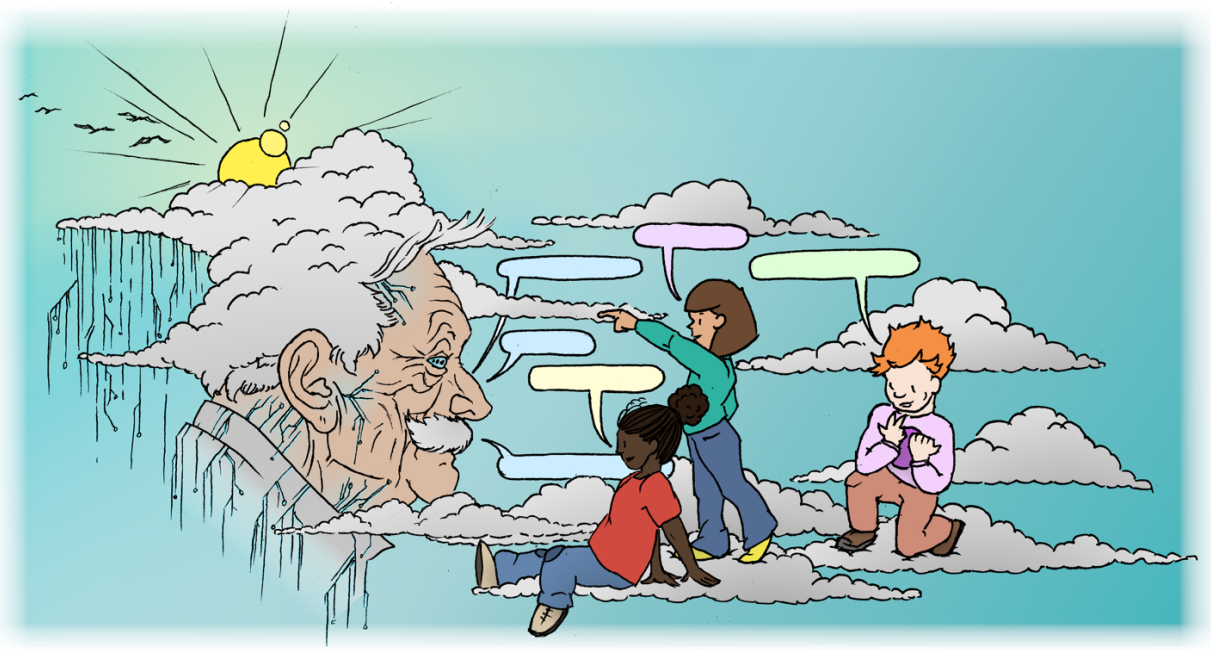

# PEUT-ON FAIRE CONFIANCE À UN TEXTE GÉNÉRÉ PAR INTELLIGENCE ARTIFICIELLE ?

*Charles Ollion[1] et Anna Bonnet[2]*

[1]*Association Surfrider Foundation Europe, Biarritz, France*
[2]*Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, Paris, France*

ChatGPT et d'autres applications d'intelligence artificielle (IA) semblent tout comprendre et nous donner des réponses convaincantes, même pour des questions d'examen difficiles. Mais peut-on toujours se fier à ces réponses ? Dans cet article, nous allons examiner de plus près la manière dont ces systèmes d'IA fonctionnent avec du texte : comment ils sont entraînés, ce qui les rend intelligents et comment nous pouvons discuter avec eux. Nous verrons qu'ils produisent des réponses qui semblent très probables, mais qu'ils ne se soucient pas de nous dire la vérité ou de nous fournir des informations factuelles.

## INTRODUCTION

Lorsque tu utilises un moteur de recherche sur Internet (Google ou Bing par exemple), les résultats que tu obtiens sont des pages web ou des réponses qui ont été créées au préalable, par exemple une page Wikipédia, généralement écrite par des humains. Lorsque tu poses une question à des plateformes comme ChatGPT, le processus est très différent : le résultat est généré en direct par un moteur d'intelligence artificielle (💻). Essaie toi-même : si tu poses la même question plusieurs fois (✍), il est possible que les réponses soient différentes. Voici un exemple réel de deux réponses de ChatGPT à ma question :

✍ Qu'est-ce que la science ? Donne une réponse courte et amusante.

