ANKIT KUMAR

Baccalaureate in Technology (B.Tech.)
Computer Science and Engineering
Sershah Engineering College, Sasaram
New Delhi, India

+91-9430490287
ankits1802@gmail.com
Website
GitHub | LinkedIn
Medium | Leetcode

Education

Degree/Certificate	Institute/Board	CGPA/Percentage	Year
B.Tech. Major: CSE (Current)	Bihar Engineering University, India	9.56	2021-205
B.Tech. Major: CSE (Cumulative)	Bihar Engineering University, India	9.16	2021-2025
Diploma: Electronics Engineering	State Board of Technical Education, Bihar	9.27	2019-2022
AISSE: Matriculation (10^{th})	CBSE Board, New Delhi	97.60%	2019

Experience

- **Research Intern – Generative AI (Ragamala Imagery)** *May 2025 – June 2025*
  IIT Kharagpur, under Prof. Priyadarshi Patnaik via NPTEL On-site
  - **Fine-tuned Stable Diffusion XL (SDXL 1.0)** using **LoRA** and **QLoRA** adapters to generate culturally grounded Ragamala paintings, improving stylistic coherence by **31%** over base models.
  - **Designed few-shot and multi-shot RAG workflows** to enhance visual-textual alignment with Indian classical musical emotions and iconographic motifs.
  - **Benchmarked outputs** against **DALL·E 3**, **MidJourney v6**, and **Kandinsky 3.0** using **FID**, **CLIPScore**, and human evaluation, achieving **24% higher perceptual relevance**.
  - **Optimized training pipelines** on **AWS EC2 (g5.2xlarge)** and **SageMaker**, reducing fine-tuning time by **18%** via mixed-precision training and efficient data loaders.
  - **Curated a dataset of ~2,000 annotated Ragamala artworks** with poetic metadata, enabling cross-modal learning for aesthetic and symbolic fidelity.
  - **Applied classifier-free guidance and prompt engineering** to steer generation toward semantically rich, culturally contextualized outputs.
  - **Collaborated with digital humanities scholars and AI researchers** to ensure interpretability and ethical integrity in generative outputs.
  - **Used Tools/Frameworks: Python | PyTorch | Hugging Face Diffusers | FastAPI | AWS SageMaker | EC2 | LoRA | QLoRA | FAISS | CLIP | NumPy | PIL | Matplotlib | Weights & Biases | ONNX**

- **Machine Learning Internship** *Dec. 2024 – Jan. 2025*
  Internshala Trainings, IIT Madras Pravartak, and NSDC Remote
  - **Developed and deployed scalable predictive models** for real-world applications, including **California Housing Price Prediction**, **Telecom Customer Churn Prediction**, and **Early Disease Detection**, driving actionable insights and improving decision-making.
  - **Applied supervised and unsupervised learning techniques**, including **Linear Regression**, **Decision Trees**, **Random Forest**, **SVM**, **XGBoost**, and **Neural Networks**, to build, optimize, and validate models, enhancing predictive accuracy by **7%**.
  - **Engineered features and preprocessed data** using **cross-validation**, **hyperparameter tuning**, and **feature selection**, boosting model robustness and reducing overfitting by **7%**.
  - **Implemented Python scripts** for efficient **data extraction, analysis, and manipulation**, streamlining the **ETL pipeline** and improving data processing efficiency by **15%**.
  - **Enhanced model performance by 7%** through algorithm research and optimization using **SVM, ARIMA, PCA, and t-SNE**, increasing both predictive accuracy and interpretability.
  - **Leveraged cloud computing resources and MLOps tools** for scalable model deployment, enabling real-world implementation and optimizing workflow efficiency by **20%**.
  - **Addressed challenges** such as **dataset imbalance, overfitting, and missing data** using **SMOTE, regularization**, and **distributed computing** techniques, boosting model robustness and reliability.
  - **Used Tools/Frameworks: Python | Scikit-learn | TensorFlow | Pandas | NumPy | Jupyter Notebooks | MLOps | Cloud Platforms | Matplotlib | Seaborn | Statsmodels**

- **Amazon Machine Learning Summer School** *Jul. 2024 – Jul. 2024*
  Amazon Remote
  - Selected for Amazon's prestigious **Machine Learning Summer School** from a national applicant pool of over 1 million, with a selection rate below **0.275%**.
  - **Gained advanced exposure to ML theory and applications**, focusing on **Large Language Models**, **data preparation**, **feature engineering**, and **model evaluation**.
  - **Participated in hands-on sessions and industry projects** under mentorship from Amazon scientists, applying theoretical concepts to real-world ML problems.
  - **Used Tools/Frameworks: Python | Jupyter Notebooks | NumPy | Pandas | Scikit-learn | LLMs | Model Evaluation Techniques | Feature Engineering Tools**

