Arnab Mandal

+91-9560452773 | arnabmandal2912@gmail.com | LinkedIn | Github | Portfolio

EDUCATION

Shiv Nadar University — Noida, Delhi-NCR, India
B.Tech in Computer Science Engineering **(8.46/10.0)** — *Aug 2023 - Jul 2027*

Heera Lal Public School, Class XII — Madan Pur Dabas, Delhi-NCR, India
Percentage : **91%** — *May 2023*

Navy Children School, Class X — Visakhapatnam, Andhra Pradesh, India
Percentage : **97%** — *May 2021*

EXPERIENCE

AI Engineer Intern, DeepLure Research Pvt Ltd, Amravati — Dec 2025 - Jan 2026

- Optimized Llava 1.5 Pipeline serving 50+ clients by migrating inference to TensorRT alongside Triton inference server.
- Business Impact: Projected Increase in vRAM utilization by 30% and serving 8x clients with no additional resources.

AI Intern, Critical AI Pvt Ltd, Bengaluru — Jul 2025 - Sep 2025

- **GenAI Audio & Image based narrative creation using Meta's Llama-3.2-8B**
  - Built multimodal RAG pipelines (FAISS, llama-cpp, OSM, TTS, STT, MongoDB) behind FastAPI microservices.
  - Business Impact: Expanded feature coverage and improved functionality by over 50% of base provided console.

AI Intern, WESEE, Ministry of Defence, New Delhi — May 2025 - Jul 2025

- **Data Analysis and Report Generation Agent with MCP**
  - Built a dockerized SQL agent leveraging custom MCP servers for internal workflows.
  - Business Impact: Cut SME report prep time by 90% and enabled streaming responses.
- **YOLOv8 Object Detection and Tracking with automated selenium web-scraped datasets**
  - Trained YOLOv8 on 2000 images & 5 classes (selenium-scraped), achieving 94% on mAP@50 & 68% on mAP50-95.
  - Business Impact: Improved reliability of real-time usage by more than 50%.

PROJECTS

Krishi AI Sahayak - empowering farmers with smarter, data-driven agriculture

- Engineered pipeline with linear regression, result validation, market-aware recommendations, and in-memory assistant.
- Built, dockerized and deployed backend, **winning Smart SNU Hackathon** among 46 competing teams.
- Technology Stack: numpy, langchain, matplotlib, pandas

XLM-RoBERTa based Multi-Lingual sentiment analysis with auditory input support

- Developed FastAPI service that applies multi-modal analysis of input, with 5 supported languages scoring 73% accuracy.
- Architected STT service, trained transformers on text and CNNs on audio, and fused predictions via MLP.
- Technology Stack: PyTorch, scikit-learn, FastAPI, transformers

GPT-2 Style LLM creation and Meta LLaMA-3.2 8B LoRA fine-tuning for Data Science workloads

- Implemented GPT-2 (MHA, GELU) from scratch; KV-cache raised throughput 5× on RTX 4060.
- Fine-tuned Meta LLaMA-3 8B employing PEFT training using LoRA, and quantized it to Q5K-M as a GGUF.
- Technology Stack: PyTorch, TikToken, Unsloth, Transformers, PEFT, HuggingFace

Graph-RAG Text-to-SQL Agent for Enterprise Databases

- Engineered Text-to-SQL agent using LangGraph, tested over 43 tables and 62 relationships in a 1GB warehouse.
- Built graph-supervised, RAG-indexed pipeline to improve multi-join query generation by 33% on custom benchmarks.
- Technology Stack: Uvicorn, LangChain, LangGraph, Gemini, NetworkX, React.js

TECHNICAL SKILLS

Languages	Python, SQL, Rust, CUDA
AI & ML	PyTorch, Transformers, TensorRT, Optuna, vLLM, Triton, llama-cpp-python
Cloud & MLops	Docker, FastAPI, Kubernetes, Jenkins, MLflow, DVC, Comet-ML, GCP, AWS

ACHIEVEMENTS

- **Smart SNU Hackathon'25 Winner** Led a 6-member team to build and deploy the Krishi AI Sahayak app.
- **Semifinalist speaker at SMVIT Parliamentary Debate'24**
- **Dean's list: Monsoon'23**