Communications Security Establishment
Centre de la sécurité des télécommunications

CANADIAN CENTRE FOR
CYBER SECURITY

The threat from large language model text generators

Canada

Audience

Subject to standard copyright rules, TLP:CLEAR information may be distributed without restriction. You can find more information on the Traffic Light Protocol at the Forum of incident response and security teams website.

Contact

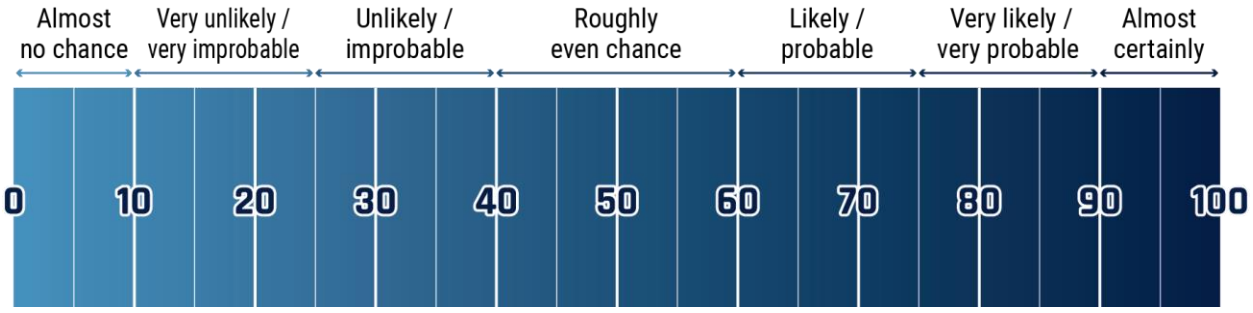
For follow up questions or issues please contact the Canadian Centre for Cyber Security (Cyber Centre) at contact@cyber.gc.ca.

Assessment base and methodology

The judgements in this assessment are based on the knowledge and expertise in cyber security of the Cyber Centre. Defending the Government of Canada's information systems provides the Cyber Centre with a unique perspective to observe trends in the cyber threat environment, which also informs our assessments. The Communications Security Establishment's (CSE) foreign intelligence mandate provides us with valuable insight into adversary behavior in cyberspace. While we must always protect classified sources and methods, we provide the reader with as much justification as possible for our judgements.

Our judgements are based on an analytical process that includes evaluating the quality of available information, exploring alternative explanations, mitigating biases and using probabilistic language. We use terms such as "we assess" or "we judge" to convey an analytic assessment. We use qualifiers such as "possibly", "likely", and "very likely" to convey probability.

The contents of this document are based on information available as of June 26, 2023.



Introduction

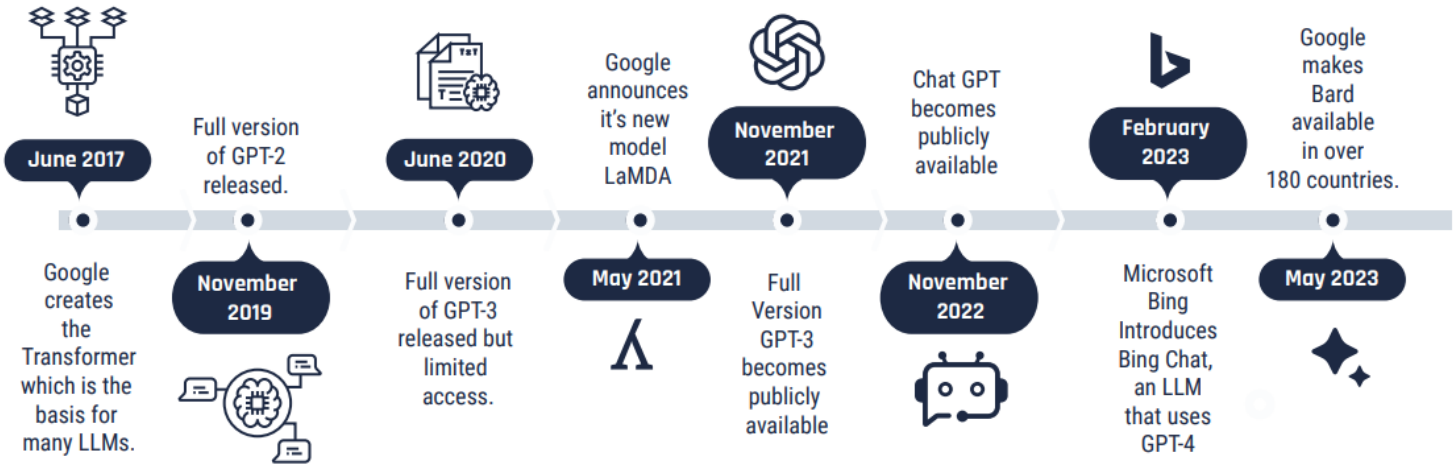
Since at least 2016, generative artificial intelligence (AI) has been able to generate synthetic content consisting of fake text, images, audio, and video documents. This synthetic content can be used in disinformation campaigns to covertly manipulate online information, and as a result, influence opinions and behaviours. Generative AI has become increasingly accessible to the public as well as to a range of cyber threat actors and state-sponsored actors. We judge that Canadians using social media have very likely been exposed to synthetic content.[1] Consequently, large language models (LLMs) represent a growing and evolving threat to Canada's information ecosystem, its media and telecommunication landscape, and the structures in which information is created, shared, and transformed.

