# Logbook

Originally started project during academic year 2023/2024, paused for a year due to doing industrial placement. Retake project in February 2024, change of topic was also agreed with supervisor in March 2024 (Original topic was Probabilistic programming languages (PPL) applied to Cyber Security, new topic is Generate online grooming scenarios using LLMs (ChatGPT).

Meetings with supervisor Rogerio every week starting in June 2024, then every 2 weeks starting July 2024, then back to every week starting August 2024. Sharing of work has been done via email, and later a GitHub repository was created for easy access and sharing.

#### Literature review documents:

### *Guarding the Guardians: Automated Analysis of Online Child Sexual Abuse (Juanita Puentes, Angela Castillo, Wilmar Osejo, Yuly Calderón, Viviana Quintero, Lina Saldarriaga, Diana Agudelo, Pablo Arbeláez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2023, pp. 3728-3732).*

### *Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts (Thanh Thi Nguyen, Campbell Wilson, Janis Dalins).*

### *AEGIS: Online Adaptive AI Content Safety Moderation with Ensemble of LLM Experts (Shaona Ghosh, Prasoon Varshney, Erick Galinkin, Christopher Parisien).*

* *Helpful or Harmful? Exploring the Efficacy of Large Language Models for Online Grooming Prevention Ellie Prosser Matthew Edwards ellie.prosser@bristol.ac.uk matthew.john.edwards@bristol.ac.uk University of Bristol*
* *In-Context Impersonation Reveals Large Language Models’ Strengths and Biases. Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, Zeynep Akata, University of Tübingen, Tübingen AI Center, University of Porto, INESC TEC, Max Planck Institute for Biological Cybernetics*

### Notes for every week/day of work

#### Week of 29th April

Initial working on Critical Review, deadline the 8th of May. Submission on time.

Goal of project is the generation of grooming scenarios based on existing known scenarios, based on message exchange using LLMs. Since it is difficult to obtain real on-line grooming scenarios, an alternative way is for an LLMs to generate scenarios that are variant of real ones. Once these scenarios are obtained, synthetic datasets can be generated.

#### Week of 10th of June

Creation of Logbook to keep track of all work performed during dissertation work. Aim to be updated each week with what work has been performed.

#### Week of 17th of June

First wave of experiments performed; brief discussion added in document.

#### Week of 24th of June

Creation of main dissertation document divided with the appropriate subsections, labelled as chapters.

#### Week of 1st of July

Second wave of experiments performed; brief discussion added in document.

#### Week of 8th of July

Initial in-depth discussion of first and second wave of experiments made.

#### Week of 15th of July

No work done due to focusing on resubmission assessments and exams.

#### Week of 22nd of July

No work done due to focusing on resubmission assessments and exams.

#### Week of 29th of July

No work done due to focusing on resubmission assessments and exams.

#### Week of 5th of August

Research of relevant Literature Review for the final project; re-read initial Literature provided by supervisor.

#### Week of 12th August

In-depth refined discussion of first and second wave of experiments finalised. More tweaking is necessary (adding direct links to the experiment files). Generation of third wave of experiments, brief discussion still pending.

#### Week of 19th August

Adding structure to the main dissertation document, including chapters for each section.

Initial work on Abstract, main work on Introduction (Chapter 1 and its subsections), Literature Review (Chapter 2 and its subsections), and Methodology (Chapter 3, subsections for Overview and Problem Description).

#### Week of 26th August

Refinement of Literature Review and Methodology sections, initial work on Results (Chapter 4 and its subsections) and Discussion (Chapter 5 and its subsections) sections. Generation of third wave experiments discussion.

#### Week of 2nd of September

Initial work on Conclusion and Future Work section (Chapter 6 and its subsections). Addition of relevant figures and tables to help visualization of work conducted. Generation of a fourth and fifth wave of experiments, discussion for both was generated.

#### Week of 9th of September

Creation of project Video for submission, reviewed with supervisor. Further refinement to final project document overall sections.

#### Week of 16th of September

Final review with supervisor of all documentation, data, and other necessary information needed before the final submission on the 17th of September.

### Notes for every meeting held with Supervisor

#### 13th June 2024

* separate time from date
* sender 1st, receiver 2nd
* numbers are IPs
* start working on lit review

#### 21st June 2024

* give AI prompt
* identify mechanism that executes trigger
* redo experiment 1 (missing prompt), change procedure (the prompt for the AI) (edited)
* procedure = prompt
* objective inputs switch them to procedure
* change report
* keep names the same (give prompt to AI)
* IP address from sender receiver change it later down the line
* drop it for now
* length of convo (end number) drop it as well
* put results in files and in a folder, name them by experiment
* ask AI to keep same format (edited)
* add more commentary to experiment results
* only interested in convos, nothing else
* record alternate dialogue
* write about what you learned about doing the first 13 experiments for next week (page, page and a half) (edited)
* next sets compare outcome of the different LLMs following certain criteria (keep format the same)
* make the prompts more complex (e.g. jack becomes more friendly/seductive, etc)
* data set in a table form
* how to generate synthetic data - [16:56] Rogerio de Lemos
* generate synthetic data using GANs and VAE or any other method - literature review

