Timer

Timer: Generative Pre-trained Transformers Are Large Time Series Models

Yong Liu [*1] **Haoran Zhang** [*1] **Chenyu Li** [*1] **Xiangdong Huang** [1] **Jianmin Wang** [1] **Mingsheng Long** [1]

Yong Liu

Haoran Zhang

Chenyu Li

Xiangdon Huang

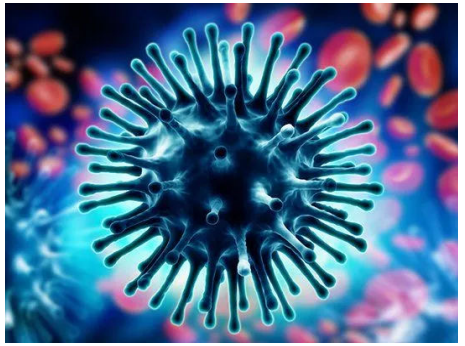
Jianmin Wang

Mingsheng Long

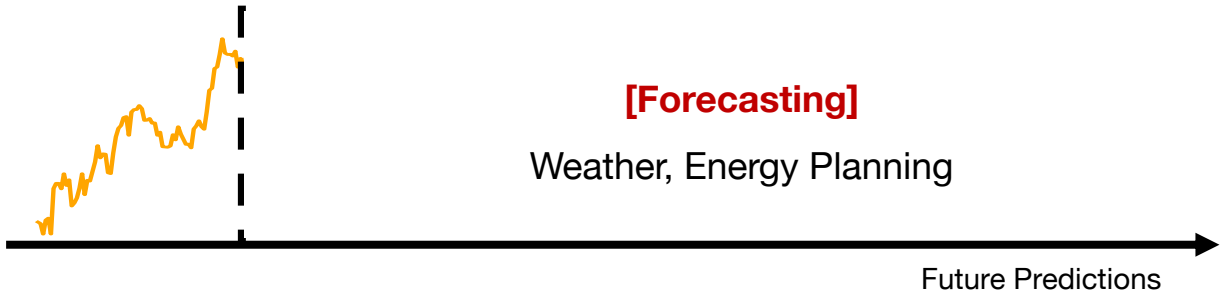
Time Series Applications

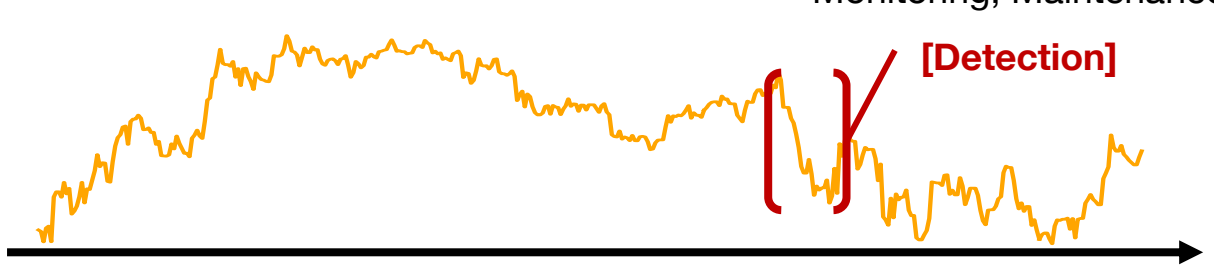


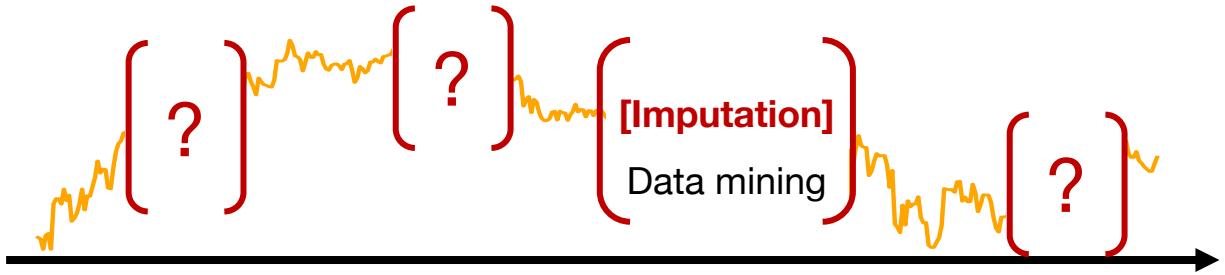


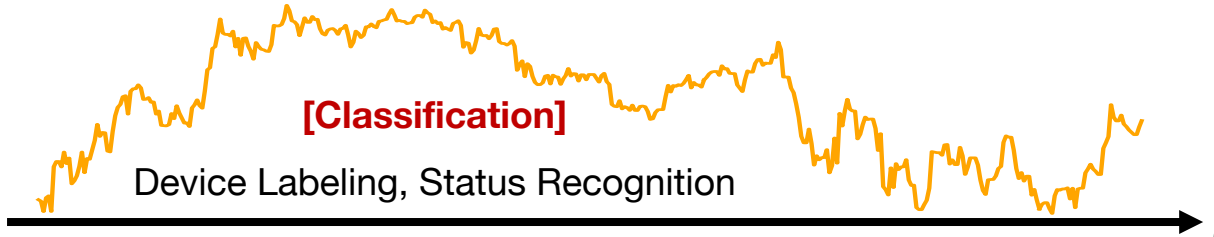


Time Series Analysis is Ubiquitous in Real World

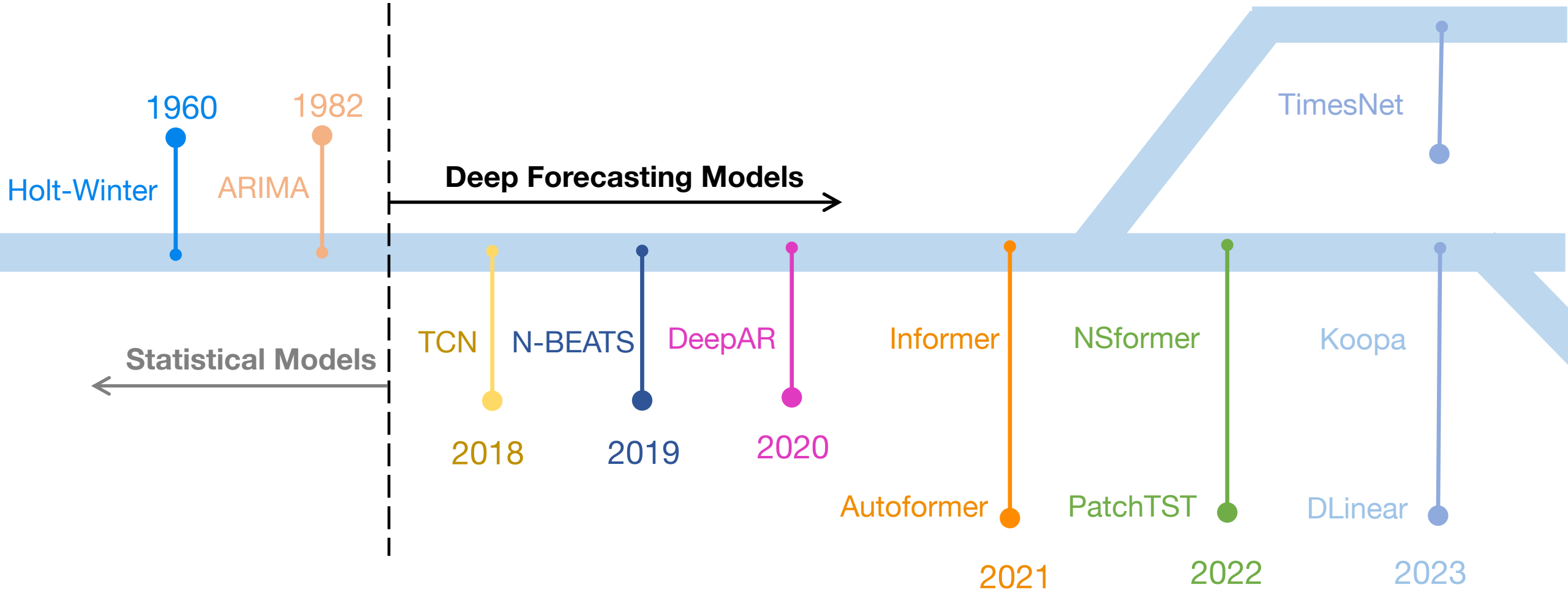




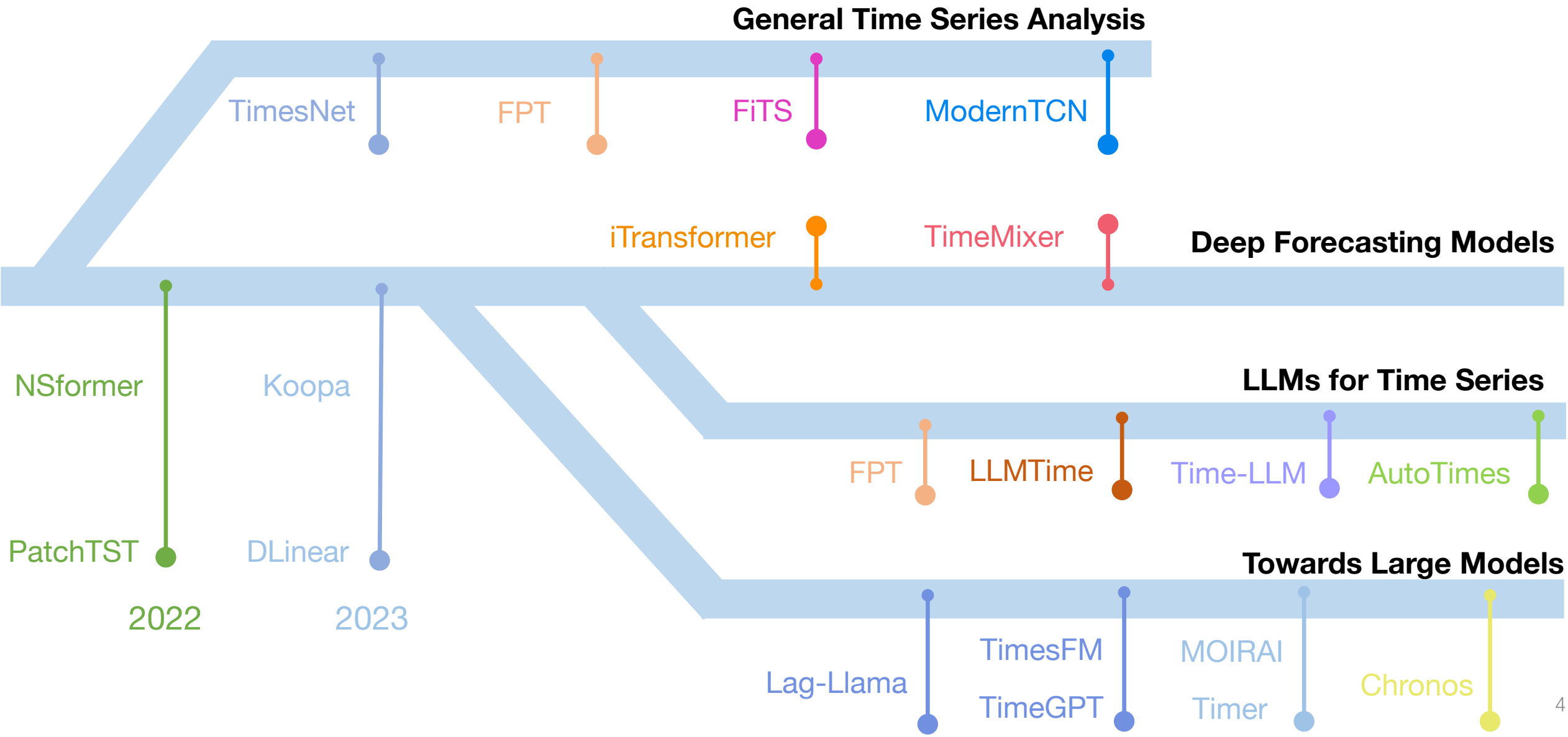




Deep Models for Time Series



Deep Models for Time Series

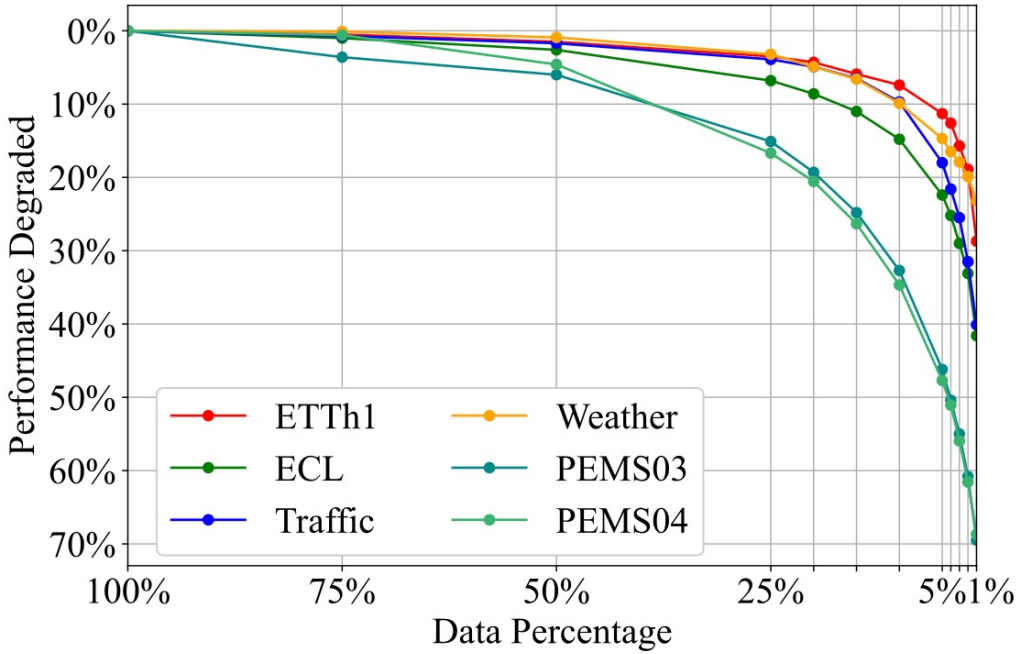


Motivations of Large Models

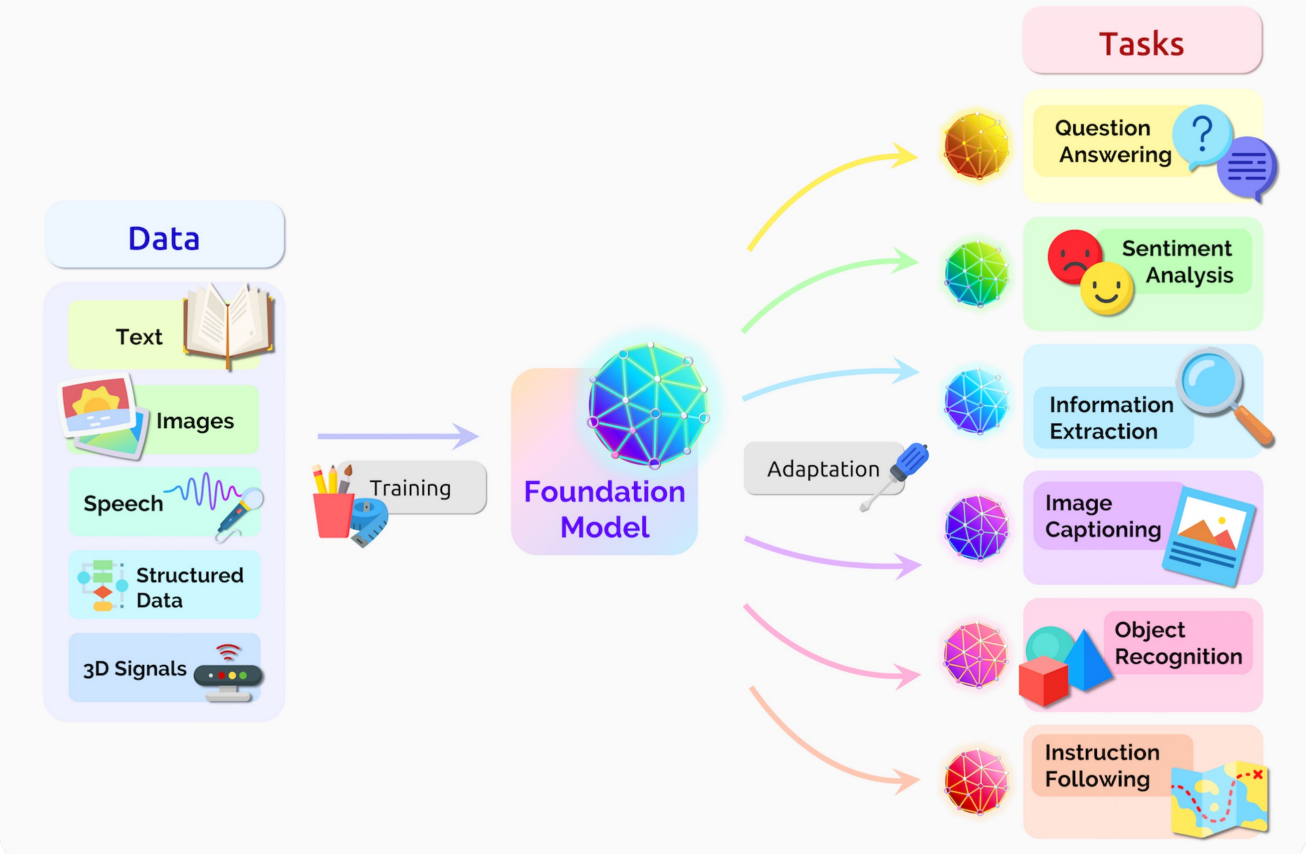
☹ **Status quo:** Costly training small models in specific scenarios (tasks, datasets, settings)

