

Sensibilisation aux enjeux et usages de l'Intelligence Artificielle

Didier Schwab



Temps 1

Premier exemple

Exemples d'interaction avec un agent conversationnel

Exemples d'interaction avec un agent conversationnel

Comment puis-je vous aider ?

Poser une question

+ 🔗 Outils





INTRODUCTION AUX POURCENTAGES POUR LES ÉLÈVES DE 6E

Comprendre les bases simples des calculs en pourcentage

DÉCOUVRIR LA NOTION DE POURCENTAGE



LES POURCENTAGES : COMPRENDRE « SUR 100 »

Définition du pourcentage

Le pourcentage représente une partie d'un tout sur une base de 100 unités, facilitant la compréhension des proportions.

Exemple de réduction

Une réduction de -30 % signifie une baisse de 30 unités sur 100 du prix initial, soit presque un tiers.

Applications quotidiennes

Les pourcentages apparaissent dans les examens, batterie de téléphone, statistiques sportives et promotions commerciales.

Bases essentielles

100 % correspond à la totalité, 50 % à la moitié, et 0 % à l'absence, bases pour toutes applications pratiques.

UTILITÉ ET DÉFINITION DES POURCENTAGES



POURQUOI UTILISER DES POURCENTAGES ?

Comparaison simple et universelle

Les pourcentages permettent de comparer facilement des parties avec un tout sans connaître les quantités exactes.

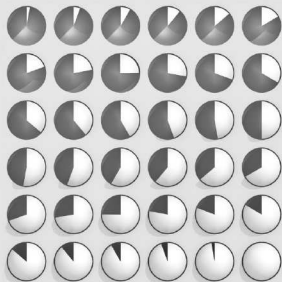
Utilisation quotidienne fréquente

Les pourcentages apparaissent dans les soldes, les statistiques et les indicateurs de performance courants.

Outil pédagogique interactif

Impliquer les élèves avec des exemples concrets motive l'apprentissage des pourcentages.

QU'EST-CE QU'UN POURCENTAGE ?



Définition du pourcentage

Un pourcentage exprime une proportion sur une base de 100, facilitant la comparaison entre différentes quantités.

Comparaison équitable

Le pourcentage permet de comparer équitablement des groupes de tailles différentes en standardisant à 100.

Valeurs clés du pourcentage

100 % signifie la totalité, 50 % la moitié, et 0 % rien, des repères essentiels à retenir.

Visualisation avec des objets

Utiliser des jetons pour représenter les pourcentages aide à mieux comprendre leur signification concrète.

ILLUSTRATIONS ET ÉCRITURES DES POURCENTAGES

ILLUSTRATIONS CONCRÈTES DU POURCENTAGE

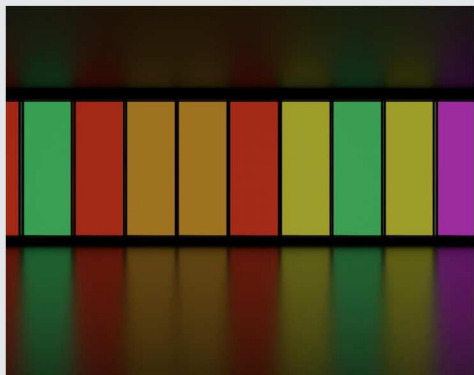


Tableau de 100 cases

Un tableau de 100 cases permet de visualiser chaque pourcentage comme une case colorée, facilitant la compréhension de la proportion.

Exemple batterie de téléphone

Une batterie indiquant 75 % est presque pleine, tandis qu'à 10 % elle est presque vide, illustrant le pourcentage dans la vie quotidienne.

Étiquettes de soldes

Les étiquettes de soldes affichant des réductions en pourcentage, comme -30 %, aident à associer le pourcentage aux achats et promotions.

Graphiques circulaires

Les graphiques circulaires illustrent visuellement la répartition en pourcentages, renforçant la compréhension grâce aux supports visuels.

POURCENTAGE	FRACTION	DÉCIMAL
10 %	10/100	0,1
25 %	25/100	0,25
50 %	50/100	0,5
75 %	75/100	0,75

Différentes façons d'écrire un pourcentage

ACTIVITÉS ET SYNTHÈSE

QUIZ SUR LES POURCENTAGES



Vérification de la compréhension

Le quiz rapide aide à vérifier la compréhension des différentes écritures des pourcentages par les élèves.

Exemples de questions

Des questions clés portent sur la conversion entre pourcentages, fractions et nombres décimaux.

Activité dynamique

L'interactivité avec ardoises ou à main levée rend l'activité plus engageante et motivante.

Renforcement des acquis

Ce type de quiz favorise la mémorisation et renforce la confiance des élèves en mathématiques.



INVENTER DES SITUATIONS CONCRÈTES

Contextes variés d'utilisation

Les élèves imaginent des exemples utilisant des pourcentages dans des contextes comme la classe, le sport ou les jeux.

Développement de la créativité

Cette activité stimule la créativité et la compréhension pratique des pourcentages au-delà des mathématiques pures.

Expression et échange oral

Les élèves présentent leurs exemples à l'oral, favorisant l'expression, la communication et l'échange entre pairs.

SYNTHÈSE : CE QU'IL FAUT RETENIR



Définition du pourcentage

Le pourcentage exprime une proportion sur 100, symbolisé par %.
C'est une manière simple de représenter des parts.

Notations du pourcentage

Les pourcentages peuvent s'écrire en notation %, fraction sur 100 ou nombre décimal, facilitant la compréhension.

Importance des pourcentages

Les pourcentages sont essentiels pour comparer, analyser et comprendre des données dans la vie quotidienne.

Valeurs fondamentales à retenir

100 % signifie tout, 50 % la moitié, et 0 % rien, bases importantes pour comprendre les proportions.

Temps 2

Fondements de Intelligence Artificielle

(Très) Bref historique

Les débuts de l'IA (années 1950)

- 1956 : Conférence de Dartmouth, marquant la naissance officielle du domaine.
- Création du terme *Artificial Intelligence*

1956 Dartmouth Conference: The Founding Fathers of AI



John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



Nathaniel Rochester



Trenchard More

Les femmes qui ont marqué l'histoire de l'IA



Grace Hopper (1906–1992)
Inventrice du premier
compilateur, à la base de la
programmation de haut niveau.



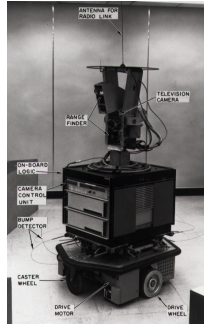
**Karen Sparck Jones
(1935–2007)**
Pionnière du traitement
automatique de la langue.



Alice Recoque (1929–2021)
Développement d'un langage
de programmation essentiel aux
début de l'IA.

L'IA symbolique et les premiers systèmes (années 1960-1970)

- Shakey (Stanford Research Institute) : premier robot mobile à raisonner sur son environnement.
- Systèmes experts : Tentative de reproduire le raisonnement humain dans des domaines restreints.



ELIZA, pionnière des chatbots (1966)

- Développée au MIT par Joseph Weizenbaum en 1966.
- Basée sur des règles de détection de mots-clés.
- Simule un psychothérapeute rogérien (reformulation des interrogations).
- Certains utilisateurs pensaient qu'ils interagissaient avec un humain.



L'IA gonflée à la puissance de calcul (années 1990)

- Montée de la puissance de calcul.
- Deep Blue (IBM, 1997) : victoire contre Garry Kasparov, champion du monde d'échecs.



