Laryn Qi

+1 (925) 336-1528 | larynqi@berkeley.edu | larynqi.com/ | linkedin.com/in/larynqi/ | github.com/LarynQi/

Education

University of California, Berkeley **GPA: 3.95/4.00**

M.S. Electrical Engineering and Computer Science *May 2024*

- **Thesis**: LLM-Based AI Tools for CS Education. Work featured by Microsoft and published on arXiv & Berkeley
- **Courses**: Generative AI & LLMs, Reinforcement Learning, Natural Language Processing, Convex Optimization

University of California, Berkeley **GPA: 3.82/4.00**

B.A. Computer Science, B.A. Music *May 2023*

- **Courses**: Statistical Machine Learning, Data Science, Probability & Stochastic Processes, Linear Algebra, Discrete Math, Operating Systems, Combinatorial Data Structures & Algorithms (Graduate), Randomized Algorithms, Computability & Complexity Theory, Programming Languages & Compilers, Security, Comp. Architecture
- **Honors**: Upsilon Pi Epsilon CS Honor Society (top 1/3 of CS majors), College of Letters & Science Honors 2020-2021

Experience

Instalily.ai New York City, NY

AI Engineer – Platform/ML Infrastructure *June 2024 – Present*

- Eng. & Architecture lead of core AI platform `pip` package consisting of multi-agent framework & internal dev tooling
- Deploying & finetuning open-source HuggingFace LLMs & embedding models on GCP for supply chain distributors
- Building robust, scalable data pipelines for RAG agents: e-commerce chatbots, customer service, and email response

Berkeley Artificial Intelligence Research (BAIR) Berkeley, CA

Graduate Researcher – LLM Applications *August 2023 – Present*

- Prompt engineered & deployed LLM AI assistant for CS students via a VS Code extension & command line integration
- Published tool's positive impact on office hour queues & homework completion times at **NeurIPS'23** & **SIGCSE'25**

Amazon Seattle, WA

Software Engineer Intern – Fraud Detection *May 2021 – August 2021*

- Built intelligence collection service to improve threat discoverability via fast searching through large datasets
- Resulted in **30%** improvement in analyst efficiency, saving **300 person-hours** a month at a cost of less than **$2/hour**
- Used serverless AWS infrastructure to implement a scalable, cost-efficient, fault-tolerant, extensible, and secure system

UC Berkeley EECS Berkeley, CA

Lecturer – Intro CS *Summer 2022, Summer 2024*

- Gave lectures, wrote exams, and hired staff of **25+** TAs/tutors and **50+** academic interns for class of **400+** students
- Taught data structures, recursion, OOP, trees, linked lists, complexity, and functional programming in Python & SQL
- Average teaching effectiveness rating of **4.52/5.00** by students, won **Outstanding Graduate Student Instructor Award (2023)**, awarded to **top 10%** of TAs university-wide, and won **Outstanding Academic Intern Award (2020)**, awarded to **top 7%** of Intro CS lab assistants

Projects

Meta (Contract Tech Lead) *February 2024 - May 2024*

- Optimizing CPU operators for ARM architecture using auto-vectorization to speed up Meta's ML workflows

San Francisco Conservatory of Music (Contract Lead Software Engineer) *May 2022 - January 2023*

- Built a dashboard for SFCM to increase concert turnout by parsing, aggregating, and visualizing historical data
- Trained **6** developers with no web dev experience to build a full-stack web app using React, Express, and PostgreSQL

Mothership (Contract Lead Software Engineer) *December 2021 – May 2022*

- Sourced & specced data science/backend project to serve carrier supply & shipment demand density in metro areas
- Led **6** developers through system architecture research, design doc, data analysis, service deployment, and testing

BlueConduit (Contract Software Engineer) *August 2021 – January 2022*

- Built web app for city officials to upload & visualize water service pipeline data for finding best replacement locations
- Part of a multimillion collaboration between BlueConduit and Google.org to support lead service line replacements
- Used Django REST framework & JSON web tokens to handle user authentication and Mapbox API for visualizations

Relativity Space (Contract Software Engineer) *February 2021 – May 2021*

- Developed web app for visualizing real-time time-series data streaming from sensors on rockets into InfluxDB
- Built APIs, sockets, React dashboards, D3 graphs with custom absolute/relative timeranges for multiple data streams
- Emphasized improved performance over Grafana through streamed data caching and client-side shared global state

Skills

Languages: Python, Java, C, SQL, Go, JavaScript, TypeScript, OCaml, LaTeX, Assembly, Lisp, HTML/CSS

Tools & Frameworks: Git, AWS, GCP, Azure, Unix, Linux, Docker, PyTorch, TensorFlow, sklearn, pandas, NumPy, MatPlot, React, Vue, Flask, Django, Express, MongoDB, InfluxDB