Web FRA

L'EXPRESS

www.lexpress.fr

Famille du média : **Médias d'information générale (hors PQN)**
Audience : **5432332**
Sujet du média : **Actualités-Infos Générales**

1 Fevrier 2023
Journalistes : **Anne Cagan**
Nombre de mots : **1289**

Visualiser l'article

ChatGPT : comment détecter un texte écrit par l'intelligence artificielle ?

Face à la gronde des enseignants, OpenAI vient lancer un outil permettant de repérer les textes écrits par son intelligence artificielle. Des protections utiles mais pas infaillibles.

L'inscription ChatGPT sur le site d'OpenAI.

Jakub Porzycki / NurPhoto / NurPhoto via AFP

Le couperet est tombé à Sciences Po. Les étudiants ont la stricte interdiction d'utiliser des IA telles que ChatGPT lorsqu'ils réalisent leurs travaux. Enfreindre cette règle les expose à "des sanctions pouvant aller jusqu'à l'expulsion de l'établissement", souligne un récent mémo de l'école . Un rappel à l'ordre qui montre le désarroi des établissements scolaires face à l'intelligence artificielle d'OpenAI dont l'usage se répand comme une trainée de poudre chez les étudiants peu scrupuleux. Enseignant à la faculté de Lyon, Stéphane Bonvallet a ainsi découvert lors d'un récent devoir que 50 % des copies de ses étudiants avaient en réalité été rédigées par ChatGPT.

"Il ne s'agissait pas de copier-coller. Mais les copies étaient construites exactement de la même manière, explique-t-il dans *Le Progrès* . On y retrouvait les mêmes constructions grammaticales. Le raisonnement était mené dans le même ordre, avec les mêmes qualités et les mêmes défauts. Enfin, elles étaient toutes illustrées par un exemple personnel, relatif à une grand-mère ou un grand-père..." Des affaires appelées à se multiplier qui soulèvent toutes une même question : comment différencier les textes écrits par des humains de ceux rédigés par des IA ?

LIRE AUSSI >>

Daniel Susskind : "L'IA, une mauvaise nouvelle pour de nombreux cadres..."

La méthode la plus solide à date est d'agir en amont et de faire en sorte que les IA "paraphent" de manière invisible tous leurs textes. "On peut insérer des signatures indécelables par des humains mais parfaitement reconnaissables par des modèles entraînés pour cela. L'idée va être, par exemple, d'insérer un certain nombre de répétitions de lettres ou de séquences de lettres dans un texte généré et de laisser une marque. Un peu à la manière des acrostiches dans certaines poésies ou des informations textuelles encodées dans des images. Grossièrement, on peut y voir une forme de stéganographie", explique Djamé Seddah, maitre de conférences à Sorbonne Université, détaché à l'Inria Paris.

Les IA "signent" leurs textes

Une équipe de l'université du Maryland a publié fin janvier une prometteuse méthode d'intégration de ce type de "watermark" aux textes générés par une intelligence artificielle. Pour la résumer à grands traits, l'idée est de rétrécir arbitrairement le périmètre de mots dans lequel l'IA pioche pour générer une phrase. Prenons par exemple la suite de termes "j'aime les fruits tels que les...". La liste de termes dans lequel un ChatGPT va faire son prochain choix comprendra sans doute des mots tels que "pommes", "poires", "fraises", etc. L'idée est alors de diviser cette liste : d'un côté, une liste "verte" de mots autorisés, de l'autre, une liste noire" de mots arbitrairement interdits. Et de forcer l'IA à choisir ses termes dans les listes vertes.

En appliquant cette règle (avec quelques aménagements pour protéger la cohérence du texte), il devient possible de distinguer les productions humaines de celles des IA : les premières contiennent plus de termes issus listes noires que les secondes puisque l'IA est contrainte de privilégier les listes vertes. Une méthode efficace, fiable "et qui fonctionne même sur des textes très courts", précise à L'Express Tom Goldstein, professeur associé de l'université du Maryland, spécialiste de l'IA et coauteur de la publication.

