IA et littérature : le logiciel Pangram est-il vraiment fiable ?

L'outil est réputé pour être le plus performant pour débusquer les textes générés par intelligence artificielle. Comment fonctionne-t-il ? En quoi se démarque-t-il des précédentes générations de détecteurs ? Et surtout, à quel point peut-on lui faire confiance ?

Pangram Labs et GPTZero affirment pouvoir identifier une prose artificielle dans plus de 99% des cas. (Pangram)

Attention : quand vous tapez «Pangram» dans Google, le site qui apparaît en tête des résultats est, le plus souvent et par la magie de la sponsorisation, un site du nom d'AI Detector, à la fiabilité pour l'heure très discutable. Un effet de manche qui, à tout le moins, témoigne du désir de certains de capitaliser sur la notoriété croissante du logiciel créé par la start-up new-yorkaise Pangram Labs.

Depuis quarante-huit heures, en effet, c'est sur la foi en l'efficacité de ce logiciel en ligne que repose l'événement littéraire de la rentrée : le (gros) soupçon d'un recours à l'intelligence artificielle pour l'élaboration du roman de Thélyson Orélien, publié chez Grasset, et jusqu'alors encensé par de nombreux critiques.

A vrai dire, Pangram n'est pas le seul logiciel du marché à promettre de distinguer la prose humaine de la prose artificielle. Avec l'essor des IA dites génératives – qui produisent au kilomètre des phrases ou des images par un jeu d'extrapolation statistique – beaucoup d'entreprises travaillent au développement de tels détecteurs.

La tâche pourrait sembler ardue : qu'est-ce qui différencie une suite de mots d'une autre suite de mots, et comment deviner si leur succession résulte des décisions d'un être humain, ou du suivi scrupuleux des règles identifiées par un «grand modèle de langage» (LLM) ? C'est la même question que se pose implicitement tout féru de littérature en lisant une page écrite par Colette, par Flaubert ou par son petit-neveu : chacun a une façon d'écrire, de rythmer son texte, d'abuser ou d'économiser les virgules, sa petite préférence pour telle ou telle figure de style, son petit penchant pour certains effets rhétoriques… C'est ainsi, d'ailleurs, qu'on peut «pasticher» un auteur, en s'efforçant de faire siennes ses marottes et ses manies.

Flairer les bizarreries à des kilomètres…

Avec l'essor des IA adossées à des LLM, Internet s'est rapidement mis à pulluler de textes débités au kilomètre par des machines. En 2022 ou en 2023, identifier ces proses assemblées en suivant *«les probabilités d'apparition du mot suivant»* ne nécessitait souvent qu'un peu d'attention et de jugeote, tant étaient fréquentes les phrases manifestement incongrues (les «hallucinations» des IA), et certains tics d'écriture hérités des contenus sur lequel les modèles étaient entraînés. L'utilisation compulsive de tirets cadratins par ChatGPT est rapidement devenue l'un de ces «drapeaux rouges» sur les réseaux sociaux – au grand dam des auteurs humains qui en faisaient depuis longtemps un usage légitime.

Les années passant, les LLM ont progressé. Mais de nombreux internautes ont appris à flairer ces proses un peu trop «artificielles», optimisées pour être claires, complètes, raisonnables. Une intuition qui tient à la configuration globale du texte, à la façon dont les idées sont organisées, le recours presque scolaire à certains connecteurs logiques pour passer d'une idée à une autre (les «*de plus», «par conséquent»…)*, ou le fait de systématiquement énoncer les objections éventuelles pour mieux les désamorcer ensuite. Si aucun de ces éléments ne saurait, isolément, démontrer l'intervention d'une IA, leur accumulation a peu de chances d'être le fruit du hasard.

«Je suis convaincu que les personnes qui travaillent beaucoup avec l'IA deviennent souvent capables de la détecter», s'amuse auprès de CheckNews le chercheur belge Filip Van Droogenbroeck. Une opinion qui peut être confortée par des travaux présentés courant 2025 dans un congrès viennois. Selon ces recherches dans une assemblée d'individus invités à classer des centaines de textes comme «humains» ou «artificiels», l'avis moyen des gros utilisateurs d'IA réussissait aussi bien cette tâche que les meilleures solutions logicielles du marché – avec un taux d'erreur inférieur à 1%. Mais les machines ont continué à progresser depuis lors…

Un entraînement sur des textes humains

Dès lors que les différentes IA adossées à des LLM ont leur style, il n'est guère difficile d'entraîner des machines à les reconnaître. Le plus simple est d'entraîner… Une intelligence artificielle, non pas spécialisée dans la génération de mots, mais dans la classification et la comparaison de textes. La première étape consiste à rassembler une vaste quantité de romans, articles scientifiques et courriers écrits avant l'ère des LLM, c'est-à-dire garantis 100% humain. Ils sont ensuite transmis, un à un, à des IA génératives missionnées pour paraphraser, réécrire ou améliorer cette prose, selon les lois statistiques apprises au cours de leur création (la fameuse probabilité qu'un mot succède à un autre dans un contexte donné). Troisième étape : transmettre chaque paire de texte (l'original humain et sa réécriture IA) à un programme qui va aléatoirement attribuer un score à différents paramètres de ces textes : longueur des phrases, règles de grammaire employées, stratégies narratives privilégiées, etc. La machine émet des verdicts d'authenticité (ou d'artificialité), et essaye de se corriger à chaque fois qu'elle se trompe. Après des millions d'essais, une cartographie virtuelle de la «façon d'écrire» des différentes IA est prête à l'emploi.

