BERT^{2}S NEURIPS 2025 WORKSHOP

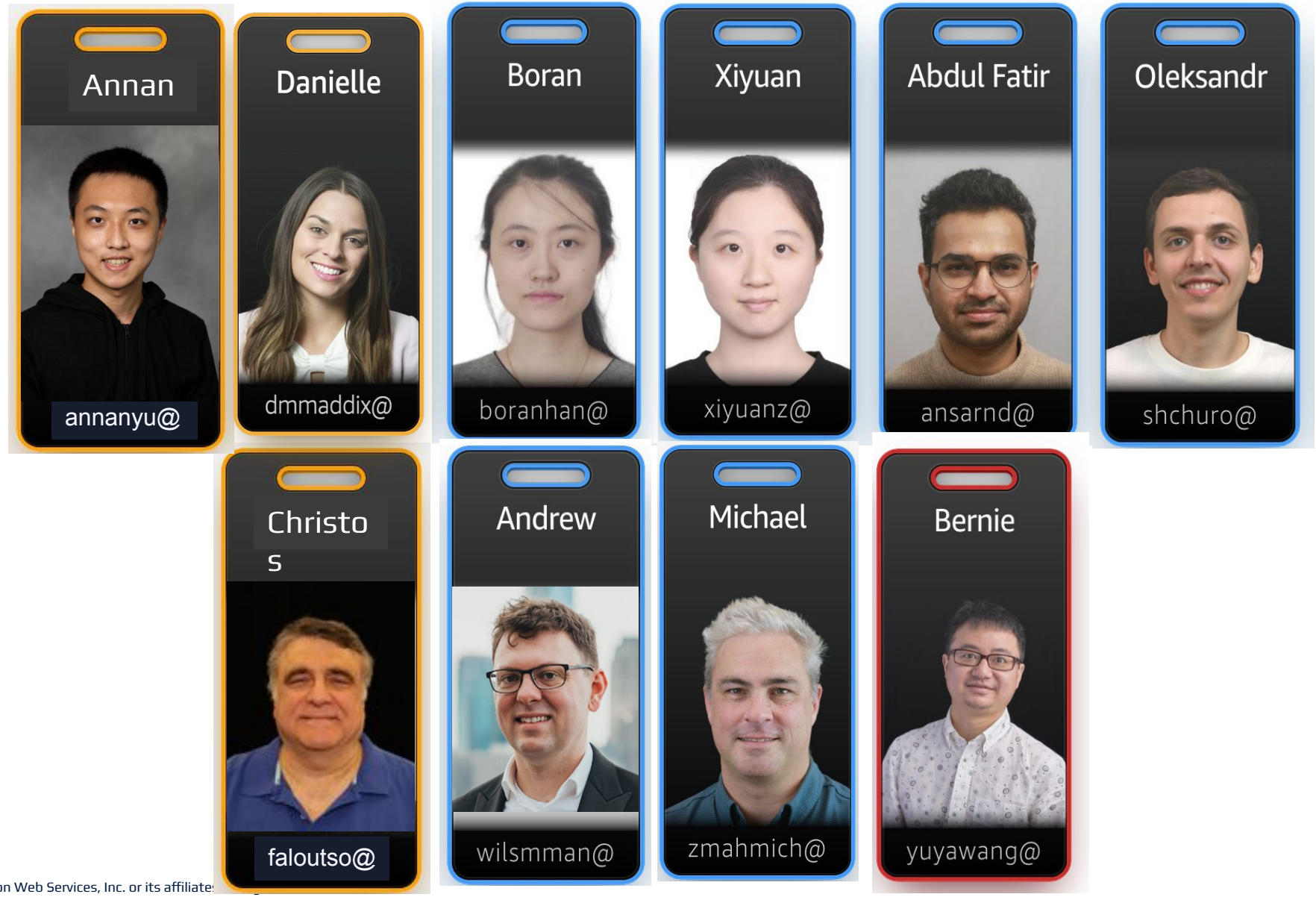
12.07.2025

Understanding the Bitter Lesson in Time Series Foundation Models

Danielle Maddix Robinson

Senior Applied Scientist, AWS AI

Collaborators

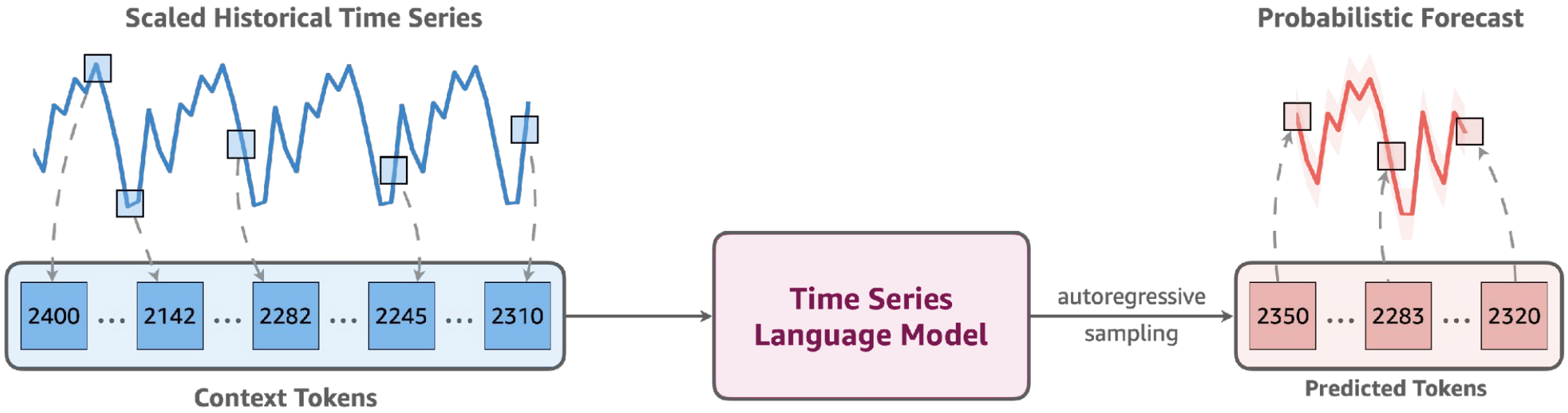


Chronos

A language modeling framework for time series data

that encodes time series into discrete tokens
and trains a language model on them

✔ probabilistic by design
✔ requires no changes to the language model architecture or training procedure



Ansari, A.F., et al., “Chronos: Learning the Language of Time Series”, TMLR, 2024.

Baselines

Pretrained Models

single model used across all tasks

- LLMTime
- ForecastPFN
- LagLlama
- Moirai

CHRONOS
goes here

Task-specific Models

separate model trained/fine-tuned for each task

- PatchTST
- DeepAR
- WaveNet
- TFT
- DLinear
- NBEATS
- NHiTS
- GPT4TS

Local Models

separate model for each time series

- Naive
- SeasonalNaive
- AutoETS
- AutoTheta
- AutoARIMA

Benchmarks

Benchmark I

15 in-domain datasets **for CHRONOS**

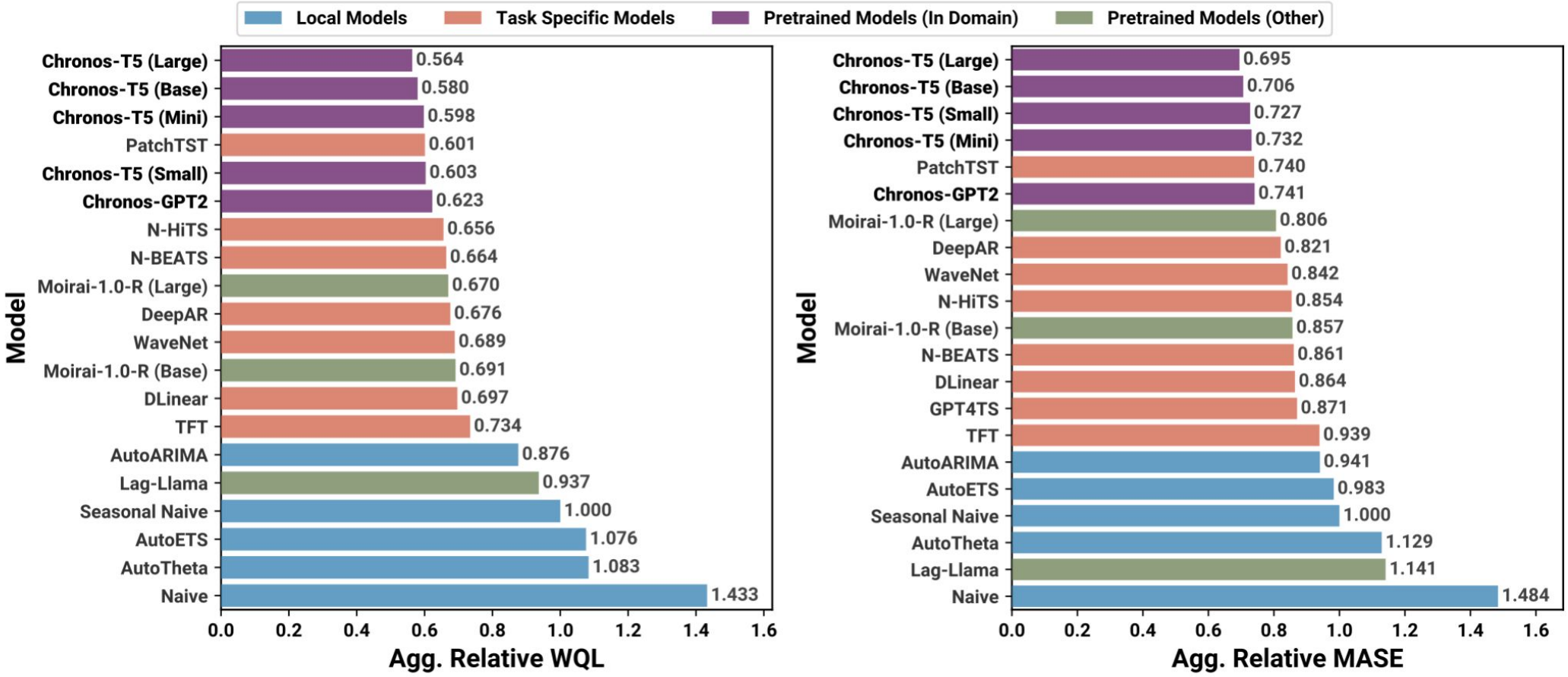
- Electricity (15 Min.)
- Electricity (Hourly)
- Electricity (Weekly)
- KDD Cup 2018
- M4 (Daily)
- M4 (Hourly)
- M4 (Monthly)
- M4 (Weekly)
- Pedestrian Counts
- Taxi (30 Min.)
- Uber TLC (Hourly)
- Uber TLC (Daily)
- Rideshare
- Temperature-Rain
- London Smart Meters

