

Understanding the Implicit Biases of Design Choices for Time Series Foundation Models

Annan Yu, Danielle C. Maddix, Boran Han, Xiyuan Zhang,
Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos,
Andrew Gordon Wilson, Michael W. Mahoney, Yuyang Wang

ICLR 2026

Neural Scaling Laws and the Bitter Lesson

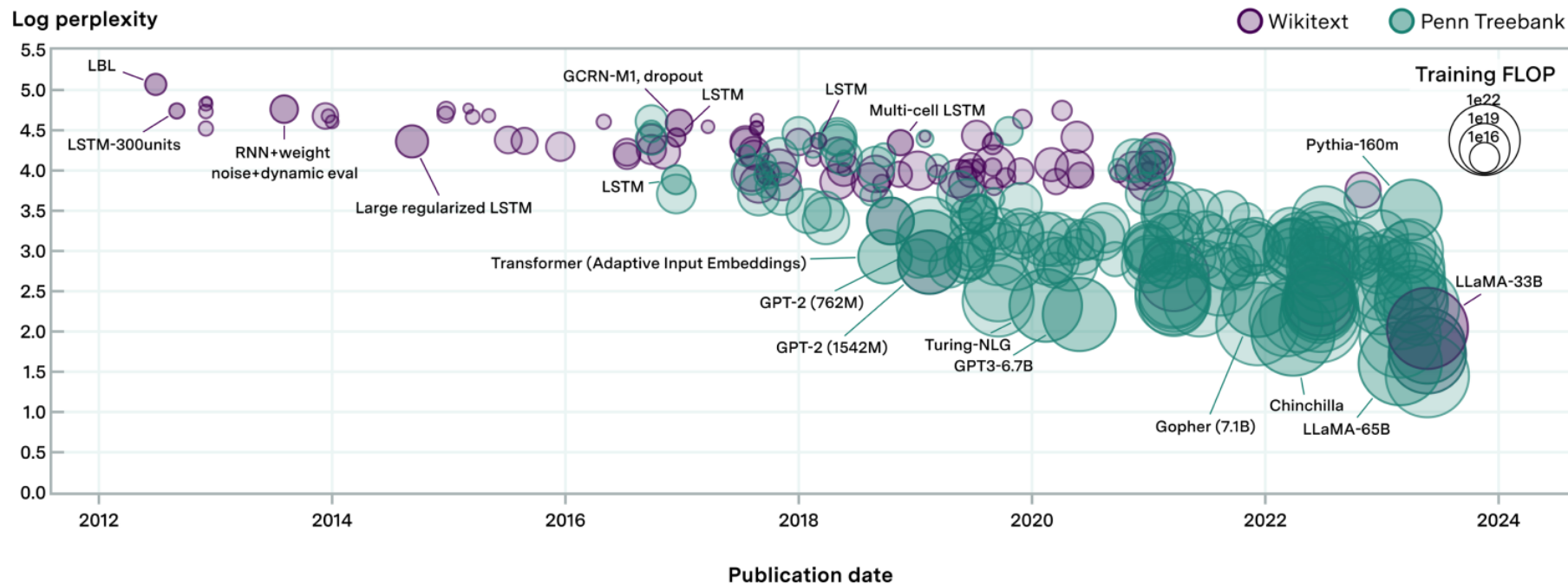
$$\mathcal{L}_{\text{validation}} \sim f(m, D, N) + \mathcal{L}_{\infty}.$$

We have to learn the bitter lesson that building in how we think we think does not work in the long run.

The two methods that seem to scale arbitrarily ... are learning and search.

- "The Bitter Lesson", Sutton 2019

Performance and scale of language models over time



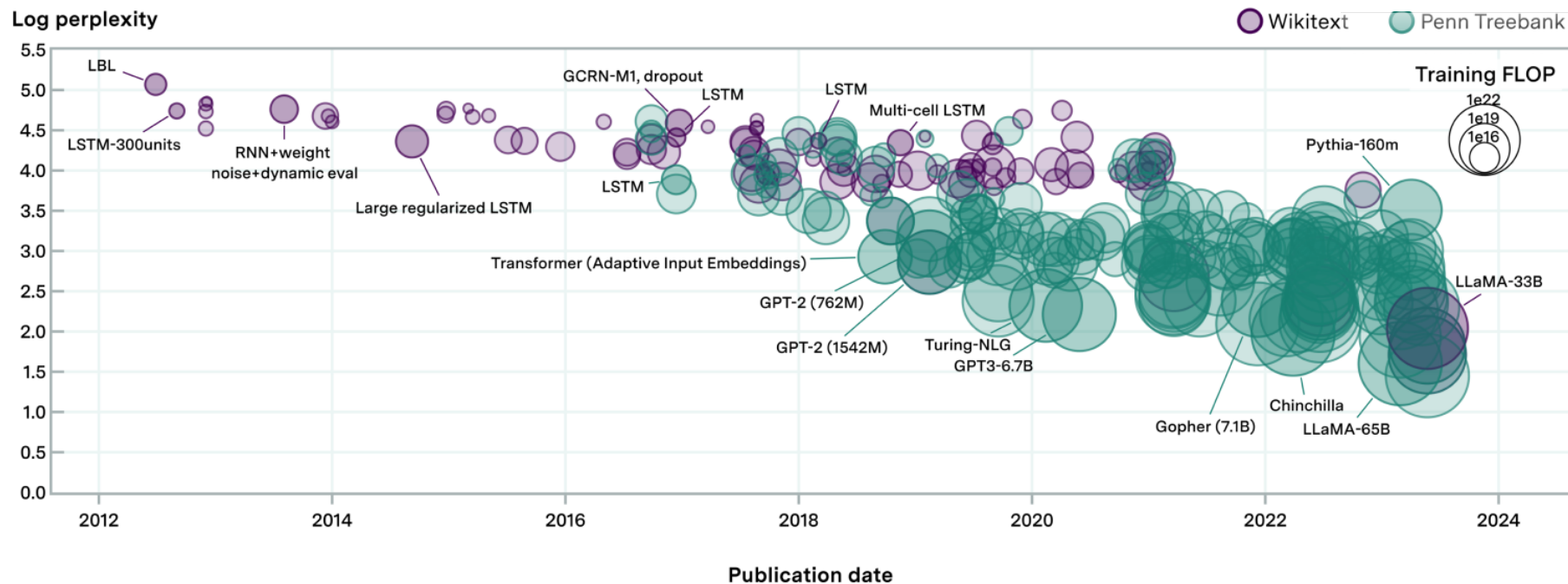
Neural Scaling Laws and the Bitter Lesson

