Project B05 of SFB 1629 is pleased to announce a talk by Frank Niu from the University of Toronto. The talk is scheduled for Monday, May 13th, 10:15 – 11:45, in room IG 1.418 (IG Farben building):

Frank Niu

(University of Toronto)

Dissecting Language Models: From Black Boxes to Interpretable AI

Transformer language models, the cornerstone of the recently popular large language models (LLMs), have revolutionised the fields of artificial intelligence (AI) and natural language processing (NLP). However, we still understand relatively little about their computational mechanisms and the reasons for their effectiveness. The internal workings of these models remain enigmatic "black boxes." Nevertheless, we are living in an exciting era of interpretation research. We are on the cusp of prying open this black box and paving the way toward the development of explainable and controllable language models and AI agents.

In this talk, I will discuss the evolution of LM interpretation research: interpreting the mechanisms of LMs with finer and finer strokes by moving from investigating output probabilities to probing representations holistically, examining individual layers, and even manipulating individual MLP neurons. Finally, I will present circuit discovery, a nascent but promising new thread of research that offers a path forward to explainable, controllable, and modular language modelling.