Forensic Science International: Digital Investigation 44 (2023) 301543

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Editorial

Digital forensic investigation in the age of ChatGPT

Large Language Models (LLMs), e.g., BERT, GPT-3, GPT-4, LLaMA, etc., have gained public notoriety in recent months with the advent of OpenAI's ChatGPT. Since its public launch in November 2022, professionals across a broad range of disciplines have evaluated its potential implications and disruptions to their respective fields. Schools and universities the world over are discussing the implications of ChatGPT for the trustworthiness of student assignment and exam submissions — and many have already opted to return to traditional "pen and paper" examinations. Elsevier has updated their publishing ethics policy to include guidance on the use of AI and AI-assisted technologies in scientific writing, advising when its use is acceptable (or not), that its use should be disclosed, and that AI and AI-assisted technologies should not be listed as an author or co-author.[1]

Optimising the use of these LLM systems depends on the accuracy of the prompts used. Prompt tone, specificity, and word choice can all influence the results returned. Cleverly constructed prompts can also be leveraged to meander around some built-in safety nets defining the scope of the allowable use cases, e.g., "pretend to do X" vs. "do X". This has resulted in the coining of the phrase *prompt engineering* in many online communities.

From a digital forensics standpoint, LLMs can certainly be used to provide a range of benefits to enhance and expedite the investigative process. However, as with everything at the intersection of artificial intelligence (AI) and digital forensics, it is vital to maintain the "AI-assisted investigation" and "human-in-the-loop" mantras when it comes to its use rather than ever approaching a world where we become overly dependent on such systems and lose the underlying understanding of the evidence. The premise of using a publicly hosted LLM is not feasible during many use cases handling sensitive or privileged data, so a locally hosted installation would be necessary. Nonetheless, publicly hosted systems can still be used for a range of use cases when its use is abstracted from any individual case. Some of the beneficial uses of this technology include:

- Automatic script generation - Perhaps one of the most useful use cases for LLMs is the ability to specify a script/program that you need to perform some evidence analysis and have it be automatically generated. Using off-the-shelf libraries and tools, instantly generated scripts can automatically parse disk images or memory dumps to find specific, pertinent information. Gone can be the days of manually creating complicated scripts, queries, and regular expressions.
- Question answering - As society has become accustomed to the question:answer interactions possible with personal assistants, e.g., Siri or Alexa, so too can this type of interaction be possible with an LLM. The ability to ask plain language questions and retrieve relevant information based on your current case could greatly expedite an investigation. This plain-language querying of forensic data can also be made available to lawyers, prosecutors, and judges to explore digital evidence without the need of any digital forensic expertise.
- Multilingual analysis - The ability to specify what you are looking for in your native language and have the system discover pertinent information no matter what language it is written in can greatly expedite the process in cross-linguistic cases.
- Automated sentiment analysis - The ability to quickly and easily identify threatening, grooming, harassment, phishing or hate speech communications can focus an investigation in an efficacious direction at the earliest stage possible.

The ability to quickly and easily leverage the above options is revolutionary. Pertinent evidence could be unearthed at an early stage of the investigation. Systems can be designed or expanded to automatically search and index a vast array of interconnected traces, which typically requires human intuition. Nonetheless, the employment of such a technology is not without its risks. Some of the risks include:

- Bias and errors - As with any AI system, the models produced are only as reliable as the data used for training. The system does not know what is right or wrong morally or ethically, and is predominantly trying to generate humanlike text.
- Hallucinations - These systems are focused first and foremost on generating humanlike text in response to a prompt. As the results are generated, the model is focused on finding the most likely or suitable word to the text that has preceded it - effectively one word at a time. As a result, they are often more focused on having *an answer* rather than *the correct answer*. This can result in inaccurate/incorrect results being presented to the end user as fact, or the tone of the response demonstrating an unfounded confidence in the information being presented. In fact, if a user requests references for any facts presented, ChatGPT will generate fake bibliographic information containing viable authors from the field, a fake title, a viable journal/conference name (including FSI: Digital Investigation), and fake volume, issue, page/article numbers, and year. Without due diligence, this could erroneously be presented as proof for a position or statement by end users.
- Legal issues - The use of an LLM during an investigation might be challenged in court. Due to its necessarily

[1] https://www.elsevier.com/about/policies/publishing-ethics.

https://doi.org/10.1016/j.fsidi.2023.301543

complicated architecture, the ability to explain the precise process followed in identifying some incriminating evidence may be lost. The prompting used can help address some of these concerns, asking the system to explain the process followed step-by-step, but to many investigators, the system itself will remain a black box environment.

- Overreliance - Having an easy to use, powerful, automated system at your disposal can naturally result in an over-reliance on its use. Investigators must not lose sight of the underlying evidence and technologies needed to manually perform the investigation.
- Ethical concerns - The employment of this technology in a forensic context raises some ethical questions surrounding transparency, privacy, fairness, non-maleficence, and trust. How much should an investigator rely on the output from the system? How can we be certain the system didn't access the information pertaining to other out-of-scope individuals, or ensure that it has not accessed privileged information?
- Lack of human judgement - Any pre-trained model may not be able to provide the same level of human judgement and insight that is needed in many investigations.
- Technical limitations - As these models are first and foremost language models, they have severe limitations on the data they can consume and process. Without suitable prompting, any results generated may not declare its limitations, what data it skipped over, or what data was not consumable.

Of course, LLMs can also be leveraged to remove some of the technical barriers to entry or increase the likelihood of success for a range of crimes, e.g., phishing, malicious code obfuscation, hacking, etc. The use of LLMs like ChatGPT to assist committing crime is already being discussed in underground criminal forums. Users of LLMs will become the focus of forensic investigation for a wide range of cases. This will be largely reliant on the preservation of traces and retention periods of the service providers. Retention of prompts, user access logs, and generated responses will be the main source of evidence. Inevitably, the models themselves will also become under forensic scrutiny. Investigating the models will certainly prove more difficult than the investigation of their users. Notably, these systems tend to be non-deterministic and return disparate answers to prompts depending on a large range of factors.

With an ever-increasing demand for expert digital forensic analysts the world over, one might reasonably envision a not-too-distant future with an enhanced digital forensic first responder model. One that leverages ChatGPT, and similar technologies, to enable the natural language querying of digital evidence by non-digital forensic experts. Indeed, this may well result in a new career specialisation: digital forensic prompt engineers.

Note: After this editorial was written, Europol published a related, relevant article on this topic from their perspective entitled *ChatGPT: The impact of Large Language Models on Law Enforcement*.[2]

Mark Scanlon, Bruce Nikkel, Zeno Geradts

[2] https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement.