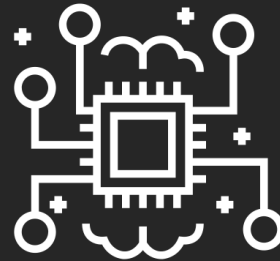


Intelligence artificielle: une nouvelle ère?

Ce que peuvent faire (ou ne pas faire)
ChatGPT, Bard, LaMDA, Galactica
et les grands modèles de langage.



«Tsunami technologique», «invention du siècle»: ChatGPT affole le monde de la tech

L ChatGPT, le robot "omniscient" qui fait trembler Google, relance les craintes sur l'intelligence artificielle

+ Faut-il craindre la révolution ChatGPT?

ChatGPT banned from New York City public schools' devices and networks

'Google killer' ChatGPT sparks AI chatbot race

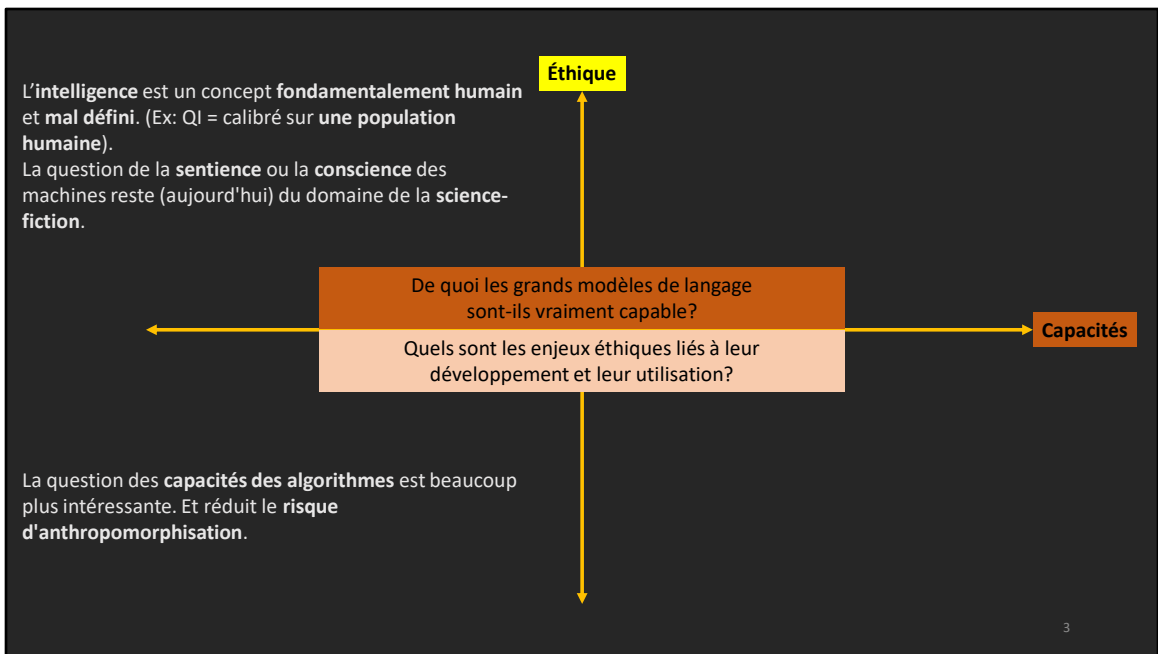
ARTIFICIËLE INTELLIGENTIE

ChatGPT maakt hoger onderwijs zenuwachtig

2

On a un tout petit peu parlé d'intelligence artificielle, ces derniers mois. Le déclencheur de toute cette frénésie médiatique est, évidemment, la sortie de ChatGPT, le "chatbot" d'OpenAI souvent présenté comme un bouleversement complet dans le monde de la technologie.

ChatGPT est une application basée sur ce qu'on appelle les "grands modèles de langage", que je vais tenter d'un peu démystifier pour vous aujourd'hui.



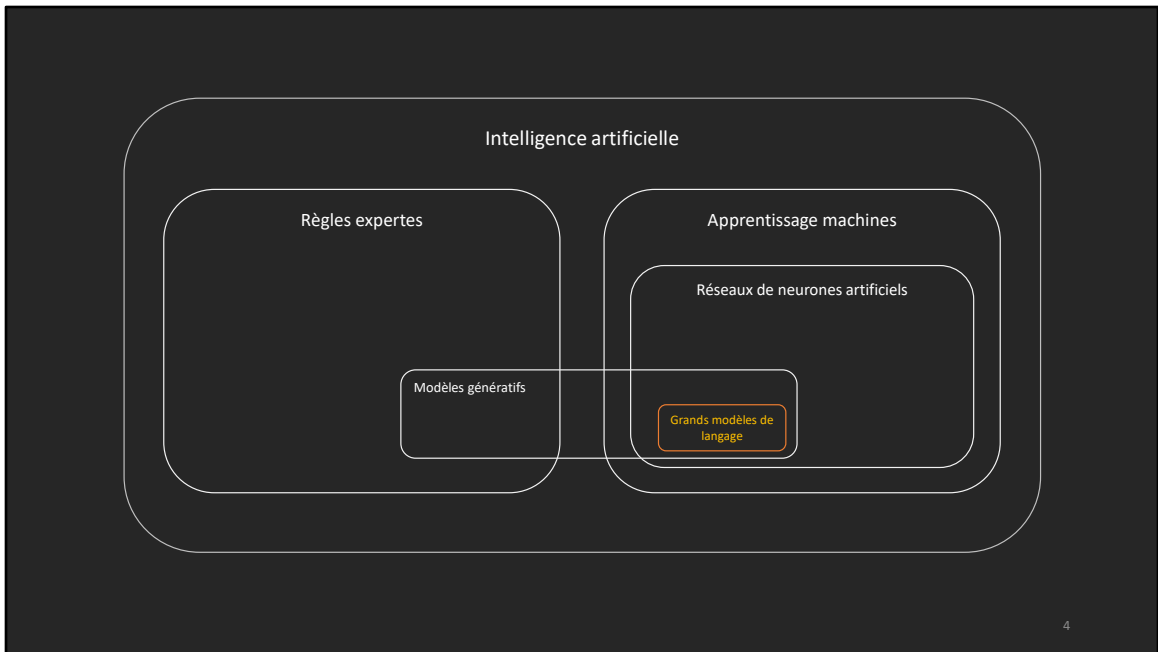
Mon objectif aujourd'hui, en vous parlant de ChatGPT et des grands modèles de langage, est de vous présenter deux axes sur lesquels on peut examiner un algorithme: d'abord, sur ses capacités – de quoi les grands modèles de langage sont-ils vraiment capable? Pour ce faire, je vais vous donner quelques bases de comment une machine "apprend" des choses, et de ce qui se passe dans la machine lorsqu'on pose une question a un outil comme ChatGPT.

Ensuite, sur le plan éthique, on a beaucoup parlé dans les médias de questions liées à la triche dans l'enseignement, mais j'aimerais aller un peu plus loin et examiner avec vous quelques enjeux éthiques liés à l'adoption massive potentielle de ces outils génératifs dans le grand public.

Une chose dont je ne vais pas beaucoup parler, c'est d'intelligence. Si je n'en parle pas beaucoup, c'est parce que le terme "intelligence artificielle" est assez mal choisi. Il date des années 50, et à l'époque on avait une bien moins bonne compréhension et de l'intelligence humaine, et des capacités des machines.

Aujourd'hui, on se rend bien compte que "l'intelligence" est un concept qui, même pour l'humain, est très mal défini. Il y a plusieurs types d'intelligence, souvent difficile

à mesurer. Le QI, par exemple, est une mesure relative, calibrée sur une certaine population. Dire, pour un humain, "j'ai 120 de QI", ne veut en soi rien dire sans savoir quel test, calibré sur quelle population, a été effectué. La notion a encore moins de sens quand on essaie de l'appliquer à des machines. Il n'y a ni sentience, ni conscience dans les "intelligences artificielles". L'idée d'une machine consciente reste aujourd'hui du domaine de la science-fiction. On restera donc focalisé sur les capacités.



Parlons-donc de ces capacités. La première chose dont il faut bien avoir conscience, c'est que les grands modèles de langage ne sont qu'une toute petite partie du domaine qu'on appelle "intelligence artificielle".

On distingue historiquement dans le domaine deux grandes approches: celles basées sur des règles expertes, explicitement écrites par des humains et appliquées par les machines, et les systèmes à apprentissage, basés sur des règles apprises à partir de set de données.

Dans les méthodes à apprentissage, on retrouve les réseaux de neurones artificiels... ainsi que pleins d'autres approches. Ces dernières années, ces réseaux de neurones sont devenus de plus en plus puissants, et une sous-catégorie a gagné en popularité: les réseaux génératifs, qui "produisent" du contenu "nouveau". On y trouve par exemple les générateurs d'image, comme DALL-E. Et dans ces réseaux génératifs, on retrouve les grands modèles de langage comme GPT et d'autres.

Si c'est un domaine très médiatisé, les grands modèles de langage restent donc une petite portion de ce qui se fait en intelligence artificielle.

L'apprentissage des machines

Entrée(s)

x

(images, textes, sons... données encodées numériquement)

Modèle

$y = ax + b$

a, b : paramètres du modèle

Sortie(s)

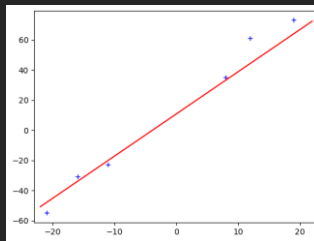
y

(décision, prédiction, action...)

Set d'apprentissage:

x	y
-11	-23
19	73
25	60
8	35
12	61
-21	-55
-16	-31

Objectif: trouver a, b
tels que $y - ax - b \rightarrow 0$
(= trouver la meilleure droite entre les points)



Méthode:

- Initialiser a, b
- Mesurer erreur $e = y - ax - b$
- Modifier a, b pour minimiser e
- Itérer jusqu'à ce que e ne diminue plus ou a, b ne changent plus.

Revenons donc d'abord à cette idée d'apprentissage des machines. Comment une machine apprend-elle?

L'idée principale derrière l'apprentissage des machines est de définir une tâche sur base de variables d'entrée – par exemple, des images, des sons, des textes... encodés sous forme numérique pour pouvoir être traités par une machine – et de variables de sortie: des décisions, prédictions, actions que la machine doit effectuer. Par exemple: sur base de photos, compter le nombre de visages; sur base d'un texte, fournir une traduction, etc.

