How Foundational are Foundation Models for Time Series Forecasting?

Nouha Karaouli[1], Denis Coquenet[2], Elisa Fromont[1], Martial Mermillod[3], and Marina Reyboz[4]

[1]Univ. Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[2]Univ. Rennes, CNRS, IRISA - UMR 6074, F-35000 Rennes, France
[3]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, Grenoble, France
[4]Univ. Grenoble Alpes, CEA, LIST, 38000 Grenoble, France

Abstract

Foundation Models are designed to serve as versatile embedding machines, with strong zero shot capabilities and superior generalization performance when fine-tuned on diverse downstream tasks. While this is largely true for language and vision foundation models, we argue that the inherent diversity of time series data makes them less suited for building effective foundation models. We demonstrate this using forecasting as our downstream task. We show that the zero-shot capabilities of a time series foundation model are significantly influenced and tied to the specific domains it has been pretrained on. Furthermore, when applied to unseen real-world time series data, fine-tuned foundation models do not consistently yield substantially better results, relative to their increased parameter count and memory footprint, than smaller, dedicated models tailored to the specific forecasting task at hand.

1 Introduction

The emergence of Foundation Models (FMs), large-scale pretrained architectures such as BERT [1] in Natural Language Processing (NLP) and Vision Transformers [2] in Computer Vision (CV), has fundamentally transformed artificial intelligence. By leveraging massive and diverse datasets during pretraining, these models exhibit strong generalization abilities, enabling zero-shot and few-shot transfer to a wide range of downstream tasks [3, 4]. This shift has allowed FMs to consistently outperform traditional task-specific models trained from scratch for narrowly defined problems [4].

Inspired by these successes, researchers have recently proposed Time Series Foundation Models (TSFMs), large pretrained models designed to capture general-purpose representations across diverse temporal data [5]. These models aim to transfer knowledge across forecasting tasks by learning temporal patterns at scale, showing promising results in a variety of domains with minimal task-specific tuning.

However, the time series domain poses unique challenges that set it apart from NLP and CV. Time series data often exhibits domain-specific structures such as seasonality, trends, irregular sampling, and high variability across applications, even within the same broad category [6]. Such characteristics introduce distribution shifts that undermine the generalization abilities of TSFMs [7]. In particular, our experiments suggest that TSFMs' zero-shot performance is highly sensitive to the alignment between the statistical properties of the pretraining and target domains. When this alignment is weak, we observe substantial drops in generalization, even across domains that might appear related.

NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S).

While TSFMs often benefit from rapid initial convergence, extended fine-tuning can lead to performance degradation, whereas task-specific models trained from scratch typically yield steady accuracy gains under longer training and limited data regimes [8].

Motivated by these challenges, we conduct a thorough empirical evaluation of the univariate forecasting capabilities of TSFMs across diverse tasks. We compare them with traditional models trained from scratch to assess whether TSFMs offer practical advantages when fine-tuned on specific, potentially domain-shifted datasets.

Our main contributions are:

- Evaluating TSFMs in zero-shot mode across both domain-related and domain-shifted forecasting datasets.
- Comparing the fine-tuning capabilities of TSFMs versus traditional models on forecasting tasks to evaluate their adaptability and effectiveness under domain shift and limited data.
- Proposing a new forecasting dataset consisting of daily electricity usage over two years, on which a small dedicated network achieves better results than a fine-tuned TSFM.

2 Related Work

Recent TSFMs such as TiReX [9], TimeGPT [10], TimesFM [11], and FEDformer [12] leverage large-scale pretraining to enable strong generalization and transfer across forecasting tasks.

To assess their practical utility, several benchmarking frameworks have emerged. GIFT-eval [13] measures cross-domain generalization using standardized protocols, OpenTS [14] offers a reproducible suite spanning datasets, horizons, and metrics, while Nixtla's Arena [15] provides a comprehensive evaluation across frequencies and domains. [16] have also pointed out that naive baselines (here, a simple auto-regressive model) can achieve competitive performance compared to TSFM on several forecasting tasks.

These efforts report promising performance on public datasets such as Monash [17] and ETT [18]. However, we had to compare the generalization performance of these foundation models on time series ensured to be completely new and not included in these benchmark databases in order to test the challenges faced in deployment.

In contrast, we evaluate TSFMs on a proprietary electricity consumption dataset with realistic and complex domain shifts not seen during pretraining. Our setup introduces explicit distributional changes, enabling a more rigorous assessment of generalization.

Contrary to standard benchmarks that primarily focus on evaluating zero-shot capabilities of TSFMs on public datasets, we further compare these models to conventional ones trained from scratch. This allows us to highlight scenarios where smaller, specialized models achieve comparable performance to large pretrained TSFMs, especially under conditions of data scarcity and nonstationarity.

