ChatGPT : Un risque de divulgation de données personnelles selon Google DeepMind

Les limites de l’IA générative : quand la sécurité est mise à mal

Les limites de l’IA générative : quand la sécurité est mise à mal

Image : Google DeepMind

Les chercheurs en intelligence artificielle (IA) trouvent de plus en plus de moyens de briser la sécurité des programmes d’IA générative comme ChatGPT. En particulier le processus dit “d’alignement”, qui vise à maintenir les programmes dans certaines limites afin qu’ils jouent leur rôle d’assistant sans avoir de comportement répréhensible. Récemment, un groupe de chercheurs de l’université de Californie avait notamment trouvé le moyen de “casser” ce processus d’alignement dans plusieurs programmes d’IA générative. Cette semaine, ce sont des chercheurs de l’unité DeepMind de Google qui ont trouvé un

moyen encore plus simple de briser l'alignement de ChatGPT.

Un nouveau type d'attaque

Pour ce faire, ils ont demandé à ChatGPT de répéter un mot indéfiniment. Ils ont ainsi réussi à forcer le programme à recracher des passages entiers de littérature contenant ses données d'entraînement. Or, ce type de fuite n'est pas censé se produire avec des programmes alignés. Plus inquiétant encore : le programme a également pu être manipulé pour livrer les noms, numéros de téléphone et adresses d'individus. Les chercheurs ont appelé ce phénomène la "mémorisation extractible". Il s'agit d'une attaque qui force un programme à divulguer les éléments qu'il a stockés en mémoire.

« Nous avons développé un nouveau type d'attaque, qui amène le modèle à s'écarter de ses générations de type chatbot et à émettre des données de formation à un taux 150× plus élevé que lorsqu'il se comporte correctement », décrit l'équipe de Milad Nasr dans l'article "Scalable Extraction of Training Data from (Production) Language Models". L'équipe a également rédigé un billet de blog, plus accessible (en anglais).

Le principe essentiel de cette attaque, c'est de faire en sorte que l'IA générative – ici, ChatGPT – s'écarte de l'alignement qui a été programmé pour revenir à un mode de fonctionnement plus basique.

Formation et alignement

Les programmes d'IA générative comme ChatGPT reposent sur un processus appelé "formation". Lors de la formation, le programme dans son état initial – plutôt informe – est soumis à des milliards d'octets de texte issus de sources diverses (pages internet telles que Wikipédia, livres publiés). L'objectif fondamental de cet entraînement est de faire en sorte que le programme reflète tout ce qui lui est donné, en compressant le texte puis en le décompressant. En théorie, une fois entraîné, un programme pourrait régurgiter les données

d’entraînement si un petit bout de texte de Wikipédia lui était soumis et déclenchait la réponse en miroir. Mais, pour éviter cela, ChatGPT et les autres programmes d’IA générative sont “alignés”. C’est-à-dire qu’ils reçoivent une couche supplémentaire de formation, de manière à ce qu’ils ne se contentent pas de recracher du texte. Ils doivent pouvoir répondre par des messages utiles. Par exemple, ils doivent savoir répondre à une question ou aider à la rédaction d’un résumé de lecture. Ce personnage d’assistant utile, créé par l’alignement, masque la fonction miroir sous-jacente.

Répéter des mots à l’infini fait diverger ChatGPT

Pour forcer ChatGPT à s’écarter de sa fonction d’assistant pratique, Milad Nasr lui a demandé de répéter des mots à l’infini. Le programme d’OpenAI a répété le mot plusieurs centaines de fois avant de diverger. Il a ensuite commencé à dériver vers des extraits de texte divers, sans queue ni tête. Mais le chercheur précise que son équipe a réussi à montrer « qu’une petite fraction de ce qui a été généré diverge vers la mémorisation : certains textes générés étaient directement copiés à partir des données de pré-entraînement » du programme.

Après un certain nombre de répétitions, ChatGPT dérive vers un texte dénué de sens qui révèle des bribes de ses données d’entraînement.

Image : Google DeepMind. Le texte dénué de sens finit par révéler des sections entières des données d’entraînement du programme (surlignées en rouge).

Les chercheurs ont ensuite dû déterminer si ces résultats étaient effectivement issus des données d’entraînement du programme. Ils ont donc compilé un énorme ensemble de données, appelé AUXDataSet, qui représente près de 10 téraoctets de données d’entraînement. Il s’agit d’une compilation de quatre ensembles de données d’entraînement différents qui ont été

utilisés par les plus grands programmes d'IA générative : The Pile, Refined Web, RedPajama et Dolma. Les chercheurs ont rendu cette compilation consultable à l'aide d'un mécanisme d'indexation efficace, afin de pouvoir comparer les résultats de ChatGPT aux données d'entraînement pour rechercher des correspondances.