$$\mathcal{L}_{\text{validation}} \sim f(m, D, N) + \mathcal{L}_{\infty}.$$

However, \mathcal{L}_{∞} captures things like

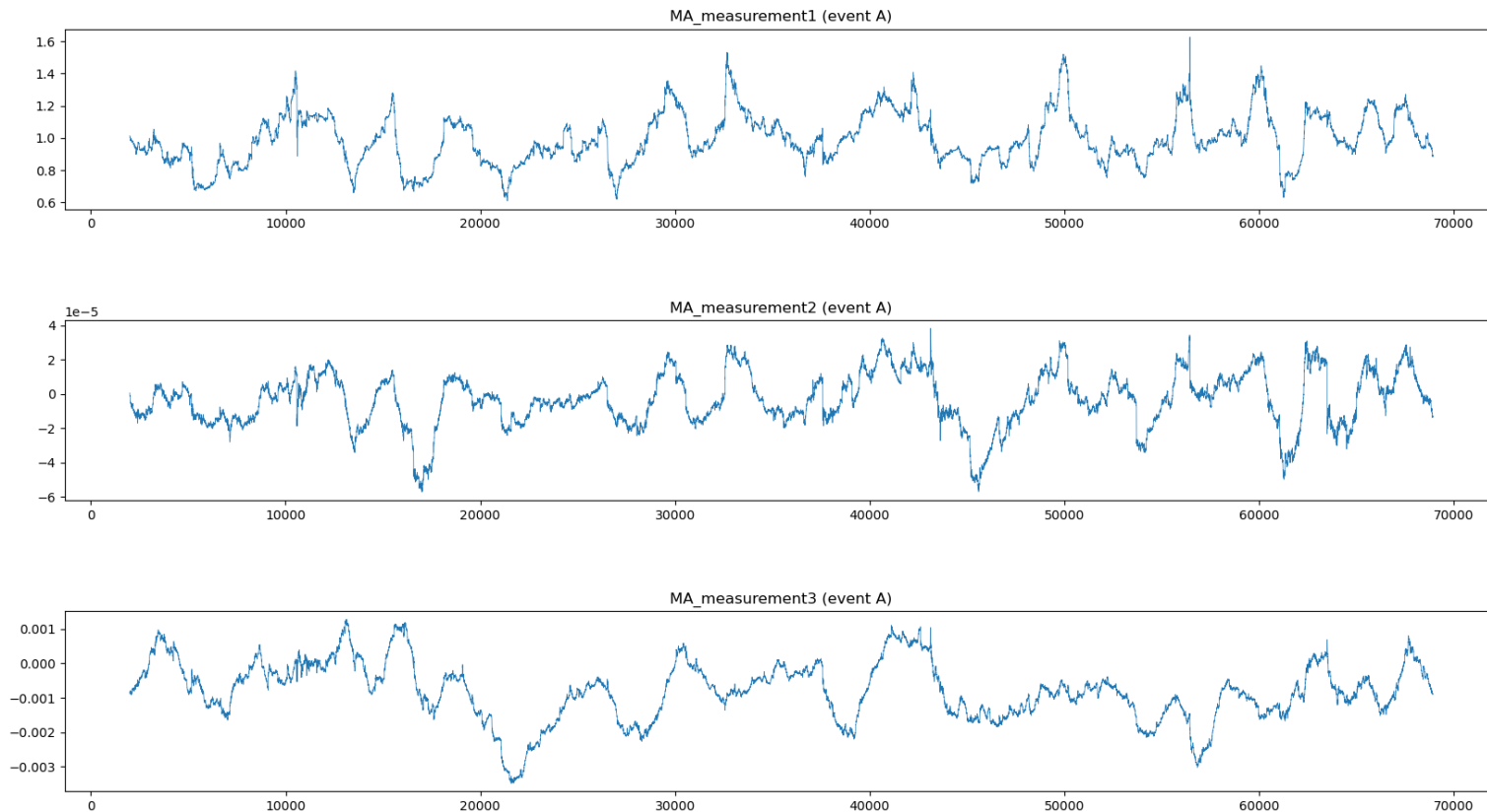
- Noises, ← implicit regularization
- Irreconcilable goals, ← inductive biases and cannot be neglected.

Performance and scale of language models over time



Irreducibility of Time Series Foundation

Time series are noisy.



Irreducibility of Time Series Foundation

There are more irreconcilable goals.

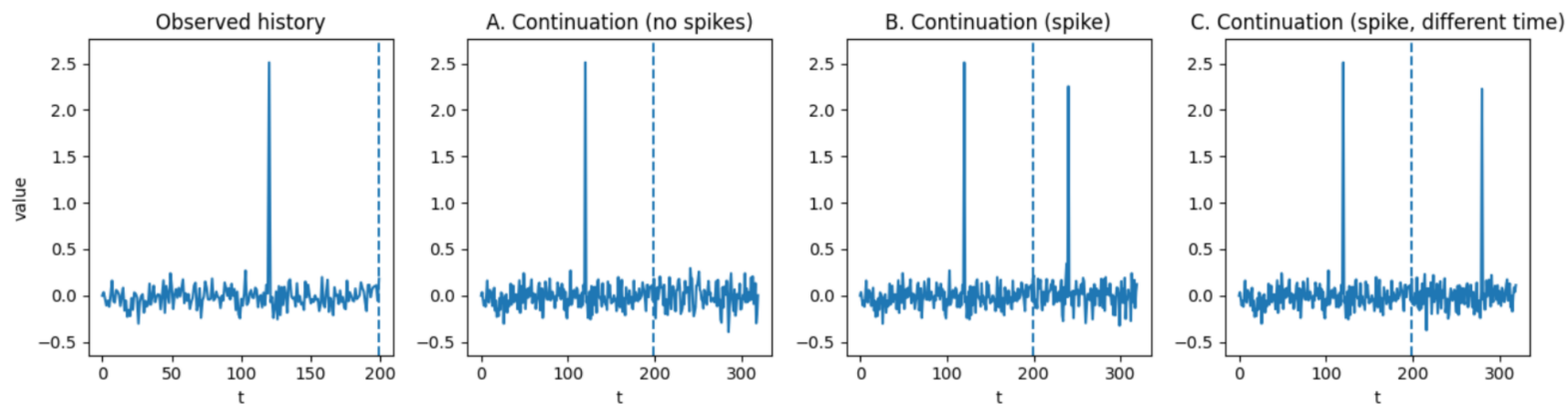
How to continue?

It was the best of times, it was ...

