Varadh Kaushik

+1 (857)-313-0854 | kaushik.var@northeastern.edu | LinkedIn | Github

Education

Northeastern University — Boston, MA
M.S. in Artificial Intelligence — *May 2024*
Narsee Monjee Institute of Management Studies — Mumbai, India
B.Tech. in Computer Engineering — *May 2022*

Professional Experience

Machine Learning Engineer — Sept 2024 – Present
WhizAI — *Somerset, NJ*

- HIPAA-Compliant LLM Pipeline: Process 1,000 NLQs/day for pharma dashboards using a quantized LLM, fine-tuned via SFT, DPO and ORPO and dynamic logit bias; fallback to a larger quantized model for under 0.5% of edge cases, ensuring over 99% valid JSON outputs.
- Custom Evaluation & Fallback Logic: Built JSON parse validator, semantic coverage checker, and complexity heuristic with automated escalation from 7B to 70B quantized model or legacy system to guarantee reliability on PHI data.
- Hierarchical RAG Knowledge Base: Deployed across 3,000 Confluence pages with LangChain + FAISS; handles 150+ internal lookups/day at under 2s latency
- Client Engagement & Solution Design: Partnered with pharmaceutical customers to gather requirements, deliver live demos, incorporate feedback, and drive iterative improvements to models and UI.

Machine Learning Intern — Jun 2023 – Aug 2023
WhizAI — *Somerset, NJ*

- RAG-Powered Help-Desk Chatbot: Built with LangChain, vector embeddings, hierarchical retrieval & reranking; deflected 200+ support tickets/quarter and onboarded 80% of new hires within 3 days.
- LLM Benchmarking & Fine-Tuning: Evaluated ChatGPT, Claude, Falcon, LLaMA, Flan-T5 under compute constraints; created PEFT datasets powering live NLQ autocomplete and context-aware SQL generation in production.
- PHI-Safe Dataset Curation: Processed 12K+ anonymized internal logs for sentence completion and data narration tasks, enabling real-time suggestion pipelines.

Machine Learning Engineer — May 2020 – Jul 2022
WhizAI — *Somerset, NJ*

- Intelligent KPI Monitoring: Engineered a multivariate forecasting + ML ensemble with domain-informed breakpoints to track and highlight anomalies across pharma KPIs, delivering on-the-fly alerts within the WhizAI platform.
- Scalable Performance: Orchestrated anomaly-detector, forecasting API, and ingestion pipelines on Kubernetes—seamlessly scaling to handle 3× spike in daily requests during peak loads with no downtime.
- Pin Recommendation Engine: Collaborative filtering & metadata ranking on 50K+ pinboard events; pilot users added 3 pins/session on average.

Software Development Intern — Jun 2019 – Dec 2019
Technology Against ALS [TAALS] — *Mumbai, India*

- Patented AAC Vision Module: Developed a lightweight CNN to detect users' eye-gaze position relative to distinct colored segments on eyewear frames, mapping gaze coordinates to specific color selections—achieving 95% command-recognition accuracy with sub-100 ms end-to-end latency on Android.
- Edge Deployment & Integration: Containerized TensorFlow Serving on Raspberry Pi via Docker and built an Android companion app, delivering seamless on-device inference under strict resource constraints.

Technical Skills

Programming Languages: Python, Java, C++, R, SQL, MATLAB, TypeScript

ML Frameworks: TensorFlow, PyTorch, Scikit-learn, ONNX

LLM Tools: HuggingFace, LangChain, FAISS, LlamaIndex, HF Accelerate

MLOps & Deployment: Docker, Kubernetes, AWS SageMaker, GCP Vertex AI, MLFlow, FastAPI, Weights & Biases, DVC

Data & Big Data: Spark, Pandas, NumPy, Vector DBs(Pinecone, Chroma), PostgreSQL

Web & Front-End: AngularJS, React, RESTful APIs