Benchmark II

27 zero-shot datasets **for CHRONOS**

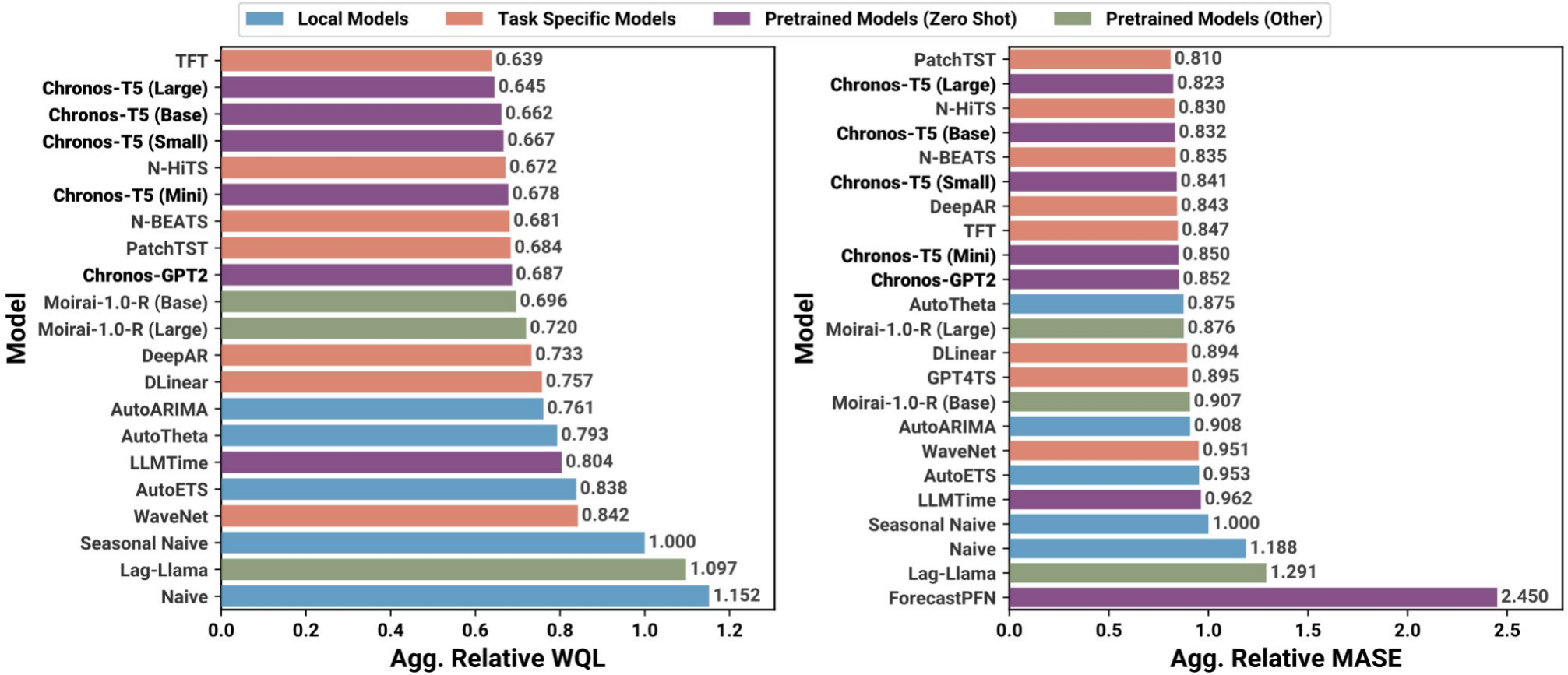
- Australian Electricity
- Car Parts
- CIF 2016
- Covid Deaths
- Dominick
- ERCOT Load
- ETT (15 Min.)
- ETT (Hourly)
- Exchange Rate
- FRED-MD
- Hospital
- M1 (Quarterly)
- M1 (Monthly)
- M1 (Yearly)
- M3 (Monthly)
- M3 (Quarterly)
- M3 (Yearly)
- M4 (Quarterly)
- M4 (Yearly)
- M5
- NN5 (Daily)
- NN5 (Weekly)
- Tourism (Monthly)
- Tourism (Quarterly)
- Tourism (Yearly)
- Traffic
- Weather

Chronos: In-domain Results



In-domain: 15 datasets that were part of the training corpus of **CHRONOS**

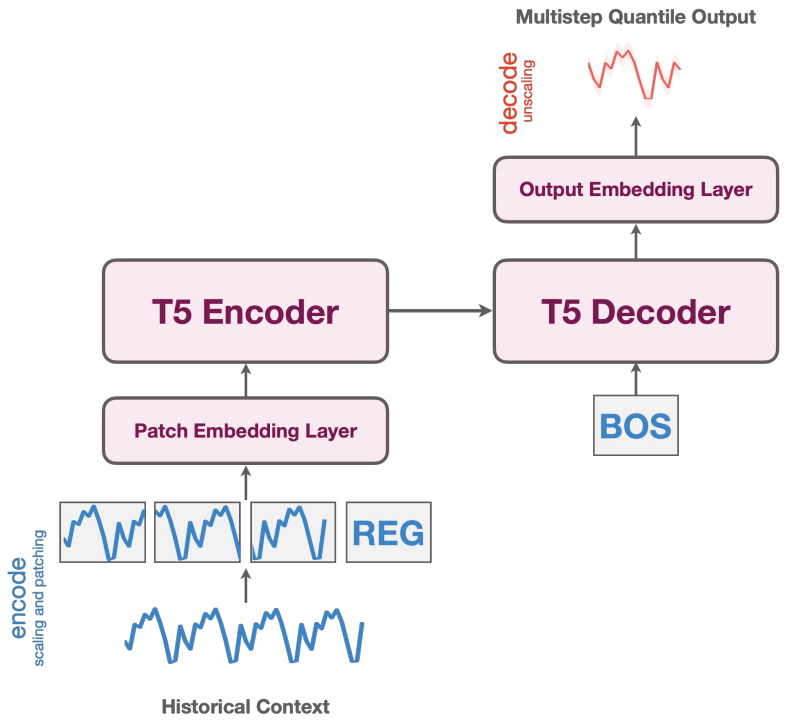
Chronos: Zero-shot Results



Zero-shot: 27 datasets not seen by **CHRONOS** during training

Chronos-Bolt ⚡

More accurate and 250x faster than the original Chronos models



	Chronos-Bolt	Chronos
Input (tokens)	Patches	Individual observations
Output (forecast)	Multi-step quantile forecast	Autoregressive sampling
Loss function	Quantile loss	Cross-entropy loss
Context length	2048	512
Inference device	CPU or GPU	GPU

Ansari, A.F., et al., "Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon", AWS Technical Report, 2025.

