Alvin Wan

Apple research scientist optimizing inference latency for server-side Large Language Models.

University of California, Berkeley

Ph.D. in Computer Science · 3000+ citations · 800+ stars · NSF GRFP 2018[1] 2018 - 2022

B.S. in Electrical Engineering and Computer Science · 3.81 Major GPA 2014 - 2018

Samuel Silver Memorial Award ('18), Summer Undergraduate Research Fellowship ('17), Leadership Award ('16), Dean's Honor List ('16), Regents' and Chancellor's Scholarship ('14)

Lecturer 2x, **Head Student Instructor** 6x · 4.83 / 5.00 rating 2015 - 2021

Taught Machine Learning, Discrete Mathematics, Data Science. Managed 70+ person staffs serving 500-1100 students. Wrote math, probability, compsci booklets (10k+ downloads)

Apple

Senior Research Scientist 2022 - now

- Reduced Apple LLM's inference latency by 2.1x in production.
- Leading production optimization effort w/ 3 researchers . Prev: UPSCALE @ ICML 2023.
- Facilitating UC Berkeley sponsorship and collaboration + Hiring and managing interns.

Tesla AutoPilot

AI Intern 2021

Deployed two of Tesla's first FSD radar-less perception models to millions of cars, improving safety-critical kinematics predictions for pedestrians, cyclists etc.

Facebook

Research Intern 2019 - 2021

Neural Architecture Search for low-latency models FBNetV{2,3} @ CVPR{2020,2021}

Software Engineering Intern 2016

Halved media requests from Android Messenger photo gallery

Startups

- Full-Time Machine Learning Engineer at REX Homes	2018 - 2021
- Machine Learning Intern at DeepScale *(acquired by Tesla)*	2017
- Software Engineering Intern at Getexp	2015

International Awards

- Microsoft Imagine Cup "Big Data" World Finals Top 6 Finalists (2018, of 40k+ entrants, 200+ countries)
- Rookies Co. "Web and Mobile" Int'l Top 16 Semifinalist (2017, ~9k entries, 80+ countries)
- Adobe Design Achievement Awards "Social Impact" Int'l Semifinalist (2017, 2018, ~7k entrants)

[1] One of 20 nation-wide recipients for Machine Learning, representing ~0.1% of 12,000+ applicants.