

Soham Bapat

Senior MLOps & AI Engineer

sohambapat26@gmail.com · (404) 519-4826 · linkedin.com/in/soham-bapat · github.com/sohambapat · soham-bapat.com

Senior MLOps and AI engineer building production ML and LLM systems at a regulated bank — model scoring at scale, retrieval-augmented assistants on AWS Bedrock, and the pipelines, monitoring, and guardrails underneath them. 4+ years across banking, supply chain, and startups. U.S. Citizen.

EXPERIENCE

Huntington National Bank — MLOps Automation Senior Engineer — Full-time (Contract-to-Hire) Feb 2025 — Present

- Own scoring, monitoring, and auditing for **57 production ML models** driving the bank's marketing and campaign decisioning — an estimated **\$100M in annual revenue** — on **Terraform**-provisioned AWS infrastructure with drift monitoring, telemetry, and monthly delivery of both scores and audit reports via **Azure DevOps** CI/CD.
- Build and operate production **retrieval-augmented generation (RAG)** pipelines behind the bank's internal AI assistants — one serves hundreds of colleagues querying internal policy documentation in natural language, returning synthesized answers with clickable source citations.
- Implement these end to end: scheduled **Step Functions / EventBridge** ingestion chunks and embeds documents with **Cohere Embed v4** into **OpenSearch Serverless**; a **Strands agent** on **Lambda** retrieves context and invokes **Claude Haiku** on **Amazon Bedrock** behind content **guardrails**. React 19 and **FastAPI** across 5 purpose-scoped Lambdas with Keycloak SSO.
- Plan the project's development roadmap — model onboarding, release cadence, automation priorities — and coordinate schema and schedule changes with the downstream teams consuming the scores.
- Delivered a **7.5M+ record** pipeline across EventBridge, Step Functions, Glue, and SageMaker Pipelines, streamed to **Kafka** with Splunk log monitoring; ETL in Glue with Pandas, PySpark, and Boto3 over Athena.
- Built internal agentic ML-ops tooling on **Bedrock and MCP servers** (n8n, Claude Code); solo speaker at the company Data Expo and SME at the company hackathon.

Vosyn.ai — Cloud / AI Platform Engineer — Contract, Deputy Team Lead Jun 2024 — Feb 2025

- Built the **audio translation pipeline** behind the product: incoming video demuxed into audio and video tracks, audio transcribed with **Whisper** speech-to-text, the transcript **machine-translated** into the target language, then re-voiced with **text-to-speech** and remuxed.
- Lead engineer on the **AWS→GCP migration** (Terraform, **GKE**); re-architected task-based Docker execution into a service-based ECS architecture.
- Built end-to-end CI/CD for the serverless microservices carrying those stages with **GitHub Actions** on Lambda and Cloud Run; migrated a SageMaker MLOps pipeline to **Vertex AI**; led a team of 5.

TELUS International — Data Analyst / GenAI Evaluation Dec 2022 — May 2023

- Rated map-service query results to build **human-aligned datasets for fine-tuning and evaluating generative AI models**.

Manhattan Associates — Software Analyst — Full-time Jun 2021 — Jul 2022

- Built and deployed client applications through Jenkins and Nexus CI/CD; wrote service-chaining scripts for Java / Spring Boot microservices and supported Supply Chain Management and Point of Sale platforms.

SELECTED PROJECT

Autonomous Self-Improving LLM Trading Agents May 2025 — Present

- Designed and operate two autonomous LLM agents running **unattended since May 2025** on self-hosted **Raspberry Pi** hardware over a private **Tailscale** network — tiered Claude routing (Haiku/Sonnet/Opus) by task complexity, an auto-reverting experiment loop validated against an **8-year backtest**, and a multi-level safety circuit breaker.
- Diagnosed a seven-loss run as a missing take-profit path rather than strategy decay; traced a silently defaulting volatility input and shipped five verified fixes across both machines in a single day.

TECHNICAL SKILLS

AI / LLM	RAG pipelines · Amazon Bedrock · OpenSearch Serverless (vector search) · Cohere Embed · Anthropic Claude API · Strands agents · MCP servers · Bedrock Guardrails · Prompt engineering · LLM evaluation · Whisper (STT) · Text-to-speech · n8n
MLOps / ML	SageMaker AI & Pipelines · Vertex AI · Model scoring · Drift monitoring · CI/CD for models · Backtesting
Cloud / DevOps	AWS (Lambda, Glue, Step Functions, EventBridge, ECS/ECR, EC2, S3, IAM, Athena) · GCP (GKE, Cloud Run) · Terraform · Docker · Kubernetes · Azure DevOps · GitHub Actions · Jenkins · Ansible · Linux · Tailscale
Data	Python (Pandas, PySpark, Boto3) · SQL · Java · Apache Spark · Kafka · BigQuery · Power BI · Splunk
Application	FastAPI · React 19 · Vite · TypeScript · Keycloak SSO · Shakudo · REST / OpenAPI

CERTIFICATIONS

Google Cloud Fundamentals for AWS Professionals, Google Cloud · 2024
AWS Certified Cloud Practitioner, Amazon Web Services · 2023 (Valid through Jan 2029)

EDUCATION

M.S. Computer Science, North Carolina State University · Aug 2022 — May 2024
B.S. Computer Science, Kennesaw State University · May 2021 · Honors