Jason Lee

jason_lee@berkeley.edu linkedin.com/in/jasonnn-lee 408-410-2915

Education

University of California, Berkeley Expected May 2027

Bachelors of Science in Electrical Engineering and Computer Science *Berkeley, CA*

- Researching Compound AI Systems under the guidance of Professor Matei Zaharia (Databricks CTO) and Jared Quincy Davis (PhD, Founder/CEO of Mithril AI)
- **Relevant Coursework:** Machine Learning, Engineering Optimization, Computer Vision, Operating Systems, Data Structures, Discrete Math, Designing Information Systems & Devices I and II, Computer Security, LLM Agents, Digital Design, NLP, Systems & Signals
- **Relevant Clubs**: HKN, Cal Ice Hockey, IEEE

Experience

Berkeley Artificial Intelligence Research/Sky Laboratory Aug 2023 – Present

Undergraduate Researcher *Berkeley, CA*

- Designed and implemented a compound AI system that utilizes GEPA RL prompt optimization to train a diverse ensemble to investigate the effectiveness of parallel inference time scaling laws
- Orchestrated a compound AI system pipeline of fine-tuned and base models; proved this can reduce mode collapse by over 30%
- Implemented batch prompting & attention masking using PyTorch for local model batch evaluation, increasing throughput by >4x
- Programmed robust experiments and evaluation pipelines to compare various LLM tuning strategies (DPO, RLHF, and Contrastive-SFT), providing insight into best practices for improving alignment and safeguarding against prompt injections

Kepler AI (AI Agents for Bioinformatics) May 2025 – Sep 2025

ML Engineering Intern *San Francisco, CA*

- Pioneered adoption of the novel GEPA framework, boosting bioinformatics benchmark scores by over 10 points (22.9% to 33.4%) to achieve industry-leading results
- Spearheaded end-to-end observability revamp: migrated all logs to Grafana, implemented 20+ user and performance metrics, and built dashboards + Slack alerting system that now catches 15+ previously undetected critical errors daily
- Led UI/UX redesign that eliminated 30% of redundant/unnecessary UI elements; post-launch, 44% of voluntary user feedback praised the improved ease of use and visuals

Apple Inc. May 2024 – Aug 2024

Software Engineering Intern *Cupertino, CA*

- Fine-tuned a vision transformer (ViT) to detect objects in PDFs, expanding detection capabilities to include **5 new object types**
- Programmed novel UI to display detected objects using Objective-C, vastly enhancing user experience interacting with PDFs
- Investigated/Analyzed the trade-offs of various AI-driven document layout analysis methods; presented findings to team
- Nominated as **one of 10 interns (out of 120)** to showcase my project to SWE Leadership, including the SVP of Software Engineering

Projects

Ember - OSS Framework for building Compound AI Systems | *Python (asyncio)* github.com/pyember/ember

- Implemented over five new evaluators to measure language model output diversity, including cosine-similarity, novelty score, etc.
- Accelerated compound AI system inference throughput by coordinating parallel LLM inference calls using thread-based concurrency, reducing end-to-end ensemble reasoning latency by 4–6× for best-of-N and verifier-judge architectures
- Developed maintainable, well-documented modules with usage examples to facilitate adoption and integration by other developers

Computer Architecture LLM Benchmark | *Python, Shell*

- Designed benchmark questions and implemented code to evaluate arbitrary LLM APIs on hand-curated computer architecture questions, covering topics such as digital logic, pipelining, and operating systems
- Automated correctness checking of LLM responses using REGEX-based parsing, storing structured results in JSON, and generating insights through custom visualization scripts

Pintos Operating System | *C, Perl, Makefile*

- Enabled user program execution on bare-bones OS, implementing argument passing, file handling, & process control syscalls
- Enhanced OS threading by implementing multithreaded user program support and a strict priority scheduler
- Built a robust MapReduce system in Rust, enabling efficient, fault-tolerant distributed data processing

Computer Vision Projects | *Python, PyTorch, NumPy, CUDA* jason-lee08.github.io/cs_180/

Skills

Programming Languages: Python, C++, C, Java, Shell, ROS/ROS2, SQL, HTML/JavaScript/CSS, x86 Assembly
Technologies: Git, PyTorch, NumPy, Matplotlib, React, Hugging Face, Pandas, Tensorflow, Cursor, VSCode, Vim
Natural Languages: Fluent in conversational Mandarin
Interests: Ice hockey, climate change, weightlifting, live music