

Raghav Dixit

website: <https://raghav-dixit.com/>

LinkedIn: www.linkedin.com/in/raghav-dixit

GitHub: <https://github.com/raghavdixit99>

Email : dixitraghav99@gmail.com

Mobile : +1 732-558-3682

New Brunswick, NJ

ABOUT ME

I am a Full-Stack ML Engineer with 3+ years of experience deploying large-scale AI solutions. My expertise spans machine learning and deep learning—including NLP, statistical modeling, video processing, reinforcement learning, generative AI, exploratory data analysis, and MLOps—covering end-to-end workflows.

EDUCATION

- **Rutgers University - New Brunswick** New Brunswick, NJ, USA
Master of Science in Data Science
CGPA: 3.92/4.0
Sep 2023 - May 2025
- **Manipal Institute of Technology** Manipal, KA, India
Bachelor of Technology in Computer & Communication Eng., Minor in Data Analytics
CGPA: 8.51/10.0
Jul 2017 - Sep 2021

JOB EXPERIENCE

- **entelligence.ai** San Francisco, CA USA
Founding AI Engineer
Jul 2024 - Present
Leading AI/Backend for entelligence.ai: an AI platform offering codebase chat, automated code reviews, automatic documentation updates, and AI-driven team insights, backed by **Mayfield Ventures**, angel investors from **Nvidia**, **Uber Freight**, **GitLab**.
 - **Personalized & Multi-Agent Systems:** Architected a production-ready multi-agent system from scratch, developed personalized agents by fine-tuning open sourced LLMs (8/4-bit quantization, PEFT/SFT).
 - **High-Performance AI Services:** Engineered an automatic documentation generation and update system; developed an interactive chat interface and built an AI-powered Code Review Bot(from scratch); created an evaluation framework that demonstrated my bot outperformed competitors by achieving a **20%** improvement in detecting critical bugs.
 - **System Architecture:** Developed scalable microservices featuring real-time data pipelines and role-based access control.
 - **CI/CD & Scalability:** Implemented automated CI/CD pipelines, auto-scaling, and load balancing with custom AWS scaling policies.
 - Collaborated with the founder/CEO to align technical execution with product growth while actively contributing to business strategy.
- **LanceDB (YC 22)** San Francisco, CA USA
AI Engineer
Jan 2024 - Jul 2024
LanceDB is an open-source AI database enabling hyper-scalable vector search, advanced retrieval for RAG, and efficient large-scale AI data management, used by **MidJourney**, **ByteDance**, **Bosch** and more.
 - Led integrations with leading open source projects to develop libraries for model loading and multi-modal data storage.
 - **Distributed Training & Fine-Tuning:** Optimized distributed training workflows using Modal Labs for on-demand GPUs, performed GPU fine-tuning on LLMs, and built PyTorch dataloaders.
 - **Collaborations & Advanced Architectures:** Partnered with HuggingFace and LlamaIndex on advanced RAG tutorials(HyDE, optimized chunking, finetuning, reranking etc.).
- **Glance, Inmobi** Bangalore, India
Machine Learning Engineer
May 2021 - Jul 2023
Glance is one of India's fastest-growing unicorns, backed by Google, Mithril Capital, and Reliance Group.
 - Built and maintained large-scale pipelines for data engineering, ML/DL model development, A/B testing and deployment, serving **250M+ DAUs**. Automated CI/CD with Kubernetes and YAML, and developed APIs/GCP container(docker) images.
 - Created a **15%** boost in user engagement via developing predictive pipelines: Utilized decision tree-based methods(LightGBM/XGBoost), graph neural networks and a graph-embedding based similarity algorithm using Node2Vec.

- Developed a shipping partner recommendation engine that cut logistics costs by **14%**.
- Built a gradient boosted classifier with a custom scoring function using weighted TF-IDF and user-item embeddings, increasing gaming platform ad conversions by **11%**.
- Achieved **3x performance gains** by applying adaptive graph contrastive learning to address sparse user data (**cold start problem**) in our deep learning recommendation system.
- Built a live stream ranking model pipeline and global feature store for Roposo's short-video platform, and developed a tool to generate short-video clips from Glance live streams using LLMs (GPT-3.5) and video models (PySlowFast).
- Enhanced code reliability and maintainability through extensive **performance, stress, security testing**, and rigorous **code reviews**. Actively participated in **triaging** and **debugging** to minimize downtime.

INDEPENDENT PROJECTS

- **Morphus: A Novel Framework for Dynamic Adaptation in Multi-Agent Systems (Ongoing)**: Developing a multi-agent framework that dynamically adapts to tasks's needs without human intervention or excessive LLM calls using Reinforcement Learning. Currently being developed and will be prepared for research publication.
- **AlgoTrading**: Developed advanced trading algorithms leveraging financial indicators, ML models (Random Forest, GBM, Ensemble Voting), and cointegration analysis. Engineered custom backtesting with walk-forward optimization and Monte Carlo simulations, achieving up to **84% returns** (beating SNP buy/hold by **22%**) with **drawdowns** as low as **19%** and **Sharpe ratios** up to **1.04**. [click here for details](#).
- **MultiModal RAG for Advanced Video Processing**: Developed a multi-modal solution using LlamaIndex, OpenCLIP, GPT-4V, and LanceDB for advanced video understanding and complex query support. Endorsed by LlamaIndex Founder/CEO Jerry Liu ([click here](#)) and detailed on LlamaIndex's Medium ([click here](#)).
- **GTE-MLX RAG CLI App**: Developed a CLI application for document-specific querying using RAG architecture. Integrated MLX LM and custom GTE model embeddings using `lancedb.embeddings.gte`, a module that I built for lancedb. ([click here to view code](#))

TECHNICAL SKILLS

- **AI/ML**: Traditional ML (regression: linear/non-linear, sampling: bootstrap, PCA, etc., decision trees, Bayesian networks, clustering); deep learning (CNNs, transformers, GANs); graph models (GNNs, Node2Vec); multi-modal video processing; LLM fine-tuning (PEFT/SFT/model quantization); reinforcement learning (MDP, PPO/MAPPO).
- **Programming**: Python, C++, JavaScript/TypeScript, Rust, R, YAML.
- **Frameworks/Tools**: PyTorch, TensorFlow, LangChain, LlamaIndex, Django, React.js, Node.js, Next.js.
- **Cloud & DevOps**: AWS/GCP/Azure, Docker/Kubernetes, CI/CD (GitHub Actions, Jenkins), MLOps (MLflow, DVC), Netlify, Vercel.
- **Data Systems**: PySpark, Kafka, Airflow, SQL (MySQL, PostgreSQL)/NoSQL (MongoDB, Redis), vector DBs (LanceDB, Pinecone).
- **Data Tools/Visualization**: Pandas, NumPy, Scikit-learn, Dask, Plotly, Superset, Streamlit/Gradio.
- **Monitoring/Security**: Grafana/Prometheus, IAM/OAuth, SOC 2 Compliance.