Domain-adapted Lag-Llama for Time Series Forecasting in the African Retail Sector.

Kelian J.L. Massa
Isazi Consulting Pty Ltd
kelian@isaziconsulting.co.za

Dario Fanucchi
Isazi Consulting Pty Ltd
DF@isaziconsulting.co.za

Abstract

Recent advancements in time series forecasting have led to the development of foundation models, but they frequently overlook domain-specific features that are crucial for accuracy, particularly in volatile markets such as African retail. Despite the African retail sector's rapid growth, there is a lack of benchmarks and models tailored to its unique conditions. We present Lag-Llama Retail, an adaptation of Lag-Llama, a state-of-the-art foundation model, capable of effectively modelling covariates like promotions and pricing. We pretrain this model on a large-scale, private dataset comprising sales data from four African retailers over two years. Our results demonstrate significant improvements in forecasting accuracy and bias, especially in capturing sales spikes caused by promotions, compared to fine-tuned Lag-Llama, DeepAR and Temporal Fusion Transformer (TFT). This work positions Lag-Llama Retail as a new baseline for time series forecasting in the African retail sector, highlighting the potential of the approach in high-volatility settings and the limitations of foundation models lacking domain-specific covariates.

1 Introduction

In recent years, foundation models have transformed machine learning with their ability to generalize across diverse domains, achieving strong zero-shot and few-shot performance (1). Foundation models such as GPT in natural language processing and Vision Transformers in computer vision have revolutionised their respective fields (2), (3). However, the application of these foundation models to time series forecasting has lagged behind, with much of the existing research focusing primarily on univariate time series prediction without sufficient consideration of domain-specific complexities (4), (5), (6), (7), (8). More recently, promising results have emerged by incorporating covariate modelling into time series forecasting foundation models (9), (10).

Time series forecasting is particularly challenging in the retail sector, due to promotions, seasonality, and rapidly shifting consumer behaviour (11), (12). Yet achieving high accuracy is essential for aligning inventory with demand, minimizing costs, and optimizing overall supply chain efficiency (13). These challenges are magnified in African retail, where economic stress, distribution inefficiencies, and political instability further complicate forecasting (14), (15), (16). As illustrated in Figure 1 from our dataset, sales volatility is pronounced during promotional periods, highlighting the limitations of models that fail to incorporate domain-specific covariates.

Lag-Llama, a recent foundation model for time series forecasting, employs a decoder-only transformer inspired by the success of LLMs like Llama (4), (17). Similarly, concurrent works like Chronos also adapt LLM architectures with minimal changes, tokenizing time series into a discrete space (6). However, we show that in domains like African retail, where promotions and local events heavily influence sales, the standard Lag-Llama model falls short. To address this, we present Lag-Llama Retail, an adaptation of the original Lag-Llama model, by incorporating retail-specific covariates such as promotions and pricing indicators. This also enables counterfactual modelling, such as predicting

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

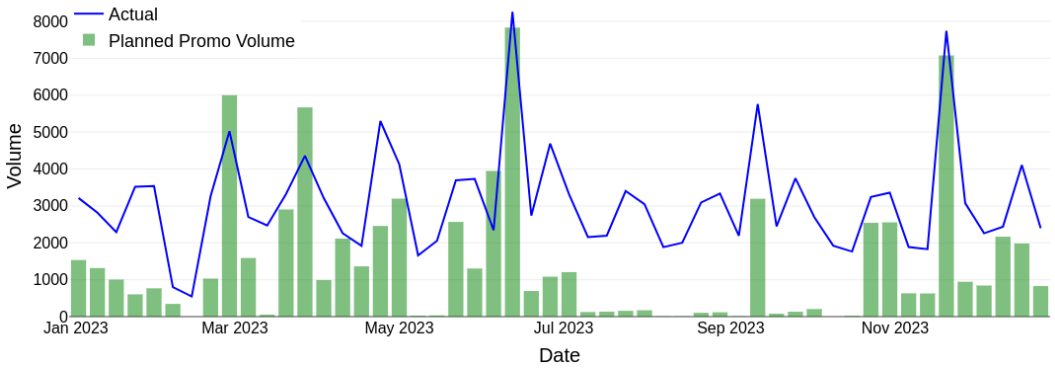


Figure 1: Time series of product sales showing volatility due to promotions.

outcomes without promotions, essential for decision-making. We adapt the model for point estimates, for both efficiency and compatibility with production retail planning systems. Retail performance is highly sensitive to point estimate forecast accuracy and bias; improved accuracy reduces working capital, and controlling bias mitigates overstocking and outages (18), (19).

