Vulnerabilities of Large Language Models to Adversarial Attacks

Yu Fu, Erfan Shayegan, Md. Mamun Al Abdullah, Pedram Zaree, Nael Abu-Ghazaleh and Yue Dong

This tutorial serves as a comprehensive guide on the vulnerabilities of Large Language Models (LLMs) to adversarial attacks, an interdisciplinary field that blends perspectives from Natural Language Processing (NLP) and Cybersecurity. As LLMs become more complex and integrated into various systems, understanding their security attributes is crucial. However, current research indicates that even safety-aligned models are not impervious to adversarial attacks that can result in incorrect or harmful outputs. The tutorial first lays the foundation by explaining safety-aligned LLMs and concepts in cybersecurity. It then categorizes existing research based on different types of learning architectures and attack methods. We highlight the existing vulnerabilities of unimodal LLMs, multi-modal LLMs, and systems that integrate LLMs, focusing on adversarial attacks designed to exploit weaknesses and mislead AI systems. Finally, the tutorial delves into the potential causes of these vulnerabilities and discusses potential defense mechanisms.

Yu Fu, Ph.D. candidate in Computer Science and Engineering at UC Riverside.
email: `yfu093@ucr.edu`
website: `https://fyyfu.github.io/`
He is a first-year Ph.D student advised by Prof. Yue Dong. His research interests lie in natural language processing and machine learning. His recent research focuses on safety alignments, watermarking text generation models, and reducing hallucinations in large language models.

Erfan Shayegani, Ph.D. Candidate in Computer Science and Engineering at UC Riverside.
email: `sshay004@ucr.edu`
website: `https://erfanshayegani.github.io/`
He is a Ph.D student advised by Prof. Yue Dong and Prof. Nael AbuGhazaleh. His research interests lie at the intersection of Generative AI and Systems, with topics spanning NLP, Alignment, and Scalability/Security/Privacy within the domain of Large (Vision) Language Models (LLMs), Multi-Modal Foundation Models, and Text-3D models such as NeRF. His keen interest lies in integrating these complex models into sophisticated systems, where he takes an adversarial approach to uncover vulnerabilities and strengthen their trustworthiness, safety, and scalability. He holds a B.Sc. in Electrical Engineering from Sharif University of Technology.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 8–9

Md. Mamun Al Abdullah, Ph.D. Candidate in Computer Science and Engineering at UC Riverside.
email: mmamu003@ucr.edu
website: https://sites.google.com/view/aamamun
He is a Ph.D. student advised by Prof. Nael Abu-Ghazaleh. His research focuses on system security.

Pedram Zaree, Ph.D. Candidate in Computer Science and Engineering at UC Riverside.
email: pzare003@ucr.edu
website: https://pedramzaree.github.io/
He is a Ph.D. student in Computer Science and Engineering at UC Riverside, advised by Prof. Nael Abu-Ghazaleh. His research focuses on the security of retrieval-based NLP systems, security of machine learning systems, and AR/VR systems. He also has a background in electrical engineering since he has completed a B.Sc program in the Electrical Engineering department at Sharif University of Technology.

Nael Abu-Ghazaleh, Professor in the Computer Science and Engineering Department at the University of California, Riverside.
email: naelag@ucr.edu
website: https://www.cs.ucr.edu/ nael/
His research is in computer systems, with emphasis on security of emerging systems. His group has developed a number of new attacks on CPUs, GPUs, AR/VR devices and operating systems, that have been reported to industry and resulted in patches and changes to consumer products. He is currently serving as the co-general chair of ASPLOS'24 and PACT'24, and as the program chair of SEED'24. He is an ACM Distinguished Member, an IEEE Distinguished lecturer, and a member of the IEEE Micro Hall of Fame.

Yue Dong, Assistant Professor in Computer Science and Engineering at the University of California, Riverside.
email: yued@ucr.edu
website: https://yuedong.us/
She leads the Natural Language Processing group, which develops NLP systems that are trustworthy, safe and efficient. She served as senior area chair for AACL'23 and area chair for EMNLP'22 & '23, and has co-organized workshops at EMNLP'21 & '23, NeurIPS'21 & 22 &'23, and tutorials at NAACL'22 and KDD'23.