Brief history of large language models (LLMs)

In June 2017, researchers at Google proposed a new artificial neural network architecture called the Transformer, a breakthrough model that trains faster and requires less training data.[2] This architecture would later become the basis for other LLMs such as OpenAI's Generative Pre-Trained Transformer (GPT) model. In August 2019, OpenAI, an artificial intelligence research and deployment company, released a partial version of its Generative Pre-trained Transformer 2 (GPT-2), a language model capable of generating paragraphs of coherent text that are virtually indistinguishable from human writing.[3] OpenAI initially released an extremely restricted version of the model due to concerns about its possible applications for "reducing the cost of generating fake content and waging disinformation campaigns".[4]

Despite concerns about LLM's malicious applications, big tech companies, such as OpenAI, Google, Meta, and Microsoft, continued to develop text generating tools.[5] On May 18, 2021, Google announced the release of LaMDA, a model trained on dialogue rather than just text compositions.[6] OpenAI released its latest GPT-3 model on November 2021 but the company was only able to launch an update to this model the following year: GPT-3.5, the LLM behind its popular ChatGPT. ChatGPT became one of the fastest-growing consumer software applications in history due to its accessibility, flexibility, and accuracy for a variety of tasks.[7] Other tech companies quickly followed suit, releasing similar LLMs, accessible with user-friendly interfaces, including Google's Bard chatbot and Microsoft Bing Chat, which uses GPT-4.

Figure 1. Timeline of Large Language Models (LLMs)



Threat landscape: Most likely threats

Online influence campaigns: Prior to LLM text generators, online influence campaigns required humans to produce content and/or spread disinformation to influence beliefs and behaviours. LLM text generators help or replace human writers to mass produce documents, comments, and discussions that disseminate or amplify misinformation/ disinformation. We assess that Canada may be particularly vulnerable to online influence campaigns using LLMs due to the high intake of social media content by Canadians.[8]

Email phishing campaigns: LLMs generate synthetic text about a particular topic in a particular style and have progressed to the point where the text they produce is often nearly indistinguishable from human-written text.[9] Cyber threat actors can enter a prompt into text generators to quickly draft targeted phishing emails to steal sensitive information, such as account credentials, or financial information from victims.

Human or machine? We assess that current machine learning detection tools are very likely unable to identify LLM generated text. We assess that, given the current lack of effective synthetic content detection tools and the increasing availability of LLM text generators, it is likely that online influence campaigns aiming to propagate disinformation will increasingly go undetected, appear authentic, or be produced on a mass scale rendering manual identification impossible. We also assess that it is very likely that as these technologies improve it will become more difficult for humans to detect their use, impacting social media companies' ability to detect and remove synthetic content.

Potential, unlikely threats

Malicious code: LLMs have the ability to write code snippets in popular programming languages, such as JavaScript, Python, C#, PHP and Java, which can help lower the level of skill needed for those aspiring to create code.[10] Some LLM text generators have a coding function that threat actors could exploit to create new malicious code.[11] However, we assess that the use of LLM text generators to create sophisticated code that could lead to a zero-day attack is unlikely.

Poisoning datasets: LLMs are trained on large language datasets. Theoretically, threat actors could inject or change data used to train newer versions of LLMs to maliciously undermine the accuracy and quality of the generated data. However, due to the large size and proprietary nature of the datasets, we judge that the poisoning of these large datasets is very unlikely.

