

# Large Language Models for Financial Aid in Financial Time-series Forecasting

Md Khairul Islam <sup>1</sup>, Ayush Karmacharya <sup>1</sup>, Timothy Sue <sup>1</sup>, Judy Fox <sup>1,2</sup>

<sup>1</sup> Computer Science Department, University of Virginia

<sup>2</sup> School of Data Science, University of Virginia  
Charlottesville, USA

Email : {mi3se, psb7wm, und2yw, cwk9mp}@virginia.edu

**Abstract**—Considering the difficulty of financial time series forecasting in financial aid, much of the current research focuses on leveraging big data analytics in financial services. One modern approach is to utilize "predictive analysis", analogous to forecasting financial trends. However, many of these time series data in Financial Aid (FA) pose unique challenges due to limited historical datasets and high dimensional financial information, which hinder the development of effective predictive models that balance accuracy with efficient runtime and memory usage. Pre-trained foundation models are employed to address these challenging tasks. We use state-of-the-art time series models including pre-trained LLMs (GPT-2 as the backbone), transformers, and linear models to demonstrate their ability to outperform traditional approaches, even with minimal ("few-shot") or no fine-tuning ("zero-shot"). Our benchmark study, which includes financial aid with seven other time series tasks, shows the potential of using LLMs for scarce financial datasets.

**Index Terms**—Financial Aid, Time Series Forecast, Deep Learning, Foundation Models, Large Language Models

## I. INTRODUCTION

The advancement of AI has taken over many domains including the field of financial market [1] and big data [2]. In particular, financial time series forecasting has improved significantly from using statistical models to machine learning [3] and then deep learning [4]. These financial forecasting areas include currency exchange rate [5], [6], stock market [2] [5] [4], commodity prices [7] [8] and more. **Pre-trained foundation models**, such as large language models (LLMs) have driven the progress in Natural Language Processing (NLP) and Computer Vision (CV). Foundation models like GPT [9], and Vision Transformer [10] can perform well on a diverse range of tasks in few-shot (little training) or zero-shot (no training) learning. This enables applications where historical data is limited or mostly missing.

**Financial time series forecasting (FTSF)** is an important domain that needs more attention among the multivariate time series forecasting tasks. Previous works on FTSF rely heavily on machine learning [3] or traditional deep learning [4] methods. With some recent works on multi-modal FTSF [11] [12]. **Financial aid (FA)** is crucial to many students' educational journeys, providing the resources needed to pursue academic dreams while fostering educational equity. However, the process, including tasks like processing the free application for Federal Student Aid (FAFSA), is often manual,

time-consuming, and susceptible to errors. Access to historical datasets is limited to yearly intervals and is subject to changes in policy.

We describe and evaluate LLM-based foundation models in the FTSF domain using 8 deep-learning models and compare especially the financial aid with 7 other financial datasets. Our research questions are,

- *Q1*: Which models are better as few-shot learners?
- *Q2*: Can pre-trained LLMs perform zero-shot learning?

Answering these questions will help us better understand the current advancement of LLMs for financial time series forecasting. In summary, our contributions are,

- Collect eight datasets from four financial domains (Stock, Commodity, Currency, Institution) for over 10 years.
- Benchmark five state-of-the-art time series deep learning models, and three LLM-based foundation models on these datasets.
- Open source code and datasets at GitHub <sup>1</sup> to facilitate full reproducibility and further research in this domain.

## II. METHODOLOGY

### A. Problem Statement

Given the input dataset,  $X \in \mathcal{R}^{F \times T}$ ,  $T$  denotes the total timesteps in days and  $F$  input features (including past targets and other features). With a lookback window of  $L$  past days, the input at time  $t$  is  $X_t = X_{t-(L-1):t}$  which contains inputs of the last  $L$  days. Given this input  $X_t$ , the model  $f$  predicts the targets  $O$  (e.g. stock prices) for the next  $\tau_{max}$  days. The target output  $y_t$  at time  $t$  can be expressed as,

$$\hat{y}_t = f(X_t), \text{ where,}$$

$$\begin{aligned} X_t &= x_{t-(L-1):t} = [x_{t-(L-1)}, x_{t-(L-2)}, \dots, x_t] & (1) \\ &= \{x_{f,l,t}\}, f \in \{1, \dots, F\}, l \in \{1, \dots, L\} \end{aligned}$$

For the financial aid data, funds allocated to each state are a time series with  $T$  years (2004 to 2020). A lookback  $L$  of 10 years is used to predict funds ( $O$ ) for the next year ( $\tau_{max} = 1$ ). For all other datasets, we have daily inputs for 10 years. With a lookback window of the past 96 days ( $L = 96$ ), we predict the targets for the next 24 days ( $\tau_{max} = 24$ ).

