Are Language Models Actually Useful for Time Series Forecasting?

Paper

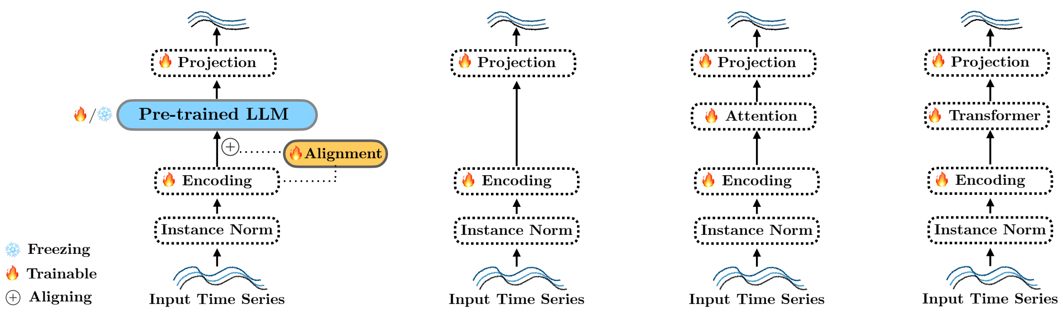
Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, Thomas Hartvigsen

University of Virginia

behavioral data science

Unfortunately, Not Yet!

RQ1:Removing LLM`s Backbone from Forecaster?



(a) w/ LLM (b) w/o LLM (c) LLM2Attn (d) LLM2Trsf

1. Forecast Performance Not Degraded, Even Improved

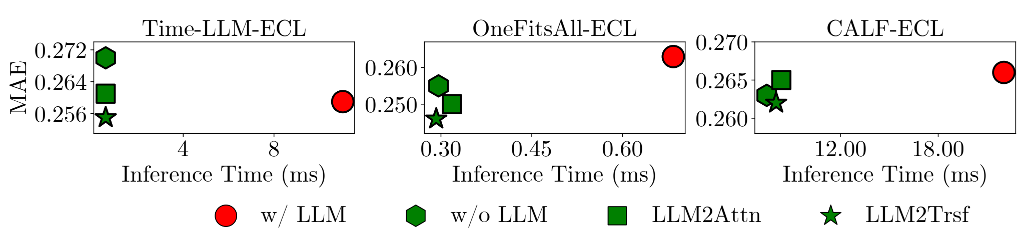
	Time-LLM	w/o LLM	LLM2Attn	LLM2Trsf
# Wins	0	12	2	12
Parameters	6651.82M	0.55M	0.55M	0.66M

	CALF	w/o LLM	LLM2Attn	LLM2Trsf
# Wins	4	7	4	11
Parameters	180.25M	8.17M	10.5M	13.68M

	OneFitsAll	w/o LLM	LLM2Attn	LLM2Trsf
# Wins	7	11	3	5
Parameters	91.36M	9.38M	10.71M	13.54M

We evaluated on **eight** commonly used time series forecasting datasets, such as ETTh, Weather, and **five** other datasets, including NN5 and FRED-MD.

2. No Gains but Significantly Increased Inference Cost



3. Substantial Training Costs Increase (A100 GPU, Weather)

Method	Time-LLM (LLaMA) # Param (M)	Time-LLM (LLaMA) Time (min)	OneFitsAll (GPT-2) # Param (M)	OneFitsAll (GPT-2) Time (min)	CALF (GPT-2) # Param (M)	CALF (GPT-2) Time (min)
w/ LLM	6642	3003	86	152	180	12
w/o LLM	0.198	1.91	4	16	8	2.32
LLM2Attn	0.202	2.22	7	21	10	2.14
LLM2Trsf	0.336	2.38	10	24	13	1.89

RQ2:Training a LLM from Scratch?