Chronos-Bolt⚡: 250x faster than Chronos

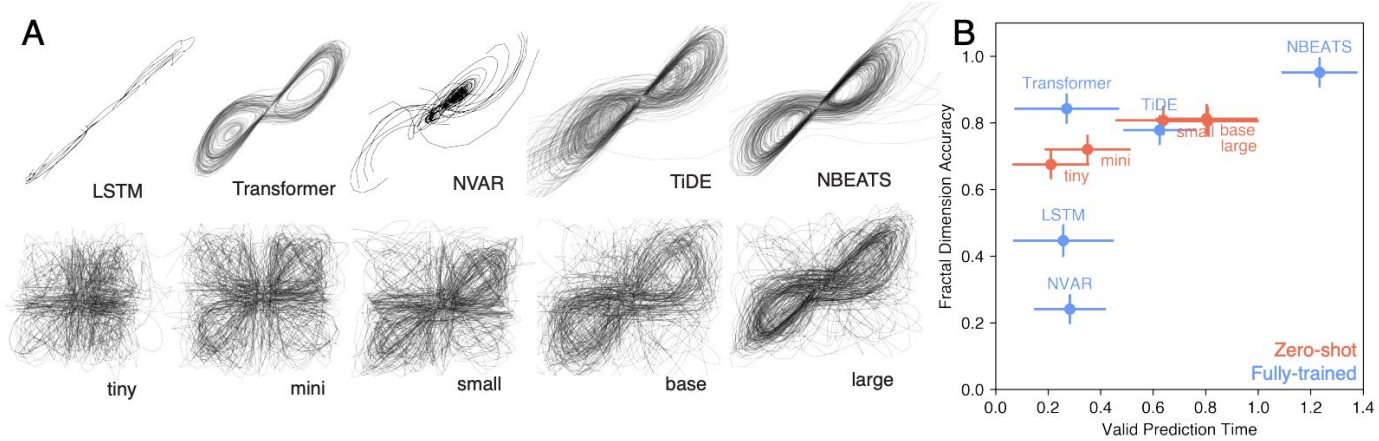
Chronos: zero-shot results

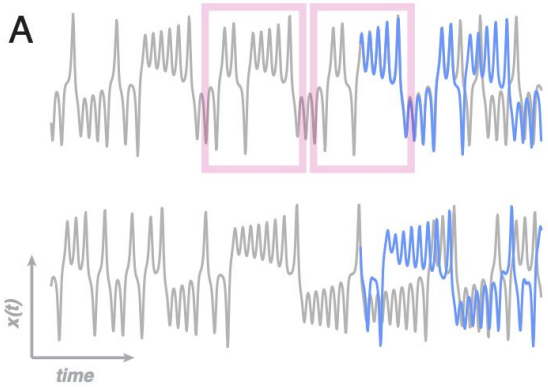
Forecast Error

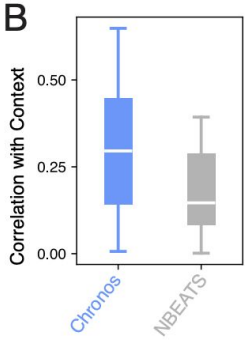
Chronos-Bolt: zero-shot results

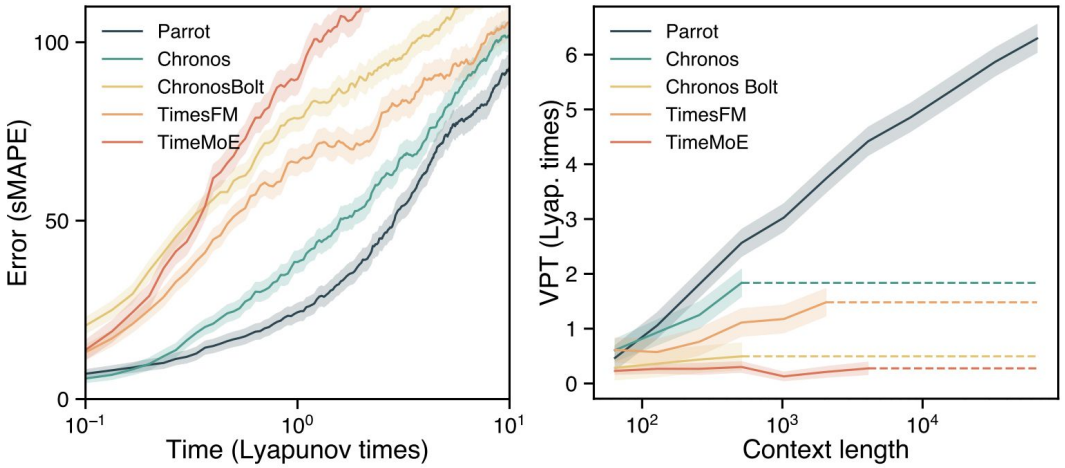
Forecast Error

Long-term Behavior on Chaotic Systems









Zhang, Y. et al. "Zero-shot Forecasting of Chaotic Systems," ICLR, 2025.
Zhang, Y. et al., "Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning", arXiv preprint arXiv:2505.11349, 2025.

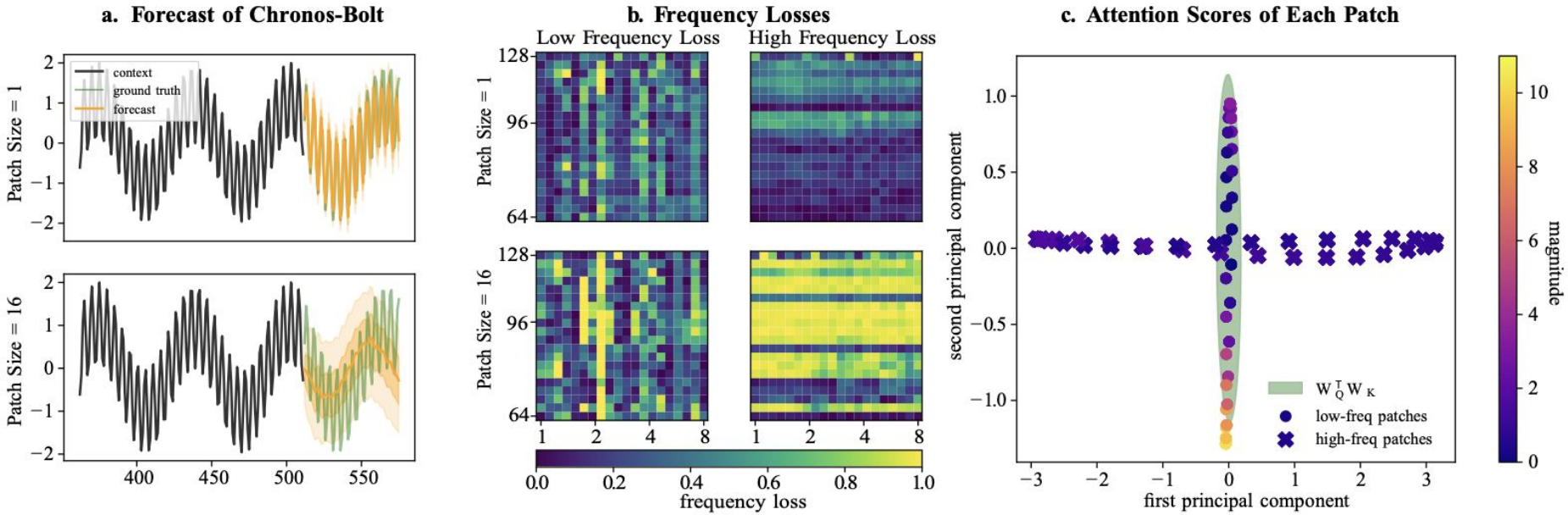
Design Choices of TSFMs

Yu, A. et al., “Understanding the Implicit Biases of Design Choices for Time Series Foundation Models”, arXiv preprint arXiv:2510.19236, Under Review, 2025.

Inductive Biases Overview

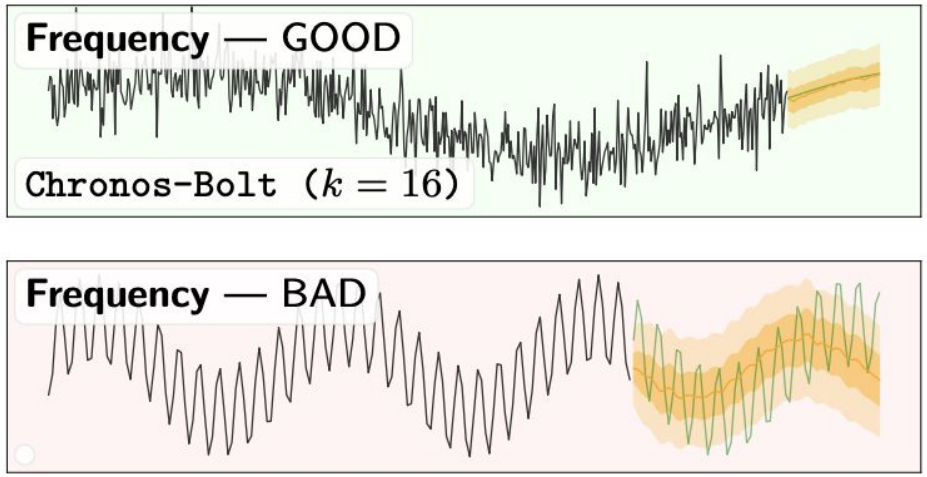
How Do TSFMs Learn Time?

