Chair of IT Security
Prof. Dr. Claudia Eckert

Fraunhofer AISEC

Student Assistant (m/f/)*

Concept Representations for Secure Deep Learning

Deep Learning models – most prominently Large Language models (LLMs) – are able to represent concepts as linear vectors in their internal representation space [1]. While much work is dedicated towards finding a comprehensive theory of concept representations, there are many unexplored areas related to potential practical applications. With the ability to control model behavior using steering vectors, one open question is the applicability to secure Machine Learning (ML), e.g. by steering LLMs towards giving truthful answers. In this position, you will support in theoretical analysis and practical application of internal representations to identify the potential of concept vectors in ML security. Given the early research stage, you have the opportunity to make an impact in the ML research community while gaining valuable experience through regular interactions with colleagues in the field.

Task Description

Your responsibilities in this role will encompass a variety of tasks, including:

- Conducting literature research
- Evaluating existing theories and finding gaps for further investigations
- Assisting in the development of algorithms to bring practical use to theoretical insights
- Contributing to scientific publications

Requirements

- Strong problem-solving skills, high motivation and ability to work independently
- Currently enrolled in a degree program in a related field
- Advanced knowledge of theoretical ML and related mathematics
- Practical experience with ML frameworks & Python (preferably PyTorch)

Contact

Please send your application with current CV and transcript of records to:

Maximilian Wendlinger
Fraunhofer Institute for Applied and Integrated Security (AISEC)
Cognitive Security Technologies
Lichtenbergstr. 11, 85748 Garching near Munich
Mail: maximilian.wendlinger@aisec.fraunhofer.de

[1] K. Park, Y. J. Choe, and V. Veitch, “The Linear Representation Hypothesis and the Geometry of Large Language Models,” (2024) https://arxiv.org/abs/2311.03658

Publication Date: 12.04.2025