

# BackProp Jackie Boscher

- [jboscher@gmail.com](mailto:jboscher@gmail.com)
- [ia@backprop.fr](mailto:ia@backprop.fr)
- <https://www.linkedin.com/in/jboscher/>
- @j\_boscher
- <https://formation.backprop.fr/>



## Groups you organize

### Generative AI Nice

👤 149 members

Create event



### Meetup Machine Learning Rennes

👤 1548 members

Create event



### TensorFlow User Group Rennes - IA par le Code

👤 550 members

Create event



# Macron annonce un nouveau fonds d'investissement dans l'IA

- Dessinant un nouveau plan d'action pour ne pas rater ce virage en cours, Emmanuel [Macron](#) appelle à un «réveil du financement européen sur ces technologies». La France organisera un sommet mondial les 10 et 11 février 2025.



Peux-tu me faire un plan pour une présentation d'une heure, sur l'intelligence générative, à destination d'un public n'ayant aucune connaissance sur le sujet ?



Bien sûr ! Voici un plan pour une présentation d'une heure sur l'intelligence générative destinée à un public novice :

### **1. Introduction (5 minutes)**

- Brève histoire de l'intelligence artificielle (IA)
- Définition de l'IA générative
- Pourquoi ce sujet est pertinent aujourd'hui

### **2. Qu'est-ce que l'IA générative ? (10 minutes)**

- Explication des modèles génératifs
- Différence entre l'IA discriminative et générative
- Exemples simples pour illustrer le concept (par exemple, la création d'images)

# ChatGPT et mon plan

---



Prototype shown

20

Bliss

---

**J'aimerais vous présenter  
le tout premier utilisateur**

**Le Parisien**

Ego Speed: 45.56 MPH  
time: 1545.441322000  
CAL P 0.60 Y 1.20 R 0.00 deg

Vision fps: 18.05 Draw fps: 17.67 Display fps: 21.34  
NL(0.00), E(0.93), F(0.07), TF(0.00), S(0.00)  
NRW: FLP(0.00), FRP(0.00)  
CufinExcited (Prb 0.56)

+0.0001 AUTO\_HIGH\_Beam  
+0.0000 BLINDED  
+0.0002 RAINING  
+0.0000 TIRE\_SPRAY  
+0.0013 WET\_ROAD  
**0.7902 RESTRICTED**  
0.0934 CONTROLLED\_ACCESS

L:0 R:0 F:2 ON:0

W:8.2 AP:1.0 I:0

VS: 46.7 MPH St: 1

merge: 1.0 1 160.2 R

AP

94 14

LC

4.0m

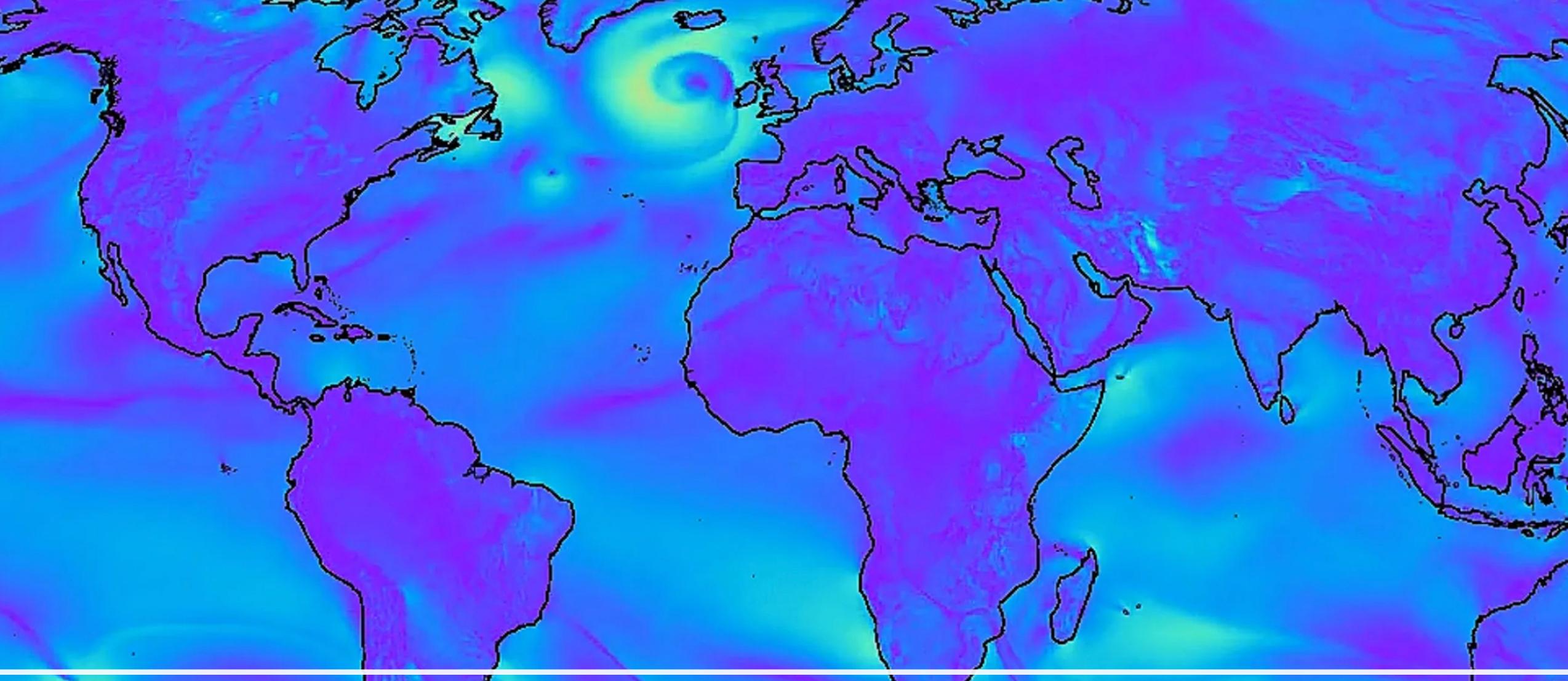
MAIN -





Line Crossing Count: 0

Entry=0



GraphCast, le modèle de Deepmind fournit des prévisions météorologiques sur 10 jours avec une précision sans précédent en moins d'une minute.

# Recherche médicale

- Grâce à l'IA, des chercheurs du MIT identifient une nouvelle classe d'antibiotiques candidats
- Ces composés peuvent tuer le *Staphylococcus aureus* résistant à la méthicilline (SARM), une bactérie responsable d'infections mortelles.





