

MOHSIN ABBAS KHAN

AI / ML Engineer · Computer Vision · NLP & LLMs · Deep Learning · MLOps

+91 9129076723 · mohsin.abb.khan@gmail.com · linkedin.com/in/mohsin-abbas-7252b126b
github.com/mohsinakh · mohsinkhan.online · Delhi, India · Immediate Joiner

PROFESSIONAL SUMMARY

AI/ML Engineer with **1+ year of production experience** building and deploying machine learning systems in a real healthcare-research environment (**ICMR HQ, New Delhi**). Work spans computer vision (YOLO anemia detection, **92% accuracy**), multimodal deep learning (Wav2Vec2 + NLP, ~87%), **RAG pipelines**, LLM fine-tuning, and agentic LLM systems — all taken from research to deployed, production-grade APIs. B.Tech CS (Data Science & AI), **CGPA 9.4, Gold Medalist**.

TECHNICAL SKILLS

AI / ML: Computer Vision (OpenCV, YOLO), NLP, Transformers, LLMs (GPT-2, fine-tuning), RAG (FAISS, BM25, LangChain), AI Agents & Tool Calling, Audio/Speech (Wav2Vec2), Deep Learning, PyTorch, Scikit-learn, Hugging Face

Data & Analysis: Data Wrangling, EDA, Statistical Analysis, Pandas, NumPy, Matplotlib, Power BI, IBM Watson

Backend & MLOps: Python, FastAPI, Django, REST APIs, WebSockets, PostgreSQL, MongoDB, Redis, Docker, AWS (EC2, S3, Lambda), GitHub Actions CI/CD, Linux

Tools: Git, Hugging Face Hub, Ollama, pytest, Playwright, Postman, Jupyter, IBM Cloud

WORK EXPERIENCE

AI / ML Engineer · GA Digital Web Word Pvt. Ltd | Feb 2026 – Present | New Delhi

Contracted on-site @ ICMR HQ (Indian Council of Medical Research) — same posting previously under Raygain Technologies Pvt. Ltd (Sep 2025 – Feb 2026)

- Built a **computer-vision anemia detection system (OpenCV + YOLO)** on conjunctiva imaging achieving **92% accuracy** — cut clinical screening time by **40%** vs manual review; deployed as a **live inference API**.
- Developed a **multimodal depression detection model** fusing audio features (**Wav2Vec2**) with NLP text signals, reaching **~87% accuracy** on held-out test data.
- Built **automated data pipelines** for large-scale research-paper retrieval from medical databases (PubMed and others), cutting manual collection effort by **35%** and feeding downstream ML training at scale.
- Designed **REST APIs (FastAPI + Django)** serving AI inference results into healthcare research pipelines at **1,000+ requests/day** with optimised query performance.
- Built **analytics dashboards** and data-visualisation tooling to surface ML model outputs to non-technical public-health researchers.

Backend Developer Intern · Rough Note | Feb 2025 – May 2025 | Remote

- Reduced API response time by **25%** through query profiling, database indexing, and **Redis caching**.
- Built and maintained REST API features in a distributed Agile team; participated in sprint planning and code reviews.

Data Science Intern · Unified Mentor | Jul 2024 – Aug 2024 | Lucknow

- Built end-to-end predictive ML models — data preprocessing, EDA, feature engineering, and time-series forecasting; presented findings to non-technical stakeholders.

KEY PROJECTS

RAG-Based Research Paper Assistant | Python · FAISS · BM25 · LangChain · FastAPI

- Built a **hybrid RAG pipeline** (FAISS dense vector search + BM25 keyword retrieval) over academic papers; hybrid retrieval cut **LLM hallucinations by 40%** on internal benchmarks.
- Full pipeline — PDF ingestion → chunking → embedding → hybrid retrieval → LLM synthesis with sub-second latency — deployed as a **production FastAPI service**.

Enterprise AI Assistant — Tool-Calling LLM Agent | Python · FastAPI · Ollama (Llama 3.1) · Pydantic v2 · Docker

- Built an **agentic LLM system with native tool calling** — **14 tools across 6 business domains** selected by the model with structured JSON arguments; multi-turn session memory plus cross-session preference memory, Pydantic guardrails; runs fully offline on local Llama 3.1 via Ollama; dockerized with pytest suites.

GPT-2 from Scratch & Fine-Tuning | PyTorch · tiktoken · NumPy — github.com/mohsinakh/gpt2

- Reimplemented the **full GPT-2 architecture from scratch in PyTorch** — embeddings, multi-head causal self-attention, transformer blocks, LM head, and weight-loading of OpenAI's pretrained parameters — with hand-built tiktoken BPE tokenization and data loaders.
- **Fine-tuned twice:** classification head for SMS spam detection and instruction tuning on ~1,000 instruction–response pairs.

Sentiment Sense — Multimodal Sentiment Analysis (Team Lead, 4 engineers) | Python · CNNs · Transformers · FastAPI · WebSockets — github.com/mohsinakh/SentimentSenseApiMain

- Led a 4-engineer team building a real-time sentiment/emotion system across **text, images, video, and live webcam** (CNN + Transformer ensemble, **~90% accuracy**) with REST + WebSocket streaming and live YouTube/Reddit ingestion; deployed on Hugging Face Spaces.

EDUCATION & CERTIFICATIONS

B.Tech — Computer Science (Data Science & AI) | Integral University, Lucknow | Jul 2021 – Jul 2025

CGPA: 9.4 / 10 · **Gold Medalist — 1st Position in Department**

11× IBM Certifications — Applied AI, Machine Learning, Data Science, Cloud, Python · IBM Data Visualization with Python (Sep 2024)