In-context Pre-trained Time-Series Foundation Models adapt to Unseen Tasks

Anonymous Authors[1]

Abstract

Time-series foundation models (TSFMs) have demonstrated strong generalization capabilities across diverse datasets and tasks. However, existing foundation models are typically pre-trained to enhance performance on specific tasks and often struggle to generalize to unseen tasks without fine-tuning. To address this limitation, we propose augmenting TSFMs with In-Context Learning (ICL) capabilities. Unlike conventional models that learn a fixed input-output mapping, an ICL-equipped model can perform test-time inference by dynamically adapting to input-output relationships provided within the context. Our framework, In-Context Time-series Pre-training (ICTP), restructures the original pre-training data to imbue the model with ICL capabilities, enabling adaptation to unseen tasks during pre-training. Experiments demonstrate that ICT improves the performance of state-of-the-art TSFMs by approximately 11.4% on unseen tasks without requiring fine-tuning.

1. Introduction

Time-series analysis serves as a critical tool in a wide range of real-world applications, including weather forecasting (Pisias & Moore Jr, 1981; Wu et al., 2021), pandemic analysis (Rodríguez et al., 2021), and imputation of missing variables or historical records (Stefanakos & Athanassoulis, 2001; Ma et al., 2019; Wu et al., 2022). Given this diversity, developping robust and accurate time-series models is essential, which is a challenge recently addressed by the emergence of time-series foundation models (TSFMs) (Das et al., 2023; Woo et al., 2024; Goswami et al., 2024; Kamarthi & Prakash, 2024). These models are pretrained on vast datasets, enabling them to perform effectively on a certain range of tasks, including forecasting (Das et al., 2023; Woo et al., 2024; Gruver et al., 2024), imputation (Goswami et al., 2024), and classification (Goswami et al., 2024; Gruver et al., 2024).

However, a key limitation of existing TSFMs is their lack of multi-task adaptation capability without fine-tuning. Current models either specialize in a single task, such as forecasting (Das et al., 2023; Woo et al., 2024), or require task-specific fine-tuning before deployment(Goswami et al., 2024; Kamarthi & Prakash, 2024; Gao et al., 2024). This limitation increases computational overhead and data requirements, hindering their real-world applicability.

To address this gap, we propose enhancing foundation time-series models with in-context learning (ICL) (Brown et al., 2020), enabling multi-task adaptation without fine-tuning. ICL operates through a test-time inference procedure: by appending input-output pairs (the context) to the original input, the model infers the task's requirements and produces the desired output.

However, integrating ICL into TSFMs presents unique challenges. In language models, ICL emerges naturally from pretraining on diverse tasks embedded in textual data (Brown et al., 2020; Gu et al., 2023). In contrast, time-series data lacks inherent task diversity, as datasets are uniformly structured in chronological order. Therefore, training on a single objective (e.g., forecasting) inherently restricts exposure to other tasks, making ICL acquisition impossible without explicit multi-task pretraining. To the best of our knowledge, this challenge is not tackled by the pre-training pipelines of existing TSFMs.

To overcome this, we introduce In-Context Time-series Pretraining (ICTP), a novel pipeline that transforms existing datasets into a multi-task format for ICL-enabled pretraining. ICTP first identifies task candidates and generates input-output examples for each task in a unified format, covering major sequence-to-sequence applications. Next, it constructs context sequences by combining examples from different tasks. Finally, it trains the model on these augmented sequences, explicitly teaching in-context reasoning. This approach generalizes to most time-series tasks, offering broad applicability.

To validate our approach, we train foundation models on datasets processed with ICTP and evaluate their perfor-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

mance on both seen and unseen tasks. Experimental results demonstrate that models pretrained with ICTP on a subset of tasks achieve significant improvements on unseen tasks. Moreover, when pretrained with ICTP on all tasks, the models exhibit further performance gains. We also conduct extensive ablation studies to analyze the mechanisms through which ICTP enhances model capabilities, providing deeper insights into its effectiveness.