Une femme élégante marche dans une rue de Tokyo remplie de néons chaleureux et de panneaux urbains animés. Elle porte une veste en cuir noire, une longue robe rouge et des bottes noires et porte un sac à main noir. Elle porte des lunettes de soleil et du rouge à lèvres. Elle marche avec assurance et nonchalance. La rue est humide et réfléchissante, créant un effet miroir des lumières colorées. De nombreux piétons se promènent







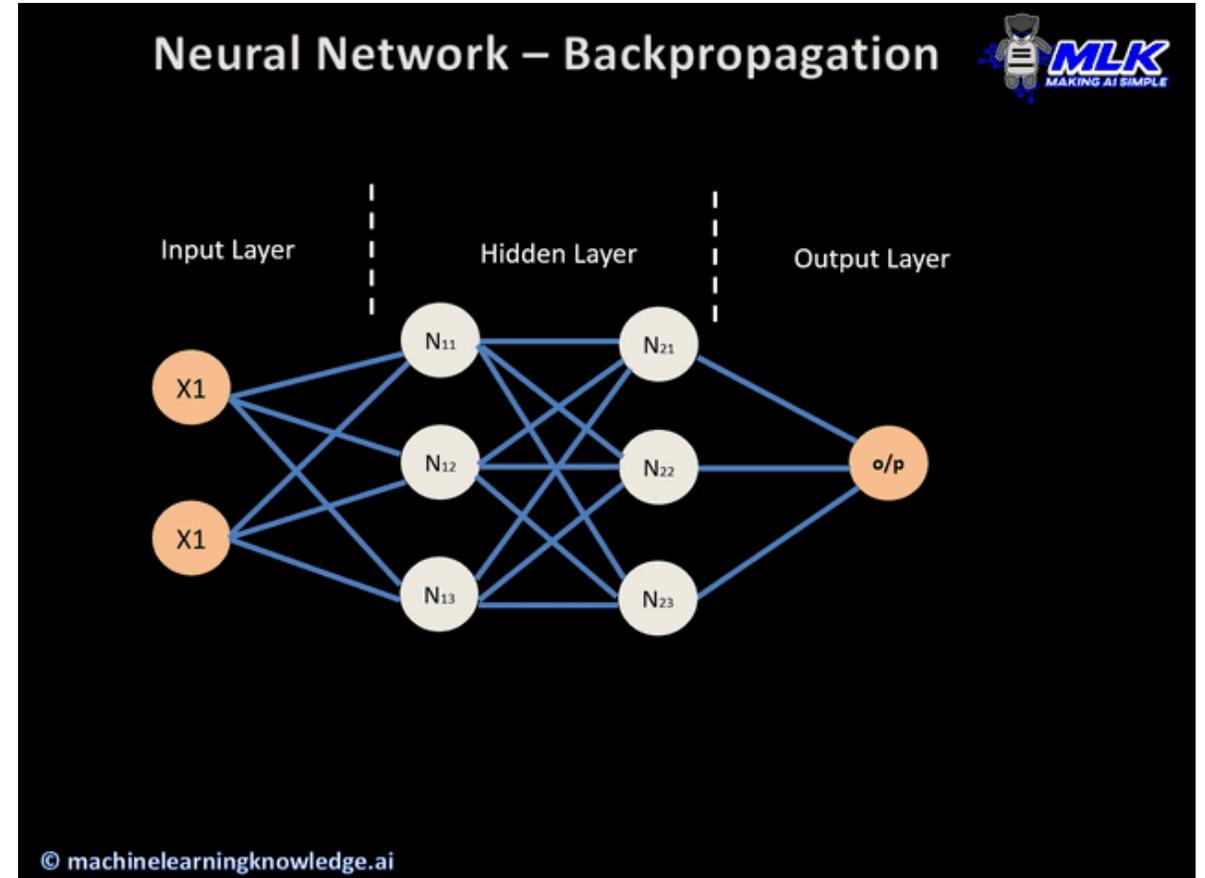


- Chihuahua ou muffin

# Apprentissage supervisé par Réseau de Neurones

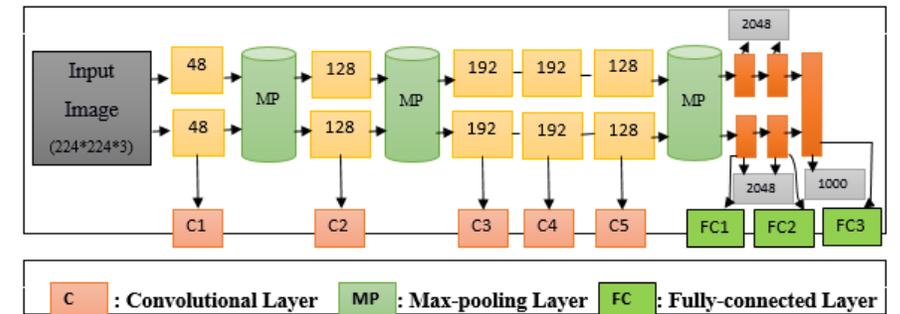
---

- La notion de paramètres



# AlexNet (2012)

- Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton
- 62 millions de paramètres
- 5 à 6 jours sur 2 GPUs



DeepMind –  
AlphaGo –  
2016





# Les acteurs

Elon Musk

# Les acteurs

---

- Open AI
- Y Combinator
- Sam Altman



# Les étapes importantes d'OpenAI



**We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs.**

**While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers.**

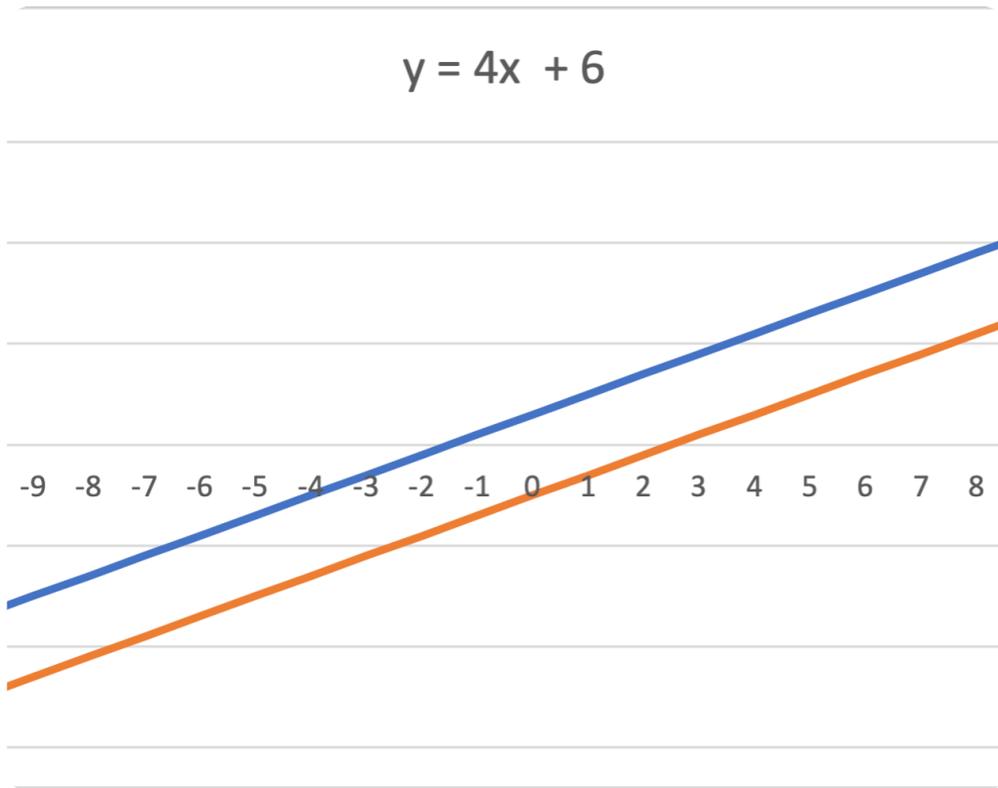
**GPT-4 is a Transformer- based model pre-trained to predict the next token in a document.**

**The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior.**

**A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales.**



# ChatGPT



- Exercice de classe de 2<sup>nde</sup>
- Une ligne parallèle à la droite d'équation :
  - $y = 4x + 6$
- passe par le point (5, 10).
- Quels sont ses paramètres ?

