Nirav Polara

Toronto, ON | (226) 350-0030 | polranirav@gmail.com | linkedin | Github

EDUCATION

University of Windsor **Windsor, Canada**
Master of Engineering: Electrical and Computer Engineering Jan 2024 – Apr 2025

Gujarat Technological University **Ahmedabad, India**
Bachelor of Engineering: Computer Engineering Aug 2015 – Sep 2019

SKILLS

- **Programming Languages :** Python (Advanced), C++, Rust, Java, TypeScript
- **AI & ML Frameworks**: PyTorch, TensorFlow, LangChain, LlamaIndex, Hugging Face Transformers
- **Generative AI & LLMs**: OpenAI API, Anthropic, Gemini, RAG (Retrieval-Augmented Generation), Fine-Tuning (LoRA/QLoRA), Prompt Engineering, Vector Embeddings
- **Vector Databases**: Pinecone, Weaviate, Qdrant, ChromaDB, FAISS
- **AI Agents & Orchestration**: AutoGen, CrewAI, LangGraph, Semantic Kernel, Agentic Workflows
- **Big Data Tech**: Apache Spark, Apache Kafka, Hadoop, Databricks, Apache Airflow, ETL/ELT Pipelines.
- **APIs & Backend**: FastAPI, RESTful APIs, GraphQL, Microservices Architecture
- **Data Storage** : PostgreSQL, MySQL, MongoDB, Oracle, Redis
- **MLOps & Deployment**: Docker, Kubernetes, AWS SageMaker, Azure AI Foundry, Vertex AI, MLflow, CI/CD for ML

PROJECTS

Enterprise Multi-Agent RAG System | Python, LangGraph, OpenAI/Cohere, Pinecone, FastAPI, Docker

- Architected autonomous agents using LangGraph to execute complex enterprise workflows without human intervention
- Engineered RAG pipelines with Pinecone to retrieve context-aware data, reducing LLM hallucinations by 40%.
- Exposed agentic capabilities via secure FastAPI endpoints, enabling seamless integration with external applications.
- Optimized inference latency by deploying quantized models via vLLM, ensuring real-time user responsiveness.
- Implemented RBAC security protocols to ensure safe data handling within generative workflows.

Real-Time Anomaly Detection & MLOps Pipeline| Python, Apache Kafka, Spark, AWS (SageMaker, EKS), Terraform, MLflow

- Built real-time fraud detection pipeline using Kafka and Spark for sub-second anomaly processing.
- Deployed scalable microservices on AWS EKS using Docker and Kubernetes for high-availability inference.
- Automated model retraining workflows using GitHub Actions and MLflow to combat data drift.
- Provisioned cloud infrastructure via Terraform, ensuring reproducible and secure production environments.
- Integrated Prometheus and Grafana for real-time system monitoring, ensuring 99.9% API uptime reliability.

EXPERIENCE

Techy Panther **Ahmedabad, India**
Software Developer Sep 2021 – Sep 2023

- Engineered high-performance Python backends for mobile applications, optimizing API response times by 30% to support real-time data processing.
- Collaborated on cross-platform architectures, integrating third-party AI/ML APIs to enhance user engagement features within iOS and Android ecosystems.
- Designed scalable database schemas in SQL, ensuring data integrity and efficient retrieval for analytics-driven application features.
- Automated deployment pipelines using CI/CD tools, reducing production deployment cycles and ensuring consistent code quality across environments

Webunity InfoTech **Surat, India**
Software Developer Apr 2019 – Feb 2021

- Built robust ETL data pipelines to synchronize inventory and sales records across e-commerce platforms, reducing manual data errors by 15%.
- Optimized server-side logic and database queries, significantly lowering latency for high-traffic e-commerce storefronts.
- Integrated Google Analytics APIs to ingest user behavior data, creating structured datasets for downstream reporting and insight generation
- Managed complex data migrations during platform upgrades, ensuring 100% data preservation and zero downtime for client business operations