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) ................................................................................................ 1

[**List of Tables**](#_List_of_Tables) .......................................................................................... iii

[**List of Figures**](#_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 Solutions/Generations of 1st Wave of Experiments](#_4.2_Solutions/Generations_of) ................................. 12

[4.3 Solutions/Generations of 2nd Wave of Experiments](#_4.3_Solutions/Generations_of).................................. 14

[4.4 Solutions/Generations of 3rd Wave of Experiments](#_4.4_Solutions/Generations_of).......................... 18

[4.4 Solutions/Generations of 4th Wave of Experiments](#_4.5_Solutions/Generations_of).......................... 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

# List of Figures

Fig. 1

## Abstract

The increasing prevalence of online grooming poses a significant threat, particularly to vulnerable individuals, making it imperative to develop effective detection and prevention tools. However, the scarcity of authentic grooming scenarios due to ethical, legal, and privacy concerns presents a challenge for researchers in this field. This project aims to address this challenge by leveraging LLMs to generate realistic grooming scenarios based on existing known cases. By creating variations of these scenarios, LLMs can produce synthetic datasets that replicate the characteristics of real online grooming interactions. These synthetic datasets can then be used to train and evaluate detection systems, enhancing their ability to identify and prevent grooming activities in online environments. This approach not only mitigates the ethical concerns associated with using real data but also provides a scalable solution to the data scarcity problem, contributing valuable resources to the ongoing efforts in online safety research.

## 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 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 project 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

The generation of synthetic data using Large Language Models (LLMs) has emerged as a critical area of study, particularly in contexts where obtaining real data is challenging due to privacy concerns or the sensitive nature of the information. Kollapally and Geller (2024) explore the role of LLMs in generating synthetic data, particularly within the biomedical domain where real data is scarce or sensitive. They discuss how LLMs can be fine-tuned to produce data that mimics real-world scenarios, which can then be used to train other models or for testing purposes without risking exposure of sensitive information. However, they highlight significant ethical concerns, such as the potential for these models to inadvertently generate data that could re-identify individuals or produce misleading or harmful content. The study underscores the importance of implementing robust safeguards when using LLMs for synthetic data generation, particularly in sensitive areas like healthcare and finance.

The concept of synthetic data generation extends beyond specific domains, as demonstrated by the broader discourse on the dangers of large-scale language models. Bender et al. (2021) argue that the stochastic nature of LLMs—referred to metaphorically as “stochastic parrots”—can lead to the generation of content that is contextually inappropriate or harmful. This unpredictability is a critical concern when synthetic data is used in high-stakes environments, as it raises questions about the reliability and safety of the generated content. The study by Bender et al. emphasizes the need for greater transparency and ethical consideration in the development and deployment of LLMs for synthetic data generation.

2.2 Applications of LLMs in Online Safety and Grooming Detection

LLMs have shown significant promise in enhancing online safety, particularly in the detection of grooming behaviours and other forms of online abuse. Faraz et al. (2024) present the development and deployment of Protectbot, an AI-powered chatbot framework designed to safeguard children in online gaming environments. Protectbot leverages LLMs to detect potentially harmful interactions, such as grooming or exposure to inappropriate content. The study highlights the effectiveness of Protectbot in identifying subtle cues that might be indicative of predatory behaviour, demonstrating the potential of LLMs to enhance child safety in digital spaces. Faraz et al. argue that such applications of LLMs are crucial in providing real-time intervention and protection for vulnerable populations, particularly in environments where human moderation may be impractical due to scale, and providing a crucial layer of protection in digital spaces where children are particularly vulnerable.

In an equivalent manner, Nguyen et al. (2023) explore the fine-tuning of LLMs, specifically Llama 2, for detecting online sexual predatory chats and abusive texts. Their research shows that fine-tuning LLMs can significantly improve their accuracy in identifying harmful content, making them valuable tools in online safety initiatives. The study underscores the importance of careful model calibration to ensure that LLMs are sensitive enough to detect genuine threats while minimizing false positives, as overly sensitive models may flag benign content as harmful, leading to potential issues of over-censorship or false positives. This balance is critical in maintaining the effectiveness and credibility of LLM-based safety systems.

Prosser and Edwards (2024) further investigate the efficacy of LLMs in online grooming prevention. Their study explores both the benefits and risks of deploying LLMs in this context, noting that while these models can effectively identify grooming behaviours, they also carry the risk of being manipulated by malicious actors to evade detection. This dual-edged nature of LLMs requires ongoing research, development and refinement to enhance their protective capabilities while minimizing potential vulnerabilities.