Through this, we uncover limitations in TSFMs' robustness and provide new insights into their practical effectiveness in real-world forecasting scenarios.

3 Methodology

Our evaluation addresses two central questions: (1) Can TSFMs generalize beyond their pretraining distributions? (2) Are they practically competitive with lightweight, specialized alternatives?

We benchmark three leading TSFMs, namely TimesFM [11], TimeGPT [10], and TiReX [9], alongside SAMFormer [19], a compact attention-based model operating over the channel dimension. Unlike the other models, SAMFormer is trained from scratch in our experiments.

Synthetic benchmarks. We construct four datasets that reflect recurring structures in TSFM pretraining, while ensuring zero data overlap.

- **D1** and **D2** are composed of harmonically aligned sine waves with full observability, probing the models' ability to recognize and extrapolate clean periodic signals.

- **D3** and **D4** consist of randomly sampled, non-harmonic sine waves, forming complex, partially observable cycles. These challenge the models to generalize from incomplete patterns.

All synthetic sequences contain 2,688 time steps (8 weeks sampled at 30-minute intervals).

Real-world evaluation. We test TSFMs on *Elec_Consumption*, a proprietary small dataset capturing daily electricity usage of a single home over two years (2023–2024). Unlike the generic, population-level datasets typically used during TSFM pretraining, this dataset reflects individual consumption behavior, introducing a clear distribution shift. This setting allows us to rigorously evaluate whether pretrained models retain strong forecasting performance when faced with personalized, unseen patterns, a crucial requirement for real-world deployment in user-specific applications.

All datasets used in this work are made publicly available for reproducibility[1].

Fine-tuning experiments. We fine-tune TimesFM on Elec_Consumption and compare it to SAM-Former trained from scratch. This setup quantifies the trade-off between the computational overhead of fine-tuning large pretrained models and the efficiency of smaller models tailored to specific domains.

Together, these evaluations dissect the *one-size-fits-all* [20, 21] promise of TSFMs, distinguishing their theoretical representational capacity (via synthetic benchmarks) from their practical effectiveness in real-world deployment. We report Mean Absolute Error (MAE) as the primary metric.

4 Results

We begin our experimental evaluation by testing all models in zero-shot mode on both synthetic and real-world datasets, Appendix A provides additional details on our models querying setup. Tables 1 and 2 report results on synthetic data using a fixed context length of 512 across three forecast horizons. Table 3 presents results on the Elec_Consumption dataset.

Table 1: Zero-shot MAE on D1 and D2 for various forecasting horizons and models. Lower is better.

Datasets		D1			D2		
Models		TimeGPT	TiReX	TimesFM	TimeGPT	TiReX	TimesFM
Horizons	**128**	0.89	**0.11**	0.13	0.80	0.29	**0.15**
	256	1.08	**0.21**	0.22	1.25	0.72	**0.35**
	512	1.09	0.37	**0.34**	1.57	1.11	**0.72**

Table 2: Zero-shot MAE on D3 and D4 for various forecasting horizons and models. Lower is better.

Datasets		D3			D4		
Models		TimeGPT	TiReX	TimesFM	TimeGPT	TiReX	TimesFM
Horizons	**128**	1.86	**1.1**	1.13	1.3	**0.78**	0.89
	256	1.43	**0.95**	0.98	1.63	**1.6**	1.62
	512	**2.29**	3.3	3.5	**2.31**	2.8	2.98

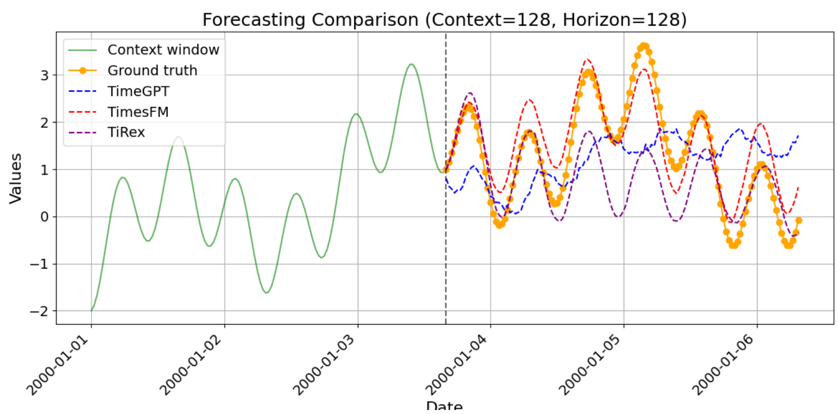
Among all five experiments, TiReX and TimesFM consistently perform best, particularly on D1 and D2, which exhibit simple and periodic sinusoidal patterns, highlighting their ability to capture repetitive temporal structures. In contrast, forecasting on D3 and D4, involving irregular and composite sinusoidal signals, is more challenging. Despite this, foundation models still generalize reasonably well, likely due to pretraining on structurally similar synthetic patterns. However, on the real-world Elec_consumption dataset, even with careful tuning of context and horizon lengths, the models struggle to accurately forecast the future values. This shows the limits of the generalization abilities of current state-of-the-art TSFMs for a real-case forecasting scenario.

