Kai Sheng Tai

kst@cs.stanford.edu
http://kaishengtai.github.io

Last updated
October 21, 2025

About

I am an AI researcher currently working on developing on-device LLMs at Meta, where I also previously led model development on text encoder foundation models for recommendation and retrieval systems. My research has explored ways to leverage structure and invariances in data to build more efficient and effective machine learning models: for instance, by exploiting the structure of language with Tree-LSTMs, the various symmetries in images with Equivariant Transformers, and the underlying physical constraints of the Earth for ML-driven earthquake detection methods.

Education

2021 Ph.D. in Computer Science, Stanford University
Thesis: *Statistical Machine Learning Under Resource Constraints*
Advisors: Peter Bailis and Gregory Valiant

2015 M.S. in Computer Science, Stanford University

2013 A.B. in Physics, *magna cum laude*, Princeton University
Thesis: *Detecting Gravitational Waves from Highly Eccentric Compact Binaries*
Advisors: Frans Pretorius and Sean McWilliams

Professional Experience

2021–present Research Scientist, Meta
- Current focus: On-device LLMs, model compression, and inference efficiency
- Previous: Technical lead for text encoder foundation model development, algorithm design for sparse model training

2016–2021 Graduate Research Assistant, Stanford University

2015–2016 Senior Data Scientist, MetaMind (acquired by Salesforce in April 2016)
- Developed 3D convolutional networks for medical MRI classification

2014–2015 Research Assistant, Natural Language Processing Group, Stanford University

Selected Publications

Kai Sheng Tai, Taipeng Tian, and Ser-Nam Lim. Spartan: Differentiable Sparsity via Regularized Transportation. NeurIPS 2022.

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Sinkhorn Label Allocation: Semi-Supervised Classification via Annealed Self-Training. ICML 2021.

Weiqiang Zhu*, **Kai Sheng Tai***, S. Mostafa Mousavi, Peter Bailis, and Gregory C. Beroza. An End-to-End Earthquake Monitoring Method for Joint Earthquake Detection and Association using Deep Learning. *Journal of Geophysical Research: Solid Earth*, 2022. (*equal contribution)

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Equivariant Transformer Networks. ICML 2019.

Vatsal Sharan*, **Kai Sheng Tai***, Peter Bailis, and Gregory Valiant. Compressed Factorization: Fast and Accurate Low-Rank Factorization of Compressively-Sensed Data. ICML 2019. (*equal contribution)

Edward Gan, Jialin Ding, **Kai Sheng Tai**, Vatsal Sharan, and Peter Bailis. Moment-Based Quantile Sketches for Efficient High Cardinality Aggregation Queries. VLDB 2018.

Kai Sheng Tai, Vatsal Sharan, Peter Bailis, and Gregory Valiant. Sketching Linear Classifiers over Data Streams. SIGMOD 2018.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks. ACL 2015.