WEI JIANG

Senior Software Engineer

AI Products • Full-Stack • Generative AI

(839) 868 - 1380 • wei.builds.ai@outlook.com • San Diego, CA • Linkedin • Github

Summary

Senior Software Engineer with **14** years of experience architecting scalable **B2B SaaS** platforms, distributed systems, and AI-native products. Delivered high-impact systems across **Stripe payments**, **Uber mobility** infrastructure and **Perplexity Agentic AI** platforms. Skilled in building intelligent applications that combine robust backend architectures, cloud platforms, and modern AI technologies.

Experience

Perplexity AI

Member of Technical Staff *Feb 2023 - Present | Remote*

Agentic AI Research Platform · Multi-Agent AI Workflows · Enterprise RAG & Retrieval Systems · 20+ Foundation Models

Built an enterprise-grade **Agentic AI Research Assistant platform** that extended **Perplexity's AI search** capabilities with autonomous research workflows, secure data integrations, and production-scale LLM infrastructure.

- Architected and delivered a production-scale **Agentic AI Research platform** using **Python, Java, AWS Bedrock, MCP** and a **custom agent orchestration runtime**, enabling multi-agent workflows for planning, retrieval, analysis, and citation-grounded answer generation.
- Built an **Orchestrator-Worker agent framework** that enabled dynamic task planning, intelligent model routing, recursive agent execution, context management, and autonomous recovery for complex, long-running research workflows.
- Developed an advanced **AI search and RAG platform** using **Vespa, hybrid retrieval, semantic ranking, and citation validation**, combining lexical search, vector retrieval, and multi-stage ranking across enterprise knowledge sources to achieve **90%+ citation-grounded answer accuracy** in internal benchmarks.
- Implemented an **MCP-based tool execution layer** that connected AI agents with enterprise applications, APIs, databases, and knowledge sources through standardized interfaces, secure authentication, and controlled context exchange.
- Engineered an **AWS Bedrock-powered LLM platform** with model routing, prompt management, guardrails, and inference optimization, orchestrating **20+ foundation models** to balance response quality, latency, and cost.
- Built **LLM evaluation and observability** infrastructure using **RAGAS, LangSmith, and production telemetry** to analyze agent trajectories, retrieval quality, citation accuracy, faithfulness, latency, cost, and task success.
- Designed scalable backend services and event-driven AI pipelines using **Python, Java, C++, FastAPI, Kafka, Vespa, PostgreSQL, Redis** and **AWS**, enabling asynchronous agent execution, workflow orchestration, telemetry processing, and reliable data persistence.
- Developed AI application key interfaces and supporting API services using **React, TypeScript** and **Node.js,** enabling users to interact with autonomous research workflows, agent execution states, and research outputs.
- Built and operated cloud-native infrastructure on **AWS EKS**, implementing container orchestration, autoscaling, observability, and reliability engineering practices to support **hundreds of millions of AI search and inference requests per month**.

***Tech Stack**: Python, Java, C++ , AWS (EKS, Bedrock), MCP, Claude/Llama/Gemini, Agentic AI, Multi-Agent Orchestration, RAG, Vespa, RAGAS, LangSmith, Vector Search, Hybrid Retrieval, Semantic Reranking, FastAPI, Kafka, PostgreSQL, Redis*

Uber

Senior Software Engineer *Jul 2018 - Jan 2023 | San Francisco, CA*

Real-Time Marketplace Systems · 40M+ Daily Trips · 1M+ Location Events/sec · ML-Powered Decision Systems

Built **Uber's Marketplace Intelligence Platform**, combining distributed backend infrastructure with ML-powered optimization systems to improve dispatch, matching, ETA prediction, and real-time transportation decisions.

