Einladung zum Oberseminar Mathematik des Maschinellen Lernens und Angewandte Analysis

Julius-Maximilians-Universität Würzburg
Professur für Mathematik des Maschinellen Lernens

Dr. Leo Schwinn

Technical University of Munich, Department of Computer Science

Adversarial Threats and Defenses in LLMs

Over the past decade, there have been extensive research efforts towards improving the robustness of neural networks to adversarial attacks, yet this problem remains vastly unsolved. Here, one major impediment has been the overestimation of the robustness of new defense approaches due to faulty defense evaluations. Flawed robustness evaluations necessitate rectifications in subsequent works, dangerously slowing down the research and providing a false sense of security. In this context, we will face substantial challenges associated with an impending adversarial arms race in natural language processing, specifically with closed-source Large Language Models (LLMs), such as ChatGPT, Google Gemini, or Anthropic's Claude. In this talk, we will discuss underexplored threat models in LLMs and possible ways to defend against them.

Ort:Humboldt-Bau, Seminarraum 41.00.006 Zeit: Mittwoch, 08.01.2025 14:15

Zu diesem Vortrag laden wir Sie herzlich ein.

gez. Leon Bungert