Copie de textes, contenu inapproprié et fuite de données

Ils ont ensuite réalisé l'expérience – répéter un mot à l'infini – des milliers de fois, et ont recherché les résultats dans l'ensemble de données AUXDataSet des milliers de fois, afin de « mettre à l'échelle » leur attaque. « La plus longue chaîne de caractères extraite dépasse les 4 000 caractères », indiquent les chercheurs à propos des données qu'ils ont récupérées. Plusieurs centaines de parties mémorisées de données d'entraînement atteignent plus de 1 000 caractères. Dans les prompts contenant les mots « livre » ou « poème », des paragraphes entiers de romans et des copies complètes de poèmes ont été retrouvés, racontent les chercheurs.

Ces derniers ont également obtenu des contenus "Not Safe For Work" (NSFW) – c'est-à-dire potentiellement choquants – en particulier lorsqu'ils demandaient au modèle de répéter un mot étiqueté NSFW. L'équipe a aussi pu récupérer « des informations permettant d'identifier des dizaines de personnes ».

Sur 15 000 tentatives, environ 17% contenaient des « informations personnelles identifiables mémorisées » comme des numéros de téléphone.

Une expérience limitée aux résultats inquiétants

Les chercheurs ont pour but de quantifier les données d'entraînement qui pourraient fuiter. Ils ont pour l'instant trouvé de grandes quantités de données, mais leur recherche est limitée par le coût de l'expérience, qui pourrait se

prolonger indéfiniment. Avec leurs attaques répétées, ils ont déjà trouvé 10 000 cas de contenu « mémorisé » dans les ensembles de données régurgités. Et ils supposent qu'il serait possible d'en trouver beaucoup plus en poursuivant ces attaques.

L'expérience consistant à comparer les résultats de ChatGPT à ceux de AUXDataSet a été réalisée sur une seule machine dans Google Cloud, avec un processeur Intel Sapphire Rapids Xeon et 1,4 téraoctet de DRAM. Il a fallu des semaines pour mener à bien ce projet. Mais un accès à des ordinateurs plus puissants pourrait leur permettre de tester ChatGPT de manière plus approfondie et de trouver encore plus de résultats. « Avec notre budget limité de 200 dollars, nous avons extrait plus de 10 000 exemples uniques », souligne l'équipe, alertant que des cyberattaquants « dépensant plus d'argent pour interroger l'API de ChatGPT pourraient probablement extraire beaucoup plus de données ».

Les chercheurs ont vérifié manuellement près de 500 exemples issus des résultats de ChatGPT en effectuant des recherches Google : ils ont alors trouvé près de deux fois plus d'exemples de données mémorisées sur le web. Ainsi, il pourrait y avoir encore plus de données mémorisées dans ChatGPT que ce qui a pu être capturé dans l'AUXDataSet, malgré la taille de ce dernier.

Certains mots sont plus efficaces que d'autres

Fun fact : les chercheurs se sont aperçus que certains mots étaient plus efficaces que d'autres pour mener à bien leur expérience. La répétition du mot « poème », présentée plus haut, était en réalité l'une des moins efficaces. Le mot « entreprise » répété à l'infini, a contrario, a été le plus efficace, comme on peut le voir dans ce graphique présentant la puissance de chaque mot (ou groupement de lettres) utilisé.

Les chercheurs ne savent pas exactement ce qui conduit ChatGPT

à révéler les textes qu'il a mémorisés. Ils supposent que le programme a été entraîné sur un plus grand nombre « d'époques » que d'autres programmes d'IA générative, ce qui signifie que l'outil passe par les mêmes ensembles de données d'entraînement un plus grand nombre de fois. « Des travaux antérieurs ont montré que cela pouvait accroître considérablement la mémorisation », écrivent-ils. Malgré tout, demander au programme de répéter plusieurs mots ne fonctionne pas comme une attaque, relatent-ils, car ChatGPT refuse généralement de continuer. Les chercheurs ne savent pas pourquoi seules les invites à un seul mot fonctionnent : « Bien que nous n'ayons pas d'explication à ce sujet, l'effet est significatif et reproductible. »

Savoir poser les bonnes questions

Les auteurs ont communiqué leurs conclusions à OpenAI le 30 août dernier, et l'entreprise semble avoir pris des mesures pour contrer cette attaque. ZDNet a ainsi testé de…