Towards Neural Scaling Laws for Time Series Foundation Models

Daichi Kimura

Paper Information

つなごう。驚きを。幸せを。
NTT docomo Business

- **Paper:** [2410.12360] Towards Neural Scaling Laws for Time Series Foundation Models
- **Authors:** Qingren Yao[1,2], Chao-Han Huck Yang[3], Renhe Jiang[4], Yuxuan Liang[2], Ming Jin[1], Shirui Pan[1]
  - [1] Griffith University
  - [2] The Hong Kong University of Science and Technology (Guangzhou)
  - [3] Nvida Research
  - [4] The University of Tokyo
- **Conference:** ICLR 2025
- **Overview**
  - Comprehensive empirical study on whether scaling laws hold for Time-Series Foundation Models (TSFMs), across in-distribution(ID), out-of-distribution(OOD), and different architectures.

Introduction: Time Series Foundation Models (TSFMs)

つなごう。驚きを。幸せを。
docomo Business

Large, pre-trained models that learn general temporal representations across diverse domains.

Examples: Timer(Liu et al., 2024), Moirai(Woo et al., 2024), Chronos (Ansari et al., 2024)

Problems: there is no systematic guideline for designing and scaling them.

- How large should a TSFM be?
- How much data diversity is enough?
- When does performance saturate?
- What architecture is the best? Encoder or Decoder?

Hence, we need a predictive principle to guide scaling:

Neural Scaling Laws can fill this gap.

Introduction:

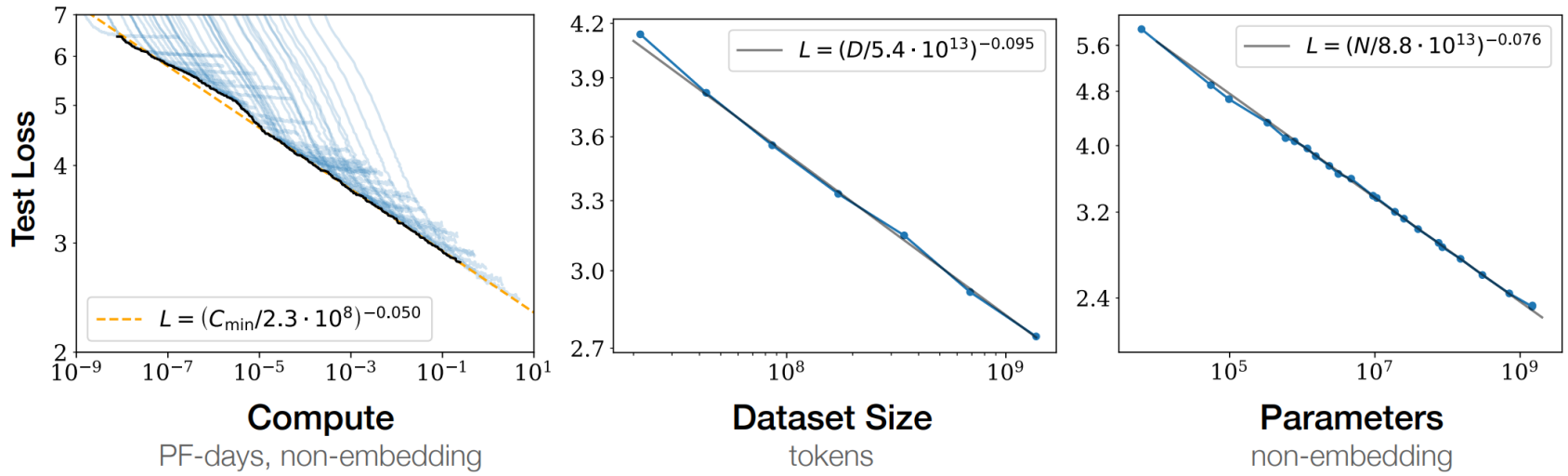
つなごう。驚きを。幸せを。
docomo Business

Neural Scaling Laws (Kaplan et al, 2020)

Empirical power-law relations between model performance and three key factors.

$$L \propto N^{-\alpha_N}, L \propto D^{-\alpha_D}, L \propto C^{-\alpha_C}$$

$\boldsymbol{L}$: Loss (e.g., NLL), $\boldsymbol{N}$: Number of model parameters, $\boldsymbol{D}$: dataset size, $\boldsymbol{C}$: Compute budget



1)

Key insight

- Model performance improves predictably as scale increases.
- Enables *compute-optimal* model design.