2. Methodology

2.1. Problem Definition

Consider a time-series dataset $\mathcal{D} = \{x_0, \dots x_n\}$, $x_i \in \mathbb{R}^{T\times m}$. We consider a multi-task adaptation scenario: for each task k from a task set $k \in K$, there's a relationship $f_k(x) = y$, $x \in D$, $y \in \mathbb{R}^{T\times h}$ between input x and the desired output y, where T and h are the input and output horizons respectively, and m is the number of channels in each vector. Our goal is to achieve *non-fine-tune adaptation* on *unseen tasks*. That is, Given a task candidate set $K = \{k_1, \dots, k_N\}$ and a input sequence x, a time-series model f_θ parameterized by θ with *non-fine-tune adaptation* capability should output $f_k(x)$ without fine-tuning θ.

2.2. Enabling Multi-task Non-fine-tune Adaptation for TSFMs

The primary challenge in achieving non-fine-tune multi-task adaptation for foundation models lies in obtaining the input-output relationship f_k for various tasks without fine-tuning the model parameters. Inspired by recent progress of test-time inference in natural language processing (Brown et al., 2020; Gu et al., 2023) and computer vision (Zhou et al., 2024), we address this challenge by equipping foundation models with in-context learning (ICL) capabilities.

ICL is an emergent capability first observed in large language models (LLMs) after pre-training on extensive textual corpora (Brown et al., 2020). Unlike traditional models that specialize in a single task during training, an ICL model learns to adapt dynamically by leveraging contextual information during inference. Specifically, when an input x is concatenated with a context sequence c_k that contains information of a task k, the ICL model f_{ICL} imitates c_k to produce task-adapted output $f_k(x)$. Formally, this can be expressed as: $f(x; c_k) = f_k(x)$, where c_k is mostly constructed from input-output pairs of task k, i.e., $c_k = \bigoplus\{x_1, f_k(x_1), x_2, f_k(x_2), \dots\}$, where $\bigoplus$ means a concatenation template.

However, extending ICL to time-series foundation models presents unique challenges not addressed by existing pipelines. Current time-series models are typically pre-trained using a single-task objective (e.g., forecasting) on large-scale datasets (Das et al., 2023; Woo et al., 2024; Goswami et al., 2024; Kamarthi & Prakash, 2024). While LLMs serendipitously developed ICL through similar uni-task pre-training (Brown et al., 2020), subsequent research (Min et al., 2022; Chen et al., 2022; Gu et al., 2023) revealed that multitask pre-training—enabled by the inherent diversity of linguistic data—is crucial for acquiring robust ICL capabilities. For instance, next-token prediction on the sentence ”The temperature dropped below zero today for the first time, so everyone is nervous” implicitly requires the model to perform sentiment analysis. In contrast, time-series data follows a strict chronological order, meaning a model trained for next-step prediction will inherently specialize in short-term forecasting but fail at tasks like imputation or anomaly detection.

Algorithm 1 In-Context Time-series Pre-training (ICTP)

Input: Time-series dataset $\mathcal{D} = \{x\}$, Task candidates K, Demonstration size m, Foundation time-series model f
Output: Reorganized dataset $\mathcal{D}_{ICL}$, model with ICL capability f_{ICL}
for $x \in \mathcal{D}$ **do**
  Sample $k \sim K$
  $y^k = f_k(x)$
  $C_x = \phi$
  for i **to** m **do**
    Sample $x_i \sim \mathcal{D},\ x_i \cap x = \phi$
    $y_i^k = f_k(x_i)$
    $C_x = C_x \oplus x_i \oplus y_i^k$
  end for
  $\mathcal{D}_{ICL} \leftarrow (C_x \oplus x, y^k)$
end for
$f_{ICL} \leftarrow$ Finetune f by $\mathcal{D}_{ICL}$

To overcome this limitation, we argue that time-series foundation models must be explicitly pre-trained on multiple tasks to acquire ICL capabilities. We achieve this through In-Context Time-series Pre-training (ICTP), a novel pipeline designed to foster multitask adaptability.

2.3. In-Context Time-series Pre-training (ICTP)

We propose In-Context Time-series Pre-training (ICTP), a novel pipeline that transforms a raw time-series dataset into a multi-task context-following dataset. Given a dataset $\mathcal{D}$ and task candidates $K = \{k_1, k_2, \dots\}$, ICTP constructs input-output pairs (x_i, y_i^k) for each data point $x_i \in D$ and each task $k \in K$, where $y_i = f_k(x_i)$. Next, ICTP assembles context pieces by concatenating pairs from the same task: $E_i^k = \{x_{i_1}, y_{i_1}^k, x_{i_2}, y_{i_2}^k, \dots\}_{seq}$. These context pieces are then augmented with the original inputs to form modified inputs $x_i' = E_i \oplus x_i$, while the corresponding outputs $y_i = f_k(x_i)$ remain unchanged. A complete pipeline is described in Fig 1. We argue that pre-training TSFMs on datasets structured by ICTP inherently equips them with

ICL capabilities, enabling dynamic task adaptation without parameter fine-tuning.

