Zero Shot Time Series Forecasting: Do Time Series FMs Outperform Cross Modal FMs?

Anonymous Author(s)

Abstract

Foundation models (FMs) have achieved major advances in language, vision, and speech. In parallel, time series foundation models (TSFMs) have been developed to address forecasting tasks. A key question is whether TSFMs truly generalize to unseen time series data, and whether they perform better than general purpose FMs from other domains in a zero shot setting. We compare four TSFMs such as Chronos, TimesFM, TimeGPT, and MOMENTs with cross domain FMs for text (GPT), audio (Whisper), and vision (ViT). For a systematic comparison, we use simple task-agnostic adapters to convert sequences into forecasts, without fine tuning or changing the backbone models. All models are evaluated on nine diverse datasets that were unseen during training. Our results show that TSFMs perform best on most datasets, highlighting the benefit of temporal pretraining and time-aware design. Overall, the strong zero shot performance of TSFMs suggests that they may represent a breakthrough comparable to BERT for time series forecasting. At the same time, large text based models such as GPT remain surprisingly competitive, in some cases even surpassing TSFMs, highlighting the ability of general purpose models to capture temporal patterns despite not being trained for this task. GitHub repository: https://github.com/anonymous4865/tsfms.

1 Introduction

Foundation models are large, general purpose neural networks trained on unlabeled data from diverse domains, and they represent a transformative shift in machine learning [1]. They have achieved major success in natural language processing and computer vision, showing strong zero shot and few-shot capabilities across tasks [2, 3]. These advances have reshaped the field by enabling a pretrained approach and have inspired time series research to move from narrow, task specific models toward foundation models that support zero shot learning and cross domain generalization [2, 4].

The emergence of TSFMs develops from the limits of traditional time series forecasting, where each dataset required a separate model, preventing the use of large pre-trained models as in other domains [5, 4]. The defining characteristic of TSFMs is their ability to perform zero shot forecasting for new datasets from limited context, without retraining or fine-tuning [6, 7]. This capability has spurred the development of notable models including TimeGPT [8], the first time series foundation model; MOMENT [7], designed to solve multiple time series tasks in one framework; TimesFM [9], a decoder-only transformer with 200 million parameters that uses patched-decoder attention and large-scale pretraining to capture temporal patterns; and Chronos [10], an encoder-decoder model based on the T5 architecture that tokenizes time series values through scaling and quantization to build a fixed vocabulary. These transformer-based designs also use patching techniques to manage long sequences efficiently. Beyond the early models, newer designs such as Lag-Llama [1], Moments [7] and Moirai [5] extend the field further, combining transformer architectures with diverse training corpora to improve forecasting accuracy and flexibility. Together, these models highlight the future of time series analysis by giving researchers and practitioners stronger tools to handle complex

Under review at the NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S). Do not distribute.

data. However, applying foundation models directly to time series remains challenging due to the heterogeneity among time series datasets, leading to the proposal of various specialized TSFM architectures.

Recent benchmark studies shows the competitiveness of the TSFM compare to the traditional time series model. For example, GIFT-Eval [11] reports that TSFMs can deliver strong results on multivariate forecasting. Similarly, FoundTS [12] finds that lightweight supervised baselines remain competitive with state-of-the-art TSFMs. Despite rapid progress, the effectiveness of TSFMs relative to cross domain foundation models remains an open question. Large language models (LLMs) such as GPT, audio models such as Whisper, and vision transformers (ViTs) have demonstrated surprising cross-modal generalization. Although pretrained on text, speech, or images, they exhibit the ability to process sequential patterns when adapted to time series tasks. This raises a critical and timely question: do domain-specific TSFMs truly outperform general purpose FMs in zero shot time series forecasting, or can cross domain FMs match or exceed their performance?

The main contribution of this paper is the **first zero shot comparison of TSFMs and cross domain models for time series forecasting**, achieved by benchmarking seven models, including four TSFMs and three cross domain models, **on nine unseen datasets**.

2 Method

In this work, we evaluate zero shot forecasting by applying time series–specific and cross domain foundation models through a standardized adapter interface, enabling a controlled comparison on unseen data.