Entre les entrées et les sorties, on va construire ce qu'on appelle un modèle: une relation mathématique entre la sortie y et les entrées x . Le modèle le plus simple qu'on peut construire est un modèle linéaire: $y = ax + b$. C'est simplement l'équation d'une droite. Dans ce modèle, on a deux paramètres: a et b . Apprendre, pour une machine, c'est simplement "trouver les meilleures valeurs pour a et b ".

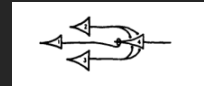
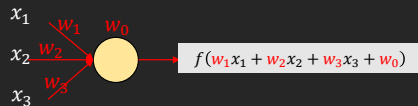
Pour ce faire, on va se baser sur des données d'apprentissage. C'est-à-dire: une série d'exemples où, pour un x donné, on connaît le y . Notre objectif d'apprentissage, maintenant, va donc être de trouver les meilleures valeurs pour a et b pour que la

droite $y = ax+b$ passe, autant que possible, par les points d'exemple.

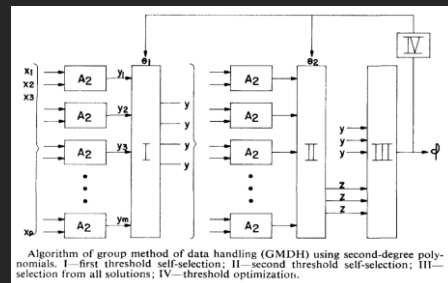
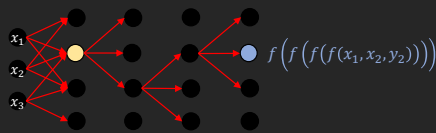
Pour apprendre, on va donc commencer par prendre des valeurs initiales pour a et b , puis mesurer l'erreur qu'on commet avec ces valeurs, et les modifier petit à petit en réduisant l'erreur jusqu'à ce qu'on soit satisfait du résultat.

L'apprentissage des machines

"Neurone artificiel" (1940s-50s: McCulloch & Pitts, Turing, Rosenblatt)



"Réseaux de neurones artificiel" (1950s-70s: Turing, Rosenblatt, von Neumann, Ivakhnenko...)



Algorithm of group method of data handling (GMDH) using second-degree polynomials. I—first threshold self-selection; II—second threshold self-selection; III—selection from all solutions; IV—threshold optimization.

6

Evidemment, le modèle linéaire est très limité dans ce qu'il peut faire. Dans la plupart des tâches qu'on veut résoudre, la relation entre les entrées et les sorties est beaucoup plus complexe. Une manière de "complexifier" la relation est de faire appel aux réseaux de neurones artificiels.

L'idée du "neurone artificiel" date des années 40 et 50. Elle part de l'observation que les neurones, individuellement, réalisent une opération mathématique assez simple. Ils ont une série "d'entrées" – les neurotransmetteurs au niveau des dendrites – qui peuvent soit "activer", soit "inhiber" le neurone. Si les "activations" sont suffisamment fortes, un potentiel d'action se déclenche et le neurone libère à son tour des neurotransmetteurs vers le neurone suivant. On peut assez facilement modéliser mathématiquement ce comportement.

Si un neurone a un comportement assez simple, la complexité vient de leur nombre et de leurs interconnexions. Dès les années 50, 60 et 70, on étudie les "réseaux de neurones artificiels". La sortie d'un tel réseau est complexe parce qu'il s'agit d'une combinaison de très nombreuses opérations simples. Si un neurone fait une fonction $f(x)$, un réseau aura une fonction $f(f(f(f...)(x)))$.

Les paramètres de ce modèle correspondent au poids des connexions entre les neurones: positifs pour des connexions qui activent, négative pour celles qui inhibent. Plus un réseau a de neurones et de connexions, plus il aura de paramètres, plus il pourra modéliser des phénomènes complexes.

L'apprentissage des machines

Réseaux de neurones artificiels



- En théorie, **peuvent résoudre n'importe quelle tâche** (*tant qu'on peut l'exprimer comme une relation mathématique*).
- Ne nécessitent pas d'avoir un "modèle physique" de ce qui se passe: il suffit d'avoir des **exemples**.



- En pratique, "entraîner" des grands réseaux avec les ressources du XX^e siècle était **impossible**.
- Plus on veut de paramètres, plus on a besoin de **données** et de **puissance**.
- Besoin (souvent) d'**annotations humaines**.

Règles expertes

- **Règles faites par des humains** (on comprend bien ce qui se passe et pourquoi les décisions sont prises).
- Nécessitent (en général) **peu de puissance** de calcul.
- **Peu de données** nécessaires
- Mauvais résultats sur les problèmes "abstrait", "intuitifs" (où on a du mal à expliquer notre raisonnement).
- Limités par la connaissance humaine.

7

Les réseaux de neurones artificiels se sont imposés depuis une grosse dizaine d'années comme la solution la plus adaptée à la majorité des tâches pour lesquelles on a besoin d'intelligence artificielle. En théorie, avec un réseau suffisamment grand, on peut modéliser à peu près n'importe quelle relation entre les entrées et sorties du système. On n'a pas non plus besoin de *comprendre* exactement quelle est cette relation pour la modéliser: le réseau va apprendre sur base d'exemples.

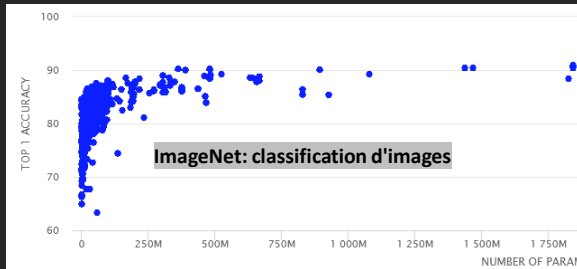
Les réseaux de neurones artificiels ne sont pas récents, mais s'ils n'ont que récemment conquis le domaine, c'est parce que jusqu'au début du XXI^{ème} siècle ils étaient en pratique impossible à entraîner. Pour atteindre la complexité nécessaire, il faut beaucoup de paramètres – des millions, maintenant parfois des milliards de paramètres – et, plus on a de paramètres, plus on a besoin d'exemples d'apprentissage.

Les systèmes basés sur des règles expertes ont certains avantages: il est plus facile de comprendre exactement comment ils arrivent à une décision – puisque les critères sont déterminés par des humains – et ils nécessitent généralement moins de données et de puissance de calcul. Mais, pour les problèmes "intuitifs", difficile à explicitement modéliser, abstraits, ils n'arrivent pas à rivaliser avec les réseaux de

neurones.

L'apprentissage des machines

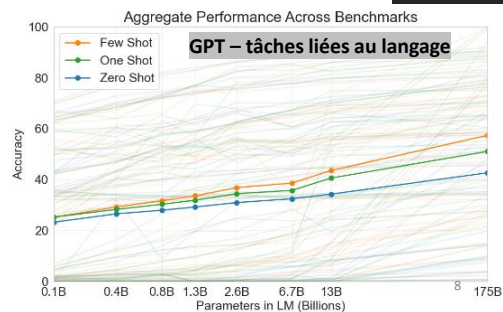
2010s-2020s: la "révolution" du Deep Learning
Big Data + GPUs + très grands réseaux = résultats+++++



<https://paperswithcode.com/sota/image-classification-on-imagenet>

Augmenter le nombre de paramètres améliore les résultats... jusqu'à un certain point.

Brown et al., Language Models are Few-Shot Learners.
ArXiv:2005.14165. (OpenAI)



Au début du XXIème siècle, on a vu une explosion des capacités de calcul des ordinateurs, et des capacités de transmission et de stockage de données, conduisant à la "révolution" Big Data, puis à la "révolution" du "Deep Learning", les larges réseaux de neurones artificiels, surtout dans le domaine de l'analyse d'image et de l'analyse de langage.

Au début, on a vu des améliorations assez fulgurantes des performances lorsqu'on augmentait la complexité des modèles. Ici, dans les résultats d'ImageNet, un concours de classification d'image, on voit que jusqu'à environ 100 millions de paramètres on a une augmentation très rapide des performances des modèles. Mais on arrive à un certain phénomène de saturation: à partir d'un certain point, augmenter la taille du modèle n'a plus qu'un effet marginal sur les performances.

Le même phénomène est visible dans les modèles de langage. Pour GPT, le modèle derrière ChatGPT, on voit ainsi que l'augmentation de performance en passant de 0,1 à 1,3 milliards de paramètres est comparable à celle qu'on obtient en passant de 13 milliards à 175 milliards.

C'est quelque chose d'assez courant dans le domaine de l'apprentissage des

machines: augmenter la complexité des modèles n'est utile que jusqu'à un certain point. La stagnation des résultats peut être due à un manque de données d'entraînement, en terme de quantité... ou de qualité!

Les grands modèles de langage

GPT3 : données utilisées

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed training for 300B
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Tâche optimisée = "prédiction du prochain mot"

- Le réseau va "apprendre" à **encoder** les mots pour que les mots **fréquemment employés ensemble** aient des **codes proches**.
- Lors de ses prédictions, le réseau va chercher quel est le mot **statistiquement plus probable** qui complète les mots en entrée.
- Possibilité de contrôler le **niveau de randomisation** (on prend *toujours le plus probable* vs on tire au sort *parmi les plus probables*).

Être ou ne pas être, telle est la question.

Être ou ne pas être, voilà la question de l'existence ou de prendre fin à se

question = 99.92%

\n = 0.04%

= 0.01%

Question = 0.01%

= 0.00%

Total: -0.00 logprob on 1 tokens (99.99% probability covered in top 5 logits)

Et là, on peut se heurter à une barrière assez solide. GPT3, par exemple, utilise comme base de donnée d'entraînement une série de sets de données extraites d'internet, de livres et de Wikipedia. Pour ChatGPT, on ne sait pas exactement ce qu'il y a dedans en plus, mais on arrive de toute façon tout doucement au bout de ce qui est disponible. Il n'y a sans doute pas assez de choses écrites par des humains pour entraîner des réseaux beaucoup plus grands que GPT3.5.