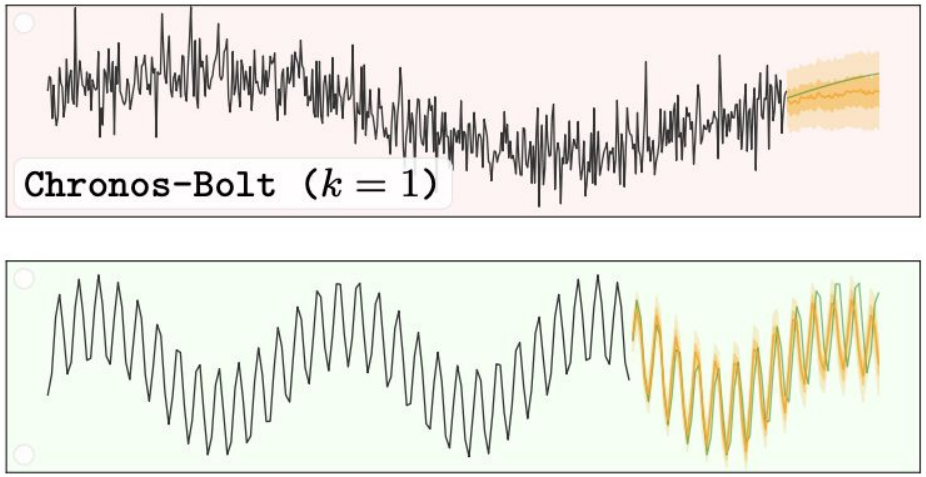
Temporal Frequency Bias



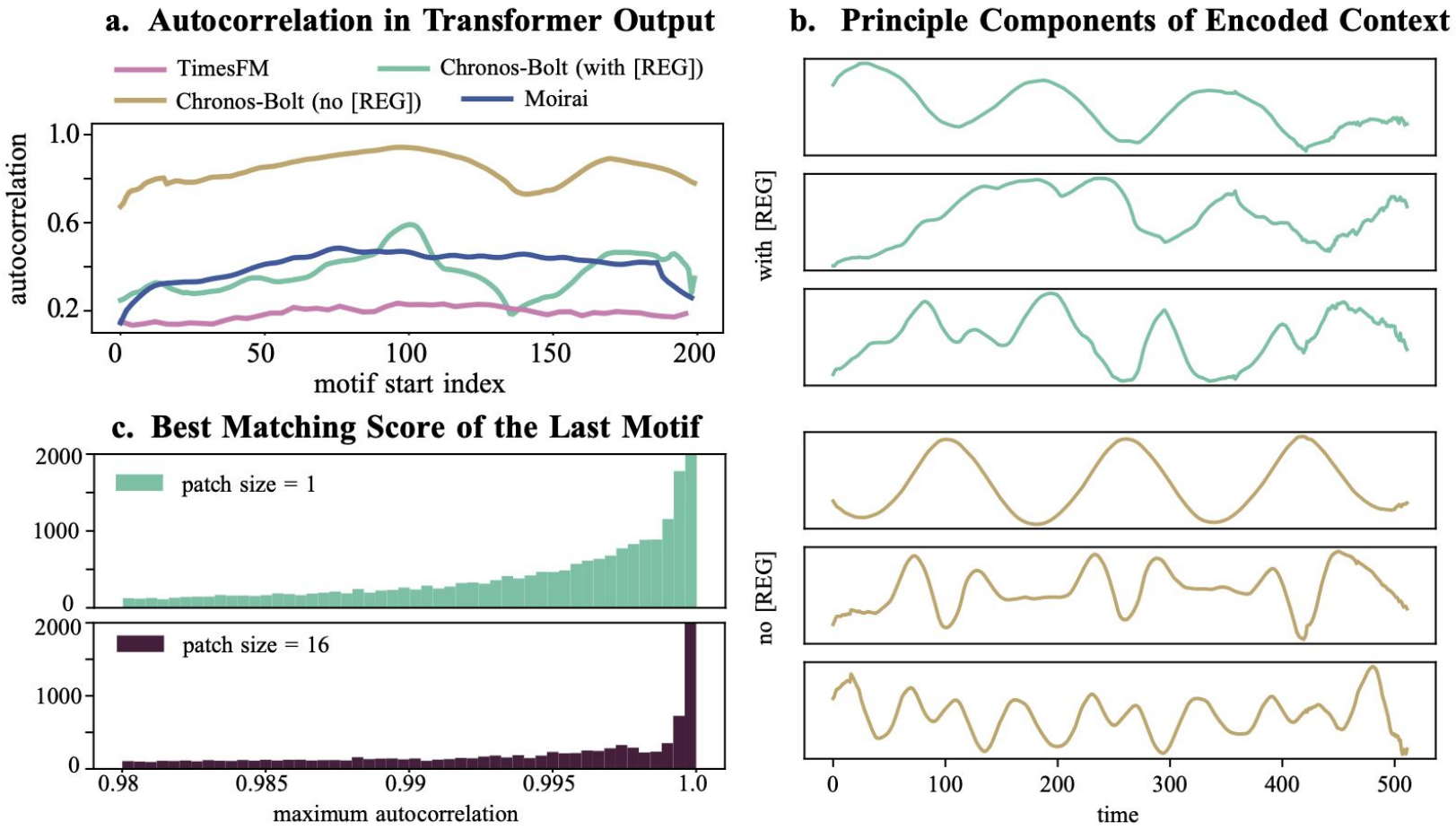
- Models with larger patch size k captures only the low frequency mode
- Chronos-Bolt $k = 16$ fails at capturing high-frequency information
- Attention scores are heavily dominated by the low-frequency patches

Frequency Bias: Good or Bad?



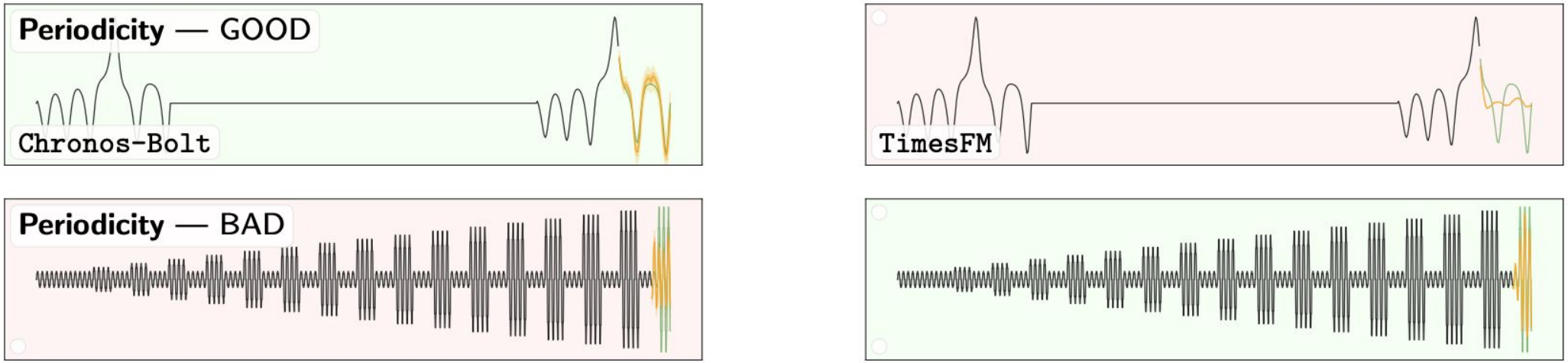


Temporal Periodicity Bias



Controlled by alignment of patch size k with underlying recurrent motifs

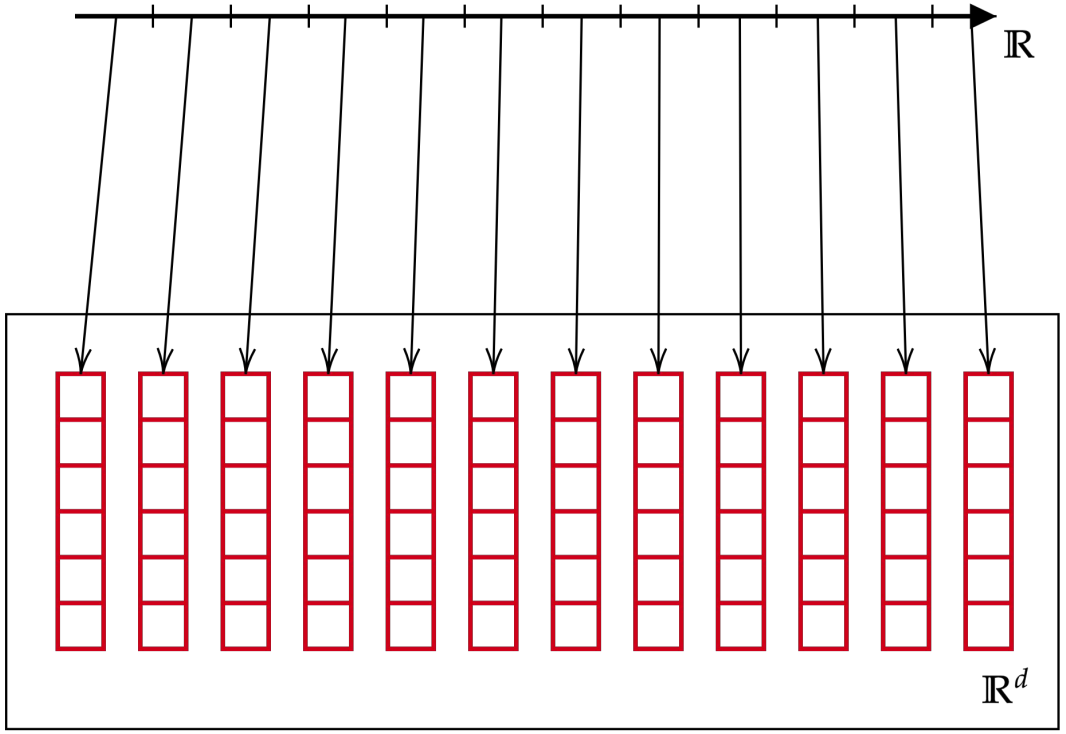
Periodicity Bias: Good or Bad?



