Newcastle University

Project title

Safe Multimodal Large-Language Models

Supervision team

Main Supervisor: Huizhi Liang <Huizhi.Liang@newcastle.ac.uk >
Co-supervisors: Varun Ojha < Varun.Ojha@newcastle.ac.uk >

Research project

The proposed PhD research focuses on creating Safe Multimodal AI for Healthcare by enhancing deep learning techniques that integrate and analyze diverse clinical data sources, such as medical texts, imaging, and sensor data. A key component of this initiative will be the utilization of Transformer-based architectures, particularly Visual-Language Models (VLMs), to facilitate combined reasoning across text and image modalities for applications like clinical decision support and automated diagnostic processes. This research will tackle significant safety concerns, including risks related to data poisoning and sensor malfunctions, which can jeopardize model reliability and patient safety. Additionally, it will address the challenges posed by heterogeneous data by developing robust fusion strategies to harmonize structured, unstructured, and temporal data within unified multimodal frameworks. The proposals will follow the AI Social, Operational, and Transparency AI Safety framework.

Applicant skills/background

This project requires skills in programming in python and machine learning research skills. The project welcomes researchers from disciplines of mathematics, engineering, physics, computer science, electronics.

References

Safe-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions (Bianchi et al., 2023)

Efficient Adversarial Training in LLMs with Continuous Attacks (Xhonneux et al., 2024)

Almost Surely Safe Alignment of Large Language Models at Inference-Time (Ji, Ramesh, Zimmer, Bogunovic, Wang, Bou Ammar, 2025)

REVAL: A Comprehension Evaluation on Reliability and Values of Large Vision-Language Models (Zhang et al., 2025)