☹ **Data scarcity is common in real-world applications**

- Real scenarios may lack training samples
- Performance of SOTA model degrades with limited data



Motivations of Large Models



[Data Universal]

Learn from various datasets

[Task Universal]

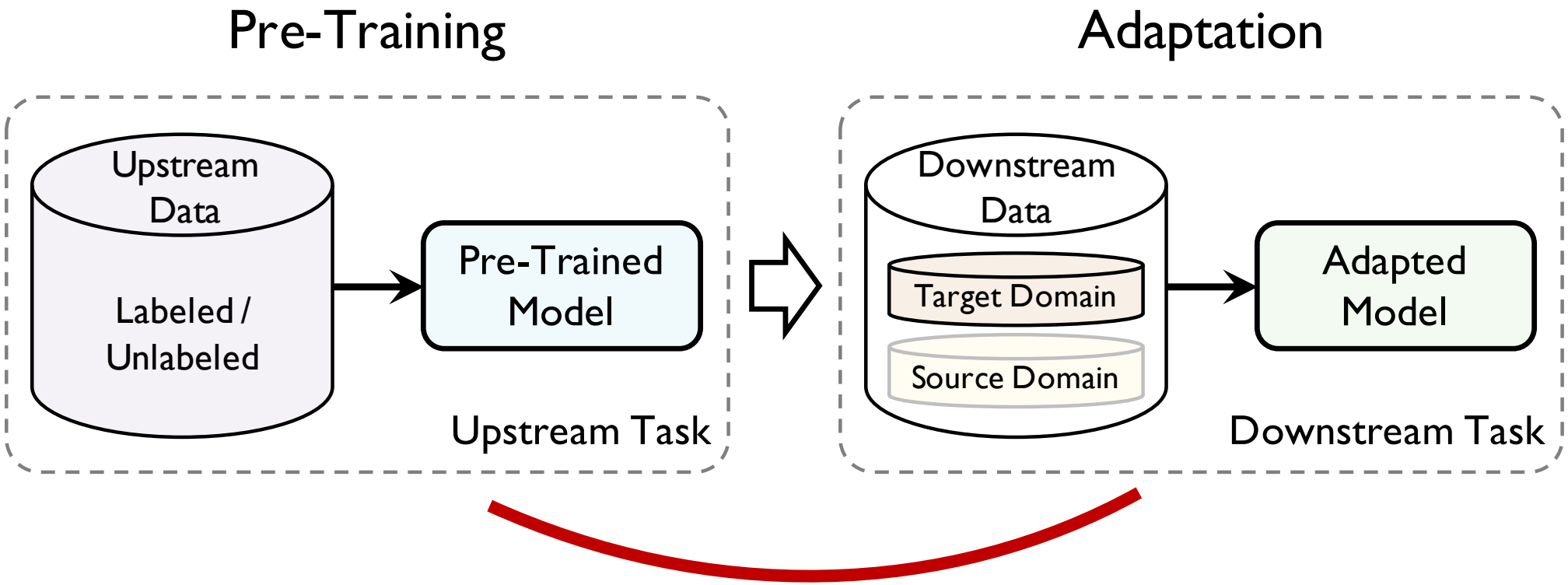
Adapt to a wide range of downstream tasks

Bommasani et al. On the Opportunities and Risks of Foundation Models. *Arxiv 2021.*

Large Time Series Models

What is Large Model

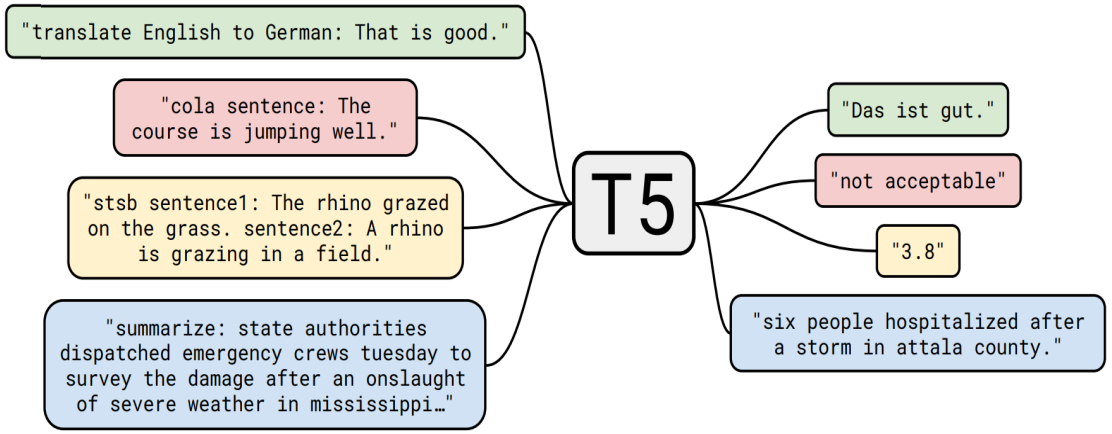
- **Generalizability**: One model fits different domains

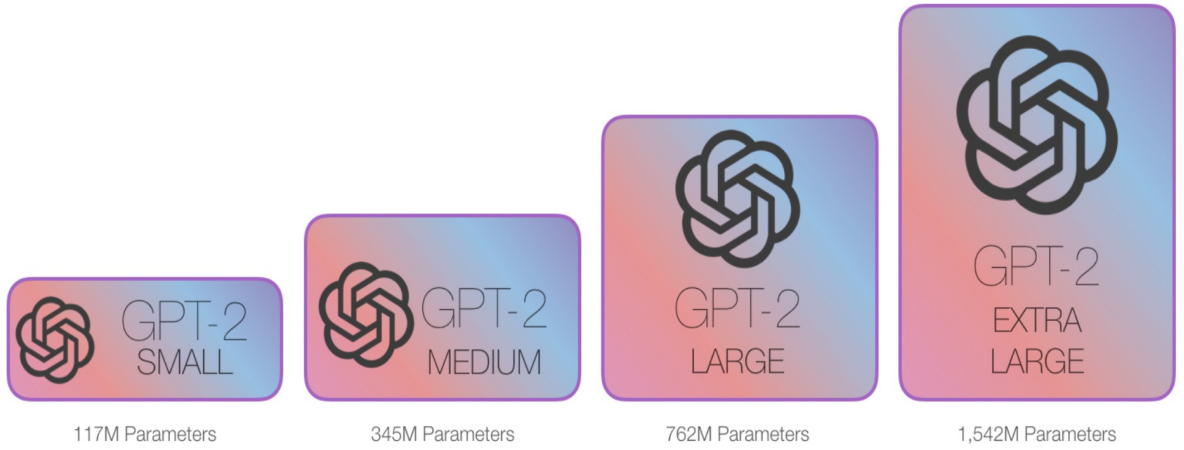


Large Time Series Models

What is Large Model

- **Generalizability**: One model fits different domains
- **Task Generality**: Versatility to cope with various scenarios/tasks
- **Scalability**: Performance improves with the scale of pre-training



