

Jeffrey Saelee

Software Engineer · Backend & AI systems

Visalia, CA | 559.936.6115 | jeffrey@jeffreysaelee.com
jeffreysaelee.com | linkedin.com/in/jeffrey-saelee | github.com/Servation

PROFESSIONAL SUMMARY

Software Engineer with 3+ years building AI-integrated, high-throughput backend systems for Apple via Wipro. Specialized in Java/Python microservices, LLM/RAG pipelines, and cloud observability across AWS and Kubernetes. Actively shipping full-stack and AI projects.

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, TypeScript

Frameworks: Spring Boot, Maven, React, Angular, Hibernate, JUnit, FastAPI

Data & infrastructure: AWS, Docker, Kubernetes, MongoDB, Cassandra, MySQL

Observability & testing: Grafana, Splunk, Prometheus, JMeter, Postman, Git

AI: LLM integration, Agent orchestration, RAG, Model Context Protocol (MCP), Claude Code, Google AI

WORK EXPERIENCE

Software Engineer | Apple Inc. (Contract via Wipro)

Aug 2022 – Sep 2025

Austin, TX · Cupertino, CA · Remote

- Built custom LLM-powered code generation tooling and validated output quality against a baseline using similarity metrics, achieving ~30–40% improvement in generated code relevance.
- Contributed to migrating a high-volume audit application (1M+ daily transactions) from Cassandra to MongoDB by rewriting queries for API changes, building new backend services, and validating data integrity between systems, reducing query latency from 10,000ms+ to under 1,000ms.
- Built custom PyCharm and IntelliJ IDE extensions in Python and Java to automate repetitive developer setup and reduce boilerplate overhead across the team.
- Configured Grafana and Splunk dashboards to monitor query performance and track bugs across backend services, supporting load testing and production observability.

TECHNICAL PROJECTS

Kindred | Google × Kaggle AI Agents Intensive capstone

2026

Python, Google ADK, MCP, RAG, Local LLM (Gemma), React, TypeScript

- Built a multi-agent LLM app that helps family caregivers of elderly relatives manage scheduling and medication, running a local LLM so private health data never leaves the machine.
- Engineered MCP tool routing with human approval gates and redaction for sensitive actions, where the model plans and deterministic Python validates and executes, backed by a RAG pipeline.

Pokémon TCG Battle Agent | Kaggle Pokémon TCG AI Battle Challenge

2026

Python, Monte Carlo search, Imitation learning, Self-play RL

- Built a CPU-only game-playing agent that competes on the live skill-rating ladder, layering determinized Monte Carlo search and policy-guided rollouts over a heuristic evaluation core, with a never-crash safety net and p99 move time under 0.2 ms.
- Ran 45+ gated experiments (paired ladder A/Bs, SPRT, Wilson bounds) and trained imitation-learned policy priors on a 640k-row dataset mined from replay episodes, hitting a peak provisional ladder rating of 972.

OTTR AI agent trading system <i>React, FastAPI, Discord.py, LM Studio, CoinGecko/Kraken APIs</i> - Designed a crypto trading desk run by 7 autonomous AI agents that meet, debate, and vote on trades in a Discord channel, featuring an LLM intent router for live CEO directives. - Engineered semantic memory recalling past meetings via local LLM embeddings, an autonomous risk engine enforcing stop-loss, drawdown breakers, and position caps, and a React terminal mirroring the trading floor.	2026
Memora Full-stack flashcard learning platform <i>Java 17, Spring Boot, React, Vite, Apache PDFBox</i> - Created a full-stack flashcard app featuring AI-powered note parsing (PDF/TXT/MD to cards) and an interactive 3D flip card UI. - Implemented multi-LLM support (OpenAI, Gemini, Ollama) and cognitive science formatting to generate atomic cards following the Minimum Information Principle, secured via HMAC-SHA256 stateless auth.	2026
Job Search Agent AI ATS crawler <i>TypeScript, React, Express, Gemini</i> - Built an AI-powered job search cockpit to directly crawl ATS platforms (Greenhouse, Lever, Ashby, Workday) and automatically discover Workday tenants. - Evaluates job listings against applicant resumes using Gemini, scoring matches and tracking the submission pipeline.	2026
Vesper Local RAG context-engineering tool <i>Python, TF-IDF, Streamlit, MCP</i> Built an offline "intent isolation" tool that uses TF-IDF to extract the most topically dense passages from large Markdown corpora into XML-tagged context packets, cutting prompt noise for downstream LLMs with zero embeddings or cloud calls (100% local).	2026
AuraSkill Hot-swappable agent skill runtime <i>Python, Subprocess isolation</i> - Implemented drop-in skill folders dynamically loaded and executed by LLM agents, using progressive context loading (metadata → instructions → resources) to save tokens. - Designed sandboxed subprocess execution based on published arXiv agent research.	2026
Lore Spinner Multi-agent narrative RPG <i>Python, Multi-agent AI, MCP, D&D 5e SRD</i> - Developed a text-based narrative RPG powered by a multi-agent AI Dungeon Master with 4 autonomous subagents (World Keeper, Faction Weaver, Encounter Architect, Lore Keeper). - Designed dynamic story spines, anti-lock loop detection, autonomous faction projects, and integrated the D&D 5e SRD via Model Context Protocol (MCP).	2026

EDUCATION

B.S., Software Engineering Western Governors University	Expected 2027
A.S., Computer Programming and Information Systems College of the Sequoias Visalia, CA	2021

ADDITIONAL INFORMATION

Languages: English (native), Mienh (conversational)
Interests: 3D printing, custom hardware design, home automation, travel, photography, sci-fi and fantasy literature