

Koushik Sarun Reddy

📍 Dallas, TX 📞 (940) 287-4359 ✉ koushikchess12@gmail.com 🌐 koushiksarunreddy 🌐 koushiksarun 🌐 koushikreddy.me

Professional Summary

AI Engineer with experience developing **LLM-powered applications**, **Retrieval-Augmented Generation (RAG)** systems, and **Computer Vision** solutions using **Python**, **PyTorch**, **TensorFlow**, and **FastAPI**. Passionate about building scalable AI software that automates workflows and solves real-world problems.

Education

University of North Texas

Denton, TX

Master of Science in Information Science

Graduated May 2026

GPA: 4.0 — Coursework: Machine Learning, Data Analytics, Information Retrieval, Database Systems, Cloud Computing

Skills

Languages: Python, C++, SQL, JavaScript, Java, TypeScript

AI & Machine Learning: PyTorch, TensorFlow, Scikit-Learn, Generative AI, LLMs, RAG, Deep Learning, Computer Vision, NLP

AI Systems: LangChain, LangGraph, AI Agents, Prompt Engineering, Embeddings, Vector Databases, Semantic Search, Fine-Tuning

Data & Cloud: PySpark, ETL Pipelines, Pandas, NumPy, BigQuery, AWS, Azure AI, Docker, Git, GitHub, CI/CD

Web & Databases: FastAPI, React.js, REST APIs, MySQL, PostgreSQL, Supabase

Experience

AI Intern

Apr 2026 – May 2026

Hirello AI

Remote

- Developed **LLM-powered applications** using **Gemini API**, **Generative AI**, and **Python** for intelligent automation workflows.
- Built **Retrieval-Augmented Generation (RAG)** systems and **Semantic Search** pipelines using **Embeddings** and **Vector Databases**.
- Engineered **Prompt Optimization** and **Conversational Memory** systems, improving response relevance by approximately **35%**.

Generative AI Intern

Dec 2023 – May 2024

Sithafal Technologies

Tirupati, India

- Developed **Generative AI** solutions using **Python** and modern AI frameworks for real-world business use cases.
- Built **Automation Tools** and intelligent workflows, reducing repetitive manual tasks by approximately **30%**.
- Performed **Data Preprocessing**, **Feature Engineering**, testing, debugging, and optimization for **Machine Learning** applications.

Projects

SLIME – AI Chatbot Platform

github

Technologies Used: Python, Gemini API, LangChain, React, FastAPI

- Built a production-ready **AI Chatbot** using **Gemini API**, **LangChain**, **FastAPI**, and **React** for context-aware responses.
- Implemented **Semantic Search**, **Embeddings**, and **RAG**, improving answer relevance by approximately **35%**.
- Developed scalable **FastAPI** backend services supporting real-time **AI inference** and frontend integration.

LLM-Powered Research Assistant

github

Technologies Used: Python, LangChain, FAISS, PyTorch

- Developed **AI-powered research workflows** using **Foundation Models**, **LLMs**, and retrieval-based search techniques.
- Built **RAG pipelines** using **Vector Databases** and **Embedding-based Retrieval** to automate knowledge extraction.
- Automated summarization and information synthesis, reducing manual research review time by approximately **40%**.

Multi-Modal Traffic Event Understanding System

github

Technologies Used: Python, PyTorch, OpenCV, TensorFlow

- Developed a **Multimodal AI** system combining visual and structured traffic data for **event classification**.
- Built **Deep Learning** models for **Object Detection**, tracking, and temporal traffic event recognition.
- Improved event classification performance by approximately **28%** through preprocessing and **model optimization**.

Certifications

Microsoft Azure AI Fundamentals (AI-900) — Google Data Analytics Professional Certificate — IBM Machine Learning Professional Certificate

Achievements

- Built and deployed multiple **AI-powered applications** using **LLMs**, **RAG**, **Computer Vision**, and **Generative AI**.
- National Chess Player with an official **FIDE Rating of 1574**.