Rajesh Kumar Bandaru

Statement of Purpose

PhD in Computer Science

I have always believed that AI models should be runnable locally. My first experience with open-source models revealed that larger models could not run on my laptop, sparking my curiosity about making AI efficient enough to operate on limited hardware without significant loss of accuracy. This led me to explore quantization and distillation. While both were interesting, quantization stood out for its direct impact on memory usage, computational cost, and inference speed. I studied data types, integer vs. floating-point formats, and advanced techniques such as GPTQ, AWQ and QLoRA, realizing that mastering quantization is essential for practical, deployable AI systems.

My Master's in Computer Science with a specialization in Computational Intelligence at Illinois Institute of Science and Technology strengthened my foundation in machine learning, and algorithmic reasoning. Courses and projects exposed me to deep learning, optimization, and data-driven problem solving, directly supporting my work with LLMs, quantization, and edge AI systems.

To build hands-on experience, I created a small LLM from scratch using PyTorch, implementing a decoder-based transformer, training it end-to-end, and exploring embeddings, attention mechanisms, and matrix operations during inference. This project gave me insight into CPU and GPU memory behavior and precision-performance trade-offs, forming the basis for my article "How Does Matrix Multiplication Happen in LLMs?"

I then focused on post-training quantization, studying GPTQ in depth and authoring "Making LLMs Smaller: The Story of GPTQ" to explain techniques such as lazy batch updates and Cholesky decomposition, which improve speed, stability, and memory efficiency. I also built a GGUF Model Explorer to inspect model metadata, tensors, and quantization formats, gaining direct experience with low-bit weight storage and deployment trade-offs.

Alongside LLM work, I explored edge AI and robotics. Using 3D-printable parts, I built a robotic arm integrated with servos and Raspberry Pi, added a camera, and deployed a lightweight object detection model using a Hailo AI module, enabling accelerated inference on the edge.

My professional experience as a Data Automation Engineer reinforced these skills. I engineered end-to-end automation workflows for multiple data portals using Python, Playwright, and Azure, converting Jira tickets into executable jobs, automating multi-stage ETL processes, and deploying containerized services at scale. These experiences honed my ability to design reliable, hands-free systems, aligning closely with the challenges of efficient AI model deployment and inference pipelines for edge and decentralized devices.

Your lab's research areas in Efficient and Secured Machine Learning and Edge AI resonate strongly with my interests. I am particularly drawn to deploying adaptive and secure AI models

on edge devices. My prior experiments with LLMs, robotics, and low-resource systems provide a strong foundation to contribute to this research.

For my next steps, I have hosted a llama.cpp server on my laptop with a quantized QWEN model, using Python scripts to query it. I plan to systematically compare this locally running model with API-based models (ChatGPT, Claude, Gemini), analyzing performance on text generation, multi-modal tasks, and PDF question answering. Simultaneously, I will experiment with different tokenization methods and hyperparameters on my self-built model to study trade-offs between model size, inference speed, and accuracy, ultimately exploring custom GPU kernels and inference engines for efficient real-time deployment of LLMs and vision tasks.

During my PhD, I aim to advance research in quantization and efficient model deployment for decentralized edge devices while maintaining robustness and reasoning ability. I plan to explore domain adaptation, 3D-vision, and edge AI holistically, developing lightweight models capable of adapting to dynamic environments. I am highly self-motivated, driven by curiosity and hands-on experimentation. I approach learning by building and testing, maintaining a consistent schedule to ensure steady progress. I am eager to bring this mindset to your lab, contributing to research on efficient and secure AI systems that bridge theory and practical deployment.