Aditya Raj

ML Applied & Research
NIT Patna (2022–2026)

adityar.ug22.ec@nitp.ac.in
+91-8797073498
linkedin.com/in/hexronus
github.com/hexronuspi
hexronuspi.github.io

Professional Experience

- **QFI Research Capital** *India*
  Applied AI Intern (Quant) | 4×A100, LLMs, Kafka, Linux/Bash, CI/CD, Data Pipelines *May 2025 – Oct 2025*
  - **Applied AI:** Built LLM-based NLP pipelines news aggregation engine for granular equity-impact modeling, and multi-hop causal reasoning across stock events; achieved strong performance with **F1 0.975**, **ROC-AUC 0.995**.
- **Research Intern (Academic Thesis)** *Hyderabad, India*
  IIIT-H | Spatial Informatics Lab *Apr 2025 – July 2025*
  - Designed a **KG-integrated RAG** framework for LLM-driven NLP tasks (retrieval, long-context reasoning), **outperforming benchmarks 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 [Blog Post]
Proposed a method for emergent behavior in LLMs and a theory that they can always be jailbroken. [docs]

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**, achieving an **improvement of 3%**, reproduced results from the **CPT (ICLR)** paper
  - Increasing model robustness and **accuracy by 1.2%** towards adversarial attack 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	Python, C++, SQL (Postgres), Triton (basic)
Libraries	TRL, vLLM, Hugging Face, PyTorch, DeepSpeed Zero, Scikit learn, Numpy, Pandas
AI	SFT, LoRA, QLoRA, RLHF, DPO, PPO, Instruction-Tuning, Quantization (INT8/FP16/Dynamic), MoE, Multi-GPU Training(upto 8X GPU clusters), bitsandbytes, Large Dataset Handling, Pytorch Compile, LLM Steering, data-driven constrained decoding and inference-time alignment
Interop & Safety	Activation Patching, Red-Teaming, Sparse Autoencoders, Jailbreaking
Concepts	Probabilty, Statistics, Machine Learing, Deep Learning, Large Language Models(LLMs), Reinforcement Learning, Alignment

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*

Notable Software

SecureLock | Anti-Cheat toolkit [C++] [Source Code]

- C++ Windows API hooks monitor thread execution - identify hypervisor (VMware, KVM), (EXE, Extension, Web).