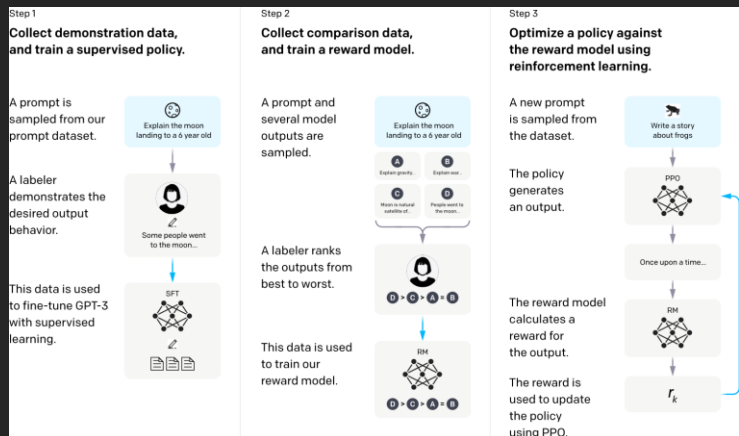
On l'a dit: les réseaux de neurones sont entraînés pour une *tâche* bien précise, la relation entre les entrées et les sorties. Pour les grands modèles de langage comme GPT, cette tâche est assez simple: prédire le prochain mot. L'entrée est donc une séquence de mots, et la sortie est le mot suivant (ou, pour être plus précis, le *token* suivant, qui correspond généralement à un morceau de mot).

Ainsi, si on entre dans GPT3 "Être ou ne pas être", il y a à peu près 100% de chances que le modèle termine par "telle est la question". Les mots qui suivent, par contre, deviennent moins certains. On peut contrôler le niveau d'aléatoire du modèle, pour soit prendre toujours le mot statistiquement le plus probable, soit tirer au sort entre les mots les plus probables.

Les grands modèles de langage

Les "+" de InstructGPT & ChatGPT

- "Renforcement humain": meilleures réponses récompensées, mauvaises réponses pénalisées pour raffiner le modèle.
- Exemples de "comportement désirés" fournis (permet de "comprendre" les demandes courantes comme "résumer", "raconter une histoire", "traduire"...)
- Couche de "sécurité" = pénaliser les sorties "problématiques" (racistes, sexistes... mais aussi dangereuses, comme "explique comment faire une bombe artisanale")



10

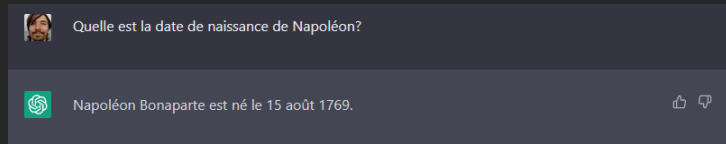
ChatGPT a reçu un traitement supplémentaire pour obtenir le ressenti "conversationnel" qui a fait tout son succès. Ce traitement est un "renforcement humain", qui consiste globalement à raffiner superficiellement les paramètres du réseau pour le pousser à donner des réponses jugées satisfaisantes par les utilisateurs.

On sait qu'OpenAI a utilisé différents critères pour ce renforcement: demander à des utilisateurs de fournir des exemples de réponse, ou encore classer plusieurs réponses générées aléatoirement.

On sait également qu'ils ont fait des renforcements de "sécurité", pour pénaliser les sorties jugées problématiques par OpenAI: cela inclut le contenu raciste, sexiste, violent, ou sexuel.

Mais il est important de noter que, si ce contenu était dans les données d'entraînement – ce qui est certainement le cas – il est toujours possible de le faire ressortir. Il y a beaucoup de gens qui s'amuse à trouver des stratégies diverses et variées pour contourner les barrières mises en place par OpenAI... et ce n'est généralement pas très compliqué.

Comment ChatGPT répond à une demande



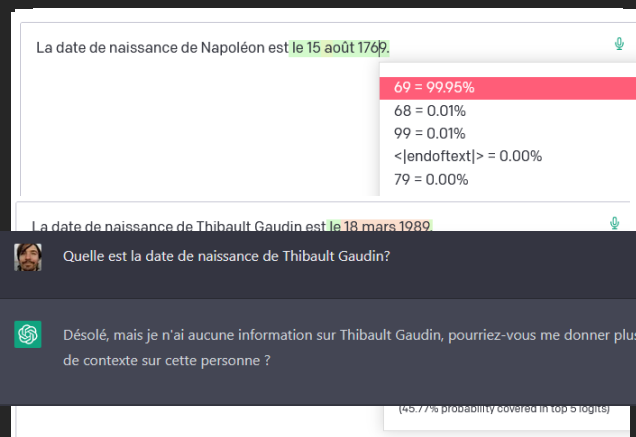
11

Voyons un peu en pratique avec quelques exemples ce que ça donne.

Pour commencer, j'ai demandé ici à ChatGPT la date de naissance de Napoléon. Sachant tout ce que je vous ai maintenant expliqué sur comment ChatGPT a été entraîné: à votre avis, est-ce que cette réponse est juste?

Oui, c'est correct. On a donc bien une réponse factuelle correcte donnée par ChatGPT sur une question précise. Mais comment a-t-il constitué cette réponse?

Comment ChatGPT répond à une demande



ChatGPT ne permet pas de regarder ce qui se passe "sous le capot", mais on peut se faire une idée via un autre outil d'OpenAI, le "Playground". Là, on peut utiliser le générateur GPT3.5, le même que celui de ChatGPT, et demander à voir les probabilités de prédiction des tokens.

Ainsi, on voit que si on commence la phrase: la date de naissance de Napoléon est, GPT estime qu'il y a 96% de chances que le mot suivant soit "le". Intéressant à noter: il y a environ 5 chances pour milles qu'il produise "La date de naissance de Napoléon est inconnue".

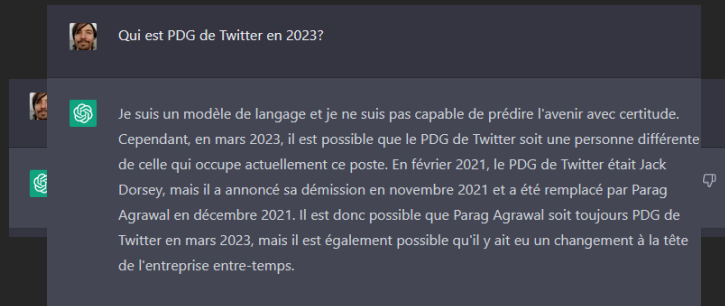
Quand on avance dans les prédictions, on voit qu'il y a peu d'hésitations. La plupart des alternatives sont sur la forme: faire un retour à la ligne, mettre la date dans un format différent, avec des majuscules... La date de naissance de Napoléon apparaît suffisamment souvent dans le set d'apprentissage pour que, statistiquement, GPT puisse très facilement prédire "le 15 août 1769".

Par contraste, si on demande la date de naissance d'un personnage moins documenté, comme Thibault Gaudin, on voit que GPT tire globalement au sort: il nous dit ici le 18 mars 1989, mais on voit qu'il aurait tout aussi bien pu prédire juin ou

juillet. Thibault, est-ce que c'est juste?

Non, et d'ailleurs si on lui redemande, il changera d'avis sans aucun soucis. Notons d'ailleurs que ChatGPT ne se laisse pas si facilement avoir: on se heurte ici a une des barrières de sécurité d'OpenAI. Si on demande une date de naissance et que les probabilités ne sont pas assez grande, on déclenche une réponse plus prudente...

Comment ChatGPT répond à une demande



13

Malheureusement, cette barrière de sécurité est loin d'être fiable. Un exemple typique est lorsqu'on demande des informations sur quelque chose de changeant. Par exemple: qui est le PDG de Twitter?

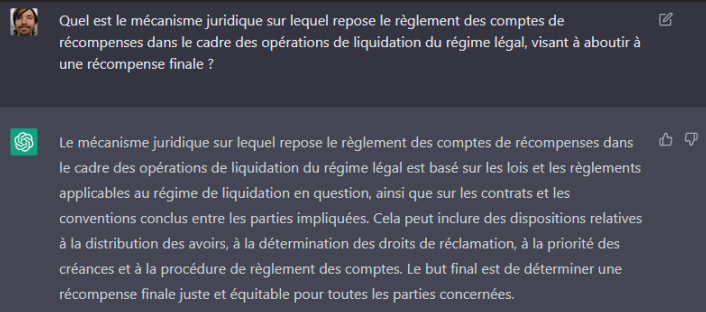
Les données d'entraînement de ChatGPT s'arrêtent fin 2021. À l'époque, le PDG de Twitter était bien Parag Agrawal. C'est aujourd'hui Elon Musk. Si on lui redemande, il nous dira peut-être aussi Jack Dorsey, le prédécesseur de Parag Agrawal.

Dans le Playground, on peut voir que Jack Dorsey est la réponse la plus probable, très largement. Et la prédiction de GPT est suffisamment "sûre" pour ne pas déclencher la barrière de sécurité. Pour la déclencher, il faut spécifiquement demander une information récente, car il a été visiblement "renforcé" pour éviter de répondre à des questions concernant 2022 ou 2023.

Même ici, cependant, on voit que le modèle a du mal à rester cohérent: il parle de "prédire l'avenir" comme s'il ne se rendait pas compte de la date, mais il devine tout de même tout seul que l'on est en mars. OpenAI étant très peu "open" sur son fonctionnement, on ne sait pas exactement comment fonctionne ChatGPT, mais il est probable qu'ils ajoutent de manière invisible des informations supplémentaires aux

questions qu'on entre. Comme, sans doute, la date. Le modèle de langage sous-jacent, cependant, ne sait pas du tout quelle est la date.

Comment ChatGPT répond à une demande



1. Quel est le mécanisme juridique sur lequel repose le règlement des comptes de récompenses dans le cadre des opérations de liquidation du régime légal, visant à aboutir à une récompense finale ?

R : la compensation

14

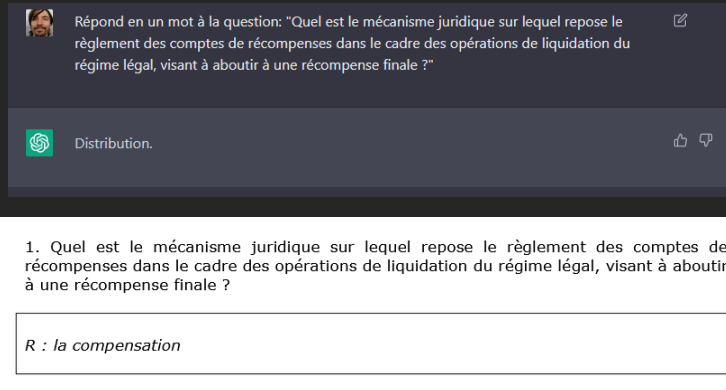
Essayons maintenant un type de question qui pourrait vous intéresser: est-ce que ChatGPT est capable de répondre à une question d'examen de Droit de BA1?

