**AI Deployment and Development Intern Positions**

**Position 1: Edge AI Deployment on NVIDIA Jetson Platform**

**Project: Private AI Model Deployment and Optimization for Edge Computing**

**Project Description:** The intern will develop and deploy private AI models on NVIDIA Jetson edge devices, focusing on creating secure, low-latency inference systems for real-time applications. The project involves optimizing AI models for edge constraints, implementing privacy-preserving techniques, and building robust deployment pipelines. The intern will work with computer vision, natural language processing, or IoT sensor data processing models while ensuring data privacy and security at the edge. This project will be performed by 1-2 students.

**Why It Matters / Global Impact:** Edge AI deployment is critical for applications requiring real-time processing, data privacy, and reduced bandwidth usage. Industries such as autonomous vehicles, smart manufacturing, healthcare monitoring, and smart cities rely on edge AI to process sensitive data locally while maintaining performance. This project demonstrates how private AI can be effectively deployed at the edge, enabling organizations to leverage AI capabilities without compromising data security or privacy.

**Objectives:**

* Deploy and optimize AI models on NVIDIA Jetson devices (Nano, Xavier, Orin series)
* Implement model quantization, pruning, and TensorRT optimization for edge performance
* Develop secure inference pipelines with privacy-preserving techniques (federated learning, differential privacy)
* Build containerized deployment solutions using Docker and NVIDIA Container Runtime
* Create monitoring and management systems for edge AI fleet deployment
* Implement over-the-air model updates and version management

**Deliverables:**

* Optimized AI models running on NVIDIA Jetson hardware with performance benchmarks
* Containerized deployment pipeline with automated model optimization
* Privacy-preserving inference framework with security documentation
* Edge device management dashboard for monitoring and updates
* Performance analysis comparing edge vs. cloud inference
* Technical documentation and deployment guidelines

**Milestones (6 months):**

* **Weeks 1-2:** Hardware setup, NVIDIA Jetson familiarization, and development environment configuration
* **Weeks 3-6:** Model optimization using TensorRT, quantization, and performance profiling
* **Weeks 7-10:** Privacy-preserving techniques implementation and secure inference pipeline development
* **Weeks 11-14:** Edge deployment automation, monitoring systems, and fleet management tools
* **Weeks 15+:** Testing, optimization, security validation, documentation, and final presentation

**Intern Background:**

* Strong programming skills in Python and C++
* Experience with deep learning frameworks (PyTorch, TensorFlow, ONNX)
* Familiarity with NVIDIA CUDA, TensorRT, or similar GPU acceleration tools
* Knowledge of containerization technologies (Docker, Kubernetes)
* Interest in edge computing, IoT, and privacy-preserving AI

**Supervisors:**

* Mahbubul Alam (LinkedIn)
* Dr. Neeli Prasad (LinkedIn)
* Dr. Albena Mihovska (LinkedIn)