

# Toshit Dwivedi

Lucknow, India

+91-9729925699 • toshitdwivedi@gmail.com • toshitdwivedi.me  
in toshit-dwivedi • ToshitDwivedi

## Summary

AI Engineer with an M.Sc. in Data Science (IIIT Lucknow) and active hands-on experience at **Zencia AI** building **GenAI and LLM-powered applications** — including voice agents, conversational AI systems, RAG pipelines, and Agentic frameworks. Proficient in **Python, FastAPI, LangChain, and LLM APIs** (Claude, OpenAI), with strong foundations in **machine learning, deep learning, and NLP**. Experienced in working across the **full AI development lifecycle** — from prototyping and experimentation to production deployment and monitoring.

## Education

### IIIT Lucknow

M.Sc. Data Science, CGPA: 7.67 / 10

Coursework: Machine Learning, Deep Learning, NLP, Statistical Modeling, Reinforcement Learning, Big Data Analytics

Lucknow

2024–2026

### Amity University Noida

B.Sc. (Hons) Mathematics, CGPA: 8.33 / 10

Coursework: Linear Algebra, Probability & Statistics, Econometrics, Operations Research, Mathematical Modelling

Noida

2021–2024

## Experience

### Zencia AI

AI Engineer Intern

Lucknow

Feb 2026–Present

- Building **LLM-powered voice agents and conversational AI systems** using Claude and OpenAI APIs, integrating speech-to-text, NLP reasoning, and text-to-speech into real-time inference pipelines.
- Implementing **RAG (Retrieval-Augmented Generation) pipelines** — covering document ingestion, text chunking, embedding generation, vector indexing, and prompt orchestration using LangChain.
- Developing **Agentic AI workflows** using LangChain and MCP (Model Context Protocol) for multi-step reasoning, tool use, and task automation across business use cases.
- Building and maintaining **FastAPI backend services** to expose AI model functionalities via REST APIs, implementing authentication, logging, error handling, and streaming responses.
- Applying **prompt engineering techniques** to optimise LLM response quality, coherence, and task-specific accuracy across diverse application requirements.
- Using **agentic coding tools** (Claude Code, GitHub Copilot) to accelerate development iteration while maintaining clean, modular, production-quality code.

### IIIT Lucknow

Teaching Assistant (Statistics & Optimization)

Lucknow

Aug 2024–Dec 2024

- Guided **220+ students** in statistical inference, optimization algorithms, and ML fundamentals — developing strong communication and problem-solving skills applicable to technical team environments.

## Projects

### Deep Learning Capstone: Computer Vision Pipeline | GitHub

- Built an end-to-end **CNN image segmentation pipeline** (U-Net / ResNet) using **PyTorch and TensorFlow** — covering data preprocessing, model training, evaluation, hyperparameter tuning, and iterative improvement.
- Applied transfer learning, data augmentation, and attention mechanisms; exported model to **ONNX** for deployment-ready inference.

### Fraud Risk Prediction System | GitHub

- End-to-end ML pipeline using **XGBoost and CatBoost** — feature engineering, model selection, cross-validation, and threshold tuning; deployed as a **FastAPI inference service** with Docker containerisation.
- Achieved **ROC-AUC of 0.91**; maintained comprehensive model documentation and experiment tracking via **MLflow**.

### Finance Analytics Lakehouse & Data Pipeline | GitHub

- Architected a scalable data lake using **Apache Spark, Hive, and Airflow** for large-scale financial dataset processing, with automated ETL pipelines and data quality validation.

## Technical Skills

**GenAI & LLMs:** Claude API, OpenAI API, LangChain, RAG Pipelines, Agentic Frameworks, MCP, Prompt Engineering, Vector Embeddings, Semantic Search

**Backend & APIs:** FastAPI, REST API Development, JWT Auth, Streaming Responses, Docker, Git, Claude Code / GitHub Copilot

**ML & DL:** PyTorch, TensorFlow, Scikit-learn, XGBoost, CNN, LSTM, NLP, Transfer Learning, ONNX, MLflow

**Data & Infra:** Python (Pandas, NumPy), SQL, Apache Spark (PySpark), Hive, Airflow, Jupyter