**Position 2: Private AI Deployment on Google Kubernetes Engine (GKE)**

**Project: Scalable Private AI Infrastructure on Google Cloud Platform**

**Project Description:** The intern will design and implement a scalable private AI deployment infrastructure on Google Kubernetes Engine (GKE), focusing on secure multi-tenant AI services, auto-scaling, and privacy-compliant model serving. The project involves building cloud-native AI pipelines that maintain data privacy while providing high-performance inference capabilities. The intern will work with microservices architecture, implement secure API gateways, and develop monitoring solutions for production AI workloads. This project will be performed by 1-2 students.

**Why It Matters / Global Impact:** Organizations need to deploy AI models at scale while maintaining strict privacy and compliance requirements. GKE provides a robust platform for deploying private AI services that can handle enterprise-scale workloads while ensuring data security, regulatory compliance, and cost efficiency. This project demonstrates how cloud-native technologies can be leveraged to build secure, scalable AI infrastructure that serves multiple applications and users simultaneously.

**Objectives:**

* Deploy private AI models on GKE with secure, scalable inference services
* Implement multi-tenant architecture with namespace isolation and resource management
* Develop auto-scaling policies based on inference load and resource utilization
* Create secure API gateways with authentication, authorization, and rate limiting
* Implement privacy-preserving techniques including encrypted inference and secure enclaves
* Build comprehensive monitoring, logging, and alerting systems using Google Cloud operations

**Deliverables:**

* Production-ready AI inference services deployed on GKE with high availability
* Multi-tenant architecture supporting isolated AI workloads
* Auto-scaling configuration with performance optimization
* Secure API gateway with comprehensive access controls
* Privacy and security compliance documentation
* Monitoring dashboards and operational runbooks

**Milestones (6 months):**

* **Weeks 1-2:** GKE cluster setup, Google Cloud services familiarization, and security configuration
* **Weeks 3-6:** AI model containerization, deployment pipelines, and basic inference services
* **Weeks 7-10:** Multi-tenant architecture implementation and auto-scaling configuration
* **Weeks 11-14:** Security hardening, privacy compliance implementation, and monitoring setup
* **Weeks 15+:** Load testing, security validation, documentation, and final presentation

**Intern Background:**

* Strong programming skills in Python and experience with containerization
* Familiarity with Kubernetes, Docker, and cloud-native technologies
* Knowledge of Google Cloud Platform services (GKE, Cloud Run, IAM, etc.)
* Experience with CI/CD pipelines and DevOps practices
* Understanding of AI/ML model serving and microservices architecture
* Interest in cloud security, privacy-preserving technologies, and scalable systems

**Supervisors:**

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