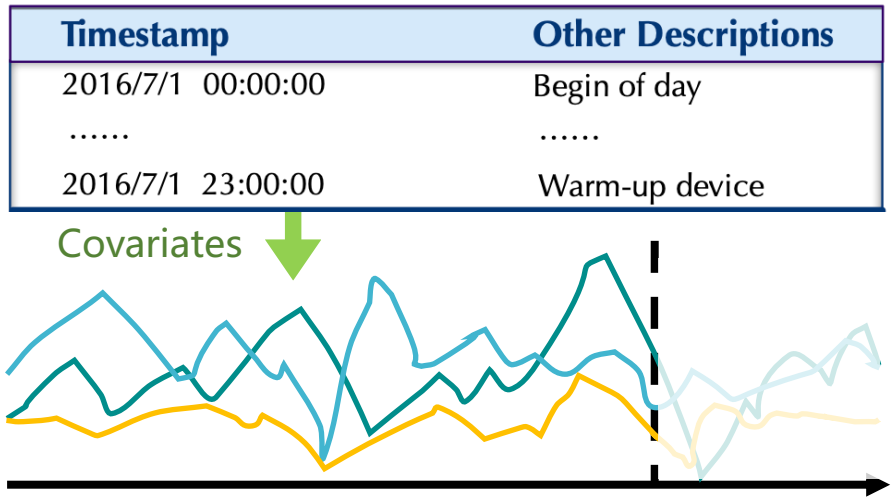


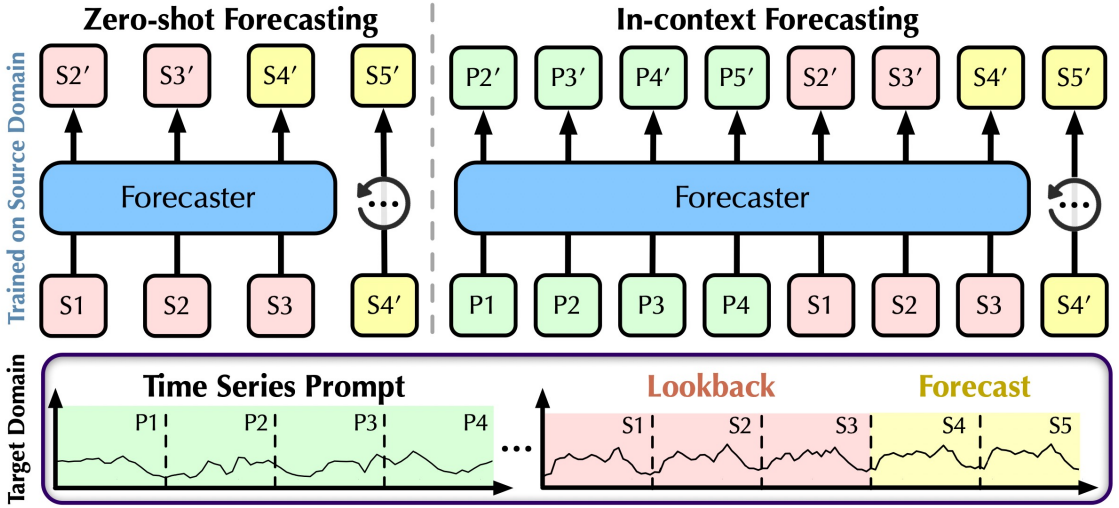
Raffle et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR 2020.*

Large Time Series Models

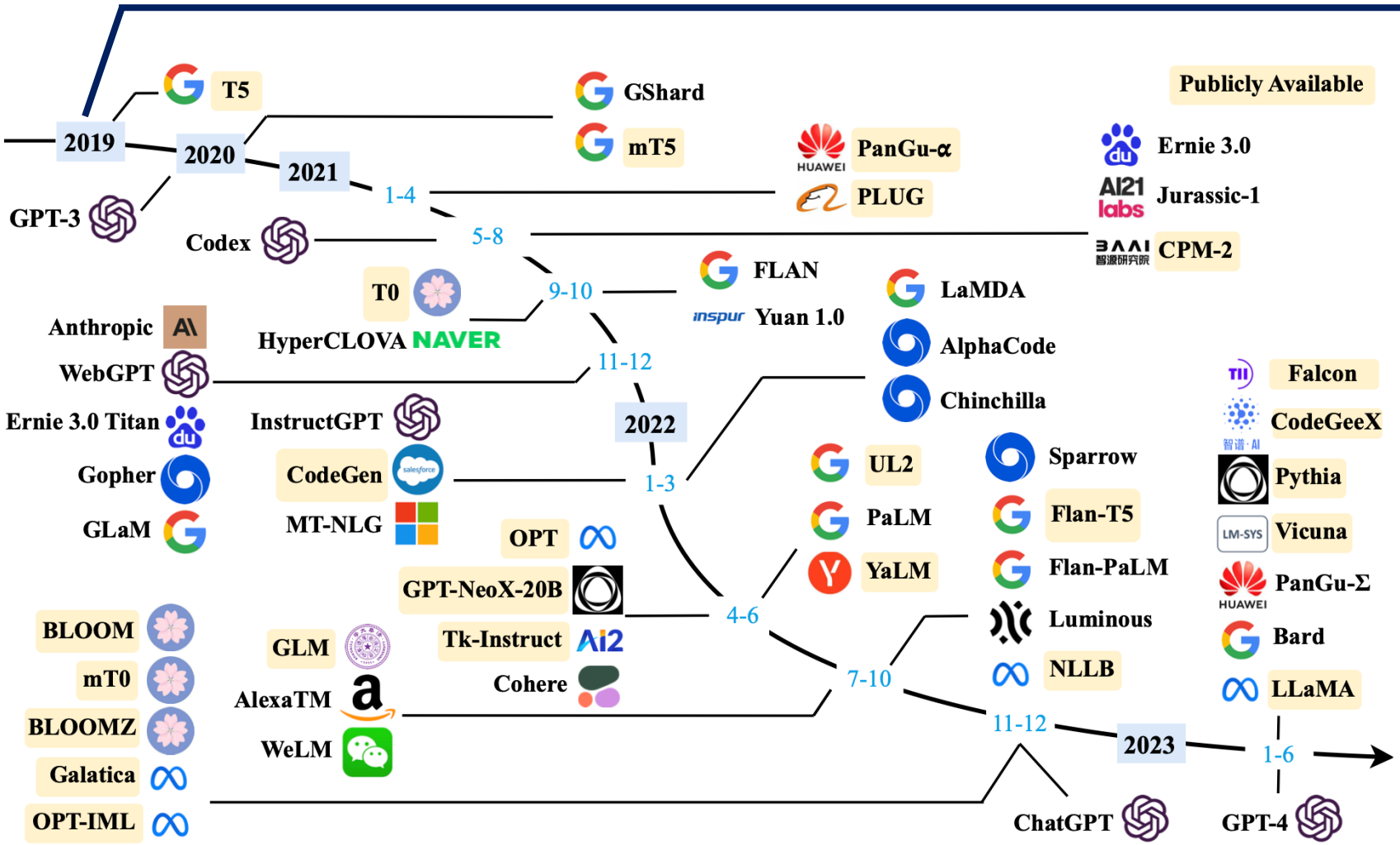
What is Large Model

- **Generalizability**: One model fits different domains
- **Task Generality**: Versatility to cope with various scenarios/tasks
- **Scalability**: Performance improves with the scale of pre-training
- **Emergence Abilities**: Multimodality, In-context Learning …





Timeline of Large Language Models



Large Time Series Models Are Still in Early Stages

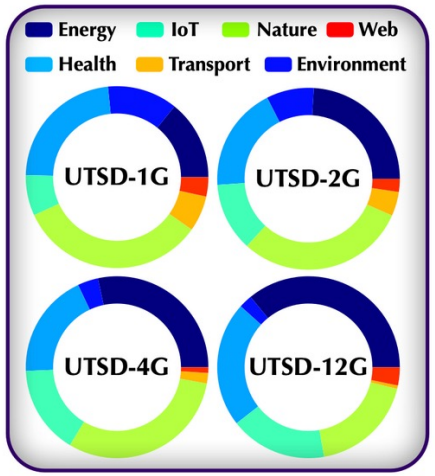
Challenges

- Data Infrastructure
- Scalable Architecture
- Task Heterogeneity

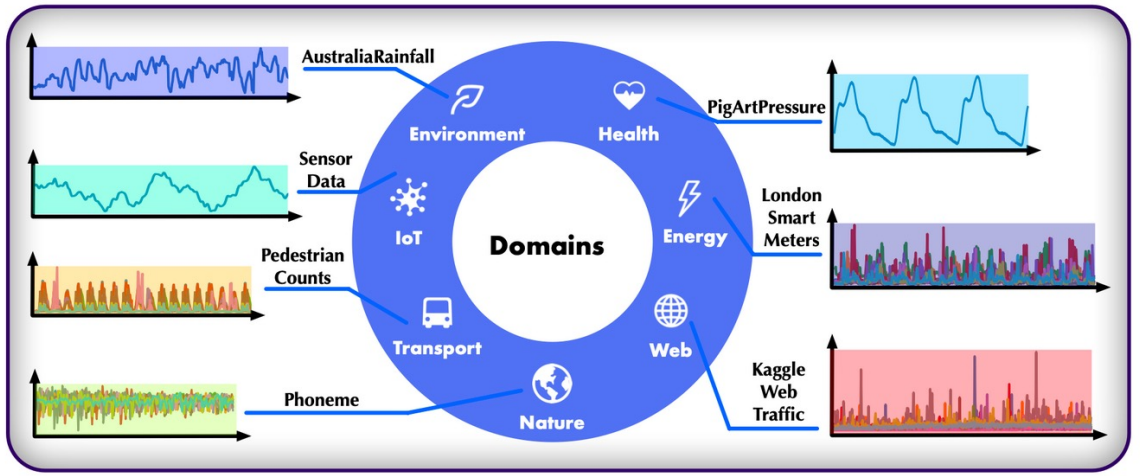
Zhao et al. A Survey of Large Language Models. *arXiv 2023.*