# Raisonnement

Caroline (une fille) a 3 frères. Chaque frère a 2 sœurs. Combien de sœurs Caroline a-t-elle ?

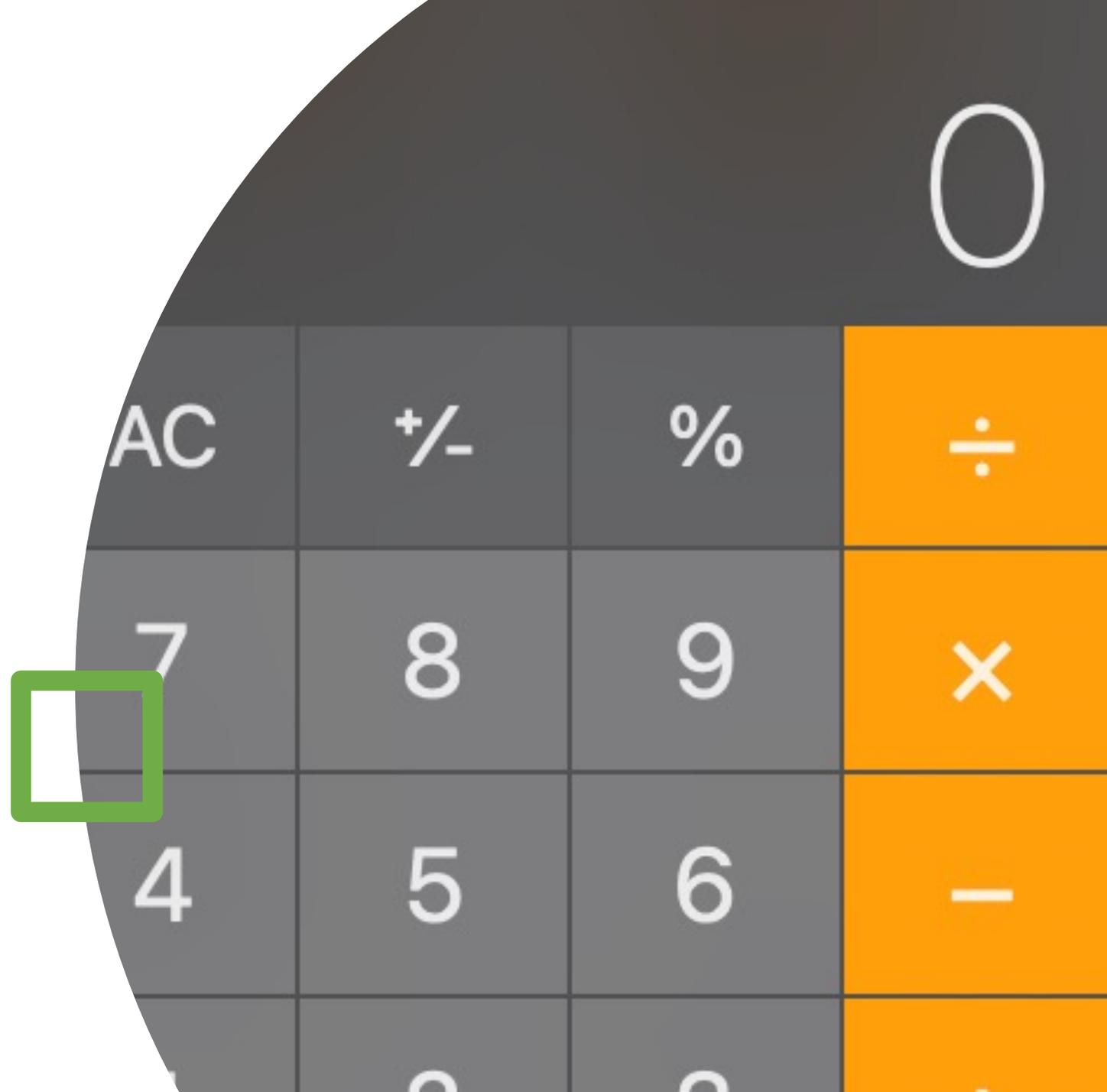


Nombre premiers  
entre eux

0.0



# Créer une calculatrice



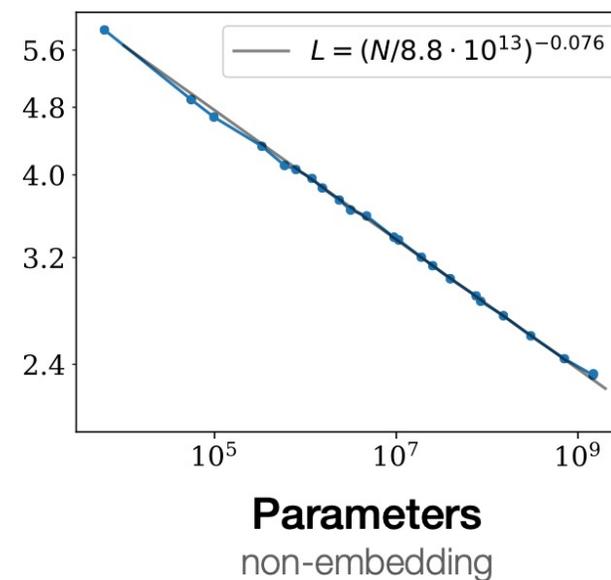
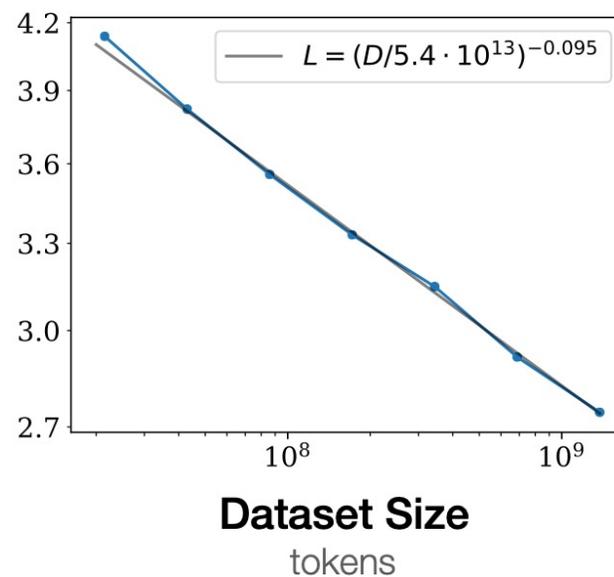
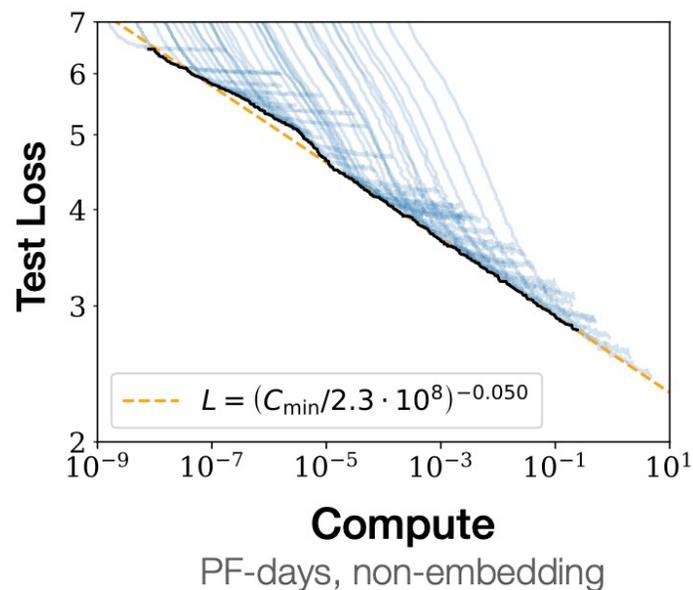
# Challenges

- Les données
- Les coûts
- Hallucinations
- Les comportements inappropriés
- Données obsolètes
- Difficulté des évaluations



# Les données

- Les performances des LLM s'améliorent progressivement à mesure que nous augmentons la taille du modèle, la taille de jeu de données et la quantité de calcul utilisée pour l'apprentissage. Pour des performances optimales, les trois facteurs doivent être développés en tandem.
- Les performances dépendent fortement de l'échelle, faiblement du modèle : plus précisément de trois facteurs : le nombre de paramètres du modèle  $N$ , la taille de l'ensemble de données  $D$  et la quantité de calcul  $C$  utilisée pour l'entraîner.



Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

[Training data mixture used in Meta's LLaMA model]

## Dataset pour LLaMA 2

Llama 2 a été pré-entraîné sur 2 000 milliards de tokens provenant de sources accessibles au public et plus de 1 million de nouveaux exemples annotés par l'homme.

GPT-4 utilise un ensemble de données plus diversifié et plus volumineux de 1 pétaoctet, tandis que GPT-3 utilise un ensemble de données plus petit de 45 téraoctets.

(nb : 1 To = 1000 Go, 1 Po = 1000 To)

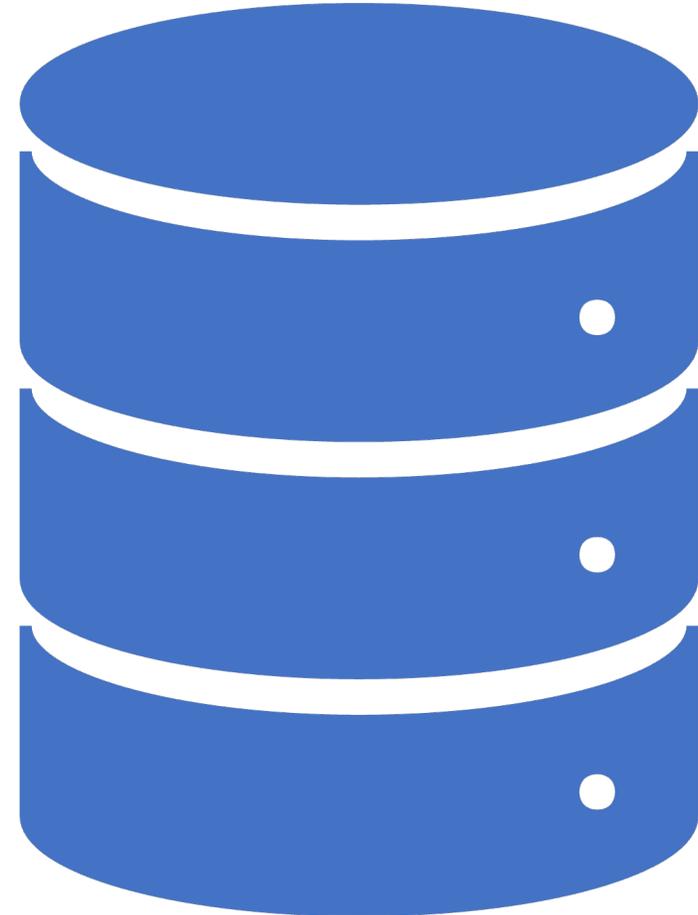
# Dataset pour Llama 3

---

- Conformément à nos principes de conception, nous avons investi massivement dans les données de pré-entraînement.
- Llama 3 est pré-entraîné sur plus de 15T de tokens qui ont tous été collectés à partir de sources accessibles au public.
- Notre ensemble de données de formation est sept fois plus volumineux que celui utilisé pour Llama 2 et comprend quatre fois plus de code.
- Pour préparer les prochains cas d'utilisation multilingues, plus de 5 % de l'ensemble de données de pré-entraînement de Llama 3 est constitué de données non anglaises de haute qualité couvrant plus de 30 langues.
- Cependant, nous n'attendons pas le même niveau de performance dans ces langues qu'en anglais.

Il est impossible de s'assurer de la qualité du dataset par un contrôle systématique et approfondi

# Datasets



# Les coûts

---

- Les coûts d'entraînement de GPT-4 seraient d'environ 63 millions de dollars, voire 100 millions selon Sam Altman





JE Which review papers discuss challenges and applications of large language models?

 As of my knowledge cutoff in September 2021, there were several influential papers discussing the challenges and applications of large language models. Here are some key resources that were available at that time:

1. "Language Models are Few-Shot Learners" by Tom B. Brown, Benjamin Mann, Nick Ryder, et al., 2020. This is not exactly a review paper, but it's a seminal work by OpenAI that discusses the potential and challenges of large-scale language models.  **Correct!**
2. "The Power and Limitations of Language Models" by Jacob Devlin, 2021. Although this is a blog post, it provides a good summary of the current understanding of the power and limitations of large language models, especially the Transformer-based models like BERT and GPT-3.  **Does not exist!**
3. "A Survey of Transformers" by Tianyu Liu, Xiang Li, Zhengyang Liu, et al., 2021.

