# ChatGPT : comment l'IA parvient à s'exprimer

Si ChatGPT (acronyme de generative pretraining transformer) donne de façon extraordinaire l'impression de savoir s'exprimer, il ne fait qu'aligner des suites de mots qui sont statistiquement les plus probables, sans intégrer la moindre dimension de leur sens. Son algorithme a été entraîné sur une base de données de 8 millions de pages Web constituée par OpenAI, WebText (articles, sites Web, discussions de forums…). OpenAI s'est également servi de romans et d'autres écrits (non divulgués) comme données d'entraînement. Cet ensemble permet à ChatGPT de maîtriser de nombreux styles : rimes, CV, fiche pratique, description, dialogue… Pour limiter la production de contenus erronés lors des interactions, ChatGPT a suivi un apprentissage par renforcement où une intervention humaine indique à une IA si elle se trompe. L'outil est également une variante d'un modèle appelé Transformeur qui lui permet de contextualiser chaque mot d'une phrase avec les autres mots, en donnant un score d'importance à chaque paire de mots. C'est la grande force de ChatGPT : il peut remonter le fil des échanges avec l'utilisateur et s'en servir pour générer ses prochaines interactions. Le tout est combiné à une fonction générative par laquelle le chatbot développe cette faculté d'expression qui captive tout le monde. Il répond, confirme, infirme, argumente, rebondit. Même s'il se trompe ou dit n'importe quoi. C'est l'un des problèmes : la puissance de l'outil réside moins dans ce qu'il dit que dans la manière dont il le dit. Il n'est pas conçu pour vérifier ses propos, ni pour éviter le plagiat (il peut recycler des bouts de textes existants). Une couche de filtrage peut lui éviter d'être ouvertement raciste ou insultant, mais dans ce cas, ce ne sera pas le fait du réseau lui-même, mais d'une intervention humaine pour pondérer a priori ses propos.


| | |
|---|---|
| **Source** | Sciences&Avenir – La Recherche n° 914 |
| **Auteur** | Arnaud Devillard |
| **Date** | 30/03/2023 |