Timer: Well-curated Datasets

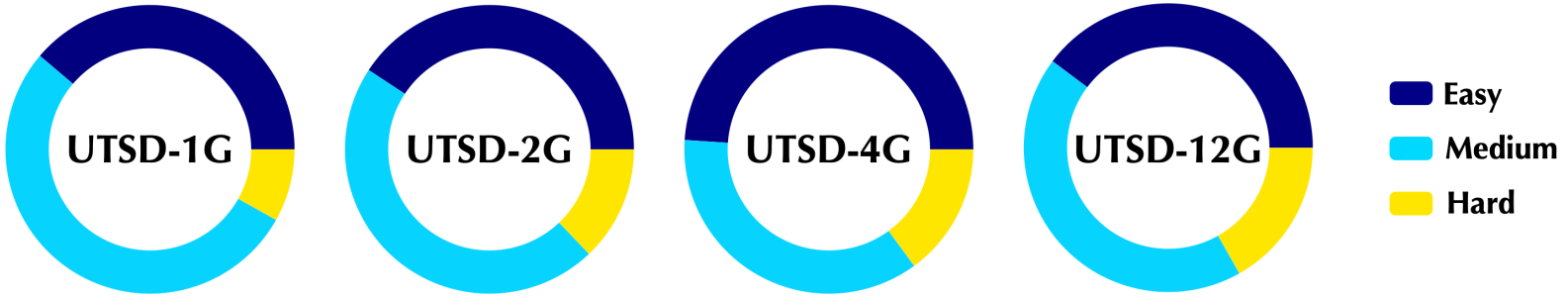
☐ UTSD: Unified Time Series Dataset



Unified Time Series Dataset



Diverse Shapes, Patterns and Domains



Dataset: https://huggingface.co/datasets/thuml/UTSD

Data Quality

- Aggregation & Filter
- Preprocess & Evaluate
- Stacking up with a hierarchy

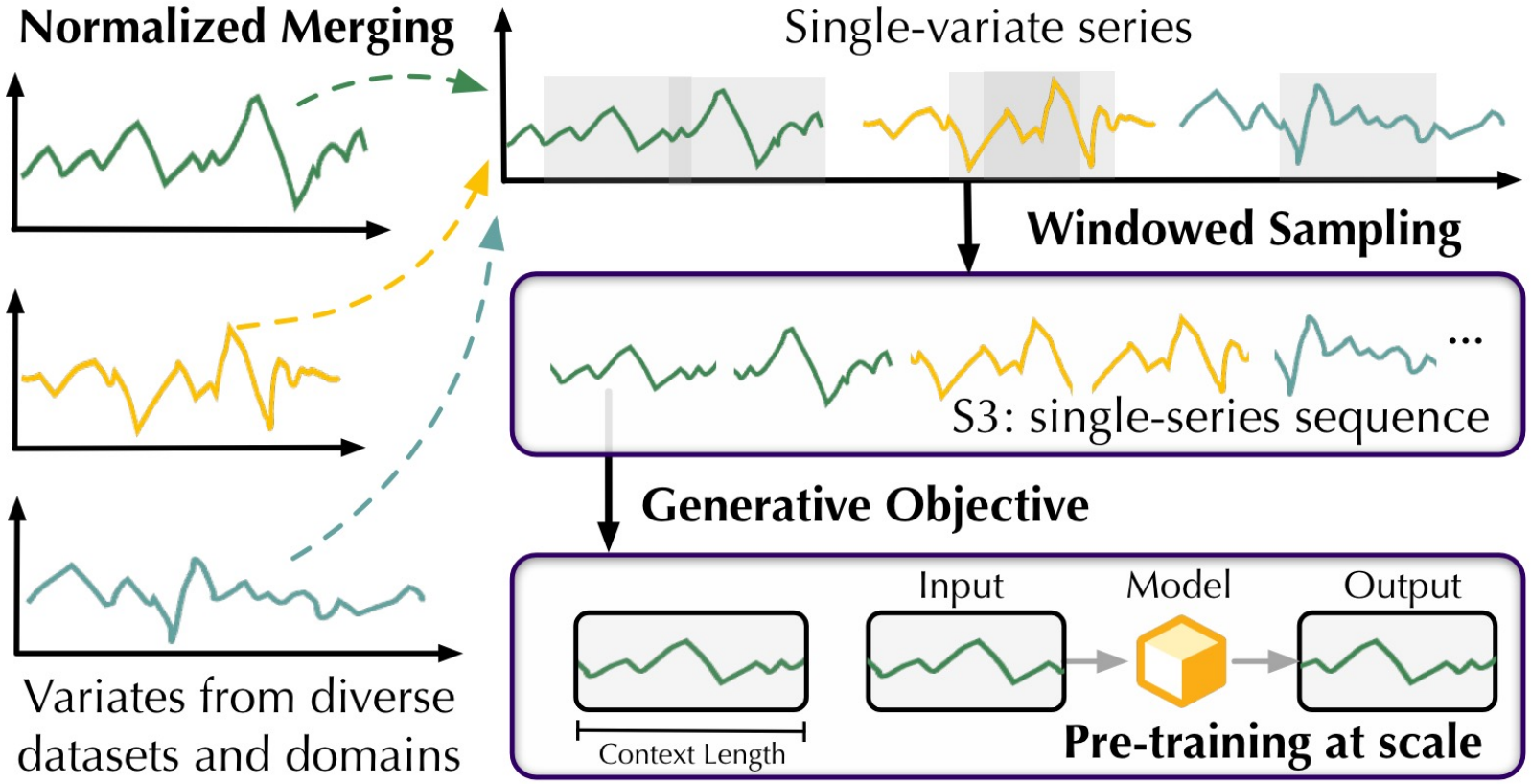
- 1 Billion Time Points
- 7 Typical Domain
- 4 Scalable Volums
- Continuous Expansion…

Timer: Single-Series Sequence

- **Unified Format to Address Data Heterogeneity: single-series sentence (S3)**

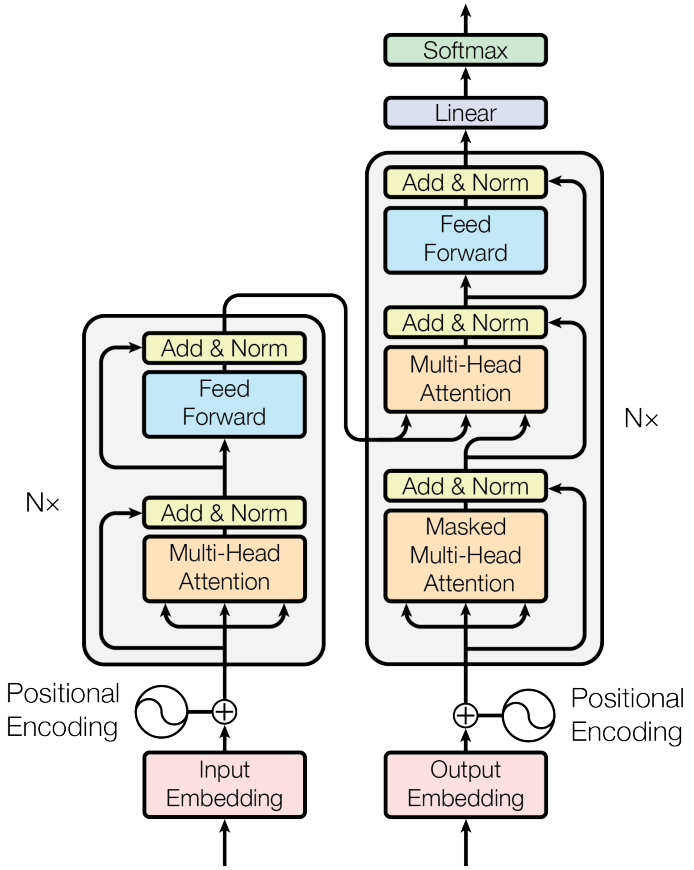
Distinct in Shape/Freq/Scale!

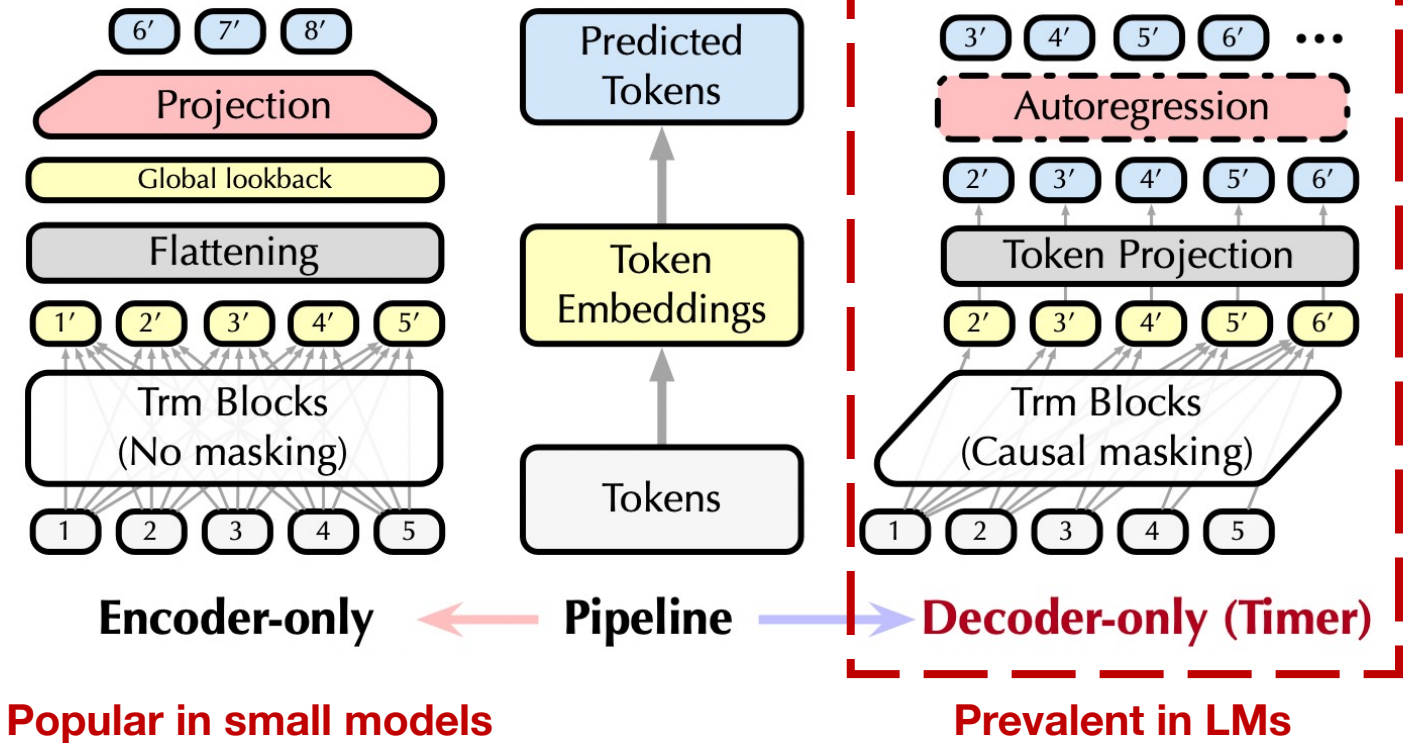
Dataset	Dim	Frequency
ETTh1, ETTh2	7	Hourly
ETTm1, ETTm2	7	15min
Exchange	8	Daily
Weather	21	10min
ECL	321	Hourly
Traffic	862	Hourly
Solar-Energy	137	10min
PEMS03	358	5min
PEMS04	307	5min
PEMS07	883	5min
PEMS08	170	5min



Timer: Explore Backbones for Large Model

- **Decoder-only Transformer with Autoregression**





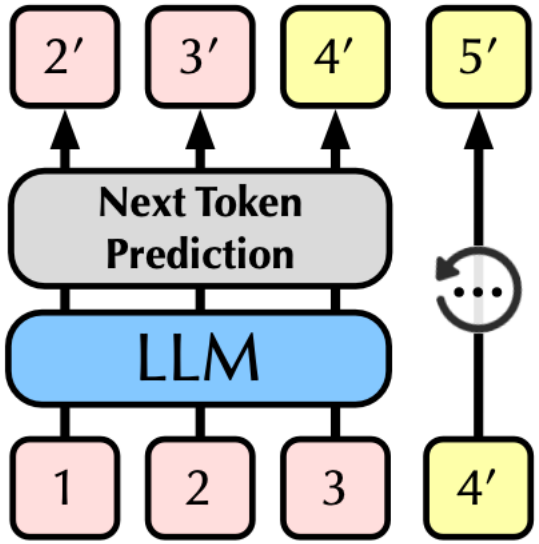
Timer: Generative Pre-training

- **Next Token Prediction (Both Training and Inference)**

$$\text{Tokenize}: \ \mathbf{s}_i = \{x_{(i-1)S+1}, \ldots, x_{iS}\} \in \mathbb{R}^S.$$

$$\begin{aligned} \mathbf{h}_i^0 &= \mathbf{W}_e \mathbf{s}_i + \mathbf{TE}_i, \ i = 1, \ldots, N, \\ \text{Forwarding}: \quad \mathbf{H}^l &= \text{TrmBlock}(\mathbf{H}^{l-1}), \ l = 1, \ldots, L, \\ \{\hat{\mathbf{s}}_{i+1}\} &= \mathbf{H}^L \mathbf{W}_d, \ i = 1, \ldots, N, \end{aligned}$$

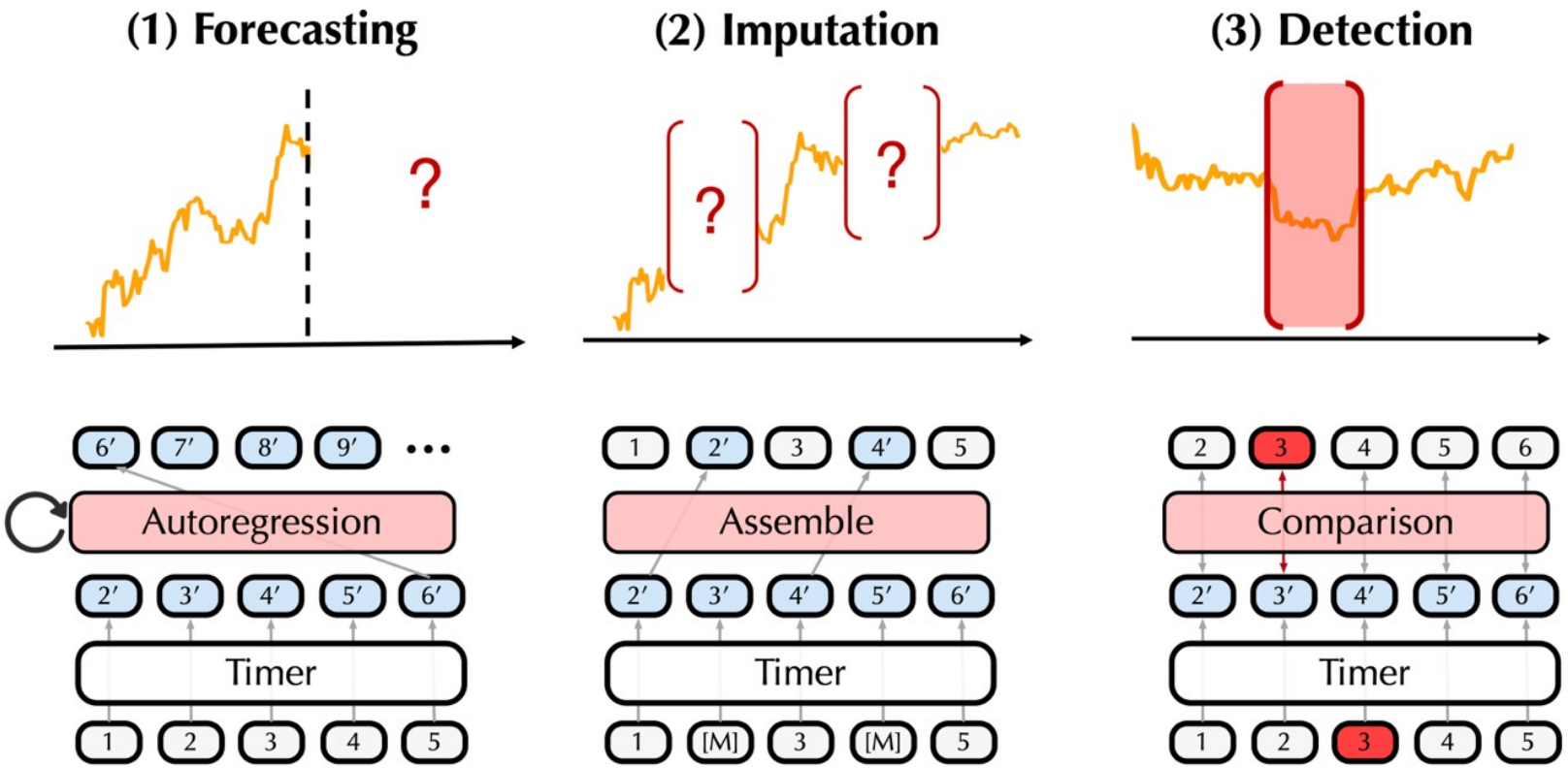
$$\text{NTP}: \ \mathcal{L}_{\text{MSE}} = \frac{1}{NS} \sum ||\mathbf{s}_i - \hat{\mathbf{s}}_i||_2^2, \ i = 2, \ldots, N+1.$$



