Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

# Evaluating and Improving the Robustness of Large Language Models against Attacks


Large language models, like ChatGPT, have demonstrated their exceptional performance in natural language processing (NLP) tasks. However, recent studies have revealed their vulnerability to adversarial attacks, which can result in incorrect predictions, posing a significant threat to the reliability and safety of these models. This study aims to evaluate the robustness of large language models against different attacks and propose techniques to improve their performance in such scenarios.

**Requirements:** Strong motivation, knowledge in deep learning, or a solid background in machine learning. Experience with Python and TensorFlow or PyTorch is an advantage as well as knowledge in graph theory, distributed computing and graph neural networks.

**Interested? Please contact us for more details!**

## Contact

- Zhao Meng: zhmeng@ethz.ch, ETZ G60.1