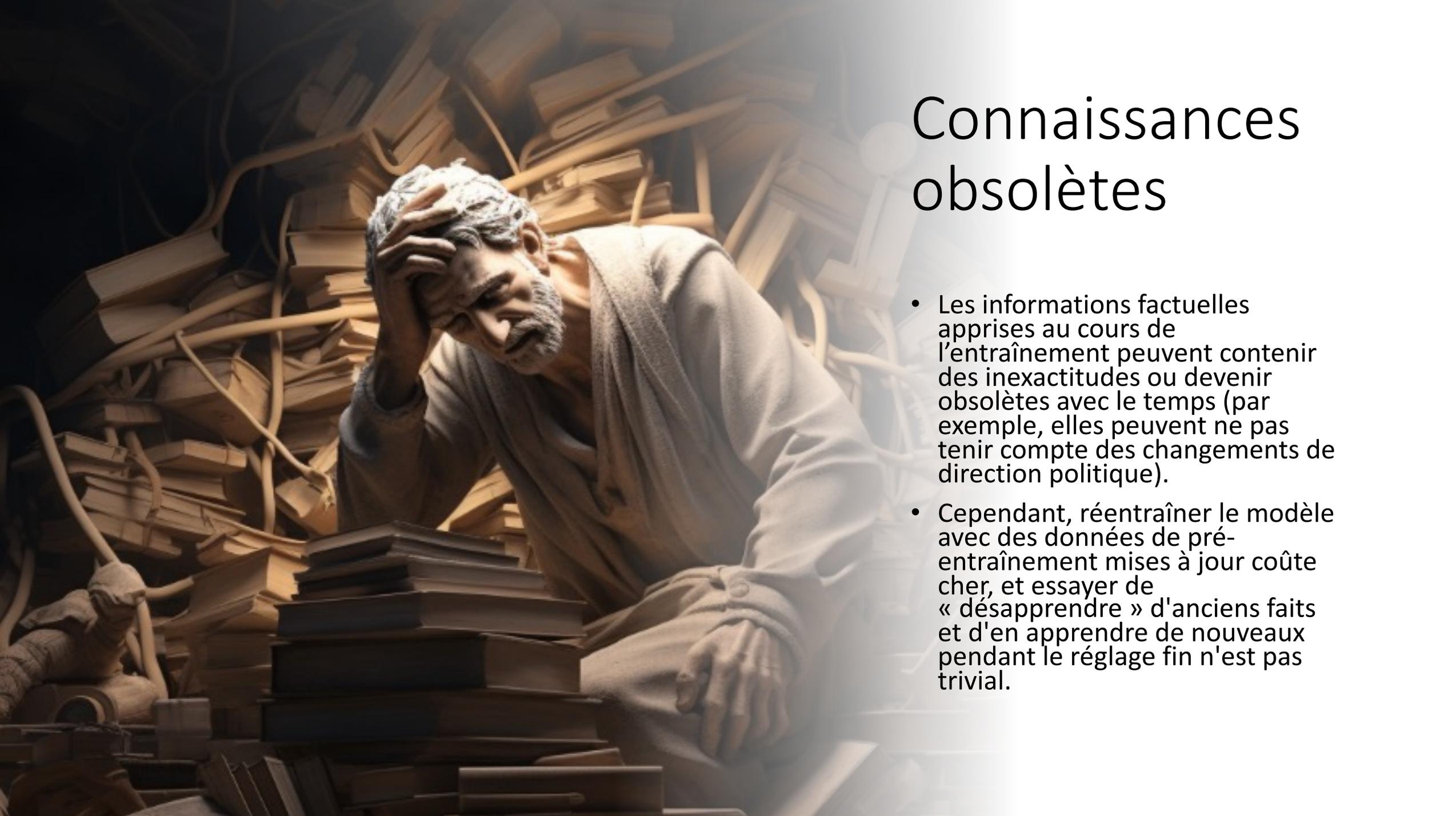
# Hallucinations

- Les LLM souffrent parfois d'hallucinations, qui contiennent des informations inexactes qui peuvent être difficiles à détecter en raison de la fluidité du texte.



# Comportement inapproprié

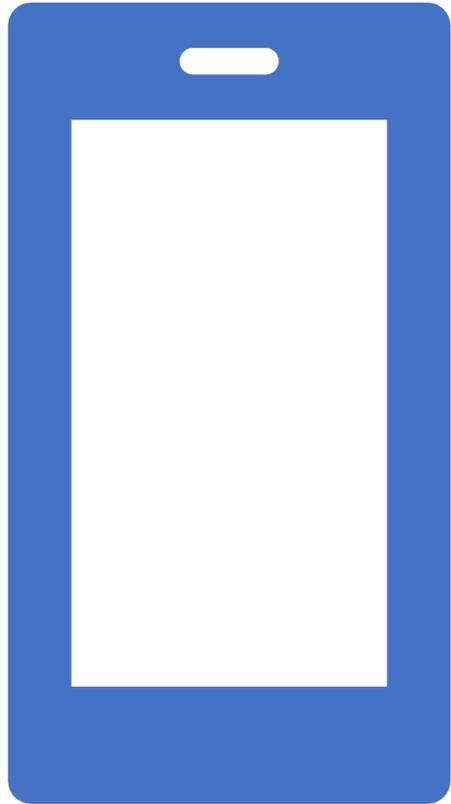
- L'alignement fait référence au défi consistant à garantir que le comportement du LLM s'aligne sur les valeurs, les objectifs et les attentes humaines et qu'il ne provoque pas de préjudices ou de conséquences involontaires ou indésirables.
- Les LLM génèrent souvent des résultats qui ne correspondent pas bien aux valeurs ou aux intentions humaines, ce qui peut avoir des conséquences inattendues ou négatives :
  - Pre-Training With Human Feedback
  - Instruction Fine-Tuning
  - Reinforcement Learning From Human Feedback (RLHF)



# Connaissances obsolètes

- Les informations factuelles apprises au cours de l'entraînement peuvent contenir des inexactitudes ou devenir obsolètes avec le temps (par exemple, elles peuvent ne pas tenir compte des changements de direction politique).
- Cependant, réentraîner le modèle avec des données de pré-entraînement mises à jour coûte cher, et essayer de « désapprendre » d'anciens faits et d'en apprendre de nouveaux pendant le réglage fin n'est pas trivial.





## Applications de l'IA générative

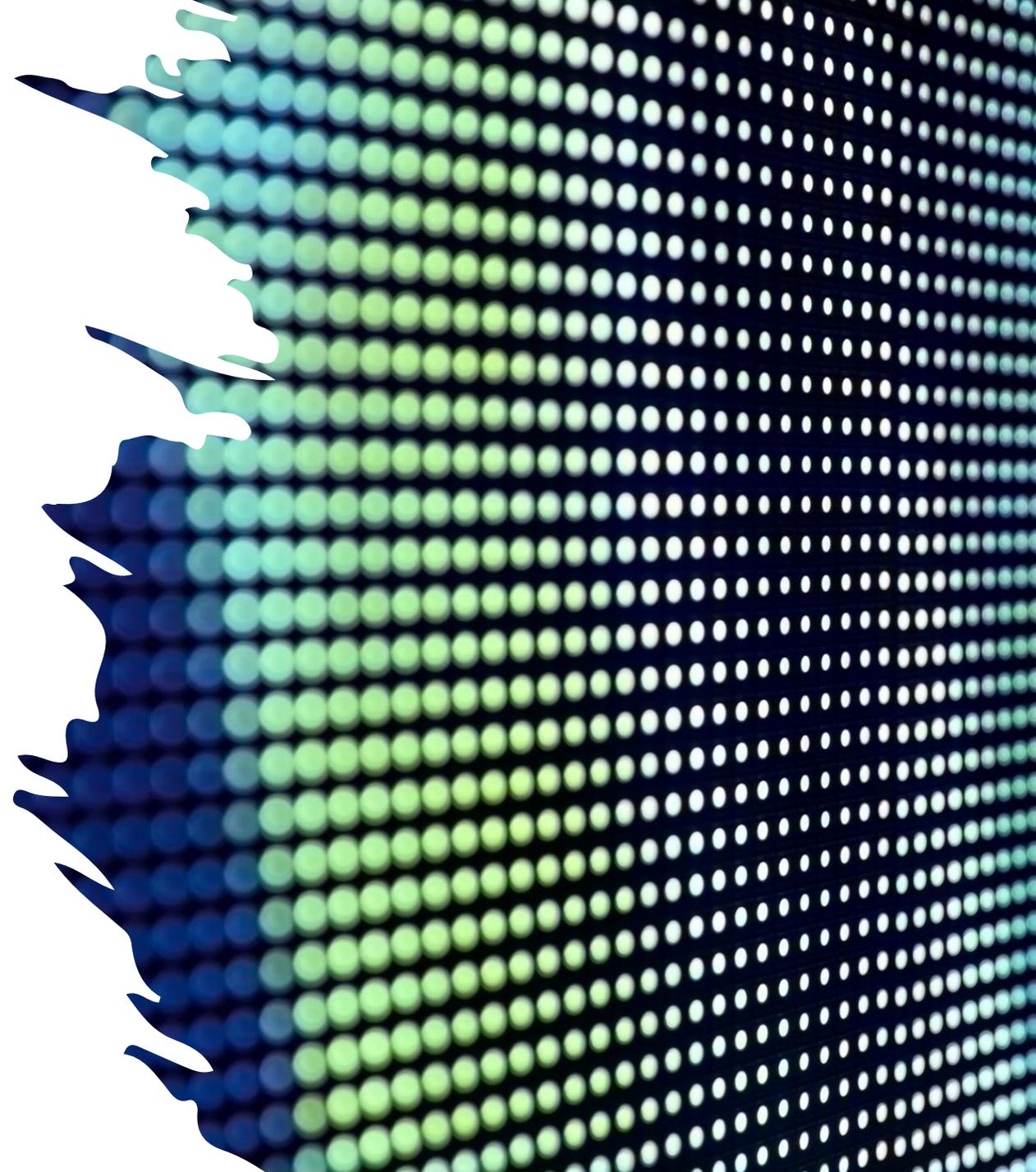
# Texte

- ChatBots
- [Résumé](#)
- Assistance pour l'écriture
- Génération de code
- Traduction
- Analyse des sentiments
- Écriture dans un style



