TIME-LLM: TIME SERIES FORECASTING BY REPROGRAMMING LARGE LANGUAGE MODELS

1023040920 Chang Wang

June 19, 2024

Abstract

Time series forecasting holds significant importance in many real-world dynamic systems and has been extensively studied. Unlike natural language process (NLP) and computer vision (CV), where a single large model can tackle multiple tasks, models for time series forecasting are often specialized, necessitating distinct de- signs for different tasks and applications. While pre-trained foundation models have made impressive strides in NLP and CV, their development in time series domains has been constrained by data sparsity. Recent studies have revealed that large language models (LLMs) possess robust pattern recognition and reasoning abilities over complex sequences of tokens. However, the challenge remains in effectively aligning the modalities of time series data and natural language to leverage these capabilities. In this work, we present TIME-LLM, a reprogram- ming framework to repurpose LLMs for general time series forecasting with the backbone language models kept intact. We begin by reprogramming the input time series with text prototypes before feeding it into the frozen LLM to align the two modalities. To augment the LLM's ability to reason with time series data, we propose Prompt-as-Prefix (PaP), which enriches the input context and directs the transformation of reprogrammed input patches. The transformed time series patches from the LLM are finally projected to obtain the forecasts. Our com- prehensive evaluations demonstrate that TIME-LLM is a powerful time series learner that outperforms state-of-the-art, specialized forecasting models. Moreover, TIME-LLM excels in both few-shot and zero-shot learning scenarios.

1 Introduction

Time series forecasting is a critical capability across many real-world dynamic systems (Jin et al., 2023a; Hu et al., 2024), with applications ranging from demand planning (Leonard, 2001) and inven- tory optimization (Li et al., 2022) to energy load forecasting (Liu et al., 2023a) and climate modeling (Schneider Dickinson, 1974). Each time series forecasting task typically requires extensive do- main expertise and task-specific model designs. This stands in stark contrast to foundation language models like GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), Llama (Touvron et al., 2023), inter alia, which can perform well on a diverse range of NLP tasks in a few-shot or even zero-shot setting.

Pre-trained foundation models, such as large language models (LLMs), have driven rapid progress in computer vision (CV) and natural language processing (NLP). While time series modeling has not benefited from the same significant breakthroughs, LLMs' impressive capabilities have inspired their application to time series forecasting (Jin et al., 2023b; 2024). Several desiderata exist for leveraging LLMs to advance forecasting techniques: Generalizability. LLMs have demonstrated a remarkable capability for few-shot and zero-shot transfer learning (Brown et al., 2020).This suggests their potential for generalizable forecasting across domains without requiring per-task retraining from scratch. In contrast, current forecasting methods are often rigidly specialized by domain. Data ef- ficiency. By leveraging pre-trained knowledge, LLMs have shown the ability to perform new tasks with only a few examples. This data efficiency could enable forecasting for settings where historical data is limited. In contrast, current methods typically require abundant in-domain data. Reason- ing. LLMs exhibit sophisticated reasoning and pattern recognition capabilities (Mirchandani et al., 2023; Luo et al., 2023b;a). Harnessing these skills could allow making highly precise forecasts by leveraging learned higher-level concepts. Existing non-LLM methods are largely statistical without much innate reasoning. Multimodal knowledge. As LLM architectures and training techniques im- prove, they gain more diverse knowledge across modalities like vision, speech, and text (Ma et al., 2023). Tapping into this knowledge could enable synergistic forecasting that fuses different data types. Conventional tools lack ways to jointly leverage multiple knowledge bases. Easy optimiza- tion. LLMs are trained once on massive computing and then can be applied to forecasting tasks without learning from scratch. Optimizing existing forecasting models often requires significant architecture search and hyperparameter tuning (Zhou et al., 2023b). In summary, LLMs offer a promising path to make time series forecasting more general, efficient, synergistic, and accessible compared to current specialized modeling paradigms. Thus, adapting these powerful models for time series data can unlock significant untapped potential.