1) Kaplan, et al. "Scaling laws for neural language models." *arXiv* (2020).

Introduction: Gaps in Existing Work

つなごう。驚きを。幸せを。
docomo Business

Current scaling studies in time series (e.g., Edwards et al., 2024; Shi et al., 2024) focus on:

✅ *In-distribution (ID)* behavior only

🚫 *Out-of-distribution (OOD)* performance remains unknown

🚫 No analysis across **model architectures** (encoder vs decoder)

Questions:

1. Do neural scaling laws generalize to OOD forecasting?
2. How does model architecture affect scalability?
3. How to design and scale TSFMs efficiently?

- [Thomas D.P. Edwards, et al.] Scaling-laws for large time-series models, NeurIPS 2024 workshop
- [Jingzhe Shi, et al.] Scaling law for time series forecasting, NeurIPS 2024

Research Questions

This paper aims to answer:

1. *Do TSFMs follow power-law scaling under both ID and OOD data?*
2. *How do encoder-only vs decoder-only Transformers differ in scaling?*
3. *What are the practical design principles for scalable TSFMs?*

Approach:

- Systematic scaling experiments on **17B-point corpus (7 domains)**
- Vary **N**, **D**, and **C** over several orders of magnitude
- Compare **baseline Transformers** and **SOTA TSFMs (Moirai, Chronos)**

Methodology & Experimental Setup

Overview of Methodology

つなごう。驚きを。幸せを。
docomo Business

Goal:

To empirically examine **scaling laws** of time series foundation models (TSFMs)

$$L(N) \propto N^{-\alpha_N}, L(D) \propto D^{-\alpha_D}, L(C) \propto C^{-\alpha_C}$$

Varying:

- **Model size (N)**
- **Dataset size (D)**
- **Compute budget (C)**

Across different **architectures** and **data distributions (ID/OOD)**.

