

NANDESHWAR GUPTA

Machine Learning Engineering - Senior Manager | Generative AI & ML | 8+ Years

+91-7011825826 • guptanandeshwar553@gmail.com • [LinkedIn](#) • [Portfolio](#)

PROFESSIONAL EXPERIENCE

Zupee Gurugram, India

India's leading skill-based casual gaming platform (2M+ DAU) & NeoAstro Chat (1M+ Play Store downloads)

Senior Manager, Machine Learning Engineering Apr 2026 – Present

- **Lead a 5–7 engineer ML team**, owning roadmap, hiring, delivery, and stakeholder alignment across Generative AI, recommendations, and platform engineering.

Generative AI & Data Science

- Built an **AI companion agentic chatbot (serving 100K DAU)** that holds natural, hyper-personalized, and deeply emotionally aware conversations with users at scale, remembering past context and guardrailed.
- Architected an **8-node LangGraph conversational AI Agent** with a 6-way parallel fan-out (mood classification, memory retrieval, guardrails, topic-exhaustion) and a bounded self-correction loop (repetition-detection + post-response critique), reducing per-turn inference latency by ~35% vs. the sequential baseline. Leading to faster, more human-feeling replies lifted user session length by aggregated 14% and D7 retention by 60 BPS.
- **Fine-tuned Qwen 3.6 35B-a3b** by curating dataset from production chat traffic to match the GPT-4.1 response generator, including prompt re-tuning for the open model and an automated validation harness to guard against quality regressions **cutting inference cost ~65% while keeping TTFT within 1s p95**.
- Built an **LLM conversational memory system** (semantic/episodic/factual/procedural) with embedding dedup (merge/contradict/insert classification) and read-time importance decay over MongoDB + Qdrant, injecting dynamic user profiles into prompts improving response personalization/relevance by 8% in human eval. Personalized continuity drove **4% higher daily active conversations** and **6% higher conversions**.
- Designed a **multi-layer LLM guardrails pipeline** (24-category multi-class classifier + LLM-as-a-judge critic with RAG-augmented few-shot), achieving **70% precision / 99% recall**. Reduced trust-and-safety incidents by 95%, protecting brand and avoiding manual moderation cost. This also led to a **60% decrease in guardrail cost**.

ML/AI Engineering

- Built a **Kafka-driven AI companion chat pipeline** in Python/FastAPI, implementing manual offset commits, partition pause/resume backoff, and a two-level debounce (inter-pod MongoDB + intra-pod in-memory scheduler) to batch bursty user messages with semaphore-bounded concurrency handling up to **500 concurrent conversations per pod**.
- Designed an **atomic ChatNode state machine** using MongoDB findOneAndUpdate operations to guarantee single-processor ownership under Kafka rebalances, plus a stuck/orphaned-node sweeper with dead-lettering after max retries to recover from pod crashes without message loss.

Manager, Machine Learning Engineering Apr 2025 – Mar 2026

- **Launched a highly scalable Agentic LLM chatbot** for Astrology with RAG, SLMs, async pollers, and queue-based autoscaling, contributing to **100% MoM user growth**, a **4-star Play Store rating**, and **500K+ downloads**.
- Architected an **RL-based recommendation engine** evolving from Multi-Armed Bandits to **Deep Q-Networks**, driving **+3.5% of organization revenue** and **+10 BPS retention** on a 2M+ DAU platform.
- Designed **user embeddings** from behavioral and demographic clustering, materially boosting recommendation accuracy and personalization engagement. Embeddings became integral across other downstream ML projects.
- Built a customer facing **LLM-driven agentic chatbot with** orchestrator agents, tool-calling, and context/nudge flows, improving **CSAT by 8%**.
- Built **SQL Smith**, a graph-based Natural Language-to-SQL agent, enabling faster self-service analytics for product and growth teams. Designed a **LLM framework** with observability and prompt management used across all GenAI use cases.