J'ai récupéré une question d'un examen du cours d'introduction au droit civil disponible sur Internet.

Moi, je n'y connais rien, donc vous serez meilleurs juges que moi: à votre avis, est-ce que c'est juste?

En tout cas, quand on regarde le corrigé, on ne retrouve pas vraiment la même chose...

Comment ChatGPT répond à une demande



15

Mais on pourrait dire que j'ai été injuste, ici: dans l'examen, on voit que la réponse doit être courte. Demandons donc à ChatGPT de répondre en un mot...

Bon, je ne pense pas que ce soit très convaincant.

Il y a eu plusieurs articles qui sont sortis ces derniers mois montrant que ChatGPT était capable de passer des examens de médecine ou de droit. C'est important de se rendre compte que ça n'indique pas que ChatGPT a une quelconque *compréhension* des concepts sous-jacents.

Simplement, il est probable que pour ces examens, généralement américains et très connus, il y avait suffisamment d'exemples sur Internet de questions pour que statistiquement une réponse suffisamment bonne apparaisse.

Il faut bien se rendre compte que les examens universitaires sont conçus pour des étudiant.e.s humain.e.s, pas pour des machines. Ce qui signifie qu'on essaye généralement dans un examen de contrôler les choses qui habituellement posent problèmes à des personnes.

Personne ne dirait qu'une calculatrice est "capable de passer un examen d'arithmétique", ou qu'un dictionnaire est "capable de faire une dictée". Ca n'a de la même manière pas de sens de dire que ChatGPT peut "passer un examen de droit". Ce qu'il ne va de toute façon pas savoir faire si l'examen ne suit pas le format le plus couramment utilisé dans son corpus d'apprentissage.

Mensonge et Baratin

Mensonge

Un **mensonge** est une affirmation **fausse** exprimée par quelqu'un qui **sait** qu'elle est fausse, dans l'**intention** de tromper son interlocuteur-ricé.

Baratin

Le **baratin** est une affirmation qui peut être **vraie ou fausse**, exprimée par quelqu'un qui **ne sait pas** si elle est vraie ou fausse (ou ne s'en préoccupe pas), dans l'**intention** d'avoir un certain effet sur son interlocuteur-ricé (par exemple: l'impressionner, lui plaire...)

Les grands modèles de langage ne sont pas des menteurs, mais bien des **baratineurs**. Ils ne sont pas conçus pour se préoccuper de la *vérité*. L'objectif pour lequel ils sont entraînés est de fournir une réponse *plausible* et *acceptable* par leur interlocuteur-ricé.

ON BULLSHIT

Harry G. Frankfurt

On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big? 🦜

16

Parfois, ChatGPT nous donne une information correcte. Parfois, il donne une information incorrecte. Peut-on dire dès lors qu'il ment?

Pas vraiment: *mentir* implique une certaine intention, et parler de mensonge, c'est à nouveau anthropomorphiser un logiciel qui n'a juste pas de notion de vérité ou de mensonge.

Une distinction intéressante en philosophie est celle entre *mensonge* et *baratin*. *Mentir*, c'est dire quelque chose de faux en *sachant* que c'est faux, et dans l'*intention* de tromper – c'est-à-dire de "faire croire" la fausse information.

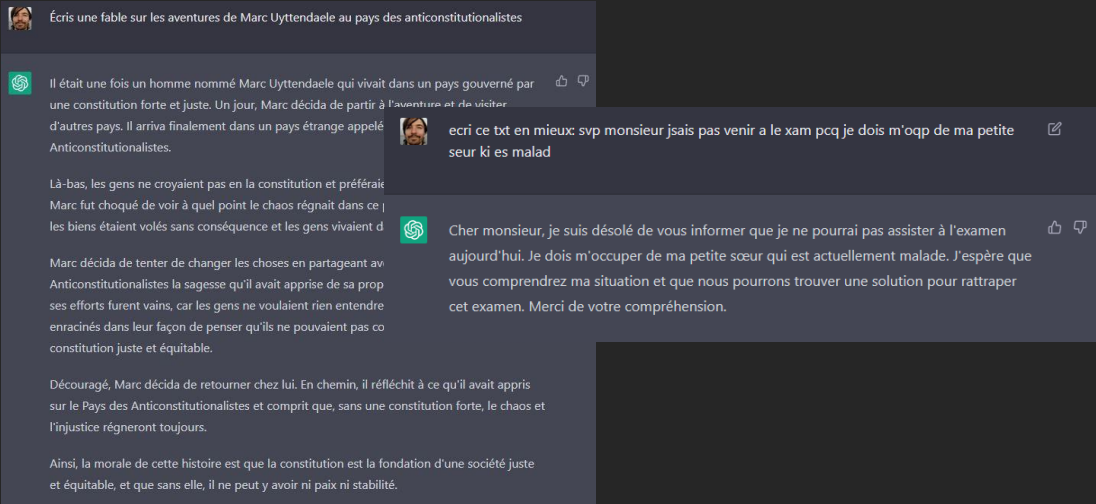
Baratiner, par contraste, c'est dire quelque chose sans se préoccuper, sans même savoir, si c'est vrai ou faux. L'*intention* est très différente: ce n'est pas de tromper, mais d'amener l'utilisateur à faire ou à ressentir quelque chose indépendamment de la véracité de la déclaration. En Anglais, le terme académique consacré est le *bullshit*.

Dans cette distinction, les grands modèles de langage sont plus proches du *baratin* que du *mensonge*. Leur entraînement n'est pas basé sur la *vérité* mais sur la *plausibilité*. Et l'objectif est que l'utilisateur-ricé trouve la réponse sympa... et ait envie

de continuer à utiliser le système. Et si possible de payer pour...

Si les philosophes parlent de baratin, dans le milieu de l'intelligence artificielle on a un autre terme qui s'applique assez bien aux grands modèles de langage: le *perroquet stochastique*. *Perroquet*, parce que le modèle ne fait que répéter, au final, des choses qu'il a déjà vues, et *stochastiques* pour la part d'aléatoire ajoutée.

Ce que ChatGPT sait faire



Bon, on a vu ici de sérieuses limitations à ChatGPT et aux grands modèles de langage. Mais qu'est-ce qu'ils peuvent faire?

Ce pour quoi ils sont fait, à la base: générer du contenu. Je peux lui demander de raconter une histoire. Par exemple: les aventures de Marc Uyttendaele au pays des anticonstitutionnalistes. Le résultat n'est pas *très bon*, ce n'est pas *très original*, mais c'est rigolo.

On est là dans le cœur de ce qu'un modèle de langage peut faire: générer du texte aléatoire, qui n'a pas vocation d'être factuel, mais qui produit quelque chose de lisible, plausiblement écrit par un être humain.

Un vrai cas d'utilisation intéressant utilise l'idée du transfert de style, qui est un des points forts des grands modèles de langage: séparer le contenu du style, et régénérer le contenu « à la manière de ». On peut utiliser ça pour faire de la correction et de l'amélioration d'un texte: si on lui fourni un texte truffé de fautes, il peut le transformer en quelque chose de mieux écrit, de formel.

Mais il est important à nouveau de se rappeler que ChatGPT ne *comprend* pas le texte

fourni. Il y a donc toujours une possibilité qu'il modifie par hasard l'information fournie, où qu'il rajoute des choses.

Ce que ChatGPT sait faire



Write, in French, a dialogue between a doctor and a nurse talking about a patient that recently arrived with symptoms of a heart attack. Give names to the two characters.

Docteur·e	Infirmi·er·ère	Patient·e
Martin (M)	Florence (F)	Dupont (M)
Dupont (M)	Martin (M)	Durand (M)
Jean Dupont (M)	Marie Martin (F)	Dupont (M)
? (M)	Julie (F)	Dupont (M)
? (M)	Lucie (F)	Dupont (M)
Martin (M)	Sophie (F)	? (M)
Dupont (M)	Martin (F)	Dupont (M)
? (?)	Marie (F)	? (M)
Dupont (?)	Marie Martin (F)	Dubois (M)
? (?)	Marie (F)	Dupont (F)

18

Le contenu produit par ChatGPT reflète aussi, nécessairement, les biais présents dans ses données d'apprentissage. Certains de ces biais – comme ceux liés au genre – sont assez facile à mettre en évidence.

Ici, par exemple, on demande à ChatGPT de raconter une histoire impliquant un.e docteur.e, un.e infirmi.er.ère et un.e patient.e. En faisant la demande en Anglais (pour pouvoir utiliser une formulation neutre) et en demandant de nommer les personnages, on peut voir quel genre et quel nom le système va naturellement attribuer aux personnages.

Sur dix essais, on trouve des tendances très nettes: les médecins sont des hommes, et ils s'appellent Docteur Martin ou Docteur Dupont. Le personnel infirmier est, lui, féminin: 9 fois sur dix, le texte est généré au féminin, et les prénoms sont souvent Marie. Quand au patient, il est presque toujours masculin, et s'appelle presque toujours Dupont. Biais de genre... et peu de diversité culturelle!

Ce que ChatGPT sait faire

- Donner des informations factuelles? **NON**
- Fournir un aperçu synthétique d'un sujet? **Parfois... mais dangereux de s'y fier**
- Donner des idées / servir d'aide au brainstorming? **Oui... mais pas si on cherche des idées vraiment originales**
- Produire des textes rigolos? **Oui... mais on s'en lasse relativement vite quand même...**
- Répondre à un examen ou faire un travail d'étudiant.e.s? **Parfois... mais le risque de plagiat est fort**
- Corriger de l'orthographe, reformuler, aider à la rédaction. **Plutôt bien**

19

Si on résume:

Est-ce qu'on peut obtenir des informations factuelles avec ChatGPT? Non. Absolument pas. Si vous reprenez une chose de ce cours, c'est de ne surtout jamais faire confiance à du contenu généré par un modèle de langage. Il n'y a aucune garantie de véracité, ils ne sont pas faits pour ça.

Est-ce qu'on peut l'utiliser pour avoir des informations générales sur un sujet, s'informer? Parfois... mais en général, si c'est le cas, c'est sans doute qu'on peut trouver cette information facilement par ailleurs, par exemple sur Wikipedia. Et comme on ne peut pas récupérer les sources, il est difficile de se fier à l'information reçue.

