Alex Lu Chen

itisalex@ucla.edu | alexluchen.github.io

EDUCATION

Stanford University — Stanford, CA
M.S. in Computer Science (incoming) — GPA: N/A, Expected June 2028

University of California, Los Angeles (UCLA) — Los Angeles, CA
B.S. in Computer Science — GPA 3.99/4.00, June 2026
Relevant courses: Machine Learning, Computer Vision, Natural Language Processing, Data Structures & Algorithms, Software Construction

- Awards and Honors: Dean's List all years, Tau Beta Pi, Upsilon Pi Epsilon

WORK EXPERIENCE

Amazon — Jan 2025 – April 2025
Software Engineer (Intern) — Seattle, Washington

- Designed and deployed a full-stack internal tool to trace which ML-based strategies led to specific Amazon.com recommendations, enabling large-scale analysis across millions of product traces
- Retrieved and transformed S3-stored model outputs using AWS Athena (SQL) to match key–value data with customer history; implemented with Java AWS Lambda, TypeScript CDK, and React frontend, deployed via AWS CodePipeline (CI/CD)
- Presented the tool to managers and engineers within Amazon's Recommendation Systems team, receiving positive feedback for improving transparency and interpretability of recommendation strategies

Physics and Astronomy Department, UCLA — Nov 2023 – Nov 2024
Software Engineer (Part-Time) — Los Angeles, CA

- Spearheaded the development of Bruin Bot, a ChatGPT-based MERN-stack chatbot, enhancing academic resources for the UCLA community
- Deployed Bruin Bot using Nginx, used by 50 research lab members and securing $35,000 in funding from the Physics department
- Oversaw the creation of a customized admin platform for Physics lab courses, designed to streamline student management for Professors and TAs

PROJECTS & RESEARCH

MINT Lab, UCLA — June 2025 – Present
Research Assistant, Advisor: Aditya Grover — Los Angeles, CA

- First author of *The Pitfalls of KV Cache Compression* (ACL main conference), revealing overlooked efficiency-accuracy trade-offs in LLM inference; project website: pitfalls-of-kv-cache-compression.github.io
- Implemented the majority of the PyTorch codebase, enabling reproducible evaluations for various KV-cache compression methods across multi-instruction LLM benchmarks

Text-to-Speech Model in PyTorch — Feb 2025 – June 2025
Remote

- Built a custom text-to-speech model from scratch in PyTorch, using phoneme-based text tokenization, Descript Audio Codec for audio compression, and an autoregressive transformer to generate audio token sequences
- Achieved functional speech generation within 5 hours of training on 4× NVIDIA T4 GPUs
- Shared a detailed write-up at alexluchen.github.io/AlexTTS with full source code at github.com/alexluchen/AlexTTS

SKILLS

Languages: Python, Java, C++, TypeScript, SQL

Frameworks and tools: React, Node.js, MongoDB, AWS, Git

ML/AI: PyTorch, CUDA, Hugging Face, Transformers