LLM-Mixer: Multiscale Mixing in LLMs for Time Series Forecasting

Yingxin Yu

December 8, 2025

Based on: [Kowsher et al., 2025]

Overview

1. Background and Challenges
2. Related Work
3. Method
4. Experiments
5. Ablation Study and Analysis
6. Limitations and Future Work

Background and Challenges

Background and Challenges

What is time series forecasting?

- A **time series** is data recorded over time (e.g., hourly demand, daily temperature).
- **Forecasting**: use the past to predict future values.
- Often **multivariate**: multiple variables evolve together. (e.g., electricity load with temperature, humidity)

Why time series forecasting matters

- Key for decision making in finance, energy, healthcare, climate ...

Background and Challenges

Why it is difficult in practice

- Real-world series are often **non-stationary**, **nonlinear**, and **multivariate**.
- Classic methods rely on strong assumptions (**stationarity / linearity**)
- Deep learning models may struggle with capturing **long-term dependencies**.
- We need to capture both **short-term fluctuations** and **long-term trends** in a model.

Given their strong generalization, recent studies have explored LLMs for time series forecasting.

Extra challenge for LLM-for-TS

- LLMs are trained for **discrete tokens**, while time series are **continuous**.
- Input context length of LLMs is a trade-off:
  - Short lookback may miss long-term trends
  - Long lookback increase computational cost and local details can be diluted.

  $\Rightarrow$ We need a good representation of the past to feed LLMs.

Paper Overview

Key Question:

- How can we present the past in a way that lets a frozen LLM understand time-series patterns better?

Core Idea:

- Introduce a new framework that forecasts time series by building **multiscale** views of the past, **decomposing & mixing** them, and feeding them into a frozen LLM.

Approach:

- **Pipeline**: multiscale inputs $\rightarrow$ PDM $\rightarrow$ prompt $\rightarrow$ frozen LLM $\rightarrow$ prediction head
- **Performance Evaluation**: long/short-term, uni/multi-variate; compare with recent SOTA
- **Ablation Analysis**: effect of downsampling and pooling methods

Related Work

Related Work

Classical forecasting (e.g., ARIMA, smoothing)

- Works well for simple tasks, but rely on strong assumptions e.g. **stationarity** & **linearity**

Deep learning for time series

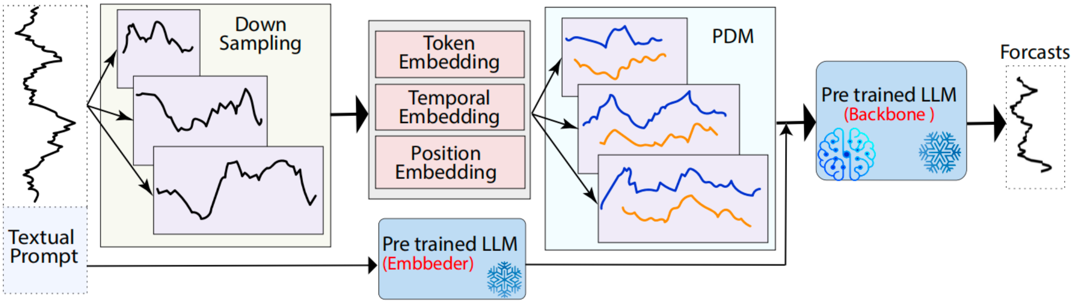
- CNNs/RNNs can model local patterns or temporal state transitions, but can struggle with **long-range dependencies**
- Transformers show strong capabilities in both local and long-range modeling

LLMs for time series

- Main challenge: handling **continuous, multi-scale** time series

Method

Method Overview: Pipeline



- **Multiscale views:** downsample X into multiple resolutions (fine $\rightarrow$ coarse).
- **PDM:** decompose (trend/seasonal) and mix information across scales.
- **Prompt embedding:** concat all scales and prepend a **text prompt**.
- **Forecast head:** feed into a frozen LLM; a **linear decoder** outputs the next K steps.

Preliminaries: Problem Formulation

Forecasting task

- Given a past window of a multivariate time series, predict the next K time steps.

Notation

- Past: $X = \{x_1, \ldots, x_T\} \in \mathbb{R}^{T \times M}$ (# of time steps T, # of features M)
- Future: $Y = \{x_{T+1}, \ldots, x_{T+K}\} \in \mathbb{R}^{K \times M}$ (i.e. predict the future K time steps)

LLM setting

- Use a text prompt P and a pre-trained LLM $\mathbb{F}$ with **frozen** parameters Θ.
- Train only specific parameters Φ (e.g., a linear decoder) for forecasting tasks

$$\hat{Y} = \mathbb{F}(X, P; \Theta, \Phi)$$

Key point: with a frozen LLM, performance depends on how we represent X before feeding it into the model (next: multiscale + PDM).

