THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Job Description: AI Application Engineer

Smart Lab
Large and Trustworthy AI for Healthcare

Role Summary

Smartlab is seeking a skilled and motivated AI Application Engineer to build clinical-grade AI applications in healthcare. You will translate cutting-edge deep learning (including LLMs/VLMs) into robust, user-centric solutions for medical imaging and medical NLP, with a strong focus on algorithm packaging, model pr high-performance inference, interoperability, and regulatory compliance. You will work closely with our PhD researchers and clinical stakeholders to move models from research to real-world use.

Key Responsibilities

- **System Testing and Validation:** Design and execute comprehensive testing plans to ensure the reliability, scalability, and robustness of AI-driven systems.
- **Integration & Downstream Adaptation**: Collaborate with researchers and clinicians to fine-tune and adapt models for medical imaging/NLP, integrating them into backend services or user-friendly GUI applications.
- **Full-Stack Delivery**: Develop and maintain front-end (React/Vue/Svelte), backend/API (FastAPI/Flask), and deployment infrastructure (Docker, Kubernetes) for clinical workflows.
- **High-Performance Deployment**: Optimize model serving pipelines (e.g., TensorRT, vLLM, Hugging Face TGI) for low-latency, scalable inference in resource-limited circumstances.

Required Qualifications

- Bachelor’s or Master’s degree in Computer Science, Software Engineering, or a related technical field.
- Strong programming skills in Python and experience with common AI/ML libraries (e.g., PyTorch, TensorFlow, Hugging Face).
- Practical knowledge of GPU/accelerator optimization (e.g., TensorRT, Triton, ONNX Runtime EPs) and model export/optimization (e.g., ONNX).

- Full-stack development experience: user-centric UI/UX front-end (e.g., React, Vue, or Svelte), back-end/API (e.g., FastAPI or Flask), and containerization/deployment with Docker.
- Proven experience taking deep learning models from research to production environments.

Preferred Qualifications

- Advanced skills in high-performance serving (e.g., vLLM, TensorRT-LLM, Hugging Face TGI) and orchestration frameworks (e.g., KServe, Ray Serve, NVIDIA Triton).
- Familiarity with Large Language Models (LLMs) and Vision-Language Models (VLMs).
- Knowledge of encryption and secure model handling practices, especially for healthcare data.
- Experience maintaining production AI applications on cloud platforms (e.g., AWS, GCP, Azure) or Kubernetes clusters.
- Prior work in research-oriented or healthcare settings, with interest in solving complex clinical problems.

How to Apply

Interested candidates are welcome to send their resumes, transcripts and other relevant materials to Prof. Chen's email: **jhc@ust.hk**. *Please indicate the application position in the subject of email.*