

# Suraj Kumar Jha

Pune, India | +91 8849083319 | surajjha21103@gmail.com  
LinkedIn | GitHub | Hugging Face

## TECHNICAL SKILLS

**Languages & Systems:** C/C++, Python, SQL, Linux/Unix, GPU Computing

**AI Engineering:** PyTorch, CUDA, Triton, Transformers, RAG Systems, LLM Fine-tuning (QLoRA), Quantization

**Infra & Tools:** FastAPI, Django, Docker, Elasticsearch, Milvus, FAISS, Git

## EXPERIENCE

### Data Scientist

Oct 2025 – Present

Solytics Partners – Nimbus AI Platform

Pune, India

- Engineered backend services in **FastAPI** and **Django** handling high-throughput analytical workloads on distributed Unix servers.
- Built hybrid **information retrieval pipelines** (Elasticsearch + VectorDB) reducing semantic query latency for LLM applications.
- Implemented **LLM guardrails**, evaluation workflows, and deployment pipelines for fine-tuned SLMs in production environments.
- Designed tooling to improve reliability, observability, and traceability across LLM-powered data pipelines.

### Founding AI Intern

May 2024 – Sep 2024

Gov Solutions

India

- Built a **computer vision system** for Indian Railways to detect structural defects in railway tracks using deep learning models.
- Worked on dataset curation, annotation pipelines, and model training for real-time defect identification from track imagery.
- Implemented edge-optimized inference pipelines for field deployment under hardware and latency constraints.

## PROJECTS

### TinySeek – High-Performance Transformer

2025

Systems Optimization

GitHub

- Architected a compact Transformer model in PyTorch following DeepSeek design principles.
- Wrote custom **Triton GPU kernels** for quantized attention, improving memory efficiency and inference speed.
- Implemented RoPE, multi-head latent attention, and optimized tensor memory layouts.

### CUDA Documentation Search Engine

2024

Retrieval Systems

GitHub

- Built a large-scale IR system achieving 87% relevance on NVIDIA technical queries.
- Designed hybrid sparse+dense retrieval (BM25 + Milvus) for high-accuracy semantic search.
- Integrated LLaMA 2 for grounded response generation over 1,000+ indexed documents.

### Transformer Translation System

2025

Algorithm Implementation

GitHub

- Reimplemented “Attention Is All You Need” achieving 93% translation accuracy.
- Optimized multi-head attention and encoder-decoder training efficiency.

### Multi-Agent Log Intelligence System

2025

AI Automation

GitHub

- Built a multi-agent system automating distributed server log analysis, reducing debugging effort by 60%.
- Applied structured reasoning workflows for anomaly detection and summarization.

## ACHIEVEMENTS

1st Place – Inerve Hackathon (AI Tax Analysis)

1st Place – MachLearn Competition (ML Algorithms)

3rd Place – Startup Saga (EdTech Innovation)

1st Place – GDSC ML Hackathon (Crop Disease Classifier)

Leadership – Joint Secretary, AI Titans Club; mentored 50+ students

## EDUCATION

### B.E. Electronics & Telecommunication

2022 – 2026

Army Institute of Technology

Pune, India