Demand Forecasting with Foundation Models: Zero-Shot vs. LoRA Fine-Tuning on Intermittent Retail Demand

Fevzi KILAS
Hacettepe University
fevzikilas@hacettepe.edu.tr

Abstract

This project studies demand forecasting under highly intermittent and zero-inflated retail demand using time-series foundation models. We compare a strict zero-shot configuration of Chronos-2 against a parameter-efficient LoRA fine-tuned variant on the FOODS_3 subset of the M5 Forecasting Dataset. Evaluation is performed using Mean Absolute Scaled Error (MASE) and probabilistic uncertainty metrics over a 28-day horizon. Results show that Chronos-2 already achieves strong zero-shot performance (MASE < 1), while fine-tuning yields marginal gains in point accuracy and sharper prediction intervals, with limited improvement in coverage.

1 Introduction

Intermittent demand forecasting is a long-standing challenge in retail and supply chain management. Classical approaches such as Croston-style methods [3, 15] explicitly model zero-demand intervals but struggle to capture complex temporal dependencies and uncertainty.

Recent time-series foundation models trained on large and diverse corpora enable zero-shot forecasting without dataset-specific training [1, 4]. Chronos-2, in particular, has demonstrated strong generalization across domains. This project evaluates whether such pretrained representations can handle highly intermittent retail demand and whether lightweight LoRA fine-tuning [7] provides meaningful benefits beyond zero-shot inference.

2 Dataset and Preprocessing

2.1 M5 Forecasting Dataset and FOODS_3

We use the M5 Forecasting Dataset [11], accessed via the Hugging Face datasets interface [10]. The analysis focuses on the FOODS_3 department, which exhibits frequent zero-sales days and sparse demand spikes, making it a realistic proxy for perishable and agricultural supply chains.

2.2 Data Preparation

The original wide-format data is converted into a standard long format by exploding timestamps and targets. The resulting dataset contains approximately 12 million observations and is converted into an AutoGluon `TimeSeriesDataFrame` [2]. A fixed train/test split is applied, reserving the final 28 days for evaluation.

Preprint.

3 Experimental Setup

3.1 Experiment 1: Zero-Shot Chronos-2

Chronos-2 is evaluated in a strict zero-shot setting using pretrained weights only [12]. Forecasts are generated via AutoGluon TimeSeries [2], producing both point forecasts and probabilistic quantiles.

3.2 Experiment 2: LoRA Fine-Tuned Chronos-2

We fine-tune the same pretrained model using parameter-efficient LoRA adaptation [7]. The base model remains frozen, and only low-rank adapters are trained on FOODS_3. This setup allows domain adaptation under limited computational budget.

4 Evaluation Metrics

4.1 Point Forecast Accuracy

Point forecasts are evaluated using Mean Absolute Scaled Error (MASE) [8], defined as:

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|}.$$

MASE is particularly suitable for intermittent demand, where naive baselines are meaningful.

4.2 Probabilistic Forecast Evaluation

Uncertainty quality is evaluated using prediction interval width and empirical coverage [5, 6]. We focus on the 0.5–0.9 quantile interval to assess sharpness and reliability.

5 Results

5.1 Point Forecast Results

Zero-shot Chronos-2 achieves a MASE of 0.8962, indicating performance superior to a naive baseline. LoRA fine-tuning yields a slightly improved MASE of 0.8948 (-0.15%), showing no degradation but limited absolute gain.

5.2 Uncertainty Analysis

Fine-tuning reduces the average prediction interval width (3.29 $\rightarrow$ 3.19), indicating sharper forecasts. However, coverage remains around 20%, well below the nominal 40%, indicating systematic under-coverage. This behavior is consistent with known challenges in probabilistic forecasting under zero-inflated demand [14].

6 Discussion

The results indicate that Chronos-2 already captures the core structure of intermittent retail demand in a zero-shot setting. Due to the dominance of zero-demand periods, MASE quickly plateaus, and improvements from fine-tuning are inherently limited. Fine-tuning primarily refines uncertainty representation rather than substantially improving point accuracy.

From a decision-making perspective, such refinements may still be valuable for inventory control and service-level optimization [13, 16], even when average error improvements are small.

7 Limitations and Future Work

This study focuses on a single dataset subset and a single quantile range. MASE may mask poor spike timing, and prediction intervals remain under-calibrated. Future work includes alternative quantile configurations [9], post-hoc calibration, incorporation of exogenous signals, and evaluation using decision-oriented metrics.

References

[1] Abdul Fatir Ansari et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2307.03724*, 2023.

[2] AutoGluon Contributors. Autogluon-timeseries: Automl for probabilistic time series forecasting. `https://auto.gluon.ai`, 2023.

[3] John D. Croston. Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 1972.

[4] Abhimanyu Das et al. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[5] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting with neural networks. *Annual Review of Statistics*, 2014.

[6] Tilmann Gneiting and Adrian E. Raftery. Evaluating probabilistic forecasts. *JASA*, 2007.

[7] Edward J. Hu et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[8] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 2006.

[9] Roger Koenker and Gilbert Bassett. Quantile regression. *Econometrica*, 1978.

[10] Quentin Lhoest et al. Hugging face datasets. `https://github.com/huggingface/datasets`, 2021.

[11] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Results, findings, and conclusions. 2022.

[12] Amazon Science. Chronos forecasting library. `https://github.com/amazon-science/chronos-forecasting`, 2024.

[13] Edward A. Silver et al. Demand forecasting for inventory control. *Production and Operations Management*, 1998.

[14] Ralph D. Snyder et al. Modeling zero-inflated count time series. *International Journal of Forecasting*, 2012.

[15] Aris A. Syntetos and John E. Boylan. Forecasting intermittent demand: A comparative evaluation. *International Journal of Forecasting*, 2005.

[16] Bryan Wilder et al. Decision-focused learning. *IJCAI*, 2019.