Alex Weaver

Senior AI Software Engineer

Round Rock, TX alexweaver0666@gmail.com (321) 386-0666 linkedin.com/in/alex-weaver-7128a1449

SUMMARY

Senior AI Software Engineer with 8+ years of experience designing, developing, and deploying production-grade AI applications, Agentic AI systems, and scalable cloud-native microservices across healthcare, enterprise SaaS, observability, and logistics domains. Experienced in building multi-agent architectures using Python, FastAPI, LangGraph, CrewAI, AWS Bedrock AgentCore, and Anthropic Claude, alongside OpenAI API integrations and enterprise RAG solutions. Strong expertise in MCP integrations, vector search, LLM evaluation, AI security, distributed systems, and cloud infrastructure automation. Proven ability to architect high-performance microservices, implement production observability using Grafana, OpenTelemetry, Langfuse, and Datadog, and optimize AI application performance, reliability, and scalability.

TECHNICAL SKILLS

Programming Languages: Python, Go, TypeScript, JavaScript (ES6), SQL

AI & Machine Learning: Generative AI, Large Language Models (LLMs), Agentic AI, Multi-Agent Systems, OpenAI API, Anthropic Claude, AWS Bedrock, AWS Bedrock AgentCore, LangChain, LangGraph, CrewAI, LlamaIndex, Vision-Language Models (VLMs)

RAG & Knowledge Retrieval: Retrieval-Augmented Generation (RAG), Pinecone, PostgreSQL, pgvector, Vector Embeddings, Hybrid Search, Reranking, Metadata Filtering, Document Processing, Model Context Protocol (MCP), Agent2Agent (A2A)

AI Evaluation & Security: DeepEval, Ragas, LangSmith, LLM-as-Judge, Agent Trajectory Evaluation, Golden Test Cases, Prompt Injection Detection, Jailbreak Detection, PII/PHI Redaction, AI Guardrails, Responsible AI, AI Governance, Red-Team Testing

Backend Engineering & APIs: FastAPI, Async Python, Go, Node.js, Express.js, RESTful APIs, Microservices Architecture, Distributed Systems, MCP Servers, OAuth-Based Authentication, Structured Outputs, Tool Calling, WebRTC

Frontend Development: React, Next.js, TypeScript, JavaScript (ES6), Grafana UI Components, Interactive Dashboards, HTML5, CSS3, Bootstrap

Cloud, DevOps & Infrastructure: AWS, Microsoft Azure, Amazon EKS, AWS Lambda, AWS EventBridge, Amazon SQS, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Infrastructure as Code (IaC), Automated Testing, Cloud-Native Deployments

Data Engineering & Databases: Apache Kafka, Databricks, Snowflake, PostgreSQL, MySQL, Redis, SQL Server, Data Pipelines, Real-Time Data Processing, Data Modeling

Observability & Monitoring: Grafana, OpenTelemetry, Langfuse, Datadog, Metrics Monitoring, Log Aggregation, Distributed Tracing, Telemetry Exploration, Application Monitoring, Dashboard Development, Production Troubleshooting, Performance Monitoring

AI Performance & Optimization: Semantic Caching, Prompt Caching, Model Routing, Token Optimization, Streaming Responses, Latency Optimization, Inference Cost Optimization, Performance Optimization

PROFESSIONAL EXPERIENCE

Senior Software Engineer | **Asurion**

February 2024 – October 2026, Houston, TX

- Architected and delivered scalable, cloud-native AI applications and microservices using Python (async), FastAPI, TypeScript, React, Next.js, and AWS, powering intelligent solutions across customer care, connected health, logistics, and enterprise operations while serving 2M+ monthly requests.
- Designed and orchestrated multi-agent AI systems using LangGraph, CrewAI, AWS Bedrock AgentCore, and Claude, implementing tool calling, structured outputs, short- and long-term agent memory, human-in-the-loop approvals, and retry/fallback strategies for reliable autonomous task execution.
- Developed Model Context Protocol (MCP) servers to securely expose internal APIs, databases, and enterprise knowledge sources using scoped permissions and OAuth-based authentication, enabling Agent2Agent (A2A) communication, cross-team collaboration, and intelligent task delegation across agent frameworks.