Step 1: Multiscale Downsampling

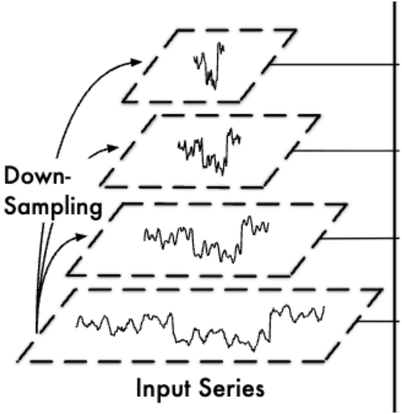
What it does

- Create τ downsampled versions using **average pooling**:

$$X = \{x_0, x_1, \ldots, x_\tau\}, \quad x_i \in \mathbb{R}^{\frac{T}{2^i} \times M}$$

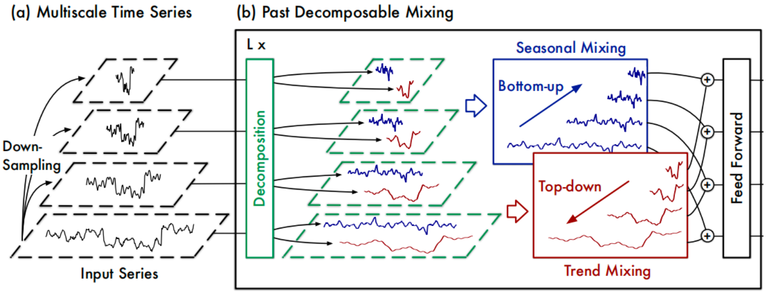
Why it helps

- Small i: keep **local details**; large i: expose **global trends**.
- Capture both local and global patterns **without** using an extremely long single window.



Source: [Wang et al., 2024]

Step 2: PDM (Past-Decomposable Mixing)

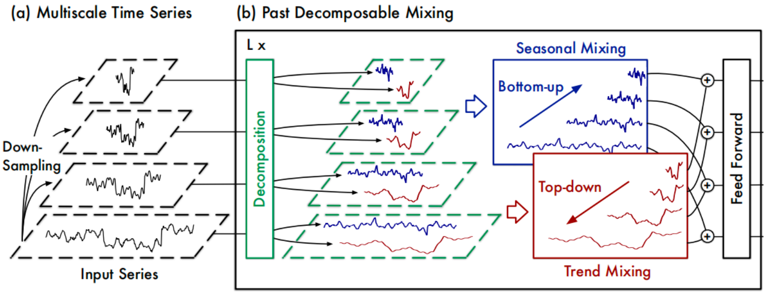


Source: [Wang et al., 2024]

What it does

- Input: multiscale past $\{x_0, \ldots, x_\tau\}$ (after downsampling)
- Decompose each scale x_i into **trend** t_i and **seasonal** s_i
- Mix information **across scales separately**: seasonal (bottom-up), trend (top-down)
- Output: updated multiscale past $X^L = \{x_0^L, \ldots, x_\tau^L\}$

Step 2: PDM (Past-Decomposable Mixing)



Source: [Wang et al., 2024]

Why it helps

- Fine scale: captures **local details**.
- Coarse scale: reveals **global trends**.
- PDM lets them interact **before** feeding to the frozen LLM.

Step 3: Prompt Design and Feeding to LLM

Prompt design

- Use a **dataset-level** text prompt P (features / structure / basic statistics).
- Embed the prompt with the LLM word embeddings, denoted by $E(P)$.

How it is used

- Feed (prompt + past) to the frozen LLM $\mathbb{F}$:

$$\mathbb{F}(E(P) \oplus \mathrm{Concat}(x_0^L, \ldots, x_\tau^L))$$

- Predict with a linear trainable decoder with parameters Φ.

Experiments

Experiments: Setup

Tasks & datasets

- **Long-term multivariate:** ETT (ETTh1/2, ETTm1/2), Weather, Electricity, Traffic
- **Short-term multivariate:** PeMS (PEMS03/04/07/08)
- **Univariate long sequence:** ETT benchmark

Backbones & metrics

- Frozen LLM backbone: **RoBERTa-base** (medium) and **LLaMA2-7B** (large)
- Metrics: **MSE** / **MAE** (lower is better)

Evaluation protocol

- Long-term horizons: {96, 192, 336, 720}; Short-term horizons: {12, 24, 48, 96}
- Compare with strong baselines (e.g., TimeMixer, iTransformer, TIME-LLM, PatchTST, DLinear, TimesNet, TiDE, ...)

