

# Alvin Wan

OpenAI research scientist optimizing inference for Large Language Models

## University of California, Berkeley

<b>Ph.D. in Computer Science</b> · <a href="#">4k+ citations</a> · <a href="#">1.8k+ stars</a> · NSF GRFP 2018 <sup>1</sup>	2018 - 2022
<b>B.S. in Electrical Engineering and Computer Science</b> · 3.81 Major GPA Samuel Silver Memorial Award ('18), Summer Undergraduate Research Fellowship ('17), Leadership Award ('16), Dean's Honor List ('16), Regents' and Chancellor's Scholarship ('14)	2014 - 2018
<b>Lecturer 2x, Head Student Instructor 6x</b> · 4.83 / 5.00 rating Taught Machine Learning, Discrete Mathematics, Data Science. Managed 70+ person staffs serving 500-1100 students. Wrote <a href="#">math</a> , <a href="#">probability</a> , <a href="#">compsci</a> booklets (10k+ downloads)	2015 - 2021

## OpenAI

<b>Member of Technical Staff</b>	Oct '24 - now
Deployed multimodal inference optimizations for OpenAI's largest model to date, <a href="#">GPT 4.5</a> . Enabled doubled context length for <a href="#">GPT 4.1</a> .	

## Apple

<b>Senior Research Scientist</b>	May '22 - Oct '24
Trained and deployed <a href="#">Apple Intelligence</a> 's first low-precision foundation models. Deployed M*-specific optimizations. Facilitated internships, UC Berkeley sponsorship. <a href="#">ICML'23</a>	

## Tesla AutoPilot

<b>AI Intern</b>	Mar - Aug '21
Trained and 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 Reality Labs

<b>Research Intern</b> - Neural Architecture Search, <a href="#">FBNetV2</a> , <a href="#">FBNetV3</a> @ CVPR	May '19 - Mar '21
<b>Software Engineering Intern</b> - Halved media requests for Android messenger	May - Aug '16

## Startups

- <b>Full-Time</b> Machine Learning Engineer at REX Homes	May '18 - Aug '21
- Machine Learning Intern at DeepScale ( <i>acquired by Tesla</i> )	May - Aug '17
- Software Engineering Intern at Getexp	May - Aug '15

## International Awards

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

<sup>1</sup> One of 20 nation-wide recipients for Machine Learning, representing ~0.1% of 12,000+ applicants.