

# Supplementary Material

## Anonymous submission

### Datasets characteristics

**ETT**(Zhou et al. 2021) contains two sub-datasets: ETT1 and ETT2, collected from two electricity transformers at two stations. Each of them has two versions in different resolutions (15min & 1h). The ETT dataset contains multiple series of loads and one series of oil temperatures.

**Electricity Consuming Load (ECL)**<sup>1</sup> corresponds to the electricity consumption (Kwh) of 321 clients.

**Weather**<sup>2</sup> contains 21 meteorological indicators, such as air temperature, humidity, etc, recorded every 10 minutes for the entirety of 2020.

**Exchange** collects the daily exchange rates of 8 countries (Australia, British, Canada, Switzerland, China, Japan, New Zealand, and Singapore) from 1990 to 2016.

**National Illness (ILI)**<sup>3</sup> corresponds to the weekly recorded influenza-like illness patients from the US Center for Disease Control and Prevention.

### Baseline Methods

Our baseline methods are described as follows:

- CALF is a novel framework that enhances multivariate time series forecasting by aligning the distribution discrepancy between textual and temporal inputs through cross-modal matching, feature regularization, and output consistency.
- Time-LLM is a reprogramming framework that reprogrammes frozen large language models for general time series forecasting by aligning time series and language modalities through text-based input transformation and a novel Prompt-as-Prefix (PaP) mechanism.
- GPT4TS is a framework that leverages language models pre-trained on large-scale data for general time series analysis, achieving strong performance across diverse tasks without modifying the core architecture of the original transformer.
- iTransformer is a Transformer-based approach which innovatively introduces an inverted perspective to model

time-series data for excellent performance and versatility.

- PatchTST is a Transformer-based method by patching and channel independence strategies for multivariate time series forecasting.
- TimesNet is a CNN-based method which transforms 1D sequence into 2D tensor to capture the periodicity of the time series.
- FEDformer is a forecasting model that integrates seasonal-trend decomposition and frequency-domain enhancement to effectively and efficiently capture both global trends and detailed patterns in time series, achieving superior accuracy with linear computational complexity.
- DLinear is an MLP-based approach which challenges the Transformer-based methods with a one-layer linear layer.

### Detailed Experimental Settings

We use ADAM as the default optimizer and report the mean squared error (MSE) and mean absolute error (MAE) as the evaluation metrics. A lower MSE/MAE indicates a better performance. All experiments are implemented using PyTorch and conducted with a fixed random seed on two NVIDIA H20 GPUs (96GB each). For models (Zhou et al. 2023; Jin et al. 2024; Nie et al. 2023; Zeng et al. 2023) whose default input length is not 96, we modify only the input length to 96 while keeping other settings unchanged. To ensure a fair comparison, we adopt the official implementations of the backbone models and follow their default configurations. Most of the baseline results are taken from the original papers of CALF (Liu et al. 2025) and iTransformer (Liu et al. 2023).

### More Visualization on Local and Global Representations

Our motivation is inspired by the hierarchical representation structure observed in pre-trained large language models. We begin by visualizing the similarity matrices of hidden representations across different layers in our proposed Logo-LLM, as illustrated in Figure 1. To systematically analyze this phenomenon, we further visualize layer-wise similarity matrices for two representative baselines, CALF (Liu

<sup>1</sup><https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014>

<sup>2</sup><https://www.bgc-jena.mpg.de/wetter/>

<sup>3</sup><https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>

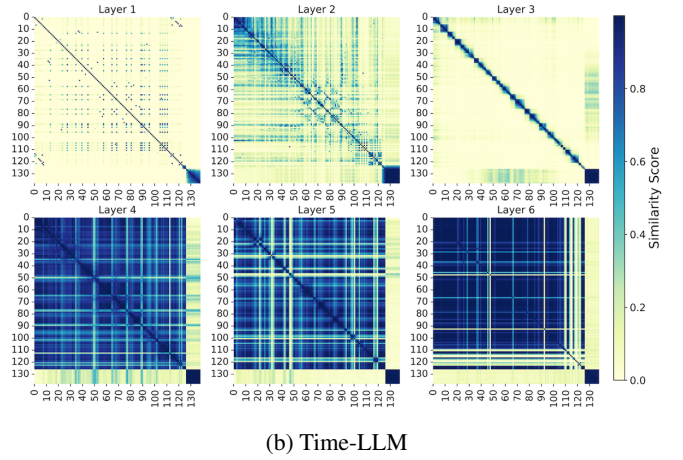
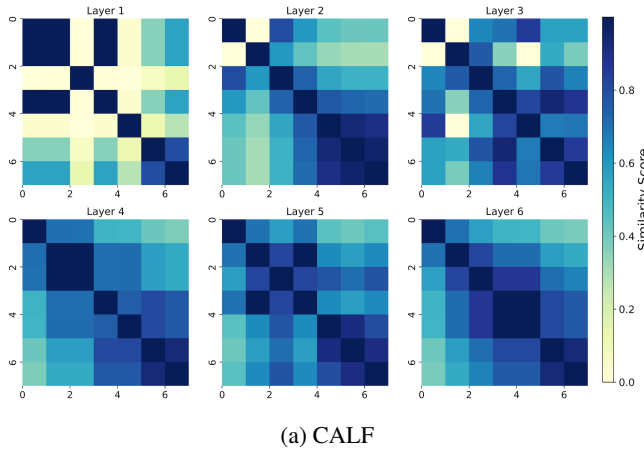