Results: Long-term Multivariate (Table 1)

Methods		LLM-Mixer (llama2)		LLM-Mixer (roberta)		TIME-LLM		TimeMixer		iTransformer		RLinear		DLinear		PatchTST		TimesNet		TiDE		TimesNet		Crossformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.368**	**0.395**	0.372	0.399	0.369	0.397	0.375	0.400	0.386	0.405	0.386	0.395	0.397	0.412	0.460	0.447	0.384	0.402	0.479	0.464	0.384	0.402	0.423	0.448
	192	**0.406**	**0.417**	0.439	0.470	0.411	0.428	0.429	0.421	0.441	0.436	0.437	0.424	0.446	0.441	0.512	0.477	0.436	0.429	0.525	0.492	0.436	0.429	0.471	0.474
	336	0.446	**0.444**	0.458	0.467	**0.440**	0.447	0.484	0.458	0.487	0.458	0.479	0.446	0.489	0.467	0.546	0.496	0.638	0.469	0.565	0.515	0.491	0.469	0.570	0.546
	720	**0.461**	**0.475**	0.465	0.480	0.462	0.477	0.498	0.482	0.503	0.491	0.481	0.470	0.513	0.510	0.544	0.517	0.521	0.500	0.594	0.558	0.521	0.500	0.653	0.621
	Avg	**0.420**	**0.433**	0.434	0.454	0.421	0.437	0.447	0.440	0.454	0.447	0.446	0.434	0.461	0.457	0.516	0.484	0.495	0.450	0.541	0.507	0.458	0.450	0.529	0.522
ETTh2	96	**0.274**	**0.334**	0.284	0.347	0.278	0.338	0.289	0.341	0.297	0.349	0.288	0.338	0.340	0.394	0.308	0.355	0.340	0.374	0.400	0.440	0.340	0.374	0.745	0.584
	192	0.339	0.384	0.343	0.389	**0.338**	**0.384**	0.372	0.392	0.380	0.400	0.374	0.390	0.482	0.479	0.393	0.405	0.402	0.414	0.528	0.509	0.402	0.414	0.877	0.656
	336	0.380	**0.408**	**0.375**	0.409	0.389	0.411	0.386	0.414	0.428	0.432	0.415	0.426	0.591	0.541	0.427	0.436	0.452	0.452	0.643	0.571	0.452	0.452	1.043	0.731
	720	**0.390**	**0.431**	0.394	0.438	0.393	0.432	0.412	0.434	0.427	0.445	0.420	0.440	0.839	0.661	0.436	0.450	0.462	0.468	0.874	0.679	0.462	0.468	1.104	0.763
	Avg	**0.345**	**0.389**	0.349	0.395	0.349	0.391	0.364	0.395	0.383	0.407	0.374	0.398	0.563	0.519	0.391	0.411	0.414	0.427	0.611	0.550	0.414	0.427	0.942	0.684
ETTm1	96	0.294	0.346	0.304	0.348	**0.293**	**0.343**	0.320	0.357	0.334	0.368	0.355	0.376	0.346	0.374	0.352	0.374	0.338	0.375	0.364	0.387	0.338	0.375	0.404	0.426
	192	**0.348**	**0.367**	0.350	0.377	0.350	0.368	0.361	0.381	0.377	0.391	0.391	0.392	0.382	0.391	0.390	0.393	0.374	0.387	0.398	0.404	0.374	0.387	0.450	0.451
	336	0.387	0.392	0.395	0.409	**0.382**	**0.391**	0.390	0.404	0.426	0.420	0.424	0.415	0.415	0.415	0.421	0.414	0.410	0.411	0.428	0.425	0.410	0.411	0.532	0.515
	720	0.439	0.442	0.448	0.450	**0.443**	0.451	0.454	**0.441**	0.491	0.459	0.487	0.450	0.473	0.451	0.462	0.449	0.478	0.450	0.487	0.461	0.478	0.450	0.666	0.589
	Avg	**0.367**	**0.387**	0.374	0.396	0.367	0.388	0.381	0.395	0.407	0.410	0.414	0.407	0.404	0.408	0.406	0.407	0.400	0.406	0.419	0.419	0.400	0.406	0.513	0.495
ETTm2	96	**0.160**	**0.251**	0.160	0.253	0.160	0.251	0.175	0.252	0.180	0.264	0.182	0.265	0.193	0.293	0.183	0.270	0.187	0.267	0.207	0.305	0.187	0.267	0.287	0.366
	192	0.226	**0.290**	0.229	0.297	**0.220**	0.292	0.237	0.299	0.250	0.309	0.246	0.304	0.284	0.361	0.255	0.314	0.249	0.309	0.290	0.364	0.249	0.309	0.414	0.492