2.1 Datasets

We evaluate model performance on nine publicly available univariate time series datasets spanning a diverse range of domains, including weather, transportation, retail, and healthcare. Table 1 lists the nine publicly available univariate time series datasets used in this study, along with their domains and forecast horizons. Each dataset consists of sequences of scalar observations $\{x_t\}_{t=1}^{T}$, where T varies across datasets and represents the length of the time series. By focusing on univariate data, we aim to assess the models' ability to capture temporal patterns and forecast future values without relying on additional contextual information. This setup provides a clear view of each model's inherent capacity to learn and generalize temporal dependencies purely from the sequence itself.

Table 1: Publicly available univariate time series datasets, their domains, and forecast horizons.

Dataset	Domain	Prediction Horizon (H)
Air Passengers	Transportation	24
Sunspots	Astronomy	120
Temp	Weather	240
Temperature	Weather	24
Humidity	Weather	24
Relative Humidity	Weather	24
Birth	Healthcare	24
Store	Retail	24
Hospitality	Service	24

2.2 Dedicated time series foundation models

We selected TimesFM, Chronos, MOMENT, and TimeGPT because they represent open-source state-of-the-art models for zero shot and multi-horizon forecasting. These models were chosen for their strong generalization capabilities across diverse time series, their ability to leverage large-scale pretraining or tokenization strategies, and their suitability for evaluating cross domain forecasting performance. Their availability as open-source implementations ensures reproducibility and facilitates direct comparison under consistent experimental settings.

2.3 Modality-adapted foundation models:

GPT-OSS (text), Whisper (audio), and ViT (images) were originally trained on sequences or structured data from their respective modalities. To adapt them for numeric time series, we modify only the

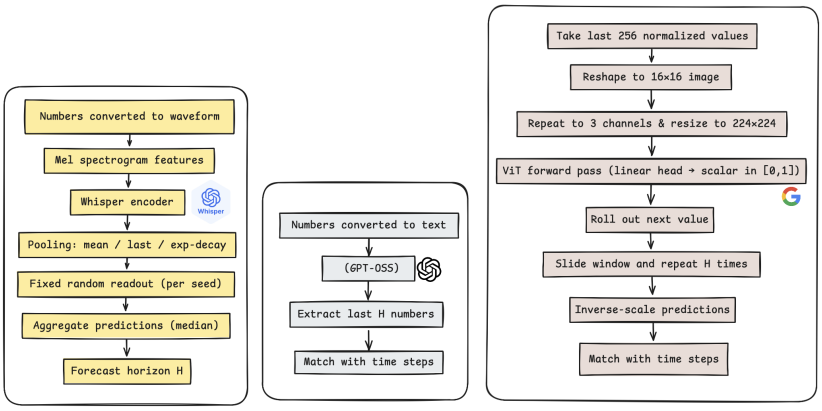

Figure 1: Adaptation of cross domain foundation models for time series forecasting. Numbers are mapped into alternative modalities: (left) audio sequences through Whisper, (middle) text sequences through GPT-OSS, and (right) images through Vision Transformers (ViT). Each pipeline illustrates the preprocessing, model encoding, and prediction steps used to generate forecasts.

input embedding layers to accept scalar sequences, while keeping all pretrained weights and attention mechanisms intact. Figure 1 illustrates this adaptation, showing how numeric values are mapped into alternative modalities and processed through each model's encoding pipeline to generate forecasts.

3 Result

Table 2 summarizes the zero shot forecasting performance of all models across the nine univariate datasets, measured using MAE, RMSE, and MAPE. We observe several noteworthy patterns.

