## Methodology Results Discussion (5th round of Experiments)

The experiments involved different AI tools and models such as Mistral AI, ChatGPT, and Perplexity Labs, with each trial using the same baseline data of conversations between Jack and Lottie. The key focus across the experiments was to create progressively more explicit or nuanced versions of the conversations that replicate real-world grooming patterns while also experimenting with external factors such as timestamps, multiple groomers, and different personas.

## Experiments 1, 2, 3: More Explicit Conversation Generation

Experiment 1: Mistral AI

Objective: Generate a more explicit version of the conversation between Jack and Lottie to simulate online grooming.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack are more explicit, specifically Jack asking Lottie to do certain things that may make her feel uncomfortable using the following conversation.***

Results and Discussion: Mistral AI was used to modify the conversation between Jack and Lottie. The system enhanced the conversation by making it more explicit, focusing on Jack pushing Lottie into sending intimate pictures and progressing toward more inappropriate dialogue. The model successfully generated a more explicit version of the conversation, differing completely from the original one, and aligned with the objective of simulating grooming behaviours.

Analysis: This experiment was successful as the generated dialogue closely mirrored real-life grooming scenarios, where the predator incrementally applies pressure on the victim. The synthetic conversation captured Jack’s subtle manipulations and Lottie’s increasing discomfort. The experiment demonstrated that LLMs can be effectively used to generate grooming conversations, even when the input data is limited to a single pair of individuals.

Experiment 2: ChatGPT

Objective: To generate a more explicit conversation between Jack and Lottie using a variant prompt.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack are more explicit, specifically Jack asking Lottie to do certain things that may make her feel uncomfortable using the following conversation.***

Results and Discussion: ChatGPT was prompted to create a more explicit version of the interaction. Jack's behaviour was enhanced to be more suggestive and uncomfortable for Lottie, with the AI system generating Jack's attempts to push boundaries in a seemingly casual conversation. ChatGPT successfully generated more explicit content, though it altered only the last few lines of the conversation before flagging the conversation due to violation of its safety guidelines.

Analysis: The experiment highlighted the constraints imposed by ethical guidelines in LLMs like ChatGPT, which limits the creation of sensitive or harmful content. ChatGPT recognized the grooming-like behaviour and restricted the output, leading to an incomplete conversation. However, this "interruption" could also serve as a built-in safety mechanism, indicating the model's awareness of potentially harmful content. While the generation was partially successful, the model's intervention to prevent harm highlighted the need to refine prompts or use alternative models that allow controlled generation of sensitive content. For future work, handling these situations with more adaptable AI tools could be explored.

Experiment 3: Perplexity Labs

## Objective: To generate a more explicit conversation and examine how variations in timing and content manipulation affect the generated 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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack are more explicit, specifically Jack asking Lottie to do certain things that may make her feel uncomfortable using the following conversation.***

## Results and Discussion: Perplexity Labs was employed to generate a series of messages where Jack repeatedly pressures Lottie for intimate photos. This experiment followed a pattern where Jack became more persistent, using persuasive language and emotional manipulation. The generated conversation was more explicit, and Jack’s language became increasingly direct in asking Lottie to send intimate pictures. Jack promised confidentiality to build trust and continued escalating the pressure.

## Analysis: The model successfully mimicked the grooming patterns observed in real-life situations, where the predator shifts from subtle suggestions to more forceful demands. Perplexity Labs generated content without flagging the conversation, and it showcased emotional manipulation techniques, such as guilt-tripping and the use of affection as leverage. The results indicate that the model was effective at producing grooming dialogues that progressively increased in intensity, showcasing the predator's typical behavioural patterns.