	336	**0.283**	0.339	0.299	0.346	0.284	**0.337**	0.298	0.340	0.311	0.348	0.307	0.342	0.382	0.429	0.309	0.347	0.321	0.351	0.377	0.422	0.321	0.351	0.597	0.542
	720	0.392	0.398	0.399	0.405	0.391	0.397	**0.391**	**0.396**	0.412	0.407	0.407	0.398	0.558	0.525	0.412	0.404	0.365	0.359	0.558	0.524	0.408	0.403	1.730	1.042
	Avg	0.265	0.320	0.272	0.325	**0.264**	**0.319**	0.275	0.323	0.288	0.332	0.286	0.327	0.354	0.402	0.290	0.334	0.291	0.333	0.358	0.404	0.291	0.333	0.757	0.610
Weather	96	0.149	0.202	0.151	0.203	**0.148**	**0.202**	0.163	0.209	0.174	0.214	0.192	0.232	0.195	0.252	0.186	0.227	0.172	0.220	0.202	0.261	0.172	0.220	0.195	0.271
	192	**0.197**	**0.239**	0.209	0.249	0.199	0.242	0.208	0.250	0.221	0.254	0.240	0.271	0.237	0.295	0.234	0.265	0.219	0.261	0.242	0.298	0.219	0.261	0.209	0.277
	336	0.270	0.282	0.310	0.281	0.262	**0.279**	0.251	0.287	0.278	0.296	0.292	0.307	0.282	0.331	0.284	0.301	**0.246**	0.337	0.287	0.335	0.280	0.306	0.273	0.332
	720	0.323	0.332	0.339	0.342	0.330	0.334	0.339	0.341	0.358	0.347	0.364	0.353	0.282	0.331	0.356	0.349	0.365	0.359	0.287	0.335	**0.280**	**0.306**	0.379	0.401
	Avg	**0.235**	**0.264**	0.252	0.269	0.235	0.264	0.240	0.271	0.258	0.278	0.272	0.291	0.265	0.315	0.265	0.285	0.251	0.294	0.271	0.320	0.259	0.287	0.264	0.320
Electricity	96	0.143	**0.233**	0.150	0.241	**0.142**	0.234	0.153	0.247	0.148	0.240	0.201	0.281	0.210	0.302	0.190	0.296	0.168	0.272	0.237	0.329	0.168	0.272	0.219	0.314
	192	**0.151**	0.242	0.166	0.259	0.152	**0.241**	0.166	0.256	0.162	0.253	0.201	0.283	0.210	0.305	0.199	0.304	0.184	0.322	0.236	0.330	0.184	0.289	0.231	0.322
	336	**0.178**	0.267	0.180	0.281	0.180	**0.263**	0.185	0.277	0.178	0.269	0.215	0.298	0.223	0.319	0.217	0.319	0.198	0.300	0.249	0.344	0.198	0.300	0.246	0.337
	720	**0.213**	**0.305**	0.221	0.311	0.218	0.308	0.225	0.310	0.225	0.317	0.257	0.331	0.258	0.350	0.258	0.352	0.220	0.320	0.284	0.373	0.220	0.320	0.280	0.363
	Avg	**0.171**	**0.253**	0.174	0.273	0.173	0.261	0.182	0.272	0.178	0.270	0.219	0.298	0.225	0.319	0.216	0.318	0.193	0.304	0.251	0.344	0.192	0.295	0.244	0.334
Traffic	96	**0.380**	**0.264**	0.394	0.274	0.382	0.268	0.462	0.285	0.395	0.268	0.649	0.389	0.650	0.396	0.526	0.347	0.593	0.321	0.805	0.493	0.593	0.321	0.644	0.429
	192	0.396	0.269	0.399	0.276	**0.394**	**0.267**	0.473	0.296	0.417	0.276	0.601	0.366	0.598	0.370	0.522	0.332	0.617	0.336	0.756	0.474	0.617	0.336	0.665	0.431
	336	**0.423**	**0.274**	0.439	0.280	0.425	0.281	0.498	0.296	0.433	0.283	0.609	0.369	0.605	0.373	0.517	0.334	0.629	0.336	0.762	0.477	0.629	0.336	0.674	0.420
	720	**0.458**	**0.296**	0.460	0.298	0.460	0.300	0.506	0.313	0.467	0.302	0.647	0.387	0.645	0.394	0.552	0.352	0.640	0.350	0.719	0.449	0.640	0.350	0.683	0.424
	Avg	**0.414**	**0.265**	0.433	0.282	0.415	0.279	0.484	0.297	0.428	0.282	0.626	0.378	0.625	0.383	0.529	0.341	0.620	0.336	0.760	0.473	0.620	0.336	0.667	0.426

