Defending Against Generative AI Threats in NLP

1 Description of tutorial topic

Generative AI and, in particular, Large Language Models (LLMs) have seen unprecedented advancements in the last few years. Given the ease of access of many state-of-the-art LLMs, these models have been heavily adopted and have entered workflows among professionals, academics and even enthusiastic lay users. With impressive performance on natural language and benchmarks, and even more complex tasks involving math and reasoning, LLMs are capable of being used as either generalist or topic-specific chatbots. While larger and more capable LLMs are being developed and used for a variety of high-impact use cases, these models are still susceptible to being misused and attacked by malicious entities. In this tutorial, we dive into the current state of LLM research and development, before exploring the types of threats and attacks that LLMs are susceptible to, and finally exploring the various defense methods that have been developed to tackle these threats, along with challenges and research directions that need attention from the community.

2 Tutorial Structure

- **Introduction to Generative AI and Large Language Models (LLMs)**
  1. History & recent trends.
  2. What are LLMs, how they work.
  3. Overview of models and example use-cases.
- **Threats in the context of LLMs**
  1. Attacks *using* LLMs (Case study: mis/disinformation with LLM-generated text)
  2. Attacks *on* LLMs (Jailbreaks, unalignment, data poisoning, etc.)
- **Defense Mechanisms against Attacks on LLMs**
  1. Detection of attacked/compromised models.
     - Garak: LLM vulnerability scanner
  2. Realignment
     - NeMo Guardrails: Programmable rails for LLM content moderation & safety
     - Model editing / in-flight safety steering
- **Challenges & Future Research Directions**

3 Expected Audience

This tutorial is intended for students, researchers, practitioners interested in using generative AI technologies such as LLMs in a variety of use cases. Audiences from a wide range of backgrounds keen on learning about effectively and responsibly leveraging LLMs can greatly benefit from this tutorial.

4 Short Bio and Contact Information of Authors

- **Amrita Bhattacharjee** is a fifth year Ph.D. student in the School of Computing and Augmented Intelligence at Arizona State University. Her research interests broadly include Machine Learning, Natural Language Processing, AI Safety, and in particular LLM Safety. Contact: abhatt43@asu.edu.
- **Dr. Raha Moraffah** is an Assistant Professor in the Department of Computer Science at Worcester Polytechnic Institute (WPI). Prior to joining WPI, she earned her Ph.D. in computer science from Arizona State University. Raha's research spans machine learning, data mining, artificial intelligence, and causal inference, with a specific focus on developing trustworthy and responsible machine learning and generative AI algorithms. Raha is the organizer of the first tutorial on causal responsible ML at KDD 2023 and organizer of the tutorial on causal inference at SBP 2022. Contact: rmoraffah@wpi.edu.
- **Dr. Christopher Parisien** is a Senior Manager of Applied Research at NVIDIA, leading the development of NeMo Guardrails, a toolkit for safety and security in Large Language Models. Chris holds a Ph.D. in Computational Linguistics from the University of Toronto, where he used AI models to explain the strange ways that children learn language. His current focus at NVIDIA is to bring trustworthy language models to large enterprises. Contact: cparisien@nvidia.com.
- **Dr. Huan Liu** is a Regents Professor of Computer Science at the School of Computing and Augmented Intelligence at Arizona State University. His research interests are in data mining, machine learning, social computing, and artificial intelligence, investigating interdisciplinary problems that arise in many real-world, data-intensive applications with high-dimensional data of disparate forms such as social media. Contact: huanliu@asu.edu.