

NIRBHAY SINGH

Jamshedpur, Jharkhand, India | nirbhayksingh14@gmail.com | +91 8210481293

LinkedIn | GitHub | Portfolio

Professional Summary

Software Engineer and AI/ML Developer with hands-on experience building production-grade Generative AI systems, including a custom 6.5M-parameter LLM, RAG pipelines, and real-time NLP agents. Skilled in system design, scalable backend architecture, and cloud deployment (AWS, Docker). Internship experience at TATA BlueScope Steel delivering FinBERT-powered financial intelligence tools adopted in production. B.Tech Computer Science graduate (May 2026) seeking Software Development Engineer or AI/ML Engineer roles.

Experience

Software Engineer | Anantron International Pvt. Ltd

Feb 2026 – Present

Jamshedpur, India

- Develop and maintain production REST APIs using Node.js, Express.js, FastAPI, and React.js; integrated AI-powered endpoints supporting LLM-based workflows.
- Designed and architected scalable system designs for production services; deployed containerized applications on AWS (EC2, S3) using Docker, ensuring high availability and streamlined CI/CD pipelines.
- Improved frontend performance by 30% through lazy loading, optimized API communication, and component-level caching while following Agile development practices.

AI/ML Intern | TATA BlueScope Steel Pvt. Ltd

May 2025 – Jul 2025

Jamshedpur, India

- Built an end-to-end financial intelligence system ingesting live stock prices and Google News RSS feeds, running FinBERT sentiment classification across 20+ stocks at ~82% accuracy, processing 500+ articles daily.
- Stored news embeddings in ChromaDB and surfaced insights via a conversational Groq LLM agent, reducing manual analyst effort by 60%.
- Delivered an interactive Plotly + Streamlit dashboard for real-time sentiment trend visualization, adopted by the finance team for daily market monitoring.

Projects

Rudra — Custom-Trained Autonomous AI Agent & MiniGPT

[GitHub]

Python, PyTorch, FastAPI, Next.js, TypeScript

- Architected "Rudra," a custom 6.5M-param Transformer LLM with ReAct agent framework for autonomous tool use (web search, math), trained on self-generated data and deployed via FastAPI + Next.js with optimized inference on a 4GB GPU and Trained the Data on GTX 1650.

Movie Recommendation System

[GitHub]

Python, Scikit-learn, FastAPI, Streamlit, NLTK, Pandas

- Engineered an end-to-end recommendation system, from an ETL pipeline (lemmatization, TF-IDF on raw metadata) to a FastAPI REST API with async poster fetching, to a Streamlit UI with real-time similarity-based movie suggestions.

FinSight AI — RAG-Powered Financial Analyst

[GitHub]

Python, FinBERT, Groq API, ChromaDB, Streamlit, Plotly

- Engineered an end-to-end financial analysis system ingesting live stock prices and Google News RSS feeds, running FinBERT sentiment classification, storing embeddings in ChromaDB, and surfacing insights through a conversational LLM agent and interactive Plotly dashboard.

Technical Skills

AI / GenAI	LLM Integration, RAG Pipelines, ReAct Agent Framework, Prompt Engineering, Vector Embeddings, Semantic Search
Deep Learning	PyTorch, Transformer (Decoder-only), Multi-Head Attention, Model Training & Evaluation
ML / NLP	Scikit-Learn, HuggingFace Transformers, FinBERT, TF-IDF, NLTK, NumPy, Pandas
Frameworks	FastAPI, LangChain, Streamlit, Next.js, React.js, Node.js
Vector DBs	FAISS, ChromaDB
Cloud/MLOps	AWS (EC2, S3), Docker, CI/CD, Git
Databases	PostgreSQL, MongoDB
CS Fundamentals	Data Structures, Algorithms, System Design, OOP, REST API Design
Languages	Python, JavaScript, TypeScript, SQL, Java

Education

B.Tech in Computer Science and Engineering | CGPA: 7.7/10

Nov 2022 – May 2026

RVS College of Engineering and Technology, Jamshedpur

Certifications

- Oracle Generative AI Professional Certification – Oracle
- Oracle Cloud Infrastructure (OCI) Foundations Associate – Oracle
- Fundamentals of Large Language Models – HuggingFace
- Introduction to Generative AI – Google Cloud
- TensorFlow Deep Learning Fundamentals – DeepLearning.AI
- How Large Language Models Work – IBM
- Natural Language Processing and Computer Vision – IBM