Figure 1: Similarity matrices of each patch across Transformer layers in (a) Logo-LLM (b) CALF and (c) Time-LLM, illustrating that shallow layers exhibit pronounced local patterns while deeper layers capture broader global dependencies.

et al. 2025) and Time-LLM (Jin et al. 2024), with the number of layers standardized to 6. As illustrated in Figure 1a and 1b, the visualizations reveal a consistent hierarchical pattern across all models: shallow layers tend to focus on local variations and fine-grained patterns, while deeper layers progressively capture global trends and long-range dependencies. These findings suggest that the pre-trained LLM inherently develops hierarchical semantics when applied to time series forecasting, enabling the LLM to extract multi-scale temporal information effectively.

### Ablation on Input Length

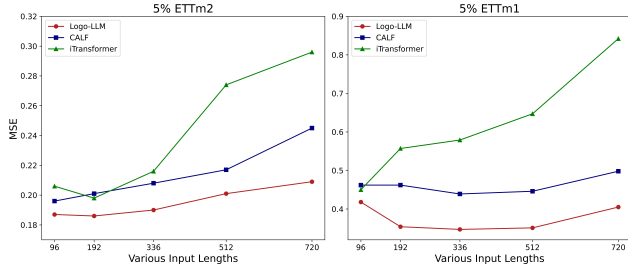


Figure 2: Ablation study on various input lengths. We conducted experiments with the input length  $L = \{96, 192, 336, 512, 720\}$  and the prediction length  $T = 96$  on 5% training data of ETTm1 and ETTm2 datasets. The figures demonstrate the strong adaptability and robustness of our model when processing temporal contexts of varying lengths.

Figure 2 presents the performance comparison of Logo-LLM, CALF (Liu et al. 2025), and iTransformer (Liu et al. 2023) across different input lengths on ETTm1 and ETTm2 datasets using MSE as the evaluation metric. From the figures, Logo-LLM consistently achieves the lowest MSE under all input length settings. This highlights the benefit of our layer-wise local-global modeling strategy, better exploit-

ing LLM’s semantic hierarchy and filtering useful temporal cues from long contexts. For all models, increasing the input length does not always lead to better performance, indicating that simple extension of input horizon may introduce more noise or redundant dependencies in few-shot learning scenarios.

### Full Results of Few-shot Learning

Following the findings of Zeng et al. (2023) and Nie et al. (2023), which demonstrate the effectiveness of the channel-independence strategy for time series data, we treat each multivariate time series as a collection of independent univariate series. Consistent with standard experimental protocols (Liu et al. 2025), each time series is divided into three subsets: training, validation, and testing. For the few-shot forecasting setting, only 5% of the training timesteps is used, while the validation and test sets remain unchanged. The evaluation metrics are the same as those used in conventional multivariate time series forecasting. As shown in Table 1, each experiment is repeated three times, and we report the average results.