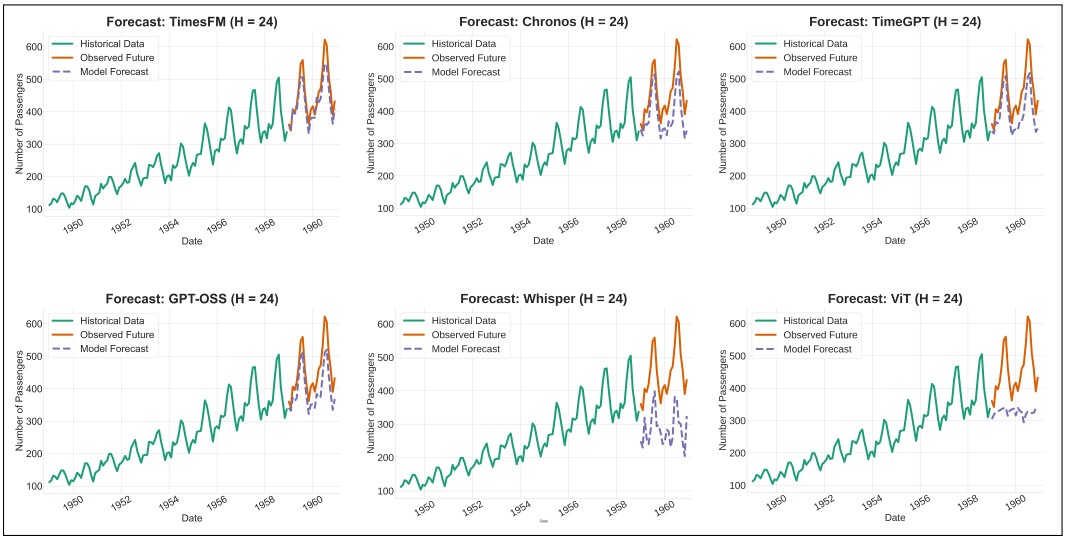

Figure 2: Comparison of different models forecasting monthly air passenger counts (H=24).

Among the modality-adapted models, GPT-OSS consistently demonstrates strong performance. Across most datasets, it outperforms MOMENT and performs competitively with TimeGPT, a dedicated time series foundation model. GPT-OSS outperforms specialized models like TimesFM and Chronos in several cases. On the Birth dataset, it achieves an RMSE of 6.46, compared to 6.59 for TimesFM and 7.03 for Chronos. On the Passengers dataset, it achieves lower errors than Chronos across all metrics: MAE 55.58 vs. 71.21, RMSE 60.91 vs. 77.25, and MAPE 12.00 vs. 15.49. For Sunspots, GPT-OSS also surpasses Chronos consistently with MAE 39.16 vs. 42.89, RMSE 47.60

Table 2: zero shot Forecasting Performance of Modality-Adapted and Dedicated Time Series Foundation Models Across Nine Univariate Datasets

Metric	Model	Air Passengers	Sunspots	Temp	Temperature	Humidity	Relative Humidity	Birth	Store	Hospitality	Avg. Rank
MAE	TimesFM	**28.98**	21.59	**1.89**	1.88	1.96	10.29	**5.45**	12.22	47.02	**2.00**
	Chronos	71.21	42.89	3.39	**1.48**	**1.61**	**5.42**	5.72	**9.06**	**16.35**	**2.00**
	TimeGPT	58.10	**6.24**	4.61	2.48	1.80	11.58	5.97	19.47	17.36	3.11
	MOMENT	206.34	83.48	3.51	5.98	4.16	16.48	5.83	13.62	550.21	5.11
	GPT-OSS	55.58	39.16	5.09	3.73	2.55	28.43	5.75	15.21	19.56	3.89
	ViT	59.38	255.05	9.98	8.84	4.02	16.55	15.22	47.93	183.02	5.78
	Whisper	165.74	423.16	12.12	4.55	4.06	16.91	8.21	25.08	265.14	6.11
RMSE	TimesFM	**34.08**	27.64	**2.54**	2.15	2.37	12.43	6.59	14.38	55.15	**2.11**
	Chronos	77.25	49.63	4.37	**1.82**	**1.99**	**7.61**	7.03	**11.61**	**20.27**	2.22
	TimeGPT	63.56	**8.36**	4.96	3.18	2.21	13.43	7.48	22.28	21.44	3.11
	MOMENT	219.43	85.18	4.16	6.92	4.79	18.38	6.85	16.36	553.32	4.78
	GPT-OSS	60.91	47.60	5.67	4.52	2.97	33.30	**6.46**	17.78	24.21	3.67
	ViT	77.90	257.88	10.46	10.97	4.86	20.52	16.80	49.21	185.74	6.22
	Whisper	173.81	477.02	15.07	5.79	4.54	19.54	11.32	29.43	275.20	5.89
MAPE	TimesFM	**6.08**	**454.31**	**23.36**	15.02	14.32	15.58	**12.13**	1.46	2.37	**1.89**
	Chronos	15.49	1229.25	32.78	**11.65**	**12.46**	**8.08**	12.76	**1.09**	**0.83**	2.00
	TimeGPT	12.57	753.09	35.92	22.46	12.54	16.57	12.86	2.32	0.89	3.11
	MOMENT	44.25	9224.93	47.31	53.80	32.97	26.85	13.15	1.63	27.91	5.44
	GPT-OSS	12.00	708.45	40.22	28.77	18.70	42.09	13.76	1.82	0.99	4.0
	ViT	12.39	7204.23	125.43	97.50	37.30	31.12	37.91	5.79	9.27	5.89
	Whisper	36.25	14441.09	126.02	24.50	26.08	23.14	17.30	3.01	13.48	5.67

