Jiaqi Duan

https://victoria-duan.vercel.app/ jd.victoria.work@gmail.com 408-438-7247

Education

University of California, Santa Cruz

Computer Science and Engineering, Master of Science (M.S.) — *Expected Dec 2025*

Computer Science and Engineering, Bachelor of Science (B.S.); Psychology, Bachelor of Art (B.A.) — *Dec 2022*

Experience

Founding Engineer @ Ripplet | *June 2024 – Present*

- Partnered with therapists and domain experts to co-design **user-facing LLM** features and define product goals, translating clinical insights into actionable UI workflows and system boundaries
- Refactored backend schema and query logic for **40% faster retrieval latency**, and integrated responsive frontend updates to streamline therapist access to real-time client insights during live sessions
- Architected a **HIPAA-compliant, multi-agent** AI system with **multi-modal retrieval augmented generation (RAG)** pipelines to surface client narratives and **evidence-based** psychology

Full Stack Engineer @ Tech4Good Lab | *June 2023 – Jan 2025*

- Led a **cross-functional** team of 10 engineers and designers to develop Pathways, an AI self-directed learning platform
- Engineered a high-performance full-stack architecture (**Solid.js** + **Express.js** + **Firebase**) for real-time **LLM-driven** recommendations, reducing latency and improving frontend responsiveness for seamless user interaction
- Achieved **25% weekly active user growth** in two months through dynamic UI workflows and personalized content
- Designed and tested prompt variants through **iterative A/B testing and user research**, improving LLM-generated content quality and **increasing recommendation relevance by 15%** based on engagement metrics

Coding Instructor @ Code For Fun | *Feb 2023 - Feb 2024*

- Taught programming to cohorts of 1–300 students (ages 6–18), fostering an interactive and inclusive environment
- Developed adaptive, project-based curriculum with **90%** parent satisfaction, guiding students to build websites, data-driven applications, and automation scripts to strengthen **real-world problem-solving skills**

Projects

Large Language Models (LLMs) are Autonomous Cyber Defenders (ACD) | Python

Research project on Explainable AI (XAI) for cybersecurity (supervised by Prof. Alvaro A. Cardenas)

- Co-authored **IEEE CAI 2025** paper *Large Language Models are Autonomous Cyber Defenders*, presented at the conference and published on arXiv: 2505.04843
- Extracted and embedded action-reason statements using **OpenAI's Embeddings API**, converting LLM-generated rationales into high-dimensional vectors for downstream clustering
- Applied **unsupervised machine learning (K-Means, DBSCAN, PCA)** with feature standardization and dimensionality reduction to uncover interpretable behavioral clusters in agent decision-making
- Built a reasoning summarizer driven by **OpenAI GPT-4o** that converts clustered behavior into human-readable defense strategies via **advanced prompting strategies**, advancing explainability and transparency in LLM-driven autonomous systems

Travel Agent | React Native, NativeWind, Redux, FastAPI, PostgreSQL

Master's capstone project on full-stack mobile travel planner powered by multi-agent LLM pipeline (supervised by Prof. Yi Zhang)

- Architected a modular **multi-agent LLM pipeline** using **AutoGen** to generate travel itineraries based on content that is accurate, travel domain specific, and aligned with user preferences and constraints in real-time
- Built a robust **agentic web scraping** module by integrating **Perplexica** for search-based discovery, **Playwright** for dynamic content rendering, and **Trafilatura** for clean content extraction
- Implemented an evaluation layer using **LLM-as-Judge** with **rule-based filtering** to check and rank itinerary suggestions, improving content relevance by **40%** and reducing low-quality recommendations by **20%** compared to a prompt-refined baseline.

Skills

- **LLM Systems & AI Tooling:** AutoGen, OpenAI APIs, Gemini APIs, Google AI Studio, Hugging Face Transformers, Ollama, LlamaIndex, Pinecone, Chroma, Weights & Biases (W&B), LastMile AI, Scikit-learn, PyTorch, TensorFlow, Keras
- **Web & Mobile Development:** React Native (Expo), React, Next.js, SolidJS, Node.js, FastAPI, Express.js, Django, Django REST Framework, Vite, Tailwind CSS, NativeWind, HTML, CSS/Sass
- **Backend & Infrastructure:** PostgreSQL, Firebase (Firestore), Supabase, MongoDB, Docker, NGINX, Google Cloud Platform (GCP), AWS
- **Programming Languages:** Python, TypeScript, JavaScript, Java, C, C++
- **Data & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Plotly, Chart.js
- **DevOps & Tools:** Git, GitHub Actions, CI/CD, Postman, NPM, Jira, Vercel
- **Experimentation & Evaluation:** A/B Testing, Prompt Evaluation, User Feedback Analysis