



Forum **TERATEC** **23**

Unlock the future

31 MAI & 1^{er} JUIN 2023 • Au Parc Floral, Paris

Un événement organisé par

 **infoprodigital**

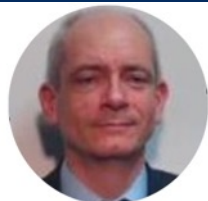




**Unlock
the future**

Atelier

« Les IA génératives passent à la vitesse supérieure »



Patrick Fabiani
DASSAULT AVIATION
Directeur Intelligence artificielle



Stéphane REQUENA
GENCI
Directeur Technique



Atelier

« Les IA génératives passent à la vitesse supérieure »

Les IA génératives proposées par les principaux acteurs du domaine sont passées à la vitesse supérieure en termes de tailles des modèles, d'exposition médiatique et de débat public.

Les questions qui se posent sur l'utilisation de ces outils vont des principes éthiques aux questions de sécurité, ou de souveraineté :

- « selon l'usage a-t-on le droit de les utiliser ou pas ? » « à l'école ? » « à l'université ? »
- « sont-elles un danger ? » « sont-elles une chance ? »
- « quelles garanties présentent elles ? » « peut-on les détourner ? »
- « sont-elles un outils utile ou dangereux selon l'usage ? »
- « aide concrète dans certains travaux évitant des tâches répétitives ou laborieuses ? »

... etc.

« Les IA génératives passent à la vitesse supérieure »

... un peu comme dans toute l'histoire des machines inventées par l'homme pour soulager sa peine et démultiplier ses efforts ?

1. Introduction « [IA Génératives ?](#) » par Stéphane Requena
2. Meriem Bendris (nVIDIA) : « [Les défis techniques de l'IA générative](#) »
3. Antoine Jacquot (Dassault Aviation) : « [Découvrez les possibilités avancées de ChatGPT pour les développeurs](#) »
4. Thomas Wolf (HuggingFace) : « [Libérer la puissance de l'IA en libre accès : Transformer les Industries Créatives](#) »

