Opposition à ChatGPT : l’énigmatique dispositif anti-fraude refusé par OpenAI

Discussion autour de l’outil anti-fraude de ChatGPT

Discussion autour de l’outil anti-fraude de ChatGPT

Depuis l’avènement de ChatGPT, de nombreux enseignants luttent pour différencier les authentiques dissertations de celles produites par une machine. OpenAI a la capacité de résoudre ce problème, mais hésite à le faire.

Crédit photo : Emiliano Vittoriosi sur Unsplash

Un débat crucial divise les équipes d’OpenAI depuis près de deux ans. Alors que ChatGPT est désormais utilisé dans divers domaines, y compris à l’école, la nécessité de détecter efficacement les textes générés par ChatGPT se fait de plus en plus pressante. Selon des sources du *Wall Street Journal*, la société éditrice du célèbre chatbot dispose d’un outil conçu spécifiquement pour ce besoin, mais ne sait pas encore comment l’utiliser.

Des documents obtenus par le média anglophone révèlent l’existence d’un outil capable de repérer les textes générés

par ChatGPT “avec une précision de 99,9 %” depuis près d’un an, mais que la société garde actuellement dans ses réserves. “Il pourrait être déployé en un clic de souris”, a déclaré une source.

Un balisage invisible sur les textes

Cependant, un tel outil risquerait de réduire considérablement l’intérêt du chatbot d’OpenAI. Un sondage cité par le *Wall Street Journal* indique en effet qu’un tiers des utilisateurs de ChatGPT pourraient être dissuadés par l’ajout d’une technologie “anti-triche”. D’autres questions animent également les équipes d’OpenAI. Par exemple, la crainte qu’un tel outil puisse pénaliser particulièrement les personnes utilisant ChatGPT pour des formulations en anglais, n’étant pas leur langue maternelle.

En réalité, un tel outil repose sur l’insertion discrète d’un “marquage” aux textes générés par ChatGPT. Invisible à l’œil nu, ce marquage serait facilement détectable par un outil de détection spécialisé. Cependant, cette approche pose des défis, car la diffusion d’un tel outil à un large public pourrait révéler comment le marquage est intégré dans le texte, rendant ainsi l’outil inefficace.

Certains soutiennent également qu’il serait simple de contourner ce marquage en convertissant un texte via Google Translate ou en ajoutant puis supprimant des emojis avec ChatGPT. De telles manipulations rendraient l’outil anti-fraude inefficace.

Profit ou transparence ?

Face à ce défi, certains enseignants usent déjà de stratagèmes pour détecter les textes générés par l’IA. Josh McCrain, professeur de science politique à l’université de l’Utah aux

États-Unis, raconte au Wall Street Journal qu'il a discrètement ajouté une demande d'intégrer des références à Batman dans une dissertation. Les élèves ayant copié-collé la requête de ChatGPT ont immédiatement été démasqués.

Au-delà de ces anecdotes amusantes, la question de l'outil anti-fraude de ChatGPT met en lumière la fracture entre transparence et profit qui divise les équipes d'OpenAI depuis un certain temps, rappelant en écho lointain ce qui avait conduit au licenciement de Sam Altman l'année précédente.

Pour en savoir plus : ChatGPT aurait volé la voix de Scarlett Johansson et le démenti fait grincer des dents