Abhinav Singh Chauhan

Machine Learning Engineer

abhinavschauhan14@gmail.com
https://www.linkedin.com/in/abhinavsingh9714/
https://abhinavsingh9714.github.io/
https://github.com/abhinavsingh9714
2407916598
College Park, MD

SUMMARY

Machine Learning Engineer with 4+ years of experience building production-grade GenAI and applied ML systems end-to-end across startups and research. Expertise in Agentic AI, finetuning LLM and building multi-agent workflows, RAG and vector databases. Strong in taking projects from data pipelines and model training (PyTorch, TensorFlow, Spark) to low-latency APIs and cloud deployment (FastAPI, Docker, AWS).

WORK EXPERIENCE

Founding ML Engineer Aug 2025 - Present

TAAI Labs

- Spearheaded development and launch of **Neuron**, a multi-tenant **ML/GenAI platform** on AWS that powers knowledge-base ingestion, infrastructure graphing, and an enterprise Q&A assistant across all company products.
- Architected an **end-to-end knowledge ingestion pipeline** that chunks, embeds, and indexes 1000+ internal documents using DynamoDB, Textract, Lambda, Step Functions, and Pinecone, enabling RAG search and Q&A across the product suite.
- Designed and shipped an **infrastructure graphing service** that converts Terraform scripts into data-flow and dependency graphs, **cutting architecture and security review time by 90**%.
- Fine-tuned and integrated a Qwen-based NIST-compliant **Q&A model with the RAG pipeline** to deliver grounded, audit-ready responses for security and compliance workflows.

Machine Learning Intern Jun 2025 – Aug 2025

EdTech Tulna - IIT Delhi

- Designed and deployed a **fault-tolerant multi-agent AI system** with LangGraph to evaluate educational videos on **20+ learning indicators**, cutting manual review time by **95%** and delivering evidence-backed structured docx reports.
- Engineered **2 LLM critic models** with structured chain-of-thought prompting and a **conflict resolution agent** with advanced prompt chaining, boosting rating consistency and accuracy, achieving **over 90% alignment with expert human evaluations.**

Founding Machine Learning Engineer Jan 2024 - Aug 2024

KradleJoy Pvt Ltd

- Led end-to-end development of a **real-time multimodal baby monitoring system** by integrating live video and audio into a FastAPI-based ML inference API on AWS EC2, achieving **sub-2s latency** and delivering synchronized safety alerts with continuous activity insights.
- Built a **multimodal spatial understanding module** using a fine-tuned YOLOv8 model for baby detection, sequential pose analysis for sleep tracking under **80–90% occlusion**, and a YAMNet audio model for real-time cry detection.

Machine Learning Engineer (Freelance) Jul 2023 - Dec 2023

Self-employed

- Developed a **real-time visual similarity engine** using CLIP + FAISS to retrieve lookalike images from a **60K+ image library in <50 ms**, exposing the capability as an API backend for an interactive visual search UI.
- Architected a **low-latency sentiment analysis pipeline** for a social media analytics dashboard by scaling tweet preprocessing with PySpark and fine-tuning a Hugging Face BERT model, achieving **85%+ accuracy** with **sub-second inference** via FastAPI deployment.

Software Engineer Oct 2019 - Jul 2021

Rapid Global School

- Engineered a **modular RESTful API** with **Java Spring Boot**, integrating multiple workflows for school management system over a unified PostgreSQL-backed system for **2000+ users**.

Software Engineer Jan 2019 - Sep 2019

Exicom Tele-Systems Limited

- Collaborated with a cross-functional team to build an **embedded OCPP-based communication system** connecting EV chargers with a central server, delivering reliable server–client communication for **25,000+ users**.

PROJECTS

- **Intelligent Jira Ticket Generator** Jun 2025 - Jul 2025
  Devised and deployed an LLM planning agent that converts product descriptions into structured Jira tickets with real-time Jira Cloud integration, schema-enforced Pydantic output, and Langchain, reducing planning overhead by 70%.
- **Slo-Mo Video Generation with Generative Adversarial Networks** Apr 2025 - Jun 2025
  Developed a U-Net-based frame interpolation model to generate 480fps slow-motion videos from 60fps inputs by inserting 2–8 intermediate frames in 0.9s clips, enabling seamless playback of unpredictable motion events.
- **Multi-Time Series Risk Prediction** Jan 2025 - Apr 2025
  Built a deep learning system to forecast 15-day volatility across 30 financial time series by combining entity embeddings with an attention-enhanced Temporal Convolutional Network, resulting in a 14% F1 score improvement over LSTM baselines.

EDUCATION

- **University of Maryland, College Park, MD** Aug 2024 - May 2026
  Master of Science in Data Science GPA: 3.89
  Relevant Courses: Advanced Machine Learning, Big Data Systems, Natural Language Processing, Algorithms for Data Science
- **Indian Institute of Science, Bangalore, India** Apr 2023 - May 2024
  Postgraduate Certification in Computational Data Science Grade: A
  Relevant Courses: Deep Learning, MLOps, Data Engineering, Big data Analytics, Probability and Statistics
- **Thapar University, Patiala, India** Jul 2015 - Jul 2019
  Bachelor of Engineering in Electronics and Computer Engineering GPA: 8.13

SKILLS

- **Languages:** Python, R, Matlab, Java, C++, C
- **ML & AI:** Machine Learning, Deep Learning, Generative AI (LLMs, Diffusion Models), Multimodal Learning, Computer Vision, NLP, Time Series Forecasting, Reinforcement Learning
- **Frameworks & Libraries:** PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face, LangChain, LangGraph, XGBoost, pandas, NumPy, Matplotlib, Plotly, FAISS, Chroma, CUDA
- **Data & Cloud:** SQL (PostgreSQL, MySQL), NoSQL (MongoDB, DynamoDB, Pinecone, Neo4j), Apache Spark, AWS (EC2, S3, Lambda, Step Functions, Bedrock)
- **Backend, MLOps & Tools:** FastAPI, Spring Boot, Docker, Git, CI/CD, Linux
- **Software Development:** Agile, System Design, Operating Systems, Software development life cycle