Lead Machine Learning Engineer Jan 2024 – Mar 2025

- Built and deployed **real-time tournament recommendations** on **Apache Flink + Aerospike**, cutting inference from **hours to seconds** and driving **+4.15% revenue** in partnership with Product and CRM.
- Improved recommendation **accuracy@top4 by 6%** and redesigned the architecture to resolve **cold-start** challenges, accelerating time-to-business-impact. **Parallelized model retraining** with Airflow, reducing training cycles from hours to minutes and enabling daily refreshes.
- Shipped a **multi-armed bandit pricing optimization** system, delivering **+8% revenue uplift**.

- Drove **Generative AI initiatives**: avatar generation, gender/NSFW detection, and **celebrity voice cloning** for IVR campaigns (+6% call completion, +4% reactivation).

Senior Machine Learning Engineer *Sept 2022 – Dec 2023*

- Spearheaded the **end-to-end ML lifecycle framework** and real-time prediction stack (training, tracking, deployment, monitoring) from scratch.
- Delivered the XGBoost based **tournament recommender engine**, boosting **revenue by 4%** and **accuracy@4 by 10%**.
- Led **Generative AI PoCs**: LLM-powered quiz generation, Indian-dialect TTS for game commentary, marketing video generation, and SageMaker GPU training optimization. Led cross-functional CRM/marketing campaigns, lowering customer acquisition cost and lifting reactivation rates.

V-Mart Retail *Gurugram, India*

Senior Manager, Data Science *Feb 2021 – Sept 2022*

- Built a **recommendation engine (Matrix Factorization)**, lifting platform revenue by **11%**.
- Developed **churn prediction (LightGBM)** and **customer segmentation** models, improving retention and personalization.
- Created **ARIMA forecasting** models for supply chain, materially improving demand planning during COVID-19 volatility.
- Designed scalable **ML pipelines on AWS** using Airflow and MLflow for reproducible training and deployment.

Vasitum *Gurugram, India*

Manager, Data Science *Sept 2020 – Feb 2021*

- Built recommendation engines and an **NLP chatbot (RASA, NER, intent classification)** to lift engagement and conversion.
- Optimized job search with **Elasticsearch** and shipped NLP/DNN models for spam-job classification.
- Delivered end-to-end ML pipelines including database modeling and migration to **serverless AWS (Lambda, Fargate)**.

Quick Company *Gurugram, India*

Data Scientist *May 2018 – Aug 2020*

- Built large-scale **NLP & CV models**: NER (spaCy + heuristics), legal text classification (Word2Vec), captcha recognition (CNN, OpenCV), and image classification (PHOG).
- Developed automated, **event-driven scraping** frameworks and used **Neo4j** for graph-based data linking across legal entities.

EDUCATION

B.Tech, Computer Science Engineering *2014 – 2018*

Manav Rachna International University, Faridabad

SELECTED HIGHLIGHTS

- Cumulative **revenue impact > 25%** across recommendation, pricing, and personalization systems (>\$50 million).
- Speaker / mentor on applied Generative AI and Reinforcement Learning in production systems.

TECHNICAL SKILLS

Generative AI & LLM: RAG, Agentic Frameworks & Agent Harness, LangChain, LangGraph, LlamaIndex, Prompt Engineering, Fine-tuning (LoRA/PEFT), SLMs, Vector DBs (FAISS, Pinecone, Elasticsearch), Tool-calling, Guardrails, LLM Observability

Machine Learning: Recommendation Systems, Reinforcement Learning (DQN, MAB, RL-Lib), Deep Learning, NLP (NER, Intent, Classification), Computer Vision (CNN, OpenCV), Forecasting (ARIMA), Matrix Factorization, LightGBM, Clustering & Embeddings

ML Engineering / MLOps: MLflow, Kubeflow, Airflow, SageMaker, Feature Stores, Model Registry, A/B Testing, CI/CD for ML, Real-time Inference (Flink, Aerospike), Monitoring & Drift Detection

Cloud & Infra: AWS (SageMaker, Lambda, Fargate, EKS, S3, Athena), Docker, Kubernetes, Jenkins, Bitbucket Pipelines

Languages & Frameworks: Python, SQL, Bash, JavaScript, PyTorch, TensorFlow, Scikit-learn, Hugging Face, FastAPI, Flask

Leadership: Team Management (5–7 engineers), Hiring, Roadmapping, Cross-functional Partnership (Product, CRM, Marketing), Mentorship