

Sai Likhith Kanuparthi

860-620-4718 | sailikhithcse@gmail.com | linkedin.com/in/sailikhithk | netflix.sailikhith.me | github.com/sailikhithk

Summary

Senior Software Engineer (AI/ML Platform) with **7+ years** owning end-to-end architecture of **multi-tenant ML platforms**, **foundation model orchestration**, and **agentic AI frameworks**. At Airbnb, owned architecture and production rollout of **BPI Virtual Analyst**: a multi-tenant GenAI platform abstracting **30+ foundation models** via **FacadeDriver** with routing, retry, fallback, and graceful degradation, serving **128+ concurrent internal devs** across **4 partner engineering teams**. At Southwest, owned real-time pipelines on **Kafka streaming** for flight tracking and booking notification routing. **Python**, **distributed systems**, **LLM deployment**, **MLOps**, **production AI reliability**. MS CS (ML and AI focus) NYU.

Technical Skills

Languages: **Python** (primary, 7+ years), Java (Spring Boot 3.2 / Java 21), SQL (Spark SQL, Hive/Trino), TypeScript (React 18 / TS 5.4), Bash

Streaming & Batch: **Kafka** (idempotent partition-keyed pipelines, DLQ, backpressure), RabbitMQ, Airflow DAG (5-layer, async tool calling), Celery, Redis, PySpark, Databricks, Azure Data Factory

Generative AI & LLMs: **AWS Bedrock**, **OpenAI**, **Anthropic Claude**, Llama, **Hugging Face Transformers**, **RAG architectures**, **vector databases** (Pinecone, Weaviate, Chroma), **FacadeDriver** model-agnostic abstraction over 30+ LLMs with routing, retry, fallback, graceful degradation

Agentic AI: Agent loop design, tool/function calling, multi-step reasoning, agentic memory (semantic, episodic, procedural), multi-turn and multi-session orchestration

LLM Deployment & MLOps: **LLM deployment patterns** (routing, retry, fallback, graceful degradation, versioning), MLflow, **drift detection**, CI/CD for ML, **AWS SageMaker**, Docker, Kubernetes, monitoring, rollback workflows

LLM Evaluation: **23+ agent versions**, **1,690 versioned ground-truth samples**, dual-model A/B testing, flip-analysis, retrieval quality evaluation

Observability: **Loki eval gates**, Datadog, OpenTelemetry tracing, Grafana, drift detection, failure-mode analysis, MTTR reduction (45 to 12 min)

Cloud & Infra: **AWS** (SageMaker, ECS, Lambda, S3, DynamoDB, API Gateway, Kinesis), Docker, Kubernetes, OpenShift, Azure (Data Factory, Databricks), Terraform, GitHub Actions. **AWS Solutions Architect Professional** certified

Backend & APIs: **FastAPI/Flask**, Django, Spring Boot, REST APIs, SQLAlchemy, Temporal workflow orchestration

Experience

Airbnb

Sep 2024 - Present

Senior AI Infrastructure & Systems Engineer, GenAI Platform (Contractor)

San Francisco, CA (Remote)

- Owned end-to-end architecture and production rollout of **BPI Virtual Analyst**, a multi-model GenAI orchestration platform abstracting **30+ foundation models** (AWS Bedrock, OpenAI, Anthropic Claude, vLLM) behind **FacadeDriver** with routing, retry, fallback, and graceful degradation. Platform processes **10K rows per run** and **40MB uploads** with PII-safe inference via Microsoft Presidio redaction across 12 entity types, serving **128+ concurrent internal devs** across **4 partner engineering teams**
- Designed **RAG architectures** with semantic retrieval and **agentic AI systems** with multi-step reasoning, tool calling, and agentic memory (semantic, episodic, procedural) for multi-turn and multi-session workflows. 5-layer **Airflow DAG** with async tool calling and retry logic, targeting **80% runtime reduction** and **99.5%+ data consistency**
- Established **LLM evaluation standards** adopted across 4 partner teams: 23+ agent versions, 1,690 versioned ground-truth samples, dual-model A/B testing, and flip-analysis integrated with production systems. Built **observability stack** with Loki eval gates, Datadog, OpenTelemetry tracing, Grafana, and **drift detection**
- Owned **MLOps and LLM deployment pipeline**: model versioning, zero-downtime deployment, monitoring, and rollback. Scaled throughput **16x (600 to 10K rows per run)** and upload capacity **4x (10MB to 40MB)**. Root-caused IAT token-expiry auth failure and shipped permanent fix

Southwest Airlines

Jan 2023 - Jan 2024

Senior Software Engineer (Contractor), Backend and Application Platform

Dallas, TX (Remote)