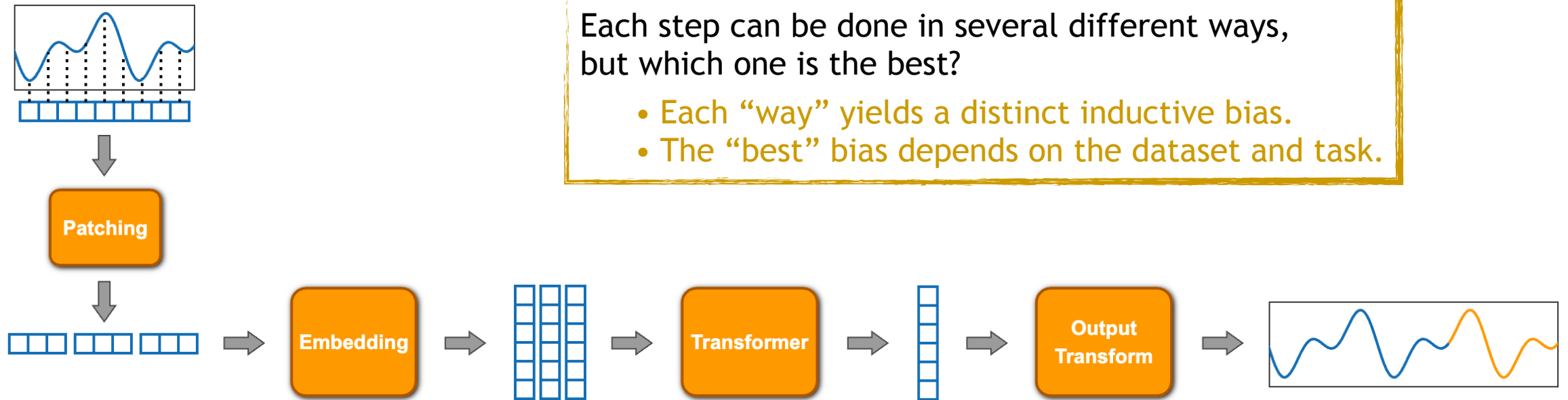
A. the worst of times.