Risks to organizations

Organizations using LLM text generators for their work duties may undermine their responsibilities for data stewardship or sidestep the frameworks that protect sensitive information. LLM text generators could be used by organizations for:

- research
- drafting findings
- collating statistics
- writing emails
- producing internal reports

The specific risks associated with using LLM text generators for organizations information include the following:

Data governance: LLM text generators require an input or prompt from the user, such that text provided by an employee could contain information under the governance of an organization. The input data is transferred outside of that organization's control, into the custody of the service provider, in order to generate the desired output text. This data can also be fed back into the LLM or stored for other uses. Unauthorized use of online tools exposes information to third parties and fails organizational data governance requirements.

Security of protected information: Members of an organization providing input into LLM text-generators may also unwittingly leak sensitive information, including personal information and confidential commercial information, outside of approved security and policy frameworks. For example, using an LLM text generator to compose a response to a client inquiry may use their personal information outside approved networks or beyond the purposes for which the information was collected. This increases the potential for a leak of that information via the third party.

Key terms

Artificial neural networks are flexible models that can be trained to perform and automate very specific and complex tasks, such as generating realistic videos of events that never occurred (commonly referred to as deepfakes). They can identify and learn the relationships and patterns that exist within extremely large datasets and build up complex representations of this data. This makes them an essential component of large language models that can generate convincing synthetic media.

Large language models (LLMs) are artificial neural networks that are trained on very large sets of language data using self- and semi-supervised learning. LLMs initially generated text via next word prediction but can now take prompts that enable users to complete sentences or generate entire documents on a given topic. Training on exceptionally large datasets allows the model to learn sophisticated linguistic structure, but also the biases or inaccuracies found in that data.

Machine learning (ML) is a field of research into methods that allow machines to learn how to complete a task from given data without explicitly programing a step-by-step solution. ML models can often approach or exceed human performance for certain tasks. Ss such, machine learning is considered a sub-discipline of AI research.

Synthetic content refers to content that is machine-generated with little to no human assistance.

Online influence campaigns occur when threat actors covertly create, disseminate or amplify misinformation or disinformation to influence beliefs or behaviours.

[1] Most Canadians have viewed some form of synthetic content on social media due to 1) the large amounts of synthetic content circulating on social media and 2) Canadians' high intake of social media content: Researchers at the Queensland University of Technology found that, on average, over 3.2. billion photos and 720,000 hours of video are created daily and available online. They note that plenty of this online content consists of synthetic media shared on social media. In 2018, 78% of Canadians used at least one social networking account and as of January 2021, the estimated number of Canadian users on social media platforms Facebook, Instagram, Twitter, TikTok, WeChat and Youtube totalled 67.1 million. See Sebastien Charlton and Kamille Leclair. Digital News Report: Canada 2019 Data Overview. Centre d'études des médias, Département d'information et de communication, Université Laval. February 2019; Schimmele et al. Study: Canadians' assessments of social media in their lives. Statistics Canada. March 24, 2021; T.J. Thompson et al. Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities. Journalism Practice, Research Gate. October 2020.

[2] Ashish Vaswani et al. Attention Is All You Need. Google Brain and Google Research. June 12, 2017.

[3] OpenAI Blog. Better Language Models and Their Implications. February 14, 2019.

[4] OpenAI Blog. Better Language Models and Their Implications. February 14, 2019.

[5] Eray Eliaçık. The role of large language models in the AI war. Data Economy. February 27, 2023.

[6] Eli Collins. LaMDA: our breakthrough conversation technology. Google: The Keyword. May 18, 2021.

[7] The ChatGPT app is estimated to have reached 100 million active users in January 2023, only two months after its launch. UBS Wealth Management. Let's chat about ChatGPT. February 23, 2023.

[8] In January 2021, the estimated number of Canadian-owned accounts on social media platforms Facebook, Instagram, Twitter, TikTok, WeChat and YouTube totalled 67.1 million. In 2019, almost 50% of Canadians aged between 18 and 24 relied on social media as their main source of news. See Schimmele et al. Study: Canadians' assessments of social media in their lives. Statistics Canada. March 24, 2021.

[9] Nguyen et al. Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv: 1909.11573v3. April 26, 2021; Ian J. Goodfellow et al. Generative Adversarial Net. Département d'informatique et de recherche opérationnelle Université de Montréal. June 10, 2014.

[10] Amber Isrelsen. How to use ChatGPT to write code. Pluralsight Blog. March 22, 2023.

[11] CheckPoint Research. OPWNAI : Cybercriminals Starting to Use ChatGPT. January 6, 2023.