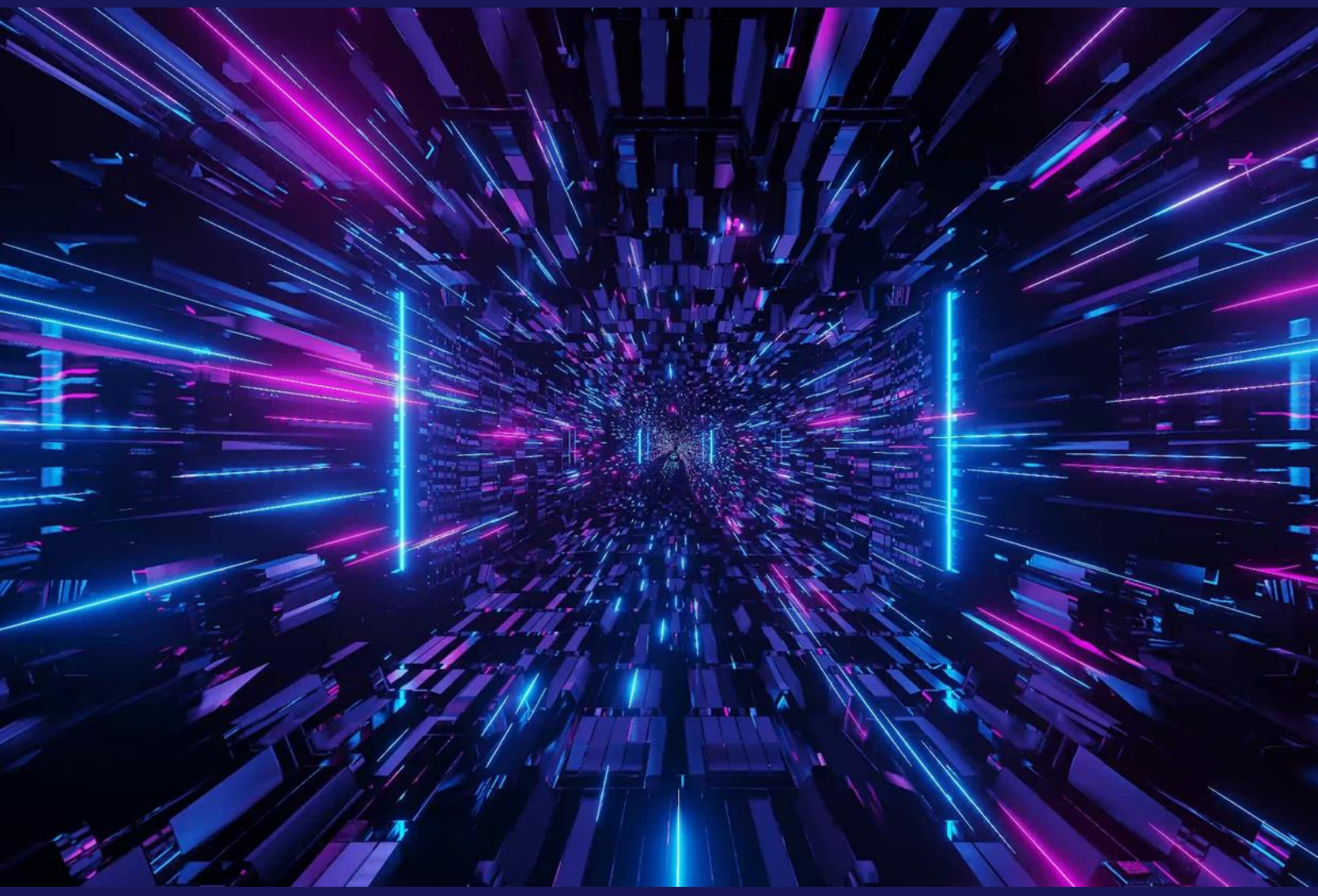
EUROPOL

TECH WATCH FLASH

ChatGPT

The impact of Large Language Models on Law Enforcement



27/03/2023

CONTENTS

INTRODUCTION

The release and widespread use of ChatGPT – a large language model (LLM) developed by OpenAI – has created significant public attention, chiefly due to its ability to quickly provide ready-to-use answers that can be applied to a vast amount of different contexts.