Table 1: Full long-term multivariate forecasting results. **Red:** the best, Blue: the second best.

⇒ LLM-Mixer **consistently ranks top1/2** in avg. MSE/MAE across many datasets and horizons.

Results: Short-term Multivariate (Table 2)

Methods		LLM-Mixer (llama2)		LLM-Mixer (roberta)		TIME-LLM		TimeMixer		iTransformer		RLinear		PatchTST		Crossformer		TiDE		TimesNet		DLinear		SCINet	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	12	0.069	0.173	0.082	0.190	0.092	0.201	0.082	0.189	0.071	0.174	0.126	0.236	0.099	0.216	0.090	0.203	0.178	0.305	0.085	0.192	0.122	0.243	**0.066**	**0.172**
	24	0.090	0.200	0.092	0.201	0.095	0.207	0.090	0.199	0.093	0.201	0.246	0.334	0.142	0.259	0.121	0.240	0.257	0.371	0.118	0.223	0.201	0.317	**0.085**	**0.198**
	48	**0.123**	**0.232**	0.126	0.237	0.127	0.237	0.125	0.235	0.125	0.236	0.551	0.529	0.211	0.319	0.202	0.317	0.379	0.463	0.155	0.260	0.333	0.425	0.127	0.238
	96	0.165	**0.274**	0.166	0.276	0.165	0.274	0.167	0.275	0.164	0.275	1.057	0.787	0.269	0.370	0.262	0.367	0.490	0.539	**0.028**	0.317	0.457	0.515	0.178	0.287
	Avg	**0.112**	**0.220**	0.116	0.226	0.120	0.230	0.116	0.225	0.113	0.221	0.495	0.472	0.180	0.291	0.169	0.281	0.326	0.419	0.147	0.248	0.278	0.375	0.114	0.224
PEMS04	12	0.072	**0.177**	0.076	0.183	**0.071**	0.177	0.073	0.179	0.078	0.183	0.138	0.252	0.105	0.224	0.098	0.209	0.182	0.324	0.087	0.195	0.148	0.272	0.073	0.177
	24	0.095	0.201	0.105	0.211	0.106	0.214	0.097	0.205	0.108	0.205	0.258	0.387	0.150	0.266	0.135	0.250	0.309	0.454	0.099	0.217	0.225	0.367	**0.084**	**0.193**
	48	**0.099**	**0.216**	0.103	0.220	0.101	0.218	0.099	0.217	0.120	0.233	0.572	0.544	0.229	0.339	0.205	0.353	0.470	0.539	0.135	0.253	0.355	0.437	0.099	0.217
	96	0.120	**0.225**	0.130	0.232	0.121	0.225	0.121	0.225	0.150	0.267	1.159	0.947	0.309	0.520	0.299	0.467	0.656	0.637	**0.043**	0.317	0.550	0.541	0.129	0.227
	Avg	**0.097**	**0.205**	0.104	0.211	0.100	0.209	0.098	0.207	0.111	0.221	0.526	0.491	0.195	0.307	0.209	0.314	0.353	0.475	0.129	0.245	0.329	0.395	0.119	0.234
PEMS07	12	**0.065**	**0.165**	0.072	0.180	0.068	0.166	0.070	0.168	0.067	0.165	0.118	0.235	0.097	0.226	0.093	0.209	0.155	0.324	0.081	0.185	0.118	0.272	0.068	0.174
	24	**0.087**	**0.105**	0.091	0.198	0.088	0.192	0.087	0.105	0.088	0.190	0.271	0.449	0.153	0.276	0.138	0.251	0.338	0.475	0.096	0.223	0.207	0.381	0.087	0.180
	48	**0.106**	**0.215**	0.117	0.224	0.109	0.219	0.106	0.217	0.110	0.215	0.596	0.621	0.343	0.459	0.309	0.401	0.532	0.547	0.145	0.264	0.593	0.484	0.149	0.233
	96	0.147	0.266	0.146	0.265	0.150	0.269	0.151	0.269	**0.141**	**0.245**	1.096	0.795	0.346	0.490	0.329	0.443	0.674	0.650	0.203	0.307	0.789	0.531	0.191	0.267
	Avg	**0.101**	**0.204**	0.107	0.217	0.104	0.212	0.104	0.218	0.101	0.204	0.504	0.478	0.213	0.303	0.215	0.326	0.355	0.499	0.129	0.245	0.529	0.387	0.119	0.234
PEMS08	12	0.082	0.186	0.086	0.190	0.080	0.184	0.083	0.183	**0.079**	**0.182**	0.133	0.247	0.168	0.232	0.152	0.267	0.215	0.367	0.154	0.276	0.172	0.291	0.087	0.184
	24	0.107	0.213	0.109	0.217	0.105	0.208	0.109	0.218	0.105	0.209	0.242	0.360	0.189	0.321	0.174	0.314	0.258	0.430	0.178	0.307	0.290	0.346	**0.104**	**0.193**
	48	0.187	0.235	0.192	0.240	0.193	0.242	0.192	0.241	0.125	0.237	0.596	0.556	0.269	0.389	0.247	0.388	0.433	0.512	0.210	0.345	0.418	0.422	**0.124**	**0.216**
	96	**0.140**	**0.245**	0.142	0.249	0.147	0.256	0.143	0.252	0.167	0.275	1.043	0.841	0.344	0.470	0.326	0.459	0.534	0.571	0.283	0.387	0.593	0.514	0.155	0.253
	Avg	0.119	0.226	0.132	0.224	0.131	0.223	0.132	0.224	0.119	0.226	0.503	0.501	0.243	0.353	0.225	0.357	0.360	0.470	0.206	0.329	0.368	0.393	**0.118**	**0.212**

Table 2: Full short-term multivariate forecasting results. **Red:** the best, Blue: the second best.

⇒ LLM-Mixer achieves **low MSE/MAE** specifically on PeMS03, PeMS04 and PeMS07.

⇒ Competitive against strong baselines (e.g., TIME-LLM, TimeMixer, PatchTST).