B. a moderately sized refrigerator doing taxes.

C. three left turns and a whispered pineapple.

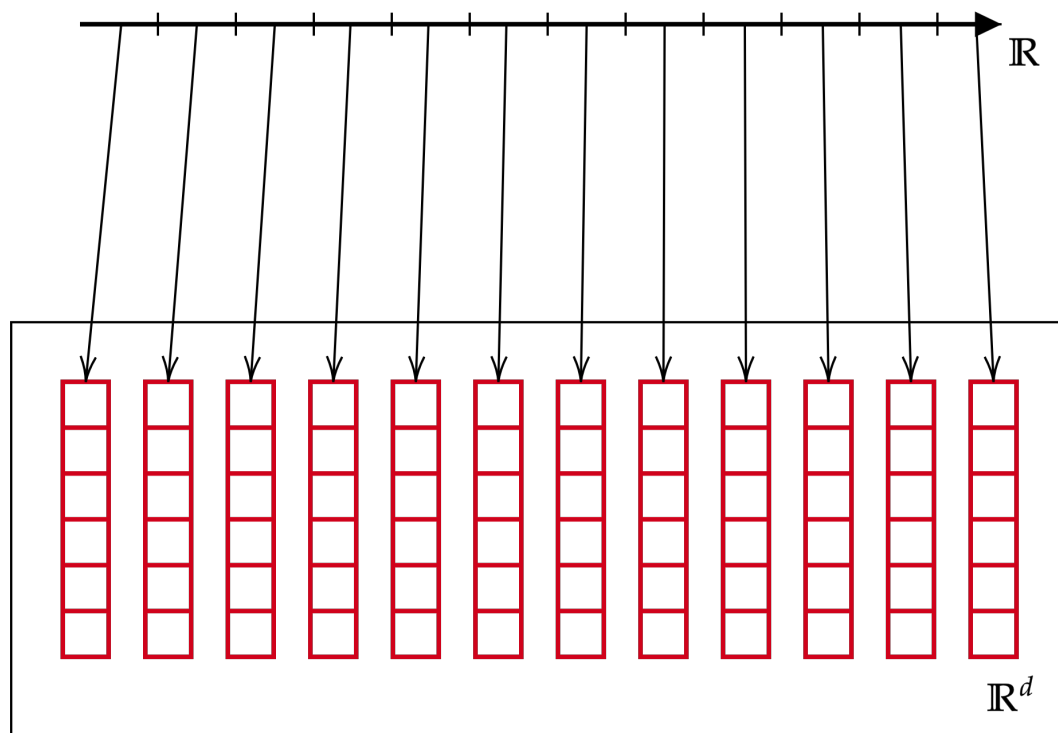


Time Series Foundation Models



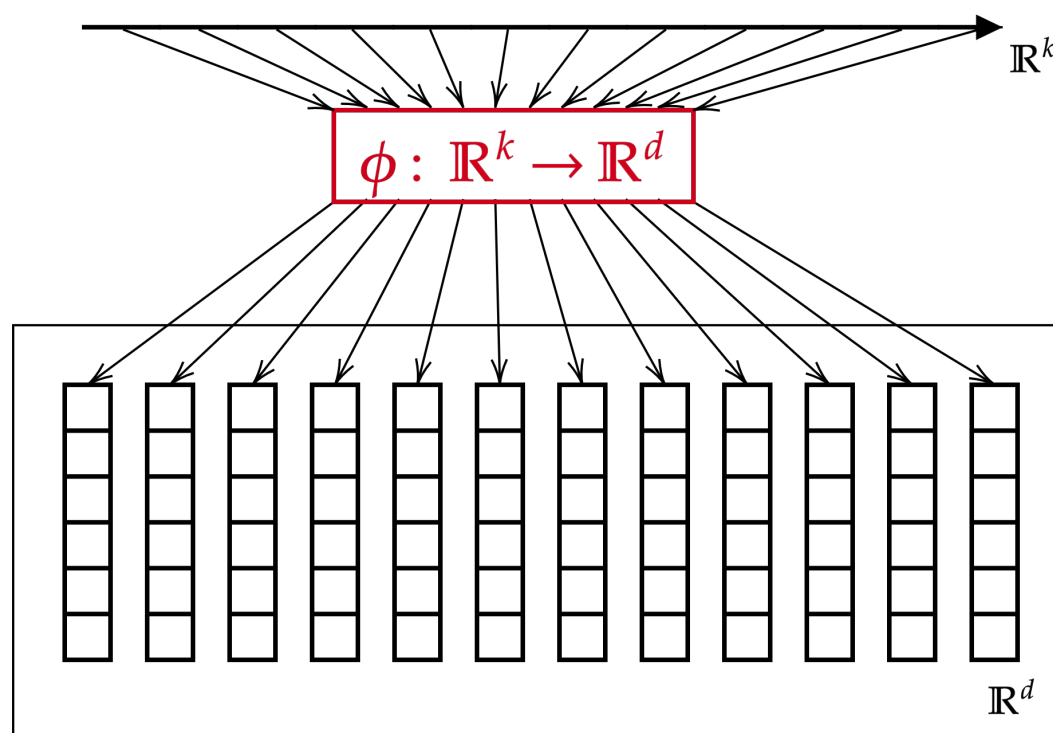
Input Embedding and Geometry Bias

Quantization



- Geometry learned via training

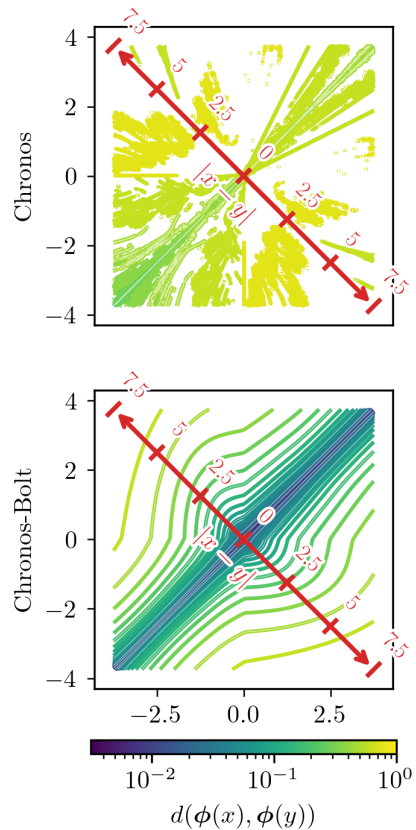
Continuous Embedding



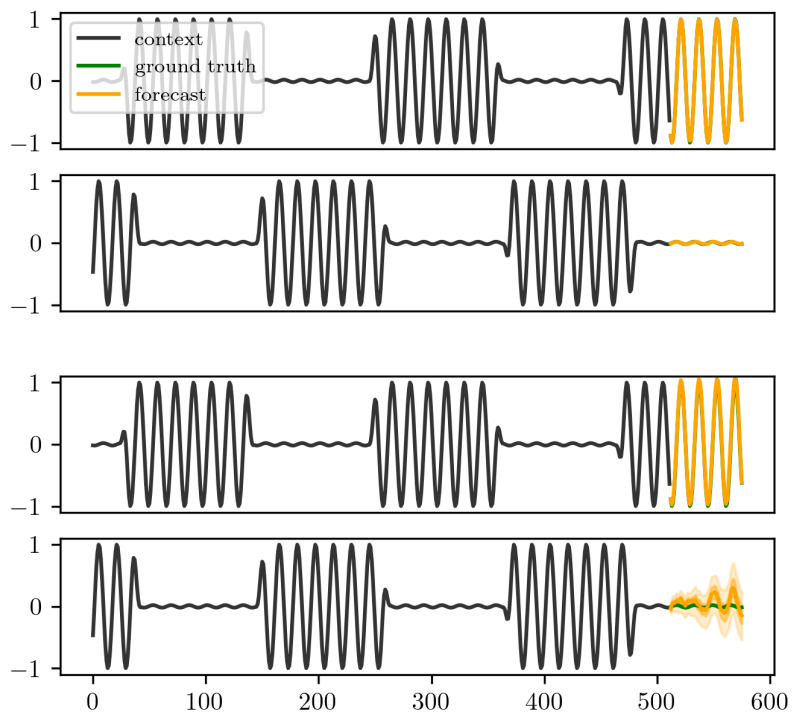
- Geometry naturally preserved by ϕ

Geometry Bias – Multiscale