These models hold masses of potential. Machine learning, once expected to handle only mundane tasks, has proven itself capable of complex creative work. LLMs are being refined and new versions rolled out regularly, with technological improvements coming thick and fast. While this offers great opportunities to legitimate businesses and members of the public it also can be a risk for them and for the respect of fundamental rights, as criminals and bad actors may wish to exploit LLMs for their own nefarious purposes.

In response to the growing public attention given to ChatGPT, the Europol Innovation Lab organised a number of workshops with subject matter experts from across the organisation to explore how criminals can abuse LLMs such as ChatGPT, as well as how it may assist investigators in their daily work. The experts who participated in the workshops represented the full spectrum of Europol's expertise, including operational analysis, serious and organised crime, cybercrime, counterterrorism, as well as information technology.

Thanks to the wealth of expertise and specialisations represented in the workshops, these hands-on sessions stimulated discussions on the positive and negative potential of ChatGPT, and collected a wide range of practical use cases. While these use cases do not reflect an exhaustive overview of all potential applications, they provide a glimpse of what is possible.

The objective of this report is to examine the outcomes of the dedicated expert workshops and to raise awareness of the impact LLMs can have on the work of the law enforcement community. As this type of technology is undergoing rapid progress, this document further provides a brief outlook of what may still be to come, and highlights a number of recommendations on what can be done now to better prepare for it.

Important notice: The LLM selected to be examined in the workshops was ChatGPT. ChatGPT was chosen because it is the highest-profile and most commonly used LLM currently available to the public. The purpose of the exercise was to observe the behaviour of an LLM when confronted with criminal and law enforcement use cases. This will help law enforcement understand what challenges derivative and generative AI models could pose.

A longer and more in-depth version of this report was produced for law enforcement consumption only.

BACKGROUND: LARGE LANGUAGE MODELS AND CHATGPT

Artificial Intelligence

Artificial Intelligence (AI) is a broad field of computer science that involves creating intelligent machines that can perform tasks that typically require human-level intelligence, such as understanding natural language, recognizing images, and making decisions. Al encompasses various subfields, including machine learning, natural language processing, computer vision, robotics, and expert systems.

Neural Networks

Neural Networks, also known as Artificial Neural Networks (ANN), are computing systems inspired by the structure and function of the human brain. They consist of interconnected nodes or neurons that are designed to recognize patterns and make decisions based on input data.

Deep learning

Deep Learning is a subfield of machine learning that involves training artificial neural networks, which are computing systems inspired by the structure and function of the human brain, to recognize patterns and make decisions based on large amounts of data. Deep Learning has been particularly successful in fields such as image recognition, natural language processing, and speech recognition.

Supervised/unsupervised learning

Supervised Learning is a type of machine learning that involves training a model using labeled data, where the desired output is already known. The model learns to make predictions or decisions by finding patterns in the data and mapping input variables to output variables.

Unsupervised Learning is a type of machine learning that involves training a model using unlabeled data, where the desired output is unknown. The model learns to identify patterns and relationships in the data without being given specific instructions, and is often used for tasks such as clustering, anomaly detection, and dimensionality reduction.

Definitions provided by ChatGPT.

ChatGPT is a large language model (LLM) that was developed by OpenAI and released to the wider public as part of a research preview in November 2022. Natural language processing and LLMs are subfields of artificial intelligence (AI) systems that are built on deep learning techniques and the training of neural networks on significant amounts of data. This allows LLMs to understand and generate natural language text.

Over recent years, the field has seen significant breakthroughs due in part to the rapid progress made in the development of supercomputers and deep learning algorithms. At the same time, an unprecedented amount of available data has allowed researchers to train their models on the vast input of information needed.

