KEVIN STEVERSON

Staff AI Engineer | Machine Learning Engineer | Full-Stack Architect

United States | +1 (315) 522-1306 | kevin.steverson1026@outlook.com

PROFESSIONAL SUMMARY

Staff AI Engineer and Full-Stack Architect with **11 years of experience** designing, building, and deploying intelligent software products and scalable enterprise applications. Extensive experience across traditional machine learning, deep learning, generative AI, natural language processing, voice AI, speech technologies, and intelligent automation.

Strong expertise in developing end-to-end AI solutions, including predictive models, recommendation engines, conversational voice agents, real-time speech processing systems, RAG pipelines, and autonomous AI agents. Combines advanced AI engineering capabilities with hands-on full-stack development, cloud architecture, distributed systems, and production infrastructure.

Experienced in leading technical initiatives, designing AI architectures, optimizing model performance, and delivering reliable software products from initial concept through production deployment.

TECHNICAL SKILLS

AI & Machine Learning	Machine Learning, Deep Learning, Neural Networks, Supervised & Unsupervised Learning, Classification, Regression & Clustering, Predictive Modeling & Recommendation Systems, Computer Vision & Image Recognition, NLP, Feature Engineering & Model Optimization, Model Evaluation & Statistical Analysis, PyTorch, TensorFlow, Scikit-learn, XGBoost
Generative AI & Agentic Systems	Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), AI Agents & Multi-Agent Architectures, OpenAI, Anthropic Claude, Google Gemini, LangChain, LangGraph, LlamaIndex, Embeddings, Semantic Search & Vector Databases, Fine-Tuning, Prompt Engineering & Guardrails, Pinecone, Weaviate, Chroma, pgvector, AI Evaluation & Observability
Voice AI & Speech Engineering	Automatic Speech Recognition (ASR), Speech-to-Text (STT) & Text-to-Speech (TTS), Real-Time Voice Agents & Conversational AI, Audio Processing & Speech Analytics, Voice Activity Detection (VAD), Streaming Audio & WebRTC, Whisper, Deepgram, ElevenLabs, Twilio Voice & Telephony Integrations
Backend & Full-Stack	Python, FastAPI, Django, Flask, Node.js, Express.js, TypeScript, React, Next.js, Vue.js, REST APIs, GraphQL, WebSockets, Microservices & Event-Driven Architecture, PostgreSQL, MySQL, MongoDB, Redis, Scalable SaaS & Distributed Applications
Cloud, Infrastructure & MLOps	AWS, Google Cloud Platform, Azure, Docker, Kubernetes, Terraform, MLflow, Weights & Biases, CI/CD, GitHub Actions, Model Deployment & Monitoring, Data Pipelines & Workflow Orchestration, Langfuse, Sentry, OpenTelemetry

PROFESSIONAL EXPERIENCE

Staff AI Engineer

2024 – 2026

MILL5 · United States

- Architected and delivered production AI platforms integrating LLMs, traditional machine learning, voice AI, and autonomous agents, reducing manual operational workloads by **65%**.
- Built an AI evaluation framework using golden datasets, automated regression testing, and quality metrics, reducing model regression detection time from **9 days to under 4 hours**.
- Implemented intelligent model routing and semantic caching, moving **70% of AI traffic to smaller models** without measurable quality degradation and reducing inference costs by **52%**.
- Developed RAG pipelines combining hybrid search, embeddings, reranking, and contextual retrieval, improving answer relevance by **38%**.
- Engineered real-time conversational voice agents using ASR, LLM orchestration, and TTS, reducing average voice interaction latency by **45%**.
- Designed multi-agent automation workflows integrating external APIs and browser-based execution, reducing repetitive business process execution time by **60%**.
- Built scalable Python and FastAPI services supporting **10,000+ daily AI requests** with asynchronous processing and distributed execution.
- Established AI engineering standards, architecture reviews, and automated quality gates across **4 engineering workstreams**.

Senior AI / Full-Stack Engineer

2021 – 2024

WellSky · United States

- Developed predictive machine learning models for operational forecasting and classification, improving prediction accuracy by **24%** over baseline models.
- Architected NLP and document intelligence pipelines processing **1.5 million+ healthcare documents annually**, reducing manual data extraction efforts by **68%**.
- Built semantic search and knowledge retrieval capabilities using embeddings and vector databases, improving information retrieval relevance by **35%**.
- Implemented machine learning classification workflows using Python, PyTorch, and Scikit-learn, reducing manual review requirements by **42%**.
- Integrated speech recognition and conversational AI capabilities into enterprise application workflows, reducing average information processing time by **30%**.
- Developed high-performance APIs using FastAPI, Node.js, and PostgreSQL, improving average API response times by **40%**.
- Redesigned asynchronous data processing workflows, increasing processing throughput by **3.2x**.
- Improved production reliability through automated testing, monitoring, and deployment improvements, reducing critical production incidents by **38%**.

Machine Learning / Software Engineer

2018 – 2021

Health Catalyst · United States

- Developed predictive analytics models using Python, Scikit-learn, and TensorFlow, improving model performance by **22%** compared with existing statistical baselines.
- Engineered data pipelines processing **5 million+ records per month**, reducing data preparation time by **55%**.
- Implemented classification and anomaly detection models to identify operational patterns, improving detection precision by **27%**.
- Developed NLP pipelines for extracting structured information from unstructured documents, reducing manual data processing by **48%**.
- Optimized machine learning feature engineering and model evaluation workflows, reducing model experimentation time by **35%**.
- Built backend APIs and data services using Python, Django, Flask, and PostgreSQL, improving API throughput by **45%**.
- Increased automated test coverage from **41% to 88%**, contributing to a **61% reduction in year-over-year production incidents**.
- Collaborated with data science and engineering teams to deploy machine learning models into production applications.

Software Engineer

2015 – 2018

PayNearMe · United States

- Developed payment processing APIs and backend services handling **100,000+ monthly transactions** across financial technology workflows.
- Optimized transaction processing services and database queries, reducing average processing latency by **35%**.
- Implemented transaction monitoring and fraud detection workflows, improving suspicious activity identification by **26%**.
- Developed data validation and risk assessment components, reducing payment processing errors by **32%**.
- Built responsive React applications and reusable frontend components, reducing feature development time by **30%**.
- Integrated third-party payment gateways and external services, reducing payment integration implementation time by **40%**.
- Improved backend reliability through automated testing, logging, and error handling, reducing recurring production defects by **28%**.
- Collaborated with product and engineering teams to deliver **12+ major application features** across payment and transaction management platforms.

EDUCATION

Bachelor of Science (B.S.) in Computer Science

2010 – 2015

University of Arizona · Tucson, Arizona, United States

ADDITIONAL STRENGTHS

End-to-End AI Product Architecture • Traditional ML & Generative AI Integration • Voice AI & Real-Time Conversational Systems • Enterprise Software Architecture • Technical Leadership & Engineering Mentorship • Cloud-Native Application Development • AI Infrastructure & MLOps • System Design & Performance Engineering