Despite the growing importance of African retail (20), tailored benchmarks and models remain scarce. To address this, we validate our approach on a large-scale time series dataset comprised of sales data from four African retailers collected over two years. Although this sensitive dataset cannot be published, it highlights the unique dynamics of African retail, where forecasting requires covariates to capture high sensitivity to promotions, events, and seasonality.

Our evaluation compared several forecasting models, including zero-shot and fine-tuned versions of the original Lag-Llama (4), DeepAR (21), and TFT (22). Results show that our adapted Lag-Llama consistently achieved the highest accuracy and lowest bias, particularly in promotional scenarios where covariates are essential. While TFT performed well on non-promotional daily data, it struggled on weekly data and exhibited less reliable bias patterns. These findings establish Lag-Llama Retail as a new baseline for forecasting in African retail, as well as demonstrating the model's potential in high-volatility environments.

2 Method

2.1 Overview of Lag-LLama

Our approach builds upon the Lag-Llama model introduced by Rasul et al. (4). Lag-Llama is a foundation model for univariate probabilistic time series forecasting, using a decoder-only transformer architecture. It employs lag features and date-time information for tokenization, and outputs the parameters of a Student's t-distribution for probabilistic forecasts. The model also utilizes robust standardization to handle varying scales of data (23), as well as frequency based data augmentations for better generalisation (24). Pretrained on a diverse corpus of time series datasets, Lag-Llama demonstrates strong zero-shot generalization and few-shot adaptation capabilities on unseen datasets, often outperforming dataset-specific models when fine-tuned.

2.2 Tokenization: Incorporating retail-specific covariates

The most crucial aspect which Lag-Llama lacks for retail forecasting is the ability to consider covariates in the time-series. In particular, we have a number of promotional and price-related covariates which are available both in the past and future. For a given time series i at time step t, Lag-Llama creates a token embedding by performing a linear projection on a vector describing the time step, including the standardized target variable and its lags, the standardization factors, and time features. We incorporate additional covariates into this initial vector before linear projection, as demonstrated in Figure 2.

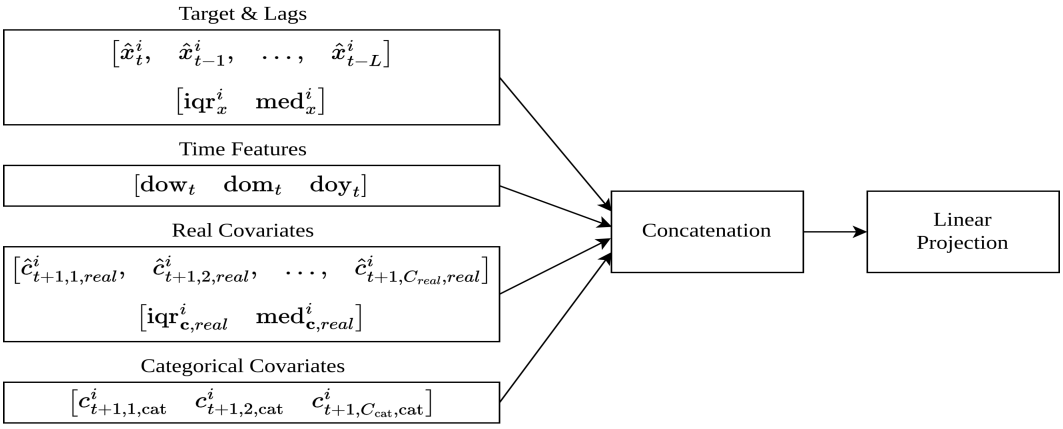


Figure 2: Tokenization scheme for incorporating additional covariates. $\hat{x}$ is the scaled target variable; dow, dom and dot are the day-of-week, day-of-month, and day-of-year time features respectively; $\hat{c}_{\text{real}}$ and c_{cat} are the scaled real-valued covariates and categorical covariates respectively; and iqr and med are the Interquartile Range and median of a feature's time series respectively.

As in the original Lag-Llama approach, standardization is performed using the robust standardization method described in (23), by removing the median and scaling by the Interquartile Range. Similarly, we scale our real-value covariates using the same robust-scaling method used in Lag-Llama, and include the scaling factors as static features. We also leverage known future covariates, such as promotions, by using the lead of the feature (e.g., c_{t+1} instead of c_t), incorporating future information into the model predictions.

Incorporating covariates in our model not only enhances accuracy but also allows us to estimate the counterfactual - what sales would have been without promotions - by setting promtional covariates to 0. This is crucial for retailers in assessing the true impact of promotions and determining how much of their sales are directly driven by these campaigns.

