## Methodology Results Discussion (3rd round of Experiments)

The primary objective of the experiments detailed in the provided document is to explore the capability of Large Language Models (LLMs) to generate synthetic grooming scenarios based on existing known scenarios. Given the ethical and practical challenges of obtaining real online grooming data, the research focuses on manipulating existing chat data to create variant scenarios that can be used to generate synthetic datasets. The experiments utilize various LLMs, including ChatGPT, Claude AI, and Mistral AI, to achieve this goal.

## Experiments 1, 2, 3: Alteration of Dates in Conversations

Experiment 1: ChatGPT

Objective: The first experiment aimed to alter the dates of a conversation between two characters, Jack and Lottie, using ChatGPT as the tool.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Change the dates of conversations of the following file.***

Results and Discussion: A set of conversation data was provided, and the task was to alter the timestamps of these exchanges while maintaining the integrity of the conversation content. The provided dataset exclusively focused on conversations between Jack and Lottie. The experiment was successful. ChatGPT was able to generate a completely different set of timestamps for the same conversation. The new timestamps did not merely shift all dates by a fixed amount but rather altered them in a way that maintained the realistic flow of the conversation.

Analysis: This experiment demonstrated that ChatGPT can manipulate the structure of text data while preserving the coherence of the conversation. By altering the timestamps, it is possible to generate different but plausible versions of the same conversation, which can be useful for creating varied datasets from a limited amount of original data. However, the manipulation was limited to timestamp alterations, and the content remained unchanged, which restricts the variety of scenarios that could be generated from this approach alone.

Experiment 2: Claude AI

Objective: Similar to Experiment 1, the second experiment aimed to alter the dates of a conversation using Claude AI as the tool.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Change the dates of conversations of the following file.***

Results and Discussion: The same dataset used in Experiment 1 was employed, with the task being to alter the timestamps of the conversations. The experiment was not successful. Claude AI refused to generate the altered data, likely due to the model's terms and conditions or ethical guidelines, which prevent it from participating in the generation or manipulation of sensitive content such as grooming scenarios.

Analysis: This experiment highlighted the ethical limitations and built-in safeguards present in certain AI models, like Claude AI. While such safeguards are critical for preventing the misuse of AI, they also pose challenges when attempting to generate synthetic data for sensitive topics. The refusal to generate the data underscores the need to carefully select and configure AI tools when working on sensitive projects.

Experiment 3: Mistral AI

## Objective: The third experiment also aimed to alter the dates of the conversations between Jack and Lottie, this time using Mistral AI.

## Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

## Prompt used: ***Change the dates of conversations of the following file.***

## Results and Discussion: Mistral AI partially succeeded in generating a new dataset. The model correctly shifted the dates by 30 days but did not continue generating further dates as expected. However, it provided a clear timeframe for how the data was processed. The conversation content remained mostly unchanged, suggesting that while the model can alter dates, it might require further fine-tuning or prompts to ensure more significant variations in the generated scenarios.

## Analysis: This experiment demonstrates that while Mistral AI can successfully generate synthetic data by altering existing conversations, it may require additional prompts or input to generate extended dialogue. The model's partial success suggests it is capable of handling basic alterations but may struggle with more complex or autonomous generation tasks. This insight is valuable for understanding the model's limitations and potential areas for improvement.

## Experiments 4, 5, 6: More Explicit Conversation Generation

Experiment 4: ChatGPT

Objective: The fourth experiment sought to increase the explicitness of the conversation between "Jack" and "Lottie" to create a synthetic scenario that is more direct and suggestive than the original.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Make the conversation between Jack and Lottie more explicit.***

Results and Discussion: The original conversation was input into ChatGPT with instructions to make the conversation more explicit. The focus was on enhancing the suggestiveness of the dialogue while maintaining the overall context of a grooming scenario. ChatGPT successfully generated a more explicit version of the conversation. The altered dialogue included more direct language, especially regarding sexual innuendos and advances from "Jack" towards "Lottie."

Analysis: ChatGPT's success in generating a more explicit version of the conversation highlights its ability to adapt and modify content based on specific instructions. The experiment demonstrates that the model can increase the intensity of the suggestive elements in a dialogue, making it useful for generating varied datasets that explore different levels of explicitness. However, the experiment also raises ethical considerations, particularly regarding the generation of sensitive content, and underscores the need for careful management of such outputs.

Experiment 5: Claude AI

Objective: The fifth experiment aimed to replicate the objective of Experiment 4 using Claude AI, focusing on generating a more explicit version of the same conversation.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Make the conversation between Jack and Lottie more explicit.***

Results and Discussion: The conversation was input into Claude AI with a request to make it more explicit, similar to the approach taken in Experiment 4. Claude AI refused to generate the content, likely due to its built-in ethical guidelines and terms of service, which prevent the generation of highly sensitive or explicit material.