vs. 49.63, and MAPE 708.45 vs. 1229.25. Figure 2 further illustrates the comparative forecasting behavior of all models on the Air Passengers dataset, showing that GPT-OSS provides forecasts closely aligned with the observed future and competitive with specialized time series models. These results demonstrate GPT-OSS's ability to generalize from a pretrained text sequence model to numeric time series forecasting. By contrast, Whisper and ViT, while able to generate forecasts, show higher error rates overall, particularly on datasets with longer horizons or high variability such as Sunspots and Temp. This suggests that the sequence modeling capabilities of audio and image foundation models can transfer to numeric time series to some extent, but not as effectively as a text-based pretrained model in this zero shot setting. As expected, models designed specifically for time series forecasting generally achieve strong results across most datasets. TimesFM and Chronos maintain the lowest average ranks across MAE, RMSE, and MAPE, demonstrating robust performance. TimeGPT shows competitive results and is often closely matched by GPT-OSS, highlighting impressive cross domain adaptability. MOMENT, while effective on certain datasets, exhibits higher errors on datasets with smaller horizons or sharp fluctuations, indicating sensitivity to dataset characteristics. Overall, GPT-OSS achieves an average rank of 3.89 in MAE and 3.67 in RMSE, placing it above dedicated time series models such as MOMENT and close to TimeGPT. Its stronger performance on datasets like Birth, Passengers, and Sunspots illustrates the potential of modality-adapted models for zero shot forecasting and suggests that pretrained sequence modeling can generalize effectively to numeric time series under certain conditions. Whisper and ViT demonstrate moderate performance, and their higher errors on some datasets indicate that modality alignment plays a critical role in transferability. These results suggest that while adaptation is feasible, the type of pretrained knowledge and sequence characteristics significantly influence forecasting accuracy.

4 Conclusion

This work presented the first systematic comparison of time series foundation models and modality adapted cross domain foundation models in a zero shot forecasting setting. Specialized models such as Chronos and TimesFM generally achieve strong performance, while cross domain models such as GPT OSS also demonstrate surprising competitiveness. These results highlight both the benefits of temporal pretraining and the transferability of general purpose sequence models to time series tasks. As future work, we will extend this evaluation to multivariate forecasting, broaden the scope to classification and anomaly detection, and incorporate additional datasets and baselines to establish a more rigorous and comprehensive benchmarking framework.

References

[1] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Bilovs, Hena Ghonia, N. Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and I. Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting. 2023.

[2] Xiaobin Hong, Jiawen Zhang, Wenzhong Li, Sanglu Lu, and Jia Li. Unify and anchor: a context-aware transformer for cross-domain time series forecasting. *arXiv.org*, 2025.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma-teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, I. Sutskever, and Dario Amodei. Language models are few-shot learners. *Neural Information Processing Systems*, 2020.

[4] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *International Conference on Machine Learning*, 2024.

[5] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *International Conference on Machine Learning*, 2024.

[6] Yuanzhao Zhang and William Gilpin. Zero-shot forecasting of chaotic systems. *International Conference on Learning Representations*, 2024.

[7] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: a family of open time-series foundation models. *International Conference on Machine Learning*, 2024.

[8] Nina Zukowska, Mononito Goswami, Michał Wiliński, Willa Potosnak, and Artur Dubrawski. Towards long-context time series foundation models. *arXiv.org*, 2024.

[9] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, April 2024. arXiv:2310.10688 [cs].

[10] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. 2024.

[11] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation, 2024. Version Number: 2.

[12] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Qingsong Wen, Christian S. Jensen, and Bin Yang. FoundTS: Comprehensive and Unified Benchmarking of Foundation Models for Time Series Forecasting. *FoundTS*, 2024. Publisher: arXiv.