RAF: A Model Agnostic Framework for Retrieval Augmented Zero Shot Time Series Forecasting

Md Younus Ahamed
LCSEE
West Virginia University
Morgantown, WV 26505
ma00087@mix.wvu.edu

Abstract

Foundation models have recently advanced zero shot time series forecasting, offering the ability to generalize without task specific training. However, in healthcare settings, where data are highly heterogeneous, exhibit regime shifts, and often contain rare but clinically critical events, these models frequently underperform. We propose Retrieval Augmented Forecasting (RAF), a model agnostic framework that strengthens foundation model predictions. RAF constructs a bank of past trajectories, retrieves nearest neighbor continuations using Euclidean or Dynamic Time Warping similarity, and blends them with foundation model forecasts through a data driven weighting scheme. The method requires no architectural changes, making it readily deployable within existing clinical forecasting pipelines. Across physiological and epidemiological datasets, including vital signs and hospital admission series, RAF consistently improves zero shot accuracy for four state of the art foundation models (Chronos, Lag Llama, MOMENT, and Toto). These gains highlight retrieval augmentation as a lightweight yet effective strategy for enhancing the robustness and clinical utility of time series foundation models in health applications. GitHub repository: https://github.com/shuvo14051/TS4H-Workshop-NeurIPS-2025.

1 Introduction

Time series forecasting is a longstanding challenge in machine learning, central to domains such as healthcare, finance, and environmental monitoring [1, 2, 3]. Traditional approaches typically learn predictors from scratch for each dataset, limiting transferability and requiring substantial task specific tuning [3]. Recent advances in *foundation models* have enabled zero shot forecasting directly generating predictions for previously unseen series or tasks without retraining [4, 5]. These models, informed by large scale pretraining across heterogeneous domains, promise robustness to data variety, simplified operational workflows, and the ability to generalize across environments [6, 4]. Notable examples, including Chronos[7], Lag Llama[8], MOMENT[9], and Toto[10], have demonstrated strong performance on diverse benchmarks [6]. However, zero shot foundation models can struggle with dataset specific dynamics, abrupt regime shifts, and rare events, particularly when in domain training data diverges from pretraining distributions [5, 11]. To bridge these gaps, retrieval augmentation has emerged as a powerful paradigm. Retrieval based methods ground predictions in relevant historical analogs, effectively blending the inductive biases of statistical and neural models with nonparametric memory [12, 11, 13, 14]. By supplementing model forecasts with continuations of similar past patterns, retrieval mechanisms have improved generalization and robustness, especially in settings with distributional shifts or unforeseen behaviors [12, 11].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

In this work, we propose **Retrieval Augmented Forecasting (RAF)**, a model agnostic framework that enhances foundation model forecasts by blending them with neighbor based analogs retrieved from the training data. RAF requires no modifications to model architecture, is easily integrated into existing pipelines, and achieves consistently improved zero shot performance across multiple benchmarks. Our contributions add to the growing evidence that retrieval augmentation is a key ingredient for robust and adaptive time series foundation models [12, 13].

2 Methodology

2.1 Problem Setup

We study the univariate time series forecasting problem. Let

$$y = (y_1, y_2, \ldots, y_T)$$

denote the historical series, where $y_t \in \mathbb{R}$. The objective is to forecast the next H steps,

$$\hat{y}_{T+1:T+H} = (\hat{y}_{T+1}, \ldots, \hat{y}_{T+H}).$$

Forecast accuracy is measured using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

2.2 Dataset

We evaluate RAF's performance on a broad collection of real world datasets covering medical, epidemiological, and physiological domains. The physiological group includes Heart Rate, Respiration, Temperature, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Glucose 1, and Glucose 2, which capture cardiovascular, respiratory, and metabolic activity at temporal resolutions ranging from seconds to minutes. The epidemiological group comprises Hospital Admission and ICU Admission, which track patient inflows during the COVID 19 pandemic. Collectively, these datasets exhibit heterogeneous temporal dynamics, including periodic rhythms, abrupt regime changes, and variable noise levels. Such diversity provides a rigorous and comprehensive basis for assessing RAF's capacity to generalize across distinct application domains.

2.3 Retrieval Augmented Forecasting (RAF)

Our central contribution is a model agnostic retrieval augmentation mechanism that improves zero shot performance of foundation models. RAF supplements the forecast from a base model with information retrieved from similar historical patterns.

2.3.1 Window Bank Construction

Given a historical sequence y, we construct a window bank $\mathcal{B}$ of input output pairs using sliding windows of length L and horizon H:

$$\mathcal{B} = \{(x_i, z_i) \mid x_i = y_{i:i+L1},\ z_i = y_{i+L:i+L+H1}\}.$$

Each input window is standardized via z score normalization,

$$\tilde{x}_i = \frac{x_i \mu(x_i)}{\sigma(x_i)},$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote mean and standard deviation.