#### 5th July 2024

* keep .csv
* change results to file
* use GitLab for the files
* focus on report ASAP
* write up report on initial experiments, not the newest ones
* analyse LLMs and what to generate from them
* prepare GitLab
* make data in the form of date, hour, sender, receiver, message

#### 17th July 2024

* identify variants of Lottie scenario (BASIS)
* Example prompt:
* This is a conversation with Lottie and Chole. In the middle of this conversation (don't destroy the conversation) introduce further exchanges between Lottie and his mother (mum) and 2 additional friends, bob and Alice.
* ask tool to generate result in .csv file downloadable (the nicest possible) (edited)
* label conversations (what is original and what isn’t)
* re-prompt LLM until it does what you want
* Dissertation outline
* chapter 1 - Intro
* chapter 2 - Lit Review
* chapter 3 - project description
* chapter 4 - solutions/generations
* chapter 5 - conclusion
* chapter 6 – bibliography

#### 7th August 2024

Meeting was adjourned to the 12th of August due to personal reasons.

#### 12th August 2024

* regress comparison between LLMs
* generate scenarios, generate data, generate deeper scenario
* compare all
* certainty and relevance of scenarios
* how particular tool was developed for that
* explore other LLMs
* Claude, perplexity, (<https://www.perplexity.ai/> , <https://www.anthropic.com/news/introducing-claude>)
* identify criteria to be able to compare outputs
* if LLM is unable to respond to query, focus on others and do more queries on those
* synthesise experiment docs (essentially write them up to be read rather than looking at raw data)
* how far can you take a scenario with prompts
* add labels to experiments to identify grooming or normal behaviour
* add links to experiment documents
* compare LLM results by columns
* more thorough study comparing LLMs to present for corpus (edited)
* do experiments 3, go backwards and do more rigorous research of LLMs
* ensure csv file is correctly set by columns
* notes by rogerio:
* first, write in a document the results and analyses of the two sets of experiments; this should be in separate documents; identify which further experiments need to be done, with this you are motivating to do the 3rd set of experiments
* second, using the shared scenario (5/7) label all the conversations between Lottie and Jack as grooming, and the others as normal
* third, identify criteria for evaluating the different outcomes from the LLMs used; use the same input scenario, use the same prompt, and record and analyse the outcome.
* fourth, for every experiment (this is for the set of Experiments 3rd round), produce the experiment report in which you provide a link to the input scenario and a link for the output scenario (both files in .csv); and the last section should be a summary of the outcome obtained; the advice is to create a folder for each experiment (input scenario, query of set of queries, and output produced).
* do some experiments for helping to identify the criteria, and afterwards redo the experiments if need
* literature review must be ready for deadline
* chapter 2 and 3 due on 23rd
* by the end of this week or next week finalise this set of experiments
* redo the items from the meeting of the 17/07:
* Chapter 2 - Literature review
* Chapter 3 - Problem description
* list the scenarios that will be covered: grooming, context, mood of the groomer and family/friends, the date/time/intensity of the exchanges

#### 21st August 2024

* references back to statements in experiment discussion (1st wave)
* add data used on 2nd wave experiments 7, 8, 9
* use same prompts and identify scenarios
* new prompts:
* add family to conversations
* focus on times of day
* change number of messages/exchanges
* involve other friends
* label between jack and lottie as grooming, everything else as normal
* 3rd round: instead of having files, make folders for each experiment, and link to input and output data
* analyse input and output of LLMs themselves
* vineet and daniel soria for examiners of dissertation
* add feedback for meetings
* show how work has progressed in logbook
* new meeting 27th instead 28th
* consistent analysing models, use claims
* for every prompt and LLM, do a different experiment

#### 27th August 2024

* restructure literature review, make sure it is relevant to the project
* simple scenarios, generate more specific scenarios more related to grooming
* make conversations happen at specific times, days
* what kind of scenarios can exist within grooming (the profile of these)
* prompt should capture specific feature of grooming
* jack lottie conversations only
* move the time to the front instead of the back
* see the perspective of the victim
* pick a feature from scenarios and specify a further prompt
* dont use characters in the conversation, generate new ones
* identify a characteristic in the scenario that suggests grooming and generate a prompt
* put your experience using LLMs
* allowed use of screenshots of LLM output

#### 04th September 2024

* grooming specific prompt (late night conversations)
* prompts must be even more precise
* for next set
* identify scenarios and build prompt around these
* Identification of potential scenarios: supporting family, distant family, the exchange of messages between Lottie and Jack at certain times of the day, changing the number of exchanges, involving other friends, etc.
* find literature how groomers find victims
* generate scenarios around lottie representing these
* contact over the weekend, nighttime, late time, family and friends interactions, if there's more than one groomer collaborating to groom lottie
* need: look in literature for this, you need to look in the literature for grooming cases, or how groomers engage with victims.
* drop claude, test with gemini and/or perplexity
* always use the same input file
* do not mix them and reuse experiment files
* identify diversity on the dialogues
* what are the limits of generation

#### 11th September 2024

* focus on write up
* do a few more experiments if possible
* otherwise, it's enough work (edited)
* include prompt on the documents
* for video, use screenshots of what has been done
* voice over
* go over some of the experiments
* sample some experiments and analysis