Centre de la sécurité des télécommunications
Communications Security Establishment

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ

La menace posée par les générateurs de texte basés sur des modèles de langage de grande taille

© Gouvernement du Canada
Le présent document est la propriété exclusive du gouvernement du Canada. Toute modification, diffusion à un public autre que celui visé, production, reproduction ou publication, en tout ou en partie, est strictement interdite sans l'autorisation expresse du CST.

Canada

Auditoire

Tout en étant soumise aux règles standard de droit d'auteur, l'information TLP:CLEAR peut être distribuée sans aucune restriction. Pour obtenir de plus amples renseignements sur le protocole TLP (*Traffic Light Protocol*), prière de consulter le site Web du Forum of Incident Response and Security Teams (en anglais seulement).

Coordonnées

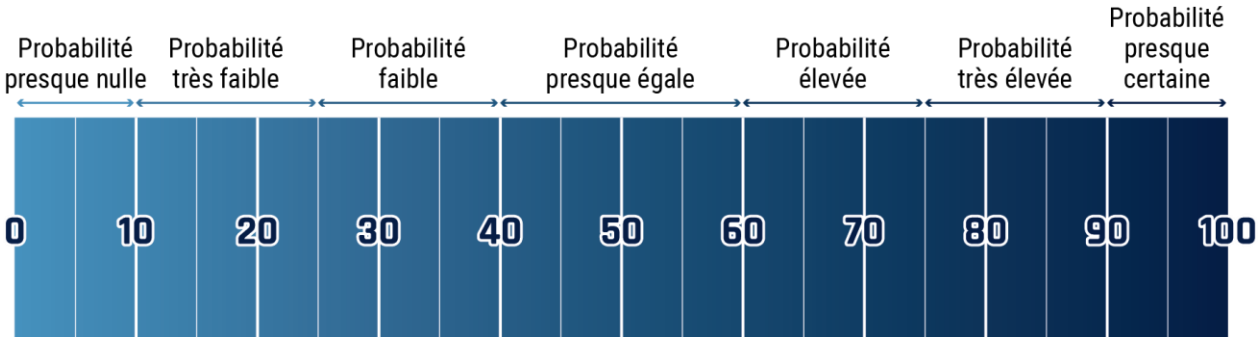
Prière de transmettre toute question ou tout enjeu relatif au présent document au Centre canadien pour la cybersécurité (Centre pour la cybersécurité) à contact@cyber.gc.ca.

Méthodologie et fondement de l'évaluation

Les avis énoncés dans la présente évaluation sont fondés sur les connaissances et l'expertise du Centre pour la cybersécurité en matière de cybersécurité. En défendant les systèmes d'information du gouvernement du Canada, le Centre pour la cybersécurité bénéficie d'une perspective unique lui permettant d'observer les tendances dans l'environnement de cybermenaces et d'appuyer ses évaluations. Le volet du mandat du Centre de la sécurité des télécommunications (CST) touchant le renseignement étranger procure au Centre pour la cybersécurité de précieuses informations sur le comportement des adversaires dans le cyberespace. Bien que le Centre pour la cybersécurité soit toujours tenu de protéger les sources et méthodes classifiées, il fournira au lectorat, dans la mesure du possible, les justifications qui ont motivé ses avis.

Les avis du Centre pour la cybersécurité sont basés sur un processus d'analyse qui comprend l'évaluation de la qualité de l'information disponible, l'étude de différentes explications, l'atténuation des biais et l'usage d'un langage probabiliste. Le Centre pour la cybersécurité utilise des formulations telles que « nous évaluons que » ou « nous jugeons que » pour présenter une évaluation analytique. Les qualificatifs tels que « possiblement », « probable » et « très probable » servent à évoquer une probabilité.

Le présent document est basé sur des renseignements disponibles en date du 26 juin 2023.



Introduction

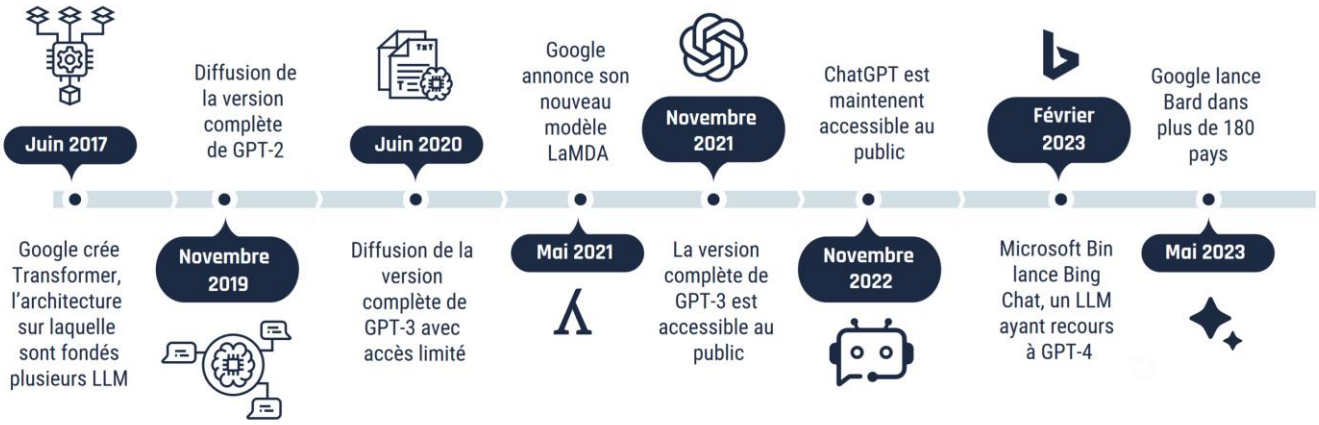
Depuis au moins 2016, l'intelligence artificielle (IA) générative permet de générer du contenu synthétique réunissant du texte fictif, de fausses images, ainsi que des fichiers audio et vidéo falsifiés. Ce contenu synthétique peut servir dans le cadre de campagnes de désinformation pour manipuler secrètement l'information en ligne et, du même coup, influencer les opinions et les comportements. L'IA générative est de plus en plus accessible au public et à une multitude d'auteures et auteurs de cybermenace parrainés ou non par des États. Nous estimons que les Canadiennes et Canadiens qui utilisent les médias sociaux seront fort probablement exposés au contenu synthétique[1]. Par conséquent, les modèles de langage de grande taille (LLM pour *Large Language Model*) représentent une menace grandissante pour l'écosystème d'information du Canada, les secteurs canadiens des médias et des télécommunications, et les structures dans lesquelles l'information est créée, partagée et transformée.

La petite histoire des modèles de langage de grande taille (LLM)

En juin 2017, les chercheuses et chercheurs de Google ont proposé une nouvelle architecture de réseau de neurones artificiels appelée Transformer. Il s'agissait d'un modèle révolutionnaire que l'on pouvait entraîner plus rapidement et qui exigeait moins de données d'entraînement[2]. C'est sur cette architecture que sont basés les autres LLM qui sont apparus plus tard, comme le modèle de transformateur génératif pré-entraîné (GPT pour *Generative Pre-Trained Transformer*) d'OpenAI. En août 2019, l'entreprise de recherche et de développement en intelligence artificielle OpenAI a publié une version partielle de son transformateur génératif pré-entraîné 2 (GPT-2 pour *Generative Pre-trained Transformer 2*), un modèle de langage capable de générer des paragraphes de texte cohérent qu'il est pratiquement impossible de distinguer d'un texte rédigé par un être humain[3]. OpenAI a d'abord lancé une version très restreinte du modèle, craignant que ce dernier serve possiblement à réduire les coûts liés à la génération de faux contenu et à la conduite de campagnes de désinformation[4].

Malgré les préoccupations liées aux applications malveillantes des LLM, les géants du numérique comme OpenAI, Google, Meta et Microsoft ont poursuivi le développement des outils de génération de texte[5]. Le 18 mai 2021, Google a annoncé le lancement de LaMDA, un modèle entraîné au dialogue capable de faire beaucoup plus que de composer du texte[6]. OpenAI a publié son plus récent modèle GPT-3 en novembre 2021, mais l'entreprise n'a pas été en mesure de lancer, l'année suivante, une mise à jour de la version GPT-3.5, le LLM derrière le populaire ChatGPT. ChatGPT est devenu l'un des logiciels grand public à la croissance la plus rapide en raison de son accessibilité, de sa polyvalence et de son exactitude dans l'exécution d'une multitude de tâches[7]. Les autres entreprises de technologie ont rapidement emboîté le pas, publiant des LLM similaires et accessibles à partir d'interfaces conviviales, comme l'agent conversationnel Bard de Google et Bing Chat de Microsoft, qui font appel à GPT-4.

Figure 1 – Chronologie des modèles de langage de grande taille (LLM)



Contexte des menaces : Les menaces les plus probables

Campagnes d'influence en ligne : Avant les générateurs de texte basés sur des LLM, les campagnes d'influence en ligne exigeaient que des êtres humains produisent le contenu et propagent la désinformation pour influencer les croyances et les comportements. Les générateurs de texte basés sur des LLM comme ChatGPT aident ou remplacent les rédactrices et rédacteurs humains afin de produire en masse des documents, des commentaires et des discussions qui visent à diffuser ou à amplifier la mésinformation ou la désinformation. Nous estimons que le Canada peut être particulièrement vulnérable aux campagnes d'influence en ligne faisant appel aux LLM en raison de la grande consommation de contenu de médias sociaux des Canadiennes et des Canadiens[8].

Campagnes d'hameçonnage par courriel : Les LLM génèrent du texte synthétique à propos d'un sujet en particulier et dans un style particulier. Les progrès réalisés sont tels qu'ils arrivent maintenant à produire du contenu qu'il est souvent quasi impossible de distinguer d'un texte rédigé par un être humain[9]. Les auteures et auteurs de cybermenace peuvent taper du texte dans les générateurs de texte à l'invite pour rédiger rapidement des courriels d'hameçonnage ciblés en vue de voler, entre autres, l'information sensible, les justificatifs d'identité ou les données financières des victimes.

Humain ou machine : Nous estimons que les outils de détection de l'apprentissage automatique actuels sont très probablement incapables de reconnaître le texte généré par des LLM. Selon nos observations, étant donné l'absence d'outils de détection de contenu synthétique efficace et la plus grande disponibilité des générateurs de texte basés sur des LLM, il est probable que les campagnes d'influence en ligne visant à répandre la désinformation soient de plus en plus difficiles à détecter, paraissent authentiques ou soient produites à large échelle, rendant ainsi impossible la reconnaissance manuelle. On estime également qu'il est très probable que les améliorations apportées à ces technologies fassent en sorte que les êtres humains aient plus de difficulté à les détecter, ce qui affectera du même coup la capacité des entreprises de médias sociaux à détecter et à supprimer le contenu synthétique.

Menaces potentielles improbables

Code malveillant : Comme les LLM peuvent écrire des extraits de code dans les langages de programmation populaires, dont JavaScript, Python, C#, PHP et Java, les personnes voulant créer du code peuvent y arriver sans avoir à démontrer de solides compétences[10]. Certains générateurs de texte basés sur des LLM disposent d'une fonction de codage que les auteures et auteurs de cybermenace pourraient exploiter pour créer du nouveau code malveillant[11]. Nous estimons toutefois qu'il est improbable que les générateurs de texte basés sur des LLM puissent être utilisés pour créer du code sophistiqué donnant lieu à une attaque du jour zéro.

Empoisonnement des jeux de données : Les LLM sont entraînés à partir de jeux de données linguistiques de grande taille. En théorie, les auteures et auteurs de cybermenace pourraient injecter des données ou modifier les données utilisées pour entraîner les versions plus récentes des LLM de manière à altérer l'exactitude et la qualité des données générées. Par contre, en raison de la grande taille et de la nature exclusive des jeux de données, nous jugeons que l'empoisonnement de ces jeux de données volumineux est très peu probable.

Risques pesant sur les organisations

Les organisations qui utilisent des générateurs de texte basés sur des LLM ou d'autres générateurs basés sur l'apprentissage automatique pour mener leurs activités pourraient affaiblir leurs responsabilités en ce qui a trait à l'intendance des données ou éluder les structures qui protègent l'information sensible. Elles pourraient utiliser les générateurs de texte basés sur des LLM pour :

- effectuer des recherches;
- commenter les résultats;
- compiler des statistiques;
- rédiger des courriels;
- produire des rapports internes.

Vous trouverez ci-dessous quelques exemples de risques particuliers associés à l'utilisation de ces générateurs de texte.

Gouvernance des données : Les générateurs de texte basés sur des LLM exigent une saisie ou une intervention de la part de l'utilisatrice ou de l'utilisateur. Le texte saisit par une employée ou un employé pourrait donc contenir de l'information relevant d'une organisation. Dans le but de générer le texte de sortie voulu, les données saisies sont transférées sur des systèmes sur lesquels l'organisation n'a aucun contrôle et qui sont sous la garde du fournisseur de service. Ces données peuvent également être réintroduites dans le LLM ou stockées à d'autres fins. Une utilisation non autorisée des outils en ligne expose l'information de tiers et va à l'encontre des exigences en matière de gouvernance des données organisationnelles.

Sécurité de l'information protégée : Les membres d'une organisation qui font des saisies dans des générateurs de texte basés sur des LLM pourraient également divulguer sans le savoir de l'information sensible, comme des renseignements personnels et de l'information commerciale confidentielle, à l'extérieur des cadres politiques et de sécurité approuvés. Par exemple, faire appel à un générateur de texte basé sur des LLM pour rédiger une réponse à la demande d'une cliente ou d'un client pourrait faire en sorte que l'on utilise ses renseignements personnels à l'extérieur des réseaux approuvés ou à d'autres fins que celles pour

lesquelles l'information a été recueillie, augmentant du coup les risques que cette information fasse l'objet d'une fuite par l'entremise d'une tierce partie.