LIRE AUSSI >> IA "génératives" : les artistes lancent la contre-offensive

Si le concepteur d'une IA n'intègre pas de lui-même des codes de détection, d'autres méthodes peuvent être employées pour démasquer les intelligences numériques. "De même que toute personne a une manière d'écrire, un style, ces modèles en possèdent également, du moins pour le moment", souligne Christophe Cerisara, chercheur du CNRS au Laboratoire lorrain de recherche en informatique et ses applications. En analysant la fréquence d'utilisation des mots, la manière dont la ponctuation est utilisée, les longueurs de phrases, il devient donc possible de distinguer des motifs qui se répètent dans les textes d'une IA, de la même manière que l'on retrouve des schémas dans ceux d'un Rabelais ou d'une Françoise Sagan.

Marie-France Marchandise, l'économiste inventée par ChatGPT

"Si l'on a une quantité suffisante de textes générés par une IA, on peut alors entraîner un outil à identifier ce qui les différencie de textes écrits par des humains", confirme Sasha Luccioni, chercheuse au sein de l'entreprise Hugging Face, qui a développé un détecteur de ce type spécialisé sur les productions d'OpenAI.

LIRE AUSSI >> Quand ChatGPT a tout faux au célèbre test de réflexion cognitive

Reste enfin les bourdes que les réponses des IA contiennent parfois. ChatGPT a beau être bluffant à bien des égards, il lui arrive de temps à autre de faire des réponses parfaitement absurdes. A David Cayla, économiste à l'université d'Angers qui lui demande de citer des économistes keynésiens français, ChatGPT évoque ainsi avec un aplomb comique l'existence d'une Marie-France Marchandise.

Quand ChatGPT invente des économistes femmes... pic.twitter.com/iKQQ3hdwCB
David Cayla (@dav_cayla) January 29, 2023

Quand l'économie le pousse dans ses retranchements en lui demandant des précisions sur cette experte au si savoureux aptonyme, l'IA persiste et fabrique de toutes pièces une carrière à son économiste imaginaire : "Marie-France Marchandise est une économiste française spécialisée en économie politique. Elle est connue pour ses travaux sur la théorie keynésienne et la macroéconomie. Elle est professeur émérite à l'université Paris-Nanterre et a publié plusieurs ouvrages sur ces sujets", notamment un "Macroéconomie : une approche keynésienne". Mais les IA ne se trompent pas toujours de manière aussi évidente. Et les techniques de détection d'IA ne sont pas infaillibles. OpenAI qui a lancé le 31 janvier un détecteur basé sur l'approche stylistique, en réponse aux inquiétudes grandissantes, le souligne lui-même. "Dans nos tests sur des textes anglais, notre détecteur identifie correctement les textes écrits par des IA dans 26 % des cas [...] mais il produit des faux positifs en labellisant à tort des productions humaines comme si elles provenaient d'IA dans 9 % des cas". L'approche stylistique fonctionne par ailleurs mal si le texte à analyser est court.

LIRE AUSSI >> "Vous en tuez un, il en réapparait dix" : sur Internet, le fléau des bots malveillants

Les techniques de "signatures", quant à elles, peuvent être mises à mal si l'utilisateur modifie significativement l'écrit généré par l'IA, par exemple en le paraphrasant extensivement. Le recours à "la traduction automatique ou la transformation du texte à l'aide d'homographes, des lettres d'un alphabet différent du nôtre qui ressemblent 'graphiquement' à nos propres lettres mais qui d'un point de vue informatique sont totalement différentes" pourraient également abîmer ces signatures, pointe Djamé Seddah.

Des modèles similaires à ChatGPT pourraient même être construits dans le seul but de transformer suffisamment les productions d'une IA pour les rendre indétectables, en quelques clics. La société et l'économie vont donc devoir s'adapter à moyen terme cette nouvelle donne. Et la sphère éducative doit dès maintenant s'y ajuster, en repensant les devoirs et les évaluations, mais aussi la formation des élèves à ces outils qui bouleverseront la manière dont les entreprises travaillent. Professeur d'économie et de technologie de l'information de l'université de Stanford, Erik Brynjolfsson, interrogé par Bloomberg , voit dans ChatGPT "l'équivalent de la calculette pour l'écriture". Il serait absurde de faire comme si nous vivions toujours dans un monde de bouliers.