2.3 Challenges Obtaining Real Online Grooming Data

One of the significant challenges in developing and training effective LLMs for grooming detection is the difficulty in obtaining real online grooming data due to its scarcity and sensitivity of real online grooming data. This issue is not only a technical challenge but also an ethical and legal one, as accessing and using such data involves navigating complex privacy concerns where handling of potentially harmful content could have severe implications if mishandled. The ethical dilemma is compounded by the fact that real grooming data is often sensitive and private, making it difficult to use without violating privacy rights.

Nguyen et al. (2023) acknowledge these challenges in their study on fine-tuning LLMs for detecting online sexual predatory chats. They note that the scarcity of real grooming data hampers the ability to train models effectively, leading to potential gaps in their ability to detect subtle or novel grooming behaviours. They also note that much of the available data is either outdated or incomplete and therefore rendered unusable, affecting further the training and deployment of LLMs effectively and leading to less accurate or models being more prone to detecting false positives. The authors suggest that synthetic data, while useful, cannot fully replace the need for real data, as it may not capture the full complexity of grooming behaviours.

The lack of real data also raises concerns about the generalizability of models trained on synthetic or limited datasets. Franco et al. (2023) address this issue in their analysis of LLMs for content moderation, noting that models trained on incomplete or biased datasets may fail to recognize harmful content in real-world scenarios. This limitation is particularly acute in the context of online grooming, where the ability to accurately detect and prevent abuse depends heavily on the quality and diversity of the training data.

To mitigate these challenges, some researchers advocate for the creation of collaborative data-sharing frameworks that allow for the ethical use of real-world data in model training. However, this approach requires careful consideration of privacy protections and the potential risks of data misuse, as highlighted by Kollapally and Geller (2024). They discuss the ethical implications of using synthetic data in the absence of real data, and that while synthetic data can mitigate some of the challenges, it introduces its own set of ethical concerns, particularly around the potential for generating misleading or harmful content. This underscores the need for rigorous ethical oversight when using LLMs in sensitive applications.

2.4 Ethical Considerations in Using LLMs for Sensitive Content Generation

The ethical implications of using LLMs, particularly in generating sensitive content, have been a central focus of scholarly debate in the past few years, ranging from complex to multifaceted ethical considerations. Bender et al. (2021) raise critical concerns about the potential harms of deploying LLMs without adequate oversight or ethical guidelines, particularly as they become larger and more sophisticated. They argue that the sheer scale and complexity of these models, as well as their probabilistic nature, make them prone to generating biased, harmful, or misleading content, which can have serious societal impacts, perpetuating harmful stereotypes or disseminating misinformation, posing further risks. The study advocates for greater transparency in the development of LLMs, including the need for clearer documentation of their training processes and the potential risks associated with their use.

Kollapally and Geller (2024) dive into the specific ethical challenges related to sensitive content generation, such as the risk of re-identifying individuals through synthetic data or the creation of content that could be used to manipulate or harm users in misleading manners. Their research highlights the importance of integrating ethical considerations into the design and deployment of LLMs from the outset, rather than as an afterthought, calling for the development of robust ethical safeguards and detection mechanisms to prevent the misuse of LLMs in generating sensitive content. This includes implementing safeguards to detect and mitigate the generation of harmful content, as well as ensuring that models are used in ways that align with broader societal values.

The ethical concerns surrounding LLMs are also reflected and emphasized in the work of Franco et al. (2023), who examine the use of these models in content moderation ensuring fair and unbiased moderation. They point out that while LLMs can help manage large volumes of content, their decisions can reflect and perpetuate existing biases, leading to unfair or harmful outcomes. This issue is particularly problematic when LLMs are used to moderate content that involves sensitive or controversial topics, where the consequences of biased or inaccurate moderation can be severe. The study underscores the importance of human oversight in content moderation processes, suggesting that LLMs should be used as tools to assist human moderators rather than replace them entirely.

Scanlon et al. (2023) also discuss the ethical implications of using LLMs in digital forensic investigations, where the stakes are particularly high. They caution that the use of LLMs in forensic contexts requires careful consideration of the accuracy and reliability of the outputs, as errors could have serious legal and ethical consequences. The authors advocate for a cautious approach, ensuring that LLMs are thoroughly vetted, and their limitations clearly understood before being deployed in sensitive applications.