Token-wise supervision: generated token at each position is independently supervised

Timer: Unified Task Formulation

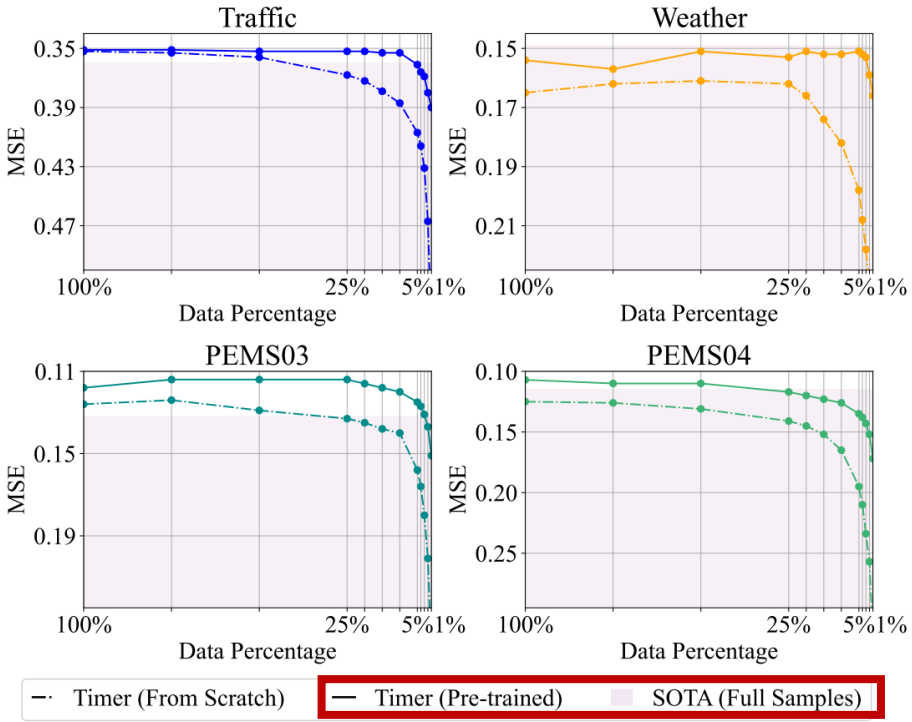
- **Unify Time Series Analysis into Generative Tasks**

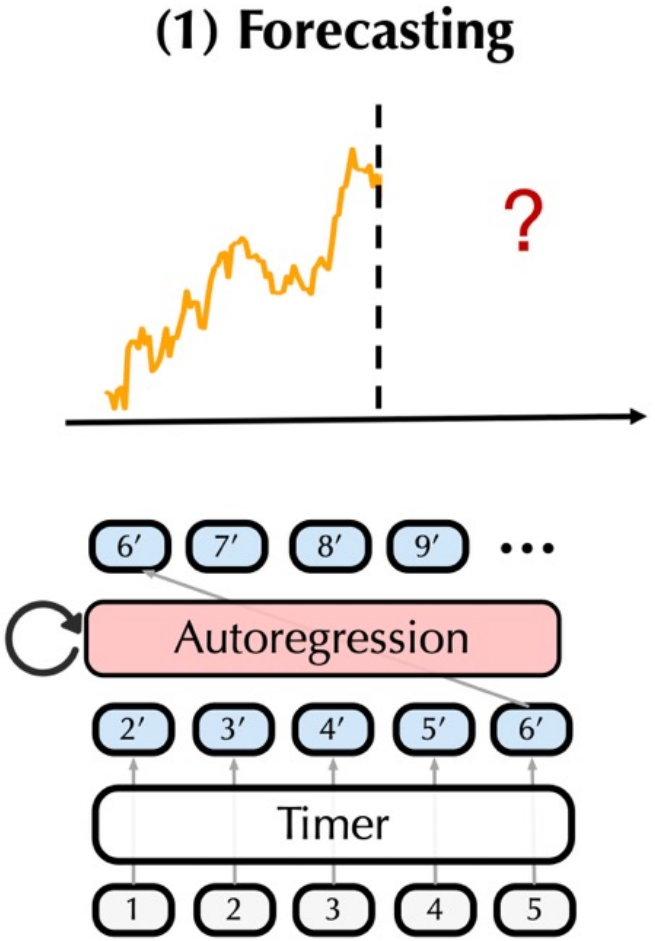


Task Generality

Time Series Forecasting

- Naturally predict the next token
- Timer trained with **1~5% samples** outperforms SOTA with **100% samples**

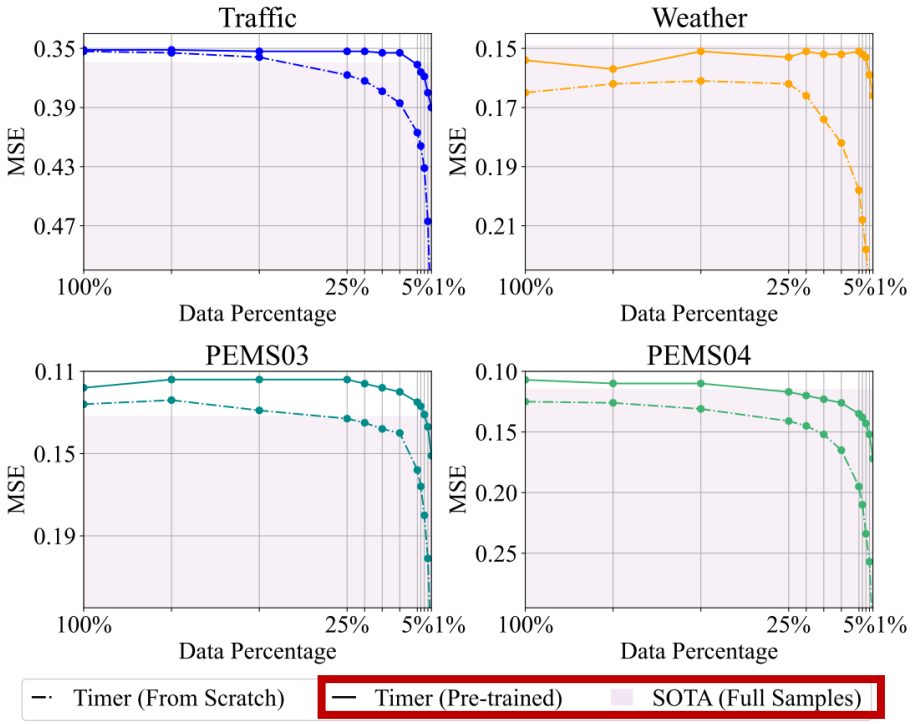




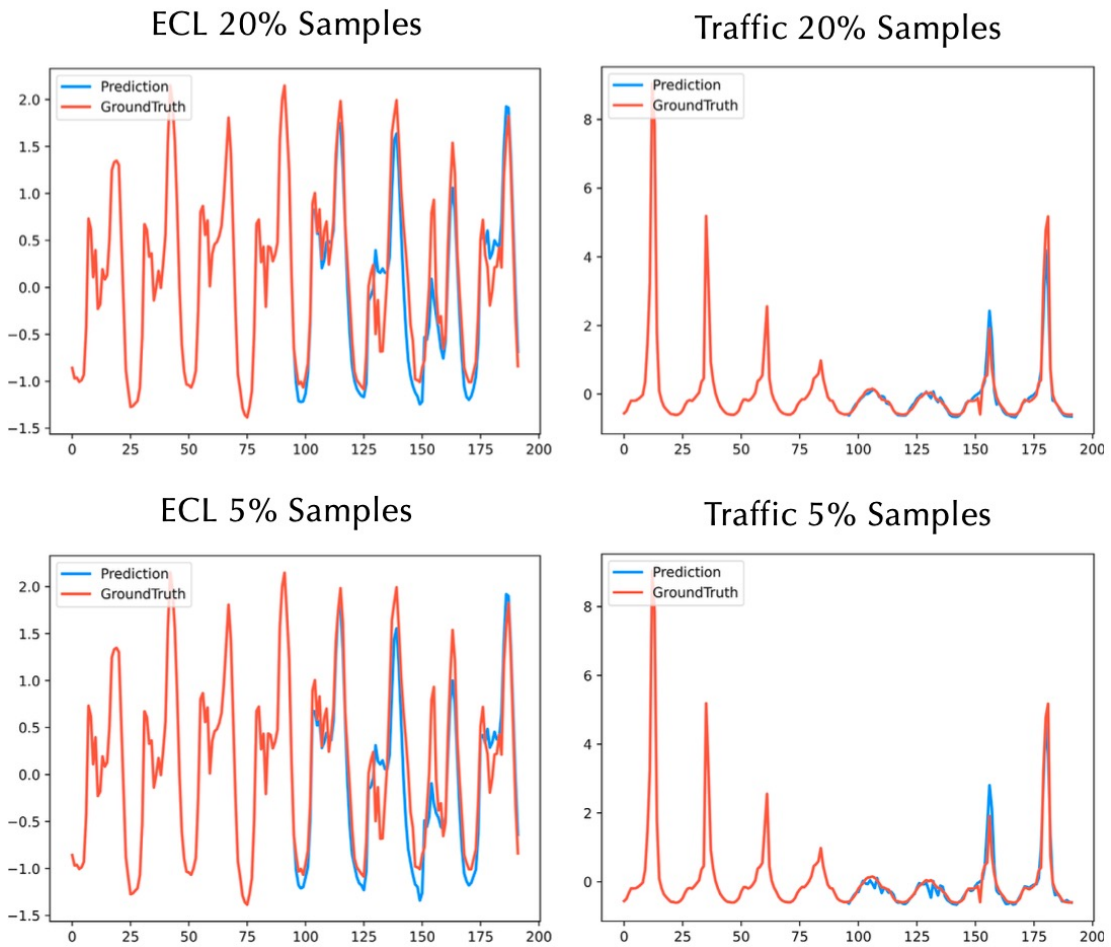
Task Generality

Time Series Forecasting

- Naturally predict the next token
- Timer trained with **1~5% samples** outperforms SOTA with **100% samples**



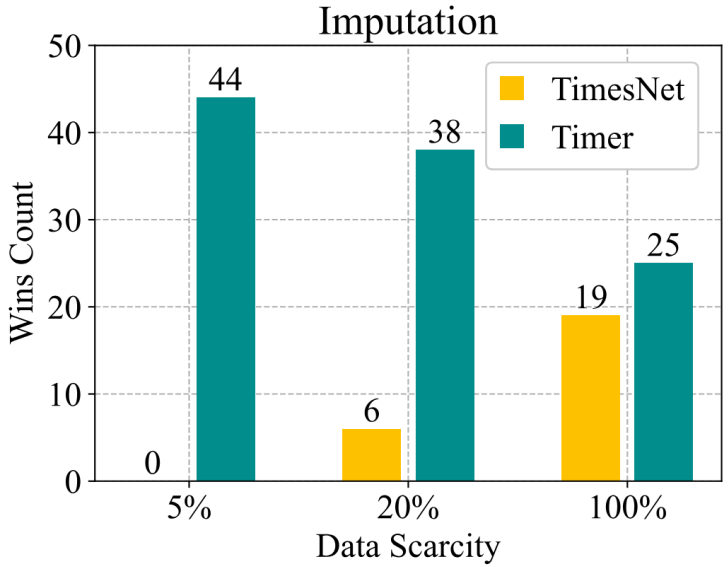
Showcases

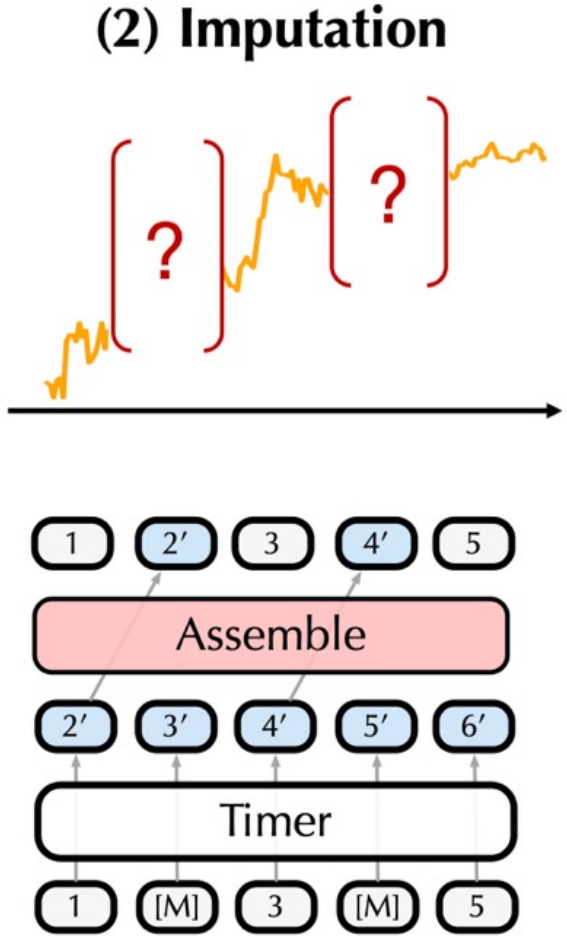


Task Generality

Time Series Imputation

- Imputation is performed by generating masked tokens with the previous context
- Surpass previous **SOTA TimesNet** in average masked cases and data scarcities.