💻 La science, c'est essentiellement poser beaucoup de questions à la nature et essayer de ne pas l'interrompre lorsqu'elle donne des réponses vraiment longues et compliquées !

INTELLIGENCE ARTIFICIELLE. Capacité d'un programme informatique à raisonner et à apprendre. Aujourd'hui, nous appelons ChatGPT "intelligence artificielle", même si les experts et les scientifiques ne le considèrent pas comme véritablement "intelligent".

La science, c'est comme essayer de résoudre un puzzle cosmique géant, mais les pièces sont parfois invisibles et le mode d'emploi est écrit sous la forme d'énigmes !

ChatGPT est une plateforme web lancée fin 2022 par une société appelée OpenAI. Elle a attiré beaucoup d'attention dans les médias ainsi qu'un grand intérêt dans le monde scientifique. Depuis, de nombreuses autres applications ont été développées (Bard, Ernie, LLaMa), en utilisant les mêmes principes fondamentaux. À propos, GPT signifie Generative Pre-Trained Transformer (transformeur génératif pré-entraîné). Qu'est-ce que cela signifie ? Nous allons nous plonger dans l'explication de chacun de ces mots pour comprendre les briques qui constituent les applications comme ChatGPT. Ensuite, nous disposerons des outils nécessaires pour répondre à la question qui nous intéresse : pouvons-nous faire confiance aux moteurs d'IA ?

## LE MOTEUR GÉNÉRATIF : LE MODÈLE DE LANGAGE

Le moteur d'IA utilise ce que nous appelons un modèle de langage [1] pour comprendre les données saisies par l'utilisateur et générer un texte en sortie. Le fonctionnement est le suivant : le moteur crée une phrase mot par mot, en prédisant quel nouveau mot s'accordera le mieux avec les précédents. Considérons un exemple de génération d'une phrase (Figure 1). La phrase d'entrée est divisée en mots (à gauche), et le rôle du moteur est de prédire quel mot, parmi un vocabulaire prédéfini de tous les mots possibles (à droite), s'adapterait le mieux. Pour ce faire, il attribue un score à chaque mot possible : la plupart des mots auront un score de 0, ce qui signifie qu'il est peu probable qu'ils soient choisis, et quelques-uns obtiendront un meilleur score (représenté par des traits bleus).

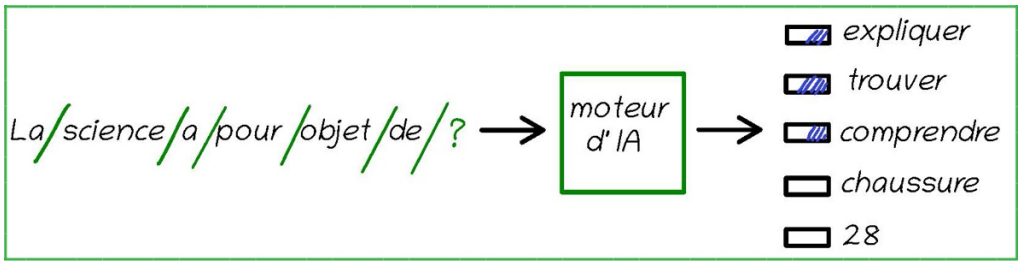


Figure 1. Le moteur d'IA cherche à prédire le mot suivant.