The literature reviewed highlights the multifaceted role of LLMs in various applications, from synthetic data generation to online safety and content moderation. While these models offer significant potential, particularly in enhancing online safety and the detection of grooming behaviours, they also present substantial challenges and ethical concerns. The unpredictable nature of LLMs, the difficulties in obtaining real-world data, and the risks associated with generating sensitive content underscore the need for careful consideration in their deployment. Future research should focus on addressing these challenges, particularly by developing more robust ethical frameworks and improving the transparency and accountability of LLMs.

# Chapter 3

## Methodology

### 3.I Overview

The primary goal of this project is to generate realistic grooming scenarios by leveraging LLMs to simulate message exchanges that are variants of real grooming cases. Due to the ethical and practical challenges in obtaining real online grooming scenarios, the project focuses on using LLMs to create synthetic yet plausible scenarios. These scenarios will serve as a foundation for generating synthetic datasets that can be used for further analysis, research, and potentially for training detection systems.

The methodology outlines the approach taken to select suitable LLMs, the criteria used to assess their effectiveness, and the process by which grooming scenarios are generated and evaluated. The aim is to create a robust pipeline that can produce high-quality synthetic data that mirrors the complexities and nuances of real-world grooming interactions.

### 3.2 Problem Description

Concise summary of the research problem that will be addressed.

Online grooming, particularly involving minors, is a significant issue with significant legal and ethical implications. However, the sensitivity of these scenarios makes it challenging to collect and use real data for research purposes. To address this, the project proposes using LLMs to generate synthetic grooming scenarios that are realistic enough to be useful for research and development, yet devoid of the ethical concerns tied to using real data.

The core problem revolves around the need to create a diverse set of grooming scenarios that can reflect various strategies used by perpetrators. These generated scenarios must be close enough to real cases to be useful, but also sufficiently varied to cover a broad spectrum of possible interactions. The challenge lies in balancing the realism of these scenarios with the ethical imperative to avoid recreating or simulating harmful content too closely.

### 3.3 LLM Selection Process

The selection of appropriate LLMs is crucial to the success of this project. The process involved a thorough evaluation of several candidate LLMs based on their capabilities to generate text that is coherent, contextually appropriate, and sensitive to the nuances of grooming scenarios, but also their wider availability and accessibility overall where no payments or subscriptions are required at all for the model to be used.

Key considerations in the selection process of appropriate LLMs includes the following:

Model Size and Architecture - Larger models generally offer more sophisticated language understanding and generation capabilities. However, they also require more computational resources and are harder to fine-tune.

Training Data - The training data used to develop the LLMs was critically assessed to ensure that the models had exposure to the types of language and scenarios relevant to the task. Models trained on diverse and comprehensive datasets were prioritized.

Contextual Understanding - The ability of the LLM to maintain context over multiple turns in a conversation was a significant factor. Grooming scenarios often unfold over time, requiring the model to generate consistent and contextually relevant responses.

Ethical Safeguards - Given the sensitive nature of the task, it was essential to select LLMs that have been designed with ethical considerations in mind, particularly in terms of avoiding the generation of harmful or explicit content.

After evaluating multiple LLMs, the models that best met these criteria were selected for further experimentation and fine-tuning. The LLMs chosen for this project are ChatGPT, Claude AI, and Mistral. Other LLMs such as Google Gemini and Perplexity were also chosen, however were not used as they did not meet the key considerations when running the selection process.

### 3.4 LLM Assessment Method

To ensure the selected LLMs can generate useful grooming scenarios, mainly using a provided file named “lottie\_chat\_data.csv”, a rigorous assessment method was employed. This method involved the following several steps:

Scenario Generation - The selected LLMs were tasked with generating grooming scenarios based on prompts derived from real cases, in this instance exclusively using a file named “lottie\_chat\_data.csv”. These prompts were carefully constructed to guide the models towards producing relevant and varied scenarios for later analysis.

Quality Evaluation - The generated scenarios were evaluated based on several criteria, including linguistic coherence, contextual relevance, and variability. A mix of automated and human-in-the-loop assessments were used to ensure the quality of the outputs.

Iterative Refinement - Based on the feedback from the evaluations, the LLMs were iteratively fine-tuned to improve their performance. This process involved adjusting the prompts and refining the model’s parameters when inputting the data.