- **Salesforce Virtual Internship** *Dec. 2023 - Jan. 2024*
  Virtual Internship through SmartInternz Remote
  - Engineered custom solutions using **Apex, Visualforce, and Lightning Web Components (LWC)** to address complex business needs.
  - Streamlined operations by implementing Salesforce Flow, Approval Processes, and Process Builder, enhancing workflow efficiency by 4%.
  - Developed **RESTful API integrations** for seamless data synchronization with external systems, **optimizing inventory management** accuracy by 3%.
  - Achieved Apex Specialist, Process Automation Specialist, and Developer Super Set Superbadges, showcasing advanced Salesforce expertise.
  - **Used Tools/Frameworks: Salesforce Lightning Platform | Apex | Visualforce | LWC | Salesforce CLI | VS Code**

- **Data Science Trainee** *Mar. 2023 – Apr. 2023*
  Internshala Trainings Remote
  - **Developed Python scripts** to extract, manipulate, and analyze structured and unstructured datasets for actionable insights.
  - **Researched and applied optimized ML algorithms**, achieving **7% reduction in runtime** through efficient model tuning.
  - **Practiced predictive modeling and supervised learning**, improving understanding of regression, classification, and model validation.
  - **Used Tools/Frameworks: Python | Pandas | NumPy | Scikit-learn | Matplotlib | Seaborn | Predictive Modeling | Data Cleaning | EDA | Jupyter Notebooks**

- **Embedded Systems & Robotics Intern** *Oct. 2021 – Nov. 2021*
  Ansoz Creations Pvt. Ltd. On-site
  - **Built Arduino-based embedded systems** using C and C++ to interface sensors and actuators with custom-built UI components.
  - **Reduced code size and improved performance** by optimizing memory usage and processing speed in embedded programs.
  - **Designed functional robotics prototypes** integrating sensor feedback and microcontroller logic for automation.
  - **Used Tools/Frameworks: Arduino IDE | Embedded C | C++ | Circuit Design | IR Sensors | Ultrasonic Sensors | Microcontrollers | Breadboard Prototyping**

Projects

- **AutoSQL: Text-to-SQL Query Generation** *Aug. 2024 – May 2025*
  Fine-tuned LLM with RAG for Complex SQL Queries over Large Databases GitHub
  - Fine-tuned a **6.7B parameter deepseek-instruct-coder model** using **LoRA** and **QLoRA** adapters, achieving a **23% accuracy boost** on complex SQL queries over the **Spider dataset** with **10,181 samples**.
  - Implemented **RAG with multi-hop retrieval**, combining **dense (FAISS)** and **sparse (BM25)** embeddings, leading to a **31% improvement in query precision**.
  - Applied **data augmentation** and **synthetic query generation** to increase the dataset size by **45%**, improving generalization.
  - Enhanced SQL accuracy by integrating a **self-refining loop**, reducing syntax errors by **19%** through iterative validation and correction.
  - Optimized a **32-layer architecture** with **2048-dim embeddings** and **10 attention heads**, improving inference latency by **15%** through **quantization and pruning**.
  - Deployed as a **FastAPI service** with an **interactive SQL editor**, featuring **syntax highlighting**, query history, and execution time visualization.
  - Reduced inference costs by **35%** using **ONNX quantization** and **TorchScript**, making the model real-time capable.
  - Executed **20 epochs of training** with **AdamW optimizer at 2e-5 learning rate**, reducing loss by **11%** through **dynamic learning rate scheduling**.
  - Incorporated **SQL validation and error correction**, increasing execution accuracy by **27%** on benchmark queries.
  - Applied **LLMOps with MLflow** for continuous evaluation and retraining, automating the model improvement lifecycle.
  - **Tools/Frameworks: Python | PyTorch | Transformers | deepseek-instruct-coder | LoRA | QLoRA | RAG | FAISS | BM25 | FastAPI | ONNX | TorchScript | SQL | NumPy | SentencePiece | MLflow**

- **TransLingua: English-To-French Machine Translation** *Apr. 2024 - Aug. 2024*
  End-to-End Machine Translation System with a Transformer Model GitHub
  - Designed and implemented a **Transformer-based** machine translation model to convert English text into French with high accuracy.
  - Processed and preprocessed **250K+ English-French sentence pairs** from the **Tatoeba dataset**, ensuring diverse linguistic coverage for training.
  - Utilized **HuggingFace BERT tokenizer** to refine text segmentation, improving tokenization accuracy and translation fluency.
  - Configured a **6-layer, 8-head attention Transformer** with **84M+ parameters** and **512-dimensional embeddings** for robust translation.
  - Implemented **beam search decoding**, **positional encoding**, and **masked self-attention**, significantly improving fluency and context retention.
  - Trained the model using **Adam optimizer** with **cross-entropy loss** and a **0.0001 learning rate**, achieving high BLEU score of 29.7 and METEOR score of 33 on validation data.
  - Optimized GPU acceleration with **CUDA**, reducing training time by **40%** and enabling real-time inference deployment.
  - Developed a **FastAPI** backend to serve the trained model via a RESTful API, ensuring efficient, low-latency text translation requests.