Dans l'exemple ci-dessus, le mot " trouver" aura le meilleur score, suivi par " comprendre" et "expliquer" tandis qu'un mot comme "chaussure" aura un score de 0. Ensuite, le moteur choisit un mot au hasard parmi ceux qui ont obtenu les meilleurs scores : dans ce cas, le mot "trouver" a le plus de chances d'être choisi, mais d'autres mots ayant un score différent de 0 peuvent également l'être. Ce mot est ajouté à la phrase, et nous pouvons répéter tout le processus pour trouver le suivant, jusqu'à ce qu'une réponse complète soit produite !
Lorsqu'une question est posée à ChatGPT, ou lorsque nous discutons pendant des heures avec lui, le moteur d'IA suit toujours le principe du

modèle de langage : essayer de prédire le mot suivant. Certains mots ont été écrits par toi, d'autres ont été générés auparavant par le moteur, ce qui fournit à un instant donné tout un contexte que ChatGPT va utiliser pour continuer à générer les mots les plus probables. Il obtient ainsi les phrases ou les réponses les plus probables, compte tenu de l'ensemble du contexte de la conversation.

## COMMENT (PRE-)ENTRAÎNER LE MOTEUR DE RECHERCHE ?

Pour que le moteur puisse prédire les bons mots suivants, il doit avoir été entraîné au préalable. Ce "pré-entraînement" a été réalisé par des ingénieurs et des chercheurs pendant plusieurs mois. Le moteur est ensuite figé et utilisé comme décrit ci-dessus. Cela signifie que ChatGPT et les autres systèmes d'IA n'apprennent plus rien lorsque tu leur parles ! Mais que signifie exactement entraîner un moteur d'IA ? Au début, avant l'entraînement, le moteur n'est pas très intelligent. À cette étape, il produit généralement les mêmes scores faibles pour chaque mot.

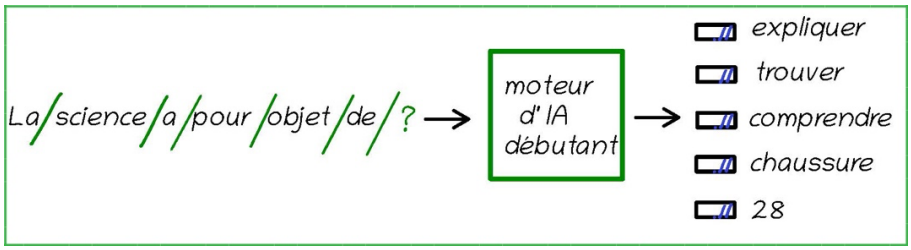


Figure 2. Un moteur d'IA débutant (non entraîné) n'est pas capable de prédire le mot suivant correctement.

Si tu poses une question à un moteur d'IA débutant (Figure 2), il produira une séquence de mots au hasard qui n'aura aucun sens.

ALGORITHME. Ensemble d'instructions à suivre pour accomplir une tâche. Par exemple, une recette peut être un algorithme pour faire un gâteau. Les algorithmes informatiques sont créés par programmation à l'aide d'un code.

Des ingénieurs et des chercheurs ont créé des algorithmes d'entraînement pour former les moteurs en leur donnant de nombreux textes existants [2]. L'algorithme est un programme informatique qui se présente comme suit :

1. choisir une phrase à montrer au moteur non entraîné, par exemple la phrase suivante : "La science est une façon de trouver des explications pour comprendre le fonctionnement du monde et de ses phénomènes."
2. demander au moteur de donner un score aux mots qui suivent une partie de la phrase "La science est une façon de".
3. Plus le score du mot juste - ici "trouver" - est faible, plus l'erreur du moteur est grande. Au début, lorsque le moteur n'est pas entraîné, cette erreur est importante. L'algorithme utilise cette erreur pour modifier légèrement le moteur, de sorte que la prochaine fois que nous demanderons des scores avec ce type d'entrée, il donnera un meilleur score à "trouver", et fera donc moins d'erreurs

Lorsque nous répétons ce processus de nombreuses fois, avec des milliards de phrases différentes, le moteur apprend et commence

progressivement à faire de moins en moins d'erreurs et donc à produire des scores plus élevés pour les mots probables. Il faut de nombreux ordinateurs rapides et des semaines d'entraînement pour entraîner les gros moteurs qui alimentent ChatGPT.

