L'intelligence artificielle générative

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ
Juillet 2023 | ITSAP.00.041

Plusieurs organisations font appel à l'intelligence artificielle (IA) pour optimiser leurs processus, analyser les données, assurer le diagnostic et le traitement de patients et personnaliser l'expérience de leurs utilisatrices et utilisateurs.

L'**IA générative** est un type d'intelligence artificielle qui génère du nouveau contenu en modélisant les caractéristiques des données tirées des grands jeux de données qui alimentent le modèle. Alors que les systèmes d'IA traditionnels peuvent reconnaître les modèles ou classifier le contenu existant, l'IA générative peut créer du nouveau contenu sous plusieurs formes, comme du texte, une image, un fichier audio ou du code logiciel.

Les modèles de langage de grande taille (LLM pour *Large Language Model*) sont une catégorie de l'IA générative qui s'est grandement améliorée au cours des dernières années. Pour créer du contenu, un ensemble de paramètres est intégré au LLM par l'entremise d'une requête ou d'une invite. Il est plus facile pour les utilisatrices et utilisateurs de générer du contenu, puisque les outils d'IA générative interagissent avec eux en mode de conversation au moyen d'invites. Depuis le début de 2022, plusieurs LLM (ChatGPT d'OpenAI et LaMDA de Google) et services ayant recours à des LLM (Bard de Google et Bing de Microsoft) ont retenu l'attention du public à travers le monde. Nombreux sont ceux et celles qui explorent les utilisations possibles de l'IA générative dans la foulée du grand intérêt qu'elle suscite. Cette publication fournit de l'information sur les risques liés à l'IA générative et les mesures d'atténuation qu'il est possible de prendre à cet égard.

De quelle façon utilise-t-on l'IA générative?

L'IA générative est une technologie à la fois transformatrice et perturbatrice qui peut changer considérablement la façon dont les consommatrices, les consommateurs, les industries et les entreprises mènent leurs activités. Elle a le potentiel d'offrir la créativité et l'innovation nécessaires pour améliorer les services et les activités commerciales. Certaines des applications utiles de l'IA générative touchent les domaines ci-dessous:

Soins de santé: Elle aide les fournisseurs de soins de santé à poser des diagnostics plus rapides. Elle leur permet également de personnaliser les plans de traitement. Elle peut servir à atteindre des cibles thérapeutiques et à découvrir de nouveaux médicaments potentiels.

Entreprises: Elle crée des communications personnalisées avec les clients existants et potentiels, et génère des modèles de vente prédictifs visant à prédire le comportement des clients.

Éducation: Elle permet aux éducatrices et aux éducateurs de créer des plans d'apprentissage personnalisés pour les étudiantes et étudiants en fonction de leur rendement, de leurs besoins et de leurs intérêts. Elle pourrait aider le personnel enseignant à offrir un meilleur soutien à la population étudiante.

Développement logiciel: Elle facilite le débogage et permet aux développeuses et développeurs de logiciels de générer du code de produire des extraits de code. Il est ainsi possible d'accélérer le développement et la diffusion des produits logiciels.

Publication et médias: Elle permet aux créatrices et créateurs de produire du contenu unique qui pourra être utilisé dans des campagnes de marketing, des publications, des émissions de télévision et des productions vidéo. Le contenu sur demande peut être généré rapidement et avec peu de ressources, ce qui permet de réduire considérablement les coûts.

Cybersécurité: Elle facilite l'amélioration des outils de cyberdéfense contre les rançongiciels et les autres attaques. Elle aide les praticiennes et praticiens de la sécurité à analyser plus facilement les grands jeux de données afin de relever les menaces et de minimiser les faux positifs en filtrant les activités non malveillantes.

Marché en ligne: Elle permet aux agents conversationnels de fournir des réponses semblables à l'humain, ce qui aide les organisations à améliorer le service à la clientèle et à réduire les coûts de soutien.

Prière de consulter la publication Intelligence artificielle (ITSAP.00.040)

Centre de la sécurité des télécommunications
Communications Security Establishment
Canada

L'intelligence artificielle générative

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ
Juillet 2023 | ITSAP.00.041

Quels sont les risques liés à l'IA générative?

Alors que les capacités technologiques de l'IA générative présentent de grandes possibilités, elles sont également source de préoccupations. L'IA générative peut aider les auteurs de menace à développer des exploits malveillants et à accroître potentiellement l'efficacité de leurs cyberattaques. Le fait qu'elle puisse permettre aux auteurs de menace d'exercer une influence considérable suscite de grandes inquiétudes. Par exemple, une manipulation délibérée du code sous-jacent et des outils qui l'utilisent risque de permettre à une menace interne de s'attaquer à la chaîne d'approvisionnement, et ce, de l'étape de la conception jusqu'à celles de la distribution et de la correction des logiciels. Vous trouverez ci-dessous certains des risques auxquels on doit porter attention:

- **Mésinformation et désinformation**: Le contenu peut ne pas être clairement identifié comme étant généré par IA et pourrait susciter de la confusion (mésinformation) ou de la déception (désinformation). Les auteurs de menace peuvent y avoir recours pour commettre des fraudes et mener des campagnes frauduleuses contre des personnes et des organisations.
- **Hameçonnage**: Les auteurs de menace peuvent concevoir automatiquement des attaques par harponnage plus fréquentes et faire appel à un degré plus élevé de sophistication. Des courriels d'hameçonnage ou des messages d'escroquerie très réalistes pourraient mener à des vols d'identité, à de la fraude financière ou à d'autres formes de cybercrimes.
- **Confidentialité des données**: Les utilisateurs peuvent fournir par inadvertance des données organisationnelles sensibles ou de l'information nominative (PII pour *Personally Identifiable Information*) dans des requêtes et des invites. Les auteurs de menace pourraient collecter cette information sensible en vue d'usurper l'identité d'une personne ou de répandre de la fausse information.
- **Code malveillant**: Des auteurs de menace possédant des compétences techniques peuvent contourner les restrictions dans les outils d'IA générative pour créer des maliciels et les utiliser lors de cyberattaques ciblées. Ceux qui n'ont que peu ou pas d'expérience en codage peuvent faire appel à l'IA pour rédiger facilement des maliciels fonctionnels qui pourraient nuire à une entreprise ou à une organisation.
- **Code entaché d'erreurs**: Les développeuses et développeurs de logiciels peuvent introduire volontairement ou involontairement du code non sécurisé ou entaché d'erreurs dans le pipeline de développement. Ce pourrait être le cas, notamment, s'ils omettent de mettre en place des mesures de traitement des erreurs et des vérifications de sécurité adéquates, ou si ces mesures sont mises en œuvre de façon inappropriée.
- **Jeux de données empoisonnés**: Les auteurs de menace peuvent injecter du code malveillant dans le jeu de données servant à entraîner le système d'IA générative, ce qui risque d'avoir une incidence négative sur la précision et la qualité des données générées. Cela pourrait également accroître les risques d'attaques à grande échelle de la chaîne d'approvisionnement.
- **Contenu biaisé**: Une grande partie du jeu de données d'entraînement alimenté dans les LLM provient de l'Internet ouvert. Ainsi, le contenu généré fait l'objet d'un biais fondamental, puisque seule une petite partie de toutes les données à travers le monde sont accessibles en ligne et peuvent être utilisées aux fins de l'IA. Le contenu généré peut également être préjudiciable si le jeu de données d'entraînement n'offre pas une représentation équitable des points de données.
- **Perte de propriété intellectuelle**: Les outils d'IA générative peuvent permettre aux auteurs de menace dotés de moyens sophistiqués de voler les données plus rapidement et en lot. Une perte de propriété intellectuelle (c'est-à-dire, des renseignements commerciaux exclusifs, des données sujettes au droit d'auteur, du code de logiciel ou les données d'essais de médicaments) peut porter atteinte à la réputation de votre organisation, à ses revenus et à sa croissance future.

Faites preuve de vigilance

L'IA générative est une technologie qui appartient au domaine de l'apprentissage automatique plutôt qu'à celui du véritable "renseignement". Elle ne comprend pas les concepts, mais produit du contenu correspondant à la meilleure réponse possible d'un point de vue statistique afin de l'utiliser dans une invite ou une requête.

N'oubliez pas que les résultats peuvent être:

- erronés
- illogiques
- non éclairés
- biasés

Centre de la sécurité des télécommunications
Communications Security Establishment

Canada

L'intelligence artificielle générative

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ
Juillet 2023 | ITSAP.00.041

Comment peut-on atténuer les risques?

L'IA générative est un autre outil auquel les auteurs de menace peuvent faire appel pour lancer leurs cyberattaques. Comme cette technologie est de plus en plus utilisée et exploitée, il est probable qu'elle mène à une augmentation du nombre de cyberattaques sophistiquées, ce qui comprend l'hameçonnage, le piratage psychologique, la désinformation, la désinformation et le vol d'identité. Bien qu'il puisse être difficile de détecter (ou d'attribuer positivement) les cyberattaques qui tirent parti de l'IA générative, les organisations et les personnes peuvent se préparer aux défis grandissants que ces attaques pourraient poser.

Les **organisations** peuvent prendre les mesures ci-dessous pour atténuer les risques de compromission découlant de cyberattaques:

- ☐ **Mettez en place des mécanismes d'authentification rigoureux** – sécurisez les comptes et les dispositifs sur vos réseaux au moyen de l'authentification multifacteur (AMF) afin de prévenir l'accès non autorisé à vos ressources et à vos données sensibles les plus précieuses. Pour de plus amples renseignements, prière de consulter les documents Sécurisez vos comptes et vos appareils avec une authentification multifacteur (ITSAP.30.030) et Étapes à suivre pour déployer efficacement l'authentification multifacteur (ITSAP.00.105).
- ☐ **Appliquez les correctifs et les mises à jour de sécurité** – activez les mises à jour automatiques sur l'équipement de TI et corrigez les vulnérabilités exploitables connues le plus tôt possible. Il sera ainsi possible d'empêcher les maliciels générés par IA d'infecter le réseau.
- ☐ **Restez à l'affût** – restez au courant des dernières menaces et vulnérabilités liées à l'IA générative et prenez des mesures proactives pour les atténuer.
- ☐ **Protégez votre réseau** – utilisez les outils de détection du réseau pour surveiller et analyser les activités anormales sur le réseau. Il est ainsi possible d'identifier rapidement les incidents et les menaces afin de prendre les mesures d'atténuation appropriées. Explorez également les façons dont l'IA pourrait être déployée d'un point de vue défensif dans les outils de protection des réseaux et envisagez toutes les conséquences. Pour de plus amples renseignements, prière de consulter les documents Journalisation et surveillance de la sécurité de réseau (ITSAP.00.085) et Les 10 mesures de sécurité des TI: No 5, Segmenter et séparer l'information (ITSM.10.092).
- ☐ **Formez votre personnel** – enseignez à toutes les utilisatrices et à tous les utilisateurs comment reconnaître les signes précurseurs d'une attaque par piratage psychologique et avec qui communiquer pour gérer ces situations en toute sécurité. Ces procédures devraient fournir aux utilisatrices et aux utilisateurs un moyen facile de signaler les attaques par hameçonnage ou les communications suspectes.

Les **personnes** peuvent prendre les mesures ci-dessous pour protéger leurs données personnelles contre les attaques par hameçonnage:

- ☐ **Vérifiez le contenu** – comme de plus amples données sont maintenant accessibles, il pourrait être difficile d'établir qui est responsable du contenu ou dans quelle mesure il est logique et basé sur les faits. Il est important de lire le contenu et de relever toute indication qu'il a été produit par un outil d'IA générative. Passez en revue le contenu généré et prenez le temps de vérifier les faits en consultant des sources crédibles. Pour de plus amples renseignements, prière de consulter le document Repérer les cas de mésinformation, désinformation et malinformation (ITSAP.00.300).
- ☐ **Adoptez des pratiques exemplaires de base en cybersécurité** – restez à l'affût, utilisez des mots de passe robustes et activez l'authentification à deux facteurs pour protéger les comptes en ligne. Assurez-vous d'appliquer les plus récentes mises à jour logicielles, d'utiliser des antivirus et d'éviter les réseaux Wi-Fi publics.
- ☐ **Limitez l'exposition au piratage psychologique ou à la compromission de courriel d'affaires** – mettez en place des pratiques de sécurité en ligne de base, comme:
  - réduire la quantité de renseignements personnels publiés en ligne
  - éviter d'ouvrir des pièces jointes et de cliquer sur les liens dans des courriels envoyés par des sources inconnues
  - communiquer par l'entremise d'un autre canal vérifié
  - faire preuve de prudence lorsque des personnes vous appellent pour demander de l'information sensible.

Pour de plus amples renseignements, prière de consulter les documents Ne mordez pas à l'hameçon: Reconnaître et prévenir les attaques par hameçonnage (ITSAP.00.101) et Qu'est-ce que l'hameçonnage vocal? (ITSAP.00.102).

Protéger la sécurité lors de l'utilisation d'outils d'IA générative

Les mesures de sécurité ci-dessous peuvent vous aider à générer du contenu fiable et de qualité tout en abordant les préoccupations en matière de protection de la vie privée:

- ☐ **Mettez en place des stratégies d'utilisation de l'IA générative** – ces stratégies devraient comprendre les types de contenu qui peuvent être générés et comment utiliser la technologie pour éviter de compromettre vos données sensibles. Vos stratégies devraient également tenir compte des processus de supervision et d'examen nécessaires pour assurer le bon fonctionnement de la technologie. Au moment de créer des solutions employant l'IA générative, assurez-vous que les exercices mènent à un comportement fiable basé sur l'éthique. Veillez également à mettre en place les stratégies rapidement et à les communiquer au personnel.
- ☐ **Sélectionnez les jeux de données d'entraînement avec soin** – obtenez les jeux de données depuis une source de confiance et mettez en place un processus rigoureux de validation et de vérification des jeux de données, qu'ils aient été acquis à l'externe ou développés en interne. Utilisez des données diverses et représentatives pour éviter tout contenu inexact et biaisé. Mettez en place un processus visant à ce que les résultats soient passés en revue par une équipe multidisciplinaire provenant de l'ensemble de votre organisation, qui s'efforcera de relever les biais inhérents au système. Peaufinez et réentraînez constamment le système d'IA en fonction de la rétroaction externe appropriée afin d'améliorer la qualité des résultats.
- ☐ **Choisissez des outils offerts par des fournisseurs axés sur la sécurité** – assurez-vous que vos fournisseurs ont adopté des pratiques de sécurité rigoureuses pour ce qui est des processus de collection, de stockage et de transfert de leurs données.
- ☐ **Faites attention à l'information que vous fournissez** – évitez de fournir de l'information nominative ou des données commerciales sensibles dans le cadre de vos requêtes ou dans les invites. Déterminez si l'outil permet à vos utilisatrices et utilisateurs de supprimer l'historique de leur invite de recherche.

Centre de la sécurité des télécommunications
Communications Security Establishment

 No de cat.D97-1/00-041-2023F-PDF ISBN 978-0-660-48573-7

Canada