### Limitations

While Logo-LLM demonstrates promising performance in time series forecasting, several limitations remain. First, the current method primarily relies on empirical observations to guide the selection of shallow and deep layers for local and global modeling, lacking a more rigorous theoretical analysis of the layer-wise representations. Then, Logo-LLM is evaluated mainly on standard benchmarks and its effectiveness under real-world distribution shifts or domain-specific anomalies remains to be explored. Finally, Logo-LLM mainly leverages the local and global temporal patterns from the hierarchical features of LLMs, yet the rich world knowledge embedded in pre-trained LLMs remains underutilized. How to fully exploit this knowledge and design more powerful frameworks that seamlessly integrate

Methods		Logo-LLM (Ours)		CALF (2025)		Time-LLM (2024)		GPT4TS (2023)		iTransformer (2023)		PatchTST (2023)		TimesNet (2023)		FEDformer (2022)		DLinear (2023)	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.418	0.406	0.462	0.438	0.560	0.477	0.545	0.472	0.450	0.473	0.399	0.414	0.606	0.518	0.628	0.544	0.437	0.430
	192	0.517	0.446	0.516	0.468	0.574	0.495	0.548	0.473	0.592	0.508	0.441	0.436	0.681	0.539	0.666	0.566	0.496	0.462
	336	0.575	0.474	0.618	0.525	0.596	0.504	0.619	0.503	0.730	0.568	0.499	0.467	0.786	0.597	0.807	0.628	0.562	0.503
	720	0.691	0.565	0.849	0.618	0.861	0.632	0.790	0.592	1.151	0.719	0.767	0.587	0.796	0.593	0.822	0.633	0.792	0.612
	Avg	0.540	0.473	0.611	0.512	0.648	0.527	0.626	0.510	0.731	0.567	0.526	0.476	0.717	0.561	0.730	0.592	0.572	0.502
ETTm2	96	0.187	0.266	0.196	0.278	0.200	0.282	0.196	0.280	0.206	0.292	0.206	0.288	0.220	0.299	0.229	0.320	0.305	0.386
	192	0.255	0.310	0.269	0.326	0.267	0.325	0.267	0.324	0.284	0.343	0.264	0.324	0.311	0.361	0.394	0.361	0.413	0.450
	336	0.319	0.348	0.338	0.364	0.328	0.360	0.323	0.353	0.338	0.363	0.334	0.367	0.338	0.427	0.378	0.427	0.482	0.487
	720	0.465	0.434	0.454	0.431	0.478	0.445	0.525	0.472	0.543	0.476	0.454	0.432	0.509	0.510	0.523	0.510	0.839	0.646
	Avg	0.306	0.339	0.314	0.350	0.318	0.353	0.328	0.357	0.343	0.369	0.314	0.352	0.344	0.372	0.381	0.404	0.510	0.492
ETTh1	96	0.472	0.448	0.528	0.490	0.576	0.519	0.513	0.478	0.632	0.538	0.557	0.519	0.892	0.625	0.593	0.529	0.547	0.503
	192	0.572	0.508	0.992	0.670	0.818	0.635	0.732	0.592	0.917	0.669	0.711	0.570	0.940	0.665	0.652	0.563	0.720	0.604
	336	0.737	0.587	1.000	0.680	1.194	0.765	0.774	0.602	0.944	0.665	0.816	0.619	0.945	0.653	0.731	0.594	0.984	0.727
	720	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Avg	0.606	0.519	0.840	0.613	0.863	0.640	0.673	0.557	0.831	0.624	0.694	0.569	0.925	0.647	0.658	0.562	0.750	0.611
ETTh2	96	0.312	0.349	0.338	0.372	0.325	0.363	0.330	0.361	0.415	0.426	0.401	0.421	0.409	0.420	0.390	0.424	0.442	0.456
	192	0.415	0.418	0.504	0.473	0.419	0.420	0.417	0.419	0.489	0.464	0.452	0.455	0.483	0.464	0.457	0.465	0.617	0.542
	336	0.456	0.453	0.528	0.495	0.899	0.652	0.823	0.639	0.511	0.485	0.464	0.469	0.499	0.479	0.477	0.483	1.424	0.849
	720	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Avg	0.394	0.407	0.457	0.447	0.548	0.478	0.523	0.473	0.472	0.458	0.439	0.448	0.463	0.454	0.441	0.457	0.827	0.615

Table 1: Full results of Few-shot learning task on 5% data with the input length  $L = 96$ . The prediction lengths  $T$  are set as {96, 192, 336, 720}. '-' means that 5% time series is not sufficient to constitute a training set.

LLMs for time series forecasting warrants further investigation.

## References

- Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; and Wen, Q. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*.
- Liu, P.; Guo, H.; Dai, T.; Li, N.; Bao, J.; Ren, X.; Jiang, Y.; and Xia, S.-T. 2025. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18915–18923.
- Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.
- Nie, Y.; H. Nguyen, N.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.
- Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are Transformers Effective for Time Series Forecasting?
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, 11106–11115. AAAI Press.
- Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*.
- Zhou, T.; Niu, P.; Wang, X.; Sun, L.; and Jin, R. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *NeurIPS*.