Unadversarial attacks on natural language generation

Background: Natural language generation (NLG) is the production of understandable texts via machine learning models. It is used in a variety of fields and commonly in chatbots such as ChatGPT. However, the produced chatbots are easily misled and often respond with incorrect answers. Moreover, some chatbots are known to engage in improper conduct, such as Meta's “BlenderBot” repeating antisemitic and right-wing conspiracy theories. Relevant papers: “Unadversarial examples: Designing objects for robust vision”.
Project Description: In this project, we aim to reduce the harmful and incorrect outputs of chatbots by utilizing unadversarial attacks.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)