Aaron Yalong

aaronjyalong@gmail.com | San Jose, CA | U.S. Citizen | LinkedIn | GitHub

Education

San José State University | San José, CA — Expected December 2027
Master of Science, Artificial Intelligence
Relevant Coursework: Machine & Deep Learning, NLP, Cloud Technologies, Enterprise Software, Data Mining

University of California, Irvine | Irvine, CA — 2022–2024
Bachelor of Science, Computer Science & Engineering — GPA 3.8, Cum Laude

Technical Skills

Languages: Python, Scala, Java, TypeScript/JavaScript, SQL, PowerShell, C/C++, HTML/CSS
AI / LLM: RAG, hybrid retrieval, MCP, prompt engineering, LLM evaluation, embeddings, Claude API, LangChain4j, Ollama
APIs & Data: REST APIs, FastAPI, Spring Boot, OAuth2, Chroma (vector DB), DuckDB, MongoDB, pandas, scikit-learn
Cloud & DevOps: AWS (EC2, S3, Bedrock), Docker, Terraform, Git, Mercurial, GitHub Actions, CI/CD, Tomcat
JVM & Tools: GraalPy, JavaFX, Mill, Maven, Gradle, IntelliJ, VS Code, Claude Code, Ubuntu (Linux), WSL2

Experience

Intuitive Machines | Palo Alto, CA — March 2026–Present
Junior Software Engineer — *Part-Time*

- Build and maintain a production agentic AI assistant (Scala, LangChain4j, Claude) on secure internal infrastructure
- Integrated the assistant with manufacturing systems so engineers can query live production data in natural language, using LLM-generated Python run in an embedded runtime and agent skill files
- Drove adoption of per-user OAuth2 authentication for enterprise integrations after reviewing alternatives for security risks
- Own the assistant's presentation tooling via MCP: vetted open-source MCP servers for maintenance and enterprise readiness, then adopted and customized a C#/.NET server
- Leading natural-language CAD automation: proved a connection to live CAD sessions; now building model generation
- Cut time-to-window from ∼10s to ∼3s by moving tool initialization to a background thread and adding a splash screen
- Document and present integrations for code review; shipped multi-chat and persistent chat history
- Develop and support an internal Java schematic editor: build SVG/XML part definitions, implement part selection and connector UI, and troubleshoot issues for engineering users

Handshake AI | Remote — January–December 2025
AI Trainer — Computer Science & AI Domain Expert (Freelance)

- Evaluated LLM outputs for technical accuracy and reasoning quality; wrote prompts to surface edge cases and failure modes

Luminexa | Remote — Irvine, CA — August–December 2024
Junior Software Developer

- Built responsive web interfaces and Spring Boot REST endpoints supporting core application features

Projects

Oracle — Local RAG for Obsidian | *Python, Ollama, Chroma, FastAPI, TypeScript, MCP* — 2026–Present

- Fully local cited Q&A over Obsidian notes with an MCP server and plugin; hybrid vector + BM25 note search hit 0.90 top-3 on held-out questions vs. 0.80 vector-only
- Ran model × prompt experiments, then added a deterministic grounding check against cited notes, halving fabricated answers

Logbook — Grounded QA over NASA Safety Reports | *Python, Claude API, Ollama, DuckDB, Streamlit* — Fall 2026

- LLM assistant over 3,080 NASA ASRS maintenance reports: a tool call turns questions into validated JSON filters run in DuckDB, then hybrid BM25 + dense retrieval feeds answers that cite every claim by report ID
- Checks every citation against the retrieved set to flag fabrications; benchmarked keyword, TF-IDF, embedding, and Claude zero-shot classifiers on a coded field (best 0.82 F1)

Alfred — Practice Exam Generator | *Python, Pydantic, Claude API* — 2026–Present

- CLI pipeline that ingests PDF, PPTX, DOCX, and diagram-heavy pages (via Claude vision) and generates schema-validated quizzes; deterministic checks confirm every cited quote is on its page and options are well-formed

Autonomous Vehicle Cloud Platform | *AWS, SQL, MongoDB, CARLA* — Fall 2026

- Back-end lead on a 5-person team linking CARLA simulations on SJSU's HPC cluster to an AWS back end; designing the API, SQL/MongoDB data models, and alert workflow