<sup>1</sup><https://github.com/UVA-MLSys/Financial-Time-Series>

TABLE I: Datasets overview. Time series indicates the number of target time series (i.e., channels). Input features are past observations and dataset size is depicted as training, validation, and test.

Domain	Dataset	Date	Frequency	Time series	Lookback	Horizon	Size
Institution	Financial Aid	2004 - 2020	Yearly	56	10	1	(137, 92, 92)
Stock	S&P 500	Sep 1, 2014 - Aug 29, 2024	Daily	4	96	24	(1903, 231, 229)
	Apple	Sep 2, 2014 - Aug 29, 2024	Daily	5	96	24	(1893, 230, 228)
	Microsoft	Sep 3, 2014 - Aug 30, 2024	Daily	4	96	24	(1893, 230, 228)
Commodity	Crude Oil	Sep 1, 2014 - Aug 29, 2024	Daily	4	96	24	(1893, 230, 228)
	Gold	Sep 1, 2014 - Aug 29, 2024	Daily	4	96	24	(1893, 230, 228)
	Natural Gas	Sep 1, 2014 - Aug 29, 2024	Daily	4	96	24	(1893, 230, 228)
Currency	Exchange Rate	Aug 1, 2014 - Aug 1, 2024	Daily	7	96	24	(1861, 225, 224)

### B. Dataset

We use the following financial datasets: (1) **Financial Aid**: Financial aid distributed to each US state by the Government to support student education and collected from years 2004 to 2020 [13]. Details of available features are in Table II and the yearly aggregated aid in Fig 1. (2) **Stock Market** [5] [12] [4] [14]: Includes the daily stock prices (Close, Open, High, Low) and volumes for each of the following stocks up to 10 years from the NASDAQ database: S&P 500 (SPX), Microsoft Corporation (MSFT), and Apple Inc (AAPL); (3) **Commodities** [7] [8]: Contains data (Close, Open, Volume, High, Low) on different kinds of raw materials such as Natural Gas, Crude Oil and Gold. (4) **Currency Exchange Rate** [5] [6]: The currency units per U.S. dollar reported daily by the issuing central bank (rates are not recorded on the weekends and certain holidays). This data covers the following currencies: Australian Dollar (AUD), Canadian Dollar (CAD), Chinese yuan (CNY), Euro (EUR), Indian rupee (INR), Japanese yen (JPY), and the U.K. pound (GBP);

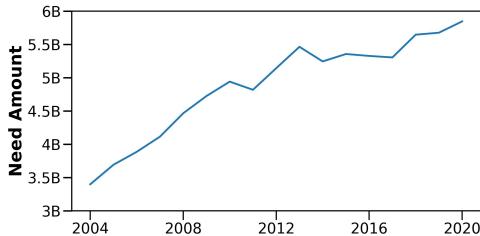


Fig. 1: Financial Aid data aggregated at the state level from 2004 to 2020 (17 years), in billions of US dollars. Access to historical datasets is limited to yearly intervals.

The statistics are in Table I. The train, validation, and test split follows the 8:1:1 ratio, the validation set follows the train set, then the test set. The data is standard normalized before passing to the model. The few missing values (<1%) are imputed using last-seen valid values.

### C. Models

We use the following time series models in our work. The models are chosen based on their popularity and recently

TABLE II: List of available features in financial aid [13]. Aid is given based on financial needs, academic merit, or both. The sub-categories are simplified and describe multiple features.

Category	Sub-category	Description
Need, Merit, both	Identifier	State id and name abbreviation.
	Number	Total students receiving the award.
	Public/Private	Whether the funds can be used for public or private sectors and how long (2 or 4 years).
	Flags	0 or 1 based on whether the aid falls in a particular category.
	Program	Aid program with the most generous eligibility criteria.
	Notes	Related text.
	Threshold	GPA, SAT, income, and other academic or financial limits to qualify for the aid.
	Time	Year
Target	Amount	Aid amount received by the students.

published work. We focus on point forecasting in our work. A high-level overview of how pre-trained LLMs are fine-tuned for custom datasets is illustrated in Figure 2.