# Image

- Image Editing
  - inpainting
  - outpainting
  - upscaling
  - super-resolution
  - deblurring
- Acteurs
  - OpenArt
  - MidJourney
  - Stable Diffusion
  - NightCafe







# Codage

- Assistant
  - compléter, expliquer, transformer et générer du code
- Création de sites
- Développement d'applications
- GitHub [Copilot](#) (1,3 M)
- [CodeWhisperer](#) (Amazon)

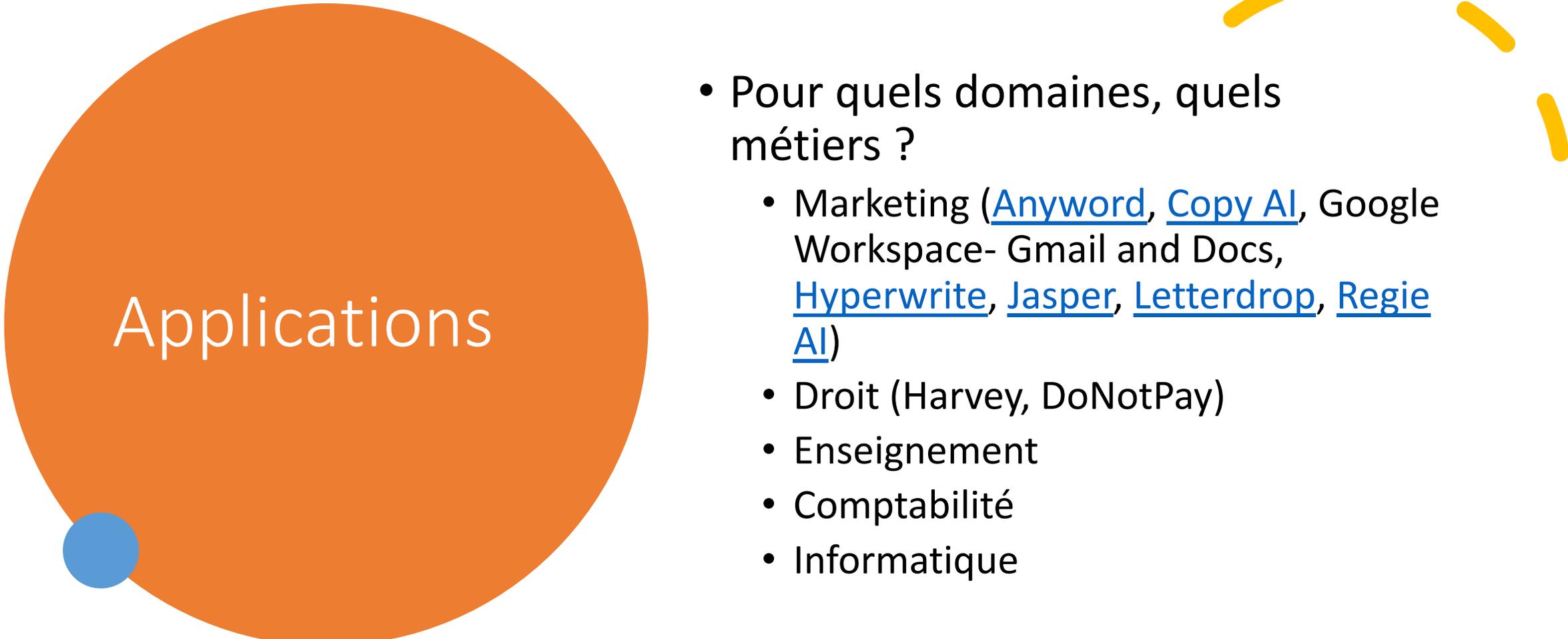
```
mirror_mod = modifier_ob.  
set mirror object to mirror.  
mirror_mod.mirror_object =  
operation == "MIRROR_X":  
mirror_mod.use_x = True  
mirror_mod.use_y = False  
mirror_mod.use_z = False  
operation == "MIRROR_Y":  
mirror_mod.use_x = False  
mirror_mod.use_y = True  
mirror_mod.use_z = False  
operation == "MIRROR_Z":  
mirror_mod.use_x = False  
mirror_mod.use_y = False  
mirror_mod.use_z = True  
selection at the end -add  
mirror_ob.select= 1  
modifier_ob.select=1  
context.scene.objects.active  
("Selected" + str(modifier_ob.  
mirror_ob.select = 0  
= bpy.context.selected_object  
data.objects[one.name].select  
print("please select exactly  
--- OPERATOR CLASSES ---  
types.Operator):  
on X mirror to the selected  
object.mirror_mirror_x"  
mirror X"  
context):  
context.active_object is not
```

# Musique / Voix / Sons

---

- [MusicLM](#) (280 heures de musique)
- AudioCraft
- [Suno.com](#)
- [Sorelec](#)

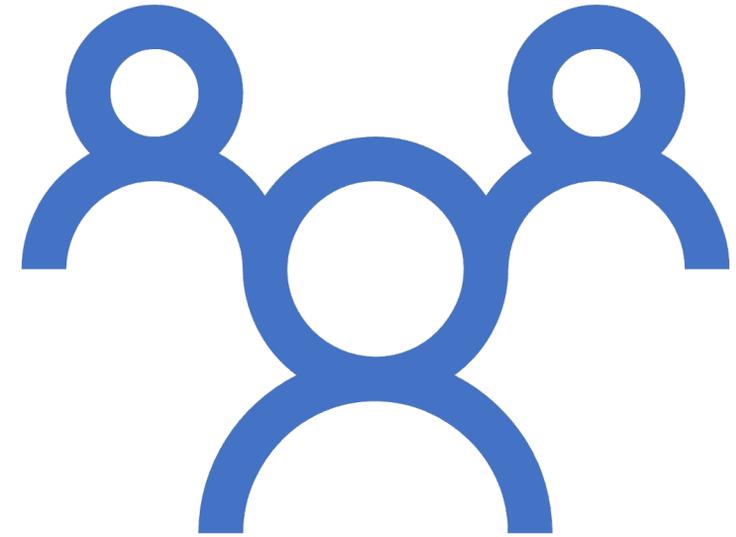




# Applications

- Pour quels domaines, quels métiers ?
  - Marketing ([Anyword](#), [Copy AI](#), Google Workspace- Gmail and Docs, [Hyperwrite](#), [Jasper](#), [Letterdrop](#), [Regie AI](#))
  - Droit (Harvey, DoNotPay)
  - Enseignement
  - Comptabilité
  - Informatique

