Forum TERATEC 23

Unlock the future

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

Un événement organisé par infoprodigital Teratec

Forum TERATEC 23

Atelier

« Les IA génératives passent à la vitesse supérieure»

Patrick Fabiani
DASSAULT AVIATION
Directeur Intelligence artificielle

Stéphane REQUENA
GENCI
Directeur Technique

Forum TERATEC

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.

Forum TERATEC

Atelier

« 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

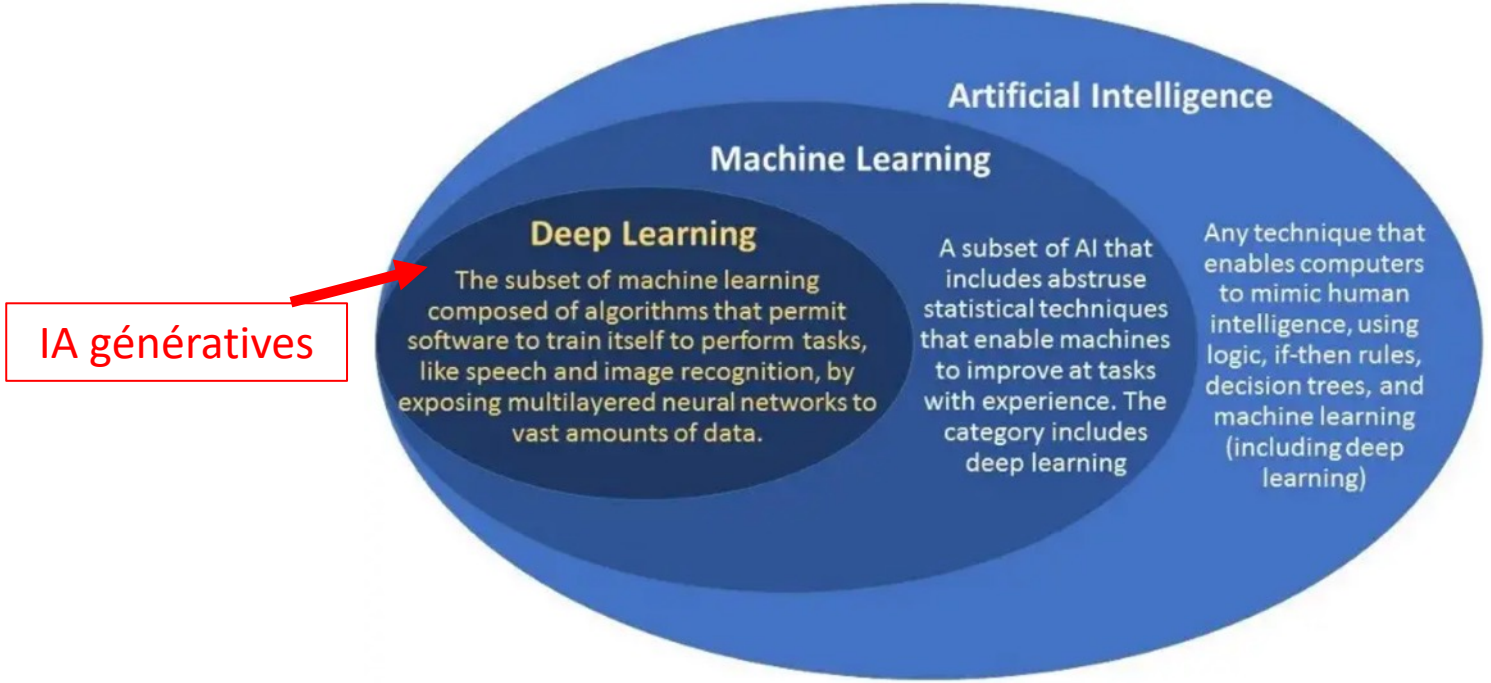
5. Daphné Marnat (UnBias): « Comment repérer les biais à risque discriminatoire (ex. sexisme, racisme) dans les modèles de langage et s'en prémunir ? »
6. Laurence Devillers (LIMSI/CNRS) : « IA, Chat GPT, quels enjeux pour demain ? »
7. Débat pour le temps restant

IA Géneratives ???