..... Pause

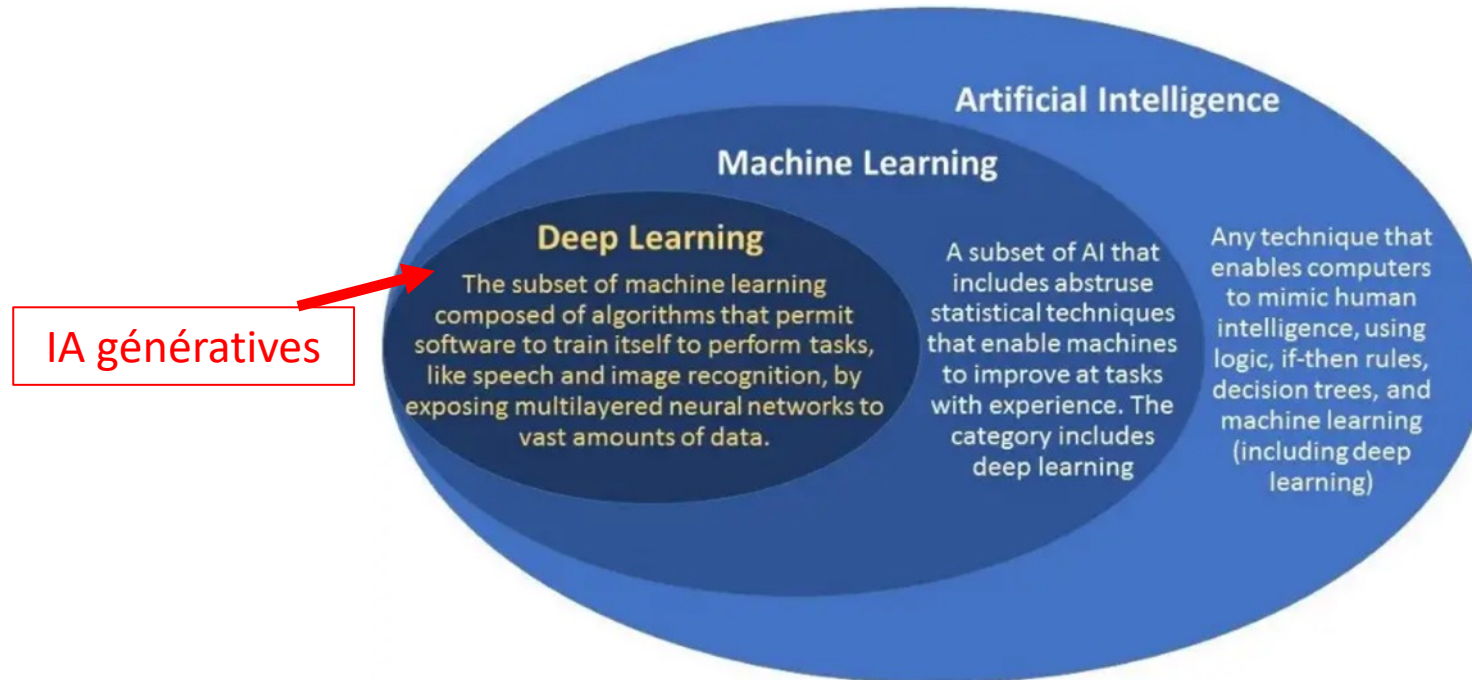
→ Reprise 11h30

5. Daphné Marnat (UnBias): « [Comment repérer les biais à risque discriminatoire \(ex. sexisme, racisme\) dans les modèles de langage et s'en prémunir ?](#) »
6. Laurence Devillers (LIMSI/CNRS) : « [IA, Chat GPT, quels enjeux pour demain ?](#) »
7. Débat pour le temps restant

Elle correspondent à des algorithmes d'IA qui utilisent des contenus existants pour apprendre et en générer de nouveaux. Il peut être question de texte, de sons, d'images, etc.

Ou

L'ensemble des algorithmes qui peuvent recevoir un message (texte, audio, images, bruit aléatoire ou rien du tout) et produire quelque chose qui semble réaliste et raisonnable pour les humains. (Zhou, 2022)