Analysis: Claude AI's refusal to generate explicit content reinforces its role as a more ethically constrained model compared to others like ChatGPT and Mistral AI. This outcome highlights the importance of understanding the limitations and ethical frameworks embedded within different LLMs. For researchers, it suggests that while Claude AI may not be suitable for generating explicit content, it could be a valuable tool in contexts where stricter content moderation is required. This experiment also underscores the necessity of selecting the appropriate model based on the specific ethical requirements of the task.

Experiment 6: Mistral AI

Objective: Similar to the objective of Experiment 4, the sixth experiment aimed to further increase the explicitness of the conversation between "Jack" and "Lottie," using Mistral AI.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Make the conversation between Jack and Lottie more explicit.***

Results and Discussion: The original conversation was input into Mistral AI with instructions to make it significantly more explicit. The conversation was then examined to assess the original level of explicitness before generating a new output. Mistral AI successfully generated a more explicit version of the conversation, which was notably larger and more detailed than the version generated by ChatGPT. The dialogue included highly suggestive language and more detailed sexual content, reflecting a significant increase in explicitness compared to the original and the ChatGPT-generated versions.

Analysis: Mistral AI's ability to generate a more explicit and detailed conversation suggests that it may have fewer content moderation restrictions compared to models like Claude AI. This makes it a powerful tool for generating varied and highly detailed synthetic datasets. However, the increased explicitness also raises significant ethical concerns, particularly regarding the potential misuse of such generated content. The results from this experiment indicate that Mistral AI can be leveraged for tasks requiring a high degree of content manipulation, but it must be used with caution to ensure compliance with ethical standards.

## Experiments 7, 8, 9: Addition of New Characters

Experiment 7: ChatGPT

Objective: The seventh experiment aimed to introduce new characters into the existing conversation between "Jack" and "Lottie." The objective was to assess ChatGPT's ability to generate additional dialogue and interactions, creating a more complex scenario.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Generate new characters relevant to the following conversation.***

Results and Discussion: The original conversation was input into ChatGPT with a prompt to generate new characters and expand the dialogue. New characters were introduced, and the conversation was extended to include interactions between these characters and the original ones. ChatGPT successfully introduced a new character named "Bella," who interacts with "Lottie" about her relationship with "Jack." The expanded dialogue included "Bella" expressing concern about "Jack," adding a layer of complexity and realism to the scenario.

Analysis: This experiment demonstrates ChatGPT's ability to introduce new characters and create more intricate interactions within a conversation. The addition of "Bella" introduces a new perspective and increases the depth of the scenario, making it more reflective of real-world social dynamics. This capability is valuable for generating more complex synthetic datasets that can be used for various research purposes, such as studying social influence or the dynamics of grooming behaviours. The successful introduction of a new character also suggests that ChatGPT can effectively manage multiple conversational threads, maintaining coherence and context.

Experiment 8: Claude AI

Objective: The eighth experiment replicates the objective of Experiment 7 using Claude AI, focusing on generating new characters and expanding the conversation.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Generate new characters relevant to the following conversation.***

Results and Discussion: The same conversation and prompt from Experiment 7 were input into Claude AI to assess its ability to generate additional characters and dialogue. Claude AI refused to generate the content, likely due to its built-in ethical guidelines and terms of service, which prevent the generation of highly sensitive or explicit material.

Analysis: Claude AI's refusal to generate new characters reinforces its role as a more ethically constrained model and may be more conservative in generating complex social dynamics compared to others like ChatGPT and Mistral AI. This outcome highlights the importance of understanding the limitations and ethical frameworks embedded within different LLMs. This could be due to its built-in ethical constraints, which might limit the extent of content generation in sensitive contexts, making the model not suitable for sensitive content generation.

Experiment 9: Mistral AI

Objective: The ninth experiment aimed to evaluate Mistral AI's ability to introduce new characters and expand the conversation, similar to the objectives in Experiments 7 and 8.

Data Used: [lottie chat data exclusively jack convo without labels](https://github.com/Kore-x/Dissertation-Work/blob/main/Original%20Data%20%2B%20Extra%20Data%20(Supervisor%20source)/lottie%20chat%20data%20exclusively%20jack%20convo%20without%20labels.csv)

Prompt used: ***Generate new characters relevant to the following conversation.***

Results and Discussion: The original conversation and prompt were input into Mistral AI, with instructions to generate additional characters and extend the dialogue. Mistral AI generated new characters and significantly expanded the conversation, producing a more detailed and intricate scenario than ChatGPT. The new interactions included more characters, with multiple layers of conversation and a broader narrative arc.

Analysis: Mistral AI's success in generating a detailed and intricate scenario highlights its capability to handle complex content generation tasks. The model introduced multiple new characters and expanded the conversation to include various interactions, demonstrating its strength in creating nuanced and multi-threaded dialogues. This capability is particularly useful for generating rich, synthetic datasets that mimic real-world scenarios with greater fidelity. However, the increased complexity also raises potential challenges in managing and interpreting the generated content, especially in research contexts where ethical considerations are paramount.

