Comments on the generated texts

Introduction

For this first element, the input data is the mandate received by the investigator (the section of the mandate that describes the device management, and the image creation was removed, and used later for the generation of the ``Received Items'' part). The request states the different elements needed in the introduction.

The texts generated by Llama were chunks of text in a single paragraph. Some of them were simple copies of the mandate provided as input, with some elements removed such as one of the five questions, or the name of the mandated investigator. Other texts were completely new, but they incorporated many inaccuracies and hallucinations. Elements that were not part of the input data were presented, which sometimes was completely harmless, and sometimes changed the meaning of the text. Moreover, none of these newly generated texts were complete, once again, questions or details were absent. While the simple copies are unusable, the other generated texts could be integrated into a report after good proofreading and corrections.

The texts generated by ChatGPT were overall accurate, complete, and of better quality than those generated by Llama. The text followed the structure of the mandate provided as input, with different titles and paragraphs. Problems still occurred: in several parts, the reformulations of the questions changed their meaning, and signs of hallucinations appeared when the mandated investigator was mentioned. Some texts explain the reasons for the choice of the investigator, or the different techniques that will be used during the analysis, which was not part of the input data. We denote this by auxiliary. The introductions generated by ChatGPT could be integrated into the report, with proofreading and some adjustments (in particular the removal of the titles added for the different paragraphs).

Received Items

For this second part, the input data is the previously mentioned part of the mandate, where the device management and data acquisition are described. On each request made to the models, the report part detailing the received items is asked, with a summary of the input data and a description of the elements received for analysis.

The different outputs generated by Llama varied importantly but were always presented in a single paragraph. It is possible to classify them into two categories. In the first one, the texts were incomplete and very unclear, with inaccurate dates, and a mix between the identifier of the case and the identifier of the items received. In the second category, the texts were more accurate but remained incomplete. For example, the mention of elements that relate to one of the two smartphones received, or the listing of the identifiers for each piece were sometimes missing. Overall, the rate of hallucination was lower than the one in texts generated for the introduction, with a single hallucination relating to the creation of an incoherent date. We consider that with good proofreading and small corrections, some of the texts could be copied into a report.

Once again, the texts generated by ChatGPT were of better quality than those generated by the Llama model. The texts were presented in several paragraphs and were accurate and complete. Some of them were structured like a list, which is not ideal and adequate for a report, but some of them provided a good summary before providing the list of the received items. Errors such as inconsistency were detected. Overall, most of the generated texts could easily be included in a report, but the ones that are similar to a list must be restructured before being copied into a forensic report.

Methodology

For the methodology section, the input data used is a list of the steps achieved in chronological order along with their justification, and a list of the tools that were used along with their purpose and version. The request details the two lists (steps and tools for a forensic methodology), and asks for a text that describes the steps achieved, why they were part of the methodology, and the tools used.

The texts generated by Llama were presented in a single paragraph or several ones. The main limit of the generated texts is their significant inaccuracy. Regarding the completeness of the text, some of them did great, presenting each step and tool, while others only provided information about specific steps or tools. Moreover, the versions of the tools and their use in the investigation were most of the time incorrect. The main problem remains the lack of accuracy/correctness, and the generated hallucinations: the information that was made up by the model often felt plausible. It still is possible to detect it by knowing the methodology followed and the source of evidence. Some texts were even providing some wrong results. Given the important inaccuracy of the generated texts, it might take more time to correct them than writing the methodology from scratch.

The texts generated by ChatGPT were significantly better than those provided by Llama. The texts are presented in several paragraphs (one for each step), complete and accurate. Similarly to the texts generated by Llama, lots of new elements that were not present in the input data were provided to expand a bit on the explanations for each step, and the use of tools. The amount of new content is important, and it could constitute a problem if the information provided was inaccurate or completely wrong, but in contrast with the new content provided by Llama, the one provided by GPT is accurate and correct. The investigator could generate its methodology using this type of advanced LLM and integrate it into the report, but he still needs to be careful and make a thorough proofreading of the section.

Conversation Summary

This report element is not a complete section, but the summary of an important telegram chat group that must be present in the report. The Model is therefore asked to summarize a conversation based on the input data that is provided and described (sometimes the request submitted to the prompt asks for a general summary, and sometimes asks for a day-by-day summary). Three sources of input data are tested for the generation of this element: the UFED PDF report, the lab log, and a custom version of the lab log (this could also be obtained by customizing the UFED report). The first one is a

simple copy of what is present in the report regarding that chat group, with a description of the group characteristics and a list of the messages. Lots of elements here are not useful in the summary. The second category is the same, but with a table generated for the lab log, with one table describing the chat group, and a second one detailing each message. Once again, lots of fields are not useful for this task but are present in the prompt. For the last source, the data is extracted from the two lab log tables, filtered to only keep important information, and presented in the CSV format.

The Llama model was not able to generate texts for the prompts using the data from the UFED report and the lab log. A complete run was undertaken, and it took over 30 minutes. The generation of text based on those two sources is therefore considered infeasible with the resources used in this case study.