2.3 Optimising for point-estimates

Lag-Llama uses a distribution-head to project logits into probability distribution parameters, along with a Monte Carlo simulation approach to simulate multiple parallel paths for the forecast. This use of multiple simulated paths is resource-intensive, and makes it cumbersome for industrial use.

To improve efficiency and better suit Lag-Llama for point-estimate forecasting, we implement a greedy-search approach, assuming the expected value of the distribution as the point estimate and adding it to the model's context. These adaptations led to two open-source contributions for forecast efficiency, these are detailed in Appendix B. Additionally, we optimise the model using loss functions focused on point-estimate accuracy, specifically the mean absolute scaled error (MASE) (25), which we selected because it is scale-independent, ensuring balanced performance across multiple datasets of varying scales during pre-training.

3 Experiments

3.1 Dataset

We evaluate our approach on a large-scale dataset comprising region-aggregated point-of-sale (POS) data from four African retailers. This dataset contains approximately 15 million data points across 19,244 unique region-product pairs over two years, with records at daily or weekly intervals depending on the retailer. The dataset includes covariates such as promotions and pricing information, a full list of features available as well as sizes of datasets from each retailer is provided in Appendix A.

3.2 Data Splitting and Evaluation

We use a time-based split instead of the original Lag-Llama's zero-shot or few-shot evaluation, due to the intermittent and unpredictable nature of our retail data. The standard in retail is to use a 4-week forecast horizon as the most crucial for decision-making, we structure our data splits accordingly. We reserve two 4-week horizons for validation and another two for testing. The dataset reflects varying product demand; to align with industry standards, we evaluate the forecasting accuracy using a volume-weighted measure, prioritising more critical, high-volume products. Specifically, we use Weighted Mean Absolute Percentage Error (WMAPE) and relative bias, both detailed in Appendix C.1, which provides a more comprehensive explanation of the evaluation procedure. These metrics focus on overall accuracy while accounting for the balance between over- and under-prediction.

Additionally, we modify the validation loss function to exclude predictions on the context portion of the sequence, as training on context combined with a time-based split would lead to data leakage. To handle different frequencies in our dataset (daily vs. weekly), we maintain the same time intervals across splits, padding weekly datasets as needed to ensure a consistent prediction length during training and early stopping.

3.3 Results

Table 1 shows the results for each approach on each dataset as well as averaged across all datasets. To align with industry standards for future forecasts and avoid overestimating performance, metrics are collected with a 3-week lag. See Appendix C for a more detailed experimental analysis, including, evaluations on promotional and non-promotional data, and a qualitative study visualising aggregated and individual product forecasts. Due to computational costs, experiments were not repeated to obtain standard deviations.

Table 1: Comparison of forecasting approaches across retailers, measured by weighted mean average percentage error (WMAPE) and relative bias (RB), shown as (*WMAPE* / *RB*) in percentages. The best approach is bolded, with ties within 1% also bolded.

Approach	Daily 1	Daily 2	Daily 3	Weekly 1	Average	Inference Time (s)
DeepAR	22.0 / -6.74	**29.1** / -13.6	26.3 / -13.5	33.8 / **3.74**	27.8 / -7.54	226
TFT	**19.6 / 0.20**	30.2 / -7.28	**22.9** / -4.37	38.5 / 9.93	27.8 / **-0.38**	**81**
Lag-Llama Zero-Shot	33.3 / -22.6	42.4 / -21.3	35.5 / -15.2	48.0 / -10.5	39.8 / -17.4	1304
Lag-Llama Finetuned	21.6 / -7.5	36.1 / -18.3	**23.5** / -4.98	39.8 / -18.1	30.3 / -12.2	1334
Lag-Llama Retail	**19.4 / -0.76**	30.6 / **-3.28**	**23.9 / -0.95**	**28.5 / -4.13**	**25.6** / -2.28	144

4 Conclusion

Lag-Llama Retail consistently outperforms DeepAR, TFT, and other Lag-Llama variants across most datasets, achieving the lowest average WMAPE and significantly reduced bias, crucial for minimizing stockouts and missed sales during peak demand. While TFT has the fastest inference time, Lag-Llama Retail remains competitively fast and far outperforms the original Lag-Llama in speed, making it highly practical for real-time retail forecasting. Although TFT shows a lower average bias, this is due to opposing biases on daily (negative) and weekly (positive) data, which is less desirable for reliable forecasting accuracy, especially given its large performance drop on weekly data.