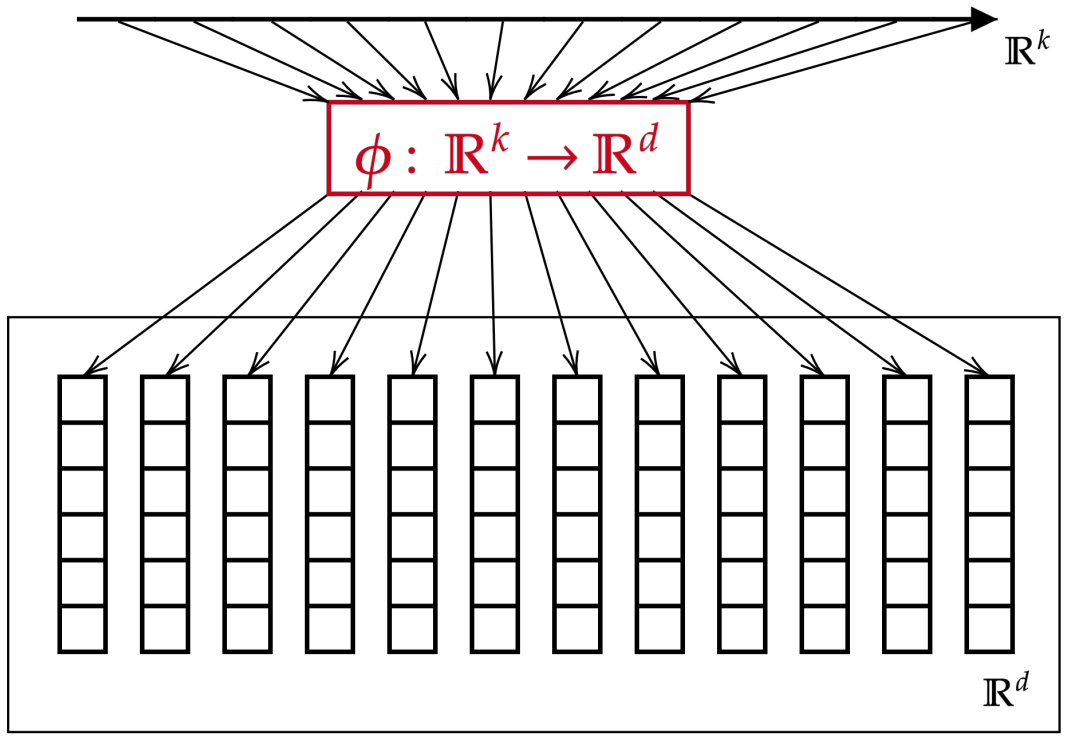
How Do TSFMs Learn Geometry?

Design Choice: Embedding Type

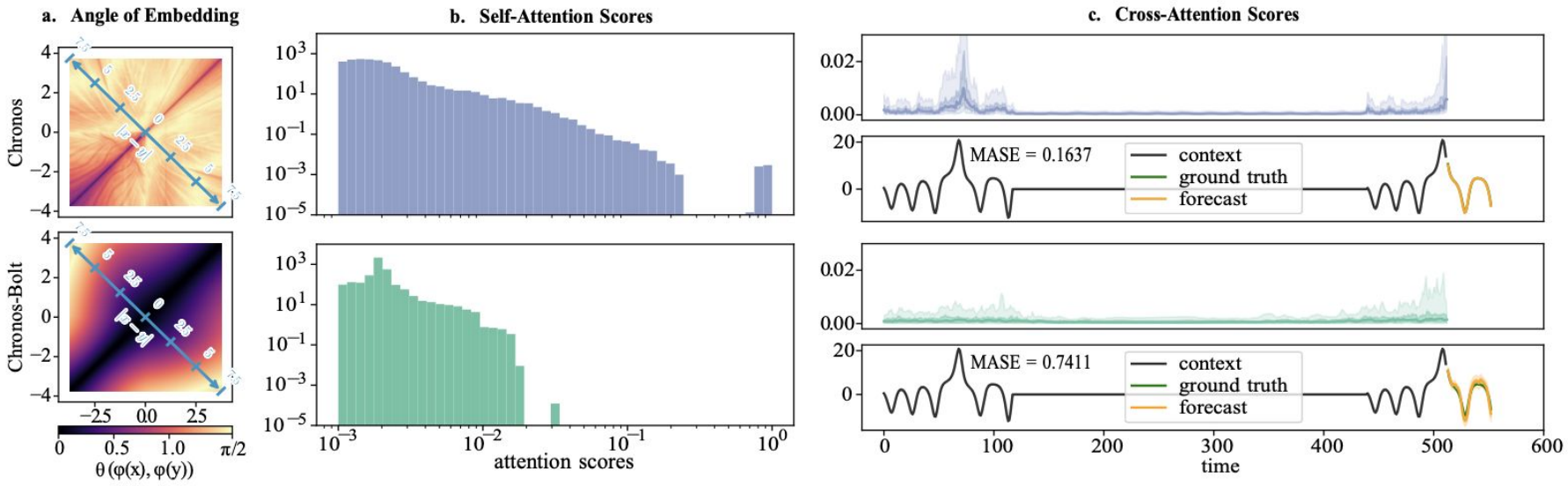
Quantization



Continuous Embedding

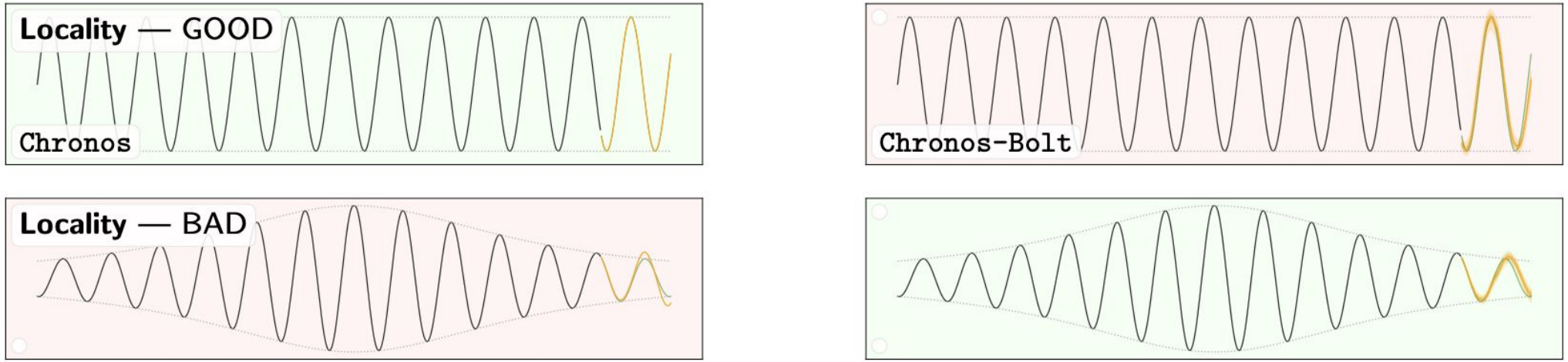


Geometric Angular Bias

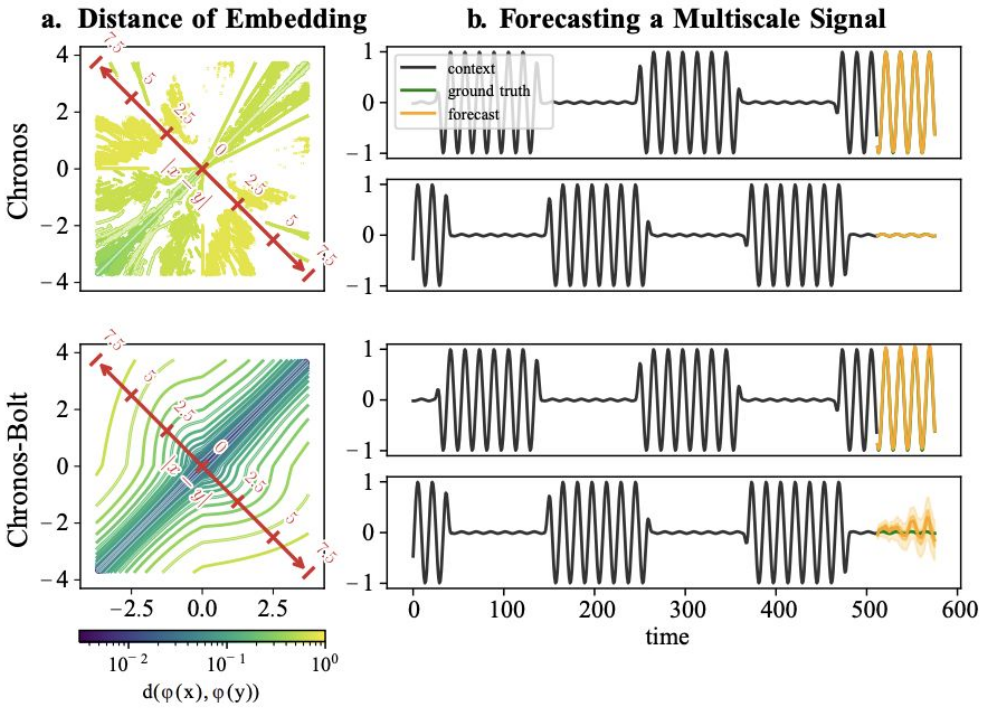


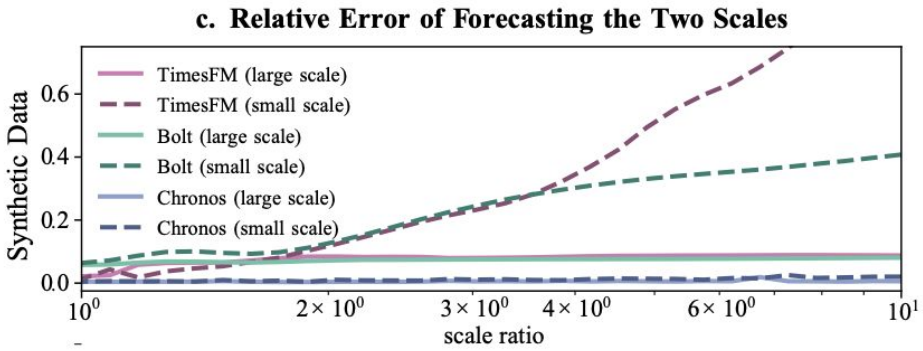
- Attention scores of Chronos are bimodal - tokens place high attention weights on their neighbors (locality)
- Attention scores of Chronos-Bolt are more evenly distributed (global)
- Chronos achieves a significantly lower MASE than Chronos-Bolt on a context formed by repeating a motif of a chaotic system by its "parroting"

Angular Bias: Good or Bad?