CHATGPT - Décryptage

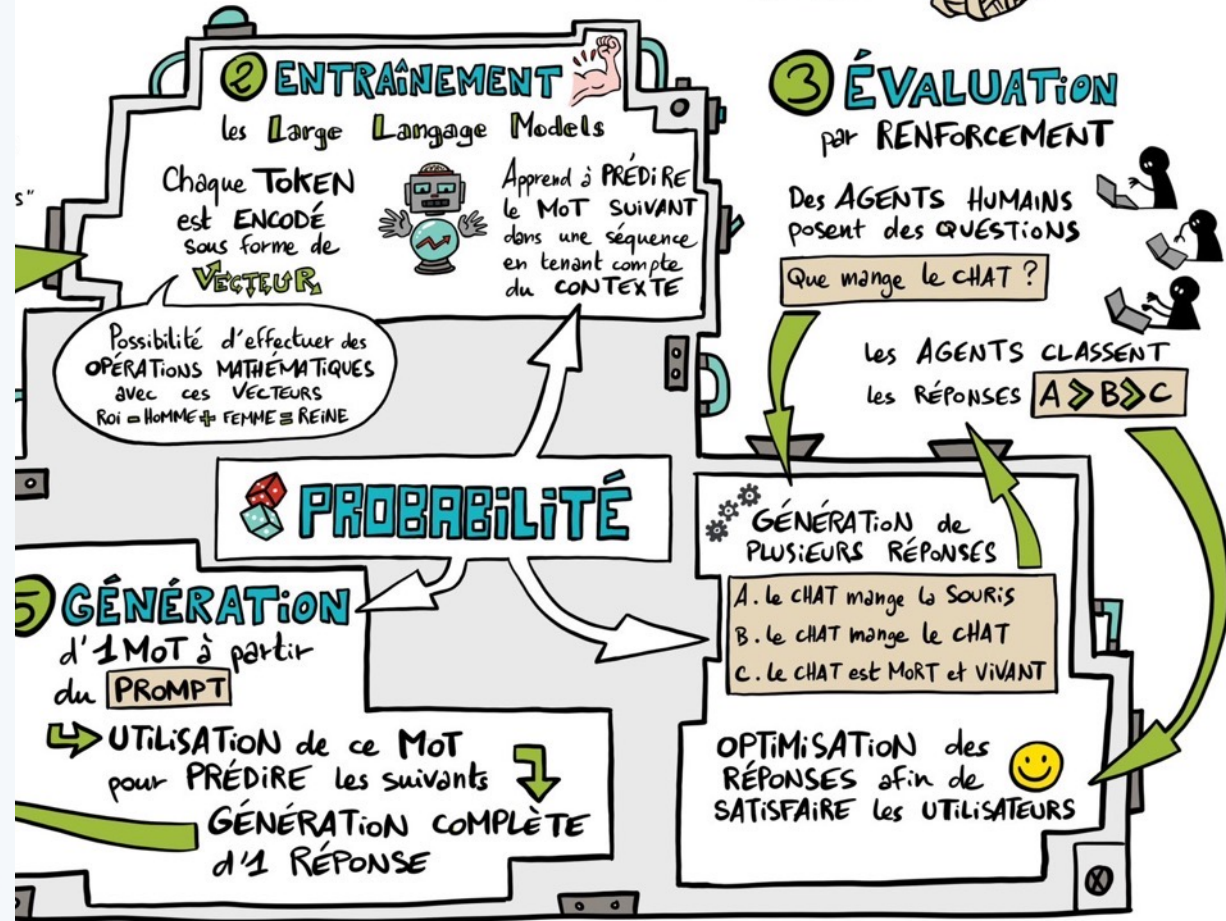


ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users

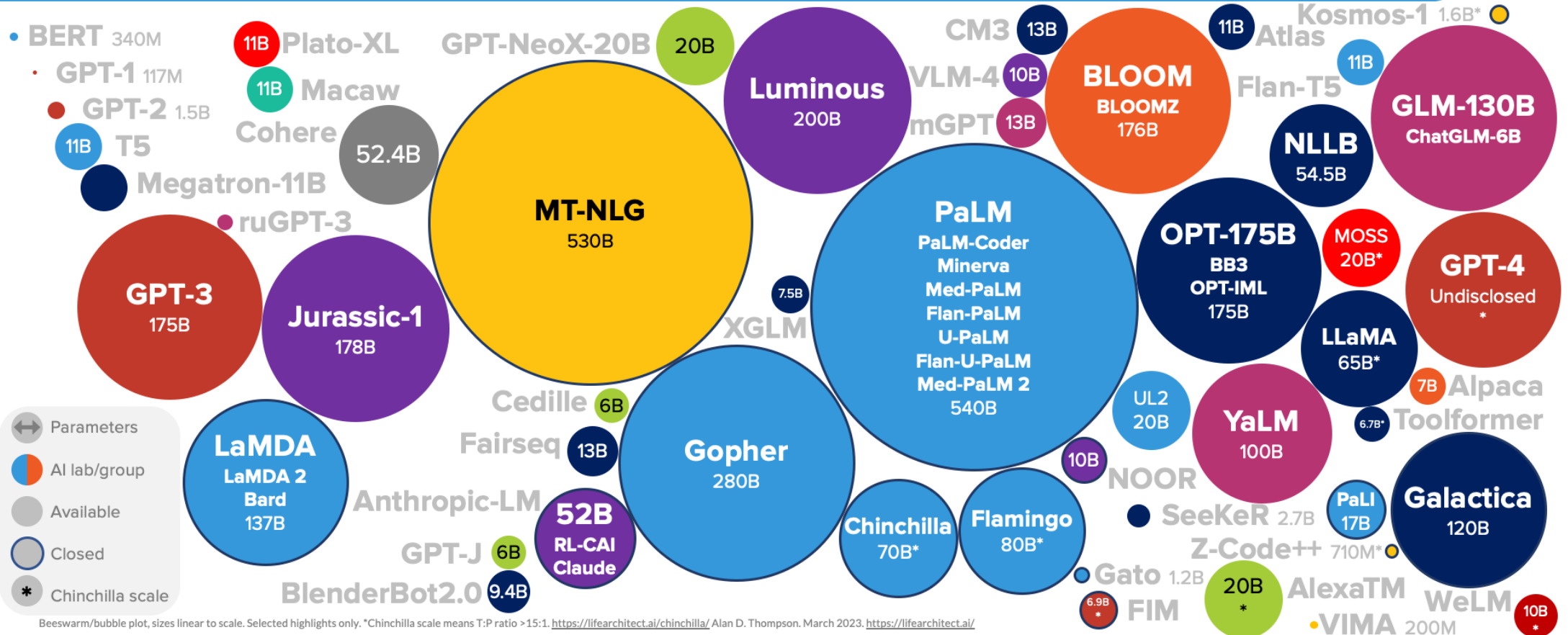


* one million backers ** one million nights booked *** one million downloads
Source: Company announcements via Business Insider/LinkedIn



→ CHATGPT ne comprend Ni le sens de la QUESTION, Ni celui de sa RÉPONSE.
→ Il est IMPOSSIBLE de remonter à la SOURCE des INFORMATIONS.

LANGUAGE MODEL SIZES TO MAR/2023



LA

- BERT
- GP
- G
- 11B

- Param
- AI lab/
- Availa
- Close
- Chin

Beeswarm/



Tweet épinglé

Guillaume Lample @GuillaumeLample · 24/02/2023

Today we release LLaMA, 4 foundation models ranging from 7B to 65B parameters.