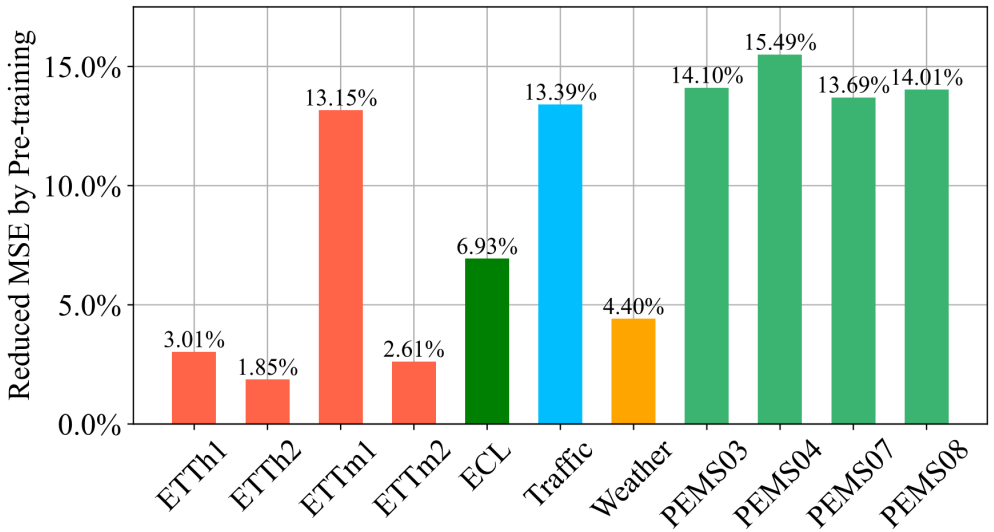




Task Generality

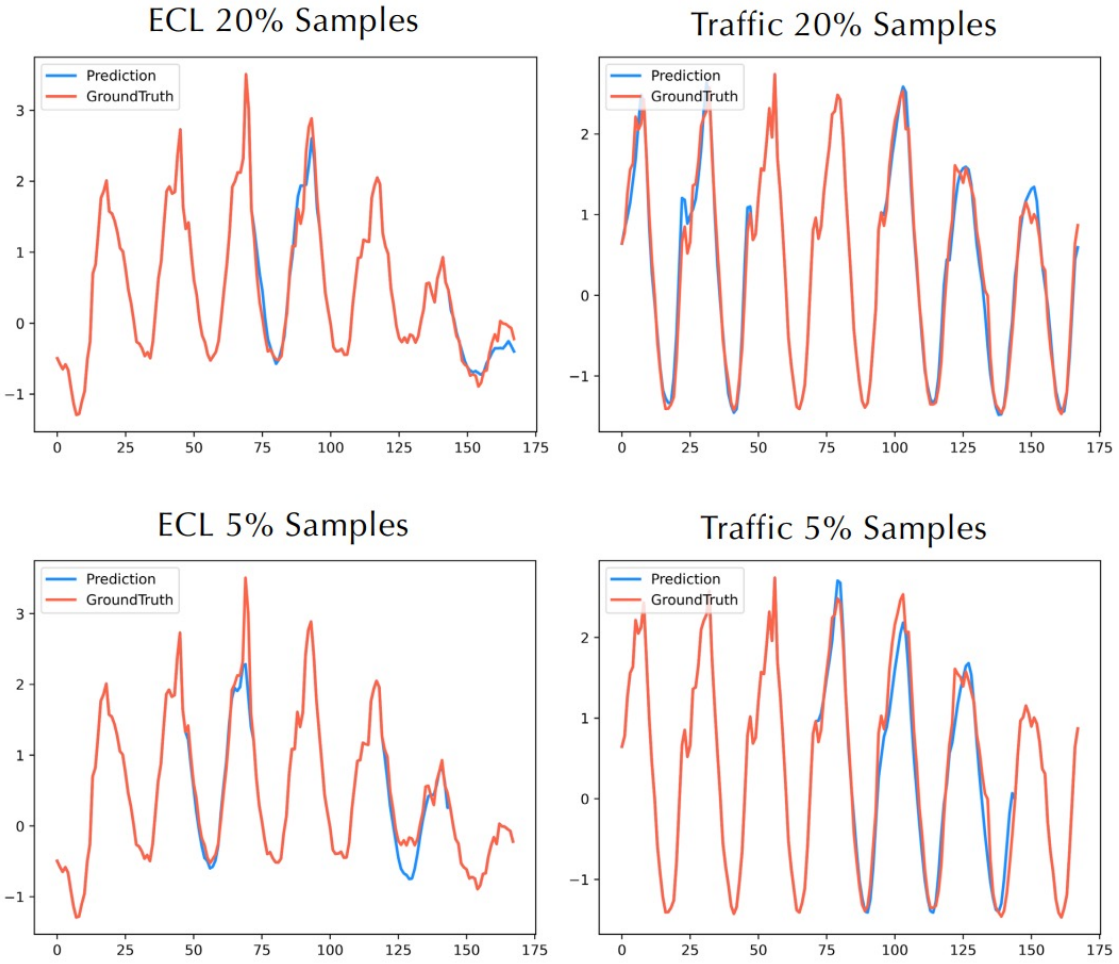
Time Series Imputation

- Imputation is performed by generating masked tokens with the previous context
- **Stable improvement** exhibited in imputation by large-scale pre-training



Compared with training from scratch

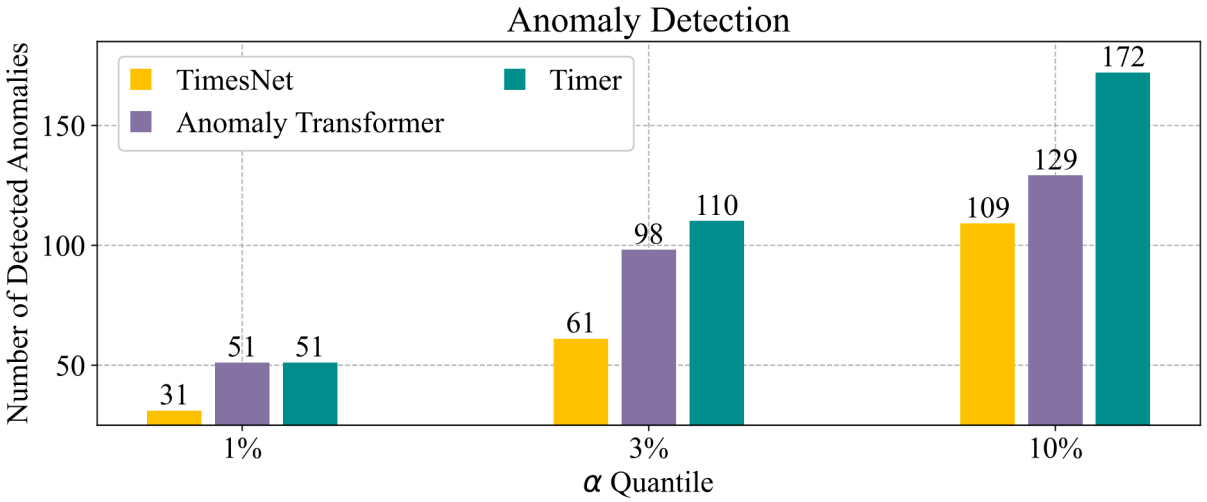
Showcases

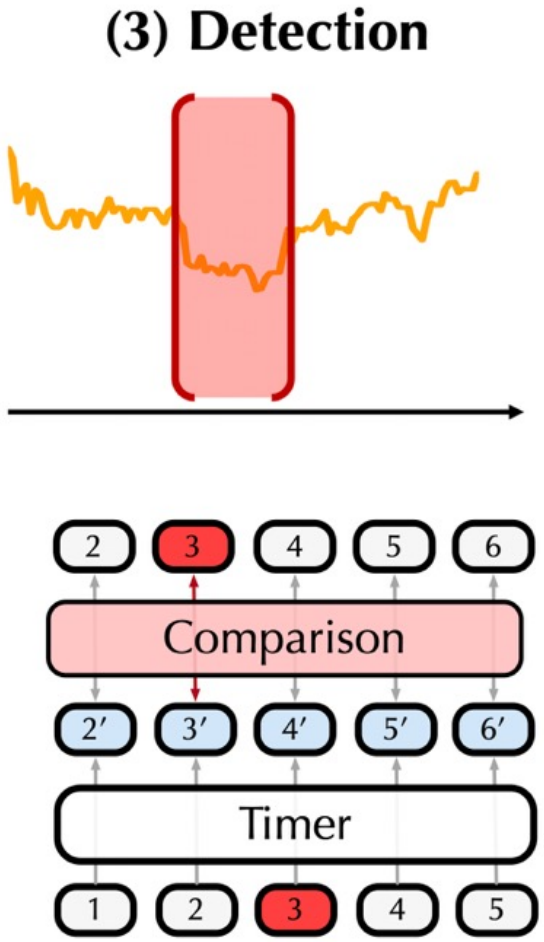


Task Generality

Anomaly Detection

- Conducted in a **predictive** approach by generating normal time series
- **Quantile** the abnormal confidence in MSE
- Surpass **task-specific SOTA models** in the challenging UCR Anomaly Archive