[1] `https://doi.org/10.5281/zenodo.17199246`

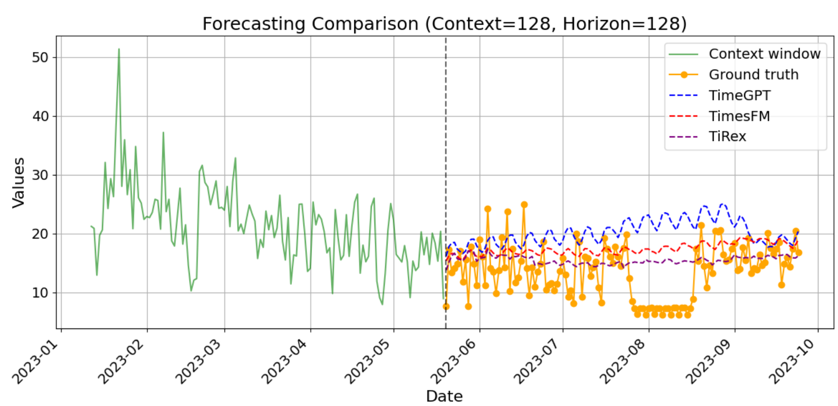
Table 3: Zero-shot MAE on *Elec_Consumption* for varying context-horizon pairs and models. Lower is better.

Models	Context – Horizon				
	15–7	30–7	60–30	128–128	365–365
TimeGPT	6.60	6.52	5.60	6.91	6.44
TiReX	6.94	**5.71**	4.61	**3.78**	5.90
TimesFM	**5.07**	5.83	**4.08**	4.63	**5.30**

This performance contrast is clearly illustrated in Figure 1. On dataset D1, the TSFMs demonstrate strong generalization, achieving MAE scores of 0.95, 0.46, and 0.71 on TimesGPT, TimesFM, and TiReX, respectively. In contrast, on Elec_consumption, the forecasts deviate more noticeably from the expected values.



(a) Forecasting results on D1.



(b) Forecasting results on Elec_Consumption.

Figure 1: Forecasting results using different models. Panel (a) shows results on D1, while panel (b) presents results on Elec_Consumption.

While zero-shot results show that TSFMs perform well on target distributions that resemble their pretraining data, their ability to adapt to small, domain-specific datasets produces high errors and low prediction ability. To investigate this, we compare fine-tuned TimesFM with SAMFormer trained from scratch on our Elec_consumption dataset. This evaluation tests whether TSFMs' learned representations and inductive biases confer advantages for personalization.

Fine-tuning and training from scratch were performed using Adam with a learning rate 10^{-4}, weight decay 0.01, and batch size 64. The choice of LR follows the default fine-tuning configuration used in the public TimesFM examples, ensuring consistency with recommended practice for this foundation model. Data were standardized and framed with a sliding window (context $= 128$, horizon $= 128$). TimesFM was fine-tuned from a fixed pre-trained checkpoint, excluding any significant source of randomness. In contrast, SAMFormer was trained from scratch, introducing natural variability in the results due to the random weight initialization. To make the evaluation robust, we computed the mean and standard deviation over 5 runs with different random seeds. Models were trained for up to 100 epochs with early stopping (patience $= 10$). Experiments were conducted on an NVIDIA Tesla V100 GPU. Results are shown in Table 4.

Table 4: MAE for TimesFM and SAMFormer with a context window of 128 and a forecast horizon of 128.

Models	MAE
TimesFM	4.49 ± 0.00
SAMFormer	**4.28** ± 0.05

As one can note, the results show that SAMFormer, trained entirely from scratch with fewer than 50K parameters, ultimately outperforms TimesFM on the target forecasting task. While TimesFM benefits from large-scale pretraining and contains over 500 million parameters, SAMFormer achieves superior accuracy while remaining extremely lightweight and efficient to train on consumer-grade GPUs. This contrast highlights a key point: massive pretrained models do not always guarantee superior downstream performance, particularly in settings where data distributions differ from the pretraining corpus or where the target domain exhibits specific structural regularities that a smaller model can exploit more effectively. Moreover, SAMFormer's compact size reduces both training time and inference cost, making it well-suited for rapid experimentation and deployment in resource-constrained environments. These findings illustrate that carefully designed, domain-adapted models can deliver competitive or even superior performance compared to large foundation models, while offering substantial advantages in efficiency, accessibility, and environmental sustainability.