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)



Forum TERATEC

L'exemple de ChatGPT

ChatGPT Sprints to One Million Users

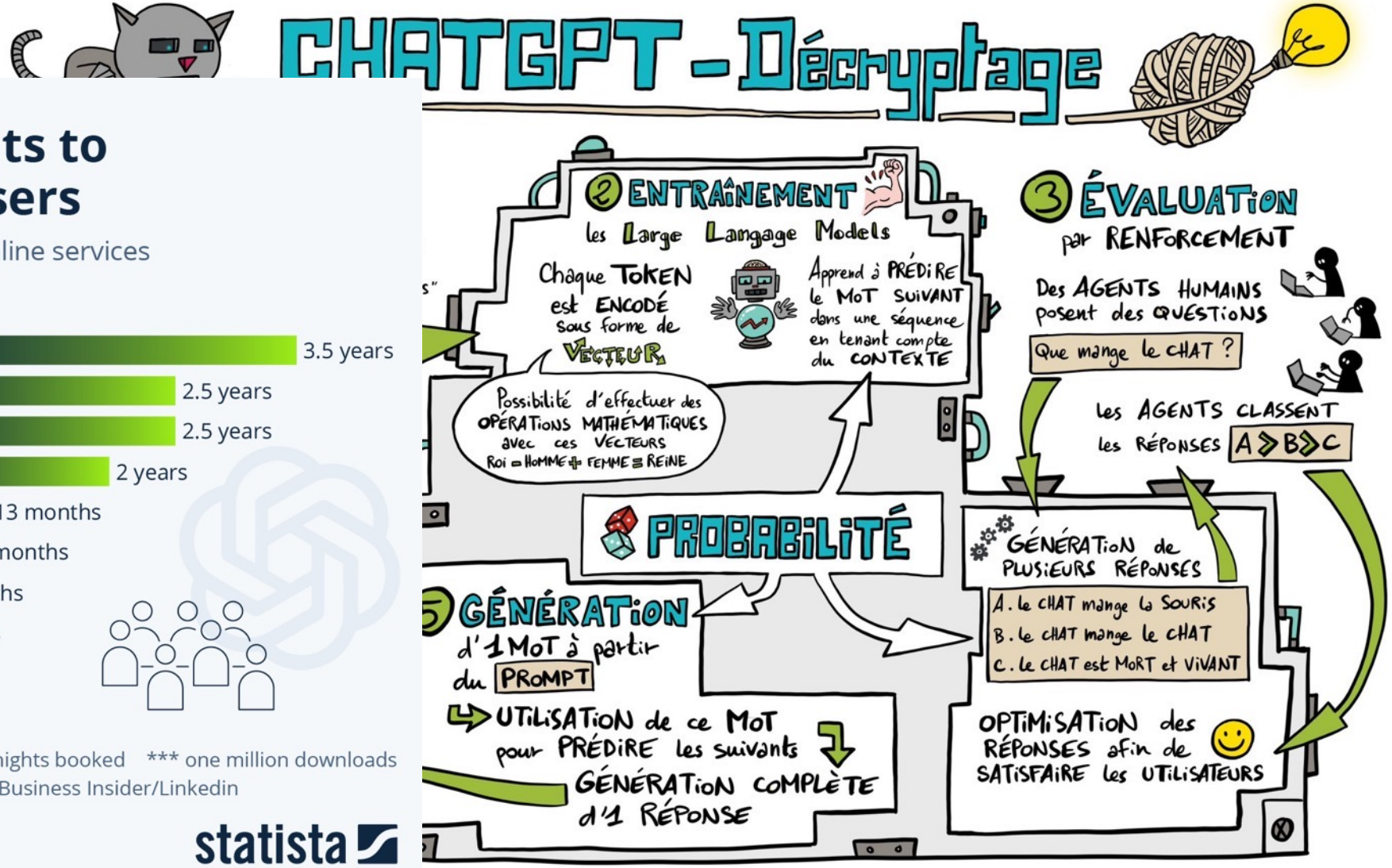
Time it took for selected online services to reach one million users

