rishi Reddy Kothinti, and Nirzar Kulkarni. 2025. Predictive analytics in supply chain management: The role of AI and machine learning in demand forecasting. *Advances in Consumer Research* 2 (2025), 142–149.

[19] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[20] Khaoula Khlie, Z Benmamoun, W Fethallah, and I Jebbor. 2024. Leveraging variational autoencoders and recurrent neural networks for demand forecasting in supply chain management: A case study. *Journal of infrastructure, policy and development* 8, 8 (2024), 6639.

[21] Stephan Kolassa. 2022. Commentary on the M5 forecasting competition. *International Journal of Forecasting* 38, 4 (2022), 1562–1568.

[22] Praveen Kumar, Divya Choubey, Olamide Raimat Amosu, and Yewande Mariam Ogunsuji. 2024. AI-enhanced inventory and demand forecasting: Using AI to optimize inventory management and predict customer demand. *World J. Adv. Res. Rev* 23, 1 (2024).

[23] A David Lainder and Russell D Wolfinger. 2022. Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies: Winning solution to the M5 Uncertainty competition. *International Journal of Forecasting* 38, 4 (2022), 1426–1433.

[24] Xin Li, Yechi Xu, Rob Law, and Shouyang Wang. 2024. Enhancing tourism demand forecasting with a transformer-based framework. *Annals of Tourism Research* 107 (2024), 103791.

[25] Bojing Liu, Mengxiang Li, Zihui Ji, Hongming Li, and Ji Luo. 2024. Intelligent productivity transformation: corporate market demand forecasting with the aid of an AI virtual assistant. *Journal of Organizational and End User Computing (JOEUC)* 36, 1 (2024), 1–27.

[26] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. 2025. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 18780–18788.

[27] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2022. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *# PLACEHOLDER_PARENT_METADATA_VALUE#*.

[28] Siwei Liu and Di Jody Zhou. 2024. Using cross-validation methods to select time series models: Promises and pitfalls. *Brit. J. Math. Statist. Psych.* 77, 2 (2024), 337–355.

[29] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*. 4095–4106.

[30] Qianli Ma, Zhen Liu, Zhenjing Zheng, Ziyang Huang, Siying Zhu, Zhongzhong Yu, and James T Kwok. 2024. A survey on time-series pre-trained models. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[31] Spyros Makridakis, Fotios Petropoulos, and Evangelos Spiliotis. 2022. The M5 competition: conclusions. , 1576–1582 pages.

[32] Arnab Mitra, Arnav Jain, Avinash Kishore, and Pravin Kumar. 2022. A comparative study of demand forecasting models for a multi-channel retail company: a novel hybrid machine learning approach. In *Operations research forum*, Vol. 3. Springer, 58.

[33] Angel E Muñoz-Zavala, Jorge E Macías-Díaz, Daniel Alba-Cuéllar, and José A Guerrero-Díaz-de León. 2024. A literature review on some trends in artificial neural networks for modeling and simulation with time series. *Algorithms* 17, 2 (2024), 76.

[34] Ali Roozbeh Nia, Anjali Awasthi, and Nadia Bhuiyan. 2021. Industry 4.0 and demand forecasting of the energy supply chain: A literature review. *Computers & Industrial Engineering* 154 (2021), 107128.

[35] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).

[36] Adedoyin Tolulope Oyewole, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and Emuesiri Ejairu. 2024. Reviewing predictive analytics in supply chain management: Applications and benefits. *World Journal of Advanced Research and Reviews* 21, 3 (2024), 568–574.

[37] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. 2023. Lag-llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278* (2023).

[38] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting* 36, 3 (2020), 1181–1191.

[39] Brian Seaman and John Bowman. 2022. Applicability of the M5 to Forecasting at Walmart. *International Journal of Forecasting* 38, 4 (2022), 1468–1472.

[40] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040* (2024).

[41] Evangelos Spiliotis, Spyros Makridakis, Anastasios Kaltsounis, and Vassilios Assimakopoulos. 2021. Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data. *International Journal of Production Economics* 240 (2021), 108237.

[42] Kritika Swaminathan and Rakesh Venkitasubramony. 2024. Demand forecasting for fashion products: A systematic review. *International Journal of Forecasting* 40, 1 (2024), 247–267.

[43] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

[44] Evangelos Theodorou, Shengjie Wang, Yanfei Kang, Evangelos Spiliotis, Spyros Makridakis, and Vassilios Assimakopoulos. 2022. Exploring the representativeness of the M5 competition data. *International Journal of Forecasting* 38, 4 (2022), 1500–1506.

[45] Juan R Trapero, Enrique Holgado de Frutos, and Diego J Pedregal. 2024. Demand forecasting under lost sales stock policies. *International Journal of Forecasting* 40, 3 (2024), 1055–1068.

[46] Diego Vallarino. 2024. A Dynamic Approach to Stock Price Prediction: Comparing RNN and Mixture of Experts Models Across Different Volatility Profiles. *arXiv preprint arXiv:2410.07234* (2024).

[47] Pradeep Verma. 2024. Transforming Supply Chains Through AI: Demand Forecasting, Inventory Management, and Dynamic Optimization. *Integrated Journal of Science and Technology* 1, 9 (2024).

[48] Arnoud P Wellens, Maxi Udenio, and Robert N Boute. 2022. Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods. *International Journal of Forecasting* 38, 4 (2022), 1482–1491.

[49] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. (2024).

[50] Xuguang Zhang, Pan Li, Xu Han, Yongbin Yang, and Yiwen Cui. 2024. Enhancing Time Series Product Demand Forecasting with Hybrid Attention-Based Deep Learning Models. *IEEE Access* (2024).

[51] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.

[52] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[53] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems* 36 (2023), 43322–43355.

[54] Florian Ziel. 2022. M5 competition uncertainty: Overdispersion, distributional forecasting, GAMLSS, and beyond. *International Journal of Forecasting* 38, 4 (2022), 1546–1554.