Geometric Distance Bias



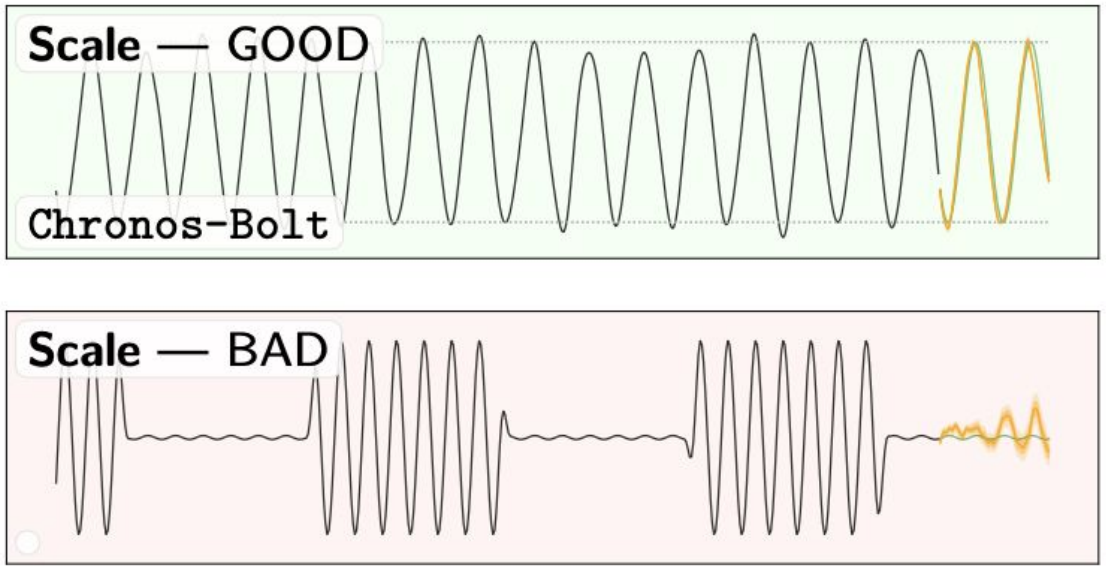


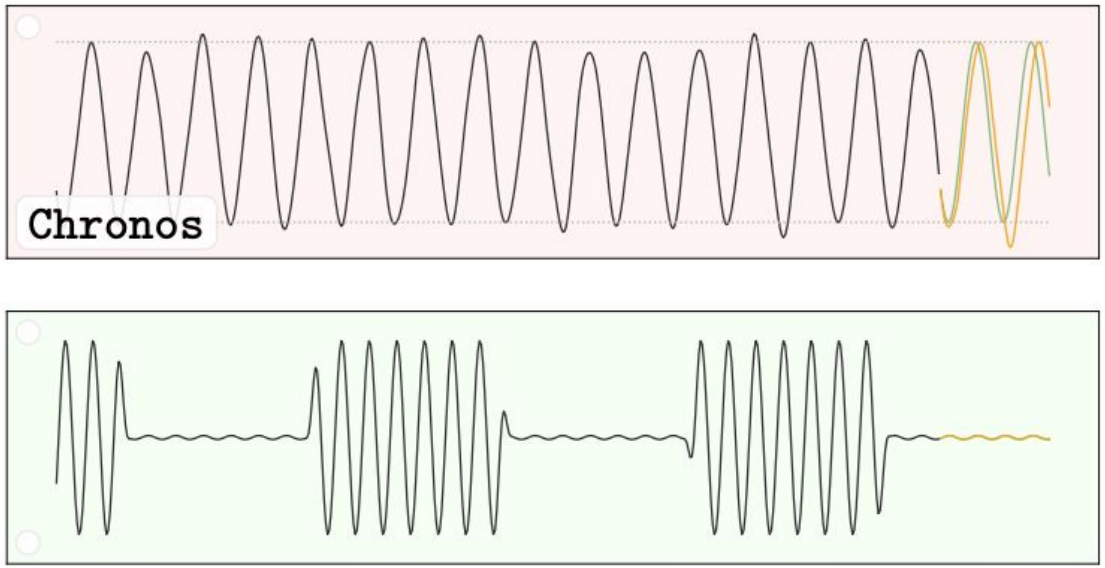
Real-World Data

	scale ratio	1	4/3	2	4	8	40/3	20	40
large scale	Chronos	1.0000	1.0057	1.0088	1.0191	1.0215	1.0234	1.0327	1.0382
	Bolt	1.0000	1.0047	1.0063	1.0140	1.0218	1.0271	1.0333	1.0341
	TimesFM	1.0000	1.0031	1.0045	1.0091	1.0150	1.0197	1.0224	1.0266
	Moirai	1.0000	1.0014	1.0050	1.0101	1.0227	1.0299	1.0329	1.0401
small scale	Chronos	1.0000	1.0291	1.0133	1.0755	1.1882	1.2749	1.3602	1.5559
	Bolt	1.0000	1.0815	1.1098	1.2207	1.3687	1.5538	1.7256	2.6792
	TimesFM	1.0000	1.1293	1.1622	1.3805	1.5120	1.8436	2.0182	3.2009
	Moirai	1.0000	1.0742	1.1131	1.1922	1.3186	1.4866	1.7705	2.4818

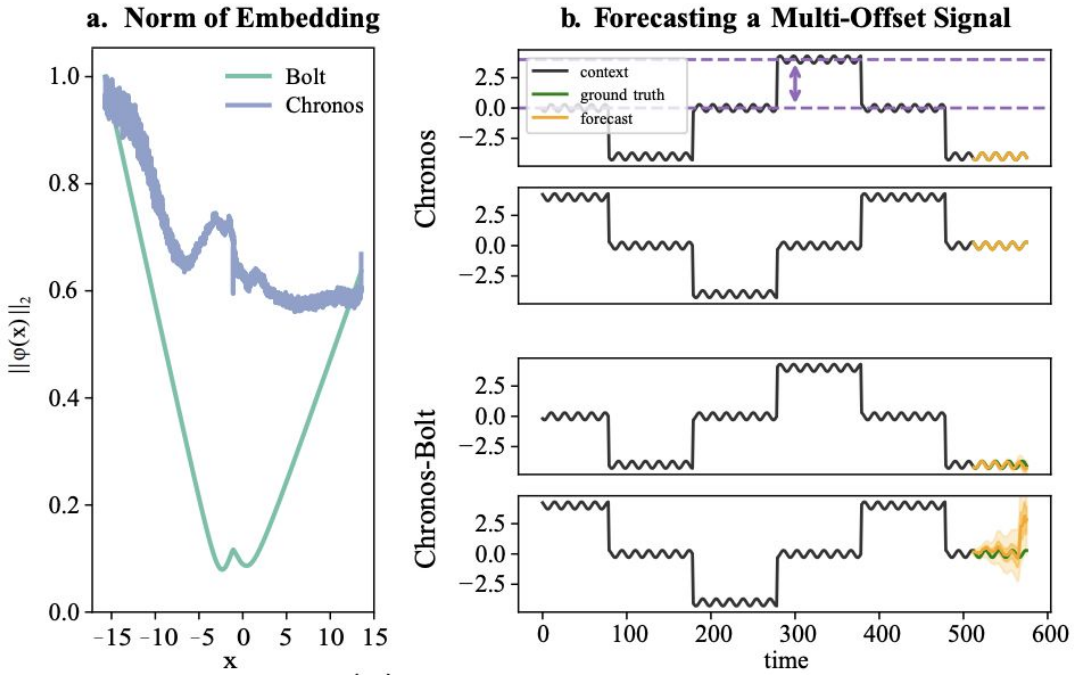
- Quantization-based embedding in Chronos magnifies smaller scales, which makes nearby numbers appear more distinct in hidden space
- Continuous-embedding in Chronos-Bolt maps nearby numbers to nearby vectors in hidden space
- For multi-scale structure, Chronos is more sensitive and better at learning fine-scale patterns than Chronos-Bolt

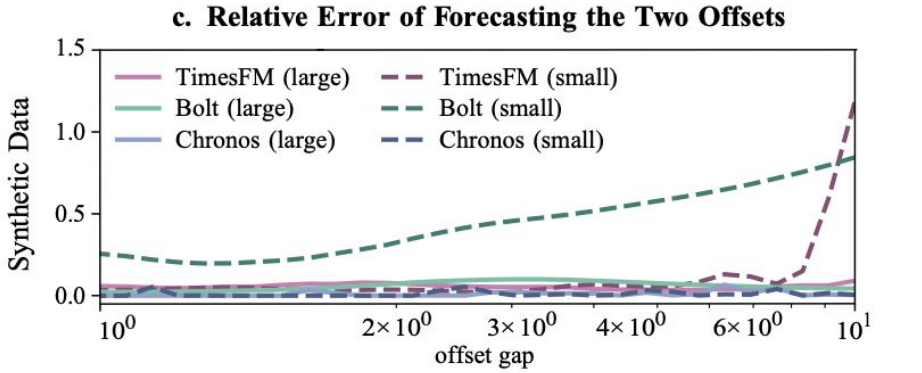
Distance Bias: Good or Bad?