	Launched	
Netflix	1999	3.5 years
Kickstarter*	2009	2.5 years
Airbnb**	2008	2.5 years
Twitter	2006	2 years
Foursquare***	2009	13 months
Facebook	2004	10 months
Dropbox	2008	7 months
Spotify	2008	5 months
Instagram***	2010	2.5 months
ChatGPT	2022	5 days

* one million backers ** one million nights booked *** one million downloads

Source: Company announcements via Business Insider/Linkedin

statista

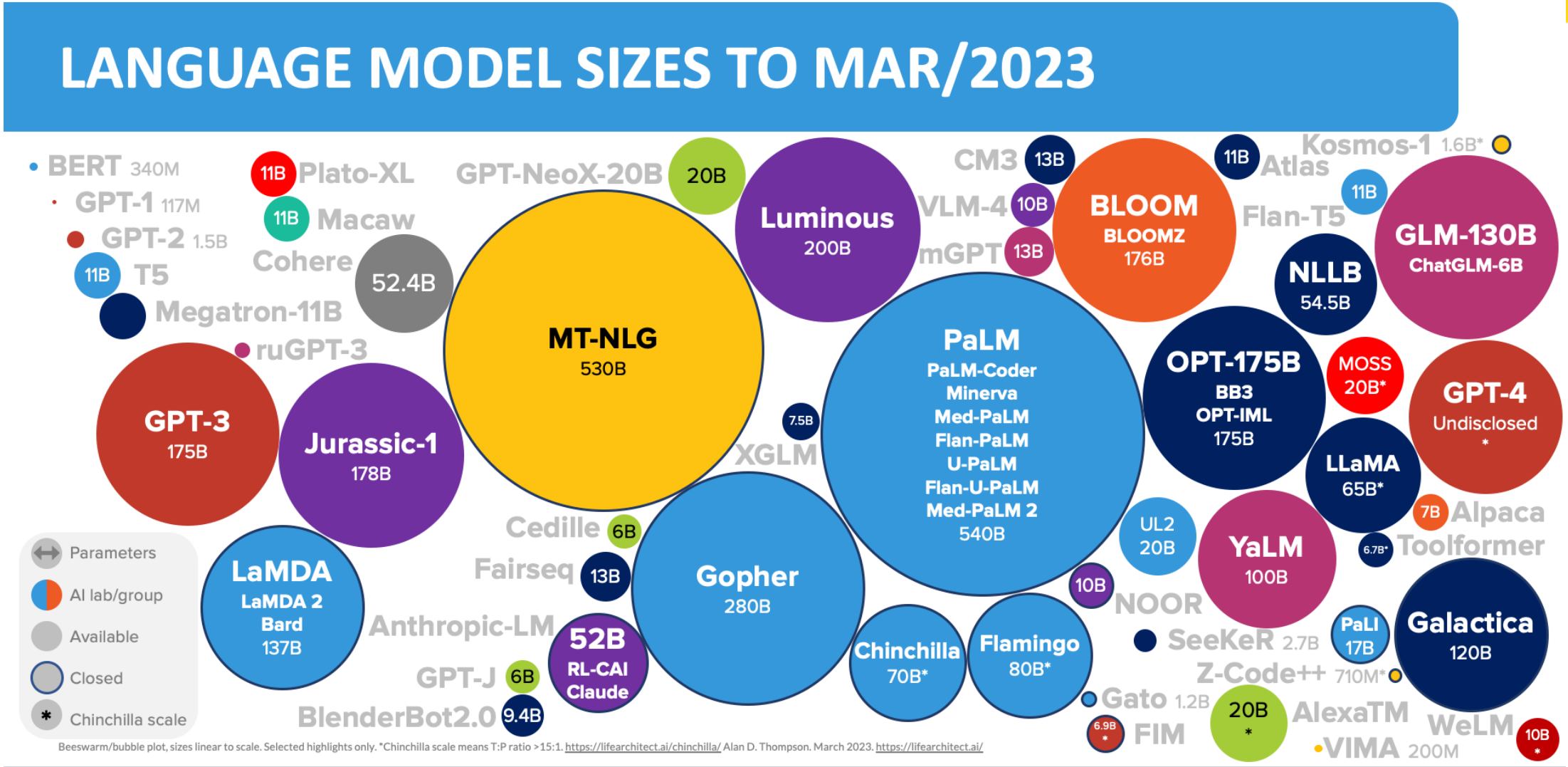


SketchNote @tewoz pour #EnjoyDigit by BNP PARIBAS

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

Forum TERATEC

Contexte : explosion cambrienne du nombre de modèles



Forum TERA

Tweet épinglé

Guillaume Lample @GuillaumeLample · 24/02/2023

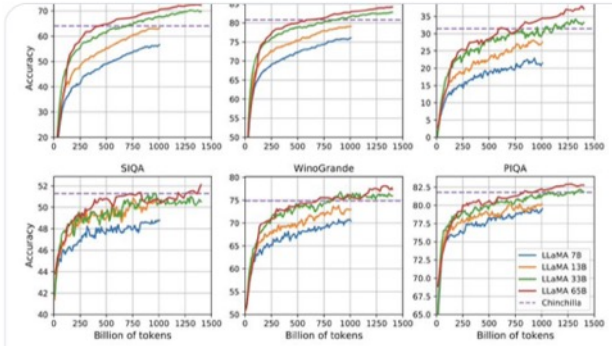
Today we release LLaMA, 4 foundation models ranging from 7B to 65B parameters.

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



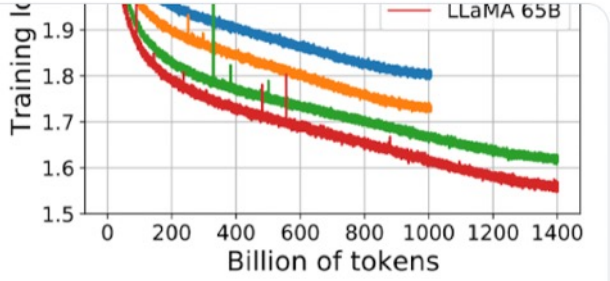


