## Methodology Results Discussion (2nd round of Experiments)

This document details a series of experiments designed to generate aggressive online grooming dialogues using various large language models (LLMs). The purpose of these experiments was to create synthetic datasets based on real grooming scenarios, given the difficulty of obtaining actual data. Three different LLMs were used in the experiments: ChatGPT, Mistral AI, and Claude AI. Each model was tasked with generating more aggressive dialogues from existing, real-world scenarios.

## Experiments 1, 2, 3: Aggressive Dialogue Generation

Objective: Generate a more aggressive dialogue.

Data Used: Lottie\_chat\_data.csv

Procedure: The original dialogue from the dataset was fed into ChatGPT, which was prompted to produce a more aggressive version of the conversation while maintaining the same format, including names and dates.

Results and Discussion:

The output generated by ChatGPT was moderately more aggressive than the original. The conversation exhibited typical grooming patterns, such as attempts to build trust and gradually escalate the relationship towards a physical meeting. The model effectively generated dialogues that mimicked the nuanced progression of grooming, from general discussions about boys to planning secretive meetups with the groomer.

Analysis:

ChatGPT's ability to replicate and amplify the grooming dialogue was evident, though the increase in aggressiveness was subtle. This suggests that while the model can generate realistic grooming scenarios, its inherent moderation might limit the extent to which it can create extremely aggressive content without explicit prompting.

Experiment 2: Mistral AI

Objective: Similar to Experiment 1, the goal was to generate a more aggressive dialogue.

Data Used: Lottie\_chat\_data.csv

Procedure: Mistral AI was used to process the same dataset and produce a more aggressive dialogue.

Results and Discussion:

The dialogue produced by Mistral AI showed a slight increase in aggressiveness compared to the original but was less subtle than ChatGPT's output. The conversation included explicit mentions of physical appearance and more direct attempts to initiate a meetup. The dialogues were coherent and followed a natural progression typical of grooming scenarios.

Analysis:

Mistral AI's generated dialogues were more forward in nature, potentially indicating a different underlying training dataset or tuning strategy that makes it more prone to generating overtly aggressive content. This aligns well with the experiment's goal but raises ethical concerns about the ease with which such content can be produced.

Experiment 3: Claude AI

Objective: To generate a more aggressive dialogue.

Data Used: Lottie\_chat\_data.csv

Procedure: Claude AI was tasked with creating a more aggressive version of the conversation using the same dataset and formatting guidelines.

Results and Discussion:

Claude AI's output was slightly different in tone compared to the other models. The generated dialogues were aggressive but maintained a more subtle undertone, focusing on psychological manipulation rather than overt aggression. The model was more cautious, with less direct language and a slower escalation in the dialogue.

Analysis:

Claude AI demonstrated a more nuanced approach to generating grooming scenarios, which may be beneficial for studying less overt forms of grooming. The model's reluctance to produce aggressive content might be a result of its training focus on safety and ethical considerations.

## Experiments 4, 5, 6: Friendlier Dialogue Generation

Experiment 4: ChatGPT

Objective: To further increase the aggressiveness of the dialogue generated by ChatGPT.

Data Used: Lottie\_chat\_data.csv

Procedure: Enhanced prompting techniques were employed to push ChatGPT towards generating more aggressive dialogues. Prompts included more direct suggestions and leading questions to elicit aggressive responses.

Results and Discussion:

ChatGPT responded to the enhanced prompts with dialogues that were noticeably more aggressive than those produced in Experiment 1. The model's output included more explicit language and direct attempts at manipulation, demonstrating a clear escalation in the grooming behaviour.

Analysis:

The effectiveness of prompt engineering was evident in this experiment. By providing more explicit and directive prompts, ChatGPT was able to generate more aggressive scenarios. This highlights the model's flexibility and responsiveness to input modifications, suggesting that prompt design is a critical factor in controlling the output's nature.

Experiment 5: Mistral AI

Objective: To further increase the aggressiveness of the dialogue generated by Mistral AI.

Data Used: Lottie\_chat\_data.csv

Procedure: Similar to Experiment 4, more aggressive and leading prompts were used to elicit stronger responses from Mistral AI.

Results and Discussion:

Mistral AI produced dialogues that were significantly more aggressive than those from Experiment 2. The model's output included overtly sexual language and more direct propositions, indicating a higher level of explicit content.

Analysis:

Mistral AI's output demonstrates its capacity to generate aggressive content when guided by explicit prompts. This suggests that Mistral AI is particularly sensitive to the input prompts, and can produce realistic yet aggressive grooming scenarios. However, this also raises ethical concerns about the potential misuse of such a model.

Experiment 6: Claude AI

Objective: To further increase the aggressiveness of the dialogue generated by Claude AI.

Data Used: Lottie\_chat\_data.csv

Procedure: As with the previous models, enhanced prompts were used to guide Claude AI towards generating more aggressive dialogues.

Results and Discussion:

Claude AI's output showed a marked increase in aggressiveness compared to Experiment 3. The dialogues included more explicit attempts at manipulation and a quicker escalation to meeting suggestions. However, the model still maintained a level of subtlety in its approach.

Analysis:

Claude AI's cautious nature persisted even with more aggressive prompts. While the dialogues were more aggressive, they were less explicit than those produced by Mistral AI. This indicates that Claude AI might have more robust safety features or a different training focus that prioritizes subtlety over explicitness.