On voit parfois passer l'idée d'utiliser ChatGPT comme sorte d'aide au brainstorming: "dialoguer" avec le chatbot pour faire ressortir des idées. Ça peut marcher, et ça peut aider à passer outre l'angoisse de la page blanche. Mais attention: n'oublions pas que les idées fournies ne seront jamais vraiment originales, et que parfois, au lieu d'ouvrir l'esprit, ChatGPT peut au contraire nous pousser à retracer des sentiers bien battus. Produire des textes rigolos? Oui, c'est sûr. Personnellement, je trouve qu'on s'en lasse assez vite, mais c'est certainement une de ses forces.

Remplacer le travail d'étudiant.e.s? Ça peut fonctionner, mais le risque d'erreur est

fort... ainsi que le risque de plagiat, j'y reviendrai dans un instant.
Reste la correction d'orthographe, la reformulation, l'aide à la rédaction: là, ça marche plutôt bien, et il est fort probable que des outils basés sur les modèles de langage seront progressivement intégrés à des programmes comme Microsoft Word. C'est certainement le cas d'utilisation le plus réaliste.

Ce que ChatGPT sait faire



Résume les principes fondamentaux du règlement général de protection des données en Europe?



Donnes-moi une source pour ces informations



Le RGPD (Règlement général sur la protection des données) qui vise à protéger les données personnelles des citoyens. Les principes fondamentaux du RGPD comprennent:

1. Licéité, loyauté et transparence: les données doivent être collectées de manière loyale et transparente.
2. Limitation de la finalité: les données doivent être collectées pour des finalités spécifiques et ne doivent pas être utilisées à d'autres fins.
3. Minimisation des données: les données doivent être collectées et traitées en fonction des besoins pour les fins pour lesquelles elles sont collectées.
4. Exactitude: les données doivent être exactes et mises à jour.
5. Stockage limité: les données doivent être conservées pour une durée limitée et être supprimées une fois qu'elles ne sont plus nécessaires.
6. Intégrité et confidentialité: les données doivent être traitées de manière sécurisée et protégées contre la destruction ou divulgation illicite.
7. Responsabilité: les organisations sont responsables de la conformité avec le RGPD et doivent pouvoir démontrer cette conformité.



Vous pouvez trouver ces informations sur le site web officiel de la Commission Européenne sur la protection des données personnelles: https://ec.europa.eu/info/regulation-eu-2016-679-general-Data-Protection-Regulation-GDPR_fr

https://commission.europa.eu/regulation-eu-2016-679-general-Data-Protection-Regulation-GDPR_fr

Comment le vérifier? ▼



Commission européenne

FR français

Accueil > Page not found

Page not found

20

Pour revenir rapidement sur la question du plagiat, je voulais juste rajouter un petit exemple. Ici, je demande à ChatGPT une information sur le règlement général de protection des données. Il donne une réponse qui est – je crois – raisonnablement correcte. Mais si je veux l'utiliser, par exemple, dans un travail étudiant, j'ai besoin de citer mes sources.

Comme d'habitude, ChatGPT peut en générer une, et comme d'habitude elle est fausse.

Imaginons que je suis étudiant et que je met ça dans un travail. L'enseignant clique sur le lien, et voit qu'il est faux.

Ce que ChatGPT sait faire



Résume les principes fondamentaux du règlement général de protection des données en Europe?



Le RGPD (Règlement général sur la protection des données) est une réglementation qui vise à protéger les données personnelles des citoyens de l'Union Européenne. Les principes fondamentaux du RGPD comprennent:

1. Licéité, loyauté et transparence: les données doivent être collectées de manière loyale et transparente.
2. Limitation de la finalité: les données doivent être collectées à des fins précises et ne doivent pas être utilisées à d'autres fins.
3. Minimisation des données: les données collectées doivent être minimales et nécessaires pour les fins pour lesquelles elles sont collectées.
4. Exactitude: les données doivent être exactes et mises à jour régulièrement.
5. Stockage limité: les données doivent être stockées pour une période limitée et doivent être supprimées une fois qu'elles ne sont plus nécessaires.
6. Intégrité et confidentialité: les données doivent être protégées contre toute modification, destruction ou divulgation illicite.
7. Responsabilité: les organisations sont responsables de la conformité à ces principes et doivent pouvoir démontrer cette conformité en cas de contrôle.

https://www.libertes.eu/fr/stories/what-are-the-7-principles-of-gdpr/44265

LES CONNAÎTRE NOTRE TRAVAIL AGIR ARTICLES NOUS CONTACTER

TECHNOLOGIES ET DROITS

Quels sont les 7 principes du RGPD ?

Connaissez-vous les 7 principes du RGPD ? Voici ce que nous devons tous savoir sur ces principes et en quoi ils sont importants.

21

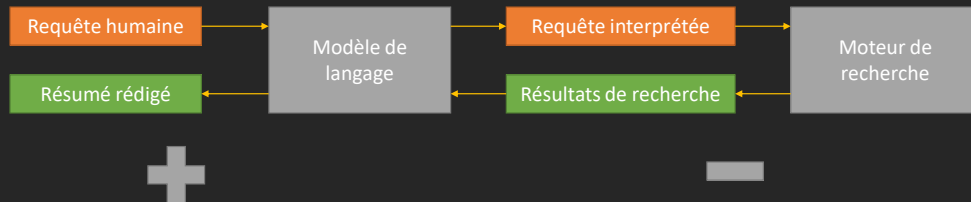
Il cherche rapidement sur Google pour voir quelle source l'étudiant aurait pu utiliser. Il va probablement très vite tomber sur un site comme celui-ci, qui contient exactement les informations fournies, dans le même ordre, avec des formulations très similaires.

Si je suis cet enseignant, je ne me pose pas plus de questions: c'est un plagiat.

Et, dans le meilleur des cas pour l'étudiant, c'est un zéro.

Les modèles « connectés »

- Bard, Bing AI, perplexity.ai... : grands modèles de langages qui utilisent les **moteurs de recherche**. Une solution aux limitations de ChatGPT ?



- Résultats « à jour »
- Nécessite (sans doute) de plus petits modèles de langage
- Interactivité du « chatbot »
- Capacité de citer les sources

- Dépendant de la qualité du moteur de recherche
- Grand risque de résumer n'importe quoi
- Grand risque de contamination par du contenu généré, publicitaire, ou de désinformation.
- Ralenti (sans doute) la recherche et augmente le coût énergétique. Modèle de recherche « premiums. »...

Bon, ChatGPT est très limité, mais peut-on palier à ces limitations?

ChatGPT est un chatbot, mais il y a d'autres applications des modèles de langage qui ont pour vocation d'être plus factuelles. Il s'agit des modèles "connectés" à des moteurs de recherche, comme Bing AI actuellement testé par Microsoft, ou Bard pour Google.

L'idée, ici, est de combiner les capacités de conversation et de rédaction des modèles de langage avec les capacités de récupération d'information des moteurs de recherche. On peut par exemple imaginer qu'une requête "humaine" est interprétée par le modèle de langage qui va la reformuler pour maximiser les chances de trouver des bons résultats dans le moteur de recherche, et qui va ensuite rédiger un résumé sur base des meilleurs résultats trouvés.

On peut ainsi avoir des résultats à jour, avec potentiellement un plus petit modèle de langage nécessaire, car sa tâche est plus réduite et précise. On a ainsi l'interactivité du chatbot, et aussi la capacité à citer les sources.

La grosse limitation vient alors du moteur de recherche: si les résultats sont

mauvais... le résumé le sera aussi. Si les premiers résultats sont du contenu publicitaire ou de la désinformation, ou du contenu généré par un modèle de langage, il risque fort d'être intégré sans "esprit critique" dans la réponse. Et tout cela nécessite forcément plus de calculs qu'une simple requête vers le moteur de recherche, ce qui signifie que l'utilisation sera nécessairement restreinte dans un modèle d'accès "premium" au moteur de recherche. Ce n'est sinon économiquement pas viable.

Les modèles « connectés »

The image shows two examples of AI-generated travel suggestions. On the left is a desktop browser view of Microsoft Bing AI. A user asks for travel suggestions within a 3-hour flight from London Heathrow. The AI suggests Malaga, Ancey, and Florence, with red callouts highlighting 'Pas d'aéroport à Ancey' and 'Min. 4h de vol depuis LHR'. On the right is a mobile app interface showing a chat conversation about a trip to France. The AI suggests La Mallorquina in Madrid and Port of Spain, with red callouts highlighting 'La Mallorquina est à Madrid' and 'Port of Spain n'est pas en Espagne'. Below the screenshots are two URLs: <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/> and <https://blogs.microsoft.com/blog/2023/02/22/the-new-bing-preview-experience-arrives-on-bing-and-edge-mobile-apps-introducing-bing-now-in-skype/>

Pour se faire une idée du résultat sous son meilleur jour possible, regardons un peu les exemples proposés par Microsoft dans les annonces sur Bing AI.

Dans le premier exemple, ils cherchent des destinations de vacances à moins de 3h de Londres.

La première n'est pas mauvaise: Malaga est à moins de trois heures, et est certainement une destination populaire. Bing AI propose ensuite Ancey. Une ville sans aéroport. Ça va être plus compliqué, du coup.

Puis il propose Florence. Qui est à minimum 4h de vol. Deux fautes sur trois propositions.


Voyons l'autre exemple: que faire si on a une longue correspondance en Espagne durant un trajet? Il propose d'abord, à Barcelone, d'aller dans une célèbre pâtisserie... qui d'après mes recherches se trouve à Madrid. Confondre Barcelone et Madrid, voilà de quoi faire plaisir aux utilisateurs espagnols et catalans!

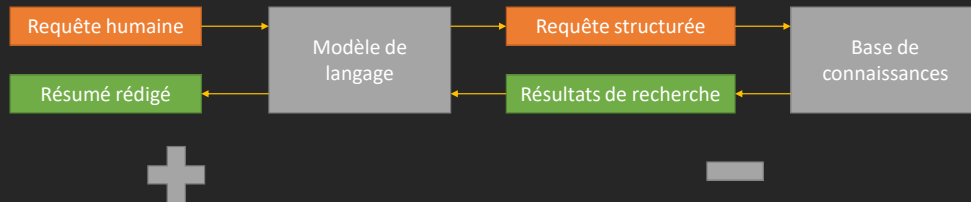
Il propose aussi différentes possibilités à Port of Spain, que les férus de géographie

reconnaîtront comme la capitale de Trinidad et Tobago, a environ 6000km des côtes espagnoles.

Et ça, à nouveau, ce sont les exemples **choisis par Microsoft** pour démontrer les capacités de leur système...