1) **LLM Foundation Models**: The LLM-based foundation models are selected based on their versatility in time series forecasting. We use the configurations from [15]. The pre-trained foundation models are frozen except for the last layer when fine-tuning. These models use a pre-trained GPT-2 [9] as the LLM backbone. We select the following recent models: (1) TimeLLM [16] (2) CALF [17] (3) GPT4TS (One Fits All, [18]).

2) **Traditional Models**: We choose the following recent non-pre-trained models: (1) DLinear [19] (2) iTransformer [20] (3) TimesNet [6] (4) PatchTST [21] (5) TimeMixer [22]. Most of these models are Transformer-based and have shown great performance in capturing temporal patterns.

### D. Implementation Details

We use the PyTorch framework and follow [6] [15] to implement our experiments. Each experiment runs three times with different random seeds (648, 506, 608), and the average results are presented. Following [18] [17] we use the **pre-trained GPT2** as the backbone for the LLMs and only fine-tune the output layer during training shown in Fig. 2. The

TABLE III: *Q1*. Few-shot learning performance with 10% training data. TimeLLM and PatchTST outperform the other models. The best and the second best results are in **bold** and underlined.

Method	DLinear [19]		PatchTST [21]		TimesNet [6]		TimeMixer [22]		iTransformer [20]		TimeLLM [23]		CALF [17]		GPT4TS [18]	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Financial Aid	2.94	1.36	2.23	<u>0.95</u>	2.17	1.04	3.04	1.36	2.22	1.14	<u>2.06</u>	1.08	<b>1.64</b>	<b>0.89</b>	2.36	1.20
S&P 500	2.08	1.14	<u>2.03</u>	<b>1.14</b>	2.43	1.18	2.49	1.25	2.19	1.19	<b>1.94</b>	<u>1.14</u>	2.37	1.19	3.07	1.40
Apple	2.78	1.30	<u>2.36</u>	<b>1.21</b>	3.22	1.41	3.44	1.46	3.05	1.39	3.00	1.36	<b>2.33</b>	<u>1.21</u>	2.80	1.29
Microsoft	2.51	1.10	<b>2.13</b>	<b>1.06</b>	3.55	1.43	2.85	1.18	2.49	1.15	<u>2.40</u>	<u>1.10</u>	2.97	1.27	3.40	1.37
Crude Oil	<b>1.66</b>	<u>1.01</u>	1.96	1.05	2.75	1.29	2.06	1.14	1.92	1.07	<u>1.75</u>	<b>1.01</b>	2.13	1.14	2.42	1.15
Gold	2.78	<u>1.14</u>	2.68	1.15	<u>2.66</u>	1.17	<b>2.57</b>	<b>1.11</b>	3.00	1.22	2.71	1.17	3.06	1.24	3.38	1.32
Natural Gas	2.20	1.16	2.48	1.24	2.55	1.23	2.91	1.34	<b>2.12</b>	<b>1.12</b>	2.36	1.20	2.30	1.17	2.17	1.13
Exchange	1.48	0.92	1.28	0.85	2.87	1.34	1.29	0.84	1.65	0.96	<u>1.20</u>	<b>0.81</b>	1.48	0.92	<u>1.22</u>	<u>0.81</u>

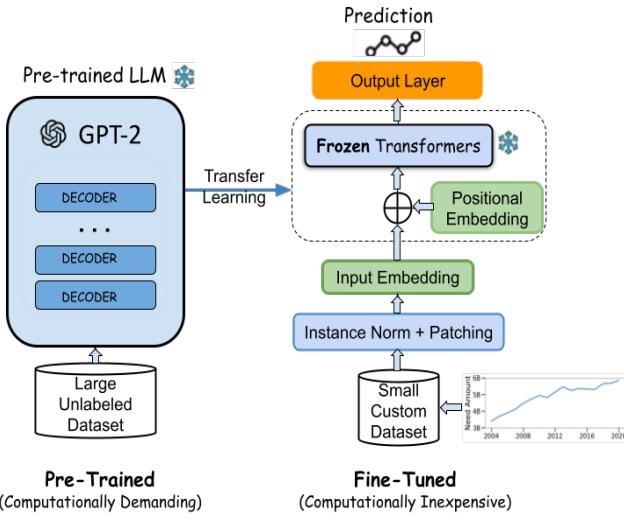


Fig. 2: A high-level overview of pre-training an LLM and fine-tuning on a custom dataset (e.g. the Financial Aid dataset) for downstream tasks.

traditional models are trained from scratch. We use the Adam optimizer with a learning rate 1e-3 and a dropout rate 0.1. The GPT4TS [18] model uses the L1 loss. The CALF [17] model uses the weighted average of task, feature, and logit loss using the L1 loss. The other models use the Mean Squared Error loss. Models are trained for 10 epochs max, with batch size 32. The experiments were run on an NVIDIA 2080Ti GPU with 11GB+ memory with 32GB RAM. Following [22] [17] [14], we use Mean Square Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics. **Lower is better** for these metrics.