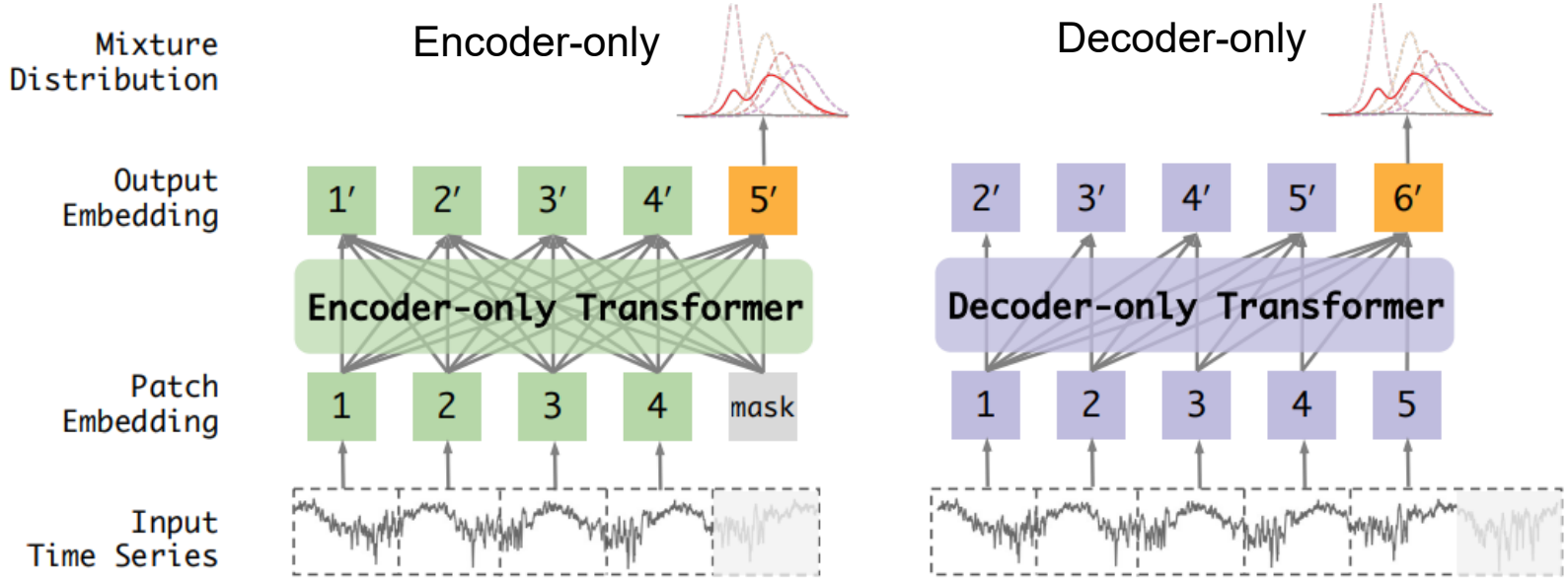
Model Architectures Compared

つなごう。驚きを。幸せを。
docomo Business

4 Models compared in scaling experiments

- **Encoder-only Transformer / Moirai**
  - Moirai adds multi-scale and any-variate attention.
- **Decoder-only Transformer / Chronos**
  - Chronos uses discrete tokenization for autoregressive forecasting.

Scaling setup: Model trained with *varying parameter size* (10^3 *to* 10^8).



Dataset Construction

つなごう。驚きを。幸せを。
docomo Business

Source:

- Built from the large-scale LOTSA achieve (Woo et al., 2024)
- Contains 17B time points across 7 domains (Energy, Transport, Climate, etc.)

Preparation for this study

- Edited LOTSA to ensure balanced domain representation
- Filtered series with signal-to-noise ratio(SNR) > 20 dB to keep high-quality signals

---

Scaling experiment subsets

- Randomly sampled 10 M, 100 M and 1 B series segments
- Split each subset: 95% train / 5% validation (In-Distribution)
- Additional test sets from Monash and LSF datasets (Out-of-Distribution)

Training & Evaluation Setup

つなごう。驚きを。幸せを。
NTT docomo Business

Training Objective:

- Maximize the log-likelihood under Student-t mixture model (robust to outliers)

Optimization:

- Optimizer: AdamW
- Batch size: 128
- Learning rate: 1e-3 (linear warm-up 10k steps + cosine decay)
- Total steps: 100,000
- Compute estimation: $C = 6NBS$

Metris

- NLL, MAPE, SMAPE, MASE, CRPS

Experimental Design Summary

つなごう。驚きを。幸せを。
docomo Business

Controlled Scaling Experiments

	Fixed Variables	Varied Variable	Purpose
Parameter Scaling	D, C fixed	N	Examine model size effect
Data Scaling	N, C fixed	D	Data-efficiency and generalization
Compute Scaling	N, D fixed	C	Efficiency w.r.t. resources

Architecture Comparisons:

- Encoder-only vs. Decoder-only Transformers
- Encoder-only vs. Moirai
- Decoder-only vs. Chronos

ID / OOD Evaluation:

- ID: train on LOTSA / test on LOSTA
- OOD: test on unseen domains (Monash, LSF datasets)

Experimental Results

Results: Overview of Findings

つなごう。驚きを。幸せを。
NTT docomo Business

Key empirical results:

1. TSFMs follow consistent **power-law scaling** in both in-distribution (**ID**) and out-of-distribution (**OOD**) data.
2. **Encoder-only Transformers** scale slightly better than decoder-only ones on ID data.
3. **Advanced architectures** (Moirai, Chronos) improve ID performance but hurt OOD
4. A **sub-linear relation between data size and model size** is observed:

$$D \propto N^{0.8}$$

5. **Emergent behaviors** appear when model size exceeds ~10 M parameters.