Lag-Llama Retail's ability to capture promotion-driven sales spikes further demonstrates its strength during volatile periods, as detailed in Appendix C. In non-promotional scenarios, the performance gap narrows as all models benefit from steady patterns, yet it continues to exhibit balanced accuracy and bias. These findings establish a robust baseline for time series forecasting in the African retail sector and suggest that future research should focus on refining foundation models to better understand covariates, which are essential for handling volatile cases.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2022.

[2] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2024.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[4] K. Rasul, A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, M. Biloš, S. Garg, A. Schneider, N. Chapados, A. Drouin, V. Zantedeschi, Y. Nevmyvaka, and I. Rish, "Lag-llama: Towards foundation models for probabilistic time series forecasting," in *Workshop at the Neural Information Processing Systems (NeurIPS)*, 2023.

[5] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 2024, pp. 10 148–10 167.

[6] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. Gordon Wilson, M. Bohlke-Schneider, and Y. Wang, "Chronos: Learning the language of time series," *arXiv preprint arXiv:2403.07815*, 2024.

[7] C. Chang, W.-Y. Wang, W.-C. Peng, and T.-F. Chen, "Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters," 2024.

[8] H. Liu, Z. Zhao, J. Wang, H. Kamarthi, and B. A. Prakash, "LSTPrompt: Large language models as zero-shot time series forecasters by long-short-term prompting," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 7832–7840.

[9] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified training of universal time series forecasting transformers," in *International Conference on Machine Learning (ICML)*, 2024.

[10] V. Ekambaram, A. Jati, P. Dayama, S. Mukherjee, N. H. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam, "Tiny Time Mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series," *arXiv preprint arXiv:2401.03955*, 2024.

[11] R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283–1318, 2022, special Issue: M5 competition.

[12] J. Wolters and A. Huchzermeier, "Joint in-season and out-of-season promotion demand forecasting in a retail environment," *Journal of Retailing*, vol. 97, no. 4, pp. 726–745, 2021, sI: Metrics and Analytics.

[13] M. Fisher and A. Raman, "Using data and big data in retailing," *Production and Operations Management*, vol. 27, no. 9, pp. 1665–1669, 2018.

[14] D. Kuteyi and H. Winkler, "Logistics challenges in sub-saharan africa and opportunities for digitalization," *Sustainability*, vol. 14, no. 4, 2022.

[15] M. Nieuwenhuyzen, W. Niemann, and T. Kotzé, "Supply chain risk management strategies: A case study in the south african grocery retail industry," *Journal of Contemporary Management*, vol. 15, pp. 784–822, 2018.

[16] D. Essers, "Developing country vulnerability in light of the global financial crisis: shock therapy," *Review of Development Finance*, vol. 3, no. 1, pp. 61–83, 2013.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[18] F. Petropoulos, D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. Cyrino Oliveira, S. De Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. Trapero Arenas, X. Wang, R. L. Winkler, A. Yusupova, and F. Ziel, "Forecasting: theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705–871, 2022.

[19] R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283–1318, 2022, special Issue: M5 competition.

[20] African Development Bank, *African Economic Outlook 2024: Driving Africa's Transformation: The Reform of the Global Financial Architecture*. African Development Bank, 2024. [Online]. Available: https://www.afdb.org/en/documents/african-economic-outlook-2024

[21] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[22] B. Lim, S. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[23] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaa, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer, 2005, vol. 488.

[24] M.-H. Chen, Z. Xu, A. Zeng, and Q. Xu, "Fraug: Frequency domain augmentation for time series forecasting," *arXiv preprint arXiv:2302.09292*, 2023.

[25] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *FInternational Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

A Dataset Characteristics

The dataset characteristics for each retailer are presented in Table 2.

Table 2: Dataset characteristics for each retailer

Dataset	Number of Data Points	Number of Promotions	Number of Regions	Number of Products	Number of Time Series
Daily 1	12,359,444	1,948,720	29	992	15,851
Daily 2	280,815	127,675	1	375	375
Daily 3	1,953,656	91,617	4	888	2,611
Weekly 1	44,029	25,234	1	407	407
Total	**14,637,944**	**2,193,246**	**35**	**2,662**	**19,244**

The full list of dataset features utilised in forecasting, available for each retailer:

