CS 5914 Security Risks of Generative AI

Fall 2023, Instructor: Dr. Bimal Viswanath

1 Course description

Generative AI is starting to disrupt many technological domains. Advances made in Large Language Models (LLM), Generative Vision models, and Multi-Modal Generative models have enabled new ways for us to retrieve, communicate and process information online. This course will look at the safety and security risks associated with cutting-edge Generative AI technologies and cover the following topics:

- Understanding failures of Generative AI that can cause harm to its users. This includes:
  - Toxicity in LLM-based applications, e.g., chatbots.
  - Hallucination by LLMs and their inability to provide factual, accurate information.
  - Challenges with filtering undesired content from Text-to-Image generative models.
- Attacks that violate the integrity of Generative AI to cause harm. We will focus on prompt-injection attacks against LLM-based applications. We will also cover new threats impacting LLM plugins.
- Misuse of Generative AI to create harmful content. We will study the threat of synthetic media or "deepfakes" produced using Generative AI. We will look at both text and vision modalities and focus on synthetic media detection schemes.
- Protecting real media (i.e., not synthetic) from (harmful) Generative AI use cases. We will study proactive defenses that aim to protect user images from deepfake manipulations. We will also cover recent work that aims to protect media from being successfully used in training Generative AI, e.g., protecting artists to prevent their artistic styles being learned by Generative AI.

2 Reference materials

Most reading material will be drawn from research papers published at venues such as IEEE S&P, Usenix Security, CCS, NDSS, IMC, WWW, ICML, AAAI and NeurIPS.

3 Prerequisites

Students are expected to have a basic understanding of deep learning and machine learning in general. Familiarity with frameworks like TensorFlow and PyTorch is required. Students who enroll for the course are expected to be highly motivated to learn and work hard and be ready to make up for any prerequisite deficiencies they may have.

4 Grading

This course will be project-based. Grading will be based on class participation, project presentations, project report, and other project deliverables.