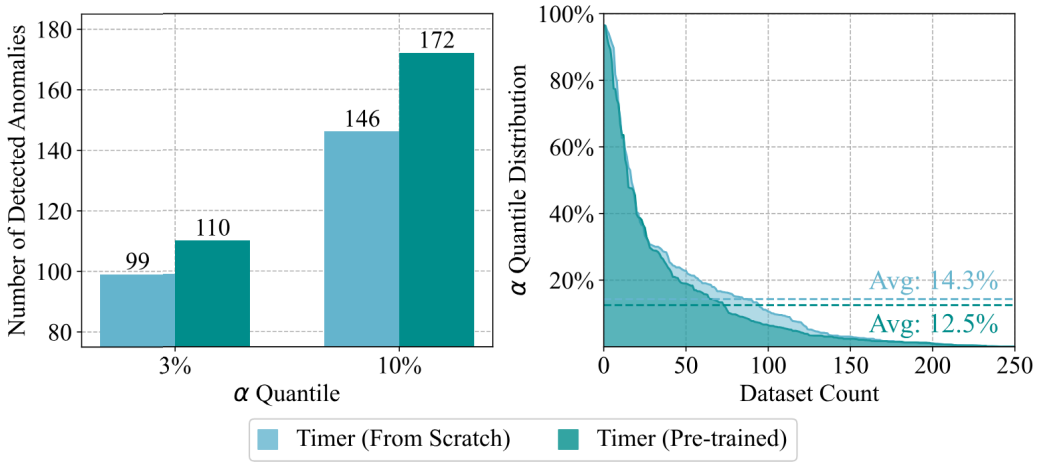




Task Generality

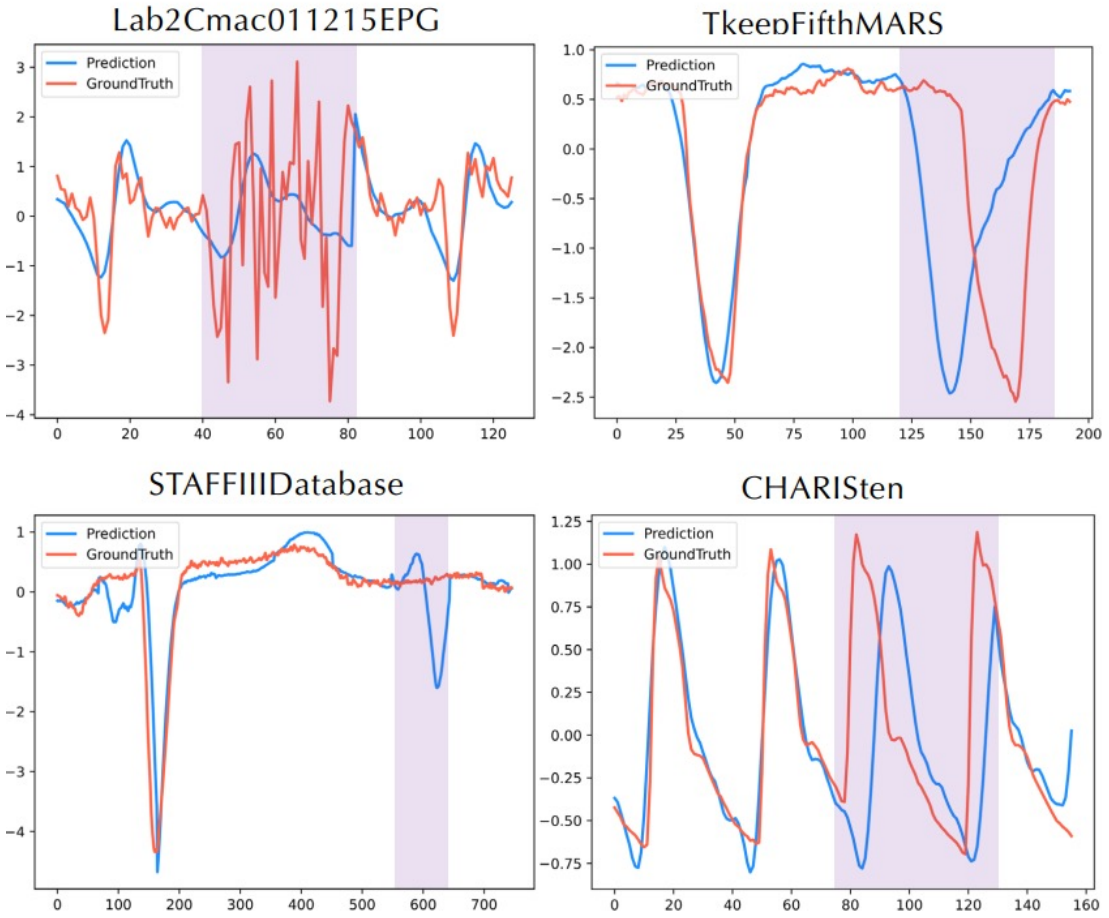
Anomaly Detection

- Conducted in a **predictive** approach by generating normal time series
- **Quantile** the abnormal confidence in MSE
- **Stable improvement** exhibited in anomaly detection by large-scale pre-training



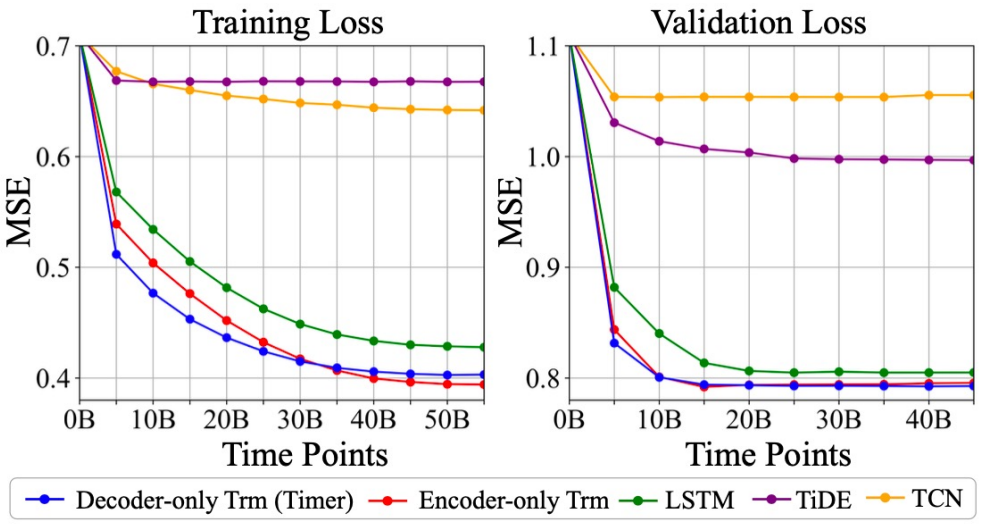
Smaller α indicates better performance

Showcases



Scalability

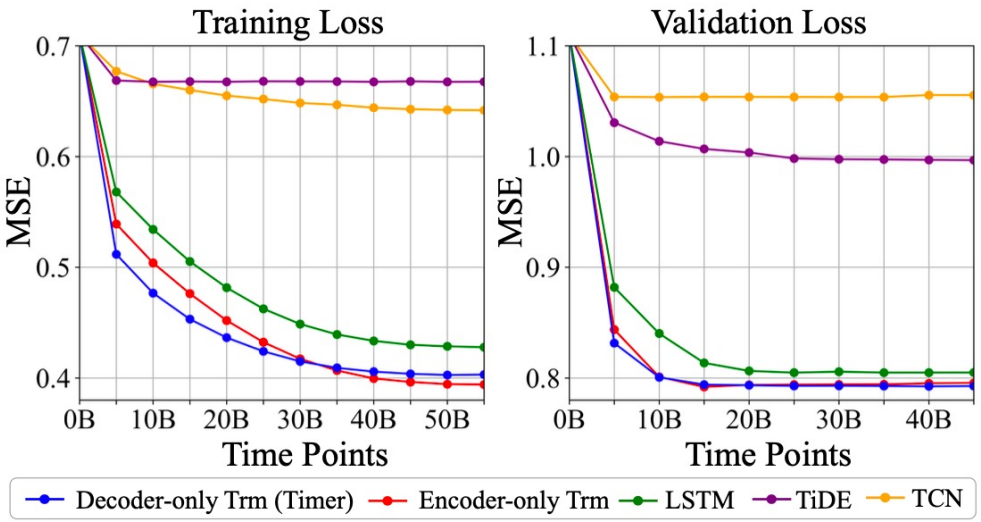
Loss Curve of Sequence Models on UTSD



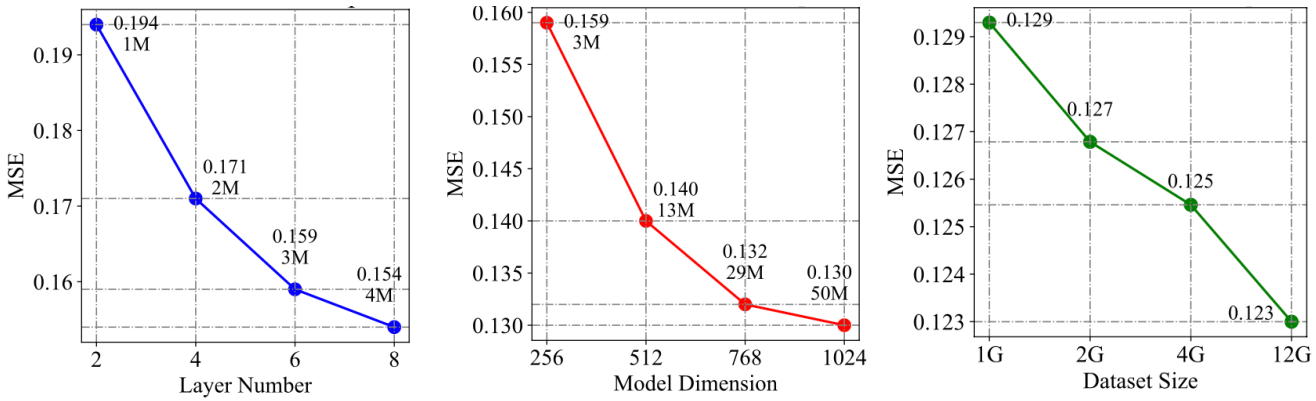
Transformer exhibits model capacity as the scalable architecture for LTSM

Scalability

Loss Curve of Sequence Models on UTSD



Scaling Model/Data Consistently Improves Performance



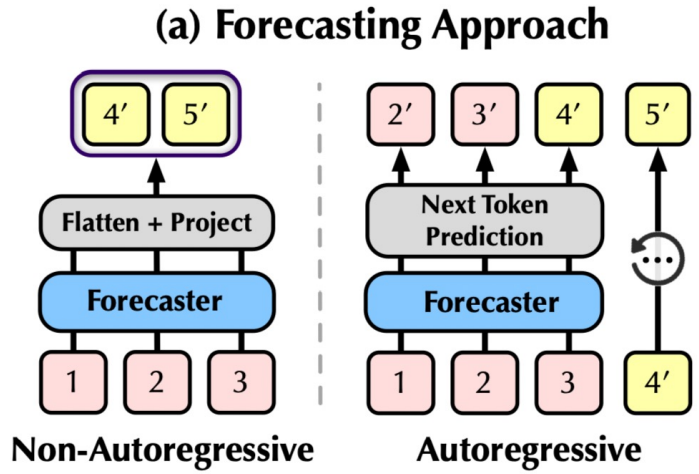
Transformer exhibits model capacity as the scalable architecture for LTSM

Scaling Timer achieves MSE: 0.194 → 0.123 (–36.6%) under data scarcity, surpassing the state-of-the-art (0.129) model with full samples

Autoregressive Model

Variable Lookback Length

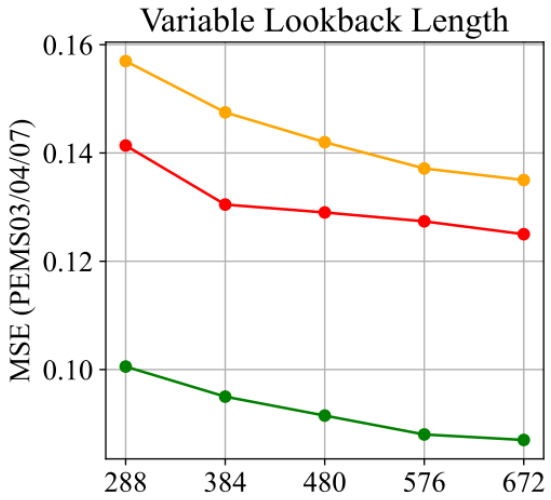
- Small models are constrained on fixed input/output lengths
- Similar to LLMs, Decoder is flexible on the context length



Autoregressive Model

Variable Lookback Length

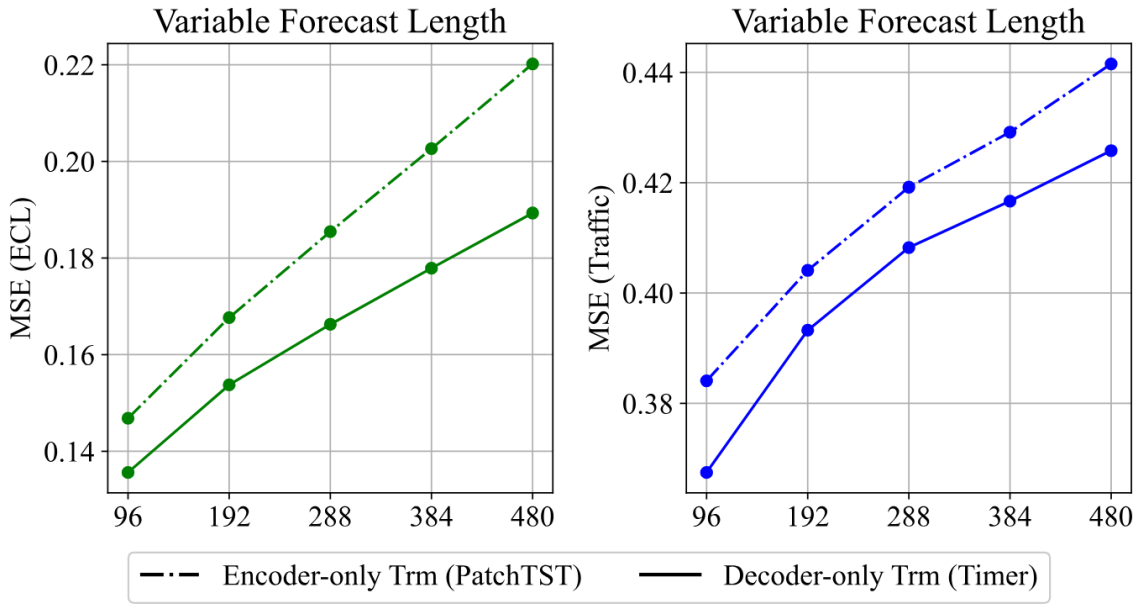
- Small models are constrained on fixed input/output lengths
- Similar to LLMs, Decoder is flexible on the context length
- Increasing the lookback window leads to stable accuracy growth



Iterative Multi-step Prediction

- Token-wise supervision alleviates error accumulation

$$\mathcal{L}_{\text{MSE}} = \frac{1}{NS} \sum ||\mathbf{s}_i - \hat{\mathbf{s}}_i||_2^2, \ i = 2, ..., N+1.$$



Few-shot Generalization

Superwisedly Trained from Scratch

Table 1. Downstream forecasting results under different data scarcity of the encoder-only and decoder-only Transformer respectively pre-trained on UTST-12G. Datasets are ordered by the oversaturation in Figure 1. Full results of PEMS and ETT can be found in Table 14.