- Engineered and optimized production-grade RAG pipelines using LlamaIndex, Pinecone, PostgreSQL/pgvector, and embedding models, combining hybrid search, reranking, metadata filtering, and optimized chunking strategies to improve retrieval accuracy by 35% across 5M+ enterprise documents.
- Created LLM evaluation frameworks using DeepEval, Ragas, and LangSmith with 1,000+ golden test cases, LLM-as-judge scoring, and agent trajectory evaluation. Added CI/CD gates, raising answer faithfulness from 78% to 94% and preventing quality regressions.
- Implemented enterprise AI security, governance, and responsible AI controls, including prompt injection and jailbreak detection, PII/PHI redaction, output schema validation, tool-level permission policies, and red-team testing to support HIPAA-aligned security and compliance requirements.
- Implemented end-to-end LLM observability using OpenTelemetry, Langfuse, and Datadog, optimizing prompt engineering, semantic caching, model routing, token budgets, and streaming responses to reduce inference costs by 40% and latency by 45% compared with the pre-implementation baseline.
- Delivered AI-powered health coaching applications, digital health twins, and real-time voice assistants using RAG, LLM tool calling, streaming STT/TTS, and WebRTC, enabling personalized wellness recommendations and conversational experiences with sub-800ms latency.
- Automated multimodal document processing and logistics workflows using vision-language models, LLM-based extraction, Kafka, AWS EventBridge, SQS, and business rules engines, achieving 95%+ extraction accuracy, reducing manual processing by 80%+, and exception-handling time by 60%.
- Led AWS production deployments and scalable data pipelines using Databricks, Snowflake, and Python, reducing deployment time by 70% while mentoring engineers in agentic architecture, evaluation-driven development, and AI security.

Software Engineer | **Grafana Labs**

January 2020 – February 2024, New York, NY

- Maintained observability platforms and developed backend services using Go, React, TypeScript, and RESTful APIs, supporting telemetry exploration, application monitoring, and production troubleshooting across distributed systems.
- Optimized telemetry query workflows and backend integrations for metrics, logs, and traces, improving query responsiveness and enabling engineering teams to investigate service performance and resolve production issues.
- Built interactive dashboards and configuration interfaces using React, TypeScript, and Grafana UI components, enabling users to visualize operational data, configure monitoring workflows, and identify application issues.
- Launched conversational AI applications using Python, FastAPI, React, Next.js, and OpenAI APIs, integrating LLMs and Retrieval-Augmented Generation (RAG) with LangChain, Pinecone, PostgreSQL, and vector embeddings to improve enterprise information retrieval accuracy by 20%.
- Streamlined intelligent document-processing pipelines and AI-powered workflow automation using Python and LLM-based extraction, automating classification, summarization, and structured data generation while reducing processing time by 30% and repetitive manual tasks.
- Engineered scalable backend services and cloud-native AI microservices using FastAPI, PostgreSQL, Redis, Docker, Kubernetes, Terraform, AWS, and GitHub Actions, implementing CI/CD, Infrastructure as Code (IaC), automated testing, and deployment improvements that increased engineering efficiency by 25%.

Full Stack Developer | **National Instruments**

January 2018 – January 2020, Austin, TX

- Developed responsive web applications using React, JavaScript (ES6), HTML5, CSS3, and Bootstrap, delivering interfaces optimized for desktop and mobile platforms.
- Connected RESTful APIs and backend services using Node.js, Express.js, and Python, enabling reliable communication between frontend applications, backend systems, and databases.
- Modeled relational databases using MySQL and PostgreSQL, implementing efficient SQL queries, data models, and database operations to improve application performance and data integrity.
- Accelerated cloud-based application deployments through containerization and CI/CD pipelines using Microsoft Azure, Docker, and automated testing, reducing deployment time by 40% while improving application performance and platform reliability with cross-functional Agile teams.

EDUCATION

Bachelor's degree in Computer Science | **Texas Tech University**

August 2013 – December 2017, Lubbock, TX