Kevin Shah

Houston, TX (Open to Canada & Remote) | linkedin.com/in/kevin-shah-2207/ | (602) 653-7420 | kevinjshah2207@gmail.com

PROFESSIONAL SUMMARY

AI Engineer with 3.5+ years of experience building and operating production-grade LLM systems, agentic workflows, and cloud-native ML platforms. Specialized in RAG pipelines, LLM agents, prompt engineering, and generative AI infrastructure on AWS. Former Amazon SDE with strong backend foundations and a track record of taking AI systems from prototype to production. Passionate about building reliable, context-aware AI that solves real-world problems at scale.

WORK EXPERIENCE

DNV **Houston, TX**

Software Developer, Machine Learning Operations *May 2023 - Current*

- Designed and owned a production-grade LLM orchestration platform enabling internal teams to build, deploy, and manage agentic workflows with persistent state, multi-step reasoning, and fault tolerance using Python, FastAPI, LangChain, and LangGraph.
- Built a persistent knowledge synthesis system using the LLM Wiki pattern — LLM-compiled, cross-referenced repo documentation maintained incrementally as source repos change — reducing token usage by 60%+ and LLM response latency by 70%+ compared to naive full-context injection.
- Built a prompt management system with an LLM-based prompt optimizer, enabling systematic versioning, evaluation, and iterative refinement of prompts — reducing hallucinations, improving output reliability, and cutting a 20-hour manual task to 3–4 hours (~85% reduction).
- Built a long-term memory system for agent infrastructure using MongoDB vector embeddings, retrieving the top-k semantically similar user memories per query to provide personalized, context-aware LLM responses at scale.
- Deployed AWS Bedrock Guardrails across multiple environments using CloudFormation IaC, enforcing content safety, PII filtering, and hallucination mitigation policies at the LLM inference layer.
- Implemented end-to-end observability for LLM systems (distributed tracing, token usage metrics, agent step dashboards), significantly reducing incident triage time and improving visibility into model behavior in production.
- Led backend design discussions and architectural reviews for new platform features, making trade-offs around scalability, cost, and reliability to support long-term system growth.
- Architected serverless backend services using AWS Lambda, ECS, and S3, reducing infrastructure costs by ~50% while improving scalability and operational reliability.
- Built a platform-agnostic workflow execution system inspired by state-machine architectures, integrating ML workloads via SageMaker and lightweight backend workers, resulting in ~40% lower execution latency.
- Implemented end-to-end observability (distributed tracing, metrics, dashboards), significantly reducing incident triage time and improving production visibility.
- Developed and maintained RESTful APIs connecting ML models to internal products, improving data retrieval performance and system throughput.
- Improved deployment reliability by standardizing build and deployment workflows for backend services, reducing rollout issues and improving developer productivity.
- Partnered with QA and product teams to debug complex production issues, accelerating release cycles and improving service reliability.
- Mentored junior engineers and interns through code reviews and sprint planning, contributing to improved delivery quality and team velocity.
- Contributed frontend features using Vue.js and Nuxt.js to support internal tools, primarily focused on enabling backend-driven workflows and improving usability.

Amazon.com Services LLC **Austin, TX**

Software Development Engineer *Jun 2022 - Mar 2023*

- Led backend changes supporting customer experience analysis across 21 global marketplaces, contributing to a zero-downtime migration impacting 550M+ users.
- Built runtime monitoring and alerting for latency and customer-impact metrics, enabling real-time detection of production issues at global scale.
- Designed and implemented backend components for a large-scale Order Summary system, integrating with multiple critical services and legacy systems.
- Participated in on-call rotations, independently diagnosing and resolving high-severity production incidents and contributing to operational reviews for a global team.

SKILLS

Languages & Scripting: Python, JavaScript, SQL, Bash
Backend & APIs: FastAPI, Node.js, RESTful APIs
Cloud & DevOps: AWS (Lambda, ECS, ECR, S3, SageMaker, CloudWatch), Docker, Git
Databases: MongoDB, PostgreSQL, Redis
MLOps / AI: LangChain, LangGraph, MCP, LiteLLM, RAG Pipelines, Prompt Engineering & Optimization, AWS Bedrock, Bedrock AgentCore, Bedrock Guardrails, Agentic Workflows, Embeddings & Context Retrieval
Testing & Reliability: Unit testing, Functional testing, Production Observability
Frontend Frameworks: Vue.js, Nuxt.js, D3.js

PROJECTS

AlgoTrader (In Progress)

Alpaca API, Python, React.js

- Developing a backend-focused algorithmic trading application using Python (FastAPI), React, and the Alpaca API to backtest and simulate quantitative trading strategies.
- Engineering a data pipeline to ingest historical market data via the Alpaca API, enabling the calculation and analysis of technical indicators like MACD, RSI, and Bollinger Bands.
- Designing a simulation engine to execute trades in a paper account based on strategy signals, with performance visualized through a custom React and TradingView charting interface.

EDUCATION

Arizona State University **Tempe, AZ**
Master of Science in Computer Science (3.81/4.00) *May 2022*
Sardar Vallabhbhai National Institute of Technology (NIT Surat) **Surat, India**
Bachelor in Computer Engineering *Jul 2020*