

KUSHAGRA GUPTA

📞 +91 9269972395 📩 kggupta.work@gmail.com 💬 [LinkedIn](#) 🌐 [Github](#)

Education

JK Lakshmiपत University, Jaipur

B.Tech in Computer Science & Engineering

Expected 2026

7.6/10 CGPA

IIIT, Hyderabad

Semester Exchange Student — B.Tech in Computer Science & Engineering

Monsoon 2025

7/10 CGPA

IIT, Gandhinagar

Semester Exchange Student — B.Tech in Computer Science & Engineering

December 2023

Technical Skills

Languages: JavaScript (ES6+), Python, C++, SQL

Core Competencies: Data Structures & Algorithms, Object-Oriented Programming (OOP), Analytical Problem-Solving

Frontend: React.js, HTML5, CSS3, Figma

Backend: Node.js, Express.js, REST APIs

Databases: MongoDB, Elasticsearch, Firebase

Artificial Intelligence & Machine Learning: Deep Learning, Retrieval-Augmented Generation (RAG), Langchain, Computer Vision (CV), Vector Databases, Statistical Analysis, PyTorch, Scikit-learn, NumPy, Pandas, Matplotlib

Tools & Technologies: Git, Docker, GitHub, Postman, Cypress.io, Prometheus, Grafana, LaTeX, Firebase, FAISS, Google Cloud

Work Experience

Full Stack Development Intern

May 2025 – July 2025

Dobby Ads

Remote

- Built & deployed two **MERN-based** web applications with **MongoDB** for storage, designing scalable architecture for client workflows.
- Automated 30+ Cypress test cases, reducing post-release bugs by 25% and regression testing time by 40%.
- Implemented Prometheus + Grafana dashboards for real-time API monitoring, improving incident resolution time by 20%.

Android Development Intern

May 2024 – July 2024

Vedic Bodhi Pvt. Ltd.

Remote

- Developed secure user authentication with Firebase and integrated REST APIs for real-time updates.
- Enhanced app performance and user experience through optimized API calls and responsive UI design.

Projects

Legal Redline Sandbox (GenAI Exchange Program Hackathon)

[Source Code](#)

Python, FastAPI, React.js, Google CloudSQL, Google Gemini API, Document AI (OCR), JWT, Docker

- Advanced to the Top 20 teams for our problem statement track in a global, multi-stage GenAI hackathon.
- Spearheaded the project's entire Google Cloud infrastructure, managing the complete lifecycle from initial setup and configuration to full-stack integration.
- Engineered the complete data persistence layer using Google CloudSQL; designed a relational database schema from scratch to manage users, chat sessions, messages, and clause rewrites.
- Developed and integrated the database across the full stack to enable persistent user accounts, session management, and chat history.
- Implemented an asynchronous notification system to update users on the real-time status of their contract analysis.

Subject Matter Expert (SME) AI Agent

[Source Code](#)

Python, FastAPI, RAG, Langchain, Elasticsearch, FAISS, Transformers, Docker

- Built a production-grade RAG system with 95%+ retrieval consistency, integrating dense + sparse search (Elasticsearch, FAISS, BM25).
- Engineered a document ingestion pipeline (PDF/DOCX/HTML) with hierarchical chunking, indexing 50k+ chunks into Elasticsearch.
- Implemented cross-encoder reranking, cutting irrelevant retrievals by 40% and boosting top-3 relevance by 25%.
- Developed an agentic workflow engine using langchain with tool-use (PDF/DOCX generation, email automation) for multi-step query handling.

- Designed a FastAPI-based server with async pipelines, achieving \sim 800ms end-to-end retrieval latency.
- Containerized services (API, embedding worker, Elasticsearch, FAISS) with Docker for seamless deployment.

Byte-Latent Transformer (BLT) vs Traditional Tokenization

[Source Code](#)

PyTorch, Natural Language Processing

- Developed and compared two sequence-to-sequence Transformer models in PyTorch: a novel Byte-Latent Transformer (BLT) utilizing custom tokenization, against a standard character-level baseline, to solve a string reversal task.
- Quantified the trade-off between computational efficiency and model accuracy: Achieved a 91% reduction in average input sequence length (from 70.9 characters down to 6.4 patches) using the BLT's entropy-based patching and hash n-gram embedding techniques, demonstrating its potential efficiency benefits.
- Analyzed model performance through rigorous evaluation, finding the baseline model reached 97.8% token accuracy, whereas the BLT plateaued at 8.4% accuracy; diagnosed this performance gap by examining training curves and identifying potential limitations including pre-processing overhead and model capacity constraints for this specific task.