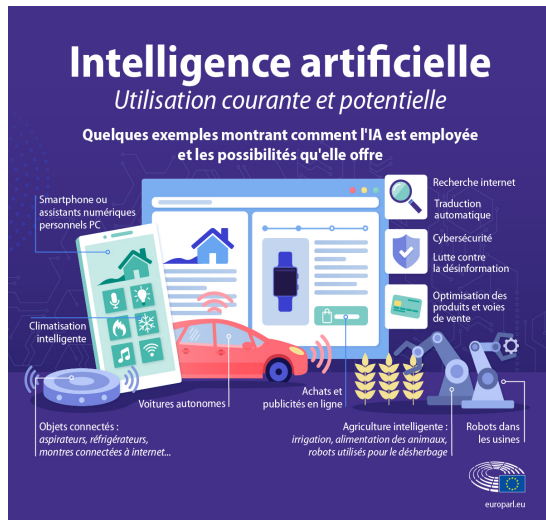
L'âge de l'apprentissage automatique (années 2010-2020)

- Réseaux de neurones profonds, GPU, apprentissage par renforcement
- AlphaGo (2016) : victoire contre Lee Sedol, joueur de go professionnel.



L'Intelligence Artificielle autour de nous

Exemples de Systèmes d'Intelligence Artificielle



Systèmes d'Intelligence Artificielle

Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle

Systèmes d'Intelligence Artificielle

Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (IA Act)

Article 3 (Définitions)

Alinéa 1 : « Système d'IA » « un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels; »

Qu'appelle-t-on Intelligence Artificielle prédictive ?

Des programmes informatiques conçus pour prédire des résultats à partir de données

Qu'appelle-t-on Intelligence Artificielle prédictive ?

Des programmes informatiques conçus pour prédire des résultats à partir de données



Qu'appelle-t-on Intelligence Artificielle Générative ?

Qu'appelle-t-on Intelligence Artificielle Générative ?

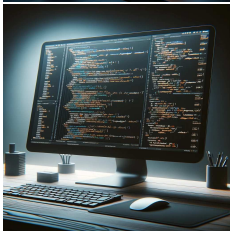
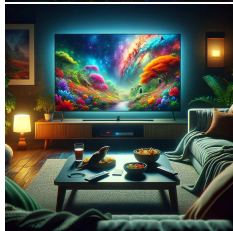
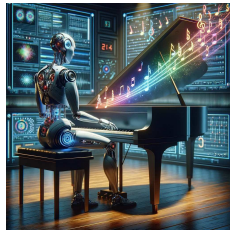
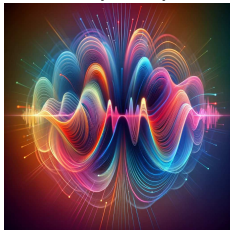
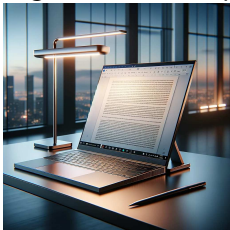
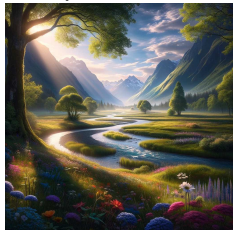
Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains

Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.

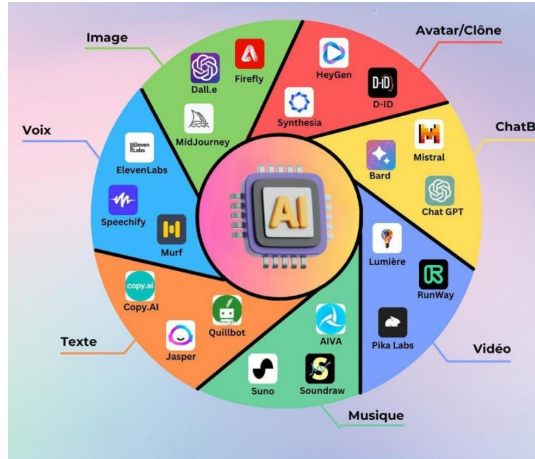
Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



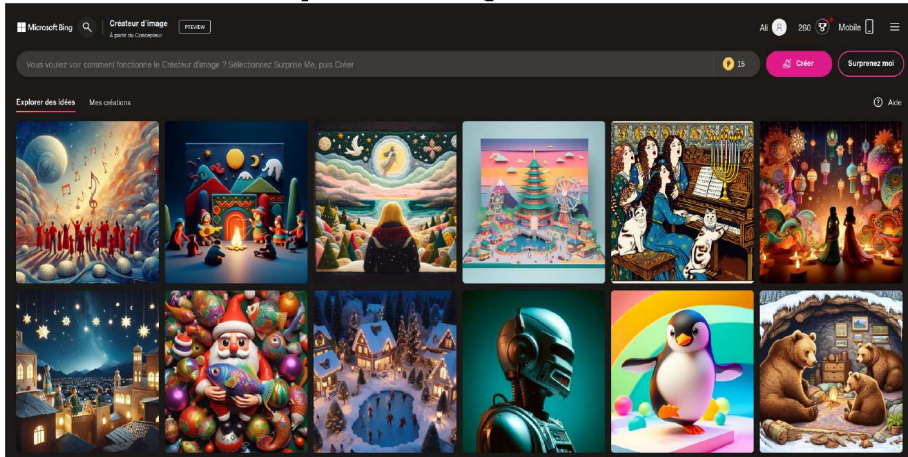
Panorama des IA génératives

<https://www.rapidops.com> [Mars 2023]



IA génératives d'images : Bing image

<https://www.bing.com/create>

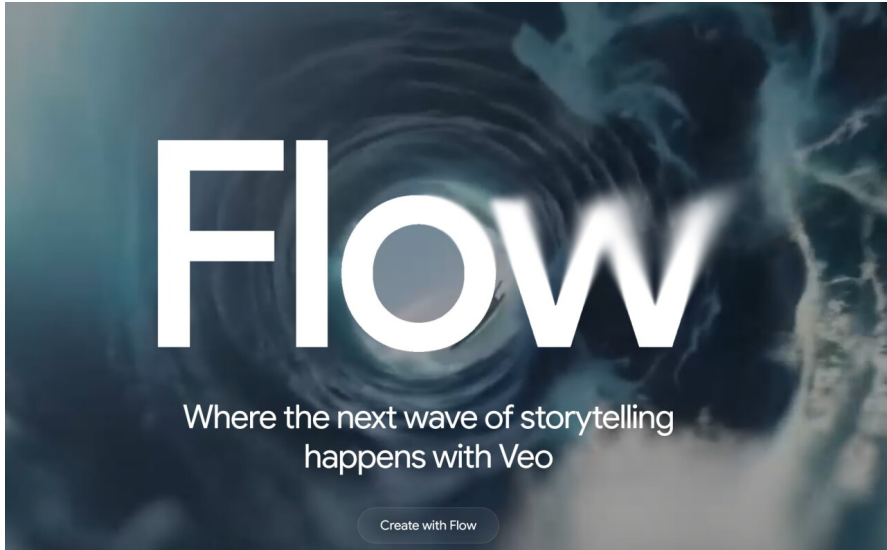


IA génératives d'images : Adobe Firefly

Rendu 3D d'un bébé perroquet, grands yeux adorables, dans un jardin avec des papillons et une végétation luxuriante, fantaisie et douceur, magie, poussière de fée



Génération de vidéos : Google Flow (20 mai 2025)



IA génératives pour la voix

- Text-2-speech : synthèse de la parole
- Existe depuis longtemps
- Avantages :
 - ▶ Apprentissage des langues
 - ▶ Accessibilité (personnes en situation de handicap)

IA génératives pour la voix

- Text-2-speech : synthèse de la parole
- Existe depuis longtemps
- Avantages :
 - ▶ Apprentissage des langues
 - ▶ Accessibilité (personnes en situation de handicap)
- MAIS la nouveauté
 - ▶ Voix très réaliste
 - ▶ Voix imitée à partir d'un enregistrement d'une minute seulement

Des avatars réalistes ?