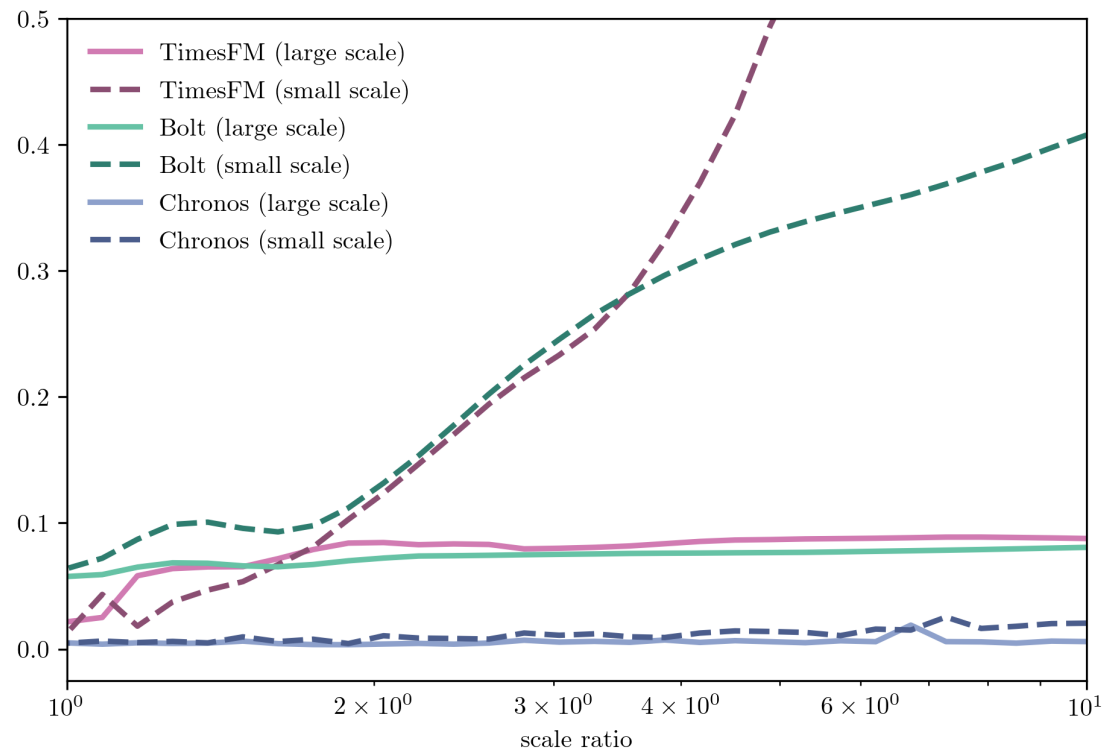
Distance of Embedding



Forecasting a Multiscale Signal

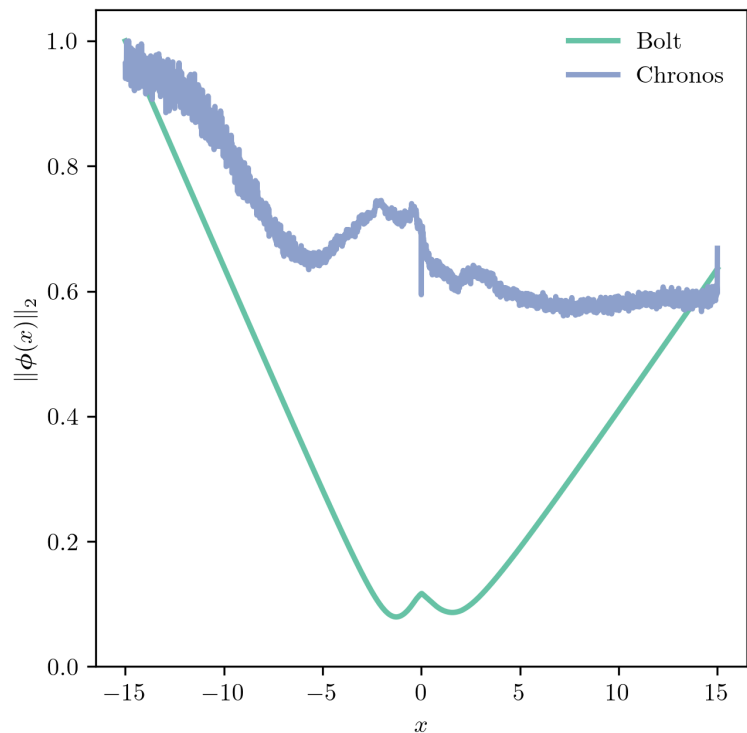


Relative MSE of Forecasting the Two Scales

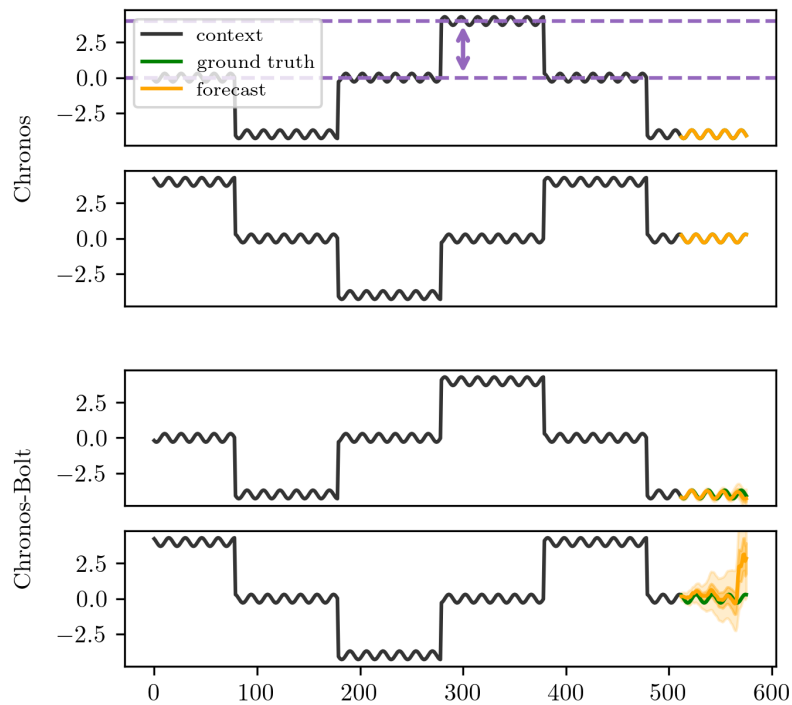


Geometry Bias – Multi-offset

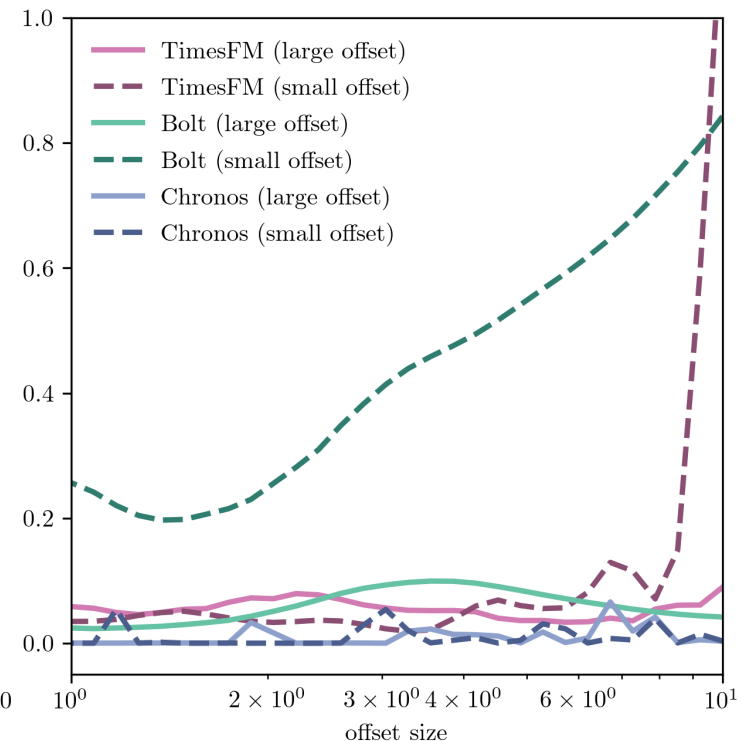
Norm of Embedding



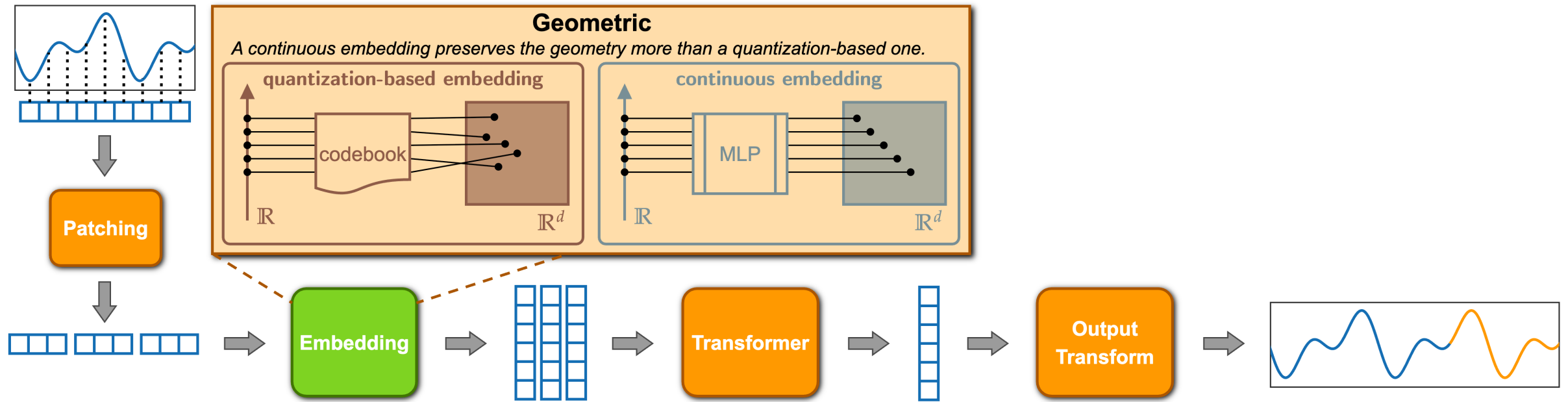
Forecasting a Multi-Offset Signal



Relative MSE of Forecasting the Two Offsets

