LLaMA-2 Extensions LLaMA-2 models are recently-released, strong LLM base models. The largest model, 70B, has reported performance on par with GPT3.5 (ChatGPT). In Figure 1, we compare LLaMA-2 70B with time series transformers. In Figure 2, we extend our real-world evaluations, and in Figure 3 we extend our analysis of scaling and the effects of model alignment.

Autoformer Datasets and Models The Autoformer repo contains many real-world multivariate time series datasets and baseline implementations of Reformer, Informer, Autoformer, and vanilla Transformers. FEDformer extends the Autoformer repo with an implementation of the FEDFormer method. We run all of these methods on the datasets shown in Figure 1. The smallest of these datasets has 7 covariates while the largest has 862. We use the last 96 or 192 timesteps of each dataset as the test set (depending on the prediction length) and split the remaining timesteps into the train and validation. We then run the code provided in the repo. For LLaMA-2 70B, the prediction of each variable in the multivariate time series is based solely on the past values of that variable and not other covariates.

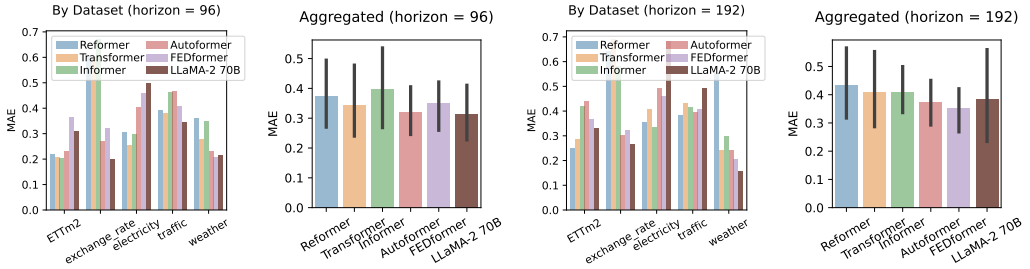

Figure 1: Zero-shot LLaMA-2 70B is competitive with transformer variants designed for time series data. LLaMA numbers are obtained by breaking the multivariate series into individual channels and forecasting the median of 20 samples. All sampling is performing with temperature 1.0 and nucleus 0.9, and no other hyperparemeter tuning. While the baselines use multiple covariates as inputs, we limit the LLaMA input to the history of the given evaluated trajectory and omit the other channels. Error bars show two standard deviations.

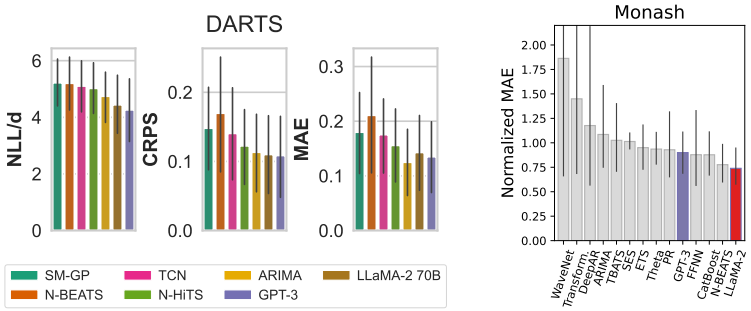

Figure 2: We update our DARTS and Monash numbers to include LLAMA-2 70B. On DARTS LLaMA-2 performs slightly worse than GPT-3 but better than almost all baselines, and on Monash LLAMA-2 70B is the best method in aggregate performance. Error bars show standard errors.

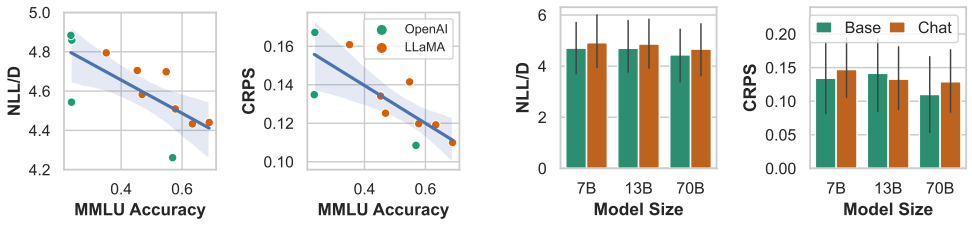

Figure 3: Scaling of OpenAI (GPT) and LLaMA-2 family models (**left**) and comparisons of base and aligned LLaMA-2 models (**right**). Results mirror those in the original submission: forecasting improves with overall reasoning ability and aligned models (e.g. ChatGPT & GPT4) perform worse than base model. Error bars show standard errors.