Kyle Ming Zhang

669-226-8281 | kylemzhang@gmail.com | linkedin | github

Education

Santa Clara University Santa Clara, CA
Master of Science in Computer Science and Engineering *2023 – 2025*

University of California, Irvine Irvine, CA
Bachelor of Science in Computer Engineering *2020 – 2022*

Experience

Founding AI Engineer March 2025 – Present
Revola AI *San Jose, CA*

- Engineered a real-time, multi-agent meeting infrastructure capable of sustaining 1,000+ concurrent autonomous conversations, **driving a 12% increase in client website traffic to demo bookings**
- Architected and deployed a continual learning system that automatically analyzes and persists meeting context, resulting in a **31% increase in prospect return rates**, and built a corresponding Analytics Dashboard (React, TypeScript) for performance tracking
- **Secured first enterprise customers** by designing and implementing scalable Python onboarding services, which automated knowledge base ingestion, FAISS index generation, and customized agent setup
- **Tripled company website traffic to account creations** by developing a GenAI-powered website scraper, auditing, and scoring system (Python/Google GenAI SDK) and integrating the system with CRM platforms (Hubspot, Zoho) to streamline qualified lead management

HCI Research Lead Sep. 2024 – June 2025
Santa Clara University *Santa Clara, CA*

- Full-stack development for app scraping SMAR research web tool; tripled site load capacity and built RESTful APIs to enable easy-to-use search and querying functionalities, **ensuring 99% uptime for 200+ concurrent users**
- Spearheaded model development for an adaptive UI browser extension aiming to predict user intent for Youtube; experimented with RAG systems, prompt engineering, and fusion models

Software Engineering Intern Jan. 2022 – July 2022
Thales *Irvine, CA*

- Led a team of 4 engineer interns to investigate and integrate third-party services on test servers to enable a fluid microservice environment, enhancing overall interoperability infrastructure
- Utilized Docker, Kubernetes, DAPR, and Bash scripting to execute technical solutions on four different platforms

Projects

Systematic App Reviews | *React, Express, Node, AWS* Sep. 2024 – June 2025

- Engineered a highly responsive web application using React, driving a significant increase in researcher engagement
- Designed a scalable backend and created RESTful APIs using Node.js and Express.js; enabled real-time retrieval of app metadata and rankings from Google Play and iOS App Store across multiple countries;
- Orchestrated the deployment of a full-stack tool on AWS EC2 instances, ensuring high availability and optimal performance

FocusMode | *Python, Pinecone, OpenAI SDK* Jan. 2025 – May 2025

- Built a RAG system with Pinecone, OpenAI GPT-4, and nomic-embed-text, reaching 76% accuracy for pilot study
- Built a fusion model that combines categorical and numerical features encoded using a DeepFM and text embeddings to fuse textual, numerical, and categorical data

Technical Skills

Languages: Python, C, C++, Java, JavaScript, TypeScript, HTML, CSS, SQL, Bash
Frameworks: PyTorch, Tensorflow, React.js, FastAPI, Express, Next.js, Lang(Chain/Graph), Google GenAI SDK
Developer Tools: Node.js, Docker, Kubernetes, Git, Redis, MongoDB, Django, PostgreSQL, MySQL, Merge
Cloud: AWS (EC2, ECS Fargate, S3, SQS, Elasticache, Lambda, CloudWatch), Pinecone