

ASHUTOSH ROY

AI/ML Engineer

☎ +91-9810599837 | ✉ ashu2003roy@gmail.com | 🌐 mitovoid.netlify.app | 🌱 ashutoshroy02 | 📍 Gurugram, Haryana

SUMMARY

AI/ML Engineer with **1+ year of hands-on experience** building AI applications using **Generative AI, LLMs, RAG, and Speech AI**. Published researcher at **RECCAP 2026** and Research Intern at **IIT Delhi**. Skilled in **PyTorch, TensorFlow, LangChain, FastAPI, Python, NLP, Transformer Fine-Tuning, and MLOps**. Built AI systems that automate tasks, improve search accuracy, reduce manual work, and speed up business processes through end-to-end ML pipelines and intelligent AI agents.

EXPERIENCE

Research Intern

Jan 2025 – Jun 2025

Indian Institute of Technology (IIT) Delhi

New Delhi, India

- Built an OCR-powered document understanding pipeline for **10,000+** archaeological documents by benchmarking Tesseract, Marker, and Gemini OCR, improving extraction accuracy and reducing manual document processing.
- Designed a hybrid retrieval system using FAISS, Pinecone, and Apache Solr for **10,000+** historical records, improving search accuracy by **35%** and enabling faster document discovery.
- Developed scalable FastAPI services for real-time LLM-based search, reducing API response time and supporting **1,000+** search requests per day.
- Co-authored the research paper **SARCH: Multimodal Search for Archaeological Archives**, helping build an AI-powered search platform that improved researcher productivity. arxiv.org/abs/2511.05667

AI/ML Engineer Intern

Oct 2024 – Nov 2024

MitoVoid AI

Remote

- Fine-tuned Mistral and LLaMA models for healthcare applications, improving medical summarization quality and chatbot performance.
- Built an AI chatbot that reduced patient query resolution time by **40%**, improved customer experience, and lowered support costs.
- Developed GenAI and RAG-based healthcare workflows that automated repetitive tasks, increased team productivity, and improved response accuracy.
- Built end-to-end ML pipelines from data preprocessing to API deployment, reducing deployment time and speeding up AI solution delivery.

PROJECTS

Exam-Helper RAG System | Python, LLaMA 3, FAISS, Pinecone, LangChain, Streamlit

Dec 2024 – Apr 2025

- Built a **RAG-powered** question answering system using LLaMA 3 with support for PDFs, images, and YouTube transcripts.
- Implemented **FAISS** and **Pinecone** retrieval, reducing LLM hallucinations, improving answer accuracy, and enabling **sub-second** semantic search.

Medical Chatbot (LLM Fine-Tuning) | Python, Mistral 7B, QLoRA, Tesseract OCR

Mar 2024 – Apr 2024

- Fine-tuned **Mistral 7B** using **QLoRA**, reducing GPU memory usage by **75%** while maintaining model performance.
- Built an OCR pipeline to process **1,000+** clinical reports, improving data extraction for downstream LLM inference.

Industrial Defect Detection System | YOLOv8, PyTorch, OpenCV, CUDA, Python

Feb 2025 – Apr 2025

- Built a real-time **YOLOv8** defect detection system for automated quality inspection, reducing manual inspection effort.
- Detected **4** defect types (scratches, cracks, dents, missing components) with GPU-accelerated inference for low-latency deployment.

TECHNICAL SKILLS

AI/ML & Research: LLMs, GenAI, Intelligent Automation, RAG, NLP, LangChain, Transformer Fine-Tuning (QLoRA/LoRA), Self-Supervised Learning, Multimodal AI, Speech Processing, Evaluation & Benchmarking, Dense Retrieval, Foundation Models

Libraries & Frameworks: PyTorch, Hugging Face, LangChain, TensorFlow, Scikit-learn, Pandas, NumPy, FastAPI, Streamlit, MLOps

Databases & Search: FAISS, Pinecone, Apache Solr, MySQL

Languages: Python, C++, SQL

Tools: Git, Docker, Linux, Jupyter Notebook, Google Colab, VS Code, OpenAI GPT, Claude, Hugging Face Transformers

ACHIEVEMENTS

- Published **2 research papers**, including one accepted at **RECCAP 2026 (IEEE)** and another co-authored with **IIT Delhi**.
- Solved **150+** DSA problems on LeetCode, strengthening data structures and algorithms.
- Won **2 AI/ML Hackathons**, building innovative AI solutions under time constraints.
- Made contributions to **open-source AI/ML projects** on GitHub and Hugging Face.
- Earned certifications in **Machine Learning, Deep Learning, Generative AI, NLP, and Python**.

EDUCATION

Chhattisgarh Swami Vivekananda Technical University (CSVTU)

B.Tech (Hons) in Computer Science and Engineering (Artificial Intelligence)

2022 – 2026

Bhilai, Chhattisgarh

St. Columbus School

Higher Secondary (Class XII - Science PCMB) - 87%

2020 – 2021

Faridabad, Haryana