3. Experiments

3.1. Settings

To demonstrate the effect of ICTP on TSFMs' performance of unseen tasks, we collect several time-series task candidates, applying ICTP to pre-train backbone TSFMs while iteratively excluding one of the task candidates, and evaluate the pre-trained TSFMs on the excluded task without fine-tuning to see if their performance on such an unseen task has been improved.

3.1.1. Backbone Models

We choose three representative foundational time-series models—MOMENT (Goswami et al., 2024), TimesFM (Das et al., 2023), LPTM (Kamarthi & Prakash, 2024)— as the backbones. Details of these backbones can be accessed in Appendix. Such a selection covers three pre-training objectives commonly adapted in time-series models: mask construction, autoregressive generation, and adaptive segmentation. While incorporating ICTP, we keep the original pre-training objective for each model, only reforming the pre-training dataset using ICTP. For all models, we use wrappers from Samay [1] to manage our experiments and datasets.

3.1.2. Tasks Candidates

we collect three time-series tasks as candidates for pre-training and evaluation: 1) Forecasting, where the model predict the consecutive future values of a given sequence; 2) Imputation, where the model rebuilds certain part of the input which was masked 3) Backtracing, where the model predict the consecutive history values of a given sequence. While evaluating ICTP on each task, we pretrain backbone models on the other tasks, keeping the evaluation task unseen. For example, while evaluating the models on backtracing, the task candidates in ICTP will be forecasting and imputation.

Specifically, as TimesFM and LPTM have accessed forecasting data during their original pre-training procedure and MOMENT has accessed imputation data, we exclude these tasks in corresponding evaluation. Implementation details can be found in Appendix.

3.1.3. Baselines

As the scope of ICTP is to adapt single-task foundation models to multiple tasks without fine-tuning, we set baseline methods as task-aware naive input reprogramming that does not require fine-tuning.

For TimesFM and LPTM, we 1) truncates imputation inputs before (after) the imputation target, depending on whether the reconstruction area surpasses the middle of the original sequence, while maintaining chronological order 2) flip the backtracing inputs and outputs.

For MOMENT, we 1) concatenate forecasting inputs and outputs as input, with mask on original outputs 3) flipped the backtracing inputs and outputs, then apply the same strategy as forecasting.

3.1.4. Dataset

We choose four datasets (ETTh1, ETTm1, Exchange Rate, Weather) from the Informer datasets (Zhou et al., 2021), and two datasets (PEMS-Bays, and METR-LA) from DCRNN datasets (Li et al., 2017) to pre-train and evaluate ICTP. Details of each dataset can be found in Appendix.

For all the datasets, we adopt Channel Independence assumption (Nie et al., 2022), conducting tasks on each channel only considering inputs from the corresponding channel. We split the train / valid / test data chronologically by 60:20:20 (that is, train data is always earlier than valid / test). We referred to implementations of TimesNet (Wu et al., 2022) to normalize the input data.

For all tasks, we consider two output lengths, 96 and 192, for all datasets. For forecasting and backtracing, the lookback window is set as 192 and 384, respectively. For Imputation, we set the input length as 192 and 384 correspondingly and randomly mask 96 (192) of the inputs as reconstruction targets so that the input/output length all aligns between different tasks. We use 4 context examples in all tasks. While building context sequences, we make sure there's no overlap between examples and target output.

3.2. Results

The results are presented in Table 1 and 2 for output lengths of 96 and 192, respectively. After adapting ICTP, the backbone TSFMs exhibited significant performance improvements on unseen tasks across most datasets, with average improvement ratios of 11.3% and 11.6%, respectively. This demonstrates that ICTP effectively enhances the capability of TSFMs to handle unseen tasks. Notably, this improvement is achieved without any prior knowledge of the downstream task, highlighting ICTP's potential for broader applications.

