Chirag Malik

+1-709-749-4598 | Email | LinkedIn | GitHub | Portfolio

Education

Memorial University St. John's, NL
Master of Artificial Intelligence *Sep. 2024 – Present*

Meerut Institute of Engineering and Technology Meerut, India
Bachelor of Technology in Computer Science and Information Technology *Aug. 2019 – Jul. 2023*

Experience

AI Solution Intern May 2025 – Aug. 2025
Genesis AI Garage *St. John's, NL*

- Created a scalable, full-stack RAG chatbot using FastAPI, React, and Llama 8B, enabling real-time, accurate semantic search across 150+ documents with minimal delays.
- Designed and built a flexible backend with PostgreSQL, Firebase Auth, and REST APIs to manage secure sessions and easy document uploads.
- Improved user experience and reliability by fine-tuning performance, reducing hallucinations, and enhancing the UI with Tailwind.

Undergraduate AI Research Assistant Sep. 2022 – Feb. 2023
Meerut Institute of Engineering and Technology *Meerut, India*

- Built a CNN based brain tumor detection tool using MRI scans, achieving 98.25% training and 97% validation accuracy.
- Improved model generalization by applying data augmentation and cross-validation, reducing overfitting by 15%.
- Published results at an IEEE conference and presented to 100+ attendees, showcasing advances in AI for medical diagnostics.

Recent Projects

Fine-Tuning Pipeline for Local LLMs | *Python, PyTorch, LoRA, PEFT, QLoRA, Unsloth*

- Developed a generalizable fine-tuning pipeline for Qwen 3 (0.6B) and other local LLMs using LoRA-based (PEFT), enabling efficient adaptation to any domain-specific dataset.
- Used the Unsloth framework with 4-bit quantization (QLoRA-style) to reduce memory footprint and training cost, integrating Hugging Face `Datasets` for instruction-style data preprocessing.
- Exported the model in GGUF format for Ollama and offline inference, supporting fast, private, and low-resource deployment of custom language models on local devices.

Local LLM RAG Chatbot | *Python, LangChain, Ollama, ChromaDB, Hugging Face, Streamlit*

- Built a modular RAG chatbot using LangChain, Ollama (for local LLMs), and ChromaDB, allowing users to upload PDFs and ask context-aware questions completely offline.
- Integrated Sentence Transformers for generating document embeddings and ChromaDB for fast, local semantic retrieval.
- Developed an interactive Streamlit UI for seamless PDF upload and real-time chat with private, offline inference.

Technical Skills

Core Programming: Python (OOP, DSA, scripting)
Frontend: HTML/CSS, basics of React.js and Tailwind CSS, Streamlit (for AI prototyping)
Backend: Python (FastAPI), basics of Flask and REST APIs
AI/ML: NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch, OpenCV, Matplotlib
GenAI: LLMs (prompt engineering, RAG, fine-tuning basics), LangChain, LlamaIndex, Hugging Face Transformers, vector databases
Data & Databases: SQL (PostgreSQL, MySQL), Excel, Power BI, NoSQL (basics of MongoDB)
Cloud/DevOps: Basic AWS (S3, EC2, Lambda, SageMaker), Docker, MLflow, GitHub Actions (CI/CD)
Developer Tools: Figma (UI/UX basics), Git/GitHub, Postman (API testing), Jira, Power Apps (basic usage)