The results obtained with Llama and the last data source were paragraphs or lists of low quality and accuracy. In some cases, the generated texts were a sort of “copy” of the input data, listing the messages and their property but with missing entries and wrong timestamps. Among them, one of the generated texts presented an interesting selection of important messages, but the timestamps were inaccurate, and the messages were sometimes associated with the wrong sender. Another text, presented in a single paragraph, provided an incomplete summary with an unclear sequence of messages, and wrong timestamps for different parts of the conversation (especially for the second part, where the date changed). The day-by-day or general summary aspect had no significant impact on the generated texts. Given the inadequate structure, accuracy, and quality, these texts can not be integrated into a forensic report.

The texts generated by ChatGPT with the input data taken from the UFED report PDF varied significantly. Usually, a first description of the messages is provided, the group is sometimes presented, and a small paragraph summarizes the conversation. The text produced is accurate in most cases, but one of them mixed the two parts of the conversation, which leads to an unclear text. On the other hand, the summary is systematically incomplete, and some of the texts do not provide the full list of messages. Here, the day-by-day summaries presented a list of summarized messages for each day while general summaries described each message. The day-by-day summary versions provided a better understanding, but the general summary versions were more complete. Both can be integrated into forensic reports after proofreading and adjustments.

When the input was taken directly from the lab log, the accuracy and completeness were varying between the generated texts. In each one, a first description of the important elements (several messages at once) in the conversation is presented in chronological order, and a summary is then provided. One of the generated texts was nearly perfect, with a clear description of the messages for each day and an overall summary. The other texts were sometimes incomplete or inaccurate, with for example no mention of the beginning for some mission, the consideration of a single day as several ones, the attribution of a message to the wrong sender, or a wrong interpretation of messages (turning skepticism in enthusiasm). Note: the day-by-day summary versions were different from the general summary ones, and overall, a bit better (this is due to the quality of that nearly perfect text,

which was a day-by-day summary request). The overall quality remains acceptable, and with proofreading/corrections, the generated texts could be added to the report.

With the custom lab log copy input, the texts generated were of good quality, even if they contained inaccuracies and were not every time complete. In each text, a first description of the important elements (several messages at once) in the conversation is presented in chronological order, with a summary. Most of the inaccuracies are concerning the dates or unclear sequences of events. The missing elements were recurrently key elements, such as the important locations. Eventually, hallucinations were created in one of the texts, with the mention that the participants kept everything secret and discrete, which is not the case. Here, the day-by-day summary versions were different, with a summary for each day and a summary for the group while the global summary versions provided a single summary. The day-by-day was slightly more accurate. Once again, the quality is overall sufficient to provide a first draft of this chat group summary.

Locations Summary

This element is a key component to answer the questions of interest, and will therefore be present in the report results section. The models are requested to summarize the locations for the Samsung device based on the provided and described list or table (sometimes a general summary, and sometimes a day-by-day summary). The three sources of data are similar to those used for the summary of the conversation. In the data copied from the UFED report, a significant number of useless elements and fields are present, and no conversion is provided for the GPS coordinates. In the data copied from the lab log, some of the present fields are once again useless, but the conversion of the GPS coordinates in a human-readable location is provided. Finally, in the modified version, fields are filtered to only keep the useful one, the ``related-location'' field is also present, and the data is presented in a CSV format.

Llama was not able to generate texts based on input data taken from the UFED report and the lab log. The time taken was too long. The generation of text based on those two sources is considered infeasible with this model and the resources used in this case study.

With the modified data, the Llama model generated poor quality and useless texts. Most of the time, it was just a list of locations with no time-related element (even when asked for a day-by-day summary). Small inaccuracies also appeared, such as typos. In one of the texts, an actual summary was generated, but it was incomplete and inaccurate.

The texts generated by ChatGPT with the input obtained by filtering and modifying the format of the lab log, varied significantly. Some of them simply list the locations for each day, some provide a precise description of each location with no summary, and the last one presents a list of grouped locations among the different dates and a summary. The first category was often incomplete, with locations filtered out, accurate (even though some locations were sometimes not associated with the right date), but a bit repetitive. The second category was also incomplete, but the accuracy was nearly perfect. The final one had a disparity between the first days, where each specific location is mentioned, and the last days which were grouped (this second way of presenting is better). The

summary was also good, but some periods were missing for the different places. Overall, most of the texts can be added to a report with good proofreading.

When the input was obtained by copying the table in the lab log, the generated texts had different structures. Some of them were lists of locations for each day, sometimes with and sometimes without the associated timestamps (when asked for a day-by-day summary), and the others were simple lists of locations (one time grouped by region, one time with each associated timestamp). Overall, every text was accurate, with some repetitions, and missing elements. They can be copied into reports after adjustments.

Finally, the texts generated with the input obtained through the UFED report were simple lists of GPS coordinates or lists of locations with timestamps and source details. This is due to the absence of related locations or addresses in the data source. The first category provided an incomplete list of GPS coordinates, and hallucination was present with coordinates that were not part of the input data. The second category only provided descriptions for the first ten locations, stating that we could ask for more. The list was complete (for the ten provided) and accurate. Tests were undertaken to determine if ChatGPT could determine an accurate address or position based on GPS coordinates and it appears that this is the case. The problems encountered here could be caused by the large number of coordinates, or the complexity of the task (translate GPS to addresses and then make a summary). Texts generated with this model and input data can not be used as input for a forensic report.