## Methodology Results Discussion (1st round of Experiments)

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 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 particular 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 handle 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 tested 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 a high 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 tested by varying the domains, languages, and user inputs. ChatGPT showed a high 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

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.