En fonction des textes que nous avons choisi de montrer au moteur pendant l'entraînement, celui-ci attribuera des scores différents aux mots lorsqu'ils seront utilisés par la suite : par exemple, si nous sélectionnons principalement des textes d'entraînement en français, il attribuera des notes très faibles aux mots anglais, mais des notes élevées à certains des mots français.

En réalité, le processus d'entraînement est un peu plus complexe et comporte plusieurs étapes.

## POURQUOI LE TRANSFORMEUR EST-IL SI PERFORMANT ?

Pourquoi ChatGPT est-il si performant ? C'est un mystère, et les scientifiques essaient toujours de l'expliquer, mais il y a plusieurs éléments dans la conception de ces moteurs d'IA qui sont très importants, examinons-en quelques-uns.

RÉSEAU DE NEURONES ARTIFICIELS. Type de programme informatique qui s'inspire vaguement du fonctionnement du cerveau des animaux. Il est utilisé pour résoudre des problèmes complexes.

L'une des armes secrètes se trouve dans la manière dont le moteur est conçu : il utilise une architecture de réseau de neurones artificiels appelée transformeur [3]. Un transformeur est comme un ordinateur efficace qui reçoit la phrase d'entrée afin de produire les scores. Il possède une caractéristique spéciale qui lui permet de se concentrer sur certaines parties de la phrase qui sont les plus importantes pour produire les meilleurs résultats possibles. Par exemple, afin de produire un bon score pour "trouver" dans l'exemple ci-dessus, le transformeur se concentrera probablement sur le mot "de", qui suggère que le mot suivant est sûrement un verbe, ainsi que sur le mot "science" qui indique le sens global de la phrase. C'est comme une loupe de détective qui aide le moteur à lire entre les lignes et à identifier quels éléments sont importants. C'est très utile, en particulier lorsqu'on donne au moteur un très long texte en entrée !

Mais l'une des principales raisons pour lesquelles il fonctionne si bien est l'échelle des données d'entraînement : il est difficile d'imaginer le nombre de phrases sur lesquelles ces moteurs sont entraînés. GPT-3 (la troisième version du modèle de langage qui alimente ChatGPT) a été entraîné sur environ 500 milliards de mots, ce qui équivaut à environ 5 millions de livres différents. Une autre raison importante est la diversité des textes sur lesquels ChatGPT a été entraîné. Non seulement les textes sont rédigés dans de nombreuses langues différentes, mais ils comprennent également des langages de programmation (par exemple Python), des instructions et des recettes, etc.

La façon dont le vocabulaire est choisi est également très importante. Rappelle-toi que le travail du moteur d'IA est d'attribuer un score à

chacun des mots suivants possibles. ChatGPT semble comprendre tous les mots français possibles, ainsi que les mots mal orthographiés comme "salu", ou tout mot d'une autre langue, comme l'alphabet japonais ひらがな, ce qui pourrait faire des milliards de possibilités ! Parmi son vocabulaire, il y a également les signes de ponctuation, qui contribuent à la structure cohérente du texte, ou encore des retours à la ligne, qui lui permettent par exemple d'écrire un poème en vers ! Pour que le vocabulaire reste relativement restreint (environ 50000 mots possibles), l'astuce consiste à utiliser non seulement des mots, mais aussi des parties de mots, appelées des tokens, et même des lettres ou des caractères, de sorte qu'en les combinant, il puisse comprendre et produire n'importe quel mot (Figure 3).

