Ronit Virwani

rvirwani@binghamton.edu | www.linkedin.com/in/ronitvirwani | https://github.com/britster03

EDUCATION

Binghamton University, State University of New York
GPA: 3.92/4
Master of Science in Computer Science *December 2025*

MIT ADT University, India
Bachelor of Technology in Information Technology *May 2024*

PROFESSIONAL EXPERIENCE

AKOS **,** *AI/LLM Engineer* **|** Scottsdale, AZ *May 2025 – Present*

- **Shipped LLM-powered AI solutions** from prototype to production, designed retrieval-augmented pipelines (RAG) and domain-specific fine-tuning using LangChain/LangGraph and FastAPI, supporting 99.9% uptime and sub-second (<800ms) API responses.
- **Built scalable data pipelines** that process and embed 30K+ records of structured and unstructured data monthly, with access to **ChromaDB Cloud *(Beta Version for Developers)*** for fast semantic search and integrating serverless, scale-to-zero architecture, reducing compute costs by 35% and improving throughput.
- **Implemented advanced retrieval techniques**—combining hybrid RAG (dense/sparse retrieval) and cache-augmented generation(CAG) to boost answer accuracy by 20%, minimize hallucinations and deliver reliable, real-time outputs
- **Managed the full AI workflow**—from microservice API orchestration, CI/CD, and monitoring to incident resolution—while working closely with product and domain experts to deliver several revenue-ready features from prototype to release.

Techpeek. **,** *AI Engineer* **|** Bengaluru, India *February 2024 – August 2024*

- **Executed** the creation and launch of a **Legal AI platform**, significantly enhancing legal services.
- **Implemented retrieval-augmented generation (RAG)** using **LangChain** and integrated open-source **Large Language Models.**
- **Designed** a customizable **NLP search service**, enhancing precision through advanced retrieval
- **Managed & created MilvusDB** and **ChromaDB collections** for **100 k+ documents**, reducing latency
- **Integrated** an **ML model** for legal case predictions with **LLM reasoning**, raising accuracy to **85 %**
- **Created** robust **Docker** configurations and simplified **Nginx** settings, achieving **99.9% system uptime**.
- **Devised** backend security with **FastAPI**, cutting the unauthorized access.

RESEARCH EXPERIENCE

SUNY Binghamton University, *Graduate Research Assistant* | Binghamton, NY *September 2024 - Present*

- Redesigned Transformer and BiLSTM models under the supervision of Prof. Sujoy Sikdar to address the Tip-of-the-Tongue phenomenon, utilizing TensorFlow and PyTorch to achieve improvement in word retrieval accuracy.
- Developed a legal-text retrieval pipeline using SentenceTransformer embeddings and ChromaDB to search 570+ Indian Penal Code sections; introduced a mask-aware snippet extractor and MiniLM cross-encoder reranker, raising Top-1 accuracy to 60% and Top-2 to 80% on held-out queries, with automated evaluation and custom data cleaning.

TECHNICAL SKILLS

Languages : Java, Python, C++, Javascript, Rust
Machine Learning Platforms : AWS Sagemaker, AWS Bedrock
Cloud Technologies : AWS (S3, EC2, Lambda), Docker, GCP
Web Development : React Js, Next Js, Svelte, Langchain, FastAPI
Libraries & Frameworks : Pytest, React, NodeJS, ExpressJS, Netflix Zuul, Nginx, Pandas
Databases : MySQL, MongoDB, ChromaDB, MilvusDB, Hadoop

PROJECT EXPERIENCE

Referral∞Inc | Bridging talent and opportunity—with real connections and real intelligence *December 2024 - Present*

- Building ReferralInc (AI SaaS platform): architected AI/LLM-powered workflows to streamline candidate–employee matchmaking, generate data-driven assessments and feedback visualizations, and automate referral status tracking—solving a real problem of hiring bottlenecks with a modern, conversational user experience.Targeting 3× faster placements and 40% higher engagement for users

teamius//AI | Your Own AI-Powered Consulting Team, On Demand *March 2025 - Present*

- Built an end-to-end AI SaaS platform that brings virtual LLM-powered consultants to automate business analysis—engineered a multi-agent workflow (React, FastAPI, Python) simulating real teams to deliver tailored solutions, dynamic dashboards, and interactive insights across 10+ business scenarios. (Pre-launch)

Applicient | Job Hunt… Supercharged! *May 2025 - Present*

- Built **Applicient** as the candidate-side sibling to **Referral∞Inc**—with LLM workflows for real-time job capture, resume analysis, fit prediction, and career coaching. Designed an Excel-style dashboard to empower job seekers, completing the vision started with ReferralInc. Pre-launch, targeting 5K+ users in year one.

ACHIEVEMENTS AND AWARDS

- Led Google Developers Program at MIT ADT University, selected among top 300 students in India.
- Won special recognition award at the Smart Pune Health Hackathon and advanced to finals in 5 national hackathons in India.