# Le business de l'IA générative



● ChatGPT  
Terme de recherche

+ Comparer

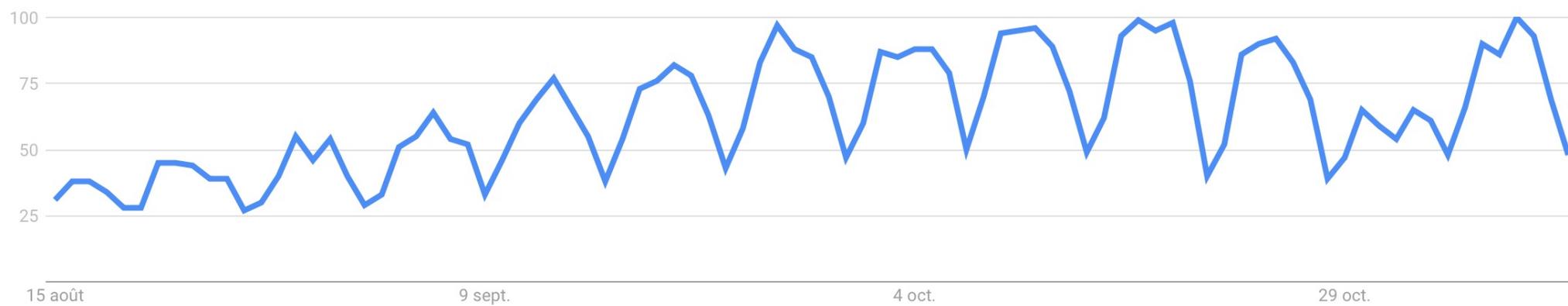
France ▼

90 derniers jours ▼

Toutes catégories ▼

Recherche sur le Web ▼

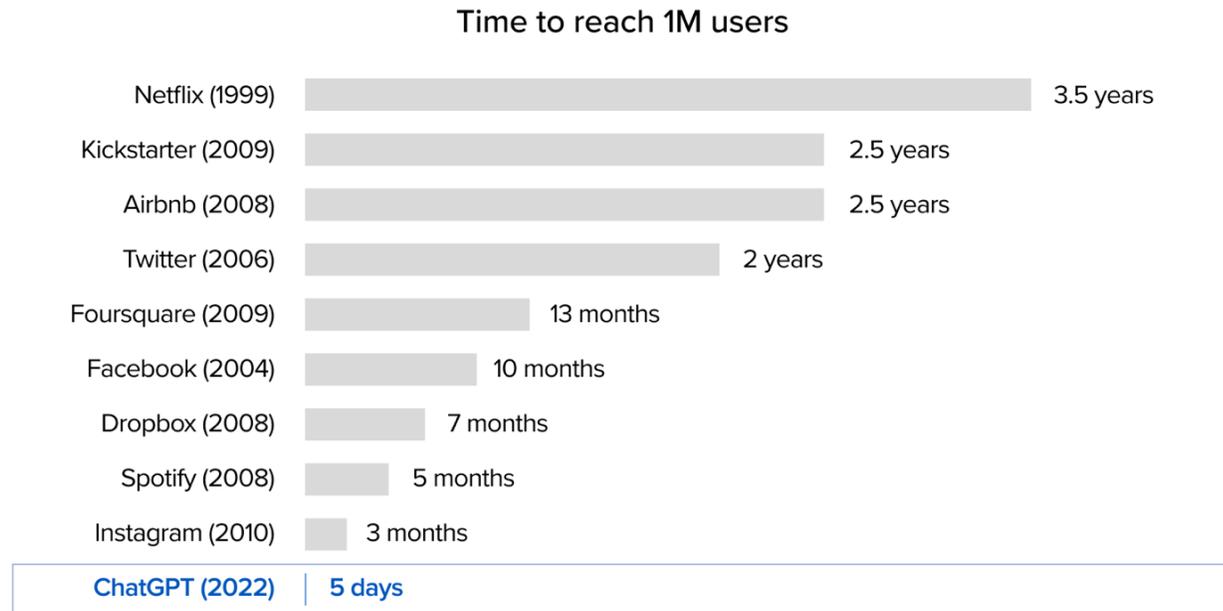
Évolution de l'intérêt pour cette recherche ?





# Arguments en faveur de l'IA générative

## Unprecedented Speed to Adoption

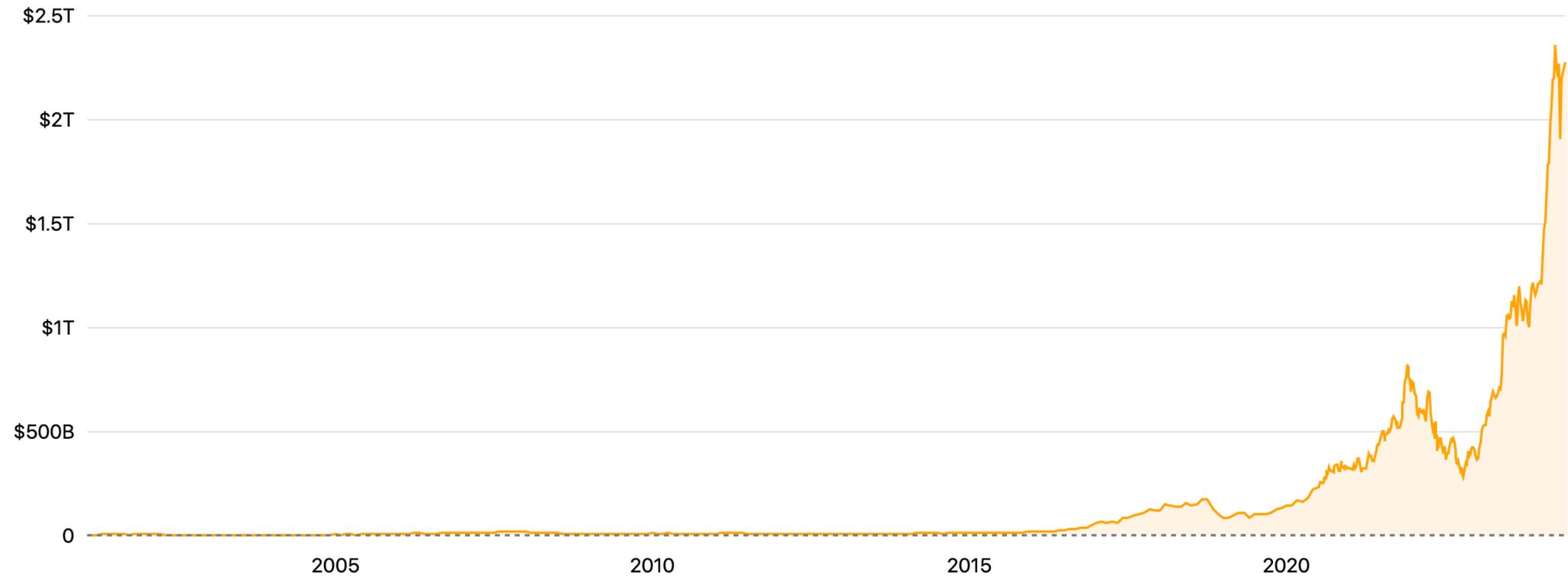


Source: Statista

ChatGPT a atteint plus de 100 millions d'utilisateurs (après 2 mois d'utilisation)