En démultipliant les variables étudiées et les corpus de textes, certains logiciels ont réussi à creuser la distance avec la concurrence. Certaines entreprises ont fait le choix de la spécialisation (littérature scientifique, prose française, etc.). D'autres aspirent à produire des détecteurs universels, efficaces sur des dizaines de langues différentes. Pangram Labs et GPTZero sont probablement, à date, les acteurs les plus proéminents du marché. Ils affirment depuis de longs mois pouvoir identifier une prose artificielle dans plus de 99% des cas. Il était toutefois difficile de prêter aveuglément foi aux rapports techniques internes de ces entrepreneurs. Mais des évaluations réalisées en 2025 sur ces outils par des équipes indépendantes ont confirmé le baratin commercial.

Les détecteurs d'images générées par IA sont-ils fiables ?

Par ailleurs, plus l'échantillon de texte est long, plus ces logiciels ont une matière conséquente pour juger s'il correspond, ou non, à ce qui ressemble à un texte «amélioré» par une IA, selon les critères de qualité implicitement encodés dans le modèle de langage utilisé. Pangram sait quel ton, quel rythme, quels arguments seront utilisés par un «Claude» s'efforçant d'empiler les briques d'un

mémoire universitaire, ou par un «ChatGPT» qui essaye de cocher les cases du roman contemporain à succès.

L'aplomb avec lequel Pangram rend ses verdicts est déroutant. *«Nous pensons que ce texte,* débute le logiciel, *a été intégralement rédigé par l'humain»,* dans le cas le plus rassurant. Mais d'autres variantes sont possibles : *«…a été rédigé principalement par une IA, avec quelques passages écrits par un humain», «…est un mélange de contenu généré par l'IA et rédigé par l'humain», «…a été intégralement généré par une IA».* Un pourcentage s'affiche également à l'écran. Il ne correspond pas à une estimation de probabilité que le texte évalué soit ou ne soit pas «artificiel», mais à la proportion de texte constitué d'un texte possédant les signes caractéristiques d'une prose inféodée à un LLM.

Comme nous l'écrivions dans l'article consacré à la polémique Thélyson Orélien, le taux de «faux positifs» de Pangram (le risque d'identifier à tort un texte humain comme généré artificiellement) est le plus bas du marché. Les évaluations disponibles sur le dernier modèle accessible, Pangram 4, estime ce pourcentage d'erreur à 0,0024% pour les textes en langue française.

Des poèmes d'Hugo jugés «100 % IA» ?

Sur les réseaux sociaux, de nombreux utilisateurs affirment pourtant avoir mis à l'épreuve Pangram avec des textes tirés de la littérature française classique, ou leur propre prose adolescente pré-LLM, et avoir été identifié comme «IA». Toutefois, les captures d'écran associées portent soit sur des échantillons de textes très courts (pour lesquels Pangram associe son verdict à un *«[niveau de] confiance limitée»*), soit… Sur un logiciel en ligne qui n'est pas Pangram. Le plus souvent, le fameux *«premier résultat sponsorisé dans Google»,* évoqué en introduction de l'article.

Sur le réseau X, le cofondateur de Pangram Labs s'est amusé de la situation : *«Quel dommage que ces faux positifs ne soient trouvés que sur du texte privé, non publié. Je mets en jeu ma prime de 100 dollars, payable à vous ou à une association caritative de votre choix, si vous pouvez partager le texte complet et fournir une preuve qu'il a été écrit en 2020.»*

Par ailleurs, les créateurs du fameux logiciel ne prétendent pas que leur solution est absolument infaillible, mais que la probabilité d'erreur a été mesurée de manière indépendante, et est extrêmement faible. Lorsque la prose signée d'un même auteur, dans différents contextes et à différentes années d'intervalle est systématiquement jugée artificielle par cette machine, douter de la qualité de sa plume est donc permis.

Il faut toutefois faire très attention au sens des mots. Un texte *«100 % écrit par l'intelligence artificielle»* n'est pas nécessairement un texte intégralement imaginé, organisé et pensé par une IA. Prenons le cas fictif d'un auteur qui aurait couché, de sa main, un long texte, entièrement né de son imagination. S'il demande à une IA d'uniformiser son style, de réécrire tout ce qui doit l'être pour gagner en clarté et en lisibilité, la machine va reprendre chaque phrase et effectuer les ajustements dictés par son «modèle de langage» de référence. Cette édition du texte va amener à la réécriture de certains mots et de passages plus ou moins longs. Tout ne sera pas intégralement modifié, mais l'ensemble va être en quelque sorte contraint de rentrer dans le «moule» jugé correct par le LLM. Pangram, alors, sera fondé à conclure que le texte a intégralement été *«généré par IA».* A l'inverse,

chaque modification d’un texte généré par IA par une main humaine va dégrader le taux d’artificialité d’un texte.

Libération - 23/09/2026