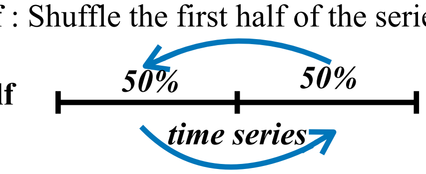
Methods	Pre+FT (GPT-2) MAE	Pre+FT (GPT-2) MSE	woPre+FT MAE	woPre+FT MSE	Pre+woFT MAE	Pre+woFT MSE	woPre+woFT MAE	woPre+woFT MSE
ETTh1	0.4312	**0.4313**	0.4284	0.4362	**0.4267**	0.4342	0.4365	0.4474
ETTh2	0.3838	0.3510	0.3839	**0.3508**	**0.3830**	0.3514	0.3872	0.3554
ETTm1	0.3910	0.3963	0.3933	0.4013	**0.3898**	**0.3954**	0.3949	0.4028
ETTm2	0.3230	0.2831	**0.3221**	0.2852	0.3221	**0.2827**	0.3224	0.2829
Illness	0.8691	1.6996	**0.8523**	**1.6146**	0.8742	1.6640	0.8663	1.6381
Weather	**0.2737**	**0.2510**	0.2760	0.2520	0.2771	0.2535	0.2776	0.2582
Traffic	0.2844	0.4438	**0.2771**	**0.4409**	0.2820	0.4446	0.2863	0.4483
Electricity	0.2660	0.1758	**0.2597**	**0.1669**	0.2635	0.1730	0.2663	0.1784
# Wins:	3		8		5		0	

woPre+FT: GPT-2 and training it from scratch yielded better performance than pre-trained model.
Pre+woFT: The frozen pre-trained GPT-2, when used as a projector, can be fitted by other MLP layers.
woPre+woFT: The frozen GPT-2, serving as a random projector, was fitted by MLPs, showing a certain level of capability.

RQ3: LLMs have TS sequential dependencies?

Dataset / Input Ablation	ETTh1 Sf-all.	Sf-half.	Ex-half	Masking
Time-LLM	51.8%	5.6%	79.6%	32.5%
w/o LLM	56.0%	4.5%	89.7%	39.5%
LLM2Attn	53.8%	3.3%	92.2%	33.8%
LLM2Trsf	50.3%	3.4%	89.2%	34.8%
OneFitsAll	62.1%	6.1%	16.6%	31.3%
w/o LLM	58.6%	6.1%	19.2%	36.1%
LLM2Attn	68.5%	9.0%	15.0%	34.4%
LLM2Trsf	58.0%	7.8%	12.6%	30.2%
CALF	50.5%	9.6%	5.6%	8.5%
w/o LLM	56.2%	12.1%	6.1%	10.4%
LLM2Attn	51.9%	10.8%	5.8%	7.3%
LLM2Trsf	50.3%	8.5%	5.5%	7.0%

Sf-All : Shuffle the whole time series.
Sf-Half : Shuffle the first half of the series.
Ex-Half



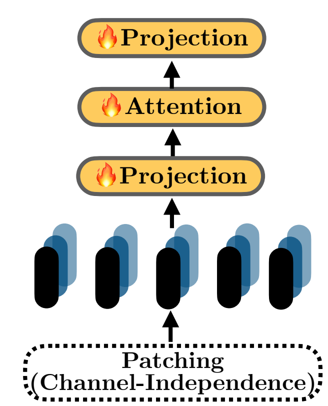
LLM-based methods show similar resilience to shuffling as their simpler counterparts.

RQ4:LLMs help with few-shot learning?

Model	GPT-2	w/o LLM	LLM2Attn	LLM2Trsf
#Wins	2	10	0	2
Model	LLaMA	w/o LLM	LLM2Attn	LLM2Trsf
#Wins	8	7	0	1

To evaluate whether this is the case we trained models and their ablations on **10%** of each dataset. The results indicate that our ablations can perform better than LLMs in **few-shot** scenarios.

RQ5:Simple models perform similarly with LLM

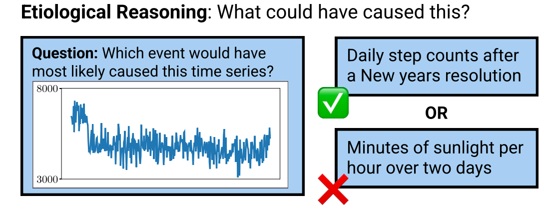


Simple Model : PAttn

Encoder plays a significant role in time series analysis. By using **Patch+Attention** for encoding and connecting it to a Linear layer, performance comparable to most LLM-based time series forecasters (Note that the CALF encoder performs better on larger time series datasets.)

Our goal is not to imply that language models will never be useful for time series.

Etiological Reasoning: What could have caused this?



LLMs hold significant potential in **reasoning** about time series with context, although their current performance still struggles in zero-shot scenarios. (Merrill et al., 2024)

Context-Aided Forecasting: What will happen if...?