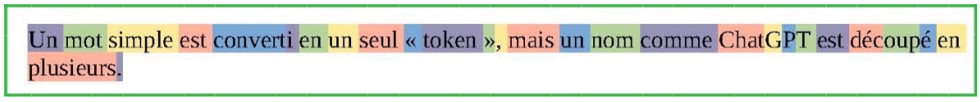


Figure 3. Cet exemple montre comment ChatGPT décompose un texte en tokens. Chaque changement de couleur indique l'endroit où ChatGPT sépare le texte. Tu peux jouer à découper un texte en tokens à l'adresse suivante https://platform.openai.com/tokenizer

On peut notamment remarquer que le mot "converti" est séparé en deux tokens "convert" et "i". Le premier contient la racine du mot, commune avec des mots de la même famille (convertir, convertisseur…) alors que la terminaison définit la nature du mot, ici un participe passé. Ceci explique pourquoi ChatGPT est capable de fournir des explications amusantes pour les mots inventés :

> Donne une définition plausible pour ce mot inventé : superhéropathie
>
> Le terme "superhéropathie" pourrait désigner une maladie imaginaire dont le principal symptôme est de faire semblant d'avoir des superpouvoirs, comme les héros de bandes dessinées. Les malades agissent comme s'ils pouvaient voler, avoir une force surhumaine ou faire des choses magiques, même s'ils ne peuvent pas vraiment le faire. Ils peuvent porter des costumes et faire des imitations en s'imaginant qu'ils sont des super-héros. Tout cela pour s'amuser et faire preuve d'imagination !

## PEUT-ON FAIRE CONFIANCE AUX MOTEURS D'IA ?

Comme les moteurs d'IA sont capables de comprendre n'importe quelle donnée et que leurs réponses sont impressionnantes, nous avons tendance à les croire "intelligents" et à leur faire confiance. Mais leur faire confiance est une grave erreur, car ils produisent des réponses qui sont simplement plausibles, mais peuvent souvent être fausses. Prenons l'exemple de calculs :

> 23 * 7 - 109
>
> Le résultat de l'expression 23 * 7 - 109 est 2.

ChatGPT fournit avec assurance une réponse fausse (la bonne réponse est 52). Plus important encore, lorsqu'on lui pose à nouveau la question, il donne une réponse différente.

> 23 * 7 - 109
>
> Le résultat de l'expression 23 * 7 - 109 est 42.

Pour comprendre ce qu'il se passe, il est essentiel de se rappeler comment ChatGPT fonctionne : le moteur d'IA n'a pas suivi les règles de l'arithmétique pour répondre. Tout ce qu'il a fait, c'est identifier qu'un petit nombre était une réponse probable à proposer, plutôt qu'un nombre plus élevé ou un token qui n'est pas un nombre.

Sachant cela, il peut quand même être utile d'interagir avec un moteur d'IA tout en connaissant ses faiblesses. En comprenant son fonctionnement, il est même possible de le rendre plus utile. Par exemple, au lieu de demander directement une réponse numérique, il est préférable de procéder comme suit :

> Combien vaut 23 * 7 – 109 ? Peux-tu décrire les différentes étapes de calcul avant de donner la réponse ?
>
> Bien sûr, je serais ravi de t'aider à faire les différentes étapes de ce calcul :
> Étape 1: Multiplie 23 par 7 : 23 * 7 = 161
> Étape 2: Soustrais 109 au résultat de la multiplication : 161 - 109 = 52
> La réponse est donc 52.

Nous avons obligé le moteur à effectuer un raisonnement intermédiaire : réfléchir étape par étape pour résoudre le problème ! Remarque que cela n'empêche pas le moteur de faire des erreurs, mais cela l'aide souvent à fournir une meilleure réponse. Cette façon de penser pour les moteurs d'IA est appelée une chaîne de pensée [4], et elle est également utilisée pour fournir des exemples intéressants pendant l'entraînement. Dans la dernière version de ChatGPT, il aura tendance à détailler par défaut son calcul, justement pour éviter de faire des erreurs. En revanche, il est toujours possible de le "piéger" en lui demandant de répondre par un nombre uniquement, ce qui le pousse souvent à se tromper…