Multilingual Language Model (English–Hindi–Bengali)

[Source Code](#)

Python, PyTorch, Transformers, SentencePiece tokenizer (Unigram), LoRA fine-tuning, multilingual data pipelines

- Trained an 18.5M parameter transformer on 400M tokens using PyTorch.
- Built custom data pipelines for cleaning, deduplication, and segmentation, reducing noise by 2–3%.
- Implemented SentencePiece tokenizer (50k vocab) achieving <0.0002% OOV rate.
- Fine-tuned Gemma-270M with LoRA on 40k-task dataset; analyzed limitations under compute constraints (Perplexity: 2857, 0% exact match accuracy).

Netflix Clone

[Source Code](#)

React.js

- Designed a scalable frontend architecture with reusable React components and managed state effectively.
- Built a fully responsive UI, optimizing for cross-device compatibility and performance.

Automatic License Plate Recognition (ALPR) — Custom CNN + EasyOCR

[Source Code](#)

PyTorch, EasyOCR, OpenCV, NumPy, pandas, Matplotlib

- Engineered an end-to-end Automatic License Plate Recognition (ALPR) pipeline as the vision module for a conceptual automated, distance-based tolling system; integrated detection and OCR into a deployable inference script.
- Designed and trained a custom 5-layer CNN (52.1M parameters) for bounding-box regression, achieving 71.48% mean IoU on validation; trained on a curated dataset of 9,600 images (3.5 GB) aggregated from 5 public sources.
- Built a robust preprocessing & data pipeline: aspect-ratio preserving resize to 416×416 , bounding box normalization, data cleaning, and an 80/10/10 train/val/test split to ensure reproducible evaluation.
- Increased inference stability using Test-Time Augmentation (TTA) (horizontal flips, brightness adjustments) and integrated detections with EasyOCR to form a complete detection→recognition flow; packaged code, notebooks and model weights for reproducibility.

Comparative Analysis and Sustainable Solutions for E-Waste Management

[Paper](#)

Python and Statistical Methods

- Analyzed 15+ datasets on e-waste from emerging and developed economies using Python and statistical methods.
- Proposed scalable, data-driven waste reduction strategies in a comprehensive academic report.

Enhanced Malware Detection using AI

[Source Code](#)

AI, Machine Learning, Wireshark, and Docker

- Engineered a machine learning-based malware classifier achieving over 90% detection accuracy in lab tests.
- Utilized Wireshark and Docker to analyze network packets for malicious patterns and built predictive models for early threat identification, demonstrating skills relevant to processing large data streams.

Spardha - College Sports Fest Website

[Source Code](#)

HTML, CSS, and JavaScript

- Collaborated with a team of 3 to develop the official Spardha website. Built responsive UI with 5+ interactive features; deployed via Vercel on the college's domain with optimized load time under 2s.

Relevant Courses

Algorithms, Object-Oriented Programming, Operating Systems, AI, DBMS, Computer Networks, Language Model and Agents, Information Retrieval and Extraction, Digital Image Processing

Leadership and Extracurriculars

- **Hackathon Participant:** Competed in HackJKLU 4.0, 3.0, and 2.0, developing solutions involving machine learning, web development, and game development.
- **Volunteer Experience:** Volunteered in Sabrang Cultural Fest, managing logistics for awards and certificates for 150+ volunteers. Smart India Hackathon (SIH) - Around 20 volunteers to organize logistics for 70+ attendees. HackJKLU - Around 100 volunteers to organize logistics for 180+ attendees

Certificates and Achievements

Google Cloud Generative AI Skill Badges

Apr 2025 – Jun 2025

Google

- Completed a comprehensive series of hands-on labs focused on the Google Cloud GenAI stack, validating expertise in building and deploying AI-powered applications.
- **Badges Include:** Explore GenAI with Vertex AI Gemini API, Inspect Rich Documents with Gemini Multimodality and RAG, Build Real World AI Applications with Gemini and Imagen, Develop GenAI Apps with Gemini and Streamlit, Prompt Design in Vertex AI.
- Demonstrated proficiency in prompt engineering, multimodal AI, and building full-stack GenAI applications using Vertex AI, Gemini, Imagen, Streamlit, and Cloud Run.

Artificial Intelligence Fundamentals

Sep 2025

IBM

- Acquired a foundational understanding of AI concepts, machine learning principles, and their real-world applications.

Gen AI Academy

Jun 2025

Google Cloud Skills Boost

- Completed foundational training on Generative AI, covering core concepts, models, and use cases within the Google Cloud ecosystem.

Data Structures & Algorithms in Java

July 2022 – July 2023

1st Alpha Batch

Apna College

- Solved 200+ Data Structures and Algorithms problems on various platforms, demonstrating a strong grasp of algorithmic problem-solving.