2 Related Work

Task-specific Learning. Most time series forecasting models are crafted for specific tasks and domains (e.g., traffic prediction), and trained end-to-end on small-scale data. An illustration is in Fig.1 (a). For example,

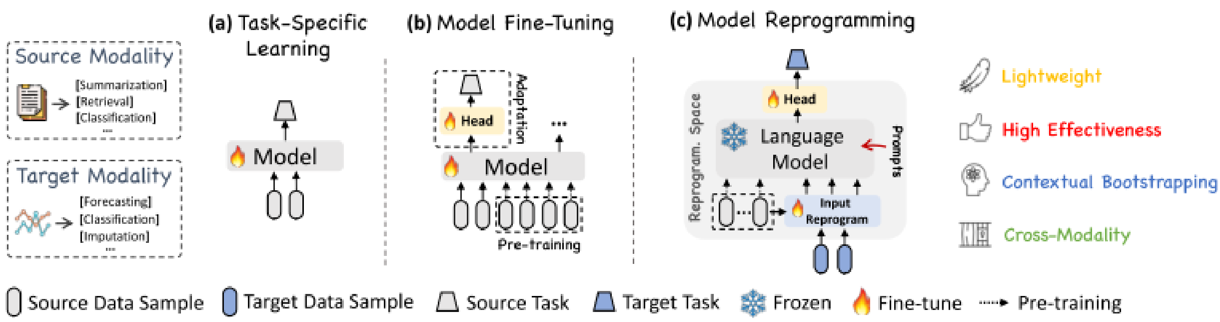

Figure 1: Schematic illustration of reprogramming large language models (LLMs) in comparison of (a) task-specific learning and (b) model fine-tuning. Our proposal investigates and demonstrates (c) how to effectively reprogram open-sourced LLMs as powerful time series learners where well- developed time series pre-trained models are not readily available.

ARIMA models are designed for univariate time series forecasting (Boxet al., 2015), LSTM networks are tailored for sequence modeling (Hochreiter Schmidhuber, 1997), and temporal convolutional networks (Bai et al., 2018) and transformers (Wen et al., 2023) are developed for handling longer temporal dependencies. While achieving good performance on narrow tasks, these models lack versatility and generalizability to diverse time series data.

In-modality Adaptation. Relevant research in CV and NLP has demonstrated the effectiveness of pre-trained models that can be fine-tuned for various downstream tasks without the need for costly training from scratch (Devlin et al., 2018; Brown et al., 2020; Touvron et al., 2023). Inspired by these successes, recent studies have focused on the development of time series pre-trained models (TSPTMs). The first step among them involves time series pre-training using different strategies like supervised (Fawaz et al., 2018) or self-supervised learning (Zhang et al., 2022b; Deldari et al., 2022; Zhang et al., 2023). This allows the model to learn representing various input time series. Once pre- trained, it can be fine-tuned on similar domains to learn how to perform specific tasks (Tang et al., 2022). An example is in Fig. 1(b). The development of TSPTMs leverages the success of pre- training and fine-tuning in NLP and CV but remains limited on smaller scales due to data sparsity.

Cross-modality Adaptation. Building on in-modality adaptation, recent work has further explored transferring knowledge from powerful pre-trained foundations models in NLP and CV to time se- ries modeling, through techniques such as multimodal fine-tuning (Yin et al., 2023) and model reprogramming (Chen, 2022). Our approach aligns with this category; however, there is limited pertinent research available on time series. An example is Voice2Series (Yang et al., 2021), which adapts an acoustic model (AM) from speech recognition to time series classification by editing a time series into a format suitable for the AM. Recently, Chang et al. (2023) proposes LLM4TS for time series forecasting using LLMs. It designs a two-stage fine-tuning process on the LLM - first supervised pre-training on time series, then task-specific fine-tuning. Zhou et al. (2023a) lever- ages pre-trained language models without altering their self-attention and feedforward layers. This model is fine-tuned and evaluated on various time series analysis tasks and demonstrates comparable or state-of-the-art performance by transferring knowledge from natural language pre-training. Dis- tinct from these approach, we neither edit the input time series directly nor fine-tune the backbone LLM. Instead, as illustrated in Fig. 1(c), we propose reprogramming time series with the source data modality along with prompting to unleash the potential of LLMs as effective time series machines.