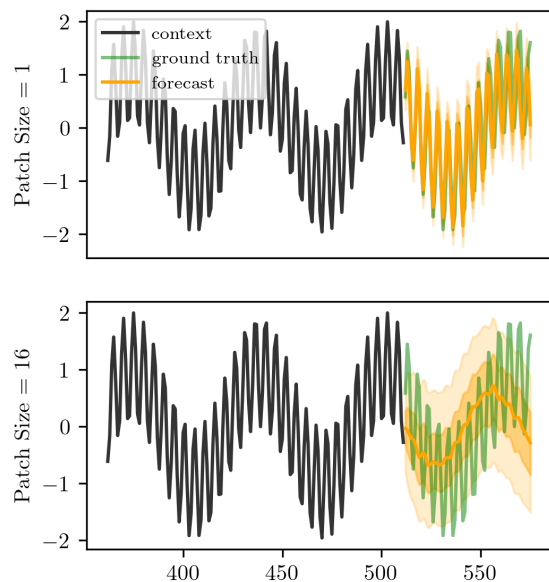


Time Series Foundation Models

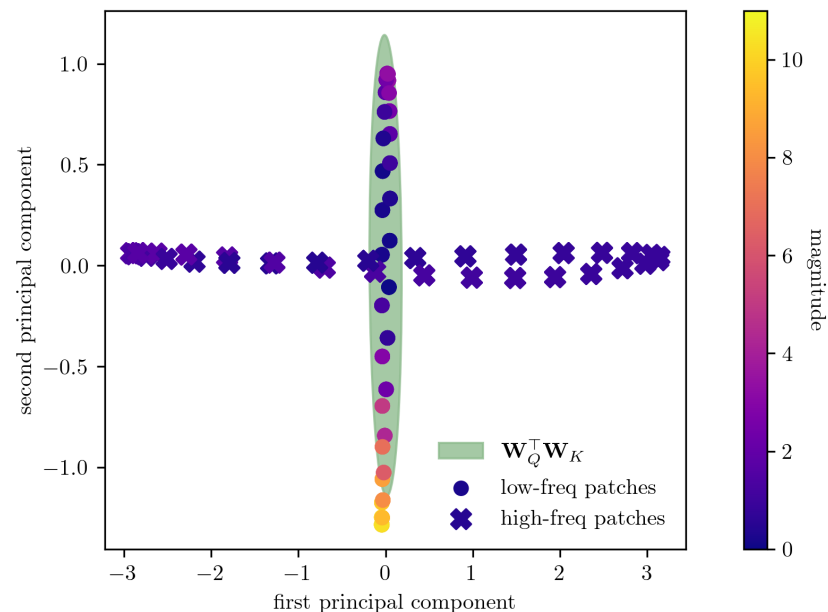


Frequency Bias

Forecast of Chronos-Bolt



Attention Scores of Each Patch



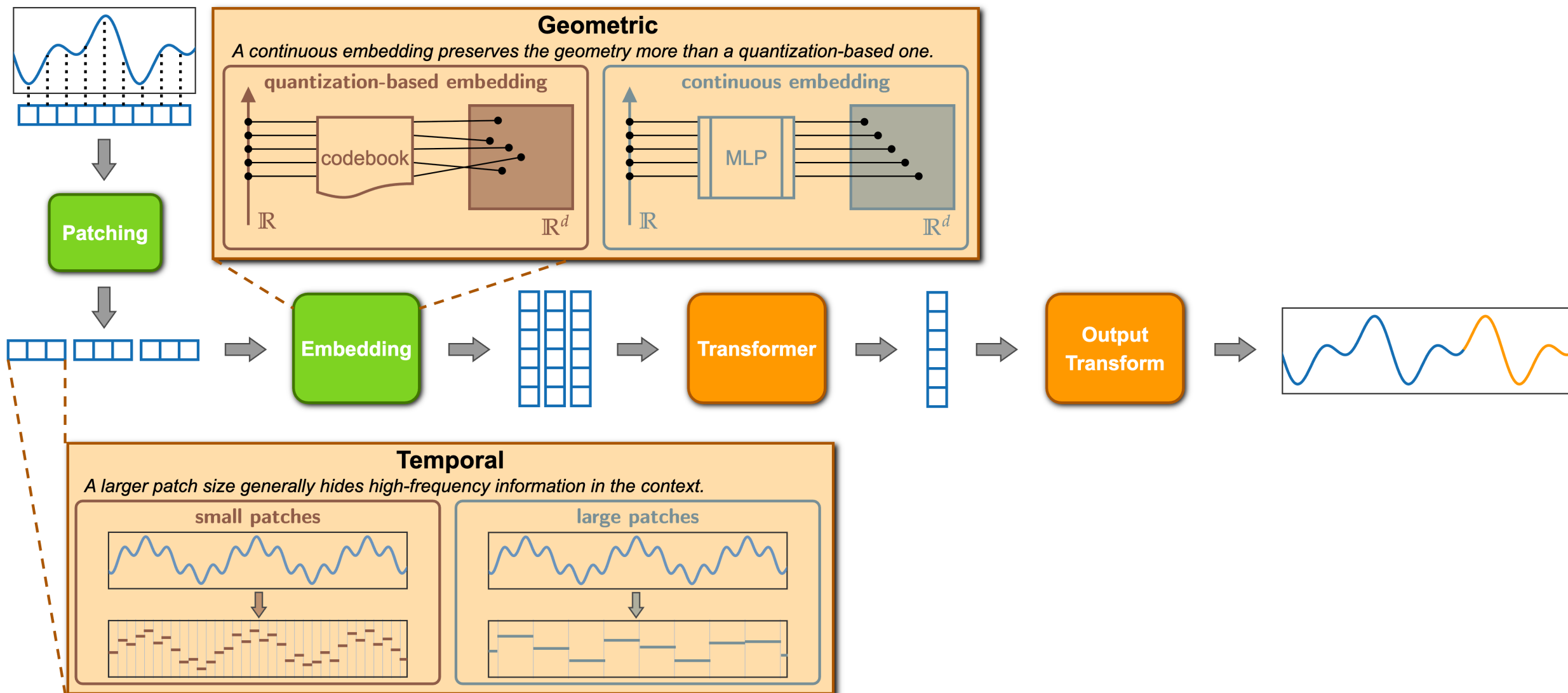
Good News.

Frequency bias prevents a NN from easily fitting high-frequency noises, making it good at generalization.

Bad News.

Frequency bias puts a curse on learning useful high-frequency information in the target.