2.3.2 Nearest Neighbor Retrieval

At prediction time, the most recent context $q = y_{TL+1:T}$ is normalized and compared with stored windows using either Euclidean distance or Dynamic Time Warping (DTW):

$$d(\tilde{q}, \tilde{x}_i) = \begin{cases} \text{DTW}(\tilde{q}, \tilde{x}_i), & \text{if DTW enabled,} \\ \|\tilde{q}\tilde{x}_i\|_2, & \text{otherwise.} \end{cases}$$

The k nearest neighbors are retrieved, and their future continuations $\{z_i\}$ are rescaled to match the mean and variance of q. The neighbor based forecast is then

$$\hat{y}^{\text{NB}} = \frac{1}{k}\sum_{i=1}^{k} \alpha_i z_i,$$

where scaling factor α_i ensures variance alignment.

2.3.3 Blending with Base Forecasts

Let $\hat{y}^{\text{FM}}$ denote the forecast from a foundation model. RAF blends it with the neighbor forecast:

$$\hat{y} = \lambda \hat{y}^{\text{FM}} + (1\lambda)\hat{y}^{\text{NB}}.$$

The blending parameter $\lambda \in [0, 1]$ is tuned by grid search on a validation slice of the training set to minimize one step ahead error:

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \sum_t \left| y_{t+1} \left(\lambda \hat{y}^{\text{FM}}_{t+1} + (1\lambda)\hat{y}^{\text{NB}}_{t+1} \right) \right|.$$

2.4 Foundation Models

We apply RAF across four state of the art foundation models. Chronos is a transformer based temporal model pretrained on large scale time series data. Lag Llama is a lag based autoregressive model designed with efficient scaling properties. MOMENT is a masked modeling architecture tailored for multiscale forecasting tasks. Toto is a foundation model specialized for telemetry and irregularly sampled sequences. Each model is wrapped into a unified interface predict_fn(y, H) that produces $\hat{y}^{\text{FM}}$, and RAF is applied without altering the internal mechanisms of these models.

3 Results

We evaluate the impact of Retrieval Augmented Forecasting (RAF) across four foundation models: Chronos, Lag Llama, MOMENT, and Toto. Tables 1-3 report mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) across ten benchmark datasets. To facilitate cross dataset comparison, we additionally compute the average rank of each method.

Table 1: MAE Outcomes: Assessing Forecast Accuracy with and without RAF

Model	Cardio	Covid Hospital	Covid ICU	DBP	Glucose 1	Glucose 2	Heart	Resp	SBP	Temp	**Avg. Rank**
Toto	3.52	321.46	43.42	8.69	8.48	7.35	9.24	1.60	13.00	0.37	3.1
RAF+Toto	3.62	315.26	43.18	9.03	8.34	7.24	9.39	1.54	12.95	0.38	**2.9**
Lag Llama	10.05	2735.99	978.39	8.90	24.56	7.40	10.15	1.67	12.38	0.39	5.4
RAF+Lag Llama	7.82	7059.47	1578.53	8.57	40.01	7.39	9.55	1.50	12.46	0.36	4.3
Chronos	4.16	238.39	34.95	10.20	12.33	8.15	9.71	1.60	13.12	0.39	4.8
RAF+Chronos	3.90	122.41	25.38	8.68	11.64	7.18	9.49	1.63	13.51	0.45	3.8
MOMENT	10.47	11580.28	1421.41	9.71	31.10	8.44	9.32	1.73	13.37	0.38	6.3
RAF+MOMENT	9.70	11533.08	1439.75	8.06	31.79	7.29	9.85	1.54	12.61	0.39	4.8

Table 1 summarizes mean absolute error (MAE) across datasets. RAF consistently improves the performance of Chronos, yielding substantial reductions on high variance series such as Hospital Admission and ICU Admission. For MOMENT, RAF provides gains on several physiological signals (e.g., diastolic blood pressure) but does not improve epidemiological series, with ICU Admission showing a slight degradation. For Toto, RAF produces modest yet consistent improvements across domains. Lag Llama also benefits in terms of average rank, though its performance on epidemiological datasets remains unstable. The rank analysis confirms overall effectiveness: RAF achieves better ranks than the baseline models, with Chronos moving from 4.8 to 3.8, MOMENT from 6.3 to 4.8, Toto from 3.1 to 2.9, and Lag Llama from 5.4 to 4.3.

Table 2 reports root mean squared error (RMSE) across datasets. The results align closely with the MAE analysis, demonstrating that RAF systematically enhances baseline models in terms of average rank. Chronos exhibits the most consistent improvement, while Toto records modest yet reliable gains. MOMENT shows marked advancement, transitioning from the weakest baseline performance to a substantially stronger position when augmented with RAF. Lag Llama remains the most variable;

