

Johnny To

☎ 669-265-7854 ✉ tojohnny1559@gmail.com 🌐 [drabblesaur](https://drabblesaur.com) 📄 [johnnyto015](https://johnnyto015.com)

Education

San Jose State University

MS in Software Engineering, Specialization in Data Science

San Jose, California

August 2024 - May 2026

Relevant Coursework: Natural Language Process, Deep Learning, Machine Learning, Enterprise Software

Platforms, Software Systems Engineering

University of California - Riverside

BS in Computer Science

Riverside, California

September 2021 - December 2024

Technical Skills

Languages: Python, JavaScript, SQL

Frameworks: React, React Native, Next.js, Express.js

AI/ML: LangChain, RAG, Prompt Engineering, LLM Benchmarking, FAISS, Ollama, OpenAI API

Data Science: NumPy, Pandas, Jupyter Notebooks

Cloud/DevOps: AWS (Lambda, API Gateway), Docker

Databases: PostgreSQL, MongoDB

Tools: Git, Github, Antigravity, Claude Code, ChatGPT

Projects

SJSU Course & Degree Virtual Assistant | *Python, Langchain, Gemini, Ollama*

[Github](#)

- Built an LLM-based agent system with a dynamic tool-routing pipeline, integrating multi-source retrieval (RAG), query planning, and orchestration across database queries, vector retrieval, and constrained web search.
- Benchmarked Qwen 1.7B vs 14B across representative course-planning queries; selected the 1.7B model for 3x faster inference (20–30s vs 60–100s) after confirming acceptable answer quality, enabling usable response times for student sessions.
- Adversarial prompts risked leaking student data; designed tool-scoped data access binding student IDs to specific tools, blocking cross-user data leakage across 27+ prompt-injection test cases.

LessGo: Smart Carpooling Matching Platform | *Python, FastAPI, Node.js, Claude Code, Antigravity*

[Github](#)

- Designed an 11-service microservice backend covering matching, payments, routing, and safety, enabling parallel team development and isolated failures.
- Proximity-only matching ignored route compatibility; built a three-stage matching pipeline (PostGIS filter, RShareForm HIN embeddings, weighted cost function), reaching 84%+ top-1 accuracy on 100-driver pools.
- Built a real-time safety service comparing live GPS to planned routes, flagging anomalies and alerts for rider safety mid-trip.

Experience

UCR Bourns College of Engineering

Undergraduate Research Assistant

Riverside, California

June 2021 – August 2021

- Tasked with testing a method to reduce matrix multiplication cost; implemented it in NumPy and validated correctness against standard baselines.
- Needed to quantify real gains; built a benchmarking harness measuring step counts across matrix sizes, demonstrating a 33% reduction in computational steps.
- Presented results to an academic audience, giving faculty a basis for applying the method to deep learning workloads.

Leadership

Developers' Guild

Vice President

Cupertino, California

September 2019 – June 2021

- Club assets looked dated and deterred prospective members; led a full redesign of branding and recruitment materials, increasing new student sign-ups by 88%.
- The GitHub organization had accumulated abandoned repositories that obscured active work; established repository standards and an archival policy, reducing junk repos by 25%.