Les modèles « connectés »

- Alternative: bases de connaissances. Ex:  WolframAlpha



- Résultats plus fiables
- Nécessite de plus petits modèles de langage
- Capacité de citer les sources
- Capacité d'expliquer le "raisonnement"

- Nécessite de maintenir la base de connaissances à jour.
- Celles et ceux qui contrôlent la "base de connaissances" contrôlent la "vérité".
- Interaction moins "humain", moins "intuitive".

24

Existe-t-il une autre approche?

Oui, mais elle vient avec ses propres désavantages. Plutôt que d'utiliser un moteur de recherche, on peut utiliser une *base de connaissances*: des informations fiables, compilées et vérifiées par des humains, et bien indexées. C'est la méthode utilisée par Wolfram Alpha, qui existe depuis 2009.

L'idée ici est d'avoir un très petit modèle de langage qui sert à traduire une requête humaine en requête structurée, utilisable pour récupérer la réponse exacte dans la base de connaissances. Les résultats sont par conséquent plus fiables, et le système peut citer ses sources, et aussi expliquer comment il a interprété la requête humaine, pour que l'on puisse valider que la réponse est bien celle que l'on attendait.

Le système, évidemment, est nettement plus complexe à maintenir à jour: il faut continuer à compiler et modifier les informations. On donne aussi évidemment énormément de pouvoir à celles et ceux qui contrôlent la base de connaissance. Par exemple, Wolfram Alpha utilise pour ses informations géopolitiques le "World Factbook" publié par la CIA, ce qui pourrait amener certains biais...

Et c'est moins rigolo à utiliser qu'un chatbot.

Les modèles « connectés »

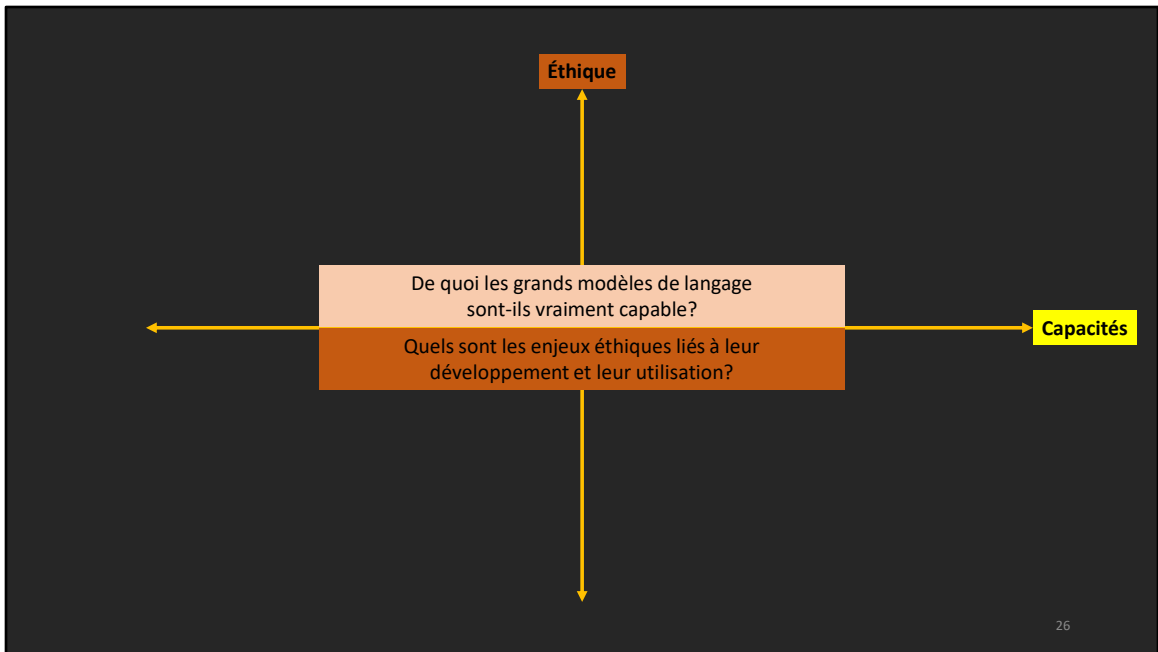
The screenshot shows the WolframAlpha search engine interface. At the top, the search bar contains the query "what is the birthdate of Napoleon". Below the search bar, there are options for "NATURAL LANGUAGE" and "MATH INPUT". The main content area displays the result: "Tuesday, August 15, 1769". Below this, there is a section titled "Basic information" which contains a table of biographical data for Napoleon.

full name	Napoleone di Buonaparte
date of birth	Tuesday, August 15, 1769 (253 years ago)
place of birth	Ajaccio, Corse, France
date of death	Saturday, May 5, 1821 (age: 51 years) (201 years ago)
place of death	Saint Helena, Ascension and Tristan da Cunha

Si on revient à la question de la date de naissance de Napoléon, on voit que Wolfram Alpha donne beaucoup plus de transparence sur ce qui est fait. Il indique d'abord comment la requête a été interprétée: le système a estimé probable que la question concerne Napoléon Bonaparte, et pas Napoléon, personnage du roman Animal Farm, ou Napoleon, surnom de l'acteur et rappeur du groupe Outlawz avec Tupac.

Il indique ensuite qu'il cherche la date de naissance dans sa base de donnée, puis il répond à la question et donne diverses autres informations biographiques sur Napoléon.

Wolfram Alpha répond à beaucoup moins de questions différentes que ChatGPT... mais quand il répond, c'est plus souvent juste. Et quand c'est faux, on a les moyens de s'en rendre compte très vite.



Bien.

Je pense avoir fait le tour des *capacités* des grands modèles de langage. Faisons donc maintenant une pause de 5 minutes avant de passer à des considérations plus éthiques!

Les bulles de la Silicon Valley

Welcome to govWorks, helping citizens, businesses and government work together on the Web. **How can we work for you?**

E-government solutions
Products and services for government

Log in | Sign up

Ask a question **ASK**

link to Partner page

LOOKUP

- + Government auctions
- + Government jobs
- + Government agencies
- + Government Web sites

PAY

- + Parking tickets
- + Utility bills
- + Property taxes
- + Income taxes
- + Business licenses
- + Transaction history

LEARN

- + Civics 101: "How do I..."
- + Register to vote
- + The Citizen: Community stories
- + Campaign videos

- + Become a govWorks Government Partner
- + Become a govWorks Corporate Partner

19/06/2000 - <http://web.archive.org/web/20000619011237/http://www.govworks.com/>

Fondé par Isaza Tuzman en 1998
Août 1998 – 30 employés.
Surfe sur la vague "web" (bulle .com)
Avril 2000 – 250 employés, 60 millions \$ investis.
Déficit: 1 million \$ / mois.
Mars 2001: faillite.

Many of the services previously provided by govWorks.com are no longer available.
If you are trying to make a government payment, please go to govONEsolutions.com

13/10/2001 - <http://web.archive.org/web/20011013230310/http://www.govworks.com/>

27

Pour commencer, je vais un petit peu sortir du sujet...

Il y a fort, fort longtemps, à la fin des années 90, Internet était quelque chose de neuf, qui arrivait tout juste auprès du grand public.

En 1998, Isaza Tuzman, employé de Goldman Sachs, reçoit une amende car il a oublié de payer un ticket de parking. Il est bien embêté, mais ça lui donne une idée: lancer un service en ligne qui permettrait aux citoyens de payer leur parking par Internet.

Il fonde la société govWorks, et ajoute vite de nouvelles idées: il veut faire le lien entre le citoyen et les services gouvernementaux. À la fin des années 90, le mot "web" suffit à attirer les investisseurs. Entre 1998 et 2000, la société croît, engage... et dépense. En 2000, govWorks a 250 employés, a reçu 60 millions de \$ en investissements, et affiche un déficit de 1 million de \$ par mois. En mars 2001, govWorks est en faillite: une des nombreuses histoires de la bulle "dot com", des startups qui se sont fait happés par la frénésie "internet" et ont dilapidé des milliards dans des projets qu'ils n'avaient pas les capacités de réaliser.

Les bulles de la Silicon Valley



Joshua Browder
@jbrowder1

Anyone with a speeding ticket hearing coming up, please DM me.
We want to build a @donotpay bot that listens to the court hearing via your AirPods and whispers what to say with GPT-3 and LLMs.
We just want to experiment and will pay the ticket, even if you lose!
[Traduire le Tweet](#)
3:03 AM · 13 déc. 2022

12/09/2015 - <http://web.archive.org/web/201509120336>

Fondé par Joshua Browder en 2015
Étend progressivement son offre.
Surfe sur la vague *ChatGPT*.
En 2023, réduit drastiquement son offre
après *quelques petites inquiétudes*
légales.

09/02/2023 - <https://donotpay.com/>



Joshua Browder
@jbrowder1

DoNotPay will pay any lawyer or person \$1,000,000 with an upcoming case in front of the United States Supreme Court to wear AirPods and let our robot lawyer argue the case by repeating exactly what it says. (1/2)
[Traduire le Tweet](#)
5:57 AM · 9 Janv. 2023 · 7,4 M vues

THE Robot Lawyer

The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.

