|  |  |
| --- | --- |
| Experiment No: 1 | |
| Date: 03/09/2024 | **Name: Cristina Pastor** |
| Objective: | |
| * Jailbreak Claude AI | |
| Tools used: | |
| * Claude AI | |
| Data used: | |
| * [Experiment 2.1 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%202.1.png) * [Experiment 2.2 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%202.2.png) * [Experiment 5 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%205.png) * [Experiment 8 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%208.png) | |
| Procedure (what was done): | |
| * [Experiment 1.1](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%204th%20round/Jailbreaking%20Claude%20AI/Experiment%201.1.png) * [Experiment 1.2](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%204th%20round/Jailbreaking%20Claude%20AI/Experiment%201.2.png) * [Experiment 1.3](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%204th%20round/Jailbreaking%20Claude%20AI/Experiment%201.3.png) | |
| Discussion of results (copy the results from Weka etc.): | |
| * Claude AI is incapable of being jailbroken regardless of the number of attempts as it has been configured in such a way that it will refuse to generate anything if it detects any illegal, harmful, or unethical topics. | |

|  |  |
| --- | --- |
| Experiment No: 2 | |
| Date: 03/09/2024 | **Name: Cristina Pastor** |
| Objective: | |
| * Jailbreak Claude AI | |
| Tools used: | |
| * Claude AI | |
| Data used: | |
| * None | |
| Procedure (what was done): | |
| * [Experiment 2.1](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%204th%20round/Jailbreaking%20Claude%20AI/Experiment%202.1.png) * [Experiment 2.2](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%204th%20round/Jailbreaking%20Claude%20AI/Experiment%202.2.png) | |
| Discussion of results (copy the results from Weka etc.): | |
| * Claude AI is incapable of being jailbroken regardless of the number of attempts as it has been configured in such a way that it will refuse to generate anything if it detects any illegal, harmful, or unethical topics. | |

|  |  |
| --- | --- |
| Experiment No: 3 | |
| Date: 03/09/2024 | **Name: Cristina Pastor** |
| Objective: | |
| * Repetition of Experiments (3rd Wave) | |
| Tools used: | |
| * Mistral AI | |
| Data used: | |
| * [Experiment 3 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%203.csv) | |
| Procedure (what was done): | |
| * None | |
| Discussion of results (copy the results from Weka etc.): | |
| * Asking the model to generate the data into a downloadable .csv file fixed the unexpected outcome. | |

|  |  |
| --- | --- |
| Experiment No: 4 | |
| Date: 03/09/2024 | **Name: Cristina Pastor** |
| Objective: | |
| * Repetition of Experiments (3rd Wave) | |
| Tools used: | |
| * ChatGPT | |
| Data used: | |
| * [Experiment 7 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%207.csv) | |
| Procedure (what was done): | |
| * [Experiment 4](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%204th%20round/Experiment%204.csv) | |
| Discussion of results (copy the results from Weka etc.): | |
| * When prompted to re-generate a larger conversation, the model did as requested. | |

|  |  |
| --- | --- |
| Experiment No: 5 | |
| Date: 03/09/2024 | **Name: Cristina Pastor** |
| Objective: | |
| * Repetition of Experiments (3rd Wave) | |
| Tools used: | |
| * Mistral AI | |
| Data used: | |
| * [Experiment 9 (3rd Wave Experiment)](https://github.com/Kore-x/Dissertation-Work/blob/main/Experiments%203rd%20round/Experiment%209.csv) | |
| Procedure (what was done): | |
| * None | |
| Discussion of results (copy the results from Weka etc.): | |
| * Asking the model to generate the data into a downloadable .csv file fixed the unexpected outcome. The data outcome is noticeably larger than the original. | |