Results: Univariate Long Sequence (Table 3)

Methods		LLM-Mixer (llama2)		LLM-Mixer (Roberta)		Linear		NLinear		DLinear		FEDformer-f		FEDformer-w		Autoformer		Informer		LogTrans	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.052**	**0.175**	0.053	0.177	0.189	0.359	0.055	0.176	0.056	0.180	0.079	0.215	0.080	0.214	0.071	0.206	0.193	0.377	0.283	0.468
	192	**0.064**	**0.200**	0.066	0.203	0.078	0.212	0.069	0.204	0.071	0.204	0.104	0.245	0.105	0.256	0.114	0.262	0.217	0.395	0.234	0.409
	336	**0.080**	**0.226**	0.081	0.226	0.091	0.237	0.084	0.228	0.098	0.244	0.119	0.270	0.120	0.269	0.107	0.258	0.202	0.381	0.386	0.546
	720	**0.075**	**0.222**	0.078	0.223	0.172	0.340	0.080	0.226	0.189	0.359	0.142	0.299	0.127	0.280	0.126	0.283	0.183	0.355	0.475	0.629
	Avg	**0.068**	**0.206**	0.071	0.207	0.133	0.287	0.071	0.208	0.104	0.247	0.111	0.257	0.108	0.255	0.106	0.252	0.199	0.377	0.345	0.513
ETTh2	96	0.125	0.274	**0.123**	0.276	0.133	0.283	0.129	0.278	0.131	0.279	0.128	**0.271**	0.156	0.306	0.153	0.306	0.213	0.373	0.217	0.379
	192	**0.166**	0.322	0.169	0.324	0.176	**0.330**	0.169	0.324	0.176	0.329	0.185	0.330	0.238	0.380	0.204	0.351	0.227	0.387	0.281	0.429
	336	**0.193**	**0.353**	0.194	0.356	0.213	0.371	0.194	0.355	0.209	0.367	0.231	0.378	0.271	0.412	0.246	0.389	0.242	0.401	0.293	0.437
	720	0.222	**0.380**	0.225	0.381	0.292	0.440	0.225	0.381	0.276	0.426	0.278	0.420	0.288	0.438	0.268	0.409	0.291	0.439	**0.218**	0.387
	Avg	**0.177**	**0.332**	0.178	0.334	0.204	0.356	0.179	0.335	0.198	0.350	0.205	0.350	0.238	0.384	0.218	0.364	0.243	0.400	0.252	0.408
ETTm1	96	**0.023**	**0.118**	0.026	0.125	0.028	0.125	0.026	0.122	0.028	0.123	0.033	0.140	0.036	0.149	0.056	0.183	0.109	0.277	0.049	0.171
	192	**0.033**	**0.145**	0.036	0.147	0.043	0.154	0.039	0.149	0.045	0.156	0.058	0.186	0.069	0.206	0.081	0.216	0.151	0.310	0.157	0.317
	336	0.053	**0.172**	0.054	0.176	0.059	0.180	**0.052**	0.172	0.061	0.182	0.084	0.231	0.071	0.209	0.076	0.218	0.427	0.591	0.289	0.459
	720	**0.071**	0.205	0.072	**0.204**	0.080	0.211	0.073	0.207	0.080	0.210	0.102	0.250	0.105	0.248	0.110	0.267	0.438	0.586	0.430	0.579
	Avg	**0.045**	**0.161**	0.047	0.163	0.053	0.167	0.048	0.163	0.054	0.168	0.069	0.202	0.070	0.203	0.081	0.221	0.281	0.441	0.231	0.381
ETTm2	96	**0.062**	**0.180**	0.064	0.181	0.066	0.189	0.063	0.182	0.063	0.183	0.067	0.198	0.063	0.189	0.065	0.189	0.088	0.225	0.075	0.208
	192	0.090	0.222	**0.089**	**0.220**	0.094	0.230	0.090	0.223	0.092	0.227	0.102	0.245	0.110	0.252	0.118	0.256	0.132	0.283	0.129	0.275
	336	**0.114**	**0.255**	0.116	0.257	0.120	0.263	0.117	0.259	0.119	0.261	0.130	0.279	0.147	0.301	0.154	0.305	0.180	0.336	0.154	0.302
	720	0.169	**0.313**	0.171	0.314	0.175	0.320	0.170	0.318	0.175	0.320	0.178	0.325	0.219	0.368	0.182	0.335	0.300	0.435	**0.160**	0.321
	Avg	**0.109**	**0.243**	0.110	0.243	0.114	0.250	0.110	0.246	0.112	0.248	0.119	0.262	0.135	0.279	0.130	0.271	0.150	0.295	0.130	0.277

Table 3: Full univariate long sequence time-series forecasting results on ETT full benchmark. **Red:** the best, Blue: the second best.

⇒ LLM-Mixer achieves the **best avg. MSE/MAE** across datasets, ranking 1/2 on most horizons.