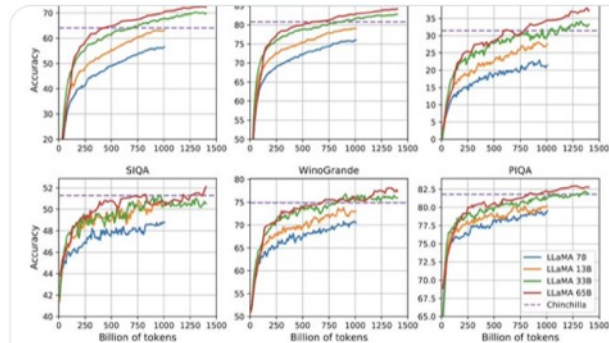
LLaMA-13B outperforms OPT and GPT-3 175B on most benchmarks.

LLaMA-65B is competitive with Chinchilla 70B and PaLM 540B.

The weights for all models are open and available at

research.facebook.com/publications/l...

1/n



| | BoolQ | PIQA | SIQA | HellaSwag | WinoGrande | ARC-e | ARC-c |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 175B | 60.5 | 81.0 | - | 78.9 | 70.2 | 68.8 | 51.4 |
| Chinchilla 280B | 79.3 | 81.8 | 50.6 | 79.2 | 70.1 | - | - |
| 70B | 83.7 | 81.8 | 51.3 | 80.8 | 74.9 | - | - |
| 62B | 84.8 | 80.5 | - | 79.7 | 77.0 | 75.2 | 52.5 |
| 62B | 83.9 | 81.4 | - | 80.6 | 77.0 | - | - |
| 540B | 88.0 | 82.3 | - | 83.4 | 81.1 | 76.6 | 53.0 |
| 7B | 76.5 | 79.8 | 48.9 | 76.1 | 70.1 | 72.8 | 47.6 |
| 13B | 78.1 | 80.1 | 50.4 | 79.2 | 73.0 | 74.8 | 52.7 |
| 33B | 83.1 | 82.3 | 50.4 | 82.8 | 76.0 | 80.0 | 57.8 |
| 65B | 85.3 | 82.8 | 52.3 | 84.2 | 77.0 | 78.9 | 56.0 |

Table 3: Zero-shot performance on Common Sense Reasoning tasks.

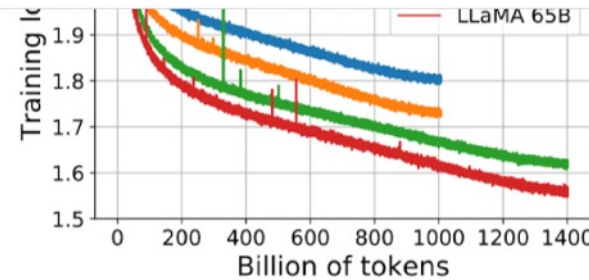


Figure 1: Training loss over train tokens for the 7B,

| | | | | | |
|------------|------|-------------|-------------|-------------|-------------|
| Gopher | 280B | 10.1 | - | 24.5 | 28.2 |
| Chinchilla | 70B | 16.6 | - | 31.5 | 35.5 |
| PaLM | 8B | 8.4 | 10.6 | - | 14.6 |
| | 62B | 18.1 | 26.5 | - | 27.6 |
| | 540B | 21.2 | 29.3 | - | 39.6 |
| LLaMA | 7B | 16.8 | 18.7 | 22.0 | 26.1 |
| | 13B | 20.1 | 23.4 | 28.1 | 31.9 |
| | 33B | 24.9 | 28.3 | 32.9 | 36.0 |
| | 65B | 23.8 | 31.0 | 35.0 | 30.0 |

175 1840 6972

gues

Kosmos-1 1.6B*

11B

GLM-130B
ChatGLM-6B

NLLB 54.5B

MOSS 20B*

GPT-4 Undisclosed *

LLaMA 65B*

7B Alpaca

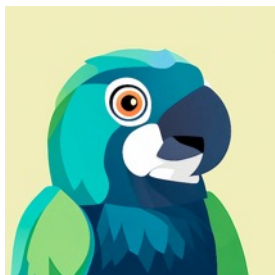
6.7B* Toolformer

PaLI 17B

Galactica 120B

AlexaTM WeLM 10B*

VIMA 200M



GUANACO
with QLoRA
Train LLMs on
IPHONE!!!