- Owned architecture and production operation of real-time flight-tracking data pipelines on **Kafka streaming** and RabbitMQ for booking and notification routing across multiple consumer services, with idempotent partition-keyed consumers, dead-letter queues, and backpressure handling for zero-data-loss delivery. Cut on-call **MTTR from 45 to 12 minutes** (73% reduction)
- Built Python backend APIs (**Flask/Django**) with Java/Gradle services, raised unit-test coverage to **95%** via test-first methodology with pytest, CI/CD, and API integration. Elasticsearch and TensorFlow/PyTorch predictive models

Shell

Senior Software Engineer (Contractor), Backend and Data Science

Jun 2021 - Dec 2022

Houston, TX (Remote)

- Owned batch ETL architecture on **Databricks and Azure Data Factory**, automating **PySpark and Spark SQL** workflows with Hive/Trino. Deployed ML models on **AWS SageMaker with Docker** and built **SHAP/LIME explainability POC** for model transparency and regulatory monitoring dashboards
- Owned **NLP text-classification system** lifting accuracy from **86% to 94%** via wildcard-regex feature engineering, tokenization, semantic similarity, entity extraction, and transformer fundamentals. Published research at SPE ADIPEC 2022 (SPE-210986-MS). API publishing service: 17M pageviews/month, 94% cache efficiency

Eli Lilly and Company

Senior Software Engineer (Contractor), Dose Management Platform

Feb 2024 - Aug 2024

Philadelphia, PA (Remote)

- Owned architecture and production rollout of **FDA-regulated compliance platform** under HIPAA, 21 CFR Part 11, and GxP, sustaining **99.9% uptime for 6 months**. Designed 8-level RBAC access layer across **14 user personas** (US and EU) with audit trails, and **Temporal workflow orchestration** for dose approval (commercial 2-step, clinical 3-step parallel) with async retries. Spring Boot 3.2 / Java 21 + React 18 / TypeScript 5.4 + PostgreSQL 16 + OpenShift on AWS

Oracle

Software Engineer, ERP Analytics and Data Engineering

Aug 2017 - Jul 2019

Bangalore, India

- Owned ERP analytics backend architecture serving **13 business units** across P2P and O2C, architected ELK + AWS Kinesis fraud-detection pipelines on **Kafka** for real-time transaction routing. Automated PIP workflows saving 600+ monthly hours. Best Performer Q3 2018

Open-Source Contributions

LiteLLM (merged): fixed Vertex Gemini `count_tokens` returning 0 by converting messages to Gemini `contents` format before token counting. 93 lines + tests

LangChain: fixed Perplexity cost tracking by propagating search query count into usage metadata (costs underreported up to 10x)

LiveKit Agents: added `ReliabilityObserver` for post-call reliability scoring of `AgentSession`

Education

New York University, Tandon School of Engineering

Aug 2019 - May 2021

Master of Science in Computer Science with focus on Machine Learning and AI, GPA 3.69

New York, NY

- Capstone: real-time credit-card fraud detection and billing pipeline on **Kafka + Spark EMR + AWS Lambda + DynamoDB/RDS + S3 + QuickSight**. Code at github.com/sailikhithk/CreditCardFraudDetectionUsingKafka

Jawaharlal Nehru Technological University

May 2017

Bachelor of Technology in Electronics and Communication Engineering, GPA 4.0

Andhra Pradesh, India

Certifications, Publications and Recognition

Certifications: **AWS Solutions Architect Professional**, AWS Solutions Architect Associate, AWS Developer Associate, GCP Data Engineer, Azure Data Scientist Associate, Machine Learning Engineer Nanodegree, Oracle Java SE 8 Programmer

Publications: Cambridge Scholars Book Chapter (2025) - State Space Models in Healthcare (Ch. 1, ISBN 9781036463335), IEEE Xplore (ICoACT 2023) - Role of AI to address Cyberbullying (DOI 10.1109/ICoACT57570.2023.10118947), SPE ADIPEC 2022 - Full-Stack ML Development Framework for Energy Industry Applications (SPE-210986-MS)

Patent: Indian Patent Filed (2025) - Modular Deep Learning Architecture for Cross-Domain Transfer and Incremental Learning (App. No. 202541010770)

Recognition: IIBT-2026 Plenary Speaker and Scientific Judge (AI-Driven Precision Therapeutics, Alzheimer's QSAR models), Drug Discovery Workshop Resource Person (QSAR Modeling, Deep Learning for Drug-Target Interaction), Multiple Airbnb peer appreciations (2026), Level 3 Google Foobar Challenge Winner