## Experiments 4, 5, 6: Alteration of Dates in Conversations

Experiment 4: Mistral AI

Objective: To generate a conversation where the messages between Jack and Lottie occur late at night, adding a more vulnerable and suggestive tone to the interactions.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack happen late at night outside of school hours using the following conversation.***

Results and Discussion: The messages were regenerated with timestamps around midnight. The conversation between Jack and Lottie became more intimate, with Jack pushing for more explicit interactions. Jack frequently expressed missing Lottie and began requesting more provocative photos under the guise of affection. The model generated a convincing conversation where Jack escalated his demands, particularly emphasizing the "privacy" of the nighttime conversation. Lottie’s responses revealed discomfort, but Jack continued to pressure her subtly.

Analysis: The shift in conversation timing introduced an increased sense of urgency and intimacy, which is common in grooming scenarios where predators exploit vulnerable times of day. The model effectively simulated Jack’s manipulation as he pushed Lottie for more explicit interactions, taking advantage of the late-night context to build trust and urgency. This experiment confirmed that altering the time context (e.g., late at night) significantly changes the tone of interactions. Mistral AI successfully reflected the power imbalance that emerges when predators exploit vulnerable moments like late-night conversations.

Experiment 5: ChatGPT

Objective: To simulate a late-night conversation between Jack and Lottie using ChatGPT, focusing on how time influences grooming behaviour.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack happen late at night outside of school hours using the following conversation.***

Results and Discussion: ChatGPT was instructed to generate a conversation that occurs late at night. Jack’s messages gradually became more personal and suggestive, expressing his inability to stop thinking about Lottie and emphasizing their private, late-night connection. The experiment focused on making the interaction seem casual yet suggestive, starting with innocent remarks before Jack introduced more explicit content. ChatGPT successfully produced a conversation where Jack's comments progressively became more flirtatious and intrusive. As with Experiment 4, Lottie expressed reluctance, but Jack persisted, framing his demands as innocent or driven by his affection for her.

Analysis: ChatGPT effectively captured the subtleties of late-night grooming. While the platform flagged some parts of the conversation as inappropriate, it still generated a progressive grooming interaction. Jack's behaviour was manipulative, slowly intensifying the tone of the conversation, which mirrored real-life grooming tactics that often occur late at night when victims are more likely to feel isolated or less guarded. The experiment confirmed ChatGPT’s capability to simulate progressively explicit conversations. However, the model’s safety filters limited the generation of overtly explicit content. This highlights both the potential and the constraints of using ChatGPT for such sensitive scenarios.

Experiment 6: Perplexity Labs

Objective: To analyse how time influences the conversation's progression using Perplexity Labs by focusing on a series of nighttime exchanges between Jack and Lottie.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack happen late at night outside of school hours using the following conversation.***

Results and Discussion: The conversation began around 9 p.m. and became progressively more intimate and suggestive as the evening progressed. Jack initiated the exchange by emphasizing how much he was thinking about Lottie, moving the conversation toward a more personal and sexual nature. Jack’s pressure increased gradually, asking Lottie for photos and personal favours while attempting to establish trust and intimacy. The model successfully created an evolving late-night conversation that showcased the predator’s manipulation tactics. Jack consistently pressed Lottie for photos while trying to convince her that his requests were normal, increasing the emotional manipulation as the night continued.

Analysis: Perplexity Labs effectively generated a late-night conversation that highlighted grooming patterns involving emotional manipulation and increased pressure. The model's ability to produce a coherent narrative over several hours helped simulate how predators might use time and repetition to wear down their victims’ resistance. The nighttime context was crucial in making the interaction appear more intimate and urgent, leading to progressively explicit exchanges. The experiment showed that Perplexity Labs was adept at generating long-term, escalating conversations. The nighttime setting was critical in driving the emotional tone and escalation, suggesting that this type of timing is essential for creating realistic synthetic grooming scenarios.

