Can a small language model forecast?

We reproduced LLMTime on a CPU using Qwen 0.5B. For each series we held out the last 96 steps. The evaluation covered all 21 weather series and 20 of the 321 electricity series.

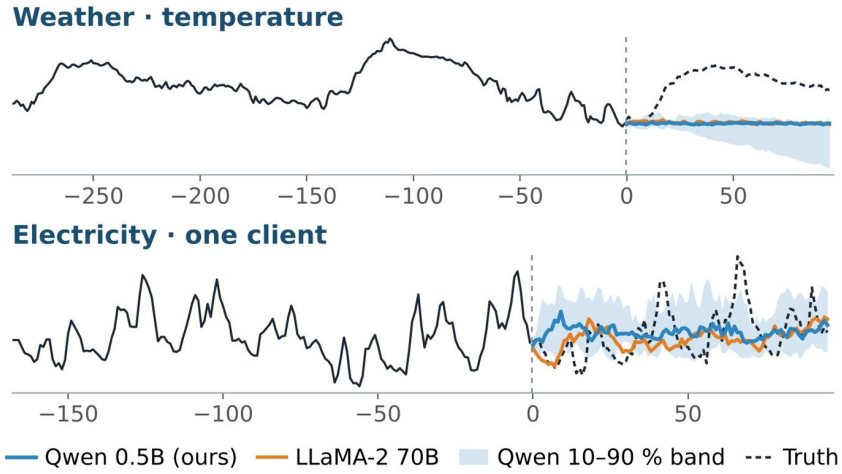
Forecast error

Lower mean absolute error (MAE) is better. We used a 2,048-token context. For LLaMA-2 70B, we used the predictions published by the paper's authors.

Method	Weather MAE	Electricity MAE
Qwen 0.5B, our run	0.195	0.790
LLaMA-2 70B, published predictions	0.214	0.539
Repeat the last value	0.240	1.024
Repeat the previous day[1]	0.15	0.27

[1] Seasonal-baseline scores are rounded to two decimals in the original comparison.

Two example series



One trace from each dataset. Blue shows Qwen, orange the published LLaMA-2 forecast and the dotted line the observations. The shaded band spans the 10th–90th percentiles of Qwen predictions.

Qwen beat the last-value baseline on both datasets. Repeating the previous day gave lower error than Qwen on both.

Paper reproduced: Gruver, Finzi, Qiu and Wilson, "Large Language Models Are Zero-Shot Time Series Forecasters," NeurIPS 2023.