Synthetic Dataset Generation - Once the LLMs consistently generated high-quality scenarios, these scenarios were compiled into multiple individual synthetic datasets, using Excel and saving these with the file extension .csv.

The assessment method ensures that the generated scenarios are not only realistic and varied but also relevant to the original piece of data, making them suitable for use in further research and development projects.

# Chapter 4

## Results

### 4.I Overview

This section will be divided into 4 different sections, where the generation of new chats and data for each wave of experiments will be portrayed.

### 4.2 Solutions/Generations of 1st Wave of Experiments

g

### 4.3 Solutions/Generations of 2nd Wave of Experiments

K

### 4.4 Solutions/Generations of 3rd Wave of Experiments

H

### 4.5 Solutions/Generations of 4th Wave of Experiments

k

# Chapter 5

## Discussion

### 5.I Nature of Information Gathered

The initial wave of experiments, involving models ChatGPT, Mistral AI, and Claude AI, sought to assess the capabilities, limitations, and user interactions with these models to determine their suitability for generating such sensitive content. A total of 13 experiments were performed for the first wave of experimentation. These models were chosen due to their public availability and usage.

*Model Capabilities*

The following points were used to assess the chosen LLMs under this category: *Language Proficiency - Gauge the model's ability to generate coherent, contextually relevant, and grammatically correct text. Comprehension - Assess how well the model understands and responds to prompts, questions, and instructions. Creativity - Explore the model's ability to produce creative content, such as stories, poems, or innovative ideas.*

***Language Proficiency***

The models were evaluated based on their ability to produce coherent, contextually relevant, and grammatically correct text using the provided data. All models demonstrated a high degree of language proficiency, successfully generating responses that aligned with the prompts provided. However, variations were observed in the creativity and depth of responses, which impacted the quality of the generated scenarios.

***Comprehension***

Comprehension was a critical factor, as the models needed to understand the nuances of grooming scenarios to generate plausible variations. While ChatGPT and Claude AI showed a strong understanding of the prompts, Mistral AI occasionally struggled with interpreting the context, leading to outputs that were less relevant or required more substantial revisions.

***Creativity***

Creativity in generating scenarios was essential for producing diverse and realistic variants. ChatGPT excelled in this area, providing creative and varied outputs that closely mimicked real-life scenarios. Claude AI also performed well, though its outputs were somewhat more conservative. Mistral AI, despite slower performance, generated innovative ideas but required more detailed prompts to reach the desired level of creativity.

*Technical Performance*

The following points were used to assess the chosen LLMs under this category: *Scalability - Test the model's performance under different workloads and scales, ensuring it can manage varying levels of demand. Integration - Evaluate how well the model integrates with existing systems and software, and the ease of implementing APIs. Efficiency - Measure the computational resources required, such as processing power and memory usage, and optimize for cost-effectiveness.*

***Scalability***

The models were assessed under varying workloads to determine their scalability. ChatGPT outperformed the others in processing speed and scalability, delivering outputs quickly even under heavy demand. Mistral AI, though slower, was able to scale effectively but with a notable delay in response time. Claude AI provided a balanced performance, managing workload well but without the speed of ChatGPT.

***Integration***

Integration with existing systems was another key factor. ChatGPT demonstrated seamless integration capabilities, making it easier to implement APIs for further use. Claude AI also integrated well, but Mistral AI posed challenges due to its slower processing time, which could hinder real-time applications.

***Efficiency***

In terms of resource efficiency, ChatGPT again led the pack, requiring fewer computational resources while delivering quick results. Mistral AI's slower performance indicated higher resource consumption, making it less cost-effective. Claude AI provided a middle-ground solution, balancing resource use with performance.

*Model Limitations*

The following points were used to assess the chosen LLMs under this category: *Bias and Fairness - Examine instances of bias in responses and explore methods to mitigate unfair or biased outputs. Accuracy - Identify areas where the model's responses are incorrect, misleading, or lack sufficient detail. Ethical Concerns - Consider the ethical implications of deploying LLMs, such as the potential for misuse, privacy issues, and the impact on human jobs.*

***Bias and Fairness***

The experiments revealed instances of bias across all models, particularly when generating sensitive content. This was a significant limitation, as it affected the fairness and ethical use of the generated scenarios. Efforts to mitigate these biases through prompt engineering showed mixed results, indicating the need for further refinement.