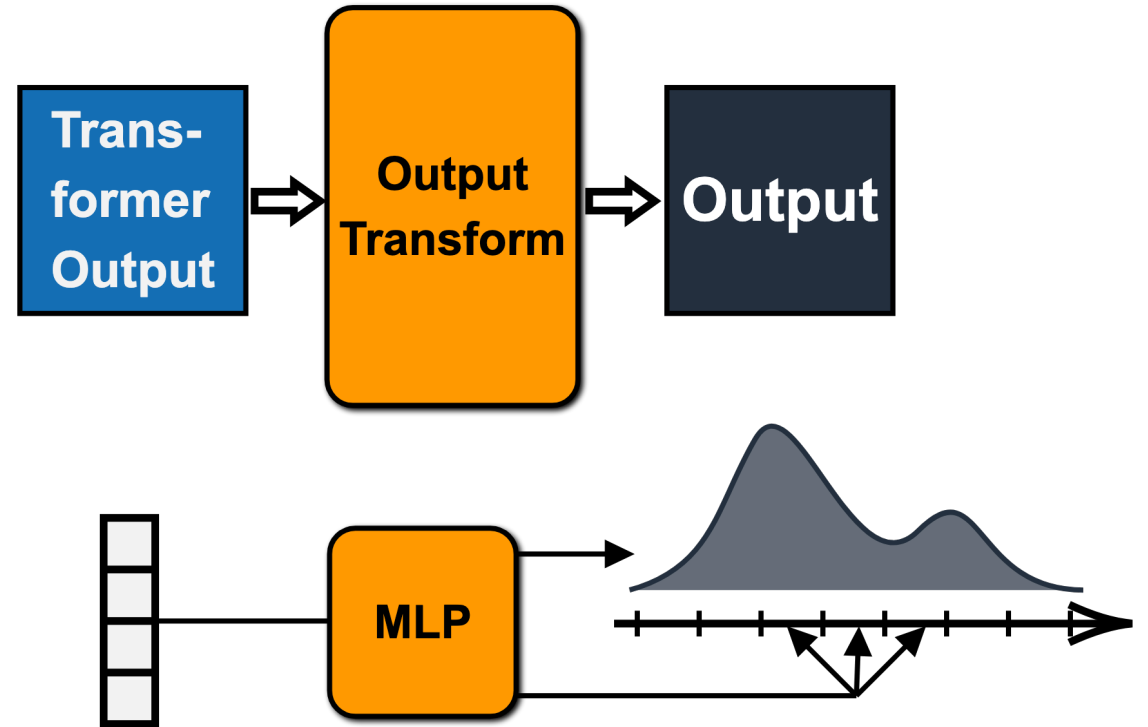
Time Series Foundation Models



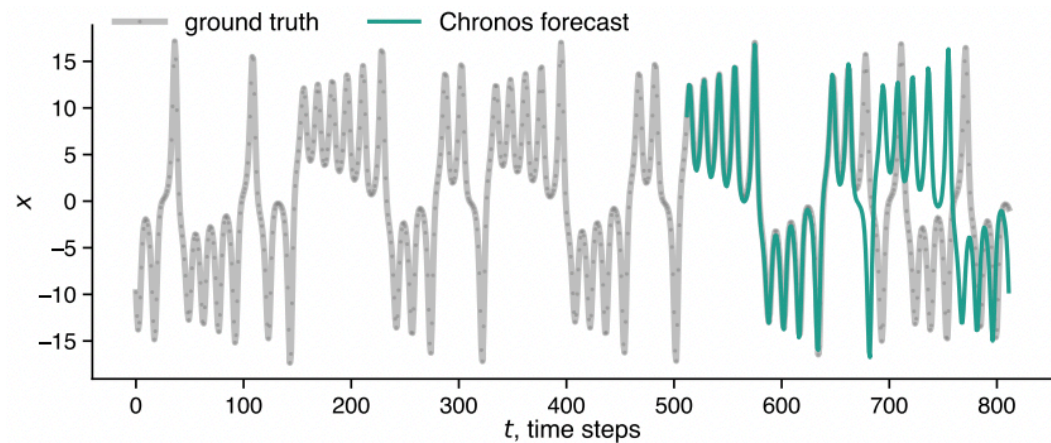
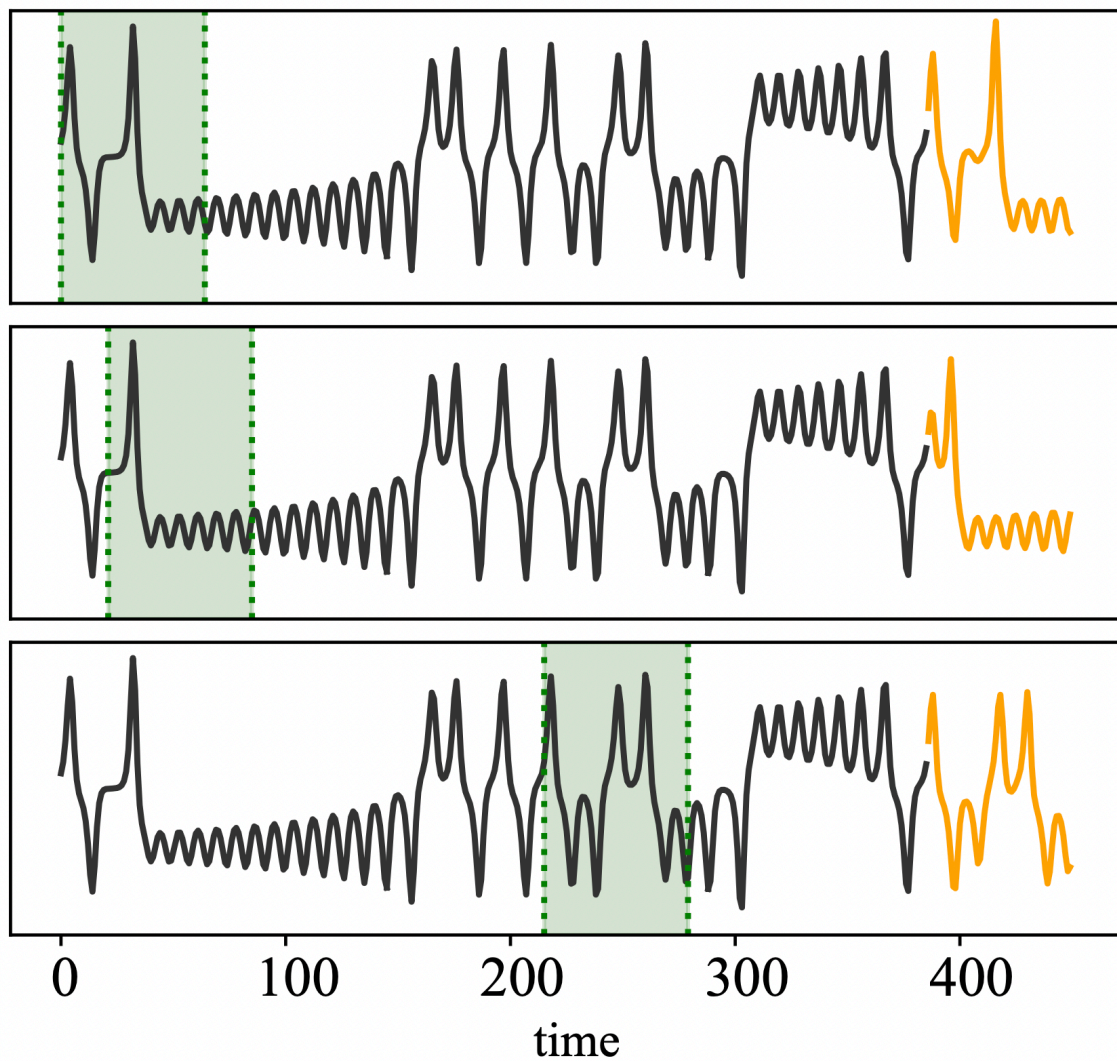
Regression Bias

