Time-LLM Experiments' Results and Analysis

I. Results

A. Long-term forecasting

Model	ETTh1 96	
	MSE	MAE
GPT2-small Time-LLM	0.387	0.412
LLama7b Time-LLM	0.362	0.392
GPT4TS	0.376	0.397
DLinear	0.375	0.399
PatchTST	0.370	0.399
TimesNet	0.384	0.402
FEDformer	0.376	0.419
Autoformer	0.449	0.459
Stationary	0.513	0.491
ETSformer	0.494	0.479
LightTS	0.424	0.432
Informer	0.865	0.713
Reformer	0.837	0.728
N-BEATS	0.496	0.475
N-HiTS	0.392	0.407
AutoARIMA	0.933	0.635
AutoTheta	1.266	0.758
AutoETS	1.264	0.756

1) MSE (Mean Squared Error): It is the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. The smaller the MSE, the higher the model's prediction accuracy.

$$\text{MSE} = \frac{1}{H}\sum_{h=1}^{H}\left(Y_h - \hat{Y}_h\right)^2$$

2) MAE (Mean Absolute Error): It is the average of the absolute differences between prediction and actual observation. Like MSE, the smaller the MAE, the higher the model's prediction accuracy.

$$\text{MAE} = \frac{1}{H}\sum_{h=1}^{H}\left|Y_h - \hat{Y}_h\right|$$

B. Short-term forecasting

Model	M4-Yearly 6		
	SMAPE	MASE	OWA
GPT2-small Time-LLM	13.555	3.038	0.797
LLama7b Time-LLM	13.419	3.005	0.789
GPT4TS	15.11	3.565	0.911
TimesNet	15.378	3.554	0.918
PatchTST	13.477	3.019	0.792
N-HiTS	13.422	3.056	0.795
ETSformer	13.487	3.036	0.795
N-BEATS	18.009	4.487	1.115
LightTS	14.247	3.109	0.827
DLinear	16.965	4.283	1.058
FEDformer	14.021	3.036	0.811
Stationary	13.717	3.078	0.807
Autoformer	13.974	3.134	0.822
Informer	14.727	3.418	0.881
Reformer	16.169	3.800	0.973

1) SMAPE (Symmetric Mean Absolute Percentage Error): It measures the percentage error between the forecast and the actual value without assigning a greater penalty to over- or under-prediction. The smaller the SMAPE, the better the model's performance.

$$\text{SMAPE} = \frac{200}{H}\sum_{h=1}^{H}\frac{|Y_h - \hat{Y}_h|}{|Y_h| + |\hat{Y}_h|}$$

2) MASE (Mean Absolute Scaled Error): It is a measure of forecast accuracy that adjusts for the variability of the time series being forecast. The smaller the MASE, the better the model's performance.

$$\text{MASE} = \frac{1}{H}\sum_{h=1}^{H}\frac{|Y_h - \hat{Y}_h|}{\frac{1}{H-s}\sum_{j=s+1}^{H}|Y_j - Y_{j-s}|}$$

3) OWA (Overall Weighted Average): It is a comprehensive metric that considers multiple evaluation metrics, often used in competitions or multi-objective optimization. The calculation method of OWA may vary depending on the application scenario.

$$\text{OWA} = \frac{1}{2}\left[\frac{\text{SMAPE}}{\text{SMAPE}_{\text{Naive2}}} + \frac{\text{MASE}}{\text{MASE}_{\text{Naive2}}}\right]$$

C. Few-shot forecasting

Model	ETTh1 96 10%	
	MSE	MAE
GPT2-small Time-LLM	0.534	0.484
LLama7b Time-LLM	0.448	0.460
GPT4TS	0.458	0.456
DLinear	0.492	0.495
PatchTST	0.516	0.485
TimesNet	0.861	0.628
FEDformer	0.512	0.499
Autoformer	0.613	0.552
Stationary	0.918	0.639
ETSformer	1.112	0.806
LightTS	1.298	0.838
Informer	1.179	0.792
Reformer	1.184	0.790

D. Zero-shot forecasting

backbone model	ETTh1 → ETTh2 96	
	MSE	MAE
GPT2-small Time-LLM	0.266	0.334
LLama7b Time-LLM	0.279	0.337
LLMTime	0.510	0.576
GPT4TS	0.335	0.374
DLinear	0.347	0.400
PatchTST	0.304	0.350
TimesNet	0.358	0.387
Autoformer	0.469	0.486

II. Analysis

The reasons for the differences from the paper's results are:

1) The main reason is that the paper used the LLAMA pre-trained model, which has significantly more model parameters and a higher model dimension than GPT2.
2) During the training process of the paper, the ZeRO-2 optimizer under the DeepSpeed framework was utilized. This is a technology optimized and accelerated for large-scale training. In this process, gradient accumulation was used to simulate larger batch sizes, but it also introduced randomness. Secondly, the paper adopted mixed-precision training, which can significantly reduce memory usage and speed up training. However, using lower precision in floating-point representation also introduces additional numerical computation errors, which is another source of result randomness. Our training process was conducted on a single GPU, which also causes some deviation in the results.
3) Different CUDA versions can also affect the model results.