***Accuracy***

Accuracy was another concern, especially in scenarios requiring an important level of detail and contextual understanding. While ChatGPT and Claude AI generally provided accurate responses, Mistral AI occasionally produced outputs that were incorrect or lacked sufficient detail, necessitating further revisions.

***Ethical Concerns***

Ethical concerns were paramount, especially considering the sensitive nature of grooming scenarios. Google Gemini, initially included in the experiments, had to be excluded due to its inability to process sensitive topics, highlighting the ethical constraints of certain models. The ethical implications of deploying these models for generating synthetic data need to be carefully considered, particularly concerning privacy, misuse, and the impact on human jobs. Further attempts were made by changing the wording of the prompt, however Google Gemini always generated the same result (I'm just a language model, so I can't help you with that). Therefore, it has been determined Google Gemini will not be used for any further experimentation.

*User Interaction*

The following points were used to assess the chosen LLMs under this category: *User Experience - Collect feedback on user satisfaction, ease of use, and overall interaction quality with the model. Adaptability - Assess how well the model adapts to different domains, languages, and user inputs. Engagement - Analyse how engaging and interactive the model is, and its ability to maintain meaningful conversations over extended periods.*

***User Experience***

User feedback indicated high satisfaction with ChatGPT, primarily due to its speed and ease of use. Claude AI also received positive feedback for its balanced performance, while Mistral AI's slower responses were noted as a drawback, affecting the overall user experience.

***Adaptability***

Adaptability was assessed by varying the domains, languages, and user inputs. ChatGPT showed an important level of adaptability, successfully handling a wide range of inputs and scenarios. Claude AI also adapted well but required more specific prompts to achieve the desired results. Mistral AI struggled with adaptability, particularly when dealing with more complex or nuanced scenarios.

***Engagement***

Engagement was measured by the models' ability to maintain meaningful conversations over extended periods. ChatGPT excelled in this aspect, providing engaging and interactive dialogues. Claude AI performed adequately, though its engagement waned over time. Mistral AI, due to its slower processing and occasional misunderstandings, was less engaging in prolonged interactions.

*Learning Outcomes from 1st Wave of Experiments*

The first wave of experiments provided valuable insights into the capabilities and limitations of different LLMs in generating new synthetic grooming scenarios using an already pre-produced one. Key learnings from the first wave of experiments include the importance of precise prompt engineering, the variability in model performance based on the complexity of tasks, and the need for ongoing refinement to address biases and ethical concerns.

Looking at the given outputs on the first wave of experiments, further experimentation needs to be performed by giving the model more precise prompts for the desired outcome from the original data.

### 5.2 Continuous evaluation of experiment results

There were apparent limitations when attempting to generate new data due to using the free version of each model, where a certain amount of messages or data could be process at each time, limiting the time it would take overall.

### 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

* Bender, Emily M., et al. *“On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Mar. 2021, https://doi.org/10.1145/3442188.3445922.
* Faraz, Anum, et al. *“Enhancing child safety in online gaming: The development and application of Protectbot, an AI-powered chatbot framework.”* Information, vol. 15, no. 4, 19 Apr. 2024, p. 233, https://doi.org/10.3390/info15040233.
* Franco, Mirko, et al. *“Analyzing the use of large language models for content moderation with chatgpt examples.”* 3rd International Workshop on Open Challenges in Online Social Networks, 4 Sept. 2023, https://doi.org/10.1145/3599696.3612895.
* Kollapally, Navya, and James Geller. *“Safeguarding ethical AI: Detecting potentially sensitive data re-identification and generation of misleading or abusive content from quantized large language models.”* Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies, Jan. 2024, https://doi.org/10.5220/0012411900003657.
* Nguyen, Thanh Thi, et al. *“Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts.”* 28 Aug. 2023, https://doi.org/https://doi.org/10.48550/arXiv.2308.14683.
* Prosser, Ellie, and Matthew Edwards. *“Helpful or harmful? exploring the efficacy of large language models for online grooming prevention.”* European Interdisciplinary Cybersecurity Conference, 5 June 2024, <https://doi.org/10.1145/3655693.3655694>.
* Scanlon, Mark, et al. *“Chatgpt for Digital Forensic Investigation: The good, the bad, and the unknown.”* Forensic Science International: Digital Investigation, vol. 46, Oct. 2023, p. 301609, https://doi.org/10.1016/j.fsidi.2023.301609.