

Aditya Raj

Systems & ML Research
NIT Patna (2022–2026) | CGPA: 7.83/10.0

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

Professional Experience

- **Software Engineering Intern (Infrastructure)** India
QFI Research Capital May 2025 – Oct 2025
 - **Backend Engineering:** Engineered a high-throughput backend service in **Golang** and **FastAPI** for backtesting toolkits.
 - **Performance:** Implemented data flow, achieving a throughput of **100,000 events** with **latency under 9ms**.
 - **Charting:** Optimized UI data loading, resulting in a 50% improvement in smoothness and reduction in latency.
- **Research Intern (Academic Thesis)** Hyderabad, India
IIT-H / Spatial Informatics Lab Apr 2025 – July 2025
 - Designed a framework integrating KG with RAG, for retrieval tasks, outperformed existing retrieval baselines by **14%**.

Research

[Knowledge Graph-Informed Query Decomposition\(KG-IQD\): Hybrid KG-RAG Reasoning in Noisy Contexts](#)
Authors: **Aditya Raj**, **Dr. Kuldeep Kurte**^{PI}

Findings

[Why Safety Constraints in LLMs Are Easily Breakable? Knowledge as a Network of Gated Circuits](#)
Proposed a method to understand emergent phenomena in LLMs and their internal representations. [\[docs\]](#)

Notable Software

- SecureLock** | Anti-Cheat toolkit [C++] [\[Source Code\]](#)
- user-mode evasion detection engine using C++ Windows API hooks to monitor thread execution and identify hypervisor signatures (VMware, KVM), a 3-tier architecture (Native EXE, Chrome Extension, Web App).

Key Projects

- **Efficient LLMs via Switchable and Dynamic Quantization** [docs](#) / October 2025
Tools: PyTorch, Hugging Face, LoRA, QAT-LLM, SQuAD Dataset
 - **Switchable quantization framework** for **GPT-2** using **LoRA modules (INT8/FP16)**, **reducing size by 29%**.
 - Implemented **CPT on SQuAD**, CPT achieved an **improvement of 3%**, demonstrating the **CPT - ICLR** claim.
 - Increasing model robustness and recovering inference **accuracy by 1.2%** using Double Win Quant - **ICML**.
- **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 **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**.

Technical Proficiency

Core Stack	C++, Python, SQL (Postgres)
AI	SFT, LoRA, QLoRA, RLHF, DPO, PPO, Instruction-Tuning, Quantization (INT8/FP16/Dynamic), Multi-GPU Training(upto 8X GPU clusters)
Frontend	Next.js, React, GSAP, TailwindCSS
Backend	FastAPI, Node.js, gRPC, WebSockets, Redis, Nginx
Infrastructure	Docker, CI/CD (GitHub Actions), Vercel

Select Achievements

M2L School	Selected (Top 10%) for global ML School summit in Split, Croatia	2025
Mathematics	Top 0.4% in Regional Mathematical Olympiad (RMO) among 250k candidates	2019
Hackathons	Winner, IIT ISM AI Challenge (Built OCR pipeline)	2024