- Built backend services for Uber's **Marketplace Fulfillment Platform**, supporting trip lifecycle management, dispatch orchestration, and rider-driver matching workflows across **40M+ daily trips** and global mobility operations.
- Architected distributed marketplace systems using **Java, Go, Python, Kafka, Redis, Cassandra, gRPC, and Kubernetes**, processing **1M+ location events/sec** to maintain real-time driver availability, trip state, and geospatial marketplace intelligence.
- Developed event-driven **dispatch and matching systems** combining real-time marketplace signals, distributed state management, and optimization logic to enable **sub-100ms matching workflows** and reduce dispatch-related failures by **35%**.
- Integrated **ML-powered prediction services** for **ETA estimation, demand forecasting, and marketplace optimization**, leveraging real-time mobility signals, geospatial features, and historical patterns to improve ETA accuracy by **15%**.
- Built real-time marketplace intelligence pipelines using **Kafka, Flink, and H3 geospatial indexing** to process mobility signals, support supply-demand analysis, and enable data-driven transportation decisions.
- Developed marketplace experimentation and simulation platforms to evaluate optimization strategies, validate ML-driven improvements, and safely test dispatch and pricing changes before production rollout.
- Designed resilient **REST/gRPC APIs and backend services** with fault-tolerant workflows, idempotency, distributed coordination, and observability capabilities to support mission-critical marketplace operations.

***Tech Stack**: Java, Go, Python, Kafka, Flink, Cassandra, Redis, PostgreSQL, gRPC, Kubernetes, AWS, H3 Geospatial Indexing, Real-Time Streaming, Distributed Systems, ML Model Integration, Marketplace Optimization*

Stripe

Software Engineer II (Promoted 2016) *Aug 2012 - Jul 2018 | San Francisco, CA*

Payments Infrastructure · Financial Systems · API Platform · Transaction Processing

Built customer-facing payment capabilities for **Stripe's platform**, enabling merchants to process payments through secure APIs, reliable workflows, and scalable financial infrastructure.

- Developed backend services powering Stripe's **payment processing platform**, supporting **$20B+ annual payment volume across 100K+ merchants** through reliable APIs, transaction workflows, and scalable financial services.
- Implemented distributed payment processing components using **Ruby, Java, Python, PostgreSQL, Redis, and Kafka**, adding idempotency handling, retries, and failure recovery mechanisms to maintain transaction correctness and reduce duplicate payment failures by **99.9%**.
- Built merchant-facing API capabilities supporting **payment authorization, capture, refunds, webhooks, and reconciliation workflows**, improving payment lifecycle reliability and reducing operational issues across Stripe's growing platform.
- Optimized backend services and financial data workflows through query improvements, caching strategies, and service enhancements, accelerating merchant reconciliation workflows by **4x** and improving overall platform efficiency.

***Tech Stack**: Ruby, Java, Python, PostgreSQL, Redis, Kafka, REST APIs, Microservices, Distributed Systems, Event-Driven Architecture, Payment Processing, Transaction Workflows*

Skills

Languages: Python, Java, Go, Ruby, TypeScript, JavaScript, SQL

Backend: FastAPI, Spring Boot, Ruby on Rails, REST APIs, gRPC, Microservices, Event-Driven Systems, API Design, Backend Services

Frontend: React, Next.js, TypeScript, HTML/CSS, Modern Web Applications

AI Technology: Agentic AI, Multi-Agent Systems, LLM Applications, RAG, MCP, AWS Bedrock, Claude, Llama, Gemini, Prompt Engineering, RAGAS, LangSmith, LLM Evaluation, AI Agents

Database: PostgreSQL, MySQL, Cassandra, Redis, DynamoDB, Vector Databases, Data Modeling, Query Optimization

Cloud / DevOps: AWS, Amazon EKS, Kubernetes, Docker, Terraform, CI/CD, Jenkins, GitHub Actions, Cloud-Native Architecture

Architecture: Distributed Systems, Real-Time Systems, Event-Driven Architecture, Streaming Systems, High-Availability Systems, Fault-Tolerant Design, System Design, Scalability Engineering

Observability: OpenTelemetry, LangSmith Tracing, Prometheus, Grafana, Datadog, Distributed Tracing, Logging, Metrics, Alerting

Others: Kafka, Apache Flink, Vespa, H3 Geospatial Indexing, Webhooks, Idempotency, Feature Flags, Agile/Scrum

Education

Bachelor of Science - BS, Engineering Mathematics and Statistics
University of California, Berkeley *2008 – 2012*