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:
Natural Language Processing, Deep Learning, Operating Systems, Machine Learning, Algorithms, Data Structures, Optimization Models, Computer Architecture, Computer Graphics, Probability and Random Processes, Linear Algebra, AI System

Selected Publication

1. **Fanjia Yan***, Huanzhi Mao*, Charlie Cheng-Jie Ji*, Ion Stoica, Joseph E. Gonzalez, Tianjun Zhang, Shishir G. Patil. Berkeley Function Calling Leaderboard.
2. Naman Jain, King Han, Alex Gu, Wen-Ding Li, **Fanjia Yan**, Tianjun Zhang, Sida Wang, Armando Solar-Lezama Koushik Sen, Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. arXiv:2403.07974

Experience

Deep Learning Software Engineer Intern, Nvidia **May 2024 – Aug.2024**

- Prototyped Nvidia Drive in-car assistant by building a multi-modal Retrieval-Augmented Generation (RAG) pipeline for retrieving car manual knowledge using NV-Embed and NeVA-22B, which was presented at monthly demo
- Fine-tuned Llama-3.1 with LoRA on efficient car control tool usages and SERP web browsing, significantly boosting in-car assistant tool calling accuracy and web search capabilities
- Implemented a multi-turn evaluation framework for in-car assistant, leveraging Monte Carlo simulation to optimize accuracy and manage large sample sizes in chat history reordering scenarios.

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, which was widely adopted by Google, Meta, and Cohere.
- 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.

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