# NVIDIA

Market cap history of NVIDIA from 2001 to 2024





# L'IA / économie

---

100 millions de personnes utilisent les produits d'OpenAI

# Mistral AI

---

- Le 13 mai 2024, par Marie-Claude Benoit.
- **En quelques mois, Mistral AI est devenu un acteur incontournable de l'IA générative en France et en Europe. Alors que la start-up s'est élevée au rang de licorne fin décembre dernier avec un tour de table de 385 millions d'euros, elle serait en train de finaliser, selon le [Wall Street Journal](#), une levée de fonds de 600 millions de dollars, ce qui triplerait sa valorisation, la portant à environ 6 milliards de dollars.**
- Note : Mistral AI a été créée en avril 2023

# H

---

- Paris, le 21 mai 2024 –
- **“H”, une entreprise de modèles de fondation et d'IA agentique, a été lancée aujourd'hui à Paris et a annoncé une levée de fonds initiale de 220 millions de dollars auprès de firmes mondiales, dont Accel, UiPath Inc., Bpifrance, via son fonds Large Venture, Eric Schmidt, Xavier Niel, FirstMark, Elaia Partners, Eurazeo, Yuri Milner et Amazon. Parmi les autres participants au tour de table figurent Aglaé Ventures - de Bernard Arnault, Creandum, Motier Ventures, Samsung, et Visionaries Club notamment.**
- Note : H a été créée fin 2023



# LANCEMENT DE KYUTAI, LE PREMIER LABORATOIRE DE RECHERCHE EUROPÉEN INDÉPENDANT DÉDIÉ À L'OPEN SCIENCE EN IA

---

- KYUTAI a été fondé conjointement par le Groupe Iliad, le Groupe CMA CGM et Schmidt Futures.
- Le Groupe Iliad et le Groupe CMA CGM ont chacun contribué à hauteur de 100 millions d'euros au financement du laboratoire de recherche.
- *The Eric & Wendy Schmidt Fund for Strategic Innovation* s'est également jointe en tant que co-fondatrice, marquant ainsi un partenariat solide et un engagement commun dans ce projet.
- Déjà doté à près de 300 millions d'euros, KYUTAI a vocation à fédérer **d'autres investissements privés**. Les membres co-fondateurs invitent aujourd'hui d'autres entités à les rejoindre pour financer sur le long terme les travaux de cette organisation à but non-lucratif.
-

# ScaleWay

- **Paris, le 5 octobre 2023**
- Suite à l'annonce du Groupe Iliad du 26 septembre dernier d'investir 200 M€ dans l'Intelligence Artificielle, Scaleway, filiale du Groupe Iliad, détaille aujourd'hui son offre autour de son supercalculateur dédié à l'IA, Nabuchodonosor 2023 Supercalculateur d'IA, un nouveau système basé sur l'infrastructure NVIDIA DGX H100.

# GPU

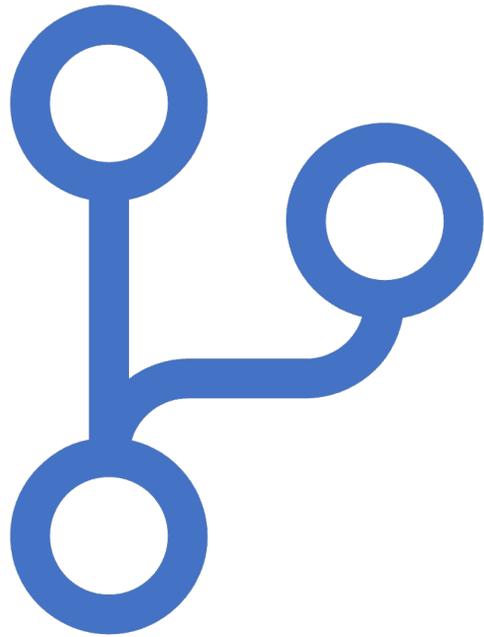
- [Meta](#) is spending billions of dollars on [Nvidia's](#) popular computer chips, which are at the heart of artificial intelligence research and projects.
- In an Instagram Reels [post](#) on Thursday, Zuckerberg said the company's "future roadmap" for AI requires it to build a "massive compute infrastructure." By the end of 2024, Zuckerberg said that infrastructure will include 350,000 [H100 graphics cards](#) from Nvidia.
- Zuckerberg didn't say how many of the graphics processing units (GPUs) the company has already purchased, but the H100 didn't hit the market until late 2022, and that was in limited supply. Analysts at Raymond James estimate Nvidia is selling the H100 for \$25,000 to \$30,000, and on [eBay](#) they can cost [over \\$40,000](#). If Meta were paying at the low end of the price range, that would amount to close to \$9 billion in expenditures.





## L'IA / économie

- Midjourney a vu son serveur Discord atteindre près de 15 millions de membres en juin 2023, moins d'un an après le lancement de sa version bêta ouverte en juillet 2022, ce qui en fait le plus grand serveur sur Discord.
- Character.AI, un fournisseur d'IA a obtenu environ 18 millions de visiteurs Web uniques actifs par mois et plus de 3 millions d'utilisateurs Web actifs quotidiens seulement 9 mois après son lancement



## L'IA / économie

- Le marché des développeurs d'IA connaît également une croissance considérable.
- Par exemple, la sortie du grand modèle d'image Stable Diffusion a fait exploser certains des projets de développeurs open source les plus réussis de l'histoire récente en termes de rapidité et de prévalence d'adoption.
- Le LLM de Meta Llama 2 a attiré plusieurs centaines de milliers d'utilisateurs, via des plateformes telles que Replicate, quelques jours après sa sortie en juillet.



## L'IA / économie

- L'exactitude est surfaite
- Applicable à un large éventail de marchés
- Bien meilleur que les humains pour les tâches à forte valeur ajoutée
- Toutes sortes de nouveaux comportements d'utilisateurs

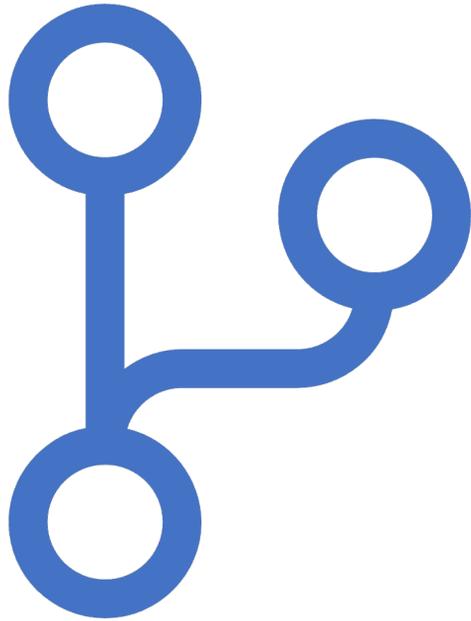


## L'IA / économie

- Faisons le calcul
- Au moment d'écrire ces lignes, le coût de calcul pour créer une image à l'aide d'un grand modèle d'image est d'environ 0,001 \$ et cela prend environ 1 seconde. Effectuer une tâche similaire avec un designer ou un photographe coûterait des centaines de dollars (minimum) et plusieurs heures ou jours (en tenant compte du temps de travail ainsi que des horaires). L'IA générative est 100 000 fois moins chère et 3 600 fois plus rapide que l'alternative humaine.
- Une analyse similaire peut être appliquée à de nombreuses autres tâches. Par exemple, les coûts nécessaires à un LLM pour résumer et répondre aux questions sur un dossier juridique complexe ne représentent qu'une fraction d'un centime, alors qu'un avocat facturerait généralement des centaines (et même des milliers) de dollars par heure et prendrait des heures ou des jours.

Avec vos données





# Questions concernant l'IA générative

- Copyright
- Emploi
- Deepfakes
- Enseignement
- L'accès