▶ 1 day ago



STABLEVICUNA
BEST OPEN
ChatGPT?

stability.ai

Stanford
Alpaca

Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks

Tiedong Liu, Bryan Kian Hsiang Low

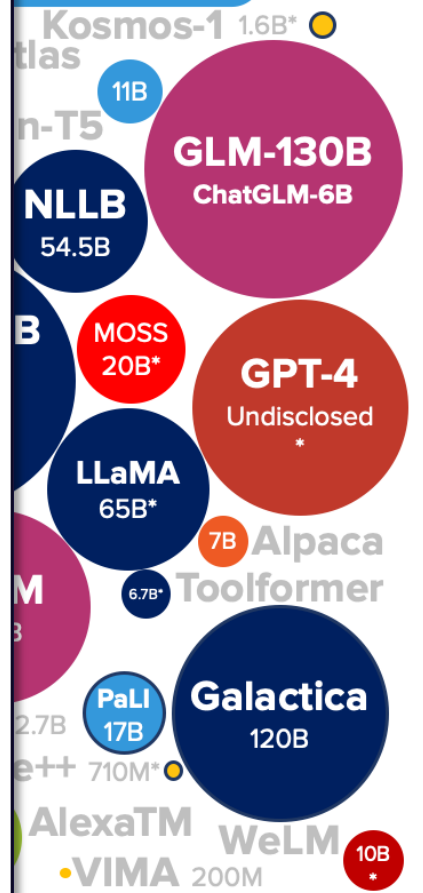
Falcon LLM

The advanced language model with 40B parameters and trained on 1T tokens

Vicuna-13B

New open-source LLM

Figures



- BERT
- GP
- G
- 11B

- ↔ Param
- AI lab/
- Availa
- Closed
- * Chinch
- Beeswarm/b





EuroHPC
Joint Undertaking



France
Universités



428 PF crête, 310 PF (LUMI-G),
4096 nœuds de calcul : 62% hybrides, 38%
scalaires (#3 top500)



Leonardo

323 PF crête, >250 PF HPL
4992 nœuds de calcul :
69% hybrides, 31% scalaires (#4 top500)



Marenostrum 5 (314 PF crête) à venir mi
2023 mais ratios similaires

IDRIS/CNRS - Ile de France

- 1^{ère} machine IA en France en réponse au plan #AIForHumanity
- Apporter une puissance **souveraine** pour la recherche française en IA
- **600 projets annuels en IA**
- **> 3100 GPUs nVIDIA** et **13 personnes dédiées** au support en IA

CINES/FU - Montpellier

- **> 70 PF** avec des CPU et GPU AMD next gen
- Dispo début 2023 avec notamment 2700 GPUs
- **Dernière grande étape** avant l'exascale pour la France

18 mois

- **FlauBERT (GC Jean Zay 2020)**
 - ✓ Un modèle de langage de type BERT (Google) en français
 - ✓ « Il fallait sauter sur cette occasion unique de créer une ressource pour le français de cette envergure. Pour cela nous avons monté l'équipe FlauBERT pour exploiter cette puissance de calcul indispensable au projet mais qui nous était inaccessible et malheureusement réservée jusqu'ici au GAFAM. » A. Allauzen
- **PAGnol, le plus grand modèle IA de langue française avec 1.5 milliard de paramètres, entraîné sur le supercalculateur Jean Zay !**
 - ✓ Collaboration entre une startup LightOn (qui fabrique entre autre des processeurs optiques) avec des chercheurs ENS et INRIA (équipe Almanach)
- **Big Science, le plus gros projet d'IA tournant sur Jean Zay (5 Mh GPU demandées)**
 - ✓ 500 partenaires (chercheurs, grands groupes, startups)
 - ✓ 1 an pour créer un mega modèle multi-langues, opensource, à l'état de l'art de l'existant en minimisant les biais



The screenshot shows a CNRS website page with the following content:

- Header:** cnrs, INSZI, Recherche, Innovation, International
- Article Title:** FlauBERT à la rescousse du traitement automatique du français
- Date:** 16 janvier 2020
- Text:** "De nombreux outils sont développés pour le traitement automatique du langage naturel, mais ils sont généralement en anglais et doivent être reconfigurés pour chaque langue. Avec FlauBERT, des chercheurs du LIG, du LAMSADE et du LIF proposent une version française de BERT, le dernier modèle de langue de Google."
- Section:** LightOn lance PAGnol, le plus grand modèle IA de langue française
- Image:** A cartoon character with a book and a speech bubble.
- Text in Image:** "Amorce fournie par l'utilisateur: A quelques kilomètres de la frontière, une nouvelle école privée propose un cursus en médecine dentaire, en partenariat avec une institution maltaise. A ce jour, dix établissements de ce type sont opérationnels à Malte. « Aujourd'hui, nous sommes les premiers à ouvrir dans notre pays et à répondre aux besoins d'une population dont le besoin en soins dentaires est très élevé », explique le docteur Alexandre Nzakaza, qui dirige l'école en lien avec le syndicat national des praticiens de santé dentaire, le Snphod." "Texte généré par PAGnol"
- Footer:** Le Monde | ACTUALITÉS | ÉCONOMIE | VIDÉOS | OPINIONS | CULTURE | M LE MAG | SERVICES
- Section:** SCIENCES · INTELLIGENCE ARTIFICIELLE
- Article Title:** Un projet géant pour faire parler une intelligence artificielle, et faire mieux que Google
- Text:** "Un consortium international de laboratoires, de grands groupes et de start-up se donne un an pour créer un modèle de langues multilingue open source, plus abouti et moins biaisé que ceux développés par OpenAI et Google. En ayant recours, notamment, à la puissance du supercalculateur français Jean Zay."

Régulation de l'usage

ARTIFICIAL INTELLIGENCE

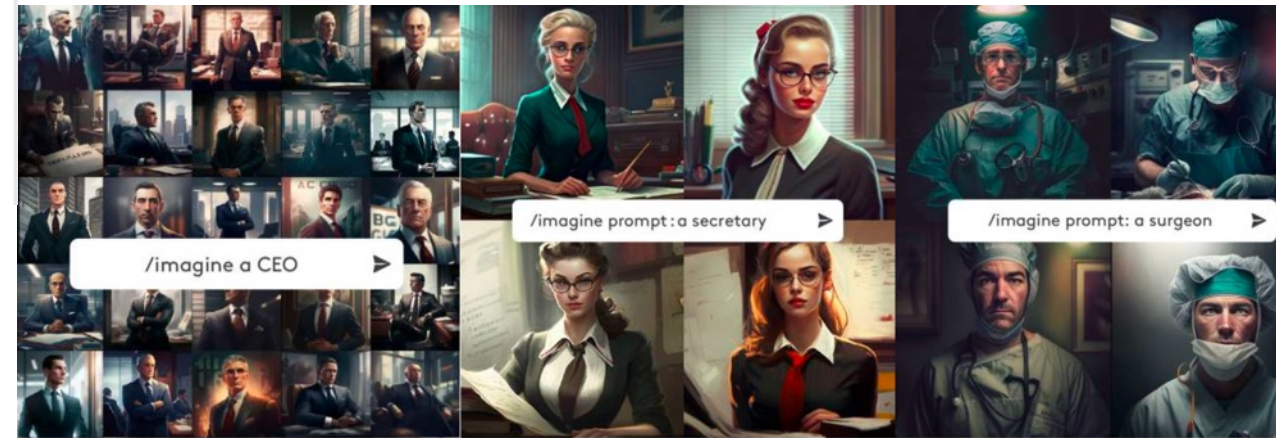
The European Union's Artificial Intelligence Act, explained

Diffusion extrême, formation et gestion de la PI

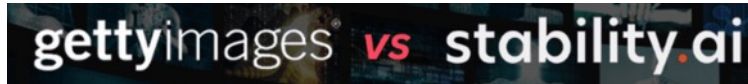
BloombergGPT: A Large Language Model for Finance



Comprendre les hallucinations et lutter contre les biais / fake



<https://www.jamaissanselles.fr/biais-intelligence-artificielle/>



Maitrise de la consommation énergétique

| Model name | Number of parameters | Datacenter PUE | Carbon intensity of grid used | Power consumption | CO ₂ eq emissions | CO ₂ eq emissions × PUE |
|------------|----------------------|-------------------|-------------------------------|-------------------|------------------------------|------------------------------------|
| GPT-3 | 175B | 1.1 | 429 gCO ₂ eq/kWh | 1,287 MWh | 502 tonnes | 552 tonnes |
| Gopher | 280B | 1.08 | 330 gCO ₂ eq/kWh | 1,066 MWh | 352 tonnes | 380 tonnes |
| OPT | 175B | 1.09 ² | 231 gCO ₂ eq/kWh | 324 MWh | 70 tonnes | 76.3 tonnes ³ |
| BLOOM | 176B | 1.2 | 57 gCO ₂ eq/kWh | 433 MWh | 25 tonnes | 30 tonnes |

Bob's wife is Amy. Bob's daughter is Cindy. Who is Cindy to Amy?