3 METHODOLOGY

We consider the following problem: given a sequence of historical observations $\mathbf{X} \in R^{N \times T}$ consisting of N different 1-dimensional variables across T time steps, we aim to reprogram a large language model $f(\cdot)$ to understand the input time series and accurately forecast the readings at H future time steps, denoted by $\hat{Y} \in R^{N \times H}$, with the overall objective to minimize the mean square errors between the ground truths $\mathbf{Y}$ and predictions, i.e., $\frac{1}{H}\sum_{h=1}^{H} \|\hat{Y}_h - Y_h\|_F^2$.

Our method encompasses three main components: (1) input transformation, (2) a pre-trained and frozen LLM, and (3) output projection. Initially, a multivariate time series is partitioned into N univariate time

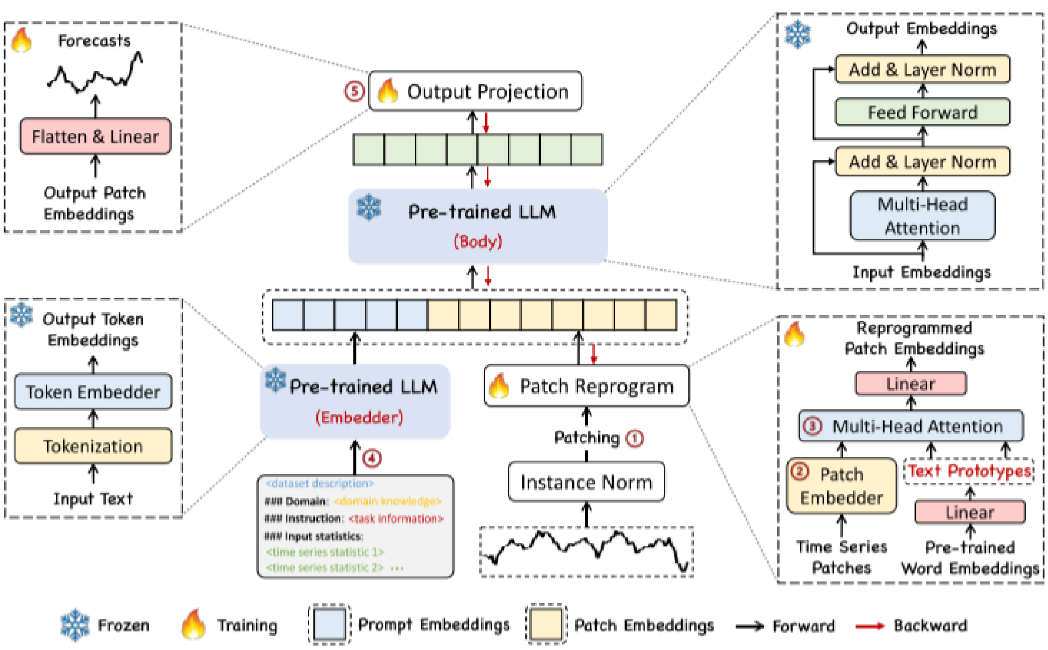

Figure 2: : The model framework of TIME-LLM.

series, which are subsequently processed independently (Nie et al., 2023). The i-th series is denoted as $\mathbf{X}^{(i)} \in R^{1 \times T}$, which undergoes normalization, patching, and embedding prior to being reprogrammed with learned text prototypes to align the source and target modalities. Then, we augment the LLM's time series reasoning ability by prompting it together with reprogrammed patches to generate output representations, which are projected to the final forecasts $\hat{\mathbf{Y}}^{(i)} \in R^{1\times H}$.

We note that only the parameters of the lightweight input transformation and output projection are updated, while the backbone language model is frozen. In contrast to vision-language and other multimodal language models, which usually fine-tune with paired cross-modality data, TIME-LLM is directly optimized and becomes readily available with only a small set of time series and a few training epochs, maintaining high efficiency and imposing fewer resource constraints compared to building large domain-specific models from scratch or fine-tuning them. To further reduce mem- ory footprints, various off-the-shelf techniques (e.g., quantization) can be seamlessly integrated for slimming TIME-LLM.

