Anchored FLoE: A Business-Guardrailed Ensemble Framework of Foundation and Local-trained Models for Demand Forecasting

Haoxian Chen* (✉), Jiangwei Wang*, Xianhui Li*, Merve Kayhan Serter*, Xiangyu Meng*, Chenhong Zu, Abinaya Ulagappa, Wenfan Yu, and Michael Behrman

Amazon
{haoxianc,wangjxz,lxianhui,mkayhan,mxiangyu,melodyzu, abinayau,vickyyu,mbehrman}@amazon.com

Abstract. Accurate demand forecasting is vital for retail supply chain efficiency, yet a persistent trust-capacity gap limits industrial production to low-capacity interpretable models that fail to capture complex market dynamics. We propose Anchored FLoE, a dual-model framework that bridges this gap by fusing high-capacity deep learning with rigorous business guardrails. The framework integrates: (1) FLoE, an ensemble merging a frozen Time Series Foundation Model (Chronos-2) for global generalization with a Local Specialist (TFT) for domain-specific patterns; and (2) a model-agnostic Anchor Layer that enforces economic monotonicity and business-logic consistency through discount-segmented confidence bands. Validated across 524 product-market segments, Anchored FLoE achieves an 11-percentage-point improvement in sales-weighted MAPE over production baselines, translating to a multi-million dollar impact on free cash flow.

Keywords: Demand Forecasting · Time Series Foundation Models · Business Guardrails · Economic Monotonicity · Trustworthy AI

1 Introduction

Global retail supply chains rely on accurate demand forecasts to coordinate procurement and inventory positioning. In industrial environments, a fundamental **trust-capacity gap** persists: while deep learning models offer superior predictive capacity, business planners often rely on inherently interpretable but low-capacity models to manage inventory risks, leading to systemic failures when encountering complex, non-linear dynamics, particularly during promotion volatility and lifecycle regime shifts. A representative case where our linear production model exhibited an overforecast issue during an end-of-life (EOL) transition (detailed in Sec. 4) highlights this risk. Adaptive learning from richer source data has improved deployment under constrained inputs [4].

* Equal contribution.

Recent work addresses this via Time Series Foundation Models (TSFMs) like Chronos-2 [1] for zero-shot generalization and interpretable deep architectures (e.g., TFT [2]). However, fine-tuning TSFMs risks catastrophic forgetting [3] and production instability, while both approaches lack explicit business-logic guardrails, often yielding forecasts that violate economic principles.

To address these issues, we propose Anchored FLoE (**F**oundation-**Lo**cal **E**nsemble), which bridges the trust-capacity gap through three contributions: (1) a non-invasive dual-stream architecture fusing a frozen TSFM Global Prior with a lightweight Local Specialist to leverage global generalization while ensuring production stability; (2) a model-agnostic Anchor Layer enforcing Economic Monotonicity and business-logic consistency; and (3) large-scale industrial validation across 524 segments, achieving an 11pp wMAPE reduction and multi-million dollar free cash flow impact.

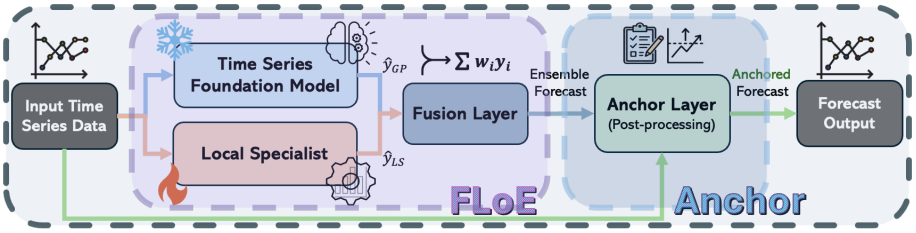

Fig. 1. The Anchored FLoE framework architecture.

2 The Anchored FLoE Framework

The Anchored FLoE framework is visualized in Fig. 1. It decouples universal forecasting intelligence from domain-specific nuances, ensuring that the high-capacity "black-box" predictions remain grounded in operational business logic.