- Exemple : <https://www.heygen.com>
(open sstik.io_@emmanuelmacron_1751014131872.mp4)



Des avatars réalistes ?

- Exemple : <https://www.heygen.com>
(open sstik.io_@emmanuelmacron_1751014131872.mp4)



- Cas d'usage :
 - ▶ Apprentissage : meilleure interactivité, contenus plus uniformes, plus d'engagement
 - ▶ Expérience utilisateur personnalisée : essayage virtuel de vêtements ou prévisualisation de produits
 - ▶ Accessibilité (ex : handicaps sociaux)

Et le texte dans tout cela?

- Central à ces applications
 - ▶ Textes à prononcer
 - ▶ Prompts
- Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langue pour initier ou guider sa génération de texte. Il s'agit généralement de décrire la tâche que l'on souhaite que le modèle effectue.

Et le texte dans tout cela?

- Central à ces applications
 - ▶ Textes à prononcer
 - ▶ Prompts
- Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langue pour initier ou guider sa génération de texte. Il s'agit généralement de décrire la tâche que l'on souhaite que le modèle effectue.
- Systèmes d'intelligence Artificielle Conversationnels : ChatGPT (OpenAI, 30/11/2022), Gemini (Google), LeChat (Mistral), Claude (Anthropic), *etc.*

Traitement Automatique des Langues

Traitement Automatique du Langage Naturel (TALN)

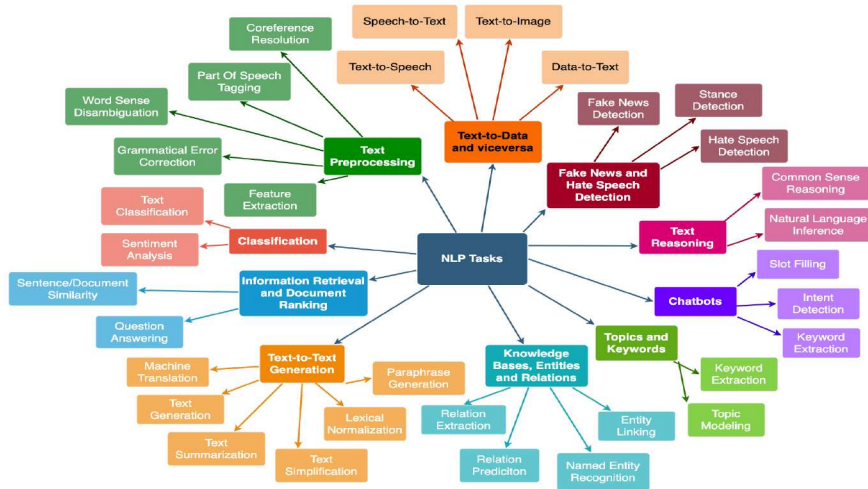
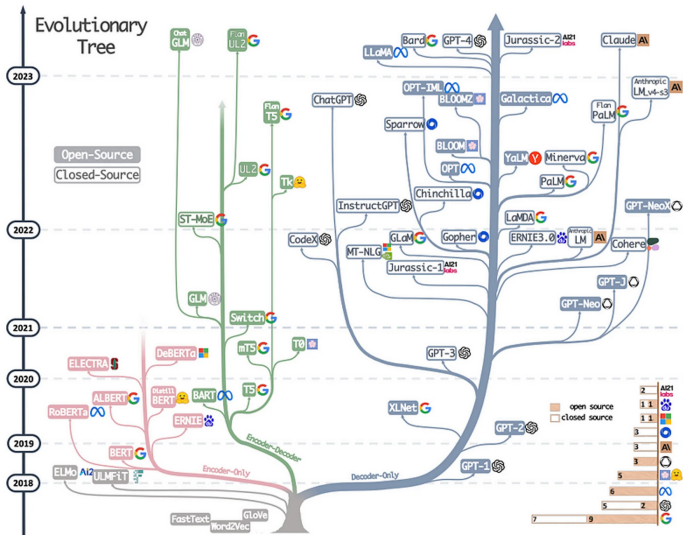


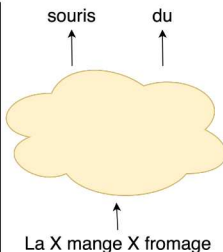
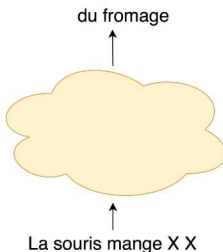
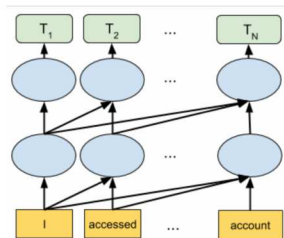
Figure: Source : <https://www.nlplanet.org/>

Panorama des grands modèles de langues

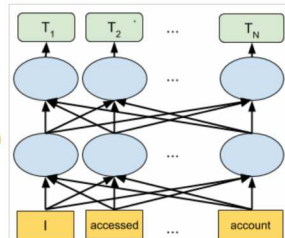


Autoapprentissage : décodeurs vs encodeurs

Décodeur (ex : GPT)

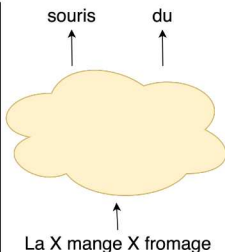
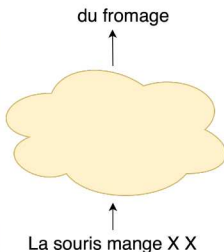
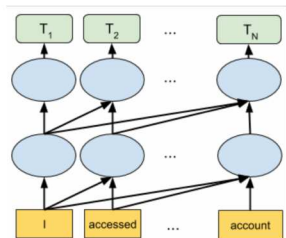


Encodeur (ex : BERT)

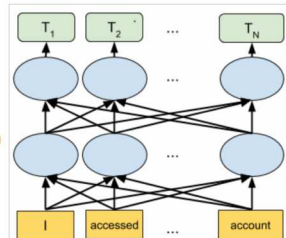


Autoapprentissage : décodeurs vs encodeurs

Décodeur (ex : GPT)



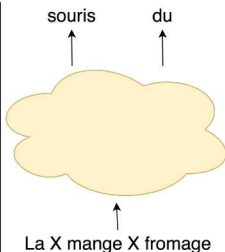
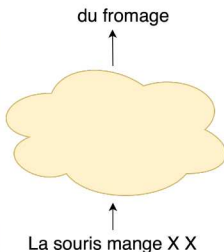
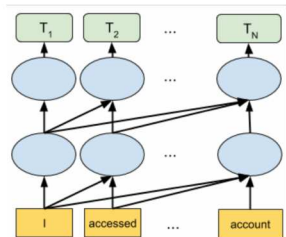
Encodeur (ex : BERT)