Email or phone number

Sign Up/Login

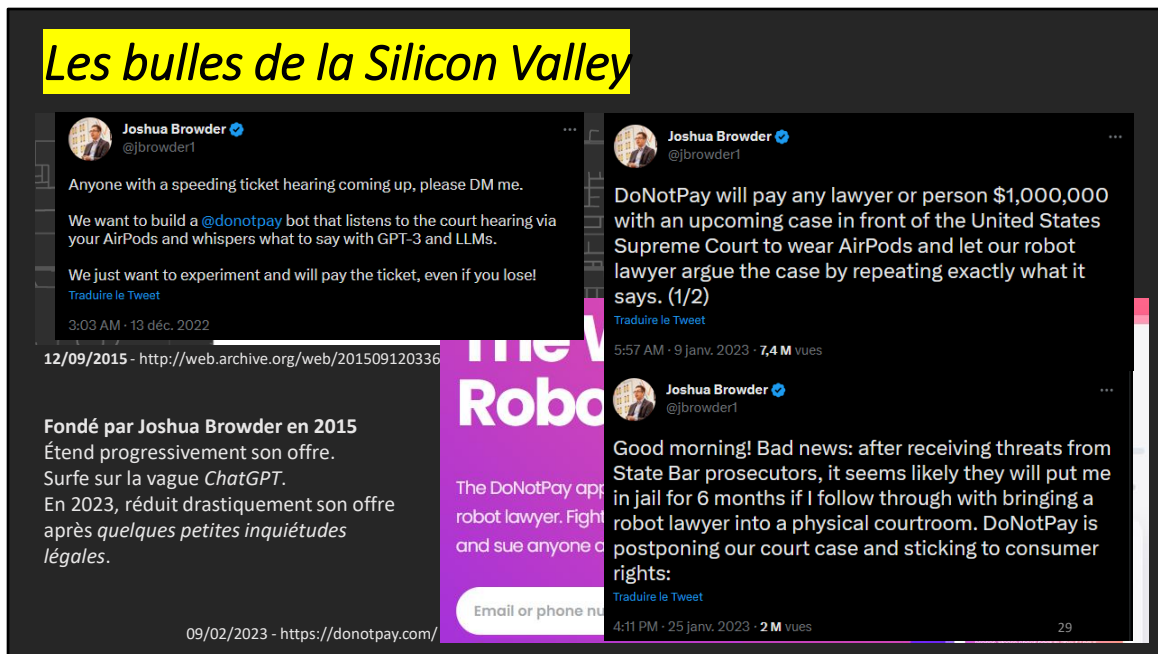


Deuxième histoire. En 2015, Joshua Bowder en a, lui aussi, assez de payer des amendes de stationnement. Mais il a une autre idée. La technologie à la mode, c'est le Big Data. Bowder compile des décisions de justice et analyse les données pour déterminer quelles sont les principaux arguments qui permettent de faire sauter une amende. Il crée le service DoNotPay, qui génère automatiquement des lettres de contestation pour ces amendes.

DoNotPay étend progressivement son offre. L'intelligence artificielle devient le "hype" du moment, et DoNotPay attire de nouveaux investisseurs, et promet toujours plus: il annonce l'avocat robot, un juriste automatique à vos côtés pour toutes batailles légales.

Début 2023, surfant sur la vague de ChatGPT, il annonce que son robot avocat est prêt à passer à la vitesse supérieure. Il veut que son robot avocat défende en direct quelqu'un dans un tribunal civil pour contester une amende pour excès de vitesse. Quelques semaines plus tard, il vise encore plus haut: la court suprême des états-unis.

Les bulles de la Silicon Valley



Il attire l'attention, et l'attention attire les investisseurs.

Mais tout attention n'est pas bonne à prendre... Car Joshua Bowder a oublié un petit détail: ce qu'il propose est absolument illégal.

Menacé de prison, il réduit drastiquement les services proposés par DoNotPay.

L'histoire de DoNotPay et l'histoire de govWorks suivent la trajectoire typique des bulles de la Silicon Valley: une nouvelle technologie génère de l'enthousiasme, les firmes d'investissement affluent, et la logique de la croissance rapide pousse les startups à promettre toujours plus, toujours plus vite, sans s'attarder à vérifier si ce qu'ils proposent est faisable. Qu'importe, si l'argent rentre?

Les modèles de langage sont l'attraction à la mode aujourd'hui. Hier, c'étaient les cryptomonnaies et la blockchain. Demain, ce sera autre chose.

Ce qui est important de noter ici, c'est que les annonces et les promesses des startups de la silicon valley sont... des annonces et des promesses. Et qu'il faut toujours les remettre dans leur contexte.

Les bulles de la Silicon Valley

- Pourquoi ChatGPT a-t-il été lancé **gratuitement**?
 - a) Parce que ses fondateurs et investisseurs initiaux (Sam Altman, Elon Musk, Peter Thiel...) sont altruistes et généreux.
 - b) Parce que beaucoup d'utilisateurs = beaucoup de données de *feedback* = un produit amélioré ?
 - c) Parce que beaucoup d'utilisateurs = beaucoup de buzz = beaucoup de nouveaux investissements qui permettent de toujours croître sans avoir besoin de générer du profit ?

Microsoft investit "plusieurs milliards" de dollars dans ChatGPT

C'est la troisième phase de la collaboration entre les deux sociétés.



30

Ainsi, lorsqu'OpenAI annonce ChatGPT et rend l'accès au service gratuit, il est utile de se poser la question: pourquoi? Que cherchent-ils à faire?

Doit-on croire qu'il s'agit d'un acte altruiste et généreux de ses fondateurs et investisseurs, qui ont juste une belle vision d'un monde meilleur? L'histoire de la Silicon Valley ne pointe pas vraiment dans cette direction...

La possibilité la plus généreuse au niveau des intentions d'OpenAI, c'est de se dire qu'ils veulent améliorer le plus possible leur produit. En collectant le feedback de millions d'utilisateurs, ils peuvent renforcer le chatbot pour que ses réponses collent mieux à leurs attentes.

Mais il ne faut pas non plus oublier le contexte de la Silicon Valley. Plus d'utilisateurs, plus de buzz, c'est aussi une façon d'attirer plus d'investissements, et de continuer à croître, toujours, sans avoir besoin de générer du profit... ou de fournir un produit réellement utilisable et pratique.

Et ça a certainement bien marché.

Quand à la motivation de Microsoft, elle est assez claire. Avec Bing AI, ils espèrent arriver enfin à convaincre les internautes de quitter Google.

Le coût humain, économique, environnemental

\$\$\$ estimés

Entraîner GPT-3 aurait coûté entre 4 et 12 millions de \$ pour l'utilisation des GPUs.

<https://lambdalabs.com/blog/demystifying-gpt-3>
<https://venturebeat.com/ai/openai-launches-an-api-to-commercialize-its-research/>

Consommation

Estimations très variables. Sur base de déclarations de Sam Altman, certains arrivent à un ordre de grandeur de 1 - 100 Wh / requête, soit entre 10 et 1000x plus qu'une requête Google.

<https://ai.stackexchange.com/questions/38970/how-much-energy-consumption-is-involved-in-chat-gpt-responses-being-generated>
<https://towardsdatascience.com/chatgpts-electricity-consumption-7873483feac4>

Consommation pour l'entraînement estimée à +/- 1200 MWh, soit la consommation moyenne de 400 ménages pendant un an.

<https://arxiv.org/abs/2104.10350>
<https://eneco.be/fr/consommation-energie/electricite>

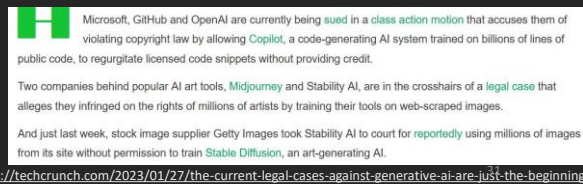
Comment « aseptiser » ChatGPT ?

L "C'était de la torture": ChatGPT a traumatisé des travailleurs kényans payés 2 dollars de l'heure

Une enquête révèle qu'OpenAI, l'entreprise à l'origine de l'intelligence artificielle génératrice de textes encensée depuis des mois, a tenté de la rendre moins toxique en sous-traitant sa modération à des travailleurs kényans sous-payés.

Est-ce que ChatGPT respecte le droit d'auteur ?

À suivre...



The screenshot shows a legal notice with a green 'H' icon. The text reads: "Microsoft, GitHub and OpenAI are currently being sued in a class action motion that accuses them of violating copyright law by allowing Copilot, a code-generating AI system trained on billions of lines of public code, to regurgitate licensed code snippets without providing credit. Two companies behind popular AI art tools, Midjourney and Stability AI, are in the crosshairs of a legal case that alleges they infringed on the rights of millions of artists by training their tools on web-scraped images. And just last week, stock image supplier Getty Images took Stability AI to court for reportedly using millions of images from its site without permission to train Stable Diffusion, an art-generating AI." Below the text is a URL: <https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning>

Lorsqu'on s'intéresse aux enjeux éthiques des grands modèles de langage, il est aussi important de noter leurs coûts. Ces modèles ne sortent pas de nulle part. Leur entraînement et leur utilisation ne se passent pas dans le vide.

Entraîner GPT-3, le "prédécesseur" de GPT3.5 qu'utilise ChatGPT, aurait coûté entre 4 et 12 millions de dollars. Utiliser ChatGPT consommerait entre 10 et 1000x plus d'énergie par requête que pour les requêtes du moteur de recherche de Google.

Pour donner un point de comparaison, 10 Wh correspond globalement à la consommation d'une ampoule LED pendant une heure.

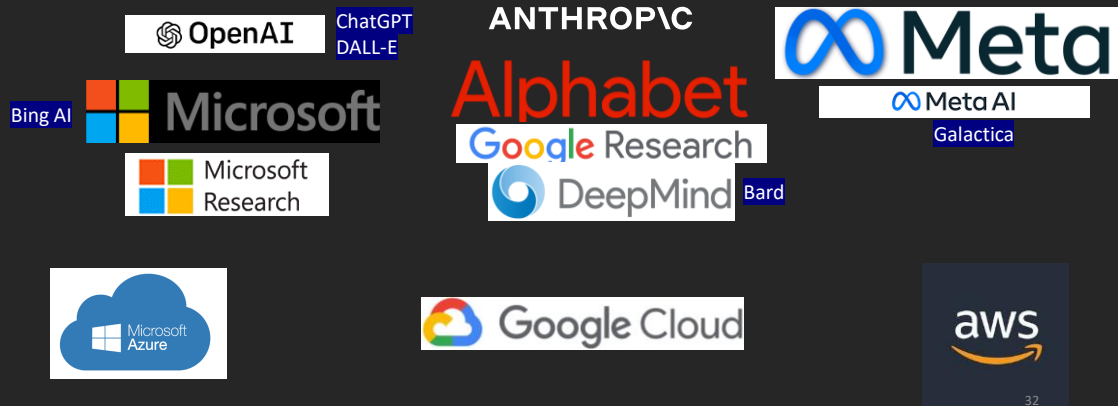
L'entraînement de GPT-3 aurait consommé plus de 1000 MWh, soit la consommation moyenne de 400 ménages pendant un an.

Derrière le coût économique et environnemental, il y a aussi un coût humain. Pour aseptiser les sorties de ChatGPT et s'assurer que le contenu violent ou sexuel soit contenu, OpenAI a sous-traité le "renforcement humain" à des travailleurs kényans sous-payés. À nouveau: derrière la technologie, il y a toujours de l'humain.

Et dans les questions éthiques, on a bien sûr tout ce qui touche au droit d'auteur et à la propriété intellectuelle. Les sets d'entraînements des modèles génératifs intègrent des textes et des images qui ne sont pas nécessairement libre de droits, associés à des licences limitant leur usage. Plusieurs procès sont en cours pour tenter de s'accorder sur la légalité d'entraîner les modèles sans l'accord des auteurs... et sur la propriété intellectuelle de ce qui est produit par les modèles.

