

Ankan Das

🌐 ankandas.online | ✉ ankan06das@gmail.com | 📄 ankan06das | 🌐 ankan06das

ABOUT

Computer Science undergraduate and ML researcher interested in **large language models and mechanistic interpretability**. I am currently interested in working on **transformer circuits and representation learning**, and have hands-on experience building ML models and research systems from scratch.

I am currently seeking a **research internship** in the field of LLMs, mechanistic interpretability, or related areas of AI research.

EXPERIENCE

Summer Research Intern

Jun 2026 – Sept 2026

University of Calcutta — Supervisor: Dr. Soumya Sen

Kolkata, India

- Analyzed **NIFTY50** price dynamics using five **statistical dependence measures**—Pearson correlation, Graphical Lasso, distance correlation, tail dependence, and market-factor residual correlation—to construct and compare stock similarity graphs.
- Built **knowledge bases** for news-driven stock movement analysis by retrieving structurally similar stocks; evaluated similarity methods using **sector coherence (NMI/purity)**, **rolling-window Jaccard stability**, and **bootstrap robustness** to identify the most reliable method for different market-event scenarios.

Undergraduate Research Assistant

Aug 2024 – Present

Heritage Institute of Technology — Supervisor Dr. Diganta Sengupta

Kolkata, India

- First-authored **MedCompress**, a modular benchmark for **structured pruning, PTQ, and QAT** across ResNet, MobileNet, EfficientNet, and DenseNet on PathMNIST, OrganAMNIST, and OCTMNIST.
- Achieved up to $8\times$ **model size reduction** (ResNet50: **89.9 MB** $\rightarrow$ **11.6 MB**), with compressed models under **2.1M parameters** retaining **AUC** > 0.97 on pathology classification tasks.
- Applied **Attention U-Net** with **GAN-based augmentation** to BraTS MRI tumour segmentation and explored **SNN-based** visual recognition through temporal encoding and **ANN-SNN accuracy-efficiency** analysis.

PUBLICATIONS

MedCompress: A Modular Framework for Benchmarking Compression Techniques in Medical Imaging

- Ankan Das**, D. Sen, P. Bhattacharya, S. Sarkar, D. Sengupta. — *DABCon 2026* - Accepted for Oral Presentation

PROJECTS

LiteGPT | *Python, PyTorch* | [GitHub](#)

2026

- Implemented a **GPT-style language model from scratch** in PyTorch, including token embeddings, causal self-attention, transformer blocks, positional encoding, and autoregressive generation.
- Built configurable **training and evaluation pipelines** for language modeling, including learning-rate scheduling, label smoothing, validation loss, and perplexity tracking.

Crucible | *Next.js, Python, FastAPI, AI/ML* | [GitHub](#) | [Live Demo](#)

2026

- Built an **AI co-founder platform** that helps hackathon teams generate, refine, and stress-test project ideas against technical feasibility, impact, and constraints.
- Developed an interactive workflow for **AI-assisted ideation and project validation**, turning unstructured ideas into actionable technical plans.

Multimodal Knowledge Retrieval | *Python, LangChain, Qdrant, VLMs* | [GitHub](#)

2026

- Built a **multimodal document retrieval system** using Docling to parse PDFs, semantic chunking to structure content, and Qdrant for vector-based retrieval.
- Integrated **vision-language models** to extract semantic information from document images and tables, enabling retrieval across both textual and visual content.
- Implemented **query rewriting and conversational context** to improve retrieval for multi-turn questions over technical documents.

BPE Tokenizer | *Python, NumPy* | [GitHub](#) | [Live Demo](#)

2026

- Implemented **Byte Pair Encoding from scratch**, covering vocabulary construction, pair-frequency computation, merge learning, encoding, decoding, and special-token handling.
- Developed a reusable tokenizer for **language-model training**, with configurable vocabulary construction and corpus-level tokenization.

EDUCATION

Heritage Institute of Technology

Bachelor's in Computer Science and Engineering

- CGPA: 9.9 / 10

Kolkata, India

2023 – 2027

TECHNICAL SKILLS

Programming: Python • C++ • C • SQL • Java • Shell • L^AT_EX

ML/AI: PyTorch • TensorFlow • scikit-learn • Hugging Face • OpenCV • LangChain

Tools: Git • Docker • Linux • CUDA • MLflow • Weights & Biases • TensorBoard • Qdrant • Jupyter

COURSEWORK

Machine Learning • Deep Learning • Artificial Intelligence • Natural Language Processing • Large Language Models • Data Mining • Linear Algebra • Probability and Statistics • Optimization