Scaling laws hold for time-series models just like LLMs

Results: Parameter Scaling (N)

つなごう。驚きを。幸せを。
docomo Business

Experiment: Vary model parameters from 10^3 to 10^8 (Encoder-only)

Findings:

- Both ID and OOD follow power-law decrease.

ID:

$$L(N) = \left(\frac{2.2 \times 10^7}{N}\right)^{0.054}$$

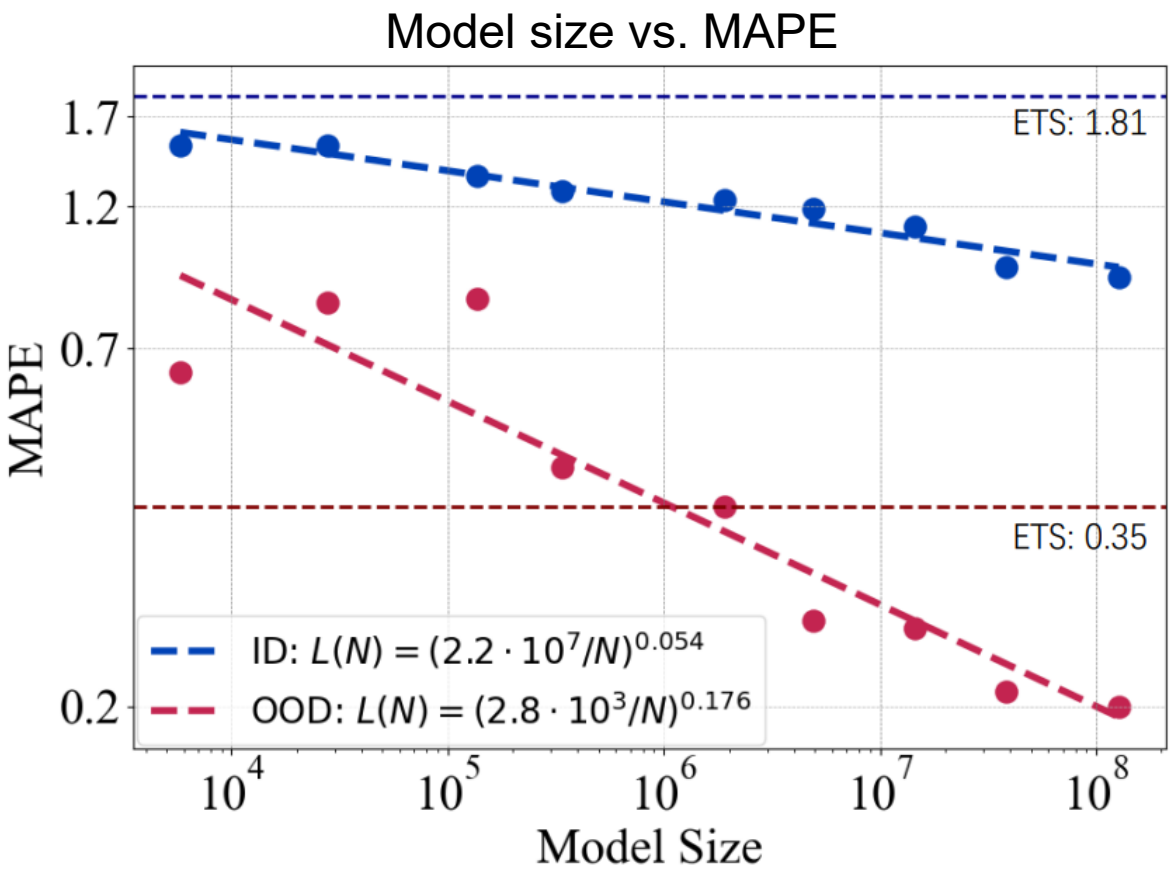
OOD:

$$L(N) = \left(\frac{2.8 \times 10^3}{N}\right)^{0.176}$$

- OOD curve has a larger slope

  ➡ **larger models benefit more for OOD generalization.**

- Pre-trained TSFMs outperform the exponential smoothing (ETS) baseline once $N > 3 \times 10^6$.