- Implemented **asynchronous processing** to handle concurrent translation requests, enhancing scalability and responsiveness.
- Integrated Firebase for **user authentication**, securing API endpoints and enforcing access control for registered users.
- Designed and built a **React (TypeScript)** frontend with a modern, chat-like UI for interactive text translation.
- Implemented real-time user feedback with loading indicators and error handling to improve usability and engagement.
- Integrated **protected routes** and session-based authentication using Firebase, ensuring secure access to translation services.
- Developed an **analytics dashboard** to visualize key metrics such as BLEU score trends, user engagement, and translation performance.
- Implemented logging and monitoring using **FastAPI middleware**, tracking API requests and optimizing performance bottlenecks.
- **Tech Stack: Python | PyTorch | Transformers | HuggingFace | BERT | CUDA | FastAPI | React (TypeScript) | Firebase | Pandas | NumPy | Scikit-Learn**

• **FineTune: LSTM-Based Piano Music Generator** *Jan. 2024 - Mar. 2024*
Innovative Music Generation with Artificial Intelligence GitHub

- Engineered a robust LSTM-based neural network model to autonomously compose melodious piano music.
- Processed and prepped the **ADL Piano MIDI dataset**, comprising 11,086 piano pieces, optimizing for sequence-based learning.
- Architected a 3-layer LSTM with 512 units per layer, employing **recurrent dropout** and **batch normalization** to enhance model generalization and stability.
- Implemented **efficient sequence generation** using 100-timestep inputs, facilitating the learning of temporal patterns.
- Trained the model on a vast corpus of 9,021 MIDI files for 200 epochs, leveraging **early stopping** and **model checkpointing**, achieving a loss reduction to 0.03.
- Developed a MIDI generation pipeline, producing high-quality, continuous piano compositions, showcasing the model's ability to emulate diverse musical styles.
- **Tools/Frameworks: Python | Keras | TensorFlow | Music21 | NumPy | h5py | RNNs | LSTMs**

• **Sign-a-Line: Real-Time American Sign Language Recognition** *December 2023*
Computer Vision for Empowerment GitHub

- Engineered a **real-time** Convolutional Neural Network (CNN) model to recognize and vocalize ASL hand signs.
- Preprocessed over 87,000 images from the **MNIST dataset**, normalizing pixel values and reshaping data for optimized input.
- Deployed a **custom CNN architecture** with 3 convolutional layers, ReLU activation, and max pooling, achieving 95% accuracy on test data.
- Integrated **MediaPipe** for accurate hand tracking, refining region of interest **(ROI) extraction**, reducing noise by 30%.
- Implemented Google Text-to-Speech (gTTS) for spoken output, enhancing user interactivity with **100% uptime** in audio playback.
- Optimized model inference time to 15ms per frame by leveraging TensorFlow, reducing latency in real-time prediction.
- **Used Tools/Frameworks: Python | TensorFlow | Keras | OpenCV | MediaPipe | gTTS | NumPy | Pandas**

Ongoing Projects/Research

• **Research Paper — Hallucination Mitigation in Large Language Models** *Ongoing*
Benchmarking, Refinement, and RAG for Reducing Hallucinations in LLMs

- Conducted an extensive **survey and analysis** of hallucination phenomena in LLMs, identifying **5 core patterns** and key challenges across **diverse NLP tasks**.
- Benchmarked **6 state-of-the-art LLMs** (ChatGPT, LLaMA, Claude, Mistral, Mixtral, Gemini) on **TruthfulQA and BIG-bench**, achieving a **7.4% hallucination reduction** through **ensemble modeling** and **multi-hop RAG retrieval**.
- Optimized LLMs using **Chain-of-Thought (CoT)**, **self-consistency**, and **iterative refinement**, reducing factual error rates by **12.8%**.
- Implemented a **retrieval-augmented verification (RAV)** step, boosting **factual accuracy by 9%** through external knowledge validation and correction loops.
- Enhanced RAG with **hybrid retrieval (FAISS + BM25)** and multi-hop lookups, improving query precision by **11%**.
- Applied **fine-tuning with LoRA and QLoRA** adapters on a synthetic fact-checking dataset, decreasing hallucination-induced inconsistencies by **15%**.
- Integrated **ONNX quantization and TorchScript**, reducing inference latency by **22%**, making the system real-time capable.
- Deployed the solution as a **FastAPI service** with an interactive interface for generating and verifying factual responses, featuring **confidence scores, syntax validation**, and **contextual error analysis**.
- Leveraged **MLflow and LLMOps pipelines** for continuous evaluation, retraining, and performance monitoring, ensuring **scalability and stability**.
- **Tools/Frameworks: Python | PyTorch | Transformers | LoRA | QLoRA | RAG | FAISS | BM25 | FastAPI | ONNX | TorchScript | TruthfulQA | BIG-bench | MLflow | LLMOps**