It's worthy noticing that the degree of improvement varies across models and datasets. Specifically, ICTP struggles to enhance imputation performance for decoder-only models (TimesFM, LPTM) but shows greater success in improving forecasting and backcasting performance for encoder-only

[1] https://github.com/AdityaLab/Samay

Backbone	Evaluated on	ICTP	ETTh1		ETTm1		Exchange		Weather		PEMS-Bay		METR-LA	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MOMENT	Forecasting	No	0.813	0.629	0.724	0.588	0.228	0.285	0.215	0.345	2.623	0.899	1.291	0.765
		Yes	**0.433**	**0.458**	**0.496**	**0.541**	0.236	0.279	**0.163**	**0.206**	**1.625**	**0.595**	**1.169**	**0.728**
	BackTracing	No	0.834	0.643	0.75	0.595	0.229	0.289	0.222	0.354	2.594	0.915	1.284	0.767
		Yes	**0.439**	**0.454**	**0.502**	**0.527**	0.241	0.305	**0.165**	**0.254**	**1.773**	**0.613**	1.315	0.780
TimesFM	BackTracing	No	0.518	0.464	0.402	0.427	0.118	0.238	0.182	0.207	2.993	0.879	1.477	0.742
		Yes	**0.438**	**0.429**	**0.382**	**0.447**	**0.097**	**0.218**	**0.175**	**0.198**	**2.167**	**0.719**	**1.283**	**0.682**
	Imputation	No	0.920	0.604	0.967	0.649	0.118	0.244	0.235	0.281	2.888	0.835	1.198	0.613
		Yes	**0.785**	**0.592**	**0.934**	**0.692**	0.134	0.265	**0.231**	**0.279**	**2.818**	**0.761**	1.412	0.723
LPTM	BackTracing	No	0.830	0.663	0.739	0.640	2.259	1.229	0.471	0.497	2.832	0.950	1.528	0.848
		Yes	**0.672**	**0.597**	**0.628**	**0.564**	**2.173**	**1.134**	**0.381**	**0.435**	**2.033**	**0.688**	**1.375**	**0.733**
	Imputation	No	1.164	0.784	1.128	0.776	1.923	1.119	0.383	0.451	2.226	0.806	1.175	0.729
		Yes	**1.141**	**0.779**	**1.096**	**0.764**	**1.873**	**1.035**	**0.331**	**0.411**	**2.122**	**0.804**	**1.097**	**0.699**

Table 1. Results of backbone TSFMs with and without ICTP of output length 96 on all tasks, all datasets. Results outlined by bold shows that ICTP improved the performance of backbone TSFMs on unseen tasks.

Backbone	Evaluated on	ICTP	ETTh1		ETTm1		Exchange		Weather		PEMS-Bay		METR-LA	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MOMENT	Forecasting	No	0.938	0.691	0.849	0.644	0.337	0.365	0.506	0.547	2.247	0.821	1.336	0.780
		Yes	**0.514**	**0.527**	**0.618**	**0.544**	**0.332**	**0.345**	**0.357**	**0.427**	**2.198**	**0.698**	**1.310**	**0.760**
	BackTracing	No	0.944	0.704	0.886	0.654	0.337	0.358	0.556	0.581	2.296	0.861	1.319	0.797
		Yes	**0.526**	**0.512**	**0.603**	**0.531**	**0.330**	**0.347**	**0.402**	**0.495**	**2.593**	**0.847**	**1.211**	**0.728**
TimesFM	BackTracing	No	0.584	0.509	0.496	0.467	0.278	0.361	0.238	0.277	3.315	0.966	1.755	0.836
		Yes	**0.512**	**0.492**	**0.415**	**0.453**	**0.181**	**0.305**	**0.231**	**0.269**	**2.249**	**0.705**	**1.379**	**0.659**
	Imputation	No	1.053	0.642	0.919	0.610	0.226	0.334	0.385	0.382	3.485	0.986	1.623	0.752
		Yes	**0.913**	**0.679**	**0.852**	**0.572**	**0.242**	0.338	**0.374**	**0.379**	**2.570**	**0.822**	**1.516**	**0.767**
LPTM	BackTracing	No	0.811	0.859	0.765	0.658	2.456	1.314	0.435	0.468	2.841	0.965	1.481	0.822
		Yes	**0.796**	**0.661**	**0.627**	**0.568**	2.645	1.359	**0.431**	**0.473**	**1.859**	**0.605**	**1.375**	**0.807**
	Imputation	No	1.244	0.818	1.403	0.861	2.016	1.156	0.515	0.497	2.618	0.909	1.330	0.774
		Yes	**1.164**	**0.785**	**1.168**	**0.784**	2.199	**1.146**	**0.343**	**0.413**	**2.508**	**0.889**	**1.294**	**0.787**

