Generating online grooming scenarios based on existing scenarios using LLMs.

Imagen que contiene Texto

Descripción generada automáticamente

University of Kent

A Dissertation submitted for the degree of

MSc. Cybersecurity

September 2024

Word count: 12,000 approx.

Cristina Pastor Pastor

# Contents

[**Contents**](#_Contents) ................................................................................................ i

**List of Tables** .......................................................................................... iii

**List of Figures** ........................................................................................ iv

[**Abstract**](#_Abstract) ................................................................................................. vi

[**Abbreviations**](#_Abbreviations) ......................................................................................... vii

[**Chapter I**](#_Chapter_1)................................................................................................ 1

[**Introduction**](#_Introduction) ........................................................................................... 1

[I.I Background and Motivation](#_I.I_Background_and) ............................................................. 1

[I.2 Research aims and objectives](#_I.2_Research_Aims) ........................................................ 3

[I.3 Chapter Overview](#_I.3_Chapter_Overview) .......................................................................... 3

[**Chapter 2**](#_Chapter_2) ................................................................................................ 4

[**Literature Review**](#_Literature_Review) .................................................................................... 4

[2.I LLMs in Synthetic Data Generation](#_2.I_LLMs_in) ........................................................... 4

[2.2 Applications of LLMs in Online Safety and Grooming Detection](#_2.2_Applications_of) .................. 5

[2.3 Challenges Obtaining Real Online Grooming Data](#_2.3_Challenges_Obtaining) ..................................... 6

[2.4 Ethical Considerations in Using LLMs for Sensitive Content Generation](#_2.4_Ethical_Considerations) …………….. 6

[**Chapter 3**](#_Chapter_3) ................................................................................................. 8

[**Methodology**](#_Methodology) ........................................................................................... 8

[3.I Overview](#_3.I_Overview) ........................................................................................... 8

[3.2 Problem Description](#_3.2_Problem_Description) .................................................................................. 8

[3.3 LLM Selection Process](#_3.3_LLM_Selection) .................................................................... 8

[3.4 LLM Assessment Method](#_3.4_LLM_Assessment) ................................................................ 10

[**Chapter 4**](#_Chapter_4) ................................................................................................ 12

[**Results**](#_Results) ..................................................................................................... 12

[4.I Overview](#_4.I_Overview) ................................................................................................ 12

4.2 UK Cyber Essentials (CE) ..................................................... 12

4.3 ISO 27001 ...................................................................................... 14

4.4 NIST Cybersecurity Framework (NIST CSF) ................................... 18

[**Chapter 5**](#_Chapter_5) ..................................................................... 21

[**Discussion**](#_Discussion) ............................................................................. 21

[5.I Nature of Information Gathered](#_5.I_Nature_of) ................................................. 21

[5.2 Continuous evaluation of experiment results](#_5.2_Continuous_evaluation) ................................. 22

[5.3 Comparison with related work](#_5.3_Comparison_with) ............................................. 22

[**Chapter 6**](#_Chapter_6) ..................................................................................... 24

[**Conclusion and Future Work**](#_Conclusion_and_Future) ........................................................... 24

[6.I Benefits and Impact](#_6.I_Benefits_and) ............................................................................... 24

[6.2 Limitations and Future Work](#_6.2_Limitations_and) ...................................................... 25

[**Appendices**](#_Appendices) .............................................................................................. 26

Appendix A ................................................................................................ 26

Appendix B ..................................................................................... 29

Appendix C ............................................................ 30

[**Bibliography**](#_References/Bibliography) ............................................................................................... 33

List of Tables

Table 1 showing the forms considered in this study ................ 9

Table 2 The percentage of forms that partially cover requirements of the security controls of ISO 27001.............. 15

Table 3 The percentage of forms that do not address any requirements in the controls required for each Function ............................................ 20

Table A. 1 Overview of the contents of each clause of ISO/IEC 27001. ........... 26

Table A. 2 Overview of the contents of each section of UK Cyber Essentials ........ 27

Table A. 3 Overview of the contents of each function of NIST CSF ...................... 28

Table B.1 Sample form mapping to ISO 27001 and the NIST CSF ................. 29

Table C.1 Full mapping of forms with UK Cyber Essentials .................................. 30

Table C. 2 Full mapping of forms with ISO 27001 ........................ 31

Table C.3 Full mapping of forms with UK the NIST CSFv1.1 ...................... 32

List of Figures

Fig. 1 The percentage of forms (n=36) that fully cover each control section of the UK Cyber Essentials ………………… 13

Fig. 2 The percentage of forms that fully cover each control section of the UK Cyber Essentials based on forms from the UK (n= 20) and the entire data set (n=36)....................... 13

Fig. 3 The percentage of forms that cover none, partial or all controls in a security clause for the ISO 27001 ..................................................... 14

Fig. 4 The percentages of forms that do not address any of the defined security control categories in the various clauses of the ISO 27001. ........................... 17

Fig. 5 The percentage of forms that cover none, partial or all controls in a security clause for the NIST Cybersecurity Framework ....................... 18

## Abstract

The overall project aims to generate grooming scenarios using Large Language Models (LLMs) based on existing known scenarios. Given the challenge of obtaining real online grooming scenarios due to their sensitive nature, this project seeks to create synthetic datasets by prompting LLMs to generate variants of real scenarios. This approach leverages the generative capabilities of LLMs to produce valuable training data for further studies or interventions in the realm of online safety and grooming detection.

