

Le énigme de l'IA: OpenAI refuse de lancer son détecteur de textes produits par ChatGPT

Article concernant l'intelligence artificielle et ChatGPT

Il y a près de un an, OpenAI a élaboré un dispositif permettant d'identifier les écrits générés par ChatGPT, son célèbre outil d'intelligence artificielle, sans pour autant le rendre accessible publiquement. Ces informations ont été dévoilées par le [Wall Street Journal](#) lors de la parution d'un article le dimanche 4 août. Un porte-parole d'OpenAI a confirmé ces faits, révélés suite à des discussions avec des employés et la consultation de documents internes de la compagnie.

Le ChatGPT, un outil d'IA générative lancé en novembre 2022 et accessible gratuitement au grand public, a la capacité de générer des textes à la demande, que ce soit pour la rédaction de courriels, de messages sur les médias sociaux ou de dissertations. Cette commodité a éveillé des inquiétudes chez certains enseignants, comme indiqué dans un article du [Monde](#). La question de l'identification de l'auteur, humain ou ChatGPT, d'un texte préoccupe de nombreux experts. Malgré la mise à disposition de plusieurs outils de détection, les résultats demeurent mitigés. En 2023, OpenAI avait introduit

un outil censé repérer les textes produits par des IA, mais, jugeant son inefficacité, avait rapidement retiré celui-ci.

Toutefois, le Wall Street Journal révèle qu'OpenAI dispose depuis environ un an d'un mécanisme fiable pour reconnaître les textes rédigés par ChatGPT, avec une efficacité de 99,9%. Ce mécanisme repose sur une légère modification du processus de sélection des termes par ChatGPT, créant ainsi un schéma détectable par un outil spécifique. Pour chaque texte analysé, cet outil fournit un degré de probabilité qu'il ait été rédigé par ChatGPT.

Découvrez également |

Article exclusif pour nos abonnés

[ChatGPT : à l'université, un outil pédagogique ou un instrument de triche ?](#)

Ajouter à vos sélections

Risque de perte d'utilisateurs

Depuis deux ans, ce projet engendre des discussions internes au sein d'OpenAI concernant la décision de rendre cette technologie publique, révèle le Wall Street Journal. La compagnie est confrontée à un dilemme, entre son objectif. OpenAI's initiative to develop accessible AI tools for all, on which they have already made concessions, and their desire to broaden their user base. A 2023 study conducted among ChatGPT users showed that 69% of them feared that the implementation of a detection tool would lead to unfounded cheating accusations. Additionally, 30% of the surveyed users stated that they would use ChatGPT less if such a system were specifically deployed for this tool. OpenAI also feared that ChatGPT's writing quality would be compromised by altering its operation, but several tests have proven the

opposite.

An OpenAI spokesperson defended a cautious approach to the Wall Street Journal, highlighting the risks assessed by the company. She mentioned a potential impact on the ecosystem beyond OpenAI, notably fearing that this tool would disproportionately penalize non-native English speakers using ChatGPT as an assistant.

Following the Wall Street Journal's revelations, OpenAI updated a May article on their blog regarding online content authentication. The company claims that their tool is very effective and accurate even in the face of localized alterations, such as paraphrasing. However, they acknowledge that the system performs less well if the entire text is paraphrased, as in automatic translation into another language before being re-translated into the original language. OpenAI therefore believes that their detection tool is vulnerable to being circumvented by malicious actors.

Lire aussi |

Article reserved for our subscribers

[Artificial Intelligence: "Will Sam Altman succeed in maintaining OpenAI's dual mission?"](#)

Add to your selections

Le Monde

[Reuse this content](#)