Table 2. Results of backbone TSFMs with and without ICTP of output length 192 on all tasks, all datasets. Results outlined by bold shows that ICTP improved the performance of backbone TSFMs on unseen tasks.

models (MOMENT). Additionally, the improvement on the Exchange dataset is the smallest among all datasets. We attribute this to the dataset's simplicity: unlike the others, Exchange consists of weekly foreign currency exchange rates, which exhibit relatively smooth patterns. Consequently, TSFMs can more easily adapt to the input-output variance gap across tasks in this case.

Furthermore, we note that the Weather dataset was already included in TimesFM's original pre-training process. Despite this, ICTP still substantially improved TimesFM's performance on unseen tasks. This supports our hypothesis that existing pre-training pipelines for TSFMs, while effective for specific tasks, do not inherently equip models with multi-task capability without fine-tuning, which is a gap that ICTP successfully addresses.

4. Conclusion and Discussion

In this paper, we present In-Context Time-series Fine-tuning (ICTP), a novel method for enhancing the non-fine-tuning adaptability of time-series foundation models (TSFMs) on unseen tasks. By restructuring pre-training datasets to incorporate multi-task coverage and explicit context paradigms, ICTP equips TSFMs with in-context learning capabilities akin to those of large language models (LLMs). Our experiments demonstrate that ICTP significantly improves performance on unseen tasks while maintaining robust performance on previously encountered tasks. Future work could explore extending ICTP to a broader range of tasks or more diverse real-world datasets.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Bhope, R. A., Venkateswaran, P., Jayaram, K., Isahagian, V., Muthusamy, V., and Venkatasubramanian, N. Optiseq: Optimizing example ordering for in-context learning. *arXiv preprint arXiv:2501.15030*, 2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, Y., Zhong, R., Zha, S., Karypis, G., and He, H. Meta-learning via language model in-context tuning. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 719–730, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.53. URL https://aclanthology.org/2022.acl-long.53/.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Gao, S., Koker, T., Queen, O., Hartvigsen, T., Tsiligkaridis, T., and Zitnik, M. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

Gu, Y., Dong, L., Wei, F., and Huang, M. Pre-training to learn in context. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Kamarthi, H. and Prakash, B. A. Large pre-trained time series models for cross-domain time series analysis tasks, 2024. URL https://openreview.net/forum?id=KJ1w6MzVZw.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Ma, Q., Li, S., Shen, L., Wang, J., Wei, J., Yu, Z., and Cottrell, G. W. End-to-end incomplete time-series modeling from linear memory of latent variables. *IEEE transactions on cybernetics*, 50(12):4908–4920, 2019.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. MetaICL: Learning to learn in context. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.201. URL https://aclanthology.org/2022.naacl-main.201/.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Pisias, N. G. and Moore Jr, T. The evolution of pleistocene climate: a time series approach. *Earth and Planetary Science Letters*, 52(2):450–458, 1981.

Rodríguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., and Prakash, B. A. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15393–15400, 2021.

Stefanakos, C. N. and Athanassoulis, G. A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data. *Applied ocean research*, 23(4):207–220, 2001.

Wang, S., Yang, C.-H. H., Wu, J., and Zhang, C. Bayesian example selection improves in-context learning for speech, text, and visual modalities. *arXiv preprint arXiv:2404.14716*, 2024.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Xu, S. and Zhang, C. Misconfidence-based demonstration selection for llm in-context learning. *arXiv preprint arXiv:2401.06301*, 2024.

Zhao, S., Nguyen, T., and Grover, A. Probing the decision boundaries of in-context learning in large language models download pdf. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Zhou, T., Niu, P., Sun, L., Jin, R., et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

Zhou, Y., Li, X., Wang, Q., and Shen, J. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024.

