Zhexiang (Bruce) Li

bruceli@ucla.edu | +1 (831) 239-2728 | Los Angeles, CA | Website | GitHub | LinkedIn

Available for internships: Jun 13 – Sep 17, 2027

Education

University of California, Los Angeles Expected Jun 2028
M.S. in Computer Science

University of Southern California May 2026
B.S. in Computer Science, *Magna Cum Laude*; GPA: 3.76/4.00; Viterbi Dean's List

Technical Skills

Languages: Python, C/C++, JavaScript/TypeScript, SQL, Java, Swift
AI/ML: PyTorch, Triton, LLM APIs, agent workflows, RAG (BGE embeddings, ChromaDB), multimodal inference
Applications: FastAPI, asyncio, SQLAlchemy, MySQL, PostgreSQL (Supabase), Next.js, Docker, Linux, Git

Engineering & Research Experience

Health Intelligence Lab, UCLA | *Graduate Researcher* Aug 2026 – Present

- Working on multimodal glucose forecasting with Prof. Yuzhe Yang: harmonized CGM data from 49 datasets with meal, insulin, and wearable context, and now building patient-disjoint benchmarks to compare forecasting models.

PMAOS / Next Big Seek | *AI Engineer Intern* Jul – Aug 2026

- Designed short-video recommendation using bidirectional user–video interest propagation, exposure budgets across multiple timescales, and fatigue controls to manage repeated recommendations.
- Built 11-language audio-visual affect inference that fuses evidence and abstains when modalities conflict.

Terlingua Threads | *Full-Stack Engineer* Remote · Jan – May 2026

- Built Shopify-integrated pages and Google Maps features with responsive, accessible layouts.

Piezolab Medical | *AI Engineer Intern* Lake Forest, CA · Jun – Aug 2025

- Built an LLM-assisted pipeline to ingest, clean, and cluster 5,000+ medical-malpractice documents for quality-of-care review; developed agent workflows for vascular-care research.

Selected AI Projects

AI Doctor WeChat | *Python, FastAPI, MySQL, RAG* 2026

- Built and deployed the FastAPI backend on WeChat Cloud Hosting for a clinic's Mini Program and web workspace covering clinical teaching and patient intake.
- Designed five-role access control so LLM-drafted visit records reach patients only after a doctor confirms them; added RAG (BGE embeddings, ChromaDB, keyword fallback) and a rate-limited asyncio queue for LLM calls.

Fridge Manager | *Next.js, TypeScript, Supabase, LLM and vision models* 2026

- Built and launched howtoeat.food, an AI kitchen web app that plans meals and grocery lists from each user's inventory, leftovers, preferences, cookware, and cooking time.
- Added photo-based inventory recognition and LLM recipe import with user review and deduplication; built a multimodal pipeline (subtitles, ASR/OCR, keyframe vision) that turns cooking videos into recipes.

Investment Analysis Agent Skill | *Python, Claude Code* Jul 2026

- Built an agent skill that researches a company from its name or pitch deck and drafts an evidence-backed investment report; a Python QA harness checks each section's sources, depth, and completeness.

Selected ML Research

- **Efficient Attention via Pre-Scoring** (first author; under review): selects informative keys before approximate attention; implemented in Triton and evaluated with HyperAttention on 131k-token contexts. arXiv:2505.11040
- **MedTokenBudget** (sole author; under review): lesion-supervised token routing that keeps the diagnostic regions of dermoscopy images under a fixed token budget; evaluated on ISIC 2019.