- **sales_date**: The date of sale, indicating when the transaction occurred.
- **product_code**: A unique identifier for each product.
- **region_code**: A unique identifier for each region, each region-product pair corresponds to unique time series.
- **volume**: The quantity of the product sold on a given day. This is the response variable that we aim to forecast.
- **relative_price**: A relative measure of the price of the product relative to other products in the dataset.
- **is_promo**: A binary flag (1 or 0) indicating whether a promotion was applied to the product on that day. A value of 1 indicates a promotion was active, while 0 indicates no promotion or the promotion was invalid.
- **is_single_price_promo**: A binary flag indicating whether the promotion was a single product discount, such as a percentage off the retail price. This is the most common type of promotion.
- **is_multibuy_promo**: A binary flag indicating whether the promotion required purchasing multiple items, such as a "buy 2 get 1 free" offer, which is designed to encourage bulk purchases.
- **rel_promo_price**: The relative price of the product during a promotion compared to the standard retail price. This feature helps to quantify the discount provided during promotional periods.
- **planned_promo_vol**: An estimate of the sales volume that the retailer planned to achieve through the promotion.
- **promo_strength**: A representation of the strength of a promotion between 0 and 1 using the relative promotional price, with 0 meaning no promotion and 1 representing a 100% discount.

B A more efficient approach to probabilistic forecasting

Lag-Llama uses a distribution-head to project logits into probability distribution parameters. A Monte Carlo simulation approach is then used for probabilistic forecasting: the underlying distribution is sampled n times, each sample is added to the model's context, and the process is repeated for p prediction steps, resulting in n parallel paths, each of length p, which represent possible forecast outcomes. While effective, this method is resource-intensive, increasing the forward passes by a factor of n for a total of $p \times n$ forwarded passes.

To address this, we propose a more efficient approach by directly leveraging the distribution-head's output. Instead of maintaining n parallel paths, we construct a single sample path using a type of greedy search, where each step represents the expected value (the mean of the Student t-distribution). At each time step, n samples are drawn from the distribution head and stored in memory, allowing for only p forward passes while maintaining n samples per step. This method, implemented as 'single-pass-sampling,' is available through our open-source contribution to the Lag-Llama repository

[1]. We also fixed the erroneous key-value (KV) cache implementation in the original repository through an additional open-source contribution, further improving efficiency [2].

C Detailed Experimental Analysis

C.1 Performance Metrics

Our evaluation uses two key performance metrics: Weighted Mean Absolute Percentage Error (WMAPE) and relative bias, both weighted by volume to reflect the significance of higher-volume products, which are critical to retailers. These metrics ensure our evaluation aligns with industry priorities, where both overall accuracy and the balance between over- and under-prediction are critical for decision-making.

One crucial factor which we also consider in our evaluation is the fact that while we collect data at a region-level, our retailers need to order stock on a weekly basis aggregated across all regions. Thus, we first aggregate volumes by week date and product code, before computing the WMAPE. This allows errors to cancel out between the same product and week in different regions. For our daily datasets, this also means that errors may cancel between different days of the same week.

Additionally, inference time is measure for a measure of computational efficiency. Inference time as reported in Tab. 1 is totaled across all forecasts and was captured on the following machine: 1x A10 (24 GB PCIe), 30 CPU cores, 205.4 GB RAM, 1.5 TB SSD.

C.1.1 WMAPE

WMAPE measures the accuracy of the forecasts by calculating the weighted average of the absolute percentage errors accross all time steps (not averaged per time series), with weights proportional to the actual sales volume. It is computed as:

$$\text{WMAPE} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}|\mathrm{A}_{i,t} - \mathrm{F}_{i,t}|}{\sum_{t=1}^{T}\sum_{i=1}^{N}\mathrm{A}_{i,t}}, \tag{1}$$

C.1.2 Relative Bias

Understanding model bias is crucial for retail applications, where stock ordering relies on accurate forecasts, and systematic bias can lead to costly inefficiencies. Relative bias directly measures the tendency of the model to over- or under-predict sales by simply dividing the total forecasted volume sold by the actual volume sold, calculated as:

$$\text{Bias} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}\mathrm{F}_{i,t}}{\sum_{t=1}^{T}\sum_{i=1}^{N}\mathrm{A}_{i,t}}. \tag{2}$$

A relative bias of 0 indicates no bias, while values greater than or less than 0 indicate consistent over- or under-prediction, respectively.

C.2 Performance On Promotional and Non-Promotional Data

While the main evaluation focused on overall performance, retailers also need insights into specific contexts like promotions and non-promotions. Tables 3 and 4 provide a comparative analysis of model performance specifically on entries with and without promotions, respectively.