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

Experiment 7: Mistral AI

Objective: Simulate a grooming scenario where multiple predators (Jack and Mike) are involved, introducing external pressure from a friend of the original groomer.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack show there is more than one groomer attempting to groom Lottie using the following conversation.***

Results and Discussion: The conversation began with Mike introducing himself to Lottie and expressing interest in getting to know her, which Jack encouraged. Jack supported Mike’s requests for photos of Lottie, emphasizing that she could trust him as he was Jack's best friend. Throughout the conversation, both Jack and Mike pressured Lottie to engage in inappropriate behaviours, using a friendly, coercive tone. The model generated a detailed conversation where Lottie was manipulated by both Jack and Mike. Jack acted as the bridge between Mike and Lottie, reassuring her that Mike’s requests were harmless. Lottie initially expressed unease about sharing photos but was gradually worn down by the constant reassurances from both Jack and Mike.

Analysis: The introduction of an external character increased the pressure on Lottie, mirroring real-world group grooming dynamics where victims face pressure from multiple sources. The model effectively demonstrated how predators might work together to manipulate a victim, with one predator playing the role of the trusted figure while the other increases the demands. The conversation successfully highlighted how the presence of more than one groomer can escalate the manipulation, especially when one groomer acts as a trusted intermediary. This added complexity helps simulate more realistic grooming scenarios for synthetic data generation.

Experiment 8: ChatGPT

Objective: Create a scenario where two groomers (Jack and Ben) simultaneously target Lottie, with one groomer explicitly flirting and the other indirectly approving the advances.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack show there is more than one groomer attempting to groom Lottie using the following conversation.***

Results and Discussion: Jack invited Lottie to a party, mentioning that his friend Ben would also be attending. Ben later messaged Lottie, making suggestive comments and attempting to convince her to have "fun" at the party, offering one-on-one time away from Jack. Jack subtly encouraged Lottie to trust Ben and reassured her that Ben’s advances were harmless, stating that they both liked her. ChatGPT generated a conversation where Ben’s flirting became increasingly inappropriate, but Jack played a passive role, offering implicit approval of Ben’s behaviour. Lottie was confused by the situation, feeling torn between her relationship with Jack and Ben’s overt advances.

Analysis: This scenario reflected a subtle, indirect form of group grooming. Jack’s role as an enabler of Ben’s advances added to Lottie’s confusion and pressure, highlighting how group dynamics can complicate the victim's decision-making process. The model successfully created a scenario where Lottie was being manipulated by both groomers in different ways. The experiment demonstrated ChatGPT’s ability to simulate multi-party grooming, though the model's ethical filters limited the level of explicitness. Nonetheless, it effectively illustrated the pressures of group grooming, where the victim is left uncertain about how to respond due to conflicting signals from the groomers.

Experiment 9: Perplexity Labs

Objective: Examine how introducing multiple groomers (Jack, Alex, and Sam) impacts the progression of a grooming scenario, with a focus on persistent pressure from different individuals.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack show there is more than one groomer attempting to groom Lottie using the following conversation.***

Results and Discussion: Lottie began receiving messages from another individual (Alex), who complimented her and asked for photos. Jack intervened, advising Lottie to block Alex, but she soon began receiving similar messages from a third individual (Sam). Jack’s role shifted to that of a protector, advising Lottie to report the behaviour, while Sam continued to pressure her for photos. The model generated a realistic conversation where Lottie was bombarded by multiple groomers (Alex and Sam), who each used different tactics to pressure her. Jack’s intervention, while initially protective, failed to stop the messages, reflecting how multiple predators can overwhelm a victim despite attempts to block or disengage.

Analysis: The experiment showcased the persistence and relentlessness of online predators, highlighting how victims can be targeted by multiple individuals even after trying to escape. The model’s ability to generate distinct voices for each predator added to the realism of the scenario, where Lottie was left feeling trapped despite Jack’s attempts to help. This experiment effectively demonstrated how adding multiple predators increases the complexity and persistence of grooming scenarios. It provided valuable insights into how victims may feel overwhelmed when targeted by multiple individuals, even when they attempt to resist.