### III. EXPERIMENTS AND RESULTS

In this section, we investigate the research questions, the setup, and the results.

#### A. *Q1. Which models are better as few-shot learners?*

Pre-trained models are preferred largely due to their generalizability and good performance in few-shot or zero-shot learning settings [24]. Since these LLMs are already trained on many datasets, they often outperform the other models when few training data are available [17] or without training [14].

Following [17], we select the last 10% training data to train the models in a few-shot learning setting.

**Results.** The few-shot learning results are shown in Table III. All model performance drops significantly after reducing the train data size. This is due to the models' inability to learn enough temporal patterns from the limited input. TimeLLM performs the best overall (three best and four 2nd best cases). While PatchTST performs the 2nd best with a close margin (three best and two 2nd best cases). DLinears performance drops significantly as it is a simple linear model. However, overall *the LLMs performed better in the few-shot learning*.

#### B. *Q2. Can LLMs perform zero-shot learning in FTSF?*

Real-world scenarios can often have no available past observations (i.e. new company stock in the market, newly launched product). Having zero-shot learning ability is crucial in forecasting those cases since traditional deep learning time series models are unable to train and forecast those cases. We investigate whether LLMs can effectively assist in those cases. We load the pre-trained LLMs and evaluate them on the test set without fine-tuning. Since no training is done in this part, we exclude the traditional time series models from this analysis.

TABLE IV: *Q2*. Zero shot performance. GPT4TS performs the best. The best and the second best results are in **bold** and underlined. The traditional models are excluded here since they are not pre-trained.

Method	TimeLLM [23]		CALF [17]		GPT4TS [18]	
Metric	MSE	MAE	MSE	MAE	MSE	MAE
Financial Aid	<b>2.99</b>	<b>1.29</b>	3.82	1.59	<u>3.17</u>	<u>1.41</u>
S&P 500	5.04	1.91	<u>3.98</u>	<b>1.74</b>	<b>3.89</b>	<u>1.76</u>
Apple	4.17	1.61	<u>3.36</u>	<u>1.44</u>	<b>3.05</b>	<b>1.36</b>
Microsoft	5.13	1.81	<u>4.12</u>	<u>1.62</u>	<b>3.96</b>	<b>1.59</b>
Crude Oil	3.05	1.39	<u>2.21</u>	<u>1.18</u>	<b>1.89</b>	<b>1.08</b>
Gold	6.15	1.95	<u>5.12</u>	<b>1.76</b>	<b>5.00</b>	<u>1.77</u>
Natural Gas	4.09	1.61	<u>3.27</u>	<u>1.43</u>	<b>2.96</b>	<b>1.35</b>
Exchange	3.25	1.47	<u>2.41</u>	<u>1.29</u>	<b>2.10</b>	<b>1.23</b>

**Results.** Table IV shows the zero-shot results of the LLMs. Compared to *Q1*, the results achieved here are significantly worse. Since each financial data may have distinct temporal patterns, without fine-tuning the LLMs fail to forecast them effectively. We conclude, *LLMs are yet not quite effective for zero-shot learning for financial time series*.

#### IV. RELATED WORKS

Deep learning for time series has significantly outperformed machine learning approaches [22] [20], also in finance [4]. [5] used RNN models to forecast stock market prices, and currency exchange rates. Many recent deep learning models have been used to forecast the stock market [2] [5] [4], commodity prices [7] [8]. Foundation models in time series have recently gained significant attention [24]. Pre-trained LLMs and vision models have been enhanced for time series. [14] [25] showed the ability of LLMs to perform in zero-shot and few-shot settings in time series tasks. GPT4TS [18] leverages pre-trained language models without altering important layers. TimeLLM [16] reprograms LLM's ability to reason with time series data by proposing a prompt-as-prefix technique. CALF [17] proposed a novel fine-tuning framework to reduce the distribution discrepancy between textual and temporal data. Chronos [26] performed significantly in probabilistic forecasting.