On promotional data, Lag-Llama Retail consistently outperforms other models, showing significantly better accuracy and reduced bias, which is expected given its ability to incorporate promotion-specific covariates. TFT and DeepAR also perform well, indicating the effectiveness of covariate integration in handling promotional variability. In contrast, the original Lag-Llama models, especially in zero-shot mode, exhibit strong negative bias during promotions, likely due to their inability to account for

[1] `https://github.com/time-series-foundation-models/lag-llama/pull/77`

[2] `https://github.com/time-series-foundation-models/lag-llama/pull/84`

Table 3: Comparison of forecasting approaches across retailers, focusing only on promotional data entries. Approaches are measured by weighted mean average percentage error (WMAPE) and relative bias (RB), shown as (*WMAPE* / *RB*) in percentages. The best approach is bolded, with ties within 1% also bolded. Metrics are collected with a 3-week lag.

Approach	Daily 1	Daily 2	Daily 3	Weekly 1	Average
DeepAR	26.2 / -3.49	**29.6** / -18.7	32.8 / -22.8	27.6 / **-2.39**	29.1 / -11.3
TFT	23.3 / **-1.58**	31.2 / -12.3	**27.1** / -17.0	34.7 / **1.62**	29.1 / -7.32
Lag-Llama Zero-Shot	40.4 / -32.2	43.8 / -38.5	46.2 / -39.0	42.3 / -30.17	43.2 / -34.0
Lag-Llama Finetuned	26.5 / -14.4	37.5 / -32.5	35.0 / -24.4	38.4 / -30.7	34.4 / -25.5
Lag-Llama Retail	**21.6 / -2.24**	**29.8 / -3.59**	31.8 / **-7.71**	**25.5** / -7.71	**27.2 / -5.31**

Table 4: Comparison of forecasting approaches across retailers, focusing only on non-promotional data entries. Approaches are measured by weighted mean average percentage error (WMAPE) and relative bias (RB), shown as (*WMAPE* / *RB*) in percentages. The best approach is bolded, with ties within 1% also bolded. Metrics are collected with a 3-week lag.

Approach	Daily 1	Daily 2	Daily 3	Weekly 1	Average
DeepAR	18.0 / -12.0	**28.3** / -4.78	26.1 / -12.4	47.8 / 18.4	30.1 / **-2.70**
TFT	**16.4 / 3.20**	**28.9 / 1.75**	**22.9 / -2.71**	47.5 / 29.2	28.9 / 7.82
Lag-Llama Zero-Shot	25.0 / -7.70	39.9 / 7.94	34.7 / -12.1	63.5 / 36.4	40.8 / 16.3
Lag-Llama Finetuned	19.2 / **4.64**	34.2 / 9.82	**23.4 / -2.50**	42.7 / 13.3	29.9 / 9.47
Lag-Llama Retail	18.8 / 5.82	32.2 / **-2.74**	**23.7 / 2.09**	**35.3 / 4.18**	**27.5 / 2.34**

promotion-driven spikes, resulting in predictions that are closer to a baseline, agnostic of these events. Consistent with the overall results, TFT's performance drops significantly on the weekly dataset for both promotions and non-promotions, highlighting its limitations in capturing multi-frequency trends, which Lag-Llama models handle more robustly.

On non-promotional data, the performance gap between Lag-Llama Retail and the original Lag-Llama models narrows, with the fine-tuned Lag-Llama achieving similar accuracy and sometimes surpassing DeepAR, particularly in terms of bias reduction. TFT achieves the best performance on non-promotional daily data, highlighting its strength in handling steady trends without promotional noise. This overall narrowing in performance suggests that while covariates are crucial for capturing promotional effects, as expected, their impact is less pronounced in non-promotional contexts across all models.

Additionally, the original Lag-Llama models demonstrate a pattern of predicting with strong negative bias during promotions and a weaker but still significant positive bias for non-promotions. This indicates that the original model isn't simply predicting a baseline; instead, it's predicting something in between promotions and non-promotions. This behaviour is undesirable for retailers, as it fails to adequately distinguish between the different contexts, leading to suboptimal stock management decisions. These results underscore the importance of domain-specific covariates in improving forecasting accuracy, particularly in contexts with significant sales fluctuations.