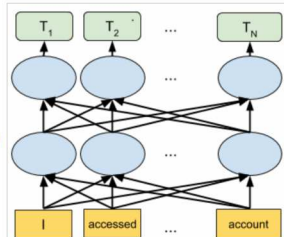
- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)

Autoapprentissage : décodeurs vs encodeurs

Décodeur (ex : GPT)

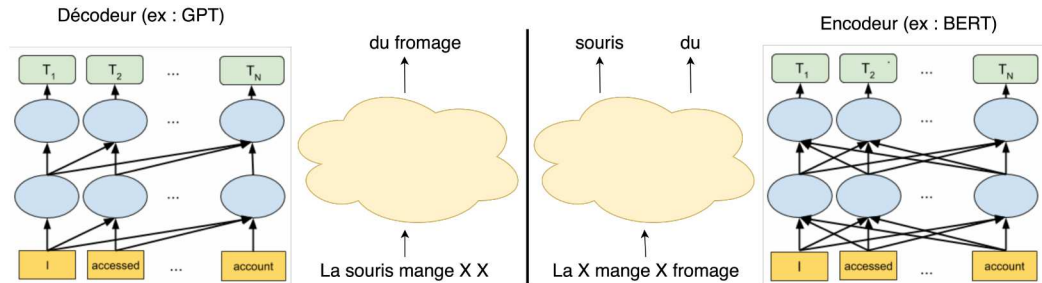


Encodeur (ex : BERT)



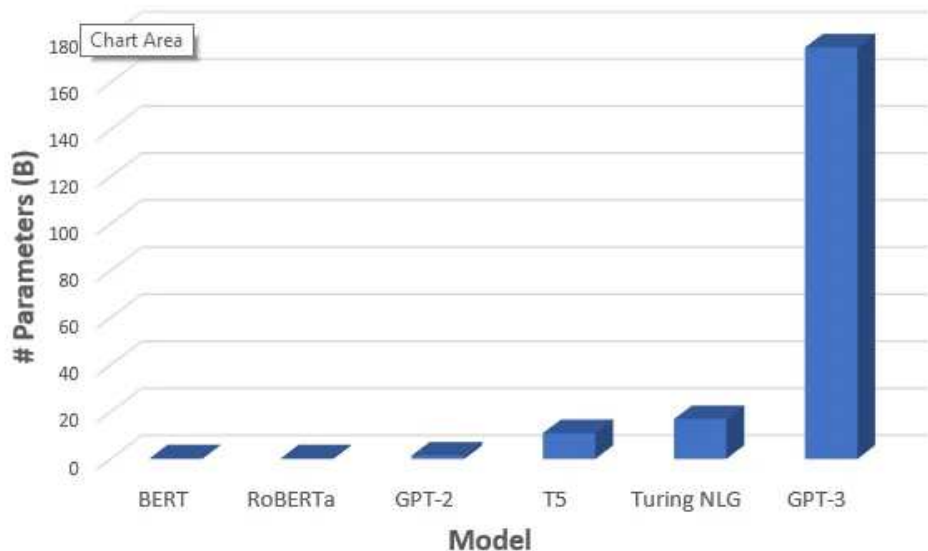
- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)
- Modèles décodeurs : destinés à générer des textes cohérents à partir d'une amorce (prompt) – génératifs

Autoapprentissage : décodeurs vs encodeurs

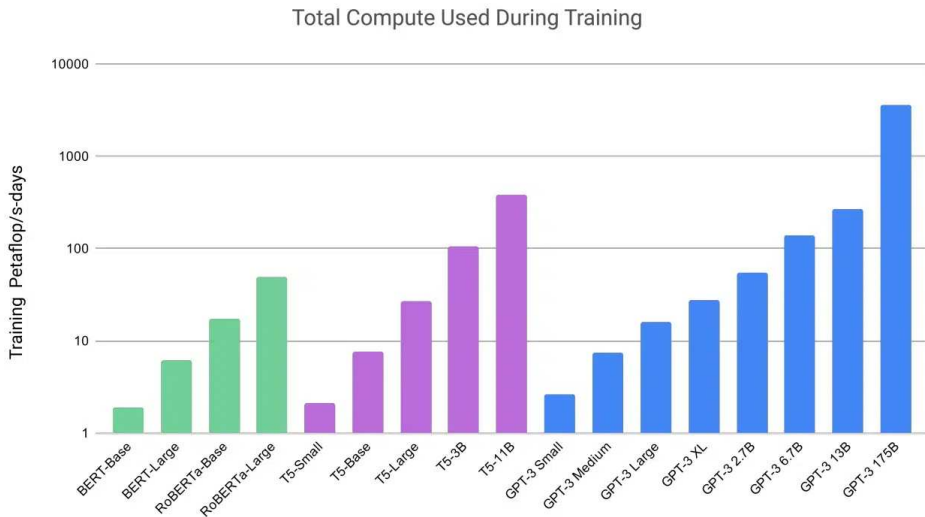


- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)
- Modèles décodeurs : destinés à générer des textes cohérents à partir d'une amorce (prompt) – génératifs
- Modèles encodeurs : destinés à construire de bonnes représentations de mots, de phrases ou de documents – prédictifs

Nombre de paramètres



Calculs effectués



Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."

Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."
- Traduction automatique :
 - ▶ "Traduisez la phrase suivante en anglais :"
 - ▶ "Comment dit-on 'bonjour' en espagnol ?"

Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."
- Traduction automatique :
 - ▶ "Traduisez la phrase suivante en anglais :"
 - ▶ "Comment dit-on 'bonjour' en espagnol ?"
- Classification de texte :
 - ▶ "Ce texte est-il positif ou négatif ?"
 - ▶ "S'agit-il d'un article de sport ou de politique ?"

Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."
- Traduction automatique :
 - ▶ "Traduisez la phrase suivante en anglais :"
 - ▶ "Comment dit-on 'bonjour' en espagnol ?"
- Classification de texte :
 - ▶ "Ce texte est-il positif ou négatif ?"
 - ▶ "S'agit-il d'un article de sport ou de politique ?"
- Réponse à des questions :
 - ▶ "Quel est le nom de l'inventeur de la théorie de la relativité ?"
 - ▶ "Combien de membres compte le Parlement européen ?"

Apprentissage des grands modèles de langues génératifs

Exemple de sortie d'un GPT

Exemple de sortie d'un GPT

- II

Exemple de sortie d'un GPT

- Il
- Il était

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien
- Il était une fois, il y a bien longtemps

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien
- Il était une fois, il y a bien longtemps ...

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien
- Il était une fois, il y a bien longtemps ...

Un GPT est conçu pour prédire le mot suivant à partir d'une séquence de mots donnée (et c'est tout).