Scenario	1% Target				5% Target				20% Target			
Architecture	Encoder		Decoder		Encoder		Decoder		Encoder		Decoder	
Pre-trained	None	12G	None	12G	None	12G	None	12G	None	12G	None	12G
PEMS (Avg)	0.286	0.246	0.328	**0.180**	0.220	0.197	0.215	**0.138**	0.173	0.164	0.153	**0.126**
ECL	0.183	0.168	0.215	**0.140**	0.150	0.147	0.154	**0.132**	0.140	0.138	0.137	**0.134**
Traffic	0.442	0.434	0.545	**0.390**	0.392	0.384	0.407	**0.361**	0.367	0.363	0.372	**0.352**
ETT (Avg)	0.367	0.317	0.340	**0.295**	0.339	0.303	0.321	**0.285**	0.309	0.301	0.297	**0.288**
Weather	0.224	0.165	0.246	**0.166**	0.182	0.154	0.198	**0.151**	0.153	**0.149**	0.166	0.151

Encoder will outperform when training samples are insufficient

Decoder necessitates substantial samples in end-to-end settings

Few-shot Generalization

Generalization on Downstream Tasks

Table 1. Downstream forecasting results under different data scarcity of the encoder-only and decoder-only Transformer respectively pre-trained on UTST-12G. Datasets are ordered by the oversaturation in Figure 1. Full results of PEMS and ETT can be found in Table 14.

Scenario	1% Target				5% Target				20% Target			
Architecture	Encoder		Decoder		Encoder		Decoder		Encoder		Decoder	
Pre-trained	None	12G	None	12G	None	12G	None	12G	None	12G	None	12G
PEMS (Avg)	0.286	0.246	0.328	**0.180**	0.220	0.197	0.215	**0.138**	0.173	0.164	0.153	**0.126**
ECL	0.183	0.168	0.215	**0.140**	0.150	0.147	0.154	**0.132**	0.140	0.138	0.137	**0.134**
Traffic	0.442	0.434	0.545	**0.390**	0.392	0.384	0.407	**0.361**	0.367	0.363	0.372	**0.352**
ETT (Avg)	0.367	0.317	0.340	**0.295**	0.339	0.303	0.321	**0.285**	0.309	0.301	0.297	**0.288**
Weather	0.224	0.165	0.246	**0.166**	0.182	0.154	0.198	**0.151**	0.153	**0.149**	0.166	0.151

In terms of Pre-training->Adaptation

Better performance can be achieved by Decoder Trm (Timer)

Evaluations of LTSMs

Quality Assessments

Method	Timer (Ours)	Moirai (2024)	MOMENT (2024)	Chronos (2024)	Lag-Llama (2023)	TimesFM (2023b)	TimeGPT-1 (2023)
Architecture	Decoder	Encoder	Encoder Decoder	Encoder Decoder	Decoder	Decoder	Encoder Decoder
Model size	29M, 50M, 67M	14M, 91M, 311M	40M, 125M 385M	20M, 46M, 200M, 710M	200M	17M, 70M, 200M	Unknown
Supported tasks	Forecast Imputation Detection	Forecast	Forecast Imputation Classification Detection	Forecast	Forecast	Forecast	Forecast Detection
Pre-training Scale	28B	27.65B	1.13B	84B	0.36B	100B	100B
Token type	Segment	Segment	Segment	Point	Point	Segment	Segment
Context length	≤1440	≤5000	= 512	≤512	≤1024	≤512	Unknown
Variable length	True	True	False	True	True	True	True
Probabilistic	False	True	False	True	True	True	True

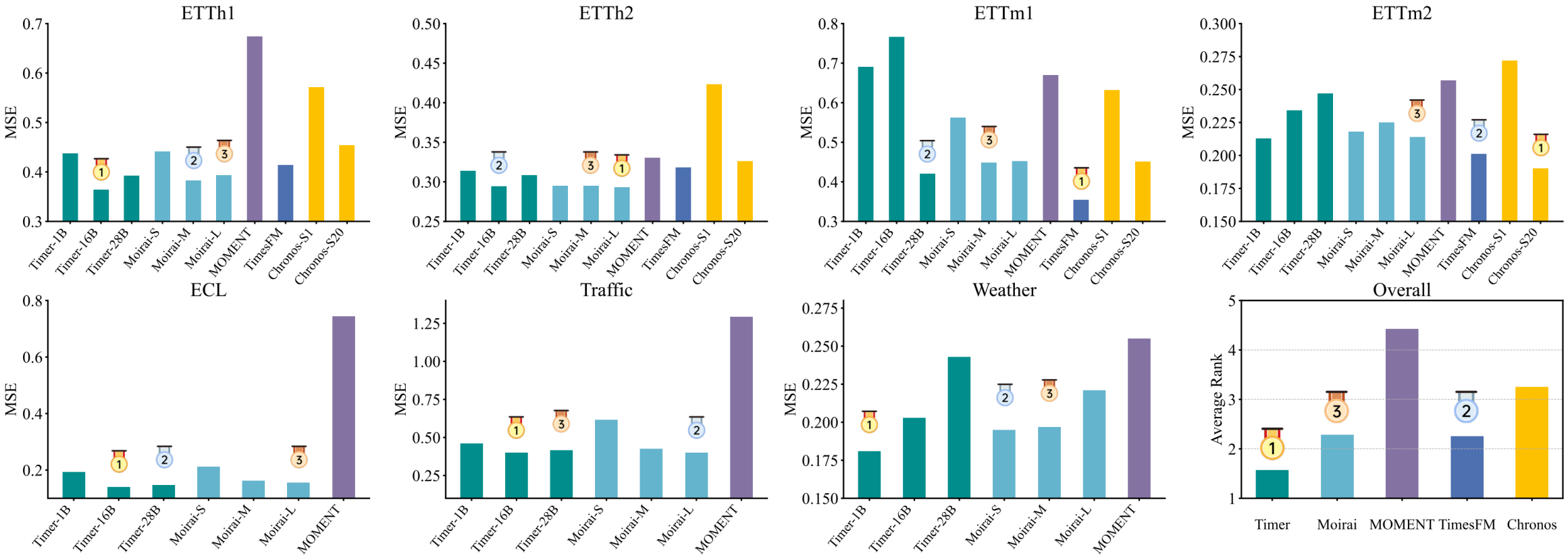
[2]https://huggingface.co/AutonLab/MOMENT-1-large
[3]https://huggingface.co/amazon/chronos-t5-large
[4]https://huggingface.co/google/timesfm-1.0-200m
[5]https://huggingface.co/collections/Salesforce/moirai-10-r-models-65c8d3a94c51428c300e0742

Future Directions

- **Generalization**
- **Longer Context**
- **Probabilistic**
- **More Tasks**
- **……**

Benchmarks of LTSMs

Quantitative Evaluations (Zero-shot Forecasting)

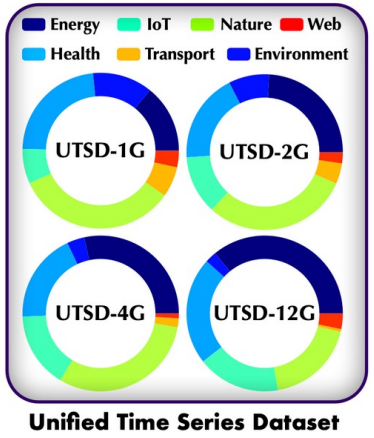


We provided the average rank, where the lower is better,

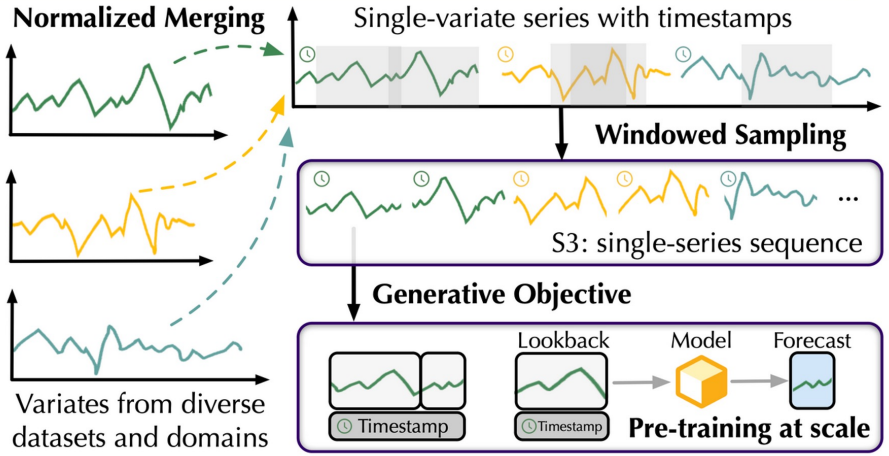
to measure LTSMs as a *general-purpose zero-shot* forecaster

Summary

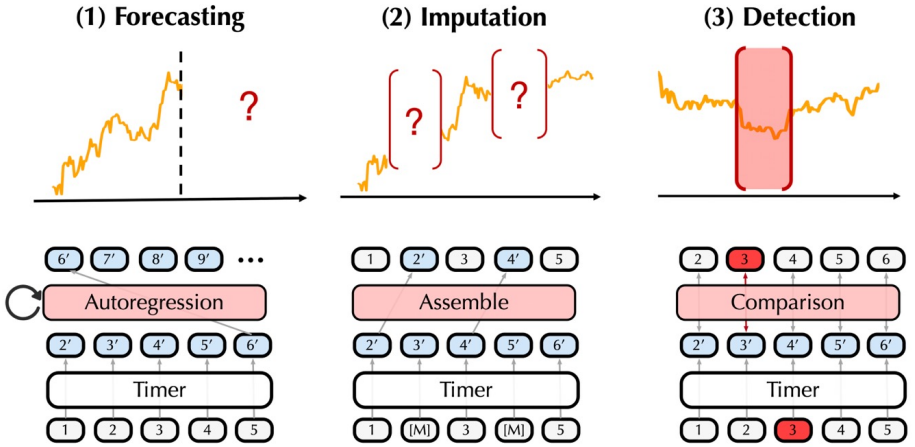
Dataset

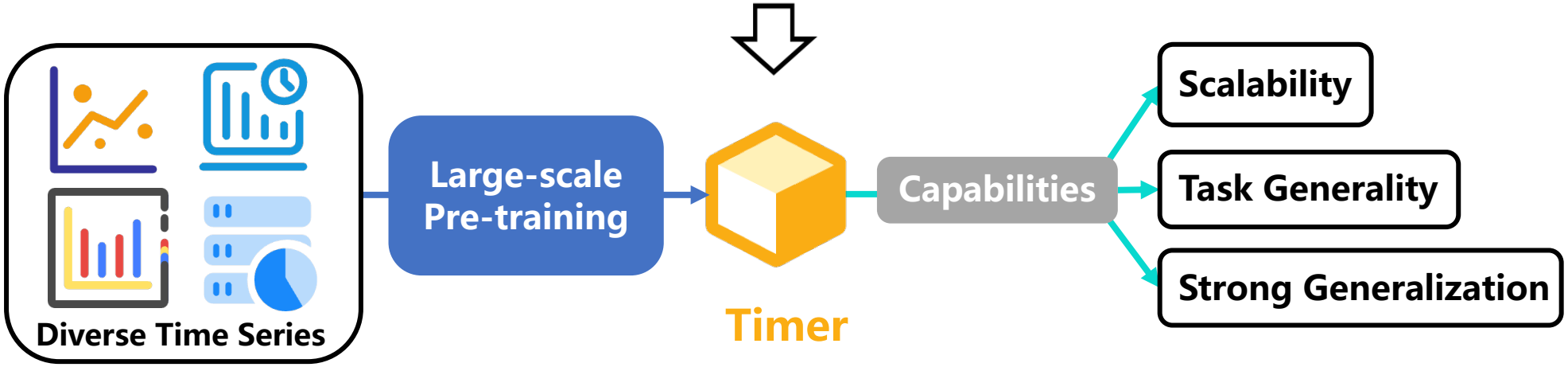


Pre-training



Adaptation





Thank You!

Timer

GitHub: https://github.com/thuml/Large-Time-Series-Model