Principaux termes

Les **réseaux de neurones artificiels** sont des modèles polyvalents que l'on peut entraîner pour exécuter et automatiser des tâches complexes très pointues, comme générer des vidéos réalistes d'événements qui ne se sont jamais produits (ce qu'on appelle communément des hypertrucages). Ils peuvent reconnaître et apprendre les liens et les modèles qui existent dans des jeux de données extrêmement grands, puis construire des représentations complexes de ces données. Cela en fait un composant essentiel des modèles de langage de grande taille pouvant générer des médias synthétiques persuasifs.

Les **modèles de langage de grande taille (LLM)** sont des réseaux de neurones artificiels que l'on entraîne au moyen de jeux de données linguistiques très volumineux en faisant appel à un apprentissage autosupervisé ou semi-supervisé. Par le passé, les LLM généraient le texte en prédisant le mot suivant, mais ils peuvent maintenant utiliser les phrases entières fournies par les utilisatrices et utilisateurs dans des invites ou générer des documents entiers sur un sujet donné. L'entraînement basé sur des jeux de données exceptionnellement grands permet au modèle d'apprendre une structure linguistique sophistiquée, mais aussi les biais ou les inexactitudes que l'on trouve dans ces données.

L'**apprentissage automatique** est un domaine de recherche axé sur les méthodes qui permet aux machines d'apprendre comment effectuer une tâche à partir des données fournies sans avoir à programmer explicitement une solution étape par étape. Comme les modèles d'apprentissage automatique peuvent souvent faire aussi bien, voire mieux qu'un être humain pour certaines tâches, on considère l'apprentissage automatique comme étant une sous-discipline de la recherche sur l'intelligence artificielle.

Par **contenu synthétique**, on entend le contenu généré par une machine sans assistance humaine ou avec une assistance humaine très limitée.

Les **campagnes d'influence en ligne** surviennent lorsque des auteures ou auteurs de menace parviennent secrètement à créer, à diffuser ou à amplifier la mésinformation ou la désinformation en vue d'influencer les croyances ou les comportements.

[1] La plupart des Canadiennes et Canadiens ont vu une forme quelconque de contenu synthétique dans les médias sociaux en raison 1) du volume important de contenu synthétique circulant dans les médias sociaux et 2) de leur grande consommation de contenu de médias sociaux. Les chercheuses et chercheurs de la Queensland University of Technology ont découvert qu'en moyenne, plus de 3,2 milliards de photos et 720 000 heures de vidéos sont créées chaque jour et mises en ligne. Ils ont remarqué qu'une grande partie de ce contenu en ligne consistait en des médias synthétiques partagés dans les médias sociaux. En 2018, 78 % de la population canadienne a utilisé au moins un compte de réseautage social et depuis janvier 2021, on estime que 67,1 millions d'utilisatrices et utilisateurs canadiens utilisent les plateformes de médias sociaux Facebook, Instagram, Twitter, TikTok, WeChat et YouTube. Consulter le rapport de Sebastien Charlton et de Kamille Leclair intitulé *Digital News Report: Canada 2019 Data Overview* (en anglais seulement), Centre d'études des médias, Département d'information et de communication, Université Laval, février 2019; Schimmele et al. *Étude : Évaluations que font les Canadiens des médias sociaux dans leur vie*, *Statistique Canada,* 24 mars 2021; T.J. Thompson et al. *Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities* (en anglais seulement), Journalism Practice, Research Gate, octobre 2020.

[2] Ashish Vaswani et al. *Attention Is All You Need* (en anglais seulement), Google Brain and Google Research, 12 juin 2017.

[3] OpenAI Blog, *Better Language Models and Their Implications* (en anglais seulement), 14 février 2019,

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

[5] Eray Eliaçık, *The role of large language models in the AI war* (en anglais seulement), Data Economy, 27 février 2023.

[6] Eli Collins, *LaMDA: our breakthrough conversation technology* (en anglais seulement), Google: The Keyword, 18 mai 2021.

[7] On estime que l'application ChatGPT comptait 100 millions d'utilisatrices et utilisateurs actifs en janvier 2023, seulement deux mois après son lancement. UBS Wealth Management, *Let's chat about ChatGPT* (en anglais seulement), 23 février 2023.

[8] En janvier 2021, on estime que 67,1 millions d'utilisatrices et utilisateurs canadiens utilisaient les plateformes de médias sociaux Facebook, Instagram, Twitter, TikTok, WeChat et YouTube. En 2019, près de 50 % de la population canadienne âgée de 18 à 24 ans utilise les médias comme principale source de nouvelles. Voir Schimmele et al. *Étude : Évaluations que font les Canadiens des médias sociaux dans leur vie*, *Statistique Canada,* 24 mars 2021.

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

[10] Amber Isrelsen, *How to use ChatGPT to write code* (en anglais seulement), Pluralsight Blog, 22 mars 2023.

[11] CheckPoint Research, *OPWNAI: Cybercriminals Starting to Use ChatGPT* (en anglais seulement), 6 janvier 2023.