Ablation Study and Analysis

Ablation Study: What and Why

Two design choices

- **Downsampling levels** τ: how many temporal scales to create.
- **Pooling rule** for multi-scale mixing: Min / Max / Avg / L2.

Questions

- **Downsampling:** Does changing τ change **how the model learns** (i.e. training behavior)?
- **Pooling:** Which pooling rule gives lower **forecast error (MSE)**?

Effect of Downsampling: Why Use NTK?

Problem: measuring "training behavior" is hard

- Final MSE/MAE tells us **what** happened, not **how** training behaved.
- We need a metrics to compare training behavior across different τ.

Neural Tangent Kernel (NTK):

- For a model $f(x;\theta)$, NTK quantifies the similarity between how sensitive $f(x)$ and $f(x')$ are to parameter changes:
$$K(x, x') = \nabla_\theta f(x;\theta)^\top \nabla_\theta f(x';\theta)$$
- Intuitively, if two inputs have a **large** $K(x, x')$, they are learned/updated in a **similar way**.

$\Rightarrow$ So NTK summarizes the model's training behavior around its current parameters.

Effect of Downsampling: NTK Distance

Goal

- Compare how training behavior changes when we change τ.

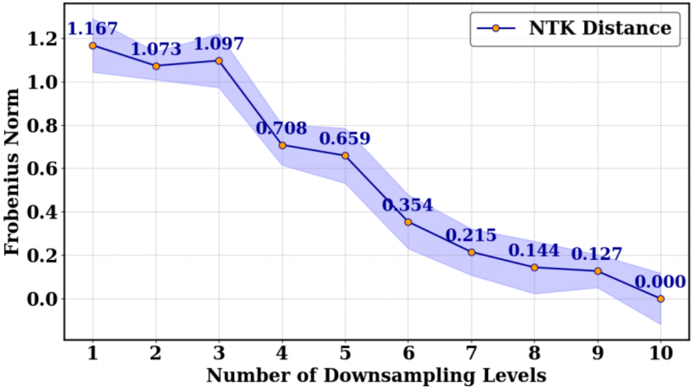
How the paper measures it

- Train 10 models with $\tau \in \{1, \ldots, 10\}$ on a synthetic multivariate dataset.
- Compute NTK matrices K_τ on 300 train/test sample pairs.
- Use $\tau = 10$ as a reference and define

$$d_{\mathrm{NTK}}(\tau) = \|K_{10} - K_\tau\|_F$$

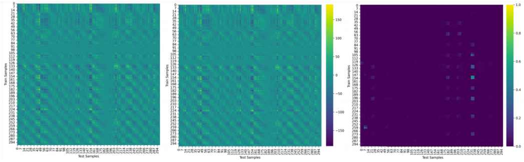
- Interpretation:
  - **Small** distance: learning behavior **similar to** the reference.
  - **Large** distance: the change of τ **significantly changes** the learning behavior.

Results: NTK distance vs. downsampling levels τ

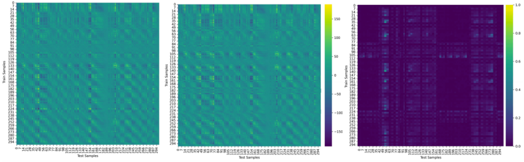


- As τ decreases, $d_{\mathrm{NTK}}(\tau)$ increases.
- $\tau = 1$ has the largest distance $\Rightarrow$ one-scale input changes training behavior the most.
- Downsampling is not always better: large τ may oversmooth local details.

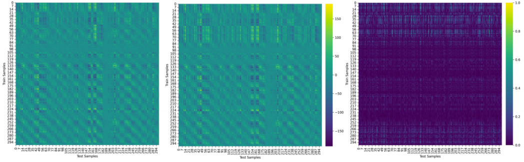
Results: NTK Visualization



(a) $\tau = 9$



(b) $\tau = 7$



(c) $\tau = 5$

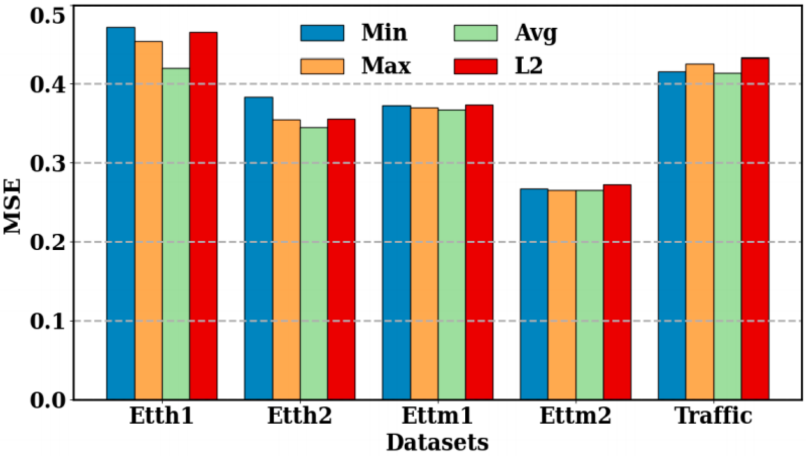
How to read the figure

- Each row compares a smaller $\tau = 5/7/9$ vs. the reference $(\tau = 10)$.
- Left: K_{ref} $(\tau = 10)$; middle: K_τ ; right: normalized $|K_\tau - K_{\text{ref}}|$.