Geometric Norm Bias



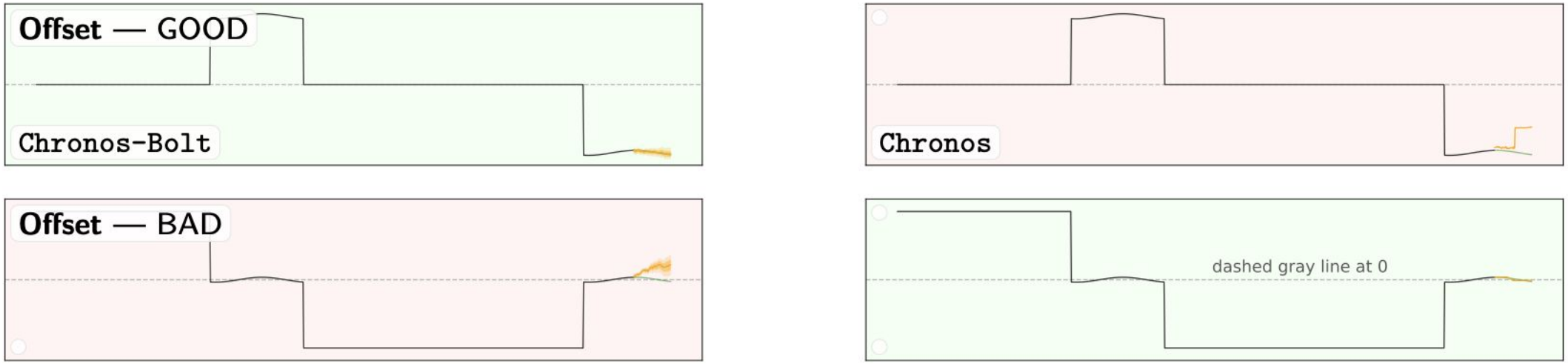


Real-World Data

	offset gap	1	4/3	2	4	8	40/3	20	40
high offset	Chronos	1.0000	1.0023	1.0047	1.0111	1.0145	1.0134	1.0227	1.0282
	Bolt	1.0000	1.0025	1.0045	1.0039	1.0079	1.0125	1.0199	1.0321
	TimesFM	1.0000	1.0019	1.0008	1.0048	1.0090	1.0121	1.0179	1.0277
	Moirai	1.0000	1.0090	1.0059	1.0099	1.0177	1.0241	1.0250	1.0328
low offset	Chronos	1.0000	1.1151	1.1032	1.0690	1.1078	1.1729	1.3423	1.6219
	Bolt	1.0000	1.1138	1.1672	1.2192	1.3566	1.6279	2.1845	3.1620
	TimesFM	1.0000	1.1781	1.2995	1.3808	1.4780	1.8239	2.6884	3.4165
	Moirai	1.0000	1.2196	1.3142	1.4234	1.5517	1.9059	2.8262	3.6152

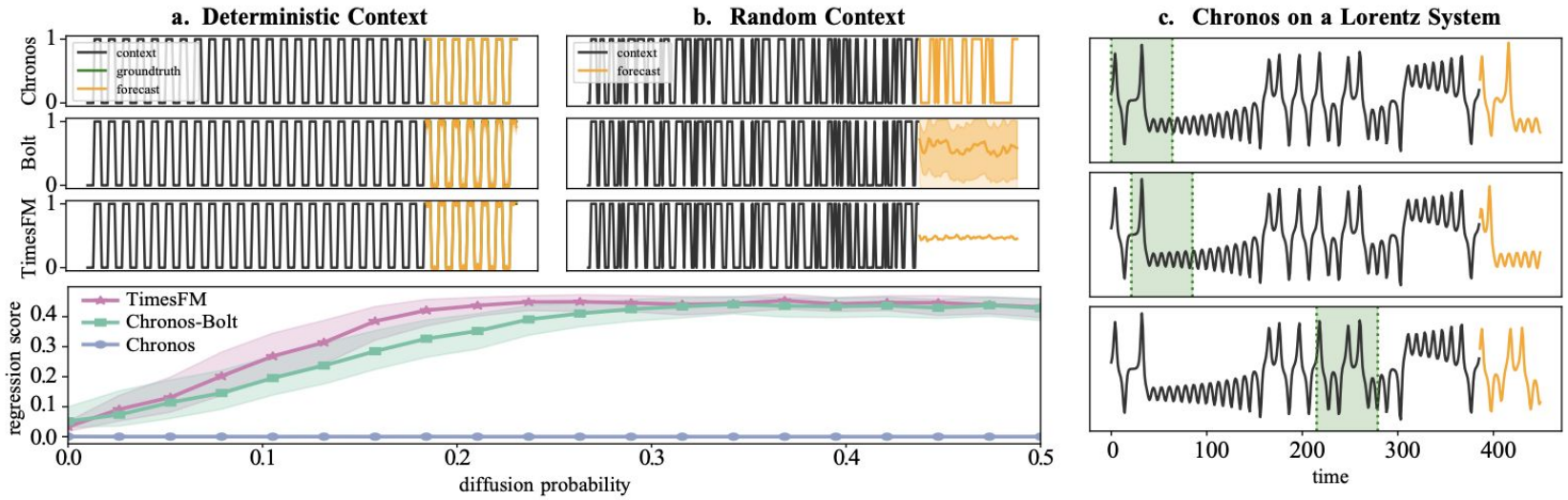
Chronos-Bolt struggles to forecast the near-zero period in the multi-offset time series

Norm Bias: Good or Bad?



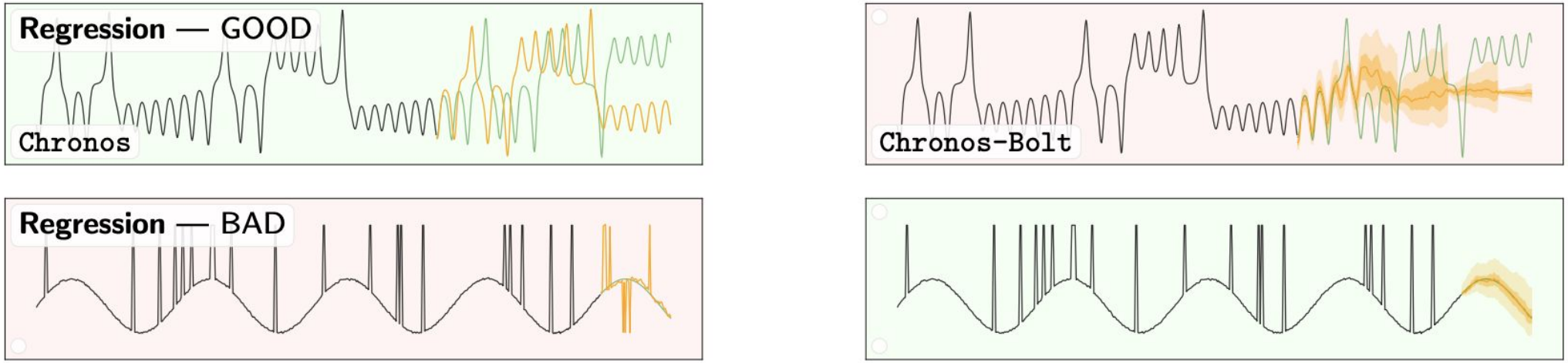
How Do TSFMs Regress to the Mean?

Regression-to-the-Mean Bias



- Models trained with L^2 or L^1 regression losses regress to mean or median
- Chronos with cross-entropy loss models full probability distribution and settles on a mode
- In chaotic systems, fractal dimension is measurement of long-term geometry of the trajectories and regressing to mean/median can severely damage it
- Chronos "parrots" three distinct outcome branches from the context of a Lorentz chaotic system

Regression-to-the-Mean Bias: Good or Bad?



Understanding Transformers for Time Series

Yu, A., **Maddix, D.C.**, et al., "Understanding Transformers for Time Series: Rank Structure, Flow-of-ranks, and Compressibility", arXiv preprint arXiv:2510.03358, Under Review, 2025.
.

