Zhexiang (Bruce) Li

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

EDUCATION

University of California, Los Angeles Sep 2026 – Jun 2028 (expected)

M.S. in Computer Science

University of Southern California May 2026

B.S. in Computer Science, *Magna Cum Laude* · GPA 3.76 / 4.00

RESEARCH EXPERIENCE

Graduate Researcher | Health Intelligence Lab (HAIL), UCLA Aug 2026 – Present

Advisor: Prof. Yuzhe Yang

- Working on multimodal foundation models for diabetes that forecast glucose from continuous glucose monitoring together with insulin, meals, activity, and patient context.
- Harmonized CGM data from 49 datasets into a common format with meal, insulin, and wearable context.
- Building patient-disjoint MetaboNet benchmarks to compare GluFormer, T1DFormer, and custom predictive-state and Transformer forecasters at 30-, 60-, and 120-minute horizons.

USC Drug Discovery Research Spring 2026

Research with a Ph.D. student in Prof. Yan Liu's group · Project poster

- Studied SARS-CoV-2 antiviral discovery through reproducible benchmarks of molecular descriptor-based models, Chemprop, and hybrid approaches across seven Therapeutics Data Commons datasets, using scaffold splits.
- Analyzed descriptor importance and SHAP attributions to examine the molecular features behind model predictions.

PUBLICATIONS & MANUSCRIPTS

MedTokenBudget: Lesion-Preserving Token Routing for Dermoscopic Image Classification 2026

Zhexiang Li (sole author). Under review

Designed lesion-supervised token routing after a frozen vision backbone so that dermoscopy classifiers keep diagnostic regions under a fixed token budget. Trained it with budget curriculum learning, diversity regularization, and attention distillation, and evaluated it on ISIC 2019.

Mass-Conservation Token Transformers for Auditable MS/MS Spectrum Retrieval 2026

Under review at AAAI 2027

Represented mass-conserving fragment pairs as sparse evidence tokens for spectrum retrieval. Used closest-mass and overlap-hard negatives, alongside evidence-removal tests, to examine whether the exported evidence affects ranking.

Efficient Attention via Pre-Scoring: Prioritizing Informative Keys in Transformers 2025

Zhexiang Li, Haoyu Wang, Yutong Bao, David Woodruff. Under review · arXiv:2505.11040

First-author work advised by Prof. David P. Woodruff (CMU). Proposed pre-scoring and implemented informative-key selection in Triton; integrated it with HyperAttention and evaluated it on ChatGLM with 131k-token contexts.

Game Theory in Companies 2023

S. Li, **Z. Li**, Z. Liu. *Advances in Economics, Management and Political Sciences*, 16(1), 357–366. DOI

Advised by Prof. Ming Gu (UC Berkeley). Led a team modeling game-release decisions as two-player normal-form games and enumerating Nash equilibria with Python and Gambit.

PROFESSIONAL EXPERIENCE

AI Engineer Intern | PMAOS / Next Big Seek
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.
- Developed 11-language audio-visual affect inference with evidence fusion and a fixed-point decision layer that can abstain or report conflicting modalities; implemented in C for Cortex-M3 and ARM926EJ-S targets.
- Screened sub-10 mm piezoelectric robot architectures and specified a PZT-steel cantilever experiment for a 9 mm candidate; work concluded at experiment and procurement planning.

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

- Built Shopify-integrated pages and Google Maps features; improved responsive layouts, accessibility, SEO, and maintainable content structures.

Co-Founder & CTO | CrAIg Capital, Los Angeles, CA
2025 – Mar 2026

- Led product and engineering for AI-assisted housing debt collection, including data pipelines and automated workflows.
- Set technical direction, hiring priorities, and delivery standards during startup development.

AI Engineer Intern | Piezolab Medical, 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 and experiment checklists for vascular-care research.

Business Development Intern | Hangzhou Haoan Jianke Technology, Hangzhou
Jun – Aug 2024

Gathered requirements for a NetEase CodeWave internal platform, helped recruit engineers, and contributed to program structure and interface planning.

Application Engineer Intern | Synaptics, San Jose, CA
Jun – Aug 2023

Validated product firmware and built an ESP32-to-Alibaba Cloud pipeline that published I²C sensor data as MQTT topics for internal dashboards.

Investment Analyst Intern | GUDU Medical, Remote
Jan 2021 – Present

Supported MedTech deal screening, market analysis, due diligence, financial modeling, and valuation.

SELECTED PROJECTS

Fridge Manager
2026

Next.js · TypeScript · Supabase · LLM and vision models · howtoeat.food

- Built and launched an AI kitchen web app that plans daily and weekly 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 corrections, admin AI review, and deduplication; isolated each user's data with Supabase Postgres row-level security.
- Built an offline multimodal pipeline (subtitles, ASR/OCR, keyframe vision) that turns cooking videos into structured recipes for the library.

AI Doctor WeChat
2026

Python · FastAPI · MySQL · RAG · WeChat Mini Program

- Built and deployed the FastAPI backend on WeChat Cloud Hosting for a clinic's Mini Program and web workspace covering clinical teaching, clinician collaboration, and patient intake.
- Designed five-role access control so LLM-organized visit records stay drafts until a doctor confirms them; added an asyncio queue that schedules LLM calls under global, per-user, and per-API limits.
- Implemented RAG with BGE embeddings, ChromaDB cosine search, and a keyword fallback.

Investment Analysis Agent Skill

Jul 2026

Python · Claude Code

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.

AI Business Card Hub

Jan 2026 – Present

Swift · SwiftUI · Vision parsing · OCR · Source code

Built photo-to-contact extraction with OCR fallback, multi-photo review, bilingual records, and searchable people and company profiles.

TEACHING

Tutor | Introduction to Machine Learning, UCSC

Spring 2023

Prof. Lise Getoor

- Held office hours and discussion sessions on logistic regression, SVMs, and tree models.
- Co-designed Jupyter Notebook assignments and contributed to the coding autograder.

Assistant Coach | Shanghai Bosidi Education

2014 – 2018

VEX Robotics & Olympiad in Informatics

Supported robot design and algorithmic problem-solving instruction, scrimmages, and events.

HONORS & ACTIVITIES

University of Southern California: *Magna Cum Laude* (May 2026); Viterbi Dean's List (Fall 2024 – Spring 2026).

TIE Hub: Co-Founder Speed Date volunteer (Sep 19, 2025); five-day startup hackathon participant, including prototype development, venture design, and pitching (Sep 2025).

Noah Business School: Entrepreneurship, financing, and Silicon Valley company visits (2025). **USC EVMA:** Member.

TECHNICAL SKILLS

Programming: Python, C/C++, JavaScript/TypeScript, SQL, Java, Swift

Machine learning: PyTorch, Triton, scikit-learn, RDKit, Chemprop; multimodal modeling, efficient attention, recommendation, and molecular prediction

LLM systems: LLM and vision-model APIs, agent workflows, retrieval-augmented generation (BGE embeddings, ChromaDB), ASR/OCR pipelines

Engineering: FastAPI, asyncio, SQLAlchemy, MySQL, PostgreSQL (Supabase), Next.js, Docker, Linux, Git; automated testing and data pipelines

---

Updated November 2026 · Full experience archive & application resumes