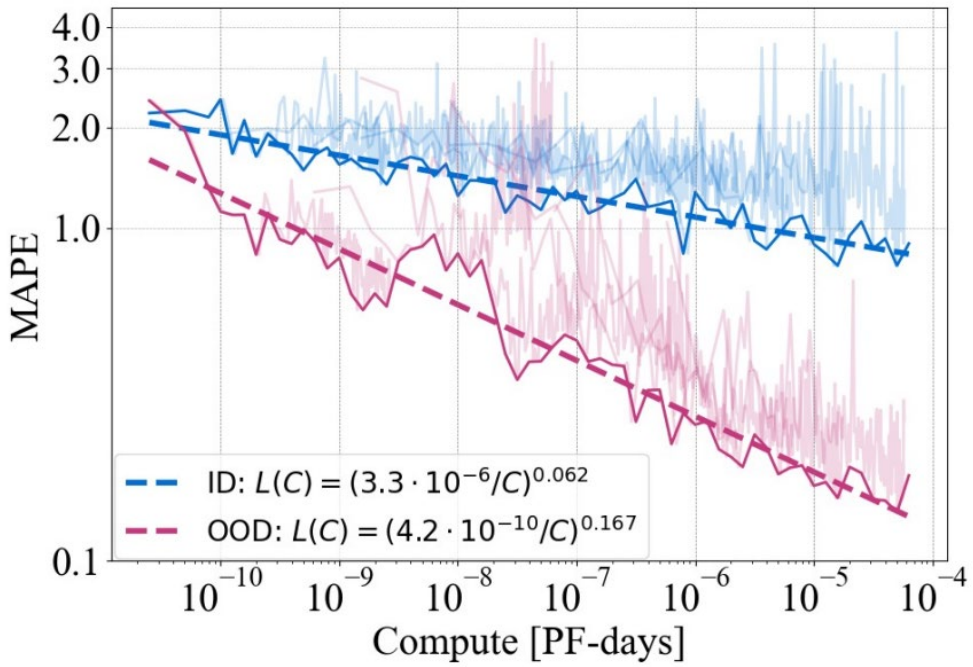
Results: Compute and Data Scaling

つなごう。驚きを。幸せを。
docomo Business

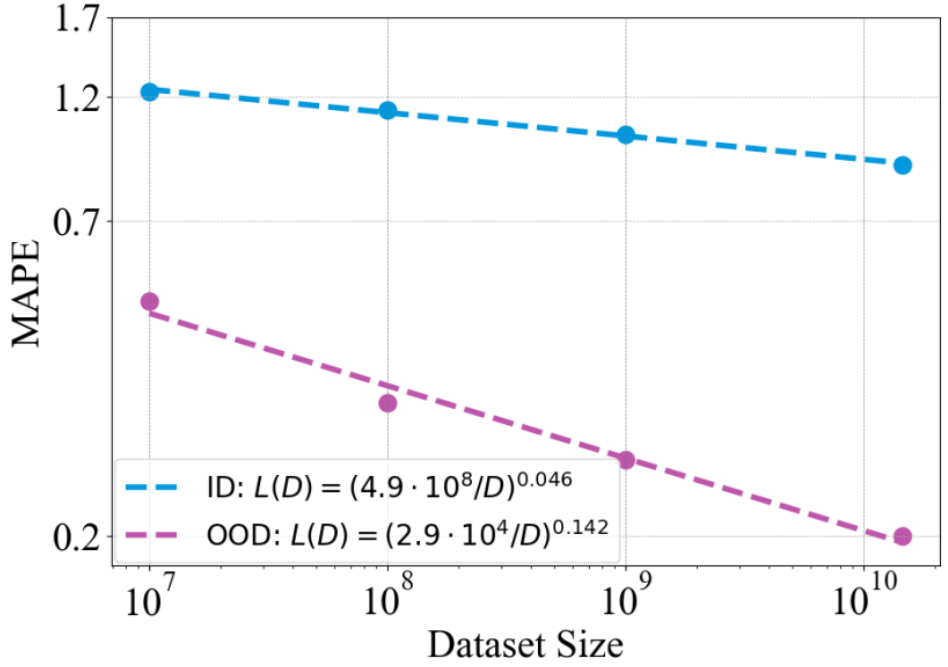
Experiment: Increase training steps/FLOPS (C), dataset size (D)

- Both follow power-law scaling: , $L(D) \propto D^{-\alpha_D}, L(C) \propto C^{-\alpha_C}$
- Larger compute improves efficiency.
- More data yields stronger gains in OOD than in ID.