Skills

• **Programming Languages**: Python, C, C++, JavaScript, SQL, Bash, Java, Kotlin, MATLAB, HTML/CSS

• **AI & ML Frameworks**: PyTorch, TensorFlow, Keras, Scikit-learn, Hugging Face, ONNX, LangChain, LangSmith, LlamaIndex, AutoGen, Semantic Kernel, MLOps, CI/CD Pipeline

• **Document & Image Processing**: OpenCV, PyPDF, PyOCR, Tesseract, Pillow

• **Vector Search & RAG Ecosystem**: FAISS, Chroma DB, Pinecone, Qdrant, Milvus, Azure AI Search, RAG Pipelines, Embedding Techniques, Prompt Engineering, Agentic AI

• **Databases & Backend**: MySQL, PostgreSQL, Firebase, Cosmos DB, FastAPI, Flask, REST APIs, GraphQL

- **Cloud & DevOps**: MS Azure (AI Studio, AI Search, Cosmos DB, ML), AWS (EC2, S3, Lambda), Docker, Nginx, Gunicorn, Git/GitHub, DeepStream
- **Data Engineering & Automation**: ETL Pipelines, Data Preprocessing, Multimodal Workflow Automation
- **Visualization & Analysis**: Power BI, Tableau, Jupyter, Matplotlib, Seaborn
- **Operating Systems**: Windows, Linux (Ubuntu, Arch, Debian)
- **Soft Skills**: Product Management, Financial Analysis, Consulting, Technical Writing, Team Leadership

Key Courses Taken/Certifications

- **Computer Science and Engineering**: Data Structures and Algorithms, Object-Oriented Programming, Operating Systems, Computer Networks, Database Management Systems, Software Engineering, Design and Analysis of Algorithms, Compiler Design, Distributed Systems, System Design, Web Technologies, Mobile Application Development, Cloud Computing
- **Mathematics and Theoretical Foundations**: Linear Algebra, Calculus & Optimization, Discrete Mathematics, Probability and Statistics, Numerical Methods, Graph Theory
- **Artificial Intelligence and Machine Learning**: Machine Learning, Deep Learning and Neural Networks, Artificial Intelligence, Computer Vision, Natural Language Processing
- **Electronics and Embedded Systems**: Electronic Devices and Circuits, Digital Electronics, Analog Electronics, Network Analysis and Synthesis, Microprocessors and Microcontroller Applications, Embedded Systems, Control Systems
- **Other Electrical and Communication Topics**: Communication Systems, Power Electronics, Electrical Machines, Measurement and Instrumentation
- **Google Data Analytics Specialization**: Completed 8-Unit course offered by Coursera with perfect score.

Positions of Responsibility

- **Startup Cell & Placement Cell Co-ordinator**, Startup Cell, Sershah Engineering College *Jan. 2024 - Present*
  - Conducted SME talks and co-ordinated placement drives benefitting over 500 collegiate students.

Achievements

- **Departmental Rank 1**: Secured and retained the top departmental rank across Bihar Engineering University through stellar and consistent academic performance. Associated with Sershah Engineering College. Jun 2025
- **Elite Leetcoder**: Solved over **1300 problems on Leetcode**, achieving an all-time global rank under 8,000 among 50 million+ users. Leetcode Account Jun 2025
- **Amazon ML Summer School 2024 Cohort**: Selected among 3000 students nationwide (<0.275% acceptance rate) for Amazon's 4-week ML Summer School program. Certificate Jul 2024
- **NPTEL Topper & Research Internship Offer**: Topped the Soft Skills MOOC (Jan–Jun 2024) on NPTEL, earning a research internship under Prof. Priyadarshi Patnaik at IIT Kharagpur. Jun 2024
- **Double Gold Medallist**: Awarded by the State Board of Technical Education, Bihar for scoring the highest cumulative GPA in the state (2019–2022). Associated with Government Polytechnic, Gaya. Aug 2022

Extracurricular

- **GoGreen**, Led Afforestration Drive In College Campus, planting 200 + saplings. *2022 - Present*
- **Student Editor**, Acted as Student Editor for the college magazine, with an enhanced throughput. *2023*