

RAJEEV JOSHI

rj1234@nyu.edu · [linkedin.com/in/rj1234](https://www.linkedin.com/in/rj1234) · github.com/rj-1234 · +1-347-216-4621

SUMMARY

Principal Machine Learning Engineer with 6+ years at Vanguard building production AI systems across LLM infrastructure, agentic AI, model inference engineering, speech analytics, and responsible AI governance. Shipped enterprise-scale platforms processing 300K+ calls daily and 80M+ API invocations, a unified model deployment framework scaling to 3M+ requests/day, a modular Responsible AI framework with real-time guardrails, and an LLM consumption SDK adopted by 6+ product teams. Published researcher (IEEE AAIML 2026) with end-to-end ownership from data labelling and model development through inference serving and production reliability.

LANGUAGES & TECHNOLOGIES

LLM & Agentic AI	NeMo Guardrails, LangGraph, Mem0, RAG, LoRA/PeFT fine-tuning, agent orchestration, LLM evaluation
ML / NLP	ASR, NER, sentiment/emotion analysis, Transformers, Keras, TensorFlow, scikit-learn
Inference & Infrastructure	Kubernetes (EKS), Docker, AWS, Serverless architectures
Frontend	React, JavaScript, HTML, CSS
Vector & Data	Pinecone, DynamoDB, PostgreSQL, Oracle
Languages	Python, JavaScript, SQL, Shell Scripting, Node.js
Dev & Tooling	Git, BitBucket, GitHub Actions & Workflows, OpenCV, Docker Compose

WORK EXPERIENCE

The Vanguard Group — Malvern, PA

Principal Machine Learning Engineer

December 2025 – Present

- Architected an enterprise Responsible AI framework with a real-time NeMo Guardrails engine inside a modular, use-case-agnostic FastAPI orchestrator — configurable rails detect and mitigate policy violations per-request across any LLM-powered product.
- Designed an AI-ready data collection pipeline for policy drift detection with automated self-healing via LoRA/PeFT fine-tuning of the base LLM when evaluation scores drop below threshold — closing the model quality loop without manual intervention.
- Built the guardrails layer to be framework-agnostic and migration-ready (mem0 + LangGraph extensible), with application teams retaining ownership of memory and agent design — enabling broad adoption without platform lock-in.
- Drove Responsible AI governance strategy: red-teaming protocols, policy evaluation datasets, and automated scoring pipelines benchmarking safety across model families.

Senior Machine Learning Engineer

September 2024 – December 2025

- Co-developed a unified model deployment framework on EKS serving LLMs, ASR, and other ML models behind a single inference gateway — scaling to 3M+ requests/day while standardizing container orchestration and autoscaling policies across all model types.
- Contributed custom model architecture configurations (30B, 8B, and SLM variants) for the vLLM serving framework, leveraging PagedAttention to achieve ~2x throughput over naive KV-cache allocation — enabling cost-effective serving of proprietary models at enterprise scale.
- Co-designed Vanguard’s enterprise speech analytics platform processing 300K+ calls/day (~150K minutes of audio) and 80M+ API invocations across retail, FAS, and IIG — with PII redaction via a custom-trained NER model, speaker diarization on mono audio, call denoising, BERT-based call intelligence classifiers, and a segment-level labelling platform to accelerate training data quality.
- Adapted LibreChat and a custom RAG API for Vanguard’s environment (2024): migrated MongoDB → DynamoDB, built a Pinecone + FastAPI retrieval backend, and containerized on EKS — giving teams a self-hosted LLM experimentation platform ahead of enterprise-grade solutions; added RLHF feedback loops and ran Behavioural Lab A/B experiments to align model responses with marketing campaign behaviours.
- Architected a unified LLM SDK (Python + Node.js) with token-based auth adopted by 6+ product teams, a shared prompt library for cross-team reuse, and mentored 1–2 engineers annually — one mentee achieving promotion.

Machine Learning Engineer

August 2019 – September 2021

- Built a multiprocessing inference pipeline on a vendor cloud platform running batch inference across 85M+ rows, cutting execution time by 70% using Python/shell scripting with integrated testing and alerting.
- Designed an event-driven database architecture (persistent + volatile memory tiers) for the backend of an in-house audio transcription engine; deployed a fund similarity API at sub-500ms latency with periodic model refresh.
- Built an in-house NLP labelling platform and an auto-scaling EDA acceleration package that reduces analysis time proportionally to dataset size — shortening the model development cycle across teams.

- Productionized a Next Best Action model integrated into the Key Insights internal advisor application, surfacing personalized recommendations to client-facing advisors at scale.

BotFactory Inc — New York City, NY
Software Engineering Intern (R&D)

June 2018 – August 2018

- Developed computer vision algorithms using OpenCV to improve the positional accuracy of the Pick-and-Place Head in an automated PCB assembly system.
- Designed UX/UI and built a critical product webpage for a new feature shipping in the next hardware version.
- Migrated the hardware abstraction layer instruction set from the previous model to the latest software architecture.

PUBLICATIONS

Bansal, A., Joshi, R., et al. — “Cost-Efficient, Model-Agnostic, Low-Latency LLM Deployment” — *IEEE AAIML 2026* | February 2026 · ieeexplore.ieee.org

AWARDS & RECOGNITION

- **IT Peer Recognition Award — Vanguard Information Technology Division (2025)** · Quarterly award granted by a panel of past recipients who review all nominations across the IT division and select a shortlist of winners — recognizing crew members whose technical contributions and collaborative impact stand out division-wide.
- **NYU Academic Achievement Award** · Recognized for academic excellence in the MS Computer Science program, New York University (2018).
- **MLH HackNYU 2018 — Third Place** · *Feed a Homeless* platform, awarded among competing teams at Major League Hacking’s HackNYU event.

EDUCATION**New York University, Tandon School of Engineering**

May 2019

MS, Computer Science · GPA: 3.77/4.0 · Academic Achievement Award

Maharaja Agrasen Institute of Technology, Delhi

May 2017

BTech, Computer Science & Engineering

COMMUNITY & MENTORSHIP

- **Hackathon Mentor — Women in Data Science, Vanguard** · Guided a competing team through problem framing, technical architecture, and AI implementation for a data-driven solution.
- **Hackathon Mentor — FAST Hackathon, Vanguard FAS (December 2023)** · Mentored participants on LLM integration patterns and production deployment strategies.
- **MLH HackNYU 2018 — Third Place** · Built *Feed a Homeless* — a real-time platform connecting surplus food donors with nearby homeless individuals via Google Cloud SQL, Maps API, and a PHP/JS frontend.
- **MLH HackNY 2018** · Built *Twitterazzi* — automated tool using the Clarifai vision API to extract image keywords and publish hashtag-rich tweets from photos with a single click.

ACADEMIC PROJECTS

- **CodeJudger** — Real-time in-class coding assessment platform for professors with CI/CD via GitHub Actions, Docker Compose, and CodeIgniter PHP.
- **Protein Residue Distance Matrix Prediction** — Keras deep neural network predicting inter-residue distance matrices from protein primary and secondary structure sequences.
- **Movie Rating Prediction** — Twitter sentiment analysis combined with IMDB features to rate and predict movie ratings; built on Google Cloud with Python.