

# 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

<b>Data Scientist</b> <i>Solytics Partners – Nimbus AI Platform</i>	Oct 2025 – Present Pune, India
<ul style="list-style-type: none"><li>Engineered backend services in <b>FastAPI</b> and <b>Django</b> handling high-throughput analytical workloads on distributed Unix servers.</li><li>Built hybrid <b>information retrieval pipelines</b> (Elasticsearch + VectorDB) reducing semantic query latency for LLM applications.</li><li>Implemented <b>LLM guardrails</b>, evaluation workflows, and deployment pipelines for fine-tuned SLMs in production environments.</li><li>Designed tooling to improve reliability, observability, and traceability across LLM-powered data pipelines.</li></ul>	
<b>Founding AI Intern</b> <i>Gov Solutions</i>	May 2024 – Sep 2024 India
<ul style="list-style-type: none"><li>Built a <b>computer vision system</b> for Indian Railways to detect structural defects in railway tracks using deep learning models.</li><li>Worked on dataset curation, annotation pipelines, and model training for real-time defect identification from track imagery.</li><li>Implemented edge-optimized inference pipelines for field deployment under hardware and latency constraints.</li></ul>	

## PROJECTS

<b>TinySeek – High-Performance Transformer</b> <i>Systems Optimization</i>	2025 GitHub
<ul style="list-style-type: none"><li>Architected a compact Transformer model in PyTorch following DeepSeek design principles.</li><li>Wrote custom <b>Triton GPU kernels</b> for quantized attention, improving memory efficiency and inference speed.</li><li>Implemented RoPE, multi-head latent attention, and optimized tensor memory layouts.</li></ul>	
<b>CUDA Documentation Search Engine</b> <i>Retrieval Systems</i>	2024 GitHub
<ul style="list-style-type: none"><li>Built a large-scale IR system achieving 87% relevance on NVIDIA technical queries.</li><li>Designed hybrid sparse+dense retrieval (BM25 + Milvus) for high-accuracy semantic search.</li><li>Integrated LLaMA 2 for grounded response generation over 1,000+ indexed documents.</li></ul>	
<b>Transformer Translation System</b> <i>Algorithm Implementation</i>	2025 GitHub
<ul style="list-style-type: none"><li>Reimplemented "Attention Is All You Need" achieving 93% translation accuracy.</li><li>Optimized multi-head attention and encoder-decoder training efficiency.</li></ul>	
<b>Multi-Agent Log Intelligence System</b> <i>AI Automation</i>	2025 GitHub
<ul style="list-style-type: none"><li>Built a multi-agent system automating distributed server log analysis, reducing debugging effort by 60%.</li><li>Applied structured reasoning workflows for anomaly detection and summarization.</li></ul>	

## ACHIEVEMENTS

1st Place – Innerve 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>B.E. Electronics &amp; Telecommunication</b> <i>Army Institute of Technology</i>	2022 – 2026 Pune, India
--	----------------------------