Fanjia Yan

fanjiayan@berkeley.edu | fanjia-yan.github.io | github.com/Fanjia-Yan | linkedin.com/in/fanjia-yan/

Education

University of California, Berkeley Class of 2025
M.S. in Electrical Engineering & Computer Science
University of California, Berkeley Class of 2024
B.A. in Computer Science and Applied Mathematics GPA: 3.7/4.0
Relevant Coursework:
Deep Learning, Operating Systems, Machine Learning, Algorithms, Data Structures, Optimization Models, Computer Architecture, Computer Graphics, Probability and Random Processes, Linear Algebra, Computability and Complexity

Experience

Deep Learning Software Engineer Intern, Nvidia **May 2024 – Present**

- Prototyped Nvidia Drive in-car assistant by building a Retrieval-Augmented Generation (RAG) device for retrieving car manual knowledge using NV-Embed and NeVA-22B, which was presented at monthly demo.
- Fine-tuned Llama-3 on efficient tool usages, significantly boosting in-car assistant tool calling accuracy and web search capabilities

Graduate Researcher, Berkeley Sky Computing Lab **Aug 2023 – Present**

- Spearheaded the implementation of **LiveCodeBench**, a live benchmark for code LLMs, ensuring real-time performance evaluations while preventing data contamination (submitted to NeurIPS 2024 arxiv.org/pdf/2403.07974.pdf)
- Designed and maintained **Berkeley Function Calling Leaderboard** (BFCL), the first comprehensive evaluation on the LLM's ability to call functions and use tools at scale
- Trained **Openfunctions**, a 6.91B parameter function calling model fine-tuned with 65K synthetic function calling question-function-answer pairs from real world usage, which received 12K deployments within a month on HuggingFace.

Software Development Intern, Amazon **May 2023 – Aug 2023**

- Designed and implemented an end-to-end debugging pipeline using Java to identify badging candidate qualifications and filtered reasons from Amazon's Choice ML Ranker, leading to a simplified debugging process and a 20% increase in debugging productivity
- Increased integration tests coverage to 96% for Amazon's Choice multi-facet pick project, resulting in a projected annualized growth adjusted profit of 200MM
- Completed the internship project within 8 weeks out of the allocated 12 weeks and provided testing and clean-up assistance for ongoing high-impact projects led by senior engineers

Software Engineer Intern, Aqueduct **Jun 2022 – May 2023**

- Built and productionized end-to-end support for On-Demand Kubernetes integration within 2 months using EKS and Golang, improving control over MLOps workflow life cycle and mitigating excessive compute resource usage that saves on an average of $1600 for individual user with idle node groups
- Implemented resource configuration for AWS Lambda and Kubernetes, which allows users to specify memory limit and GPU allocation for their workflows
- Enabled workflow parametrization using Python that allows users publish the same workflow with different parameterized inputs that are configurable at runtime

Software Engineer Intern, Ping An Good Doctor **Nov 2020 – Jan 2021**

- Deployed and optimized existing auto-filling and filtering algorithm using Java to the patient database, resulting in a 15% improvement in the daily query offline build for patient demographic information
- Performed segmentation analysis on Python to cluster existing patients based on transactions, order histories, and symptom's descriptions which is adopted as a core dashboard metric

Extracurriculars

Undergraduate Researcher, Berkeley Barsky Lab **Sep 2022 – Jun 2023**

- Worked with Professor Brian Barsky on prototyping vision correcting VR display using Google Cardboard that served as an alternative to traditional screen displays
- Reimplemented pre-filtering algorithms in OpenGL, which allowed for parallelizations and decreased pre-filter runtime by 30%

Skills

Programming Language: Python, Java, JavaScript, Go, Unix/Linux, C/C++, R, Rust, SQL, HTML
Technology: AWS, Kubernetes, Docker,Pytorch,Git Langchain, vLLM,CUDA
Frameworks: , Node.js, React, Delta Lake, Django, Hadoop,MapReduce