C.3 Qualitative Evaluation

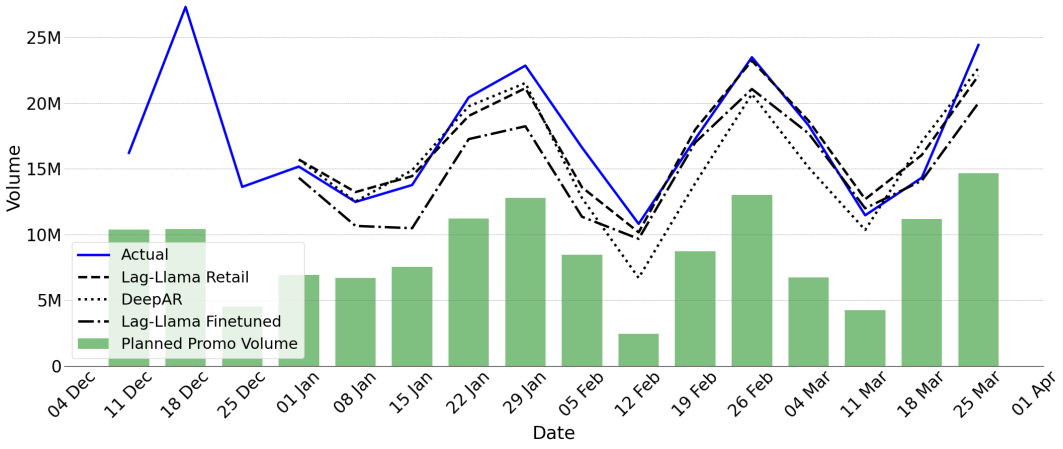
To further explore model performance, we conduct a qualitative analysis. To understand potential sources of bias, Figure 3 shows aggregated forecasts across all products and regions for each retailer. The daily forecasts are aggregated to weekly forecasts, aligning with our weekly-level metric evaluation. As in the quantitative evaluation, forecasts are shown with a lag of 3 weeks, meaning for

each predicted time step, the model had access to 3-week old actual data. The visualised forecasts contain 13 weeks of predictions over both the validation and test sets (16 weeks total, with 3 weeks ignored due to lag).

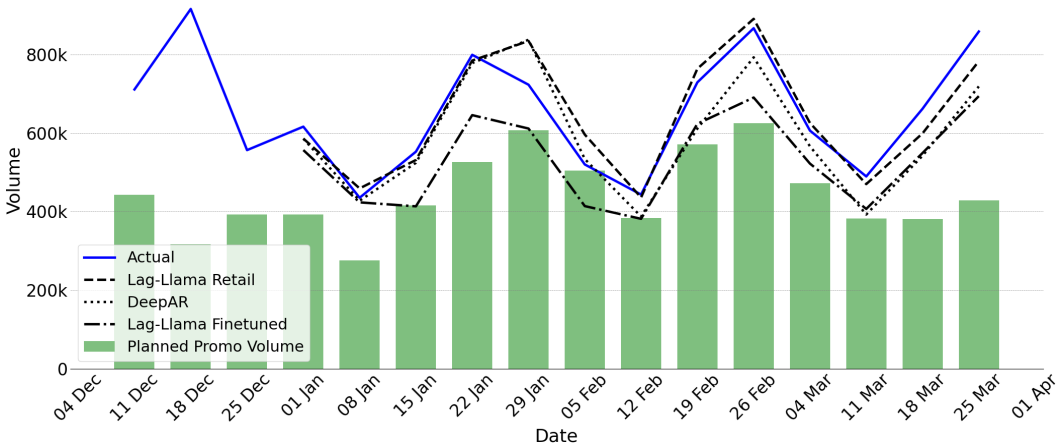
For this analysis, we selected DeepAR as a baseline and the fine-tuned Lag-Llama as a representation of the original model our adaptation builds upon, as including additional models would overly clutter the visualizations.

The aggregated forecast visualizations provide a number of insights into the performance of the models. In Figure 3c, we observe almost non-existent planned promotional volumes, which was due to the limited promotional information provided by the retailer. This explains why the original finetuned Lag-Llama achieves similar performance to our Lag-Llama Retail model, as there are few promotions to leverage. Importantly, this also demonstrates that our model remains robust even when promotional data is sparse, as indicated by the closely aligned total forecasts. This is not the case with DeepAR, which comparatively shows far more negative bias in this scenario.

Across all datasets, the performance differences between the finetuned Lag-Llama and our Lag-Llama Retail variant are most pronounced during sales spikes, typically corresponding with end-of-month promotions. The original Lag-Llama consistently underpredicts these spikes, showing a strong negative bias, whereas our model exhibits far less negative bias. Although both models predict sales spikes, the magnitude is more accurately captured by our model.

