

72 linguistes démunis face à ChatGPT



Si ChatGPT épate régulièrement ses utilisateurs par son aisance en langage naturel, qui mieux que des linguistes pour en juger ? Deux professeurs associés en linguistique des universités de Floride du Sud et de Memphis, aux Etats-Unis, ont fait l'expérience de confronter 72 experts à quatre extraits d'articles académiques pour leur demander de dire lesquels avaient été écrits par des auteurs humains et lesquels par le chatbot d'OpenAI.

Ce travail est paru dans la revue [Research Methods in Applied Linguistics](#) et le bilan est loin d'être brillant : si, parmi les linguistes sollicités, certains ont été capables d'identifier des textes générés par IA, aucun n'a réussi à juger correctement les quatre extraits qui leur étaient soumis. Leurs performances globales sont même assez médiocres : moins de 39% de

bonnes réponses pour l'ensemble des participants. Pire : 12,5% d'entre eux se sont trompés sur l'origine des quatre extraits d'articles !

Des textes générés par IA à partir de vrais articles

L'équipe avait à disposition huit extraits d'articles. Quatre ont été réellement écrits par des chercheurs et publiés en 2021 ou 2022 dans deux revues à comité de lecture de linguistique appliquée. Les quatre autres ont été générés par ChatGPT à partir des articles complets d'où proviennent les extraits cités plus hauts. Pour chaque article les différentes parties (introduction, revue de la littérature, méthode, résultats, discussion) ont été clairement identifiées pour demander explicitement à l'agent conversationnel de résumer chacune d'elles. A partir de quoi, ChatGPT a créé son texte scientifique. Aucun des huit extraits n'a ensuite été retouché par l'équipe de recherche, mais les noms des auteurs ont été retirés.

Les textes ont été soumis aux linguistes de manière aléatoire *"si ce n'est que nous avons fait en sorte qu'aucun d'entre eux ne reçoivent que des extraits générés par IA ou que des extraits écrits par des humains, pour que l'expérience garde tout son sens"*, précise J. Elliot Casal, professeur associé à Memphis et coauteur de cette étude.

Des erreurs dans les deux sens

Après quoi, il leur fallait dire, pour chacun des quatre textes, s'il était écrit par un chercheur, par une IA ou s'il n'arrivait pas à se prononcer. Dans un deuxième temps, 7 de ces 72 linguistes ont participé à un entretien où ils expliquaient sur quelles bases, quels critères, ils avaient jugé qu'un texte était plutôt humain ou plutôt artificiel.

Le bilan global est de... 38,9% de bonnes réponses. *"La plupart des répondants a eu deux mauvaises réponses ou plus sur 4, dans les deux sens [ils ont jugé un article humain comme écrit par ChatGPT et inversement, ndlr], note J. Elliott Casal. Dans l'ensemble, ils ont été meilleurs pour identifier les textes provenant d'un humain que d'une*

IA". En l'occurrence, la différence est de 44,1% de bonnes réponses contre 33,7%. Et si aucun des chercheurs n'a su donner quatre bonnes réponses, ils ne sont que 18,1% à avoir vu juste sur 3 des 4 textes et 34,7% à avoir donné deux bonnes réponses.

Le meilleur taux de bonne réponse sur un texte se monte à seulement 55,6% et concerne un extrait d'article signé d'un auteur humain. *"D'un point de vue rhétorique et linguistique, ces textes étaient très stéréotypés et brefs, mais nous pensions tout de même qu'il aurait été plus facile que cela de repérer ceux générés par IA"*, reconnaît le chercheur américain.

Des critères de jugement inopérants

Tout aussi préoccupants sont les critères utilisés par les répondants. Ceux-ci sont tout à fait logiques et appropriés à la tâche demandée, comme de juger la fluidité globale de l'extrait, le fait qu'il donne des informations précises ou vagues, le niveau d'expression... Or, malgré ces outils, les performances ont été peu brillantes. En fait, cet aspect de la recherche montre que les répondants estiment a priori que l'IA produit des textes plus vagues, moins fluides et moins faciles à lire qu'un humain. Autant de critères qui se sont avérés inopérants pour débusquer ChatGPT...