Aditya Mittal

New York, NY, USA | +1-347-206-6415 | theadityamittal@gmail.com | Website | LinkedIn | GitHub

SUMMARY

I build production LLM systems end to end; training pipelines, eval harnesses, and inference architectures, not just wrappers around foundation model APIs. Currently at NovumAI. NYU MS Computer Science, 2025.

EDUCATION

Master of Science GPA 3.90 — **Sep 2023 - May 2025**
New York University — Computer Science

Bachelor of Technology GPA 8.86 — **Jul 2018 - May 2022**
Manipal University Jaipur — Computer Science

EXPERIENCE

AI/ML Engineer — **Nov 2025 - Present**
NovumAI

- Built a 6-pass SageMaker pipeline converting 13K BPO transcripts into 30K supervised fine-tuning (SFT) and direct preference optimization (DPO) training pairs via diarization, stage labeling, and NEPQ rewriting.
- Designed an 8-dimension evaluation harness with LLM-as-Judge scoring suggestions on signal quality and delivery quality, gating releases on offline eval and champion/challenger testing on a live user cohort.
- Rewrote the monolithic coaching prompt as a dynamic stage-aware template and SFT + DPO fine-tuned Llama 3.1 8B via Bedrock CMI; improved composite eval score by 42% over the Scout 17B baseline.
- Split per-message inference into two paths off the WebSocket: a suggestion Lambda on the hot path and a reasoning Lambda that owns call state (stage, sentiment) in Redis; p50 latency 1.8s to 1.2s.

AI Engineer — **Jul 2025 - Oct 2025**
Changing The Present

- Deployed a Slack onboarding agent that personalizes a plan for new interns, assigns them to channels, and schedules meetings through tool calls. Reduced intern ramp time from about 5 days to 2 across 50 hires.
- Engineered a dual-LLM ReAct loop with four tools (search, send, schedule, assign) and a hierarchical Pinecone RAG that scores context on four factors before grounding, achieving 85% answer accuracy in manual review.
- Shipped the bidirectional Asana sync behind the HR dashboard so applicant and intern state stays consistent across both platforms for the 100-200 interns using them concurrently.

Applied ML Engineer — **Nov 2022 - Mar 2023**
Engagebud

- Implemented a Thompson Sampling bandit routing over K-means visitor clusters to relevant campaigns in real time. Client (boAt) hit 80% engagement in case study.
- Launched the widget runtime, configurator, and analytics dashboard in React and TypeScript. Integration time for new clients dropped from 30 minutes to under 5.

ML Engineer Intern — **Jan 2022 - Jun 2022**
Juniper Networks

- Developed PySpark ETL and feature extraction feeding Mist AI's LSTM predictive failure pipeline. Schema validation and quality checks improved early-failure recall by about 15%.
- Instrumented PSI-based drift monitoring with automated retrain triggers on threshold breach. False-positive alerts in Mist AI's production fell roughly 20% across firmware release cycles.

SKILLS

Programming & Scripting Languages Python, TypeScript, JavaScript, SQL, Bash
LLM Engineering SFT, DPO, QLoRA, LLM-as-Judge, RAG, ReAct Agents, MCP, Hugging Face, PEFT, PyTorch, Prompt Engineering
AI Infrastructure Amazon Bedrock, SageMaker, Pinecone, Gemini, LiteLLM
Backend & APIs FastAPI, Node.js, WebSockets, OAuth 2.0, Supabase, FHIR R4
Cloud & Infrastructure AWS Lambda, API Gateway, SQS FIFO, KMS, S3, Railway, CloudFormation
Databases PostgreSQL, DynamoDB, Redis, RDS
DevOps & Observability Docker, GitHub Actions, GitLab CI/CD, CloudWatch, New Relic

PROJECTS

Adverse Event Investigator Devpost

- Deployed a standalone A2A agent (Google ADK + LiteLLM, Railway) with native FHIR R4 connectivity to any HL7-compliant server; 14-tool MCP server handles investigate / reason / draft across patient data.
- Implemented Naranjo causality and FDA seriousness scoring as deterministic Python — zero LLM calls in clinical math; LLM handles investigation only, producing a MedWatch Form 3500 on request.

Claude Professor GitHub

- Built a Claude Code MCP plugin enforcing concept teaching before design progress via a JIT state machine — the wrong architectural path is impossible, not just discouraged; combats vibe coding.
- Integrated FSRS-5 spaced repetition as deterministic math decoupled from LLM inference; two-stage Haiku concept matcher uses thin retrieval then full metadata rerank across 407 concepts in 18 domains.