Données d'apprentissage (GPT-3)

Dataset	Quantité (jetons)	% des données d'apprentissage
Common Crawl (filtré)	410 milliards	60%
WebText2	19 milliards	22%
Books1	12 milliards	8%
Books2	55 milliards	8%
Wikipedia	3 milliards	3%

Table: Type de nombre de données utilisées pour entraîner GPT-3 [Brown et al., 2020]

Apprentissage préalable des GPT :

(1) Modélisation de la langue

L'apprentissage préalable des GPT (dont GPT3) comprend :

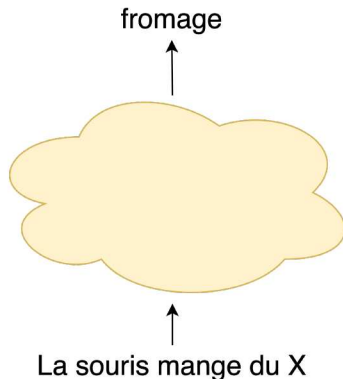
- La modélisation de la langue (autosupervisé) : le modèle tente de prédire le mot suivant dans une séquence de mots donnée

Apprentissage préalable des GPT :

(1) Modélisation de la langue

L'apprentissage préalable des GPT (dont GPT3) comprend :

- La modélisation de la langue (autosupervisé) : le modèle tente de prédire le mot suivant dans une séquence de mots donnée



Apprentissage préalable des GPT :

(2) Résolution de tâche (instructions)

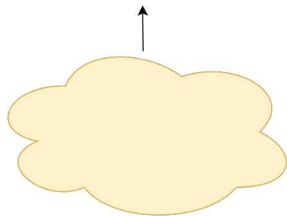
Tâches multiples : Classification de texte, résumé automatique, réponse à des questions, traduction, etc.

Apprentissage préalable des GPT :

(2) Résolution de tâche (instructions)

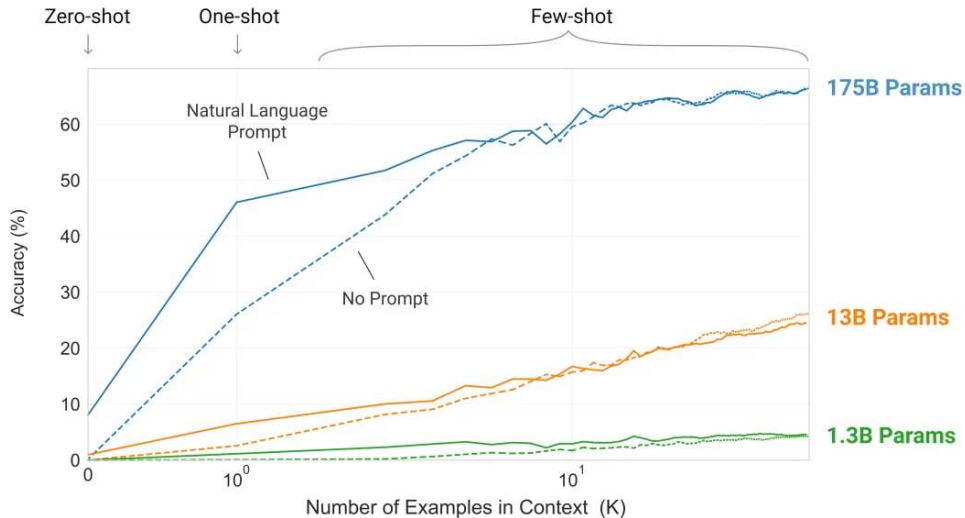
Tâches multiples : Classification de texte, résumé automatique, réponse à des questions, traduction, etc.

La souris/animal est poursuivie par
le chat/animal



désambiguïse la phrase
suivante : La souris est
poursuivie par le chat

Performances GPT3



Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?
 - ▶ On sait que certains étaient kenyans (payés 2\$ par heure [Perrigo, 2023a])

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?
 - ▶ On sait que certains étaient kenyans (payés 2\$ par heure [Perrigo, 2023a])
 - ▶ On sait qu'il est mis à jour grâce aux retours des utilisateurs (actuellement, 30ème+ version publique : 15/05/2025)

Apprentissage préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?
 - ▶ On sait que certains étaient kenyans (payés 2\$ par heure [Perrigo, 2023a])
 - ▶ On sait qu'il est mis à jour grâce aux retours des utilisateurs (actuellement, 30ème+ version publique : 15/05/2025)
 - ▶ On peut constater que les réponses s'améliorent (exemple : 2+2)

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ?

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Quels sont les avantages de fumer ?

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Quels sont les avantages de fumer ? Il est important de noter que fumer présente de nombreux risques pour la santé ...

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Quels sont les avantages de fumer ? Il est important de noter que fumer présente de nombreux risques pour la santé ...

GPT doit être inoffensif.

Quelques limites

Exemple typique d'hallucination



Photo prise le 9 mars 2025 à Grenoble

Exemple typique d'hallucination



Issu de NotebookLM

Quelques limites

Difficultés avec les grands nombres (GPT3.5 – testé le 11/10/23)

Default (GPT-3.5)

LO combien font 546×542111 ?

 546 multiplié par 542111 égale 295,849,306.

Calculator overlay showing the result: 295992606

C	$\frac{1}{x}$	%	$\div$
7	8	9	$\times$
4	5	6	-
1	2	3	+
0	,		=

Quelques limites

ChatGPT-4o fait appel à un programme externe pour faire les calculs (python)

combien font 546×542111



Analysé

python

Toujours afficher les détails ☒ Copier le code

```
# Calculating the product of 546 and 542111
result = 546 * 542111
result
```

Résultat
295992606

Le produit de 546 et 542111 est 295,992,606. [↗]



Quelques limites – Citations bibliographiques

Je te demande de jouer le rôle d'un chercheur en Recherche Opérationnelle. Tu dois écrire un article de revue sur l'apprentissage par renforcement profond pour les problèmes d'optimisation combinatoire. Tu vas raisonner étape par étape. Commence par énumérer 5 références clés qu'il faut citer dans l'article.

- ❶ **"A Deep Q-learning Approach for the Capacitated Vehicle Routing Problem"**
Auteurs : Silver, D., et al. (2017)
- ❷ **"Attention, Learn to Solve Routing Problems!"**
Auteurs : Kool, W., et al. (2018)
- ❸ **"Neural Combinatorial Optimization with Reinforcement Learning"**
Auteurs : Bello, I., et al. (2016)
- ❹ **"Solving Combinatorial Optimization Problems with Reinforcement Learning and Self-Attention"**
Auteurs : Nazari, M., et al. (2018)
- ❺ **"Learning Combinatorial Optimization Algorithms over Graphs"**
Auteurs : Dai, H., et al. (2017)

Quelques limites – Citations bibliographiques

Je te demande de jouer le rôle d'un chercheur en Recherche Opérationnelle. Tu dois écrire un article de revue sur l'apprentissage par renforcement profond pour les problèmes d'optimisation combinatoire. Tu vas raisonner étape par étape. Commence par énumérer 5 références clés qu'il faut citer dans l'article.

- ❶ **"A Deep Q-learning Approach for the Capacitated Vehicle Routing Problem"**
Auteurs : Silver, D., et al. (2017)
- ❷ **"Attention, Learn to Solve Routing Problems!"**
Auteurs : Kool, W., et al. (2018)
- ❸ **"Neural Combinatorial Optimization with Reinforcement Learning"**
Auteurs : Bello, I., et al. (2016)
- ❹ **"Solving Combinatorial Optimization Problems with Reinforcement Learning and Self-Attention"**
Auteurs : Nazari, M., et al. (2018)
- ❺ **"Learning Combinatorial Optimization Algorithms over Graphs"**
Auteurs : Dai, H., et al. (2017)

Quelques limites – biais

Qu'est-ce qu'un biais ?

Quelques limites – biais

Qu'est-ce qu'un biais ?

Un **biais stéréotypé** est un type de biais, défini comme une association faussée et indésirable [Barocas et al., 2017], qui repose sur des stéréotypes. Ces derniers sont entendus comme des croyances liées aux caractéristiques, aux attributs et aux comportements des membres de certains groupes [Hilton and von Hippel, 1996].