Cindy is Amy's **daughter-in-law**.

(a) Intrinsic hallucination

Explain RLHF for LLMs.

RLHF stands for "**R**ights, **L**imitations, **H**arms, and **F**reedoms" and is a framework for models like LLMs (Large Language Models).

(b) Extrinsic hallucination



SONDAGE IFOP POUR TALAN

LES FRANÇAIS ET LES IA GÉNÉRATIVES

Étude réalisée par l'Ifop pour Talan en mai 2023

L'enquête a été menée auprès d'un échantillon de 1008 personnes, représentatif de la population française âgée de 18 ans et plus.

ifop

Talan*



des Français qui utilisent les IA génératives dans l'entreprise ne le disent pas à leur supérieur hiérarchique



souhaitent que l'État français soutienne davantage l'émergence d'entreprises françaises qui développent des IA génératives



estiment ne pas avoir les connaissances suffisantes pour les utiliser. Un énorme effort de formation semble donc nécessaire

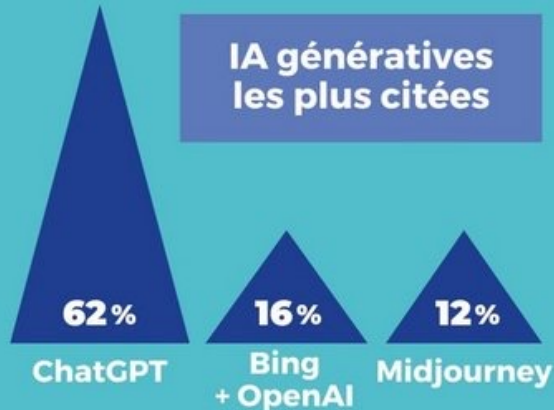


ont des craintes vis-à-vis de l'émergence des IA génératives

des Français ont entendu parler des IA génératives

71%

IA génératives les plus citées



45%

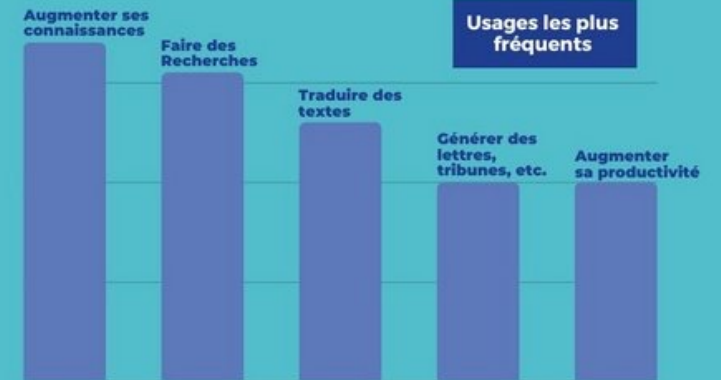
des 18-24 ans utilisent les IA génératives contre seulement 18 % des 35 ans et +

74%

de ces Français pensent qu'elles constituent une nouvelle révolution industrielle

51%

estiment que les enseignants doivent s'en saisir pour en enseigner les avantages et inconvénients



ifop

Talan*

« Les IA génératives passent à la vitesse supérieure »

... un peu comme dans toute l'histoire des machines inventées par l'homme pour soulager sa peine et démultiplier ses efforts ?

1. Introduction « [IA Génératives ?](#) » par Stéphane Requena
2. Meriem Bendris (nVIDIA) : « [Les défis techniques de l'IA générative](#) »
3. Antoine Jacquot (Dassault Aviation) : « [Découvrez les possibilités avancées de ChatGPT pour les développeurs](#) »
4. Thomas Wolf (HuggingFace) : « [Libérer la puissance de l'IA en libre accès : Transformer les Industries Créatives](#) »