Understanding Transformers for Time Series

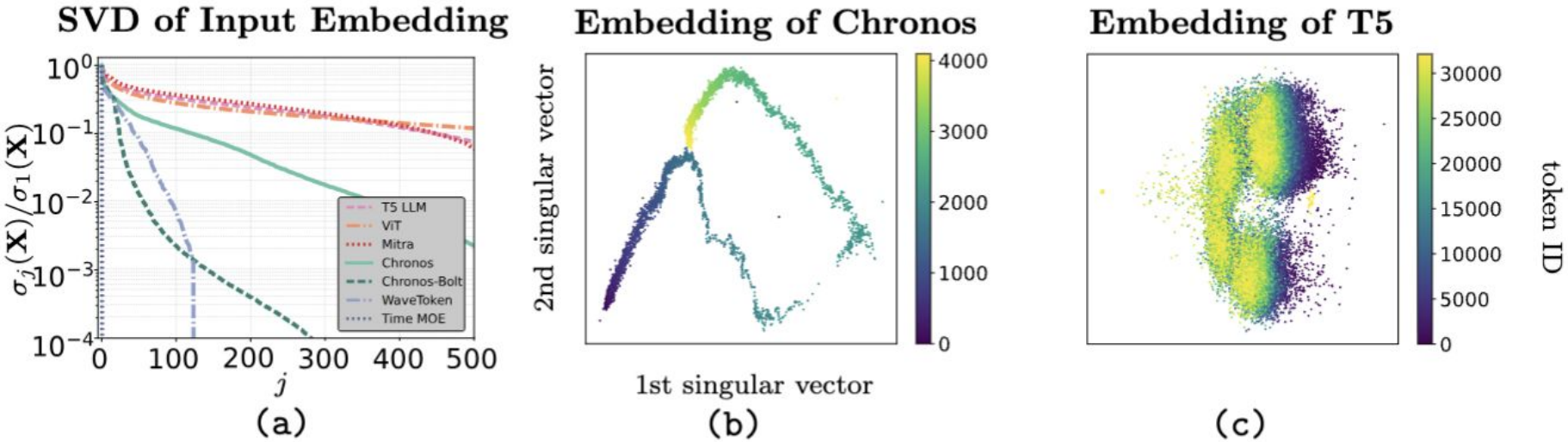


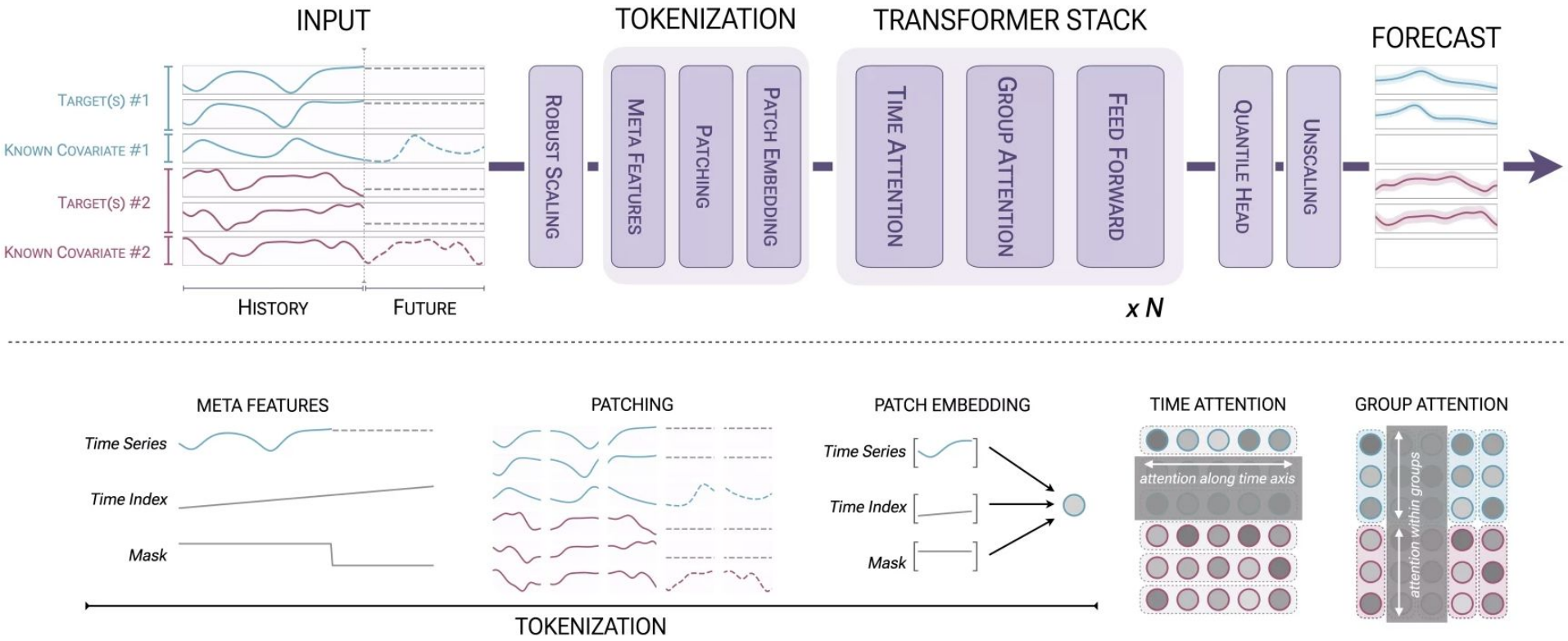
Figure 2: (a): Singular values of the embedded input matrices from many different TSFMs, a TFM, a ViT, and an LLM. (b,c): Embedding space of Chronos and a T5 LLM, respectively, visualized by projecting them onto the leading two singular vectors of the embedding matrix.

Conclusion: Have we achieved a BERT moment?

- Identify design choices in TSFMs that cause 3 inductive biases:
  - Temporal, geometric and regression-to-mean
- Careful numerical analysis and design of TSFMs is required
- Temporal data is a different data-modality, e.g., more compressible, has frequency parameters, continuity in time
- Bitter Lesson
  - Adding traditional forecasting inductive biases can help improve performance on classical benchmarks
  - But it can hurt generalization on unseen domains and tasks,

    e.g., chaotic systems

Chronos-2: From Univariate to Multivariate



Ansari, A.F., et al., "Chronos-2: From Univariate to Universal Forecasting", arXiv preprint arXiv:2510.15821, 2025.

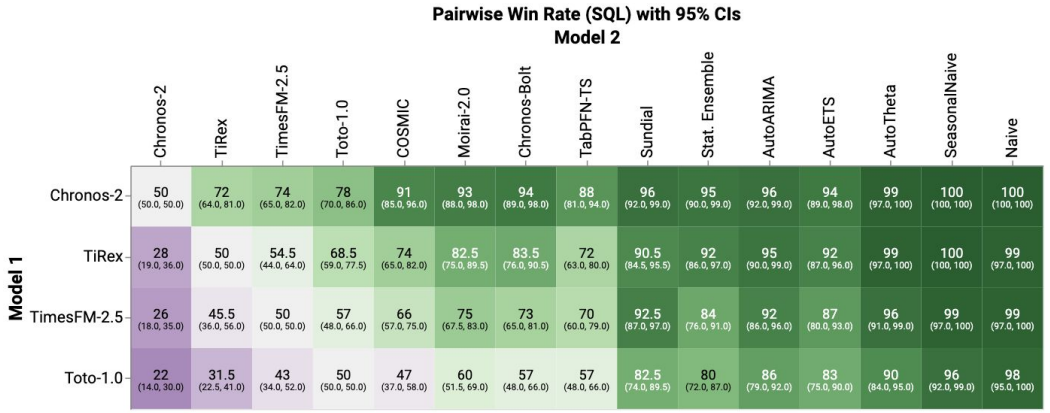
fev-bench : Realistic benchmark for time series forecasting

- Large-scale evaluation on real-world forecasting tasks
  - 100 univariate & multivariate tasks (incl. 46 with covariates)
- Statistically sound aggregation methods
  - Reliable model comparisons using bootstrap confidence intervals
- Extensible infrastructure for reproducible evaluation
  - Lightweight Python wrapper on top of 🤗 `datasets` library
- Paper: arxiv.org/abs/2509.26468
- Code: github.com/autogluon/fev
- Leaderboard: huggingface.co/spaces/autogluon/fev-bench

🏆 Leaderboard

Results for various forecasting models on 100 tasks of the **fev-bench** benchmark, as described in the paper fev-bench: A Realistic Benchmark for Time Series Forecasting.

Model Name	Avg. win rate (%)	Skill score (%)	Median runtime (s)	Leakage (%)	Failed tasks (%)	Zero-shot	Organization	Link
Chronos-2	91.4	47.3	3.6	0	0	✓	AWS	🔗
TiRex	82.7	42.6	1.4	1	0	✓	NX-AI	🔗
TimesFM-2.5	77.6	42.2	16.9	10	0	✓	Google	🔗
Toto-1.0	70.2	40.7	90.7	8	0	✓	Datadog	🔗
Chronos-Bolt	64.4	38.9	1.0	0	0	✓	AWS	🔗
Moirai-2.0	64.4	39.3	2.5	28	0	✓	Salesforce	🔗



Shchur, O., et al., "fev-bench: A Realistic Benchmark for Time Series Forecasting", arXiv preprint arXiv:2509.26468, 2025.

Chronos in the Open Source

- Inference code available on GitHub
- Model weights available on Hugging Face 🤗
- Deploy Chronos-2 on AWS using SageMaker JumpStart
- Run Chronos with 1 line of code using AutoGluon (Chronos-2 coming soon!)

Thank you!

Danielle Maddix Robinson

dmmaddix@amazon.com

https://dcmaddix.github.io