DetectGPT, le tueur de ChatGPT

Amusant au premier abord, mais plein de dangers

ChatGPT démontre à quel point l'intelligence artificielle peut être impressionnante et trompeuse, surtout dans son "raisonnement" et la production de textes. Cependant, il est bien connu que cela peut représenter un danger, notamment dans l'université et la triche. Et ce n'est même pas à mentionner son utilisation dans la création de jeux vidéo et dans d'autres domaines. C'est pourquoi de nombreuses écoles telles que Sciences Po ont décidé d'interdire son utilisation. Comment contrôler cela ?

Arme contre ChatGPT: DetectGPT

Voici un nouvel outil qui pourrait s'avérer essentiel pour beaucoup, y compris les professeurs dans les lycées et les universités : DetectGPT. Comme son nom l'indique, l'objectif est d'analyser un texte et de détecter l'utilisation d'IA. C'est développé à **l'université de Stanford**. Pour l'instant, seulement **disponible** en démo ici, les performances sont très bonnes et les résultats sont étonnants dans la détection de l'IA. Cependant, il n'est pour l'instant possible d'analyser que de petits morceaux de textes, mais la méthode s'améliorera avec le temps.

Besoin croissant de détection

La fluidité et les connaissances factuelles des modèles de langage de grande taille augmentent le besoin de systèmes

correspondants pour détecter si un morceau de texte a été écrit par une machine. Par exemple, les étudiants peuvent utiliser des LLMs pour compléter leurs devoirs, laissant les professeurs incapables d'évaluer précisément l'apprentissage des étudiants.

DetectGPT, une solution sans entraînement

Notre approche, appelée DetectGPT, ne nécessite pas l'entraînement d'un classificateur séparé, la collecte d'un ensemble de données de passages réels ou générés, ou le filigrane explicite du texte généré. Cela peut sembler compliqué à comprendre, mais en gros, DetectGPT a l'avantage de ne pas avoir besoin d'être entraîné comme une IA pour analyser correctement les textes.