MANVIK TALWAR

8800287735 • manvik.talwar@gmail.com • linkedin.com/in/manvik-talwar •
https://github.com/CodexManvik

AI/ML Engineer

Computer Science student specialising in Generative AI and MLOps. Proven ability to design, build, and deploy end-to-end applications featuring Retrieval-Augmented Generation (RAG) pipelines, local LLMs (Phi-3, Mistral-7B), and cloud services like Azure OpenAI. Proficient in Python, PyTorch, and containerization with a focus on building scalable and efficient AI systems.

SKILLS

Programming Languages: C/C++, JavaScript, Python

AI/ML: PyTorch, Hugging Face Transformers, Pandas, Xarray, TensorFlow, Scikit-learn, Retrieval-Augmented Generation (RAG), Vector Databases (ChromaDB & FAISS), LLM Quantization

Web & API Development: FastAPI, Streamlit, HTML, CSS, JavaScript

Cloud and MLOps: Microsoft Azure, Azure OpenAI Service, Docker, Git, CI/CD, Agile/Scrum

Databases: PostgreSQL, SQLite

PROJECTS

FloatChat | AI-Powered ARGO Ocean Data Explorer

- **Architected** and built a full-stack conversational AI to democratize access to complex ARGO oceanographic data, enabling natural language queries on a massive scientific dataset.
- **Implemented** an end-to-end Retrieval-Augmented Generation (RAG) pipeline using a local LLM (Microsoft Phi-3), Nomic Embeddings, and a ChromaDB vector store.
- **Engineered** a robust backend using FastAPI with a dual-database system (PostgreSQL for structured data, ChromaDB for semantic search) to ensure efficient and accurate data retrieval.
- **Developed** an interactive Streamlit frontend with Plotly visualizations (maps, profile plots) to present query results dynamically to users.

Local AI Assistant | Python, TensorFlow/PyTorch, React.js | In Development

- **Engineered** a privacy-focused, offline conversational assistant leveraging a quantized local LLM (Mistral 7B) and a FAISS-based RAG pipeline for on-device operation.
- **Optimized** and benchmarked the system on consumer hardware, achieving ≈12 tokens/sec throughput and reducing peak VRAM usage to under 4GB through INT4 quantization.
- **Designed** a modular plugin system with sandboxed commands to safely automate tasks like file system access and web searches, achieving a ≥95% success rate in deterministic tests.

WORK EXPERIENCE

Path Infotech — 06/2025 - 08/2025

AI Development Intern

- **Developed** and deployed a production-grade GenAI application using Azure OpenAI Service, achieving 99.9% uptime while scaling to support over 100 concurrent users.
- **Engineered** backend services that reduced API response time by 25% through optimized routing and payload handling.
- **Automated** deployment workflows on Azure App Service, cutting release lead time to ~12 minutes and enabling multiple weekly updates.

EDUCATION

Bachelor of Technology in Computer Science (AI/ML)
Manipal University Jaipur — Jaipur • 01/2023 - 01/2027

Relevant Coursework: Machine Learning, Computer Vision, Data Structures and Algorithms, Artificial Intelligence, Deep Learning, Cloud Computing

Higher Secondary Education
Gaurs International School — 01/2021 - 01/2023

Key Skills Developed: Web Development (HTML, CSS, JavaScript), Python Programming, Problem-Solving

AWARDS & SCHOLARSHIPS

Student Excellence Award — 05/2025
Manipal University Jaipur

CERTIFICATIONS

Agile Software Development
Programming in Python
Back-End Development
Data Structures and Algorithms
Design and Analysis of Algorithms