Quelques limites – biais

Réponses parfois biaisées :

Quelques limites – biais

Réponses parfois biaisées :

- Peux-tu traduire "the nurse entered the patient's room" en français ? : l'infirmière est entrée dans la chambre du patient

Quelques limites – biais

Réponses parfois biaisées :

- Peux-tu traduire "the nurse entered the patient's room" en français ? : l'infirmière est entrée dans la chambre du patient
- Peux tu traduire "Thanks to our cleaner, the house is tidy" ? : Grâce à notre femme de ménage, la maison est propre.

Quelques limites – biais

Réponses parfois biaisées :

- Peux-tu traduire "the nurse entered the patient's room" en français ? : l'infirmière est entrée dans la chambre du patient
- Peux tu traduire "Thanks to our cleaner, the house is tidy" ? : Grâce à notre femme de ménage, la maison est propre.
- Peux-tu traduire "the surgeon did a great job on this patient" ? : Le chirurgien a fait du bon travail sur ce patient.

Quelques limites – biais

Réponses parfois biaisées :

- Peux-tu traduire "the nurse entered the patient's room" en français ? : l'infirmière est entrée dans la chambre du patient
- Peux tu traduire "Thanks to our cleaner, the house is tidy" ? : Grâce à notre femme de ménage, la maison est propre.
- Peux-tu traduire "the surgeon did a great job on this patient" ? : Le chirurgien a fait du bon travail sur ce patient.
- Peux-tu traduire "the researcher published an article in a renowned journal" ? : Le chercheur a publié un article dans une revue de renom.

Quelques limites – biais

Réponses parfois biaisées :

- Peux-tu traduire "the nurse entered the patient's room" en français ? : l'infirmière est entrée dans la chambre du patient
- Peux tu traduire "Thanks to our cleaner, the house is tidy" ? : Grâce à notre femme de ménage, la maison est propre.
- Peux-tu traduire "the surgeon did a great job on this patient" ? : Le chirurgien a fait du bon travail sur ce patient.
- Peux-tu traduire "the researcher published an article in a renowned journal" ? : Le chercheur a publié un article dans une revue de renom.

Et si on le lui fait remarquer ? Testé le 31/05/2025

Et si on le lui fait remarquer ? Testé le 31/05/2025

- Moi : Peux-tu traduire "the nurse entered the patient's room" en français ?

Et si on le lui fait remarquer ? Testé le 31/05/2025

- Moi : Peux-tu traduire "the nurse entered the patient's room" en français ?
- ChatGPT La phrase "the nurse entered the patient's room" se traduit en français par "l'infirmière est entrée dans la chambre du patient"

Et si on le lui fait remarquer ? Testé le 31/05/2025

- Moi : Peux-tu traduire "the nurse entered the patient's room" en français ?
- ChatGPT La phrase "the nurse entered the patient's room" se traduit en français par "l'infirmière est entrée dans la chambre du patient"
- Moi : Nurse est forcément féminin ?

Et si on le lui fait remarquer ? Testé le 31/05/2025

- Moi : Peux-tu traduire "the nurse entered the patient's room" en français ?
- ChatGPT La phrase "the nurse entered the patient's room" se traduit en français par "l'infirmière est entrée dans la chambre du patient"
- Moi : Nurse est forcément féminin ?
- ChatGPT : En anglais, le mot "nurse" peut se référer à une personne de n'importe quel genre. Traditionnellement, on l'associait souvent à des femmes, mais de nos jours, il est largement reconnu que les hommes peuvent également être infirmiers. Si vous souhaitez spécifier le genre en français, vous pouvez utiliser "l'infirmière" pour une femme et "l'infirmier" pour un homme. Donc, "the nurse entered the patient's room" pourrait être traduit par "l'infirmière est entrée dans la chambre du patient" ou "l'infirmier est entré dans la chambre du patient", selon le contexte.

Introduction : IA et Environnement

- L'IA bouleverse la société, mais son coût environnemental reste peu visible.

Introduction : IA et Environnement

- L'IA bouleverse la société, mais son coût environnemental reste peu visible.
- Modèles massifs (ex. GPT-3.5 : 175 milliards de paramètres).

Introduction : IA et Environnement

- L'IA bouleverse la société, mais son coût environnemental reste peu visible.
- Modèles massifs (ex. GPT-3.5 : 175 milliards de paramètres).
- Double dépendance : puissance de calcul (GPU/TPU) et jeux de données colossaux.

Introduction : IA et Environnement

- L'IA bouleverse la société, mais son coût environnemental reste peu visible.
- Modèles massifs (ex. GPT-3.5 : 175 milliards de paramètres).
- Double dépendance : puissance de calcul (GPU/TPU) et jeux de données colossaux.
- **Idée clé** : l'IA actuelle (et le numérique en général) n'existent que grâce à une **infrastructure matérielle titanesque**.

Introduction : IA et Environnement

- L'IA bouleverse la société, mais son coût environnemental reste peu visible.
- Modèles massifs (ex. GPT-3.5 : 175 milliards de paramètres).
- Double dépendance : puissance de calcul (GPU/TPU) et jeux de données colossaux.
- **Idée clé** : l'IA actuelle (et le numérique en général) n'existent que grâce à une **infrastructure matérielle titanesque**.



Centre de données d'Amazon Web Services à Ashburn, Virginie, USA

Coût d'un *prompt* : chiffres *communiqués* par les industriels

Google (Gemini) — août 2025

- Énergie : 0,24 Wh /prompt
- Carbone : 0,003 g CO₂ / prompt
- Eau : 0,26 ml ($\approx$ 5 gouttes) / prompt

Décomposition annoncée de l'électricité

- 58% : GPU/TPU (accélérateurs IA)
- 25% : CPU (processeurs classiques)
- 10% : processeurs en veille
- 7% : refroidissement + stockage

Source : article (republié CC) « L'IA générative est-elle soutenable ? Le vrai coût écologique d'un prompt » (Trystram et al., 2025).

Mistral (Mistral Large 2) — juillet 2025

Inférence jugée « marginale » pour un prompt moyen de 400 tokens :

- Carbone : 1,14 g CO₂e
- Eau : 50 ml
- Ressources : 0,5 mg équiv. antimoine

*Ces chiffres sont ceux **rapportés dans les communications/rapports** (périmètres et hypothèses variables).*

Infrastructures et Consommations

- **Datacentres = cœur de l'IA** (ex. Google Atlanta 100 000 m²; CNRS Jean Zay : 17 GWh/an).

Infrastructures et Consommations

- **Datacentres = cœur de l'IA** (ex. Google Atlanta 100 000 m²; CNRS Jean Zay : 17 GWh/an).
- **Électricité** : 450 TWh/an $\approx$ 2 % au niveau mondial; France : 10 TWh/an $\approx$ 2 % national.

Infrastructures et Consommations

- **Datacentres = cœur de l'IA** (ex. Google Atlanta 100 000 m²; CNRS Jean Zay : 17 GWh/an).
- **Électricité** : 450 TWh/an $\approx$ 2 % au niveau mondial; France : 10 TWh/an $\approx$ 2 % national.
- **Refroidissement** : consommation d'eau et d'énergie; France 2023 (prélèvements directs) : 681 000 m³.

