Raghav Dixit

Jersey City, NJ | +1 732-558-3682 | dixitraghav99@gmail.com | LinkedIn | GitHub

AI/Machine Learning Engineer

Software engineer with 4+ years of experience building production LLM/GenAI, deep learning, and recommendation systems at scale (serving up to 250M+ users). Skilled in machine learning, NLP, generative models, data engineering and open-source development.

Education

Rutgers University – New Brunswick **2023 - 2025**
MS, Data Science & Statistics
- **GPA:** 3.85/4.0

Manipal Institute of Technology **2017 - 2021**
BTech, Computer Science & Communication Engineering
- **GPA:** 8.51/10

Skills

- **AI/ML**: LLMs/SLMs, Transformers, Graph ML (Node2Vec, GNNs), RAG/GraphRAG, Knowledge Graphs, Reinforcement Learning (GRPO, PPO, MDP), GANs, Fine Tuning (SFT, RLHF, QIoRA, LORA), Classical ML (Trees/XGBoost, Regression algos), DNNs/CNNs, quantization (4bit/8bit), Optimizations, Data Generation
- **Data Systems**: PySpark/Spark, SQL/NoSQL, Vector DBs (LanceDB, Qdrant, ElasticSearch), Kafka, Redis
- **Frameworks**: PyTorch/TensorFlow, HuggingFace, LangChain/LangGraph, LlamaIndex, Node.js
- **Infra & DevOps**: AWS, GCP, Docker, Kubernetes, Terraform, ElasticSearch, CI/CD, MLflow • **Programming**: Python, C++, JavaScript/TypeScript, SDKs

Work Experience

Glance (InMobi) **May 2021 - Jul 2023**
Machine Learning Engineer (full-time)
Contributed to Glance's ML ecosystem powering e-commerce apps, real-money gaming apps, and short-video feeds for 250 M+ Android users. InMobi(Glance) is India's first unicorn backed by Google, MithrilCapital, SoftBank and Reliance.
- Built Spark ETL pipelines, feature-engineering workflows, a unified feature store, and DL/ML models, plus the deployment, serving, and monitoring stack on GCP Vertex AI & Kubernetes, improving data processing efficiency and model deployment speed
- Contributed to in-house Deep Learning Recommendation System using PyTorch by integrating DropoutNet layers for cold-start, achieving 3x day-0 retention, and optimizing CUDA operations for fast response times. Used synthetic data generation techniques to balance data distribution
- Built Node2Vec graph embeddings with PyTorch, processing billion-edge user-item graphs; optimized random walk sampling and sparse matrix operations for latency-aware similarity serving, driving 15% engagement lift for the app.
- Designed a TF based live-stream ranker with contrastive GNN achieving 25% watch time increase and an LTV classifier with custom embedding logic boosting ad conversions 30%.
- Conducted A/B experiments and causal studies to validate every ML launch, ensuring safe, reliable rollouts to the full 250 M-user audience.

LanceDB(YC '22) **Jan 2024 - Jul 2024**
Machine Learning Engineer (CPT)
Led OSS integrations with LangChain, LlamaIndex, and 10+ RAG/LLM frameworks drove 37% increase in downloads and 2x developer engagement through technical demos on advanced Graph/LLM based information retrieval techniques.
- Built custom PyTorch data loaders optimized for LanceDB vector operations, accelerating data loading by 3x; leveraged this for multi-GPU FSDP fine-tuning of LLMs/SLMs with 4-bit quantization, reducing memory usage by 40%
- Optimized embedding model inference through PyTorch JIT compilation and ONNX conversion pipelines, reducing latency from 200ms to 50ms; shipped model registry SDK enabling seamless HuggingFace/Cohere integration.

Entelligence.ai **Jul 2024 - Apr 2025**
Founding AI Engineer (CPT)
Engineered AI-powered productivity platform featuring automated documentation, code analysis, and personalized agents, saved 10+ hours per person weekly and reduced 2X documentation effort for customers.

- Built PyTorch based QLoRa finetuning pipeline to improve (Qwen2.5) LLM-based vectorDB query filter generation accuracy, increased customer task accuracy by 35% while reducing fine-tuning compute costs by 60%.
- Built multi-agent code review system using LangGraph, achieved 41% critical bug detection (vs.13% baseline) and 58% faster review cycles.
- Optimized production infrastructure with scalable microservices and vectorDB indexing reducing latency from 10min to <2min for 1GB+ repositories.
- Architected latency aware knowledge graph-based retrieval for customer agent interactions, enabling contextualized recommendations and supporting new enterprise onboarding pilots; this contributed to raising 5M$ seed (Mayfield).

Air Labs Inc. **Apr 2025 - Jul 2025**

Senior Applied AI Engineer (contract)

Contributed to developing Air's multimodal hybrid search (text/image/audio/video) to support sub-300ms p99 latency at 100K+ DAU scale (TypeScript/Node.js/ElasticSearch/Bedrock).

- Built a modular model runtime service and eval framework (nDCG@10, MRR@10, synthetic data via SORA/GPT) to optimize/compare embedding models from HuggingFace, Amazon, Meta.

Glance (InMobi) **Jan 2021 - May 2021**

Data Science Intern

Built data visualization dashboards for company performance metrics presented to key stakeholders (Google, Microsoft, Reliance)

- Developed video genre detection model using feature extraction—achieved 14% accuracy improvement over baseline and earned early FTE conversion in 3.5 months

XaiPient Inc **May 2020 - Jul 2020**

ML Research Intern

Built interactive ML dashboard using Streamlit and deployed on cloud infrastructure

- Researched causal inference applications in explainable AI, authored a technical report on CAUSEV2 model methodology, and contributed to enhancing AI model transparency

Projects

Langroid (multi-agent LLM framework)

- Contributed to the OSS project by integrating new vectorDBs, LLM backends, and resolving issues.

Multimodal Video RAG

- Created LlamaIndex+LanceDB solution enabling time-aligned QA over videos; recognized by LlamaIndex CEO, blog published here.

GTE-MLX RAG CLI

- Built a PDF QA tool with retrieval caching and injection patterns by creating an MLX compatible version of the general text embeddings model

Glance Internal thesis – Conducted research on video genre detection

Glance

- Conducted research on video genre detection: extracted key-frame segments and spectrogram features, fine-tuned 3D-ResNeXt101 with custom dropout; boosted accuracy 14% over baseline and secured early FTE conversion.
- Due to PII data this was an internal report and was not allowed to be published.