FLoE: Complementary Knowledge Fusion. FLoE integrates two distinct intelligence streams through a non-invasive dual-stream process. To ensure production stability and avoid catastrophic forgetting, the **Global Prior** (M_{GP}) leverages a *frozen* TSFM to capture universal patterns ($\hat{y}_{GP} = f(X, Z)$), where X denotes the historical time series and Z represents exogenous features (promotions, price changes, holidays, etc.). Simultaneously, the **Local Specialist** (M_{LS}) is trained on the target dataset to capture granular domain sensitivities via $\hat{y}_{LS} = g(X, Z)$. These streams are fused via $\hat{y}_{FLoE} = \alpha\hat{y}_{GP} + (1-\alpha)\hat{y}_{LS}$. The "winning probability" of each stream during backtesting provides a principled starting point for α (e.g., $\sim$50/50 $\rightarrow \alpha = 0.5$), which we validate via ablation over backtesting versions (Fig. 2). This fusion provides mutual regularization: the Global Prior stabilizes the Specialist against noise, while the Specialist prevents over-generalization from the TSFM. A natural extension is dynamically selecting α per segment, akin to Mixture-of-Experts gating.

Anchor Layer: Business Guardrails. The model-agnostic Anchor Layer projects $\hat{y}_{FLoE}$ into a feasible space Ω defined by economic axioms. We segment sales data into market-product groups g and ordered discount bins $b_1 < b_2 < \cdots < b_n$ based on the depth of promotion (e.g., 0–10%, 10–20% off).

For each segment, we construct asymmetric bounds derived from the empirical sales distribution $y_{g,b}$ within that specific bin: $L_{g,b} = P_{10} + C_{g,b}(P_{30} - P_{10})$ and $U_{g,b} = P_{90} - C_{g,b}(P_{90} - P_{65})$, where $C_{g,b} \in [0, 1]$ is a confidence score. The confidence score averages two normalized signals—data sufficiency and demand stability—as $C_{g,b} = \frac{1}{2}\big[\min(n_{g,b}/n_{\text{full}}, 1) + (1 - \min(\text{CV}_{g,b}, 1))\big]$, where $n_{g,b}$ is the number of observed sales days (reaching full sufficiency at a threshold n_{full}) and $\text{CV}_{g,b}$ is the coefficient of variation of $y_{g,b}$. As $C_{g,b} \to 1$ (abundant, stable sales), bands tighten toward interior percentiles (P_{30}–P_{65}); as uncertainty grows ($C_{g,b} \to 0$), they widen toward the tails (P_{10}–P_{90}). To ensure operational consistency, we enforce **Economic Monotonicity** such that both lower and upper bounds are non-decreasing with deeper discounts: $L_{g,b_i} \geq \max_{j<i} L_{g,b_j}$ and $U_{g,b_i} \geq \max_{j<i} U_{g,b_j}$. Finally, the raw forecast $\hat{y}_{FLoE}$ is regularized into the final output $\hat{y}_f$ by clipping it relative to the bin-specific historical sales mean $\mu_{g,b}$: $\hat{y}_f = \text{clip}(\hat{y}_{FLoE}, \min(\mu_{g,b}, L_{g,b}), \max(\mu_{g,b}, U_{g,b}))$. This mechanism effectively grounds "black-box" neural predictions in operational reality, preventing unphysical hallucinations during extreme promotion or regime-shift events.

3 Industrial Evaluation and Impact

We evaluate Anchored FLoE on a large-scale real dataset of 5,000 products across 524 product-market segments in 10 global markets over a 0-26 week horizon. Our dataset includes noisy and irregular promotion volatility, where stochastic market shocks often lead standard architectures to produce unphysical hallucinations. We instantiate the framework using Chronos-2 [1] as the Global Prior and TFT [2] as the Local Specialist. For robust comparison, we use 19 monthly rolling backtesting versions. Performance is measured in standard industry metrics of sales-weighted Mean Absolute Percentage Error (wMAPE) and bias (wBias).