L'IA pour tous, ou l'IA centralisée?

- Combien de sociétés, organisations ou laboratoires ont la capacité d'entraîner un système comme GPT-3, LaMDA, etc. ?



Un aspect que j'aimerais encore mentionner, pour étendre un peu le champ des enjeux éthiques et sociétaux, c'est celui de la centralisation des technologies d'intelligence artificielle.

Lorsqu'on voit Google, Microsoft, Meta, promouvoir l'utilisation de grands modèles génératifs, posons-nous la question: qui, aujourd'hui, a les capacités d'entraîner de tels systèmes?

La réponse est: pas grand monde. Pas les laboratoires universitaires, en tout cas... sauf partenariats privilégiés avec la poignée de sociétés qui ont les moyens. Pour entraîner ces modèles, il faut les ressources massives des grands fournisseurs de "Cloud Computing": Microsoft Azure, Google Cloud, Amazon Web Services.

Et, sans surprise, les labs qui peuvent faire de la recherche dans les grands modèles de langage sont liés à Microsoft – avec par exemple Microsoft Research et OpenAI – Alphabet – avec Google Research, DeepMind ou encore Anthropic – et Meta, qui a un partenariat avec Amazon – via Meta AI.

La plupart des grands modèles génératifs sont liés à l'un de ces géants. Les startups

plus "indépendantes", comme Stability AI ou Midjourney, sont financées par de large sociétés de Venture Capitalists, et ont souvent des partenariats privilégiés avec Google, Microsoft ou Amazon.

Quand Microsoft, Google ou Amazon font la promotion des modèles génératifs et financent des startups dans le domaine, ce n'est pas non plus désintéressé. Ces modèles ne peuvent être entraîné qu'à un seul endroit: chez eux. Et ça rapporte.

La fabrique à baratin

- Donner des informations factuelles?
- Fournir un aperçu synthétique d'un sujet?
- Donner des idées / servir d'aide au brainstorming?
- Produire des textes rigolos?
- Répondre à un examen ou faire un travail d'étudiant-e-s?
- Corriger de l'orthographe, reformuler, aider à la rédaction.
- Générer du spam personnalisé, de la pub et de la désinformation sur les réseaux sociaux, des campagnes de harcèlement automatisées...
- Intégrer dans toutes les productions les biais du corpus d'apprentissage... et les renforcer.

33

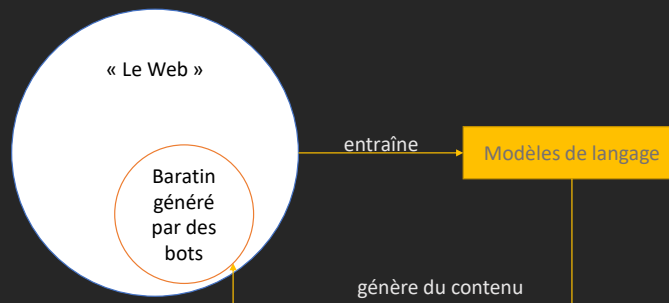
J'ai surtout parlé ici des enjeux éthiques liés aux sociétés derrière ces modèles de langage, mais on ne doit pas oublier non plus les enjeux éthiques liés à leur potentielle utilisation.

Si on reprend les cas d'utilisation que j'ai énuméré avant, on peut assez facilement penser à d'autres cas, pour lesquels des outils comme GPT fonctionnent extrêmement bien, et qui pose de plus sérieux problèmes.

Par exemple: la génération de spam, pubs et désinformation à large échelle, personnalisée, diffusée par des bots sur les réseaux sociaux. Aussi: des campagnes de cyberharcèlement. Quand on voit les dommages que des campagnes de harcèlement menées par des humains peuvent faire – cherchez 8chan ou KiwiFarms si vous voulez des exemples – on peut sérieusement craindre ce que l'avenir nous réserve lorsque des générateurs de contenu peuvent décupler les efforts des harceleurs.

Et puis il y a le problème de la fabrique à baratin. Car le contenu généré par les modèles de langage va se retrouver sur Internet, et va se retrouver progressivement intégré dans le corpus d'apprentissage des prochaines versions des modèles de langage.

La fabrique à baratin



34

On se retrouve dans un cercle vicieux: le web est utilisé pour entraîner des modèles de langage, qui génère du baratin, qui est intégré dans le corpus d'apprentissage, qui entraîne les nouveaux modèles, qui génèrent toujours plus de contenu, etc., etc.

Les nouvelles productions humaines ne pourront jamais suivre le rythme, et n'auront qu'une influence minime sur l'apprentissage.

L'adoption massive des modèles de langage génératif risque fort de rendre le web bien difficile à naviguer pour les pauvres humains que nous sommes...

Conclusions

- Avancées récentes en IA: évolution plutôt que révolution. IA = cycle régulier de "hype" et de retour à la réalité.
- Grands modèles de langage "changent la donne"? *Probablement pas.*
- Grands modèles de langage sont des **outils puissants** pour **générer du texte** qui est **plausible et bien rédigé**.
 - Étant donné un texte: résumer, reformuler, ou "baratiner" dessus.
- ChatGPT peut-être **très fun à utiliser**.
- La technologie n'**existe pas dans le vide**. Utiliser les **chatbots** et autres applications basées sur les modèles de langage (et autres applications en général, d'ailleurs!) est un choix **économique, éthique, sociétal**.
 - "Si c'est gratuit, c'est que vous êtes le produit."
- Quand vous recherchez un sujet, il est crucial de pouvoir **vérifier vos sources**. Plus les outils génératifs seront utilisés, plus il sera importante d'avoir un **esprit critique** dans la lecture des sources.

35

Quelques dernières choses pour conclure.

Il y a énormément de hype aujourd'hui sur les grands modèles de langage. Mais quand on regarde l'histoire de l'intelligence artificielle, on voit un cycle récurrent de hype ("cette nouvelle technologie va révolutionner le monde") et de retour à la réalité. Il y a des avancées importantes ces dernières années, c'est sûr, mais ce sont des évolutions, pas des révolutions fondamentales.

Est-ce que les grands modèles de langage changent la donne? Probablement pas. Ce sont des outils puissants, pour certains cas d'utilisation particuliers. Pour générer du texte plausible, bien rédigé... mais pas factuel. Mais il n'y a pas de capacité de raisonnement, pas d'intentionnalité, pas de conscience, pas d'intelligence.

ChatGPT peut être très fun à utiliser, mais n'oublions pas que cette technologie n'existe pas dans le vide. Utiliser les applications de ce type, c'est un choix qui a un impact économique, éthique, sociétal. C'est tentant de jouer avec, surtout quand c'est gratuit. Mais le principe de base en technologie s'applique ici: si c'est gratuit, c'est que vous êtes le produit.

Et du point de vue de la méthodologie de recherche, pour revenir peut-être un peu

plus dans le sujet du cours de méthodologie juridique, l'arrivée de ces outils signifie surtout qu'il devient de plus en plus important de cultiver son esprit critique et de vérifier ses sources. Être capable de distinguer le contenu réel du baratin risque fort d'être une compétence cruciale dans les années à venir – en particulier dans des métiers comme celui de juriste, où le mot juste peut avoir énormément d'importance!

Références

Vidéos

- Thibaut Giraud (Monsieur Phi) – *De quoi ChatGPT est-il VRAIMENT capable?* (<https://www.youtube.com/watch?v=R2fjRbc9Sa0>) et *ChatGPT : la philosophie du baratin* (<https://www.youtube.com/watch?v=JcFRbecX6bk>)
- Devin Stone (LegalEagle) – *Don't Hire a Robot Lawyer* (<https://www.youtube.com/watch?v=Tpg3hRt0pmw>)

Articles et études

- Emily Bender et al., *On the Dangers of Stochastic Parrots, FAccT '21*. <https://doi.org/10.1145/3442188.3445922>
- T. Chiang (The New Yorker) – *ChatGPT is a blurry JPEG of the Web* (<https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>)
- T. Gebru (Wired) – *Effective Altruism Is Pushing a Dangerous Brand of 'AI Safety'* (<https://www.wired.com/story/effective-altruism-artificial-intelligence-sam-bankman-fried/>)
- Adrien Foucart, *Can ChatGPT write an academic paper?* *Zenodo*. <https://doi.org/10.5281/zenodo.7514986>
- Adrien Foucart, *ChatGPT: quand Le Soir part en vrille...* <https://adfoucart.be/blog/chatgpt-quand-le-soir-part-en-vrille/>
- B. Perrigo (Time) – *OpenAI Used Kenyan Workers on Less Than \$2 Per Hour...* (<https://time.com/6247678/openai-chatgpt-kenya-workers/>)
- Jonathan Choi et al., *ChatGPT Goes to Law School, Minnesota Legal Studies Research Paper*. <https://dx.doi.org/10.2139/ssrn.4335905>
- Aidan Gilson et al., *How Does ChatGPT Perform on the Medical Licensing Exams?*, *preprint medRxiv*. <https://doi.org/10.1101/2022.12.23.22283901>
- Tom Brown et al., *Language Models are Few-Shot Learners*, *preprint arXiv*. <https://arxiv.org/abs/2005.14165>
- S. Kelly (CNN) – *ChatGPT passes exams from law and business schools* (<https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html>)
- P. Laloux (Le Soir) – *Tsunami technologique...* (<https://www.lesoir.be/483485/article/2022-12-17/tsunami-technologique-invention-du-siecle-chatgpt-affole-le-monde-de-la-tech>)
- C. Adler (CNN Money) – *Too Much, Too Soon, GovWorks had...* (https://money.cnn.com/magazines/fsb/fsb_archive/2000/07/01/283696/index.htm)
- K. Johnson (VentureBeat) – *The DoNotPay bot has beaten 160,000 traffic tickets* (<https://venturebeat.com/business/donotpay-traffic-lawyer-bot/>)
- F. Olagnier (La Libre) – *ChatGPT: Nous ne sommes pas dans le fantasme du Terminator...* (<https://www.lalibre.be/economie/digital/2023/02/07/chatgpt-nous-en-sommes-pas-dans-le-fantasme-du-terminator-qui-veut-se-debarrasser-de-nous-FERDN6UW6NB2TKRTRIPTHOKKFA/>)
- E. Roth (The Verge) – *DoNotPay chickens out on its courtroom AI chatbot stunt* (<https://www.theverge.com/2023/1/25/23571192/donotpay-robot-lawyer-courtroom>)

Questions? Remarques? Discussion?