BHAVEN NAIK

AI Engineer | LLM Infrastructure · RAG Systems · MLOps · AI Platform Engineering

Toronto, ON · naikbhaven11@gmail.com · (902) 318-7215
bhaven-naik.com · linkedin.com/in/bhaven-naik · github.com/bhaven123

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of production experience building and deploying end-to-end LLM applications, RAG systems, and scalable ML infrastructure. Proven track record delivering enterprise-grade GenAI pipelines — from data ingestion and embedding to LLM serving, monitoring, and cloud deployment. Deep expertise in Generative AI, LLM orchestration (LangChain, vLLM), retrieval-augmented generation (Elasticsearch, vector databases), and MLOps (Airflow, MLflow, Docker, Kubernetes, Terraform). Consistently bridges research and real-world impact in regulated and air-gapped environments.

CORE COMPETENCIES

Generative AI & LLMs: LangChain · LlamaIndex · vLLM · Hugging Face · Prompt Engineering
RAG & Search: Elasticsearch · FAISS · Sentence Transformers · Docling · GraphRAG · Neo4j
MLOps & Infrastructure: Airflow · MLflow · DVC · Docker · Kubernetes · Terraform · Ansible
Cloud & DevOps: AWS · Azure · GitLab CI/CD · FastAPI · Flask · Microservices
ML & Deep Learning: PyTorch · TensorFlow · Scikit-Learn · Reinforcement Learning · GANs
Data & Streaming: Kafka · PySpark · Postgres · MongoDB · Redis · Hadoop

PROFESSIONAL EXPERIENCE

AI Engineer · Lemay.ai

Remote (Toronto, ON) | Sep 2023 – Present

- Architected and deployed a production-grade, deterministic RAG Q&A system using Elasticsearch for retrieval, vLLM for LLM inference, and FastAPI microservices; implemented semantic search via Sentence Transformers and Docling-based document extraction — fully containerized with support for secure air-gapped enterprise environments.
- Engineered a GraphRAG-powered due diligence platform using Neo4j knowledge graphs, integrating NER, multi-document summarization, and multi-source ETL automation via LLM pipelines, reducing manual research time by an estimated 40%.
- Built and deployed full-stack GenAI applications with LLM integration using Angular (frontend) and FastAPI (backend), delivering interactive AI interfaces consumed by enterprise clients across regulated industries.
- Led an ML-based time-series forecasting project, developing and operationalizing predictive models that surfaced actionable insights for critical business planning cycles.
- Provisioned scalable, secure cloud infrastructure using Terraform and Ansible across AWS environments, reducing deployment lead time by ~50% through infrastructure-as-code standardization.
- Automated end-to-end ML training workflows and model lifecycle management using Python, Apache Airflow, and MLflow, enabling reproducible experiment tracking and one-click model promotion to production.
- Designed distributed reinforcement learning environments using multi-GPU setups, accelerating training iteration speed for experimentation pipelines.
- Contributed to a mobile application platform incorporating Kafka for real-time, event-driven messaging, supporting high-throughput AI inference pipelines at scale.

AI Research Assistant (Intern) · St. Francis Xavier University

Antigonish, NS | Sep 2021 – Apr 2022

- Led a deep learning research project applying Generative Adversarial Networks (GANs) for data augmentation in medical imaging; implemented and fine-tuned a Deep Convolutional GAN (DCGAN) using PyTorch Lightning on the HMDB51 Human Action Recognition dataset.
- Built a scalable multi-GPU training pipeline, optimizing model architecture and hyperparameters to improve augmentation fidelity; validated synthetic video quality using PyTorchVideo's pre-trained classifiers.

- Contributed novel findings to lab research on model robustness and AI-driven data augmentation techniques in computer vision.

SELECTED PROJECTS

Air-Gapped RAG Platform *[vLLM · Elasticsearch · Sentence Transformers · FastAPI · Docker]*

Production Q&A system deployable in air-gapped enterprise environments. Docling-powered document ingestion, semantic chunking, vector embedding, and vLLM-served LLM inference behind FastAPI microservices. Fully containerized with CI/CD.

GraphRAG Due Diligence Tool *[Neo4j · LangChain · Python · FastAPI]*

Knowledge-graph-powered analysis platform for M&A due diligence. Automated NER, entity resolution, multi-source ETL, and multi-document summarization via LLM pipelines, reducing analyst research cycles by ~40%.

DCGAN Video Augmentation (Research) *[PyTorch Lightning · PyTorchVideo · Python]*

Research implementation of a Deep Convolutional GAN for Human Action Recognition data augmentation on HMDB51. Scalable training pipeline with multi-GPU support and benchmarked against pre-trained classifiers.

Diabetic Retinopathy Classifier *[TensorFlow · Keras · VGG16 · Flask · AWS EC2]*

Fine-tuned VGG16 to classify 5 severity grades of diabetic retinopathy from fundus images. Built Flask web interface and deployed on AWS EC2 for clinical accessibility.

EDUCATION

Master of Applied Computer Science · St. Francis Xavier University · Antigonish, NS · 2020 – 2022

Bachelor of Computer Engineering · University of Mumbai · Mumbai, India · 2016 – 2020

CERTIFICATIONS & MEMBERSHIPS

IBM Machine Learning Essentials