Fig. 2 shows that Anchored FLoE achieves a **10–11pp wMAPE gain** over the production baseline (37.39% → 27.31%), delivering an estimated **a multi-million dollar impact on free cash flow** by optimizing safety stock and operational capital efficiency. It significantly outperforms all deep learning baselines (e.g., DeepAR, Informer), achieving a SOTA **27.31% wMAPE** and a near-perfect **0.02% wBias** – surpassing its components: Chronos-2 (30.97%) and TFT (31.62%). This synergy confirms the mutual regularization effect.

Although the Anchor Layer is primarily designed to enforce business-logic consistency, it empirically facilitates a universal "shift toward the origin," guiding models into the optimal "Target Zone" of operational reliability (low error, near-zero bias). Specifically, it improves the production baseline's wMAPE from **37.39%** to **31.89%** and refines the FLoE's wBias to a near-perfect **0.02%**.

Deep learning baselines failures (>40% wMAPE, negative wBias) stem from models training from scratch without the prior knowledge as in TSFMs or the dynamic sensitivities (VSN) in TFT to identify features like promotions or holidays. Consequently, these models misidentify noisy, irregular promotions as periodic patterns and over-smooth forecasts, a behavior reinforced by minimizing MAE.

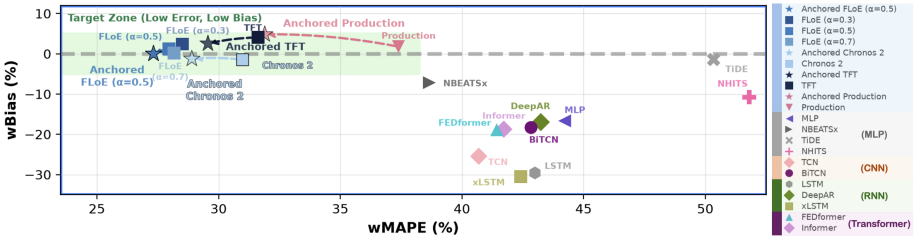

Fig. 2. 26 week horizon performance comparison across 524 product-market segments.

4 Lessons Learned and Deployment Insights

The development and deployment of Anchored FLoE in a global market provided critical insights into balancing high-capacity modeling with operational trust:

◇ **Outlier Reduction.** The framework's primary value lies in mitigating "tail risks" that erode planner trust. FLoE reduced high-error segments (wMAPE> 90%) from 14% to 9%, limiting their sales impact to 4%. This is evident in EOL transitions when deep discounts are applied to clear inventory. Our linear production models often fail to account for the regime shift, erroneously predicting a massive spike driven by the price drop, leading to a 271% wBias. Anchored FLoE successfully internalized this terminal decay, correcting the wBias to 4%. This shows that capturing non-linear dynamics is far more robust and efficient than manually tuning linear heuristics.

◇ **Model Architecture and Synergy.** Our 77k-parameter Local Specialist (hidden_size=16, n_head=2) matches Chronos-2's accuracy on internal data despite being 200× smaller, proving domain-specific models can rival massive pre-training efficiently. Moreover, empirical analysis shows their complementary failure modes exhibit a ∼ 50/50 winning probability across segments, confirming neither model subsumes the other.

◇ **Training Efficiency.** Unlike legacy models requiring frequent monthly retraining, the Local Specialist requires only annual retraining (adapting via inference-time context), with the full 524-segment pipeline completing in ∼15 mins on 4 NVIDIA L4 GPUs.

References

1. Ansari, A.F., et al.: Chronos-2: From univariate to universal forecasting. arXiv preprint arXiv:2510.15821 (2025)
2. Lim, B., et al.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int. J. Forecast. **37**(4), 1748–1764 (2021)
3. Karaouli, N., et al.: Are time series foundation models susceptible to catastrophic forgetting? arXiv:2510.00809 (2025)
4. Zhang, H., Zhan, D., Lin, Y., He, J., Zhu, Q., Shen, Z.-J., Zheng, Z.: Daily physical activity monitoring: Adaptive learning from multi-source motion sensor data. In: Proceedings of the Fifth Conference on Health, Inference, and Learning. PMLR, vol. 248, pp. 39–54 (2024)