Compute



Dataset Size



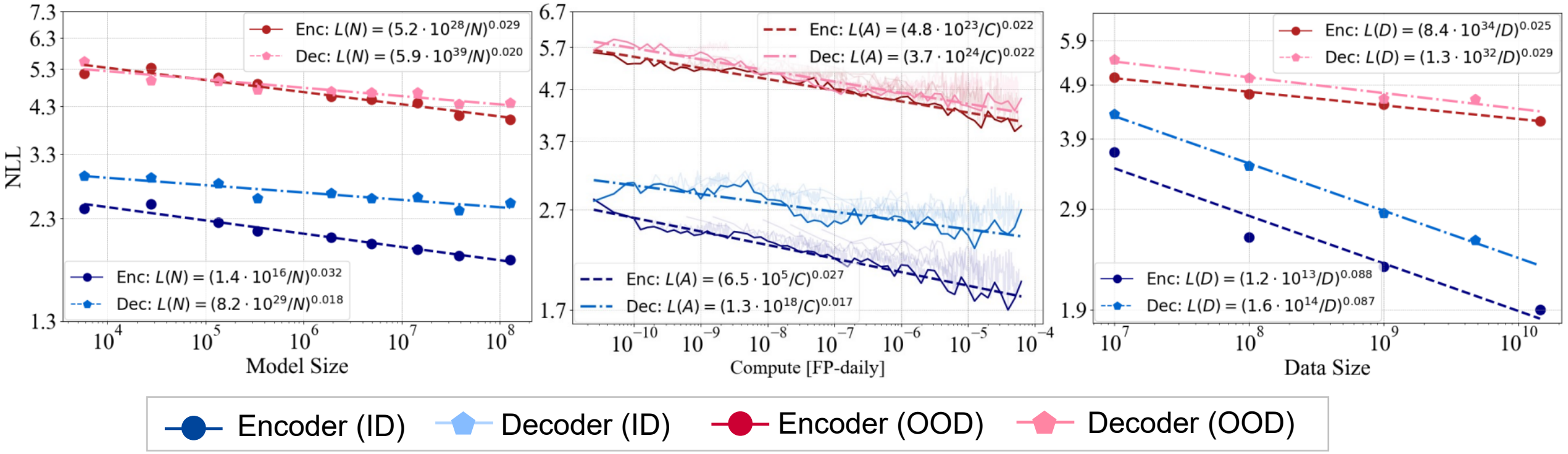
※ PF-days = petaFLOP × day（10^{15} FLOPs × 86 400 sec）

Results: Architectural Comparison

つなごう。驚きを。幸せを。
docomo Business

Encoder-only vs. Decoder-only

- Similar scalability
- Encoder-only shows a slight advantage, with a higher power-law exponent on ID data



