Jason (Junjie) Zhu

jasonjunjiezhu.com

Email: junjie.zhu.jason@gmail.com
Mobile: 650-285-7123

Summary

I am a curiosity-driven and scientifically-trained builder with experience in AI/ML, Statistics, and Graph Algorithms. Drawn to hidden patterns, scalable impact, and high-agency teams, I have continuously been applying my skills to real-world problems: multi-modal RAGs, search products, biomedical discovery, etc.

Education

Stanford University — Stanford, CA
Ph.D. in Electrical Engineering · M.S. in Statistics — *2014 – 2020*

Olin College of Engineering — Needham, MA
B.S. in Electrical and Computer Engineering — *2010 – 2014*

Experience

Nexa AI — Cupertino, CA
Head of AI/ML — *Feb 2025 – Present*

- Leadership: Leading a lean and fast-paced team to accelerate Gen-AI edge inference on CPUs, NPUs, and GPUs.
- Local RAGs: Developing privacy-preserving RAGs with small AI models and on-device vision capabilities.
- Agentic Systems: Designing action-driven applications and prototypes with new AI protocols (e.g., MCP, A2A).

Apple — Cupertino, CA
Machine Learning Engineer — *Jan 2020 – Feb 2025*

- Mentorship: Coached software and ML engineers to publish and present at internal and external conferences.
- Synthetic Data Generation: Invented methods to test model robustness via high-dimensional perturbations.
- Preference Learning: Designed cost-efficient offline A/B testing to handle user preference and distribution shifts.
- System Testing: Implemented pipelines (Java) to evaluate query understanding systems for Maps Search.
- Ranking Triage: Developed methods to intepret the impact of multi-ranker systems with linear-time algorithms.

Stanford University — Stanford, CA
Research Assistant — *Sep 2014 – Feb 2020*

- Graph Visualization: Built graph visualizations to interpret and analyze the Gene Ontology.
- Statistical Inference: Improved multiple-hypothesis testing methods to account for data snooping.
- Unsupervised Learning: Created dimension-reduction methods for stem cell and cancer model systems.
- Selective Inference: Proposed methods to study tissue-specific expression quantitative trait loci.
- Sequence Alignment: Optimized speed and accuracy of DNA sequence alignment in C/C++.

Olin College of Engineering — Needham, MA
Research Assistant — *Sep 2010 – May 2014*

- Graph Theory: Solved distance-2-based graph coloring problems for special graph families.
- Information Theory: Modeled wireless networks with stochastic geometric and interference models.

Selected Publications

1. Automatically Authoring Regression Tests for Machine-Learning-Based Systems. *ICSE, 2021*
2. Progenitor identification and SARS-CoV-2 infection in human distal lung organoids. *Nature, 2020*
3. Exploratory gene ontology analysis with interactive visualization. *Scientific Reports, 2019*
4. Visualization and analysis of sc-RNA-seq data by kernel-based similarity learning. *Nature Methods, 2017*

Full list shown on Google Scholar: https://scholar.google.com/citations?user=2EasRdEAAAAJ&hl