Table 2: RMSE Outcomes: Evaluating Model Stability across Datasets

Model	Cardio	Covid Hospital	Covid ICU	DBP	Glucose 1	Glucose 2	Heart	Resp	SBP	Temp	**Avg. Rank**
Toto	4.07	425.37	49.81	10.63	9.59	8.91	10.56	2.07	15.01	0.55	4.0
RAF+Toto	4.25	430.35	48.41	10.90	9.43	8.84	10.67	1.99	14.84	0.54	**3.6**
Lag Llama	10.92	3133.95	1082.99	10.89	26.51	9.16	11.63	2.08	14.51	0.55	6.6
RAF+Lag Llama	8.76	7118.93	1640.34	10.33	40.89	9.31	11.05	1.94	14.87	0.53	4.6
Chronos	4.86	267.90	40.83	12.11	14.36	9.36	11.13	2.04	15.39	0.56	5.6
RAF+Chronos	4.55	155.44	30.25	10.33	13.08	9.35	10.81	2.07	15.93	0.61	5.0
MOMENT	11.41	11691.31	1436.08	11.32	32.94	10.15	10.68	2.11	15.84	0.55	7.8
RAF+MOMENT	10.67	11663.08	1459.58	9.98	34.66	8.99	11.16	2.00	14.85	0.55	4.8

although its error magnitudes fluctuate across datasets, its overall rank nonetheless improves from 6.6 to 4.6. These findings suggest that RAF not only improves predictive accuracy but also contributes to greater stability across models with differing baseline characteristics.

Table 3 reports MAPE, which captures relative forecasting accuracy. RAF substantially improves Chronos (average rank 5.1 → 3.6) and MOMENT (6.4 → 4.9). Toto exhibits marginal improvements, whereas Lag Llama continues to show mixed behavior. Importantly, the consistency of average rank improvements across metrics underscores that RAF provides a net accuracy benefit across heterogeneous datasets, with the strongest relative gains observed for Chronos.

Table 3: MAPE Outcomes: Relative Accuracy Gains through RAF Integration

Model	Cardio	Covid Hospital	Covid ICU	DBP	Glucose 1	Glucose 2	Heart	Resp	SBP	Temp	**Avg. Rank**
Toto	4.52	6.14	5.74	11.29	4.59	8.29	14.56	10.04	10.64	1.01	3.3
RAF+Toto	4.63	6.07	5.71	11.78	4.51	8.15	14.83	9.82	10.60	1.04	**3.2**
Lag Llama	13.05	51.17	143.63	11.60	13.37	8.14	15.99	10.55	10.04	1.07	5.3
RAF+Lag Llama	10.20	125.49	225.84	11.13	21.82	8.12	15.16	9.45	10.32	0.99	4.1
Chronos	5.37	4.29	4.83	13.32	6.67	9.10	15.38	10.28	10.81	1.06	5.1
RAF+Chronos	4.99	2.16	3.37	11.23	6.30	7.83	15.12	10.09	10.97	1.21	3.6
MOMENT	13.59	203.39	196.52	12.61	16.89	9.51	14.69	10.85	11.00	1.04	6.4
RAF+MOMENT	12.61	203.16	201.39	10.49	17.26	8.20	15.42	9.58	10.31	1.05	4.9

Overall, the results show that RAF improves zero shot forecasting across all four evaluated foundation models. By coupling retrieval based historical continuations with foundation model predictions, RAF consistently lowers error rates and yields stronger average rankings. These relative gains highlight RAF's effectiveness as a model agnostic augmentation strategy, demonstrating benefits for every baseline model considered. RAF consistently strengthens every foundation model we tested, proving itself a powerful augmentation for zero shot time series forecasting.

4 Conclusion

We presented Retrieval Augmented Forecasting (RAF), a model agnostic framework that systematically enhances the zero shot performance of time series foundation models. By combining base forecasts with neighbor based continuations retrieved from historical trajectories, RAF achieves consistent improvements in accuracy across multiple state of the art models. These results underscore the value of retrieval as a lightweight mechanism for complementing large pretrained architectures without altering their internal design or requiring task specific training.

Future research will extend the scope of RAF along several dimensions. An immediate direction is the evaluation of RAF in multivariate forecasting settings, where dependencies among variables introduce additional challenges. Beyond forecasting, RAF can be explored for other time series learning tasks such as classification, anomaly detection, and representation learning. Another promising direction is to study the interaction between retrieval augmentation and fine tuning, assessing whether RAF provides additive benefits when foundation models are adapted to specific domains.

References

[1] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[2] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

[3] Jason Brownlee. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2022.

[4] Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Xinjiang Lu, Le Zhang, and Hui Xiong. Timefound: A foundation model for time series forecasting. *arXiv preprint arXiv:2503.04118*, 2025.

[5] Yunkai Zhang, Hyungjin Woo, Ziyang Liu, and Luke et al. Darlow. Are time series foundation models ready for zero-shot forecasting? In *ICML Workshop on Foundation Models for Structured Data*, 2025.

[6] Parseable. Zero-shot forecasting: Our search for a time-series foundation model. `https://www.parseable.com/blog/zero-shot-forecasting`, 2025. Blog Post.

[7] 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. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[8] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[9] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[10] Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024.

[11] Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Y Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. 2025.

[12] Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan Ö. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *International Conference on Machine Learning (ICML)*, 2025.

[13] Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. volume 37, pages 2766–2786, 2024.

[14] The Moonlight. Literature review: Retrieval augmented time series forecasting, 2025. `https://www.themoonlight.io/en/review/retrieval-augmented-time-series-forecasting`.