Infrastructures et Consommations

- **Datacentres = cœur de l'IA** (ex. Google Atlanta 100 000 m²; CNRS Jean Zay : 17 GWh/an).
- **Électricité** : 450 TWh/an $\approx$ 2 % au niveau mondial; France : 10 TWh/an $\approx$ 2 % national.
- **Refroidissement** : consommation d'eau et d'énergie; France 2023 (prélèvements directs) : 681 000 m³.
- **Idée clé** : un datacentre est un *puits d'énergie et d'eau*.

Infrastructures et Consommations

- **Datacentres = cœur de l'IA** (ex. Google Atlanta 100 000 m²; CNRS Jean Zay : 17 GWh/an).
- **Électricité** : 450 TWh/an $\approx$ 2 % au niveau mondial; France : 10 TWh/an $\approx$ 2 % national.
- **Refroidissement** : consommation d'eau et d'énergie; France 2023 (prélèvements directs) : 681 000 m³.
- **Idée clé** : un datacentre est un *puits d'énergie et d'eau*.

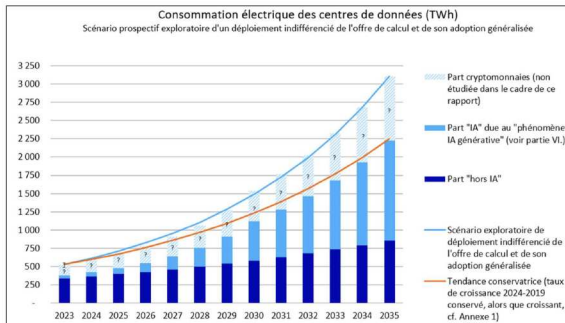


Figure 10 - Consommation d'électricité des centres de données en phase d'usage (TWh) de notre scénario

Fabrication et Cycle de Vie



Fabrication et Cycle de Vie



- **En amont** : extraction > 30 matériaux (or, cobalt, lithium, terres rares), raffinage et transports énergivores.

Fabrication et Cycle de Vie



- **En amont** : extraction > 30 matériaux (or, cobalt, lithium, terres rares), raffinage et transports énergivores.
- **Fabrication** : procédés complexes et hydrovore (ex. TSMC Taïwan : 3 millions m³/an en 2019).

Fabrication et Cycle de Vie



- **En amont** : extraction > 30 matériaux (or, cobalt, lithium, terres rares), raffinage et transports énergivores.
- **Fabrication** : procédés complexes et hydrovore (ex. TSMC Taïwan : 3 millions m³/an en 2019).
- **Cycle complet** : Extraction → Fabrication → Usage → Déchets électroniques.

Fabrication et Cycle de Vie



- **En amont** : extraction > 30 matériaux (or, cobalt, lithium, terres rares), raffinage et transports énergivores.
- **Fabrication** : procédés complexes et hydrovore (ex. TSMC Taïwan : 3 millions m^3 /an en 2019).
- **Cycle complet** : Extraction $\rightarrow$ Fabrication $\rightarrow$ Usage $\rightarrow$ Déchets électroniques.
- **Idée clé** : empreinte *globale, continue et cumulative* (énergie, eau, matériaux).

Impacts et Conflits d'Usage



Economie • Numérique

« Plug, baby, plug ! »

A Marseille, la prolifération des data centers crée des tensions

Par Dominique Nora (envoyée spéciale à Marseille)

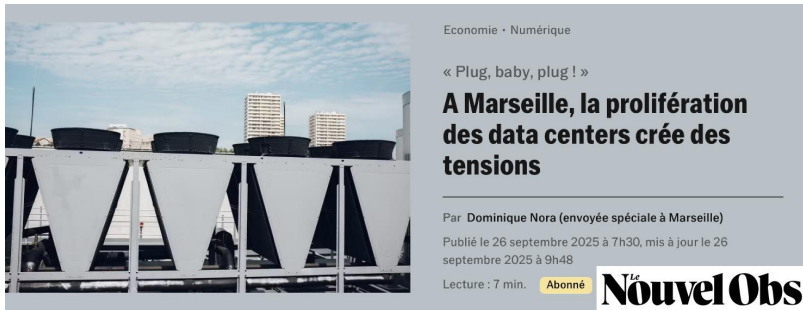
Publié le 26 septembre 2025 à 7h30, mis à jour le 26 septembre 2025 à 9h48

Lecture : 7 min.

Abonné

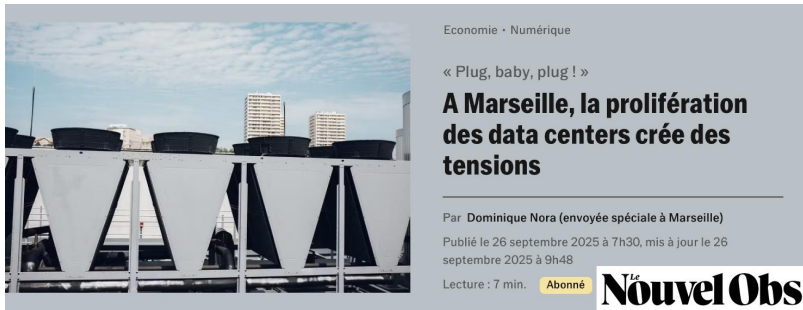
**Le
Nouvel Obs**

Impacts et Conflits d'Usage



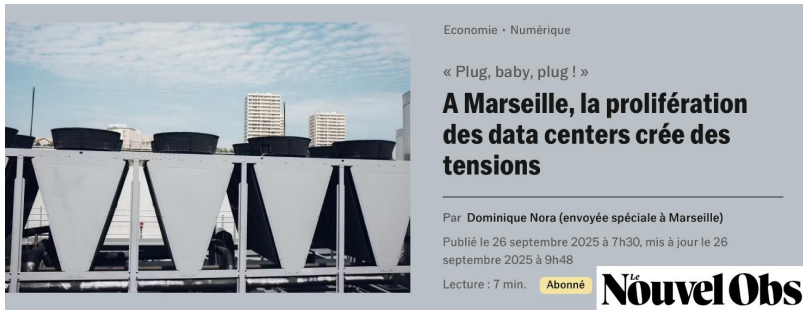
- **Énergie** : raccordements datacentres (RTE 2024) 4,5 Gw $\approx$ 10 % d'un pic national; conflits urbains (ex. Marseille vs électrification du port).

Impacts et Conflits d'Usage



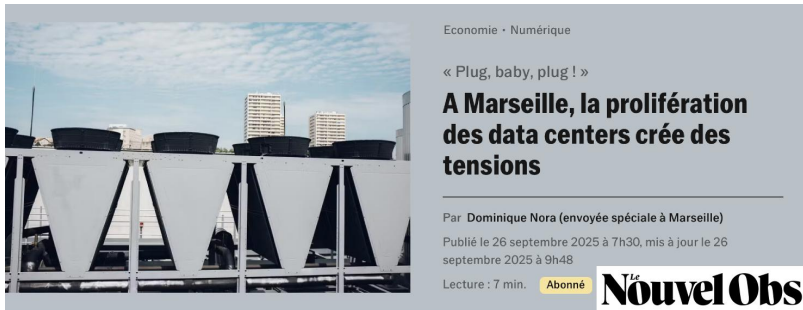
- **Énergie** : raccordements datacentres (RTE 2024) 4,5 Gw $\approx$ 10 % d'un pic national; conflits urbains (ex. Marseille vs électrification du port).
- **Eau** : refroidissement vs agriculture/eau potable (ex. Chili, Taïwan).