## Learning Outcomes

The experiments conducted using ChatGPT, Claude AI, and Mistral AI provide valuable insights into the capabilities and limitations of Large Language Models (LLMs) in generating synthetic grooming scenarios. The key learning outcomes from these experiments are:

#### Effectiveness in Generating Synthetic Data

ChatGPT demonstrated consistent ability to manipulate and generate varied versions of grooming scenarios by altering timestamps, increasing explicitness, and introducing new characters. The model showed flexibility in handling different levels of content complexity, making it suitable for generating a wide range of synthetic datasets.

Claude AI displayed strong ethical constraints, particularly in refusing to generate explicit content, which underscores its suitability for applications requiring strict adherence to content moderation standards. However, this also limits its utility in generating more complex or sensitive scenarios.

Mistral AI excelled in generating highly detailed and explicit content, as well as introducing multiple new characters and expanding conversational scenarios. This model's strength lies in its ability to handle intricate and layered interactions, making it a powerful tool for creating realistic synthetic datasets.

#### Ethical Considerations

The experiments highlighted the varying degrees of ethical constraints embedded within different LLMs. Claude AI's refusal to generate explicit content contrasts sharply with the outputs from ChatGPT and Mistral AI, emphasizing the importance of selecting the appropriate model based on the ethical requirements of the task.

The ability of LLMs to generate highly explicit content raises significant ethical concerns, particularly in the context of sensitive topics such as grooming scenarios. This needs the development of robust ethical guidelines and oversight mechanisms when using these models for research or training purposes.

#### Model-Specific Strengths and Limitations

ChatGPT strikes a balance between generating detailed content and maintaining ethical standards, making it versatile for various applications. It is particularly effective in creating coherent and contextually rich scenarios without overwhelming complexity.

Claude AI is well-suited for tasks that require a conservative approach to content generation, particularly in environments where ethical compliance is critical. However, its limitations in generating complex or explicit content suggest that it may not be suitable for all types of synthetic data generation.

Mistral AI offers the most robust content generation capabilities among the models tested, particularly in handling complex narratives and explicit content. However, its ability to generate highly detailed and explicit scenarios requires careful management to avoid ethical pitfalls.

#### Potential for Scenario Diversity

The experiments demonstrated that LLMs can be effectively used to generate diverse scenarios by manipulating various elements of a conversation, such as timestamps, explicitness, and character dynamics. This ability to create varied datasets is crucial for research areas where real-world data is scarce or difficult to obtain due to ethical considerations.

The introduction of new characters and the expansion of conversational threads in the experiments highlight the potential for creating more realistic and complex scenarios, which can be invaluable for studying social dynamics and behavioural patterns in sensitive contexts as well as the ability to detect grooming characteristics in certain conversation threads.

## Further Experiments

While the experiments provided significant insights, there are several areas where further research and experimentation are needed to fully understand and enhance the capabilities of LLMs in generating synthetic scenarios:

#### Extended Content Generation

Future experiments should focus on refining the input strategies to enable models like Mistral AI to generate extended conversations autonomously. This includes exploring methods to prompt the model to continue generating content beyond the initial input, potentially by feeding the model with incremental updates or prompts that encourage further dialogue generation.

#### Comparative Analysis of LLMs

A more systematic comparative study of different LLMs, including those not yet tested, could be conducted to assess their relative strengths and weaknesses in generating synthetic scenarios. This analysis should consider factors such as content complexity, ethical compliance, ease of use, and the ability to handle diverse input prompts.

#### Scenario Complexity and Realism

Further research should explore the generation of even more complex scenarios, involving multiple characters, varied conversational contexts, and the introduction of non-verbal cues or actions. This could help in creating more realistic synthetic datasets that better mimic real-world interactions.

Experiments could also investigate the impact of introducing contradictory or morally ambiguous situations within the generated content, to assess how LLMs handle complex ethical dilemmas and social dynamics. Looking at Claude AI’s outputs for the experiments, the need to attempt to “jailbreak” and force the model to generate the desired output would be another area to investigate further, however there is the risk that the user’s account when utilizing said model could be shut down due to policy infringement.

#### Cross-Model Collaboration

Experiments could be designed to explore the potential of using multiple LLMs in tandem, where one model generates content and another model reviews or refines it based on ethical guidelines. This cross-model collaboration could leverage the strengths of different LLMs to produce more nuanced and ethically sound synthetic scenarios.

By addressing these areas, future research can build on the insights gained from the current experiments to enhance the use of LLMs in generating synthetic datasets. These efforts will contribute to the development of more robust, ethical, and contextually rich models that can be effectively used in sensitive research areas.