Reinforcement Learning for Safe Content Generation in Sensitive Domains

Maisha Mahrin
Computer Science, NYUAD
mm10294@nyu.edu

Advised by: Primary Advisor: Christina Pöpper, Secondary Advisor:

ABSTRACT

Designing and building effective and secure generative reinforcement learning (RL) models for sensitive topics presents unique challenges in the context of cybersecurity. In this research, we propose a novel framework for training generative RL models using a combination of positive and negative reinforcement to prevent the model from generating harmful or sensitive content while promoting informative and positive content. Our approach leverages pre-training on large datasets and fine-tuning on target datasets to ensure the model learns to generate informative content while minimizing the risk of generating harmful content. We evaluate the effectiveness of our proposed framework on a large corpus of sensitive texts, including political and conspiracy theory texts. Our results demonstrate the effectiveness of our approach in training generative RL models that are effective in generating informative content while also preventing the model from generating harmful outputs.

To conclude, our research contributes to the growing body of knowledge on building generative RL models for sensitive topics and has important implications for applications in various domains, including online discourse, social media moderation and content filtering. Our work paves the way for more effective and secures generative AI models in the future.

Reference Format:
Maisha Mahrin. 2023. Reinforcement Learning for Safe Content Generation in Sensitive Domains. In *NYUAD Capstone Seminar Reports, Spring 2023, Abu Dhabi, UAE.* 1 page.

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك ابوظبي
NYU ABU DHABI

Capstone Seminar, Spring 2023, Abu Dhabi, UAE