

Tianyi Zhang

<https://www.linkedin.com/in/tyzhanguch>

<https://pttzy.github.io>

Email: tyzhang@berkeley.edu

Mobile: +1-765-637-6774

EDUCATION

University of California, Berkeley

Master in Statistics

GPA:4.0/4.0, Rank: 1/32

Berkeley, CA

2015 - 2016

Purdue University

B.S. in Mathematics, B.S. in Statistics

GPA:3.99/4.0, Rank: Top 1%, Highest Distinction, Minor: Economics, Management

West Lafayette, IN

2012 - 2015

EXPERIENCE

xAI

Member of Technical Staff - Post Training

Palo Alto, CA

May 2025 - Present

- Core contributor to Grok 4, Grok 4 Fast, Grok 4.1, Grok 4.1 Fast, Grok 4.2, and Grok 4.3.
- I lead Grok search post-training team, covering Deep Search with long reasoning/tool use, multi-agent post-training, and fast mode with minimal reasoning and tool-use. I also captain the post-train RL run for fast mode of Grok 4.2 and Grok 4.3.
- Lead search post-training/distillation for Grok models with hybrid reasoning/non-reasoning model achieving SOTA search benchmarks and #1 on LM Search Arena (Grok 4.2 and Grok 4 Fast) upon release.
- Lead factuality training for Grok 4.1/4.1 Fast, developed & pioneered agentic grading recipe that reduced factual hallucinations by 70% for instant-mode info-seeking prompts.
- Developed scalable synthetic QA pipelines for RLVR that boosted Grok 4 Browsecomp from 10% to 40+%, Grok 4.2 Multi-Agent to 80%+.
- Developed rubric generation pipeline for deep research open-ended questions improving Grok-4 10%+ improvements on internal expert bench.
- Developed Synthetic RL data/Internal benchmarks for X searches and toolsets making Grok best to browse X.
- Developed and optimize tools for search & browsing in production and RL training, including search snippet generation model, web page browsing tools, engineering optimizations of caching and speculative browsing of pages to reduce latency.

Apple

Sr Engineering Manager - ML, Siri and Search

Cupertino, CA

Dec 2019 - May 2025

- Lead the team developing completion, spell corrections, QU for spotlight/Safari/Siri Search, as well as build technologies to power those in Apple Music, Appstore, Apple TV products for most of Apple search traffic.
- Tech lead for personal question answering and follow-up questions gen powering Apple Intelligence.
- Developed production spell-correction services with LMs, user log mininig, and fuzzy inverted-index matching.
- Built low-latency NLP models powering joint auto-complete and keyboard corrections for 20+ languages spotlight/safari suggestions powering billions of searches weekly. Performed denoising target pre-training from scratch with usage data.
- Lead Siri ASR server post corrections with language model rewriting, N-best rescoring models, and phonetic fuzzy matchings through phonetic embeddings to resolve ASR errors.
- Developed the whole stack of search relevance retrieval and ranking models for Maps domain in Siri requests

Uber Technologies

Applied Scientist

San Francisco, CA

July 2017 - Dec 2019

- I built most of the real-time and offline ML systems in Uber Risk team with two other co-workers which cut the company's financial fraud loss by 60% with only a third of action rate compared to when I joined.
- Built chargeback fraud, abuse prevention, and auto-refreshing ML systems across payment add/trip request/login flows.
- Within 2.5 years, I helped transforming Uber Risk from having a human-rule dominated system into an auto-updated ML-dominated system

LeanTaaS Inc (Series D SaaS Healthcare AI company)

Research Scientist

Santa Clara, CA

June 2016 - July 2017

- As one of the earliest data scientist, I (together with 2 researchers) designed and developed the scheduling optimization algorithms currently powering the company's core product - IQueue
- Complex operational constraint-based integer programming optimizing cancer center schedules at 50+ top US hospitals
- Built non-homogeneous simulation algorithms for queuing models and staffing optimization tools for Emory University Cancer Lab and MD Anderson, and decreased the average lab patient wait time by 30 mins.

Tencent Inc

Summer Strategy Intern

Shenzhen, China

Summer 2015

Alipay Co., Ltd (Ant Financial), Alibaba Group

ML Intern, Business Intelligence

Hangzhou, China

Summer 2014

SKILLS SUMMARY

- Skills: Machine Learning, NLP, Statistics, LLM Post-training, Agent RL, Synthetic Data