Zhifei (Andy) Li

☎ (+1) 510-292-1508 ✉ zhifei.li@berkeley.edu andylizf Zhifei (Andy) Li

Research Interests

My research interests lie in **designing efficient systems for ML**, focusing on cloud resource orchestration, distributed training infrastructure, and compound AI systems addressing the growing resource demands of diverse AI applications. I am also interested in exploring how AI techniques can advance systems design methodologies.

Experience

Sky Computing Lab — University of California, Berkeley

Research Intern, advised by Prof. Ion Stoica
worked with Prof. Joseph E. Gonzalez, Prof. Matei Zaharia — July 2025 - December 2025

- **SkyNomad: Multi-Region Spot Instance Scheduling** (submmited to **OSDI '26**)
  - Designed a multi-region spot instance scheduling system, addressing single-region availability bottlenecks for offline workloads, via Unified Cost Model trading off cross-region availability and pricing vs. migration costs
  - Achieved 50% cost reduction over the SOTA, saved $1,000+ from a $2,200 training job vs. AWS SageMaker
  - Led project from research formulation to production, drove methodology design, and built simulation framework
- **LEANN: Storage-Efficient Compound AI Systems** (submitted to **MLSys '26**)
  - Co-designed a two-level recompute algorithm to cut vector index storage overhead in RAG pipelines
  - Achieved 97% storage reduction with <5% latency impact; led open-source implementation to 4,000+ GitHub stars
  - Built the system extending `FAISS` C++, contributing 70% codebase; conducted comprehensive experimental evaluation
- **AI-driven Systems Research**
  - Investigated automated systems optimization through evolutionary algorithms and LLM-guided design space exploration
  - Led case study in Barbarians at the Gate paper, which demonstrated 30% improvement over the SOTA
  - Co-developed FrontierCS benchmark with problem specifications and evaluations for 40 open-ended problems

University of California, Berkeley — Berkeley, CA, USA

Exchange Student, Computer Science — August 2024 - December 2024

- **CS294-162 Machine Learning Systems graduate seminar**
  - Optimized complex DAG workload execution through intelligent data placement and cross-cloud task scheduling
  - Achieved 45% cost reduction; select optimizations merged into SkyPilot

Renmin University of China (Ranked 23rd globally on CSRankings 2025) — Beijing, China

Bachelor's in Computer Science, **Turing Honors Class** — September 2022 - June 2026 (Expected)

- GPA: 3.8/4.0 (Top 5%)

Publications

SkyNomad: Cost-Effective Multi-Region Scheduling for Offline Workloads on Spot Instances
Zhifei Li*, *Tian Xia**, *et al.*, *Scott Shenker*, *Ion Stoica*
OSDI '26 (in submission)

LEANN: A Low-Storage Overhead Vector Index
Yichuan Wang, **Zhifei Li**, *Shu Liu*, *et al.*, *Ion Stoica*, *Sewon Min*, *Matei Zaharia*, *Joseph Gonzalez*
MLSys '26 (in submission)

Barbarians at the Gate: How AI is Upending Systems Research
*Audrey Cheng**, *Shu Liu**, *Melissa Pan**, **Zhifei Li**, *Bowen Wang*, *et al.*, *Ion Stoica*
arXiv: 2510.06189

FrontierCS: The Next Frontier of Computer Science
*Qiuyang Mang**, *Wenhao Cai**, **Zhifei Li***, *Huanzhi Mao**, *et al.*, *Ion Stoica*, *Jingbo Shang*, *Zhuang Liu*, *Alvin Cheung*
arXiv

SkyWalker: A Locality-Aware Cross-Region Load Balancer for LLM Inference
Tian Xia, *Ziming Mao*, *Jamison Kerney*, *Ethan J. Jackson*, **Zhifei Li**, *Jiarong Xing*, *Scott Shenker*, *Ion Stoica*
EuroSys 2026

Open-Source Projects

LEANN: the Smallest Vector Index in the World

LEANN (4.1k ★)

Enjoy 97% storage savings for RAG application on your personal device — September 2024 - Present

- Led research-to-production translation of LEANN from prototype to production-ready open-source Python package with CI/CD pipeline, grew to 4,000+ GitHub stars with 3 active external contributors and 40k+ community downloads
- Drove technical outreach including blog posts social media campaign achieving 600k+ views

SkyPilot: Run AI on Any Infra

SkyPilot (8.9k ★)

Framework for running ML/AI workloads across any cloud infrastructure — September 2024 - Present

- Top 10 contributor; created 70+ issues and merged 50+ pull requests; contributed 30,000+ lines of code changes
- Implemented High Availability Controller for SkyServe control plane; adopted by startups including Hypermode

Services

USENIX ATC '25 Artifact Evaluation Committee

Reviewer — May 2025

Introduction to Computer Systems (ICS)

Head Teaching Assistant — Fall 2024, Spring 2025

- Led 6 TAs in teaching systems curriculum that covered cache hierarchies and memory optimization to 200+ students

RUC Computer Association

President — July 2024 - July 2025

- Organized 10+ tech talks with 5000+ total attendees covering topics from Functional Programming to Rust ecosystem
- Led 100+ members across 6 departments, fostered a startup atmosphere and inclusive environment

Cheese Tech

Co-founder — September 2023 - October 2024

- Raised $300K seed funding, grew to 1,000+ users across 3 partner institutions within first year
- Developed AI-powered research platform with inspiration tracking, progress management, and intelligent advising

Honors and Awards

Elite Collegiate Award, China Computer Federation (<100 recipients **nationally**) — August 2025

Dean’s Scholarship, Gaoling School of AI (15 recipients out of 2000) — May 2025

National Scholarship (Top 0.2% nationally) — September 2024

First-Class Scholarship for Social Service (48 recipients out of 35,000) — September 2023

First Prize, National Olympiad in Informatics, Beijing (300 winners) — December 2019

Skills

Coding	**C++**, **Python**, **CUDA**, Rust, TypeScript
Tools	**PyTorch**, **SkyPilot**, NeMo, Ray, Kubernetes, verl, Nix, Typst
Languages	English, Chinese (native), French