The goal of the project was to generate grooming scenarios based on existing known scenarios by leveraging Large Language Models (LLMs) like ChatGPT and Mistral AI. Due to the difficulty in obtaining real online grooming scenarios, this study aimed to create synthetic scenarios that could closely mimic real-life situations. The generated dialogues were intended to explore variations of known grooming tactics, thus providing data that could be useful for developing protective measures and training datasets.

Goal of project is the generation of grooming scenarios based on existing known scenarios, based on message exchange using LLMs. Since it is difficult to obtain real on-line grooming scenarios, an alternative way is for an LLMs to generate scenarios that are variant of real ones. Once these scenarios are obtained, synthetic datasets can be generated.

Online grooming is a significant and growing threat in the digital age, with predators exploiting the anonymity of the internet to target vulnerable individuals, particularly minors. Developing effective methods to detect and prevent such malicious activities is crucial. However, the scarcity of real online grooming scenarios poses a challenge for researchers and developers working to enhance detection systems. This project addresses this issue by utilizing Large Language Models (LLMs) to generate realistic grooming scenarios based on known instances. By creating variations of these existing scenarios, LLMs can produce synthetic datasets that are critical for advancing research and improving detection algorithms.

## Abbreviations

LLM Large Language Model

## Chapter I

## Introduction

### I.I Background and Motivation

The fight against online grooming has been hindered by the limited availability of authentic grooming scenarios, which are often difficult to obtain due to privacy concerns, legal restrictions, and the sensitive nature of the content. Traditional methods of gathering data for research in this area are not only time-consuming but also fraught with ethical challenges. The advent of LLMs, which can simulate human-like text exchanges, presents a promising alternative. By generating realistic yet synthetic grooming scenarios, LLMs can help overcome the data scarcity problem, enabling the development of more robust detection systems. This project is motivated by the need to create these synthetic scenarios to facilitate ongoing research and innovation in online safety.

### I.2 Research Aims and Objectives

The primary aim of this project is to generate realistic grooming scenarios using LLMs, based on existing known scenarios, and to create synthetic datasets from these generated interactions. The objective of this project is to develop a method for generating grooming scenarios by leveraging Large Language Models (LLMs) to create variations of existing, known scenarios. Given the challenges associated with obtaining real online grooming scenarios, which are often scarce or sensitive in nature, the use of LLMs presents a viable alternative. By analysing and replicating patterns found in authentic message exchanges, LLMs can be trained to generate plausible grooming scenarios that retain the essential characteristics of real interactions while introducing variations. These generated scenarios can then be used to create synthetic datasets, which are essential for further research, training, and development of tools aimed at detecting and preventing online grooming. This approach not only mitigates the ethical concerns related to using real data but also provides a scalable method for generating diverse scenarios that can enhance the robustness of existing detection systems.

The specific objectives are:

1. To analyse and identify key characteristics of known grooming scenarios that can be replicated by LLMs.
2. To develop a framework for generating variations of these scenarios using LLMs, ensuring they maintain the authenticity and complexity of real interactions.
3. To evaluate the generated scenarios for realism and relevance, ensuring they can effectively contribute to the creation of synthetic datasets.
4. To produce comprehensive synthetic datasets that can be used for training and testing online grooming detection systems.

By achieving these objectives, this project will contribute valuable resources to the field of online safety, enabling the development of more effective tools to combat online grooming.

### I.3 Chapter Overview

This Dissertation is structured as follows:

* **Chapter I** contains an Introduction to the project, including an overview of the project and its aims and objectives.
* **Chapter 2** provides an in-depth Literature Review of related work by various authors in academia and non-academic work.
* Chapter 3 portrays an overview of the Methodology used in the study involving the different LLMs used throughout each experiment wave.
* Chapter 4 presents the Results of the study after generating several waves of experiments using different LLMs.
* Chapter 5 presents a Discussion of the Results and is centred on learning outcomes
* Chapter 6 provides a Conclusion drawn from the Results of the study with suggestions on how this work can be extended and further explored.

# Chapter 2

## Literature Review

### 2.I LLMs in Synthetic Data Generation

### 2.2 Applications of LLMs in Online Safety and Grooming Detection

### 2.3 Challenges Obtaining Real Online Grooming Data

### 2.4 Ethical Considerations in Using LLMs for Sensitive Content Generation

# Chapter 3

## Methodology

### 3.I Overview

### 3.2 Problem Description

Concise summary of the research problem that will be addressed.

### 3.3 LLM Selection Process

### 3.4 LLM Assessment Method

# Chapter 4

## Results

### 4.I Overview

### 4.2 Solutions/Generations

# Chapter 5

## Discussion

### 5.I Nature of Information Gathered

### 5.2 Continuous evaluation of experiment results

### 5.3 Comparison with related work

# Chapter 6

## Conclusion and Future Work

### 6.I Benefits and Impact

### 6.2 Limitations and Future Work

Validation and Evaluation of Synthetic Data in AI Research: Methods for validating the accuracy and usefulness of synthetic data. Techniques for evaluating the realism and applicability of AI-generated content.

Bias and Fairness in AI-Generated Content: Addressing bias in LLMs and its implications for generating sensitive scenarios. Approaches to ensuring fairness and avoiding harmful stereotypes in AI-generated data.

Use of LLMs in Simulating Criminal or Malicious Intent: Research on the use of AI to simulate scenarios involving criminal or malicious activities. Ethical and practical challenges in using AI for such purposes.

Future Directions in AI-Generated Synthetic Datasets: Emerging trends and future research opportunities in synthetic data generation using AI. Potential advancements in LLMs and their applications in creating more sophisticated datasets.

## Appendices

## References/Bibliography