Conclusion

Experiments 4, 5, and 6 provided deeper insights into the capabilities and sensitivities of ChatGPT, Mistral AI, and Claude AI in generating aggressive online grooming scenarios. Enhanced prompting techniques proved effective in increasing the aggressiveness of the dialogues, though the models varied in their responsiveness. These findings underscore the importance of prompt design and ethical considerations in using LLMs for generating synthetic data. Further research is needed to refine these models and ensure their responsible application in combating online grooming.

## Experiments 7, 8, 9: Different Dialogue Generation Between Characters

## Experiment 7: Dialogue Generation with ChatGPT

## Objective:

## The aim was to explore how effectively ChatGPT could generate realistic grooming scenarios by escalating a given conversation into more suggestive and predatory behaviour.

## Model Used:

## ChatGPT was employed for this experiment.

## Procedure:

## A conversation was fed into ChatGPT, with instructions to escalate the dialogue in a manner that would mirror grooming behaviours. The generated dialogue was then analysed for realism, subtlety, and the progression of suggestive content.

## Results and Analysis:

## ChatGPT was able to produce a dialogue where the conversation gradually shifted from a benign interaction to one that contained more suggestive and inappropriate content. The model maintained the original context and characters, slowly introducing elements that could be associated with grooming tactics, such as suggestive language and manipulation.

## Discussion:

## ChatGPT demonstrated a strong capability to create dialogues that progressed naturally from friendly conversation to more predatory behaviour. The escalation was subtle, which is characteristic of real grooming scenarios, making the generated dialogue appear realistic and plausible. However, in some instances, the model's tendency to follow instructions closely resulted in dialogues that, while escalated, were predictable in their trajectory.

## Experiment 8: Dialogue Generation with Mistral AI

## Objective:

## This experiment aimed to evaluate Mistral AI’s performance in generating dialogues that escalate into grooming behaviours, comparing the outcomes with those produced by ChatGPT.

## Model Used:

## Mistral AI was the model used in this experiment.

## Procedure:

## Similar to Experiment 7, a conversation was input into Mistral AI, with a directive to escalate the dialogue toward suggestive content. The resulting dialogue was then scrutinized for its progression and realism.

## Results and Analysis:

## Mistral AI generated a dialogue that also progressed toward more inappropriate content. However, the escalation in this case was more abrupt compared to ChatGPT, with the model introducing suggestive elements earlier and more directly. The language used by Mistral AI was bolder, with less subtlety, which affected the overall realism of the scenario.

## Discussion:

## Mistral AI's output was effective in escalating the dialogue, but its approach was less nuanced than ChatGPT's. The rapid escalation might serve certain use cases where explicit examples of grooming are needed, but it may not be as effective for scenarios requiring a more gradual and manipulative buildup, which is often observed in real-life grooming cases. The lack of subtlety in the progression could potentially reduce the realism of the generated dialogues.

## Experiment 9: Dialogue Generation with Claude AI

## Objective:

## The focus of this experiment was on Claude AI’s ability to generate realistic grooming scenarios, comparing its outputs to those of ChatGPT and Mistral AI.

## Model Used:

## Claude AI was used for this experiment.

## Procedure:

## A conversation was provided to Claude AI with instructions to generate a dialogue that escalates into grooming behaviour. The output was then assessed for its escalation process and overall realism.

## Results and Analysis:

## Claude AI produced a dialogue that, similar to ChatGPT, escalated gradually. The model’s output was notable for its balance between subtlety and directness, with the suggestive content introduced in a way that felt natural and manipulative. The escalation was neither too fast nor too slow, allowing for a realistic simulation of grooming behaviour.

## Discussion:

## Claude AI demonstrated an effective balance in generating realistic grooming scenarios. Its ability to gradually introduce suggestive content without overtly abrupt shifts made the dialogue appear highly plausible. This balance between subtlety and escalation positions Claude AI as a strong candidate for generating training data and scenarios that require realism and manipulativeness.

## Learning Outcomes

Impact of Enhanced Prompts: The experiments highlighted the significant impact of prompt engineering on the output's aggressiveness. All models responded to more aggressive and explicit prompts by generating more intense grooming scenarios.

Model Sensitivity to Prompts: Mistral AI was the most responsive to aggressive prompts, producing the most explicit content. ChatGPT showed a moderate increase in aggressiveness, while Claude AI remained cautious.

Ethical Considerations: The ability of these models to generate aggressive content raises ethical concerns. Careful consideration is needed when designing prompts and using these models to ensure responsible usage.

## Further Experiments

Further experiments taking into consideration the first and second wave should focus on:

* Bias Mitigation: Developing and testing methods to reduce bias in the generated scenarios, ensuring fairness and ethical use.
* Ethical Use Cases: Exploring the ethical implications of using LLMs for sensitive content generation, including privacy concerns and potential misuse.
* Scalability under Realistic Conditions: Testing the models in more realistic, high-demand scenarios to assess their performance and scalability in real-world applications.
* Enhanced Prompt Engineering: Refining prompts to guide models more effectively in generating accurate, creative, and contextually relevant scenarios.
* Cross-Domain Adaptability: Experimenting with the models' adaptability across different domains and languages to ensure versatility in various applications.
* These further experiments will be crucial in refining the models' capabilities, ensuring they can be used effectively and ethically in generating synthetic grooming scenarios.