4 Results

TIME-LLM consistently outperforms state-of-the-art forecasting methods by large margins across multiple benchmarks and settings, especially in few-shot and zero-shot scenarios. We compared our approach against a broad collection of up-to-date models, including a recent study that fine-tunes language model for time series analysis (Zhou et al., 2023a). To ensure a fair comparison, we adhere to the experimental configurations in (Wu et al., 2023) across all baselines with a unified evaluation pipeline1. We use Llama-7B (Touvron et al., 2023) as the default backbone unless stated otherwise.

Baselines. We compare with the SOTA time series models, and we cite their performance from (Zhou et al., 2023a) if applicable. Our baselines include a series of Transformer-based methods: PatchTST (2023), ESTformer (2022), Non-Stationary Transformer (2022), FEDformer (2022), Aut- oformer (2021), Informer (2021), and Reformer (2020). We also select a set of recent competi- tive models, including GPT4TS (2023a), LLMTime (2023), DLinear (2023), TimesNet (2023), and LightTS (2022a). In short-term forecasting, we further compare our model with N-HiTS (2023b) and N-BEATS (2020). More details are in Appendix A.

5 Conclusion

TIME-LLM shows promise in adapting frozen large language models for time series forecasting by reprogramming time series data into text prototypes more natural for LLMs and providing natural language guidance via Prompt-as-Prefix to augment reasoning. Evaluations demonstrate the adapted LLMs can outperform specialized expert models, indicating their potential as effective time series machines. Our results also provide a novel insight that time series forecasting can be cast as yet an- other "language" task that can be tackled by an off-the-shelf LLM to achieve state-of-the-art perfor- mance through our Time-LLM framework. Further research should explore optimal reprogramming representations, enrich LLMs with explicit time series knowledge through continued pre-training, and build towards multimodal models with joint reasoning across time series, natural language, and other modalities. Furthermore, applying the reprogramming framework to equip LLMs with broader time series analytical abilities or other new capabilities should also be considered.

References

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[2] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[4] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza, Max Mergenthaler, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[5] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023.

[6] Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*, 2023.

[7] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. *arXiv preprint arXiv:2202.10629*, 2022.

[8] Shohreh Deldari, Hao Xue, Aaqib Saeed, Jiayuan He, Daniel V Smith, and Flora D Salim. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv preprint arXiv:2206.02353*, 2022.

[9] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 2023.

[10] 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*, 2018.

[11] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Transfer learning for time series classification. In *IEEE International Conference on Big Data*, pp. 1367–1376. IEEE, 2018.

[12] Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 2023.

[13] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[14] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[16] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza, Max Mergenthaler, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[17] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023.

[18] Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*, 2023.

[19] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. *arXiv preprint arXiv:2202.10629*, 2022.

[20] Shohreh Deldari, Hao Xue, Aaqib Saeed, Jiayuan He, Daniel V Smith, and Flora D Salim. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv preprint arXiv:2206.02353*, 2022.

[21] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 2023.

[22] 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*, 2018.

[23] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Transfer learning for time series classification. In *IEEE International Conference on Big Data*, pp. 1367–1376. IEEE, 2018.

[24] Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 2023.

[25] Linhao Luo, Yuan-Fang Li, Reza Haf, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations*, 2023b.

[26] Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, and Xie Chen. Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition. *arXiv preprint arXiv:2309.10294*, 2023.

[27] Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476, 2000.

[28] 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.

[29] Igor Melnyk, Vijil Chenthamarakshan, Pin-Yu Chen, Payel Das, Amit Dhurandhar, Inkit Padhi, and Devleena Das. Reprogramming pretrained language models for antibody sequence infilling. In *International Conference on Machine Learning*, 2023.

[30] Suvir Mirchandani, Fei Xia, Pete Florence, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, Andy Zeng, et al. Large language models as general pattern machines. In *Proceedings of the 7th Annual Conference on Robot Learning*, 2023.

[31] Diganta Misra, Agam Goyal, Bharat Runwal, and Pin-Yu Chen. Reprogramming under constraints: Revisiting efficient and reliable transferability of lottery tickets. *arXiv preprint arXiv:2308.14969*, 2023.

[32] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[33] OpenAI. Gpt-4 technical report, 2023.

[34] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[37] Stephen H Schneider and Robert E Dickinson. Climate modeling. *Reviews of Geophysics*, 12(3):447–493, 1974.