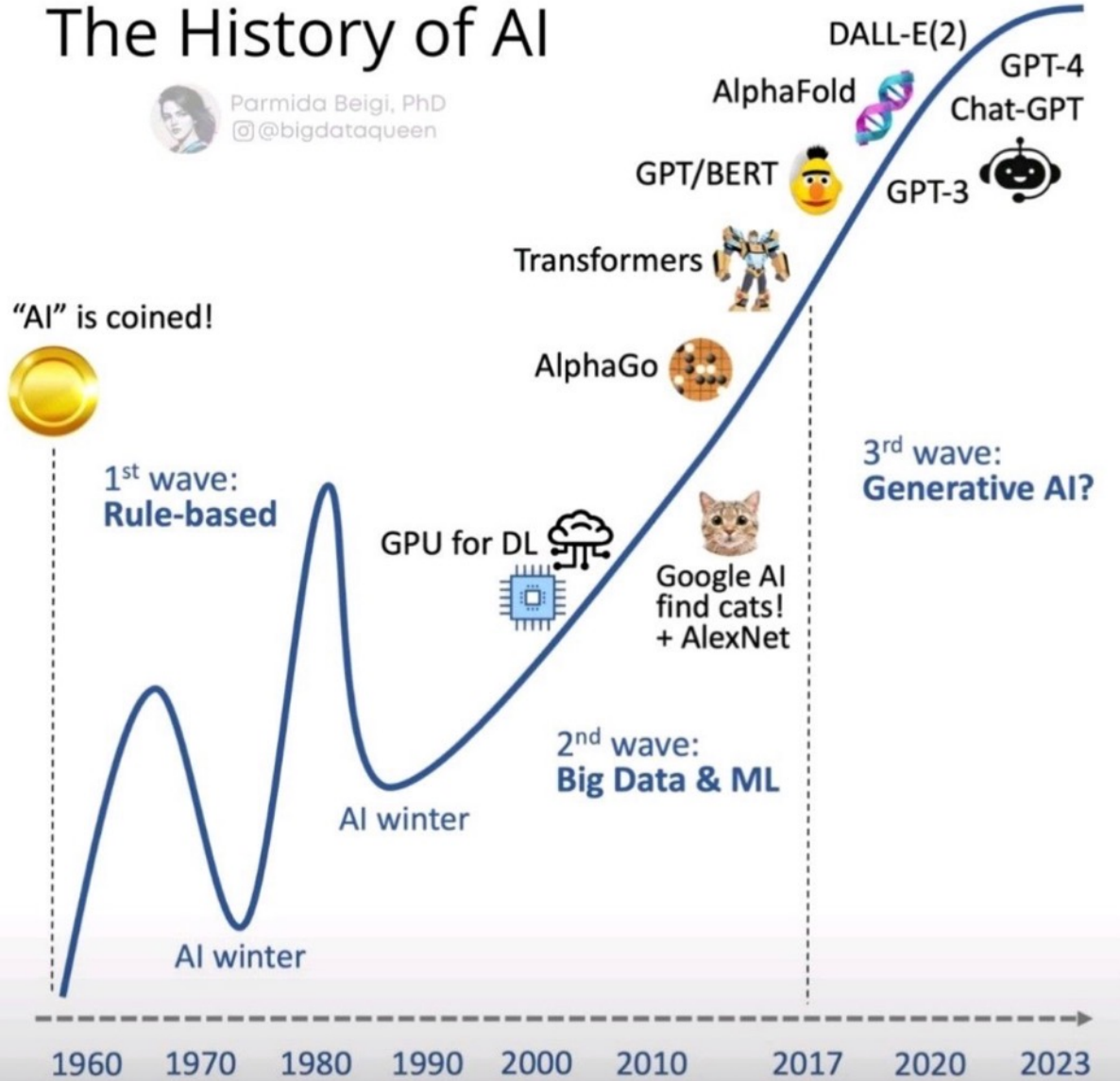
..... Pause

→ Reprise 11h30

5. Daphné Marnat (UnBias): « [Comment repérer les biais à risque discriminatoire \(ex. sexisme, racisme\) dans les modèles de langage et s'en prémunir ?](#) »
6. Laurence Devillers (LIMSI/CNRS) : « [IA, Chat GPT, quels enjeux pour demain ?](#) »
7. Débat pour le temps restant

The History of AI

 Parmida Beigi, PhD
@bigdataqueen



1950 : Test de Turing

1956 : premier usage de l'IA pour de la recherche de documents sur IBM 701 avec mécanisme de récompenses

1961 : M. Minsky concept société de machines qui coopèrent

1965 : ELIZA le 1^{er} chatbot

1982 : réseaux neuronaux récurrents (RNN)

1997 : introduction des LSTM (long-short term memory)

2003 : Y. Bengio introduit notion feed-forward network

2011 : Apple Siri

2013 : Google word2vec

2014 : I. Goodfellow introduit les GAN

2015 : D. Bahdanu introduit les mécanismes d'attention

2017 : Google introduit le modèle Transformer

2018 : A. Radford présente GPT (Generative Pre Training)

2017-2019 : OpenAI GPT1 et GPT2

2021 : OpenAI sort Dall.E

2022 : Stability AI sort Stable AI text to image model

2022 : OpenAI sort chatGPT et Midjourney passe en beta

2023 : OpenAI GPT4 et intégration chatGPT dans MS Bing