Figure 1: **Training loss over train tokens for the 7B,**

	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c
175B	60.5	81.0	-	78.9	70.2	68.8	51.4
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

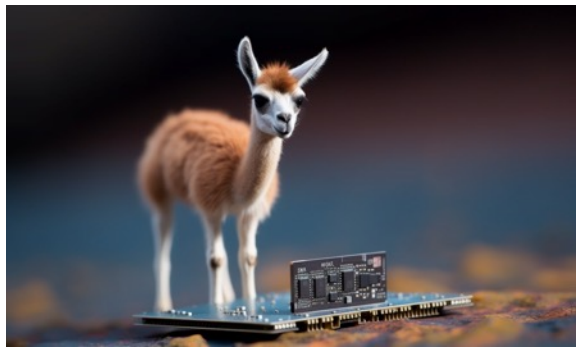
Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

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**	**39.9**

175 1840 6972

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 710M* AlexaTM WeLM 10B* VIMA 200M

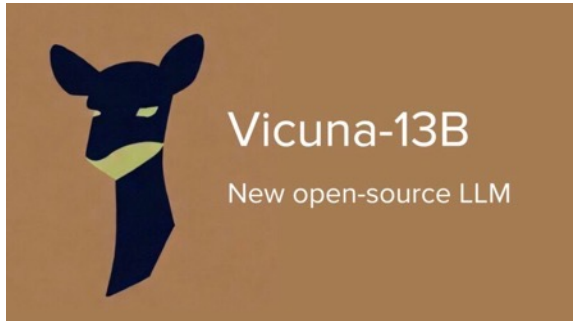
Forum TERA



Stanford Alpaca

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

Tiedong Liu, Bryan Kian Hsiang Low



L'enjeu des infrastructures de calcul publiques

Forum TERATEC

GENCI

EuroHPC Joint Undertaking

France Universités

IDRIS/CNRS - Ile de France

- **1ère 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**

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

Forum TERATEC

La preuve par l'exemple sur Jean Zay (GENCI)