A. Related Work

A.1. Time Series Foundation Models

Early approaches to time-series foundation models relied on repurposing (Gruver et al., 2024) or reprogramming (Jin et al., 2023; Zhou et al., 2023) existing large-language-models. Subsequent work shifted toward task-specific pre-training on large-scale time-series data, employing objectives such as patch-level autoregressive forecasting (Das et al., 2023), mask-then-reconstruction (Woo et al., 2024), or maximizing likelihood of tokens (Ansari et al., 2024). However, these models are designed for a single task and exhibit limited generalization to other tasks. Meanwhile, A few recent efforts have explored multi-task TSFMs. For instance, MOMENT (Goswami et al., 2024), pre-trains an encoder via masked reconstruction, while LPTM (Kamarthi & Prakash, 2024) employs adaptive segmentation. Nevertheless, both models require task-specific fine-tuning of projection heads for adaptation. To our knowledge, no existing time-series foundation model achieves task-agnostic adaptation without fine-tuning.

A.2. In-Context Learning

In-context learning (ICL), first observed in LLMs (Brown et al., 2020), enables task adaptation through dynamic prompting rather than parameter updates. While early models acquired ICL capabilities unintentionally (Brown et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023), recent studies proposed to actively enhance ICL through improved pre-training strategies (Zhao et al.; Gu et al., 2023) or optimized example selection (Wang et al., 2024; Xu & Zhang, 2024; Bhope et al., 2025). Though these advances significantly boost ICL performance in language models, their applicability to time-series domains remains unexplored.

B. Implementation Details

B.1. Introduction of Backbone Models

We introduce the backbone TSFMs introduced in our experiments.

- MOMENT[2] (Woo et al., 2024) is a family of large, pre-trained time-series foundation models based on a transformer architecture designed for diverse time-series tasks, including forecasting, classification, anomaly detection, and imputation. The model employs a masked time-series modeling approach during pre-training, where patches of time series are masked and reconstructed to learn robust representations. MOMENT is equipped with a lightweight reconstruction head, reversible normalization, and relative positional embeddings, making it highly adaptable to multivariate and univariate time-series data with varying temporal characteristics. We use their released checkpoint **AutonLab/MOMENT-1-large**.

- TimesFM[3] (Das et al., 2023) is a decoder-only time-series foundation model designed for zero-shot forecasting, leveraging a patching strategy to divide time-series into non-overlapping segments for efficient training. The model employs stacked transformer layers with causal self-attention to handle varying context lengths and prediction horizons, optimizing predictions through residual blocks and positional encodings. Additionally, it supports longer output patches compared to input patches, enabling efficient forecasting of long horizons with fewer autoregressive steps while maintaining robustness across diverse granularities and domains. We use their release checkpoint **google/TimesFM-1.0-200m-pytorch**.

- The LPTM[4] (Kamarthi & Prakash, 2024) is a foundational model designed for multi-domain time-series analysis, leveraging a transformer-based architecture with an adaptive segmentation module. This segmentation module dynamically determines optimal segment lengths during pre-training, ensuring that time-series data from diverse domains are tokenized effectively based on self-supervised learning losses. By incorporating masking-based self-supervised tasks, LPTM learns robust representations, enabling efficient transfer to various downstream tasks such as forecasting and classification.

[2] https://github.com/MOMENT-timeseries-foundation-model/MOMENT
[3] https://github.com/google-research/TimesFM
[4] https://github.com/AdityaLab/Samay

B.2. Dataset

We explain the details of our dataset here:

- The Electricity Transformer Temperature (ETT) dataset monitors the oil temperature of electricity transformers, a critical indicator in power grid management. It includes two years of data collected from two stations in China, with sampling frequencies of 15 minutes (ETTm1) and 1 hour (ETTh1). Each data point comprises the target variable (oil temperature) and six covariates representing power load features.
- The Exchange Rate dataset contains daily foreign exchange rates of eight currencies over a period spanning from 1990 to 2016.
- The Weather dataset contains local climatological data collected hourly from nearly 1,600 locations across the United States.
- PEMS-Bay The PEMS-Bay dataset contains hourly sampled traffic volume records gathered across 300+ sensors across the Bay Area. We choose the first two sensors in our experiment.
- METR-LA The METR-LA dataset contains hourly sampled traffic volume records gathered from Los Angeles. We choose the first twenty sensors in our experiment.