Training Loss:

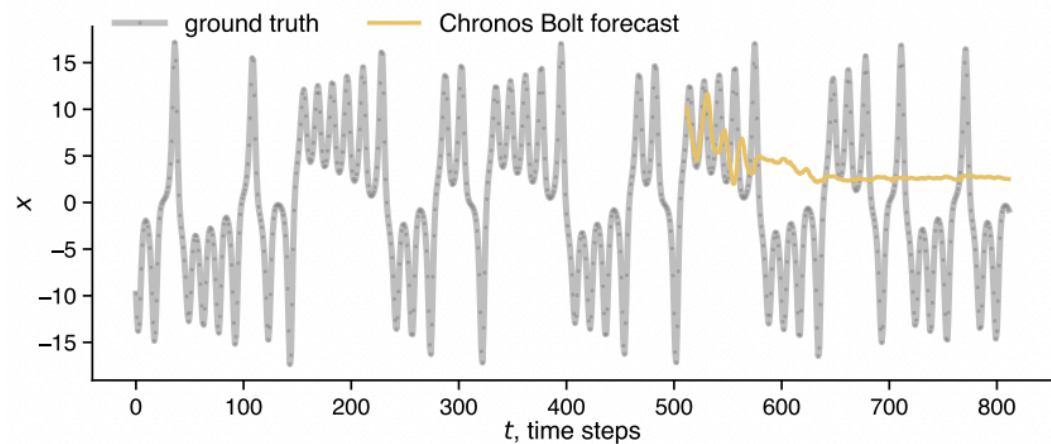
- L^2 -based (regress to mean)
- L^1 -based (regress to median)
- Cross-entropy (regress to mode)



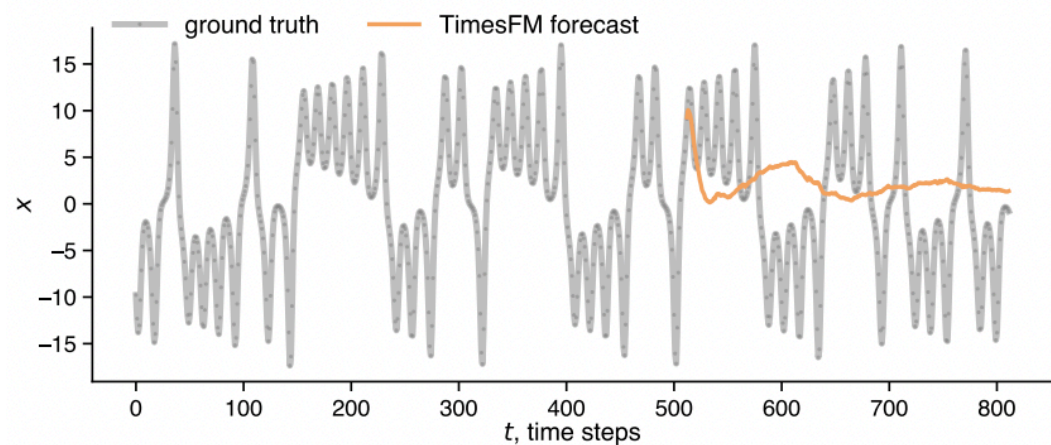
Regression Bias



cross-entropy

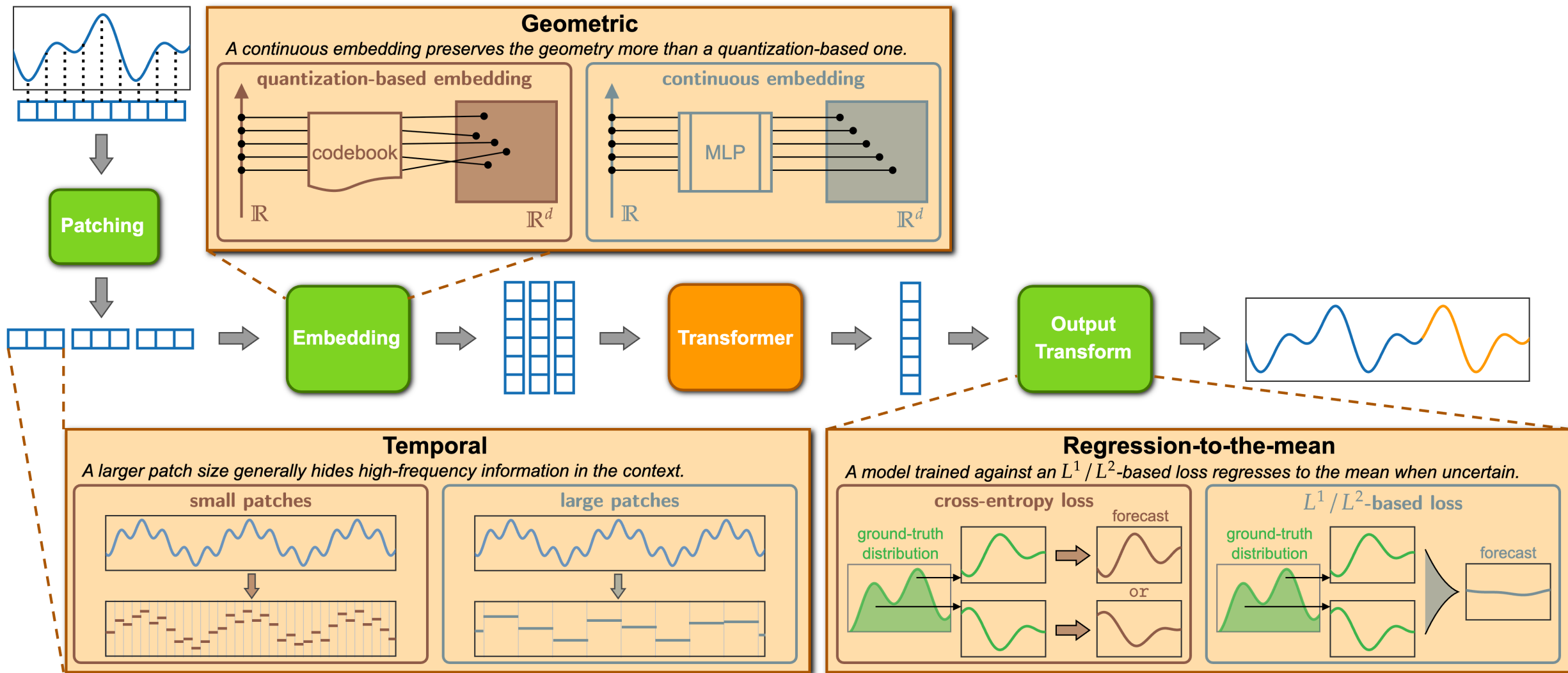


L^1



L^2

Time Series Foundation Models



The Spectrum of Biases

