Ishank Sharma

Forward Deployed Engineer | **HIPAA/PHI Deployments**

Los Angeles, CA
+1 562-341-2003 · ishankdev@gmail.com
linkedin.com/in/ishank47 · ishank-dev.github.io

Forward Deployed Engineer who owns AI deployments end to end inside the customer's environment: discovery, scope, POC, pilot, and handoff. Sole engineer on a HIPAA/PHI claims-automation deployment for a major California healthcare payer, replacing an offshore data-entry vendor with a Claude on Amazon Bedrock pipeline. 4+ years building production backend, data, and applied LLM systems. At home in the work that decides whether a deployment lands: undocumented formats, legacy SQL Server, security and BAA review, SME working sessions, and a handoff the customer's team can own.

TECHNICAL SKILLS

AI systems	Claude (Opus, Sonnet) on Amazon Bedrock (Converse and ConverseStream, prompt caching, cross-model failover), OpenAI GPT-4o, Gemini, LLaMA, LangChain, LangGraph, LlamaIndex, MCP servers and clients, agents, tool calling, structured outputs, Claude Code
Document AI and RAG	Vision and multimodal document extraction, OCR, RAG, hybrid retrieval (BM25 + embeddings, Cohere Embed v4), semantic embeddings, pgvector, FAISS, Pinecone, Text-to-SQL
LLM evals and safety	Golden-set evals, regression gating in CI, model A/B shadow runs, prompt versioning, extraction traces and replay, per-call token cost and latency tracking, guardrails, human-in-the-loop escalation, PHI-safe logging
ML and data science	PyTorch, Hugging Face Transformers, scikit-learn, pandas, NumPy, SciPy, causal inference (Difference-in-Differences), recommendation systems, synthetic data, MLflow, Jupyter
Python	FastAPI, Flask, Pydantic, SQLAlchemy 2.0, Alembic, asyncio, boto3 and aiobotocore, psycopg 3, Pillow, pypdf, openpyxl, pytest, Playwright, ruff, uv
Languages and web	Python, SQL (T-SQL, Snowflake), JavaScript, Java, C++; Node.js, Spring Boot, REST, Vue
Data	PostgreSQL (Aurora), SQL Server, Snowflake, Databricks, PySpark, Airflow, MongoDB, Redis
Cloud and delivery	AWS (Bedrock, ECS Fargate, Lambda, S3, SQS, Step Functions, Transfer Family, Aurora, RDS IAM, CloudWatch, Cognito, SSM, EC2), AWS Copilot, GCP and Vertex AI, Docker, Kubernetes, Terraform, GitHub Actions, Okta SSO, PagerDuty
Deployment	Customer discovery, scope docs and success criteria, POC sprints, pilots and parallel runs, HIPAA/BAA and security review, runbooks and knowledge transfer

EXPERIENCE

Forward Deployed Engineer, Prentis AI

Jun 2026 to present

Sole engineer on the deployment, discovery to live parallel run in 6 weeks:

California healthcare payer *Claims automation (HIPAA/PHI), retiring an offshore data-entry vendor*

- Ran discovery with claims SMEs and wrote the scope the payer signed: read scanned CMS-1500 claims, emit fixed-width EZ-CAP EPF-653 load files, graded against the incumbent vendor's output. Locked business rules over 15+ working sessions; reverse-engineered the undocumented EPF-653 spec and the payer's legacy VB scrubber.
- Built the image-to-EPF pipeline on Claude (Amazon Bedrock): vision extraction with automated feedback rounds, codified payer business rules, and a member/provider resolution ladder over EZ-CAP SQL Server using set-based lookups across ~1M-row views.
- Stood up a golden-set eval harness with regression runs gating every prompt and rule change: 96% field-level accuracy and ~90% member auto-match on 1,700-claim batches, within 2 points of the human vendor, proven in a live parallel run on production traffic.
- Engineered fault-tolerant ingestion and delivery: SFTP (AWS Transfer Family) into S3 with a SHA-256 dedup ledger, quarantine, a crash-resumable per-claim state machine, and immutable deliveries via S3 conditional writes, reconciled so arrived = terminal states.
- Shipped the FastAPI + SPA reviewer dashboard claims staff rely on: per-field provenance from scan to database to EPF, byte-level diffs against the vendor's file, and a Completed/Escalated contract giving every held claim one reason and a next step.
- Held the PHI boundary end to end (Bedrock under AWS BAA; PHI never in git, logs, or on a public port) and tracked per-claim token cost and latency. Handed off runbooks, monitoring, and knowledge transfer so the payer's team runs it without me.

Forward Deployed Engineer, CommerceIQ

Feb 2026 to May 2026

Meta, Hill's Pet Nutrition, SharkNinja *CommerceIQ AI-agent playbook, Day 1 discovery to Day 60 production*

- Owned three brand deployments through the full engagement model: discovery kickoff, workflow mapping (people, processes, tools, agent scope), prototype reviews with brand teams, UAT, and production sign-off inside 60 days.
- Designed and shipped the multi-tenant analytics portal underneath them (Flask + Vue SPA, Docker, magic-link auth), serving bespoke executive dashboards for all three brands from a single deployment.
- For Meta, built the Sales & Traffic dashboard (WoW/YoY/T4W against Best Buy and Target) on a Difference-in-Differences causal framework (scipy) that put retail-media lift at ATE ≈ +5 pp, giving the exec sponsor a defensible number rather than a correlation.
- Productionized "Ally" GPT-4o executive summaries: prompt orchestration over sales, traffic, and causal statistics producing structured, retailer-aware commentary with CI-aware null-safety guardrails.

Full Stack Engineer, Stealth Startup

May 2025 to Jul 2025

- Benchmarked Text-to-SQL pipelines across LLaMA, OpenAI, and Gemini, tuning latency, quality, and cost with timeouts, fallbacks, and caching.
- Built personalization workflows on semantic embeddings and audience signals for targeting, segmentation, and campaign scheduling; raised query QPS 85% (PostgreSQL, Docker).

Software Engineer (SDE 1 to SDE 2), CommerceIQ

Sep 2021 to Aug 2024

- Built an AWS media-budget recommendation pipeline (Step Functions, SQS, Lambda, 40 SQL modules) predicting ad spend, saving clients $200K/year and cutting processing from 1 day to 3 minutes.
- Solved cold start for no-history clients with synthetic data fusing sales and ad-performance signals: 90%-accurate recommendations.
- Migrated budget prediction from Snowflake to Databricks (infra cost –52%, latency –80%) via strategic indexing and broadcast joins; engineered the Amazon Advertising campaign platform and Criteo strategy automation, cutting manual campaign operations 60%.

EDUCATION & CERTIFICATION

M.S., Computer Science, California State University, Long Beach, May 2026. GPA 3.9/4.0. Graduate Teaching Assistant (AI), Instructional Assistant (Ethical & Explainable AI), Research Assistant (Neuromarketing & AI).

B.E., Information Science & Engineering, Ramaiah Institute of Technology, Aug 2021. GPA 8.52/10.

Certifications: Forward Deployed Engineer for AI Systems, Claude Code in Action, Mastering MCP, AI Engineer Agentic Track, GCP Fundamentals, Responsible AI Data Mgmt.

Honors: Tau Beta Pi (CA Theta). Runners-up, GE Healthcare Hackathon 2020 and India Innovation Series 2019 (Microsoft). Top 4, CodeFury 2.0.