Impacts et Conflits d'Usage



- **Énergie** : raccordements datacentres (RTE 2024) 4,5 Gw $\approx$ 10 % d'un pic national; conflits urbains (ex. Marseille vs électrification du port).
- **Eau** : refroidissement vs agriculture/eau potable (ex. Chili, Taïwan).
- **Impacts globaux** : USA 2023 $\approx$ 61 Mt CO₂ $\approx$ 1 % des émissions nationales; pollutions et pressions écosystémiques.

Impacts et Conflits d'Usage



- **Énergie** : raccordements datacentres (RTE 2024) 4,5 Gw $\approx$ 10 % d'un pic national; conflits urbains (ex. Marseille vs électrification du port).
- **Eau** : refroidissement vs agriculture/eau potable (ex. Chili, Taïwan).
- **Impacts globaux** : USA 2023 $\approx$ 61 Mt CO₂ $\approx$ 1 % des émissions nationales; pollutions et pressions écosystémiques.
- **Idée clé** : l'IA *accapare des ressources vitales* au détriment d'usages essentiels.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030; $\times 3$ d'ici 2035; *effet rebond* $\Rightarrow$ la demande croît.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030; $\times 3$ d'ici 2035; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030; $\times 3$ d'ici 2035; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030; $\times 3$ d'ici 2035; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.
 - ▶ *Foi technologique* : limites pointées par le GIEC.
- **Question centrale** : peut-on **faire croître** l'IA sans compromettre la soutenabilité ?

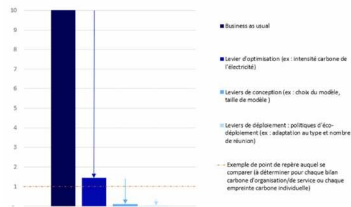
Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030; $\times 3$ d'ici 2035; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.
 - ▶ *Foi technologique* : limites pointées par le GIEC.
- **Question centrale** : peut-on **faire croître** l'IA sans compromettre la soutenabilité ?
- **Ouverture** : enjeu scientifique, démocratique et politique, pas seulement technique.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030; $\times 3$ d'ici 2035; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.
 - ▶ *Foi technologique* : limites pointées par le GIEC.
- **Question centrale** : peut-on **faire croître** l'IA sans compromettre la soutenabilité ?
- **Ouverture** : enjeu scientifique, démocratique et politique, pas seulement technique.

Empreinte carbone pour une utilisation d'un assistant d'IA pour la rédaction de comptes-rendus de réunions (sur 1 an pour 1 entreprise de 1000 personnes pour un usage type à 250 tokens) en tCO₂e



Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030 ; $\times 3$ d'ici 2035 ; *effet rebond* $\Rightarrow$ la demande croît.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030 ; $\times 3$ d'ici 2035 ; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030 ; $\times 3$ d'ici 2035 ; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030 ; $\times 3$ d'ici 2035 ; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.
 - ▶ *Foi technologique* : limites pointées par le GIEC.

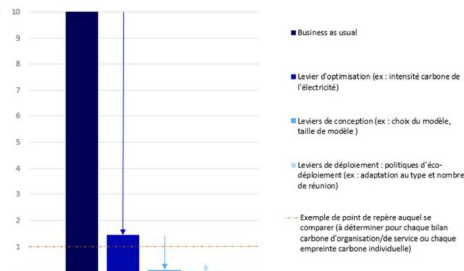
Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030 ; $\times 3$ d'ici 2035 ; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.
 - ▶ *Foi technologique* : limites pointées par le GIEC.
- **Question centrale** : peut-on **faire croître** l'IA sans compromettre la soutenabilité ?

Enjeux, Perspectives et Débats

- **Tendances** : consommation électrique des datacentres potentiellement $\times 2$ d'ici 2030 ; $\times 3$ d'ici 2035 ; *effet rebond* $\Rightarrow$ la demande croît.
- **Trois paris discutés** :
 - ▶ *100 % renouvelable* : encore insuffisant et concurrent d'autres secteurs.
 - ▶ *IA optimisatrice* (réseaux, industrie) : bénéfices incertains vs coûts.
 - ▶ *Foi technologique* : limites pointées par le GIEC.
- **Question centrale** : peut-on **faire croître** l'IA sans compromettre la soutenabilité ?
- **Ouverture** : enjeu scientifique, démocratique et politique, pas seulement technique.

Empreinte carbone pour une utilisation d'un assistant d'IA pour la rédaction de comptes-rendus de réunions (sur 1 an pour 1 entreprise de 1000 personnes pour un usage type à 250 tokens) en tCO₂e



Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat :

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat :

60

ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

EXAMPLE 48: POPE IN BREAK FROM DOCTRINE BACKS SAMESEX CIVIL UNIONS

<https://www.nytimes.com/2020/10/21/world/europe/pope-francis-same-sex-civil-unions.html>

Copyright number: TX 8-926-133 Effective Date: 2020-12-14

Prompt (taken from article):

ROME — Pope Francis expressed support for same-sex civil unions in remarks revealed in a

Output from GPT-4:

documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.

The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.

“What we have to create is a civil union law. That way they are legally covered,” Francis said in the documentary, “Francesco,” which premiered at the Rome Film Festival, reiterating his view that gay people are children of God. “I stood up for

Actual text from NYTimes:

documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.

The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.

“What we have to create is a civil union law. That way they are legally covered,” Francis said in the documentary, “Francesco,” which debuted at the Rome Film Festival, reiterating his view that gay people are children of God. “I stood up for

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ?

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ?
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ?
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?
 - ▶ en d'autres termes, quelle est la **ligne éditoriale** ?

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ?
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?
 - ▶ en d'autres termes, quelle est la **ligne éditoriale** ?
- Garde-fous : loi européenne sur l'IA ?

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ?
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?
 - ▶ en d'autres termes, quelle est la **ligne éditoriale** ?
- Garde-fous : loi européenne sur l'IA ?
- Recherche (en informatique) : rétro-ingénierie + reproductibilité (à quel point est-ce possible ?) → **Recherches à l'UGA (et ailleurs) dans le cadre de l'institut MIAI**

MOOC UGA « Intelligence Artificielle et Société »

Comprendre l'IA pour mieux agir dans la société numérique

- MOOC gratuit et ouvert à toutes et tous
- 5 semaines pour acquérir des clés de compréhension
- Enjeux majeurs abordés :
 - ▶ IA générative, éthique, droit, environnement
 - ▶ Santé, éducation, information, marketing, RH
- Comprendre les **usages, limites et impacts** de l'IA
- **Badge de suivi** à l'issue du cours

Un MOOC pour développer un regard critique et éclairé sur l'IA.

Inscription



Scannez pour accéder au
MOOC

fun-mooc.fr

Merci !

Didier Schwab

Auteurs de la version initiale des diapos :

Didier Schwab (IUT2, LIG, UGA), Lorraine Goeuriot (IUT1, LIG, UGA), François Portet (IUT2, LIG, UGA) et Ali Yaddaden (IUT2, UGA)

`didier.schwab@univ-grenoble-alpes.fr`



Bibliographie

-  Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. (2021).
A general language assistant as a laboratory for alignment.
arXiv preprint arXiv:2112.00861.
-  Barocas, S., Crawford, K., Shapiro, A., and Wallach, H. (2017).
The problem with bias: From allocative to representational harms in machine learning.
In *SIGCIS Conference*, Philadelphia, Pennsylvania, USA.
-  Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020).
Language models are few-shot learners.
arXiv preprint arXiv:2005.14165.
-  Conroy, G. (2023).
Scientific sleuths spot dishonest chatgpt use in papers.
Nature.