## CONCLUSION

Les moteurs d'IA tels que ChatGPT fournissent des résultats impressionnants, mais ils ne sont pas entraînés pour dire la vérité, et il ne faut donc jamais leur faire confiance aveuglément. Ils peuvent néanmoins être utiles, car ils sont utilisés dans de plus en plus d'applications et continueront à l'être, par exemple pour créer des assistants personnels ou améliorer les recherches sur le web. Toutefois, leur utilisation soulève de nombreuses questions éthiques : par exemple, les moteurs d'IA peuvent significativement modifier nos façons de rédiger un texte, sans donner de crédit aux auteurs des données sur lesquelles ils ont été entraînés. Par ailleurs, si nous commençons à trop utiliser ces moteurs, il

pourrait être difficile de savoir quels textes ou conversations sont rédigés par des humains et lesquels sont artificiels. Un autre aspect à prendre en compte est que ChatGPT peut écrire des histoires, des essais et d'autres choses pour nous. Mais si nous comptons sur lui pour accomplir toutes nos tâches, nous risquons de ne pas apprendre, de ne pas être créatifs, de ne pas nous développer et de ne pas penser par nous-mêmes. Et surtout, pense à remettre systématiquement en question ce que ChatGPT te dit ! Même les scientifiques ne savent pas exactement pourquoi il marche aussi bien, mais il est toujours bon de connaître sa façon de fonctionner, puisque cela nous permet de l'utiliser tout en comprenant ses forces et ses faiblesses !

## RÉFÉRENCES


[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Kukasz Kaiser, Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (NIPS 2017).

[2] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams. Learning representations by back-propagating errors. Nature 323.6088 (1986): 533-536.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent. "A neural probabilistic language model. Advances in neural information processing systems 13 (NIPS 2000).

[4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, DennyZhou. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022): 24824-24837.


## ARTICLE ORIGINAL (VERSION FRANÇAISE)


**SOUMIS** le 1er mars 2024 ; **ACCEPTÉ** le novembre 2024
**PUBLICATION :** janvier 2025

**ÉDITION** : Catherine Braun-Breton & Ula Hibner, Association Jeunes Francophones et la Science

**MENTORS SCIENTIFIQUES** : Catherine Braun-Breton & Ula Hibner

**CITATION** : Ollion C., Bonnet A. (2025). Peut-on faire confiance à un texte généré par intelligence artificielle ? Jeunes Francophones et la Science (https://www.jeunesfrancophonesetlascience.fr/).

**DÉCLARATION DE CONFLIT D'INTÉRÊT :** Les auteurs déclarent que les travaux de recherche ont été menés en l'absence de toute relation commerciale ou financière pouvant être interprétée comme un conflit d'intérêt potentiel.

## JEUNES EXAMINATEURS

### LISE, 12 ANS

Lise est une grande admiratrice d'Hercule Poirot, bien qu'elle préfère éviter les découvertes macabres en dehors des romans. Elle voit dans la science un terrain de jeu idéal pour s’attaquer à des énigmes aussi intrigantes que celles d’Agatha Christie.

### CLOTHILDE 16 ANS, SIMON 15 ANS, IASSIM 12 ANS, FAYSSAL 13 ANS

Participants de l'atelier "Ensemble, décryptons la science" de la Nuit de Chercheuses à Montpellier le 27 septembre 2024

## AUTEURS

### CHARLES OLLION

Charles fait de la recherche en intelligence artificielle (IA) et ses applications. Il est titulaire d'un doctorat en informatique et a enseigné l'IA pour la vision automatique et la compréhension du langage à l'Institut Polytechnique de Paris. Il travaille actuellement pour l'ONG Surfrider Foundation Europe où il utilise l’IA afin de détecter les déchets plastiques. S'il n'est pas derrière son ordinateur, il est probablement en train d'escalader la montagne la plus proche !

### ANNA BONNET

Anna est enseignante-chercheuse en statistique à Sorbonne Université, spécialisée en modélisation mathématique pour les sciences du vivant. Elle collabore notamment avec des écologues, des médecins et des biologistes pour développer et appliquer de nouveaux outils d'analyse de données.