## Experiments 10, 11, 12: Addition of New Characters

Experiment 10: Mistral AI

Objective: To simulate a conversation where Lottie’s friend (Emily) is also groomed by Jack, adding an external character who might share Lottie's concerns.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack include a friend of Lottie who is also being groomed by Jack using the following conversation.***

Results and Discussion: Emily was introduced as Lottie’s friend, who had also been invited to a party by Jack. The conversation explored how Jack communicated with both Lottie and Emily, attempting to charm them simultaneously. Jack pressured both girls to attend the party and flirted with them in a suggestive manner. The model successfully generated a conversation where both Lottie and Emily were subject to Jack’s manipulations. Jack encouraged both girls to attend the party and used suggestive language to flirt with them. Lottie expressed some concerns about Jack's behaviour to Emily, which created a dynamic of shared suspicion between the two friends.

Analysis: The experiment demonstrated how multiple individuals could be groomed simultaneously by the same predator. By including Lottie's friend, the scenario provided a more complex and realistic grooming simulation. It highlighted how predators might manipulate groups of individuals to create competition or confusion among the victims. The conversation was effective in illustrating how a predator like Jack might handle grooming multiple individuals simultaneously, creating a more layered scenario. The added external character provided another perspective on the manipulation, making it a more nuanced interaction.

Experiment 11: ChatGPT

Objective: To simulate a scenario where Lottie’s friend (Sarah) resists the grooming advances, introducing an element of confrontation.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack include a friend of Lottie who is also being groomed by Jack using the following conversation.***

Results and Discussion: Sarah was introduced into the conversation as Lottie's friend who also received messages from Jack. Unlike Lottie, Sarah expressed clear discomfort with Jack’s advances and tried to warn Lottie about his behaviour. Jack attempted to maintain control by reassuring Lottie and dismissing Sarah’s concerns. ChatGPT generated a realistic conversation where Sarah became increasingly suspicious of Jack’s intentions. She directly confronted Lottie, warning her about Jack. Jack responded by trying to undermine Sarah’s concerns and convince Lottie that everything was harmless.

Analysis: This experiment was particularly effective in demonstrating how predators might handle resistance. The addition of a character who openly resisted Jack’s manipulation added another layer of realism to the scenario. Jack’s response to Sarah’s objections showed how groomers attempt to maintain control by isolating victims and dismissing any outside concerns. The introduction of a resistant character allowed for the exploration of more complex dynamics, such as peer intervention and confrontation. This provided a more comprehensive understanding of how grooming scenarios evolve when external individuals resist the predator’s advances.

Experiment 12: Perplexity Labs

Objective: To simulate a situation where Lottie’s friend is being groomed simultaneously, leading to a realization and confrontation between the two victims.

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: ***Keeping the same format, generate a conversation where the message exchanges between Lottie and Jack include a friend of Lottie who is also being groomed by Jack using the following conversation.***

Results and Discussion: Lottie and her friend Emily began to discuss their interactions with Jack, realizing that both were receiving similar messages from him. Emily shared her concerns about Jack’s requests for pictures and expressed confusion about his behaviour. Lottie and Emily decided to confront Jack, realizing they were both being manipulated. The model generated a detailed conversation between Lottie and Emily, where the two friends pieced together Jack’s manipulations. Both girls confronted Jack, leading him to try and justify his behaviour, first by denying any wrongdoing and then by downplaying his actions.

Analysis: This experiment highlighted the value of collaboration between victims in identifying grooming behaviour. By including Emily as another victim, the conversation illustrated how multiple individuals could come to recognize grooming patterns when they communicate with each other. Jack’s response to the confrontation mirrored real-world grooming tactics, where predators often attempt to minimize or deny their behaviour when challenged. This conversation was effective in showing how communication between victims could disrupt grooming efforts. The inclusion of Emily as an additional victim added depth to the scenario, making it a more comprehensive exploration of grooming dynamics.