*1000 en 18 mois

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, entrainé 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

Quelques autres enjeux

Régulation de l'usage

ARTIFICIAL INTELLIGENCE

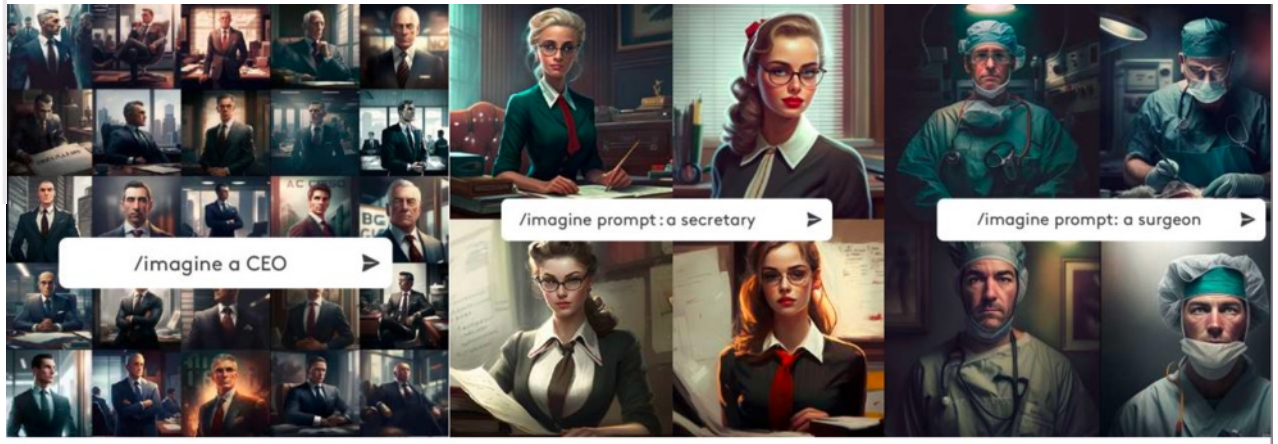
The European Union's Artificial Intelligence Act, explained

WH.GOV

MAY 04, 2023

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety

Comprendre les hallucinations et lutter contre les biais / fake



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

Diffusion extrême, formation et gestion de la PI

BloombergGPT: A Large Language Model for Finance

techradarpro Subscribe

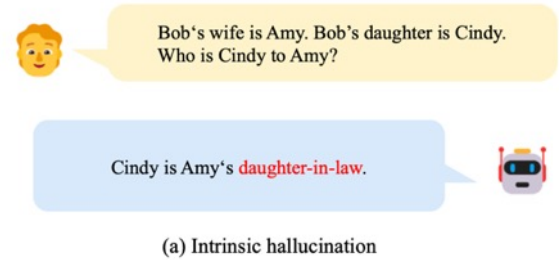
Samsung workers made a major error by using ChatGPT

By Lewis Maddison published 1 day ago

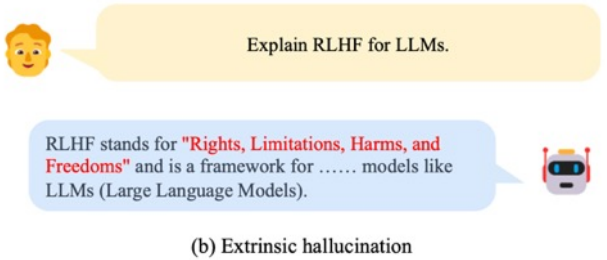
Samsung meeting notes and new source code are now in the wild after being leaked in ChatGPT

gettyimages vs stability.ai

Maitrise de la consommation énergétique



(a