Jiecheng Lu

Email: jlu414@gatech.edu **Phone: (+1) 470 929 4016**

Personal Website: https://ljc-fvnr.github.io/ **Google Scholar: https://scholar.google.com/citations?user=H_Bz5A0AAAAJ**

■ Education Background

Georgia Institute of Technology **Atlanta, U.S.**

Ph.D. in Machine Learning (Advisor: Shihao Yang) ***08.2023-05.2027***

Research Interests: Foundation Models, General Sequence Modelling, Linear Attention Mechanisms, Time Series Analysis

M.S. in Analytics ***08.2021-08.2023***

Tianjin University **Tianjin, China**

Bachelor in Logistics Engineering ***09.2016-06.2020***

■ Publications

[1] (ICLR 2026) **Lu, Jiecheng**, and Shihao Yang. "Free Energy Mixer." In The Fourteenth International Conference on Learning Representations.

[2] (NeurIPS 2025 Spotlight) **Lu, Jiecheng**, Xu Han, Yan Sun, Viresh Pati, Yubin Kim, Siddhartha Somani, and Shihao Yang. "ZeroS: Zero-Sum Linear Attention for Efficient Transformer." In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[3] (ICML 2025) **Lu, Jiecheng**, and Shihao Yang. "Linear Transformers as VAR Models: Aligning Autoregressive Attention Mechanisms with Autoregressive Forecasting." In *Forty-second International Conference on Machine Learning*.

[4] (ICML 2025) **Lu, Jiecheng**, Xu Han, Yan Sun, and Shihao Yang. " WAVE: Weighted Autoregressive Varying Gate for Time Series Forecasting." In *Forty-second International Conference on Machine Learning*.

[5] (ICLR 2025) **Lu, Jiecheng,** Yan Sun, and Shihao Yang. "In-context Time Series Predictor." In *The Thirteenth International Conference on Learning Representations*.

[6] (ICML 2024) **Lu, Jiecheng**, Xu Han, Yan Sun, and Shihao Yang. "CATS: Enhancing Multivariate Time Series Forecasting by Constructing Auxiliary Time Series as Exogenous Variables." In *Forty-first International Conference on Machine Learning*.

[7] (ICLR 2024) **Lu, Jiecheng**, Xu Han, and Shihao Yang. "ARM: Refining Multivariate Forecasting with Adaptive Temporal-Contextual Learning." In *The Twelfth International Conference on Learning Representations*.

[8] **Lu, Jiecheng**, and Shihao Yang. “HyperMLP: An Integrated Perspective for Sequence Modeling.” (ICML 2026 Under Review)

[9] Kim, Yubin, Viresh Pati, Jevon Twitty, Vinh Pham, Shihao Yang, and **Jiecheng Lu***. “StretchTime: Adaptive Time Series Forecasting via Symplectic Attention.” (ICML 2026 Under Review)

[10] Pati, Viresh, Yubin Kim, Vinh Pham, Jevon Twitty, Shihao Yang, and **Jiecheng Lu***. “CAPS: Unifying Attention, Recurrence, and Alignment in Transformer-based Time Series Forecasting.” (ICML 2026 Under Review)

(* Corresponding author; project lead supervising undergraduate researchers)

- Served as reviewer for NeurIPS 2024, ICLR 2025, ICML 2025, NeurIPS 2025, ICLR 2026, AISTATS 2026, ICML 2026, the IEEE Internet of Things Journal, etc.

■ Professional Experiences

Tencent **Shenzhen, China**

Data Scientist Intern, Tencent Medical AI Lab (JARVIS Lab) ***07.2021-07.2022***

- Led developments of deep learning and statistical models for projects in medical insurance forecasting and epidemic monitoring.
- Created tech solutions for medical facilities in 4 major cities, handling up to 10 million data points daily.

Amazon **Shenzhen, China**

Business Analyst Intern, Global Sourcing Team, Amazon Private Brands ***01.2021-07.2021***

- Boosted sourcing team performance through data-driven methods, using machine learning and causal inference for cost analysis.
- Developed AWS-based dashboard web apps, aiding in cost decision-making, saving the team over $100,000 monthly in costs.

Peking University **Beijing, China**

Research Assistant, Guanghua School of Management ***09.2020-01.2021***

Country Garden Group **Foshan, China**

Strategic and Data Analytics Intern, New business Division ***07.2020-09.2020***

■ Awards

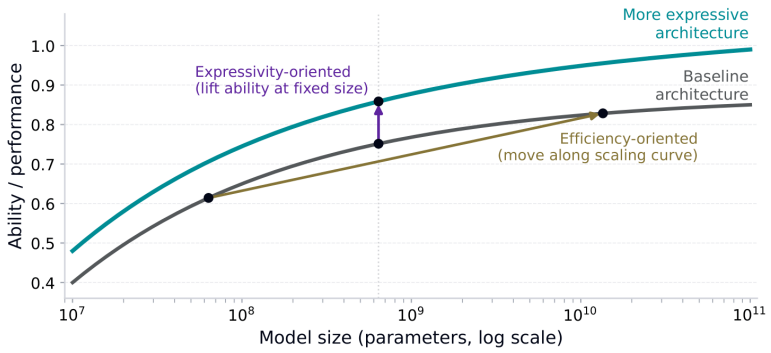
- Margaret and Stephen Kendrick PhD Student Fellowship for Research Excellence, Georgia Tech, 2025
- Named ISyE Fellowship, Georgia Tech, 2023

■ Overview of My Ongoing Research

Research Topic: Reshaping the Scaling Law with Next-Generation Sequence Models

Jiecheng Lu, Georgia Institute of Technology

Modern foundation models hinge on one core capability, sequence modeling. Existing scaling laws describe how performance improves with model size under fixed Transformer architectures. Our research asks a deeper question. Can we design fundamentally more powerful sequence models, so each unit of compute yields more capability? Our goal is not merely to scale along the size to performance curve, but to reshape the curve itself.



We view progress in sequence modeling along two complementary directions. The first is efficiency, improving attention to run faster or handle longer contexts through sparsity, factorization, or streaming. While important, these advances largely preserve the expressive limits of current architectures. Our work focuses on the second direction, expressivity, increasing what a model can represent and compute at fixed size and cost. We explore alternatives that preserve memory quality while enhancing the read mechanism, treating it as a dynamic and differentiable data structure rather than a static average. This perspective underlies our research including ZeroS (Neurips 2025 Spotlight), Free Energy Mixer (ICLR 2026), and HyperMLP (ICML 2026 Under Review).

This framework has guided our recent contributions. In time series modeling, ARM (ICLR 2026), CATS (ICML 2024), In context Time Series Predictor (ICLR 2025), WAVE (ICML 2025), and Linear Transformers as VAR (ICML 2025) align efficient sequence models with classical theory, improving interpretability and control. In core architectures, our works break key expressivity limits of attention. ZeroS relaxes constraints in linear attention, Free Energy Mixer reframes attention as value aware selection, while HyperMLP directly sees attention as a dynamic two-layer MLP for much flexible parameterization.

Looking ahead, we aim to systematically map and expand the compute to expressivity frontier of sequence models. We will unify attention, recurrent models, and state space models under a common theoretical lens, characterize their expressive limits, and scale these designs across domains including vision, audio, robotics, and scientific AI. Our central hypothesis is that the next generation of AI will be driven not only by bigger models, but by better sequence models. By improving how models store and read from memory, we can build systems that are smaller, faster, and more capable, fundamentally reshaping the scaling law itself.