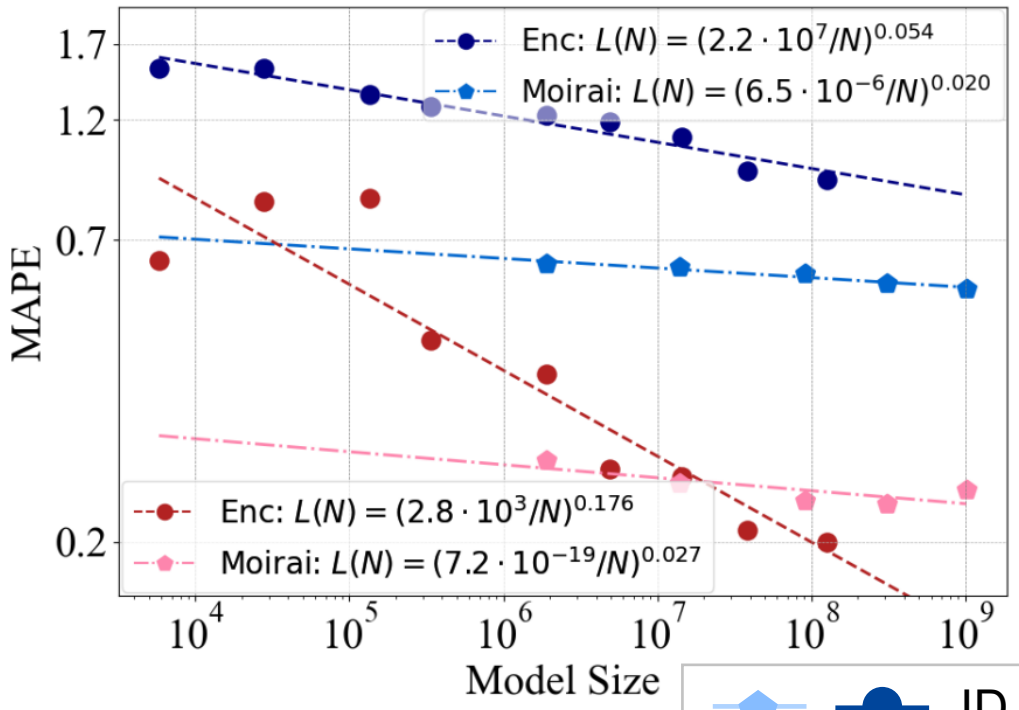
Results: Architecture comparison

つなごう。驚きを。幸せを。
docomo Business

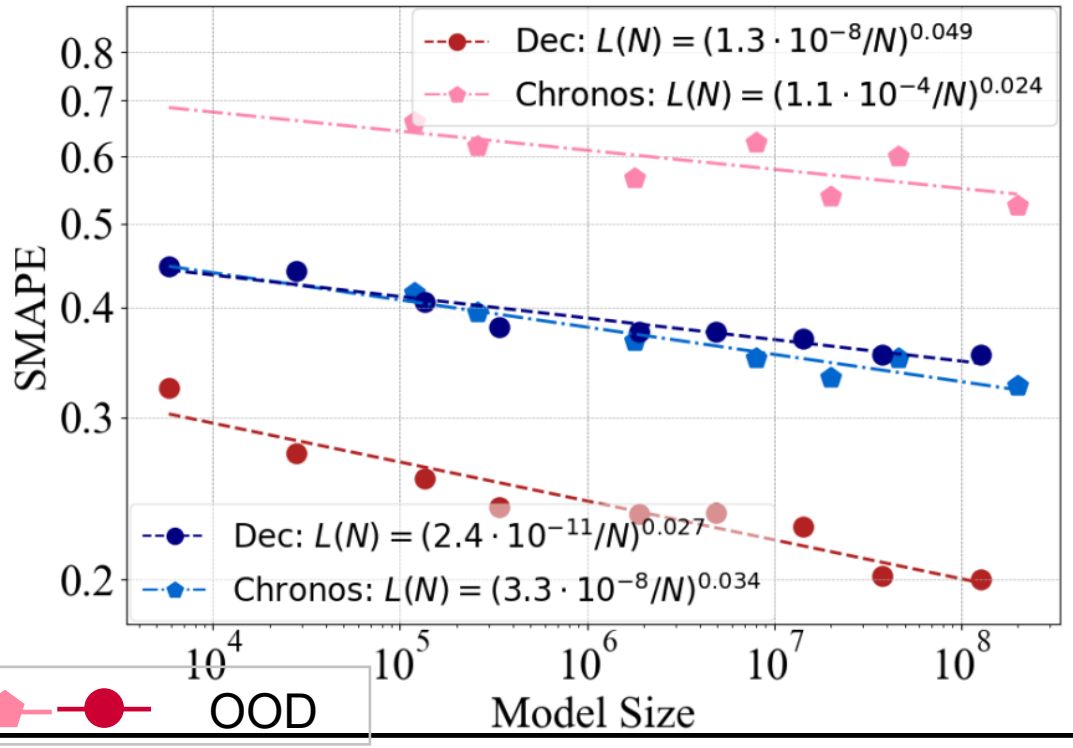
Naïve ? or Specialized modules?