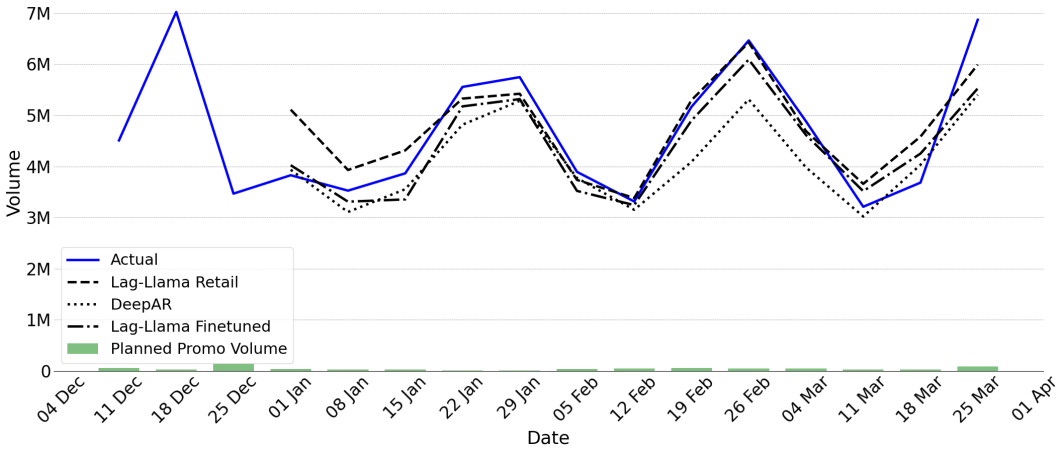


(a) Aggregated weekly forecasts for Daily Retailer 1, derived from daily forecasts and aggregated across all products and regions.

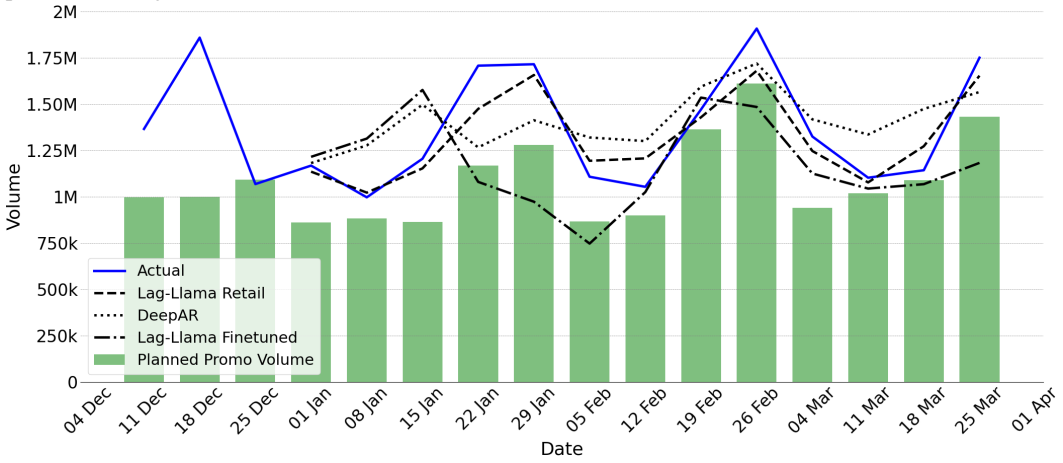


(b) Aggregated weekly forecasts for Daily Retailer 2, derived from daily forecasts and aggregated across all products and regions.

Figure 3: Aggregated weekly forecasts derived from daily and weekly forecasts across all products and regions for various retailers. Each time step is predicted at a 3-week lag.



(c) Aggregated weekly forecasts for Daily Retailer 3, derived from daily forecasts and aggregated across all products and regions.



(d) Aggregated weekly forecasts for Weekly Retailer 1, derived from weekly forecasts and aggregated across all products and regions.

Figure 3: Aggregated weekly forecasts derived from daily and weekly forecasts across all products and regions for each retailer. Each time step is predicted at a 3-week lag. (Continued)

The shape of our forecasts closely follows the planned promotional volumes, suggesting that while our model is generally accurate, it may be vulnerable to errors in promotional planning. For instance, in Figure 3b, a spike in actual sales occurs a week before the planned promotional volume spike, leading our model to forecast a spike one week late, whereas the original Lag-Llama correctly predicts the timing, though not the magnitude, of the spike.

In Figure 3d, both DeepAR and the original Lag-Llama perform poorly, likely due to the limited amount of weekly data available for training. However, models incorporating covariates, including Lag-Llama Retail and DeepAR, perform significantly better, closely following the planned promotional volumes. This suggests that our model may be more adept at handling multi-frequency data, as it provides accurate forecasts despite the scarcity of weekly data (about 90% of the time series are daily), although further data is needed to confirm this.

Finally, we observe a general trend of underprediction rather than overprediction, likely due to the spikes and volatility in the data. This may indicate that the models are learning a more conservative baseline, although the exact reasons for this behaviour require further investigation.