5 Conclusion

While TSFMs show strong zero-shot performance on synthetic and structurally similar data, their generalization ability is tightly coupled with the distribution seen during pretraining. In real-world settings involving domain shifts and limited data, a lightweight model like SAMFormer, with only 49.5K parameters and no large-scale pretraining, can still achieve better results when trained from scratch. This suggests that the "one-size-fits-all" promise of TSFMs may not hold in practice, especially under resource constraints or personalization requirements. Our findings advocate for a more nuanced deployment strategy: leveraging TSFMs when pretraining-task similarity is high, and favoring lightweight, specialized models when personalization, efficiency, or domain mismatch is critical.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, and Alexander Kolesnikov et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared Kaplan et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pages 8748–8763, San Francisco, CA, USA, 2021. PMLR. Equal contribution. OpenAI.

[5] Y. Liang et al. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6555–6565, Barcelona, Spain, August 2024. ACM.

[6] Jiexia Ye, Weiqi Zhang, Ke Yi, Yongzi Yu, Ziyue Li, Jia Li, and Fugee Tsung. A survey of time series foundation models: Generalizing time series representation with large language model. *arXiv preprint arXiv:2405.02358*, 2024.

[7] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Qingsong Wen, et al. Foundts: Comprehensive and unified benchmarking of foundation models for time series forecasting. 2024.

[8] Lifan Zhao, Yanyan Shen, Zhaoyang Liu, Xue Wang, and Jiaji Deng. Less is more: Unlocking specialization of time series foundation models via structured pruning, 2025.

[9] Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *arXiv preprint arXiv:2505.23719*, 2025.

[10] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2024. Version 3, 27 May 2024.

[11] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 4599–4623. PMLR, 2024.

[12] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, pages 27268–27286, 2022.

[13] Taha Aksu et al. GIFT-Eval: A Benchmark for General Time Series Forecasting Model Evaluation. arXiv, November 2024.

[14] DecisionIntelligence. OpenTS – A Comprehensive and Fair Benchmark for Time Series Analytics. `https://decisionintelligence.github.io/OpenTS/algorithms/methods/index.html`, October 2024. Accessed 25 October 2024.

[15] Nixtla. Foundation time series arena – benchmarking foundation models for time series. `https://github.com/nixtla/experiments/tree/main/foundation-time-series-arena`, 2024. GitHub repository.

[16] Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized foundation models struggle to beat supervised baselines. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

[18] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115. AAAI Press, 2021.

[19] Romain Ilbert, Ambroise Odonnat, Vasilii Feofanov, Aladin Virmaux, Giuseppe Paolo, Themis Palpanas, and Ievgen Redko. Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. *arXiv preprint arXiv:2402.10198*, 2024.

[20] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[21] Peiwen Yuan, Yueqi Zhang, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Beyond one-size-fits-all: Tailored benchmarks for efficient evaluation. *arXiv preprint arXiv:2502.13576*, 2025.

A Querying Foundation Models

We evaluated three pretrained time-series foundation models in a zero-shot setting: TimesFM [11], TiReX [9], and TimeGPT [10]. All models were used without architectural modifications; only the context length, forecast horizon, and sampling frequency were adapted to match our datasets. This ensures comparability while respecting each model's inference interface.

TimesFM. TimesFM is a decoder-only transformer trained on large-scale synthetic and real-world collections of time series. The model represents inputs as patches of fixed length embedded into a 1280-dimensional latent space, with autoregressive decoding over output patches. We employed the released checkpoint `google/timesfm-2.0-500m-pytorch` and queried the model with context windows of varying length (L) and horizons matched to the experimental setup (H). Forecasts were generated via the `forecast_on_df` API, with the frequency parameter aligned to the dataset resolution (30 minutes for synthetic benchmarks; daily for *Elec_Consumption*).

TiReX. TiRex is a TSFM built upon the xLSTM architecture. The pretrained checkpoint `NX-AI/TiReX` was queried directly, providing context windows of length L and requesting horizons of length H. The model outputs both quantile forecasts and mean trajectories; we report the mean predictions in all experiments. Dataset frequency was set to match the native resolution, consistent with the TimesFM setup.

TimeGPT. TimeGPT is a hosted transformer-based FM trained on over 100B time-series observations across diverse domains. We accessed the `timegpt-1-long-horizon` variant through the Nixtla Python client, using an API key for authentication. The model was queried with training windows of length L and a forecast horizon H identical to the experimental split. Frequency alignment was handled through the timestamp column in the input dataframe.

Across all models, context length (L), horizon length (H), and sampling frequency were varied according to the dataset and experimental condition. No fine-tuning or weight adaptation was applied; results therefore reflect pure zero-shot performance under consistent querying protocols.