

Aditya Raj

B.Tech, Electronics and Communication Engineering
National Institute of Technology, Patna
Expected Graduation: 2026

✉ adityar.ug22.ec@nitp.ac.in
☎ +91-8797073498
🌐 [linkedin.com/in/hexronus](https://www.linkedin.com/in/hexronus)
🐙 github.com/hexronuspi
🌐 hexronuspi.github.io

Summary

A **speed and quality** focused engineer specializing in the design of **high-performance, low-latency systems** for ML and data. Designed and built a high-throughput data processing tool during a quant internship, achieving processing speeds of **<9ms** for over **100k** data points and individual operation **latencies under 80ms**.

Education

Degree/Certificate	Institute/Board	GPA/Percentage	Year
B.Tech, Electronics and Communication Engineering	NIT, Patna	8.1 / 10.0	2022–2026

Achievements

Stanford AI Lab	Interviewed at Prof. Jiajun Wu's Lab for research position in computer vision	2025
NK Securities Research	Ranked 67th/2095 in IV Prediction ; MSE = 1.3e-5	2025
M2L Summer School	Selected from 1600+ global applicants (BTech to Industry); Split, Croatia	2025
Amazon ML Challenge	Ranked 184th/75,000+ ; F1 score = 0.4667 using fine-tuned moondream	2024
IIT ISM AI of GOD	Winner ; WER = 0.116 using TrOCR + T5 ; post-processing algorithm	2024
RMO	Top 0.4% nationwide mathematical olympiad(classes 8–12) in India , out of 250k students	2019

Research

- **Knowledge Graph-Informed Query Decomposition(KG-IQD): Hybrid KG-RAG Reasoning in Noisy Contexts**
Authors: **Aditya Raj, Dr. Kuldeep Kurte^{PI}** | Poster (ISWC 2025) - Rejected

Experience

- **AI Research Intern** India
QFI Research Capital May 2025 – October 2025
 - Worked with a team of 2 engineers to build **data pipelines** and **forecasting models** predicting **product timelines** and assessing **market impact**, integrated with a **real-time sentiment engine** for long-term **alpha capture**.
 - **Resolved critical computation bugs**, built and deployed an **internal toolkit** to manage **workflow**.
- **Research Intern** Hyderabad, India
IIIT-H / Dr. Kuldeep Kurte, Spatial Informatics Lab Apr 2025 – July 2025
 - Achieved **state-of-the-art results** on a **custom disaster QA benchmark**, outperforming **RQ-RAG** by **14%** and **KG** by **18%**, by developing a **neuro-symbolic framework** that **guides query decomposition** using **Knowledge Graphs** and **interrelates points with RAG** on sub-queries for **QA** over structured and unstructured data.

Projects

- **Efficient LLMs via Switchable and Dynamic Quantization** docs | October 2025
Tools: *PyTorch, Hugging Face, LoRA, QAT-LLM, SQuAD Dataset*
 - **Switchable and dynamic quantization - GPT-2**, per-layer bit-width control (INT8-FP32); adaptive **LoRA**.
 - Trained on **SQuAD** using **cyclic precision training** and joint bit-width optimization, achieving stable accuracy across dynamic precision configurations and demonstrating **quantized inference**.
 - Evaluated the **robustness** under random precision switching, aligning insights with **CPT (ICLR'21)** and **Double-Win Quant (ICML'21)** and found it perfectly aligned.
- **Optimizing and Quantizing FBNet Models for Edge Deployment** docs | October 2025
Tools: *Hugging Face, PyTorch, Edge Device*
 - Converted **FBNet-A/B** from **PyTorch** → **TensorFlow** → **TFLite (FP32, FP16, INT8)** for edge deployment, achieving **MSE 1e-19, 100% accuracy**, and up to **4× model size reduction**.
 - Rebuilt architectures in **Keras+TF**, used a **parser** for weight transfer, and verified using **MSE and accuracy metrics**.
 - Implemented **TFLite GPU batch resizing**, improving **conversion stability** and **edge-device performance**.
- **Serverless Web Platform** webpage | Dec 2024 – Feb 2025
Tools: *Next.js, Vercel, Supabase*
 - Replaced a \$150/month SaaS solution by building a serverless platform that migrated 10,000+ users.
 - Achieved **<1s LCP** on slow 4G networks using a Next.js/Vercel/Supabase stack with extensive frontend optimizations.

Technical Skills

Languages	Python, C++, BASH, TypeScript, MATLAB
Developer/ML AI	FastAPI, RestAPI, PyTorch, Hugging Face (Transformers, PEFT), TensorFlow, scikit-learn, NumPy RLHF, DPO, PPO, Instruction-Tuning, Supervised Fine-Tuning (SFT)
DevOps & Tooling	Docker, Git, GitHub Actions, CI/CD, monitoring, Linux (WSL/Ubuntu), W & B
Coursework	Distributed System, Operating Systems, System Design