- Simple design: scale smoothly in both ID and OOD.
- Added specialized modules: better ID accuracy but poorer OOD scalability.

Encoder-only / Moirai



Decoder-only / Chronos



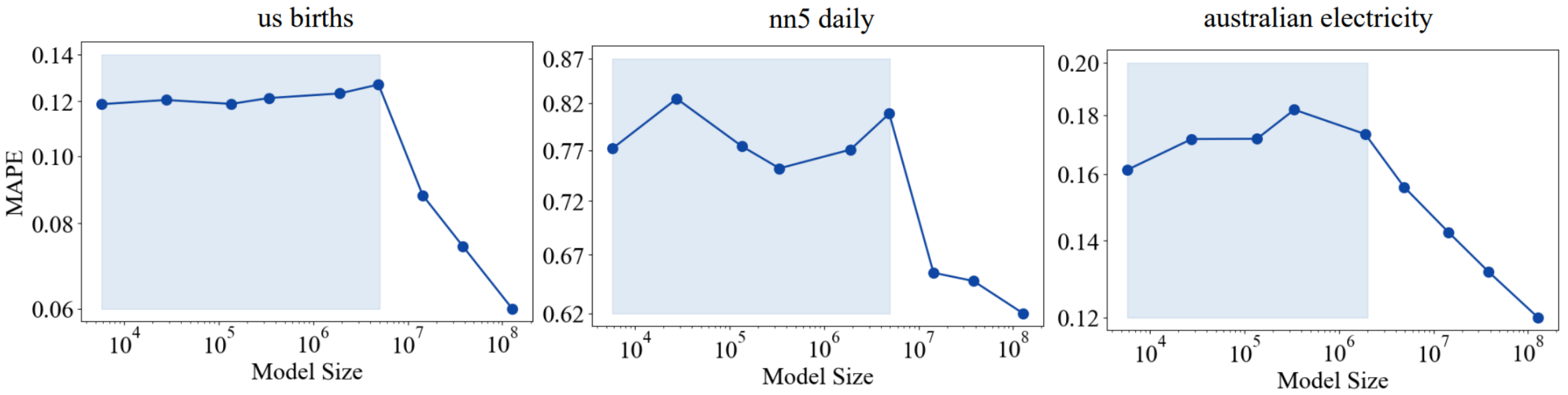
ID OOD

Results: Emergent Behaviors

つなごう。驚きを。幸せを。
docomo Business

Emergent Behaviors

- Task: Zero-shot OOD Prediction
- Above 10^6 model parameters, models show a sudden improvements in MAPE.
- Indicates emergent capability, similar to LLMs



Discussion: Design Principles

つなごう。驚きを。幸せを。
docomo Business

Training Data:

- Larger datasets improve performance, especially OOD.
- 2x data → ~10% lower OOD MAPE
- Diversity matters as much as size.

Model Size:

- Bigger models yield stronger ID/OOD gains.
- $D \propto N^{0.8}$: doubling model size needs ~1.7 x data.
- Emergent behaviors appear above 10^6 params.

Architecture:

- Simple models scale better in OOD than domain-specific ones.

Compute:

- More compute → better performance.

Conclusion

This paper presents a comprehensive empirical study on whether scaling laws hold for Time-Series Foundation Models (TSFMs), across in-distribution(ID), out-of-distribution(OOD), and different architectures.

Main findings:

- Similar to LLM studies, power-low scaling emerges with respect to model size, data size, and compute.
- Scaling laws hold consistently in both ID and OOD.
- The difference between Encoder-only and Decoder-only models is modest.
- Specialized architectural modules (e.g. Moirai, Chronos) improve ID accuracy but can harm OOD scalability.
    - Simpler architectures exhibit better scaling generalization across domains.