EPSRC DTP PhD Research Project: Advertising Form

Project Title: Lightweight Guardrails for Large Language (Vision-Language) Models

Primary Supervisor details: Gaojie (Jay) Jin, g.jin@exeter.ac.uk
personal website

Additional Supervisory team details: Yanda Meng y.m.meng@exeter.ac.uk
personal website

Department: Computer Science

Location: Innovation Centre, Streatham campus, Exeter.

PhD Programme: PhD in computer science

Project Description:
As large language models (LLMs) and vision-language models (VLMs) grow in size and capability, the demand for ensuring their safe and responsible deployment becomes paramount. Traditional guardrail mechanisms are often resource-intensive, making them impractical for real-time applications and large-scale deployments. This project aims to pioneer the development of lightweight yet highly effective guardrails tailored for LLMs and VLMs, ensuring robust safety and reliability without compromising on performance efficiency. The proposed research will integrate cutting-edge theoretical frameworks in robustness analysis and statistical learning, such as Probably Approximately Correct (PAC) learning, with advanced implementation strategies designed for scalability and low computational overhead. By innovating on both the conceptual and practical fronts, the project seeks to deliver guardrails that are not only theoretically sound but also optimized for real-world application—balancing the trade-offs between model size, speed, and safety. This work is particularly novel in its approach to minimizing the energy and time costs associated with safety mechanisms, making it feasible to implement guardrails in a wide range of AI systems, from cloud-based services to edge devices.

Entry Requirements:
Basic knowledge about Python, Pytorch, and DNN's safety properties (generalisation, robustness, …).

Project specific enquiries: Gaojie (Jay) Jin, g.jin@exeter.ac.uk