The LLM ChatGPT is based on the Generative Pre-trained Transformer (GPT) architecture. It was trained using a neural network designed for natural language processing on a dataset of over 45 terabytes of text from the internet (books, articles, websites, other text-based content), which in total included billions of words of text.

The training of ChatGPT was carried out in two phases: the first involved unsupervised training, which included training ChatGPT to predict missing words in a given text to learn the structure and patterns of human language. Once pre-trained, the second phase saw ChatGPT fine-tuned through Reinforcement Learning from Human Feedback (RLHF), a supervised learning approach during which human input helped the model learn to adjust its parameters in order to better perform its tasks.

The current publicly accessible model underlying ChatGPT, GPT-3.5, is capable of processing and generating human-like text in response to user prompts. Specifically, the model can answer questions on a variety of topics, translate text, engage in conversational exchanges ('chatting'), and summarise text to provide key points. It is further capable of performing sentiment analysis, generating text based on a given prompt (i.e. writing a story or poem), as well as explaining, producing, and improving code in some of the most common programming languages (Python, Java, C++, JavaScript, PHP, Ruby, HTML, CSS, SQL). In its essence, then, ChatGPT is very good at understanding human input, taking into account its context, and producing answers that are highly usable.

In March 2023, OpenAI released for subscribers of ChatGPT Plus its latest model, GPT-4. According to OpenAI, GPT-4 is capable of solving more advanced problems more accurately[1]. In addition, GPT-4 offers advanced API integration and can process, classify, and analyse images as input. Moreover, GPT-4 is claimed to be less likely to respond to requests for 'disallowed content' and more likely to produce factual responses than GPT-3.5[2]. Newer versions with greater functionalities and capabilities are expected to be released as the development and improvement of LLMs continues.

Limitations

Still, the model has a number of important limitations that need to be kept in mind. The most obvious one relates to the data on which it has been trained: while updates are made on a constant basis, the vast majority of ChatGPT's training data dates back to September 2021. The answers generated on the basis of this data do not include references to understand where certain information was taken from, and may be biased. Additionally, ChatGPT excels at providing answers that sound very plausible, but that are often inaccurate or wrong[3] [4]. This is because ChatGPT does not fundamentally understand the meaning behind human language, but rather its patterns and structure on the basis of the vast amount of text with which it has been trained. This means answers are often basic, as the model struggles with producing advanced analysis of a given input[5]. Another key issue relates to the input itself, as often, the precise phrasing of the prompt is very important in getting the right answer out of ChatGPT. Small tweaks can quickly reveal different answers, or lead the model into believing it does not know the answer at all. This is also the case with ambiguous prompts, whereby ChatGPT typically assumes to understand what the user wants to know, instead of asking for further clarifications.

Finally, the biggest limitation of ChatGPT is self-imposed. As part of the model's content moderation policy, ChatGPT does not answer questions that have been classified as harmful or biased. These safety mechanisms are constantly updated, but can still be circumvented in some cases with the correct prompt engineering. The following chapters describe in more detail how this is possible and what implications arise as a result.

[1] OpenAI 2023, GPT-4, accessible at https://openai.com/product/gpt-4.
[2] Open AI 2023, GPT-4 System Card, accessible at https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[3] Engineering.com 2023, ChatGPT Has All the Answers – But Not Always the Right Ones, accessible at https://www.engineering.com/story/chatgpt-has-all-the-answers-but-not-always-the-right-ones.
[4] Vice 2022, Stack Overflow Bans ChatGPT For Constantly Giving Wrong Answers, accessible at https://www.vice.com/en/article/wxnaem/stack-overflow-bans-chatgpt-for-constantly-giving-wrong-answers.
[5] NBC News 2023, ChatGPT passes MBA exam given by a Wharton professor, accessible at https://www.nbcnews.com/tech/tech-news/chatgpt-passes-mba-exam-wharton-professor-rcna67036.