#### V. CONCLUSION AND FUTURE WORKS

In this paper, we benchmark financial datasets from multiple domains using state-of-the-art time series models and LLM-based foundation models. Our results show that LLMs are more effective for few-shot and zero-shot learning. **Especially, the few-shot and zero-shot capabilities of LLMs can be effective for financial Aid practitioners who are currently unable to apply deep learning methods due to limited data availability.** We focus on point forecasting with a single modality in this work. Incorporating data from different financial modalities into time series models will be future work, and probabilistic forecasting can help the financial domain by outputting a probabilistic distribution. Our research highlights the potential of foundation LLMs in financial aid and the overall finance time series forecasting domain.

#### ACKNOWLEDGMENT

This work is partly supported by NSF grant CCF-1918626 Expeditions: Collaborative Research: Global Pervasive Computational Epidemiology, and NSF Grant 2200409 for CyberTraining:CIC: CyberTraining for Students and Technologies from Generation Z.

#### REFERENCES

- [1] L. Cao, "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022.
- [2] Y. Zu, J. Mi, L. Song, S. Lu, and J. He, "Finformer: A static-dynamic spatiotemporal framework for stock trend prediction," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 1460–1469.
- [3] T. Leung and T. Zhao, "Financial time series analysis and forecasting with hilbert–huang transform feature generation and machine learning," *Applied Stochastic Models in Business and Industry*, vol. 37, no. 6, pp. 993–1016, 2021.
- [4] X. Li, X. Shen, Y. Zeng, X. Xing, and J. Xu, "Finreport: Explainable stock earnings forecasting via news factor analyzing model," in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 319–327.
- [5] K. Sako, B. N. Mpinda, and P. C. Rodrigues, "Neural networks for financial time series forecasting," *Entropy*, vol. 24, no. 5, p. 657, 2022.
- [6] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations*, 2023.
- [7] X. Xu and Y. Zhang, "Commodity price forecasting via neural networks for coffee, corn, cotton, oats, soybeans, soybean oil, sugar, and wheat," *Intelligent Systems in Accounting, Finance and Management*, vol. 29, no. 3, pp. 169–181, 2022.
- [8] S. Zhang, J. Luo, S. Wang, and F. Liu, "Oil price forecasting: A hybrid gru neural network based on decomposition-reconstruction methods," *Expert Systems with Applications*, vol. 218, p. 119617, 2023.
- [9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [10] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [11] Z. Chen, L. N. Zheng, C. Lu, J. Yuan, and D. Zhu, "Chatgpt informed graph neural network for stock movement prediction," *arXiv preprint arXiv:2306.03763*, 2023.
- [12] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu, "Temporal data meets llm-explainable financial time series forecasting," *arXiv preprint arXiv:2306.11025*, 2023.
- [13] R. K., B. D., J. Ortagus, K. R., E. A., L. M., and C. J., "State financial aid dataset," 2023. [Online]. Available: <https://informedstates.org/state-financial-aid-dataset-download>
- [14] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [15] M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen, "Are language models actually useful for time series forecasting?" *arXiv preprint arXiv:2406.16964*, 2024.
- [16] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan *et al.*, "Time-llm: Time series forecasting by reprogramming large language models," *arXiv preprint arXiv:2310.01728*, 2023.
- [17] P. Liu, H. Guo, T. Dai, N. Li, J. Bao, X. Ren, Y. Jiang, and S.-T. Xia, "Taming pre-trained llms for generalised time series forecasting via cross-modal knowledge distillation," *arXiv preprint arXiv:2403.07300*, 2024.
- [18] T. Zhou, P. Niu, L. Sun, R. Jin *et al.*, "One fits all: Power general time series analysis by pretrained lm," *Advances in neural information processing systems*, vol. 36, pp. 43 322–43 355, 2023.
- [19] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [20] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=JePfAI8fah>
- [21] Y. Nie, N. H. Nguyen, P. Sindhong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.
- [22] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. Zhou, "Timemixer: Decomposable multiscale mixing for time series forecasting," *arXiv preprint arXiv:2405.14616*, 2024.
- [23] M. Jin, H. Tang, C. Zhang, Q. Yu, C. Liu, S. Zhu, Y. Zhang, and M. Du, "Time series forecasting with llms: Understanding and enhancing model capabilities," *arXiv preprint arXiv:2402.10835*, 2024.
- [24] Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen, "Foundation models for time series analysis: A tutorial and survey," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6555–6565.
- [25] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," *arXiv preprint arXiv:2310.10688*, 2023.
- [26] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor *et al.*, "Chronos: Learning the language of time series," *arXiv preprint arXiv:2403.07815*, 2024.