What it suggests

- Smaller τ (fewer scales) $\Rightarrow$ training behavior deviates more from the reference.

Results: Pooling Rule



- **Average pooling** yields consistently lower MSE across datasets ⇒ used as default.

Limitations and Future Work

Limitations and Future Work

Practical limitations

- **High compute cost:** using a large frozen LLM can be expensive.
- **Prompt quality:** performance may rely on prompt quality and domain knowledge.

Technical limitations

- **OOD robustness:** a fixed prompt may not work well when test distribution changes.
- **Missing classical baselines:** comparisons focus on deep models; ARIMA/smoothing are not reported.
- **Leakage risk:** prompt-based setups require careful checks to avoid using test information.

Future directions: e.g., domain generalization, adaptive prompting and efficiency optimizations

References

Kowsher, M., Sobuj, M. S. I., Prottasha, N. J., Alanis, E. A., Garibay, O., and Yousefi, N. (2025).
Llm-mixer: Multiscale mixing in llms for time series forecasting.
In *Proceedings of the 4th Table Representation Learning Workshop*, pages 156–165.

Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J. Y., and Zhou, J. (2024).
Timemixer: Decomposable multiscale mixing for time series forecasting.
arXiv preprint arXiv:2405.14616.

Thank you!

Q&A

Appendix

Appendix: 3-steps of PDM block

Step 1: Decompose each scale (trend + seasonal)

- Moving average (low-pass) gives trend:

$$t_i = \mathrm{MA}(x_i), \qquad s_i = x_i - t_i, \qquad i = 0, \ldots, \tau$$

Step 2: Mix across scales (message passing)

- Seasonal mixing (bottom-up, fine $\rightarrow$ coarse):

$$\tilde{s}_0 = s_0, \qquad \tilde{s}_i = s_i + \mathrm{BU}_{i-1\rightarrow i}\big(\tilde{s}_{i-1}\big), \quad i = 1, \ldots, \tau$$

- Trend mixing (top-down, coarse $\rightarrow$ fine):

$$\tilde{t}_\tau = t_\tau, \qquad \tilde{t}_i = t_i + \mathrm{TD}_{i+1\rightarrow i}\big(\tilde{t}_{i+1}\big), \quad i = \tau - 1, \ldots, 0$$

- BU/TD: learned projections to match lengths (e.g., $T/2^{i-1} \rightarrow T/2^i$).

Step 3: Merge and update (per scale)

- Recombine and apply FFN + residual (same scale):

$$x_i \leftarrow x_i + \mathrm{FFN}\big(\tilde{s}_i + \tilde{t}_i\big), \qquad i = 0, \ldots, \tau$$

Appendix: NTK — Definition and Intuition

Setup

- Model: $f(x;\theta) \in \mathbb{R}$, parameters $\theta \in \mathbb{R}^p$.
- **Tangent / sensitivity vector:** $g(x) := \nabla_\theta f(x;\theta) \in \mathbb{R}^p$.

Neural Tangent Kernel (NTK)

$$K(x,x') := g(x)^\top g(x') = \nabla_\theta f(x;\theta)^\top \nabla_\theta f(x';\theta).$$

Intuition (why it measures "similar learning")

- A small parameter change $\Delta\theta$ changes the output by

$$\Delta f(x) \approx g(x)^\top \Delta\theta.$$

- If $K(x,x')$ is large, then $g(x)$ and $g(x')$ are aligned $\Rightarrow$ updates that change $f(x')$ tend to change $f(x)$ in a similar way.

Appendix: NTK — Training Dynamics (Squared Loss)

Squared loss on training set $\{(x_i, y_i)\}_{i=1}^n$

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_{i=1}^{n}\big(f(x_i;\theta) - y_i\big)^2, \qquad e_i := f(x_i;\theta) - y_i.$$

One GD step

$$\Delta\theta = -\eta\nabla_\theta\mathcal{L} = -\eta\sum_{i=1}^{n} e_i\, g(x_i).$$

Effect on any input x **(local linearization)**

$$\Delta f(x) \approx g(x)^\top \Delta\theta = -\eta\sum_{i=1}^{n} e_i\, g(x)^\top g(x_i) = -\eta\sum_{i=1}^{n} e_i\, K(x, x_i).$$

Vector form on training points Let $f \in \mathbb{R}^n$ with $f_i = f(x_i;\theta)$, $e = f - y$, and $[K]_{ij} = K(x_i, x_j)$:

$$f^{(t+1)} \approx f^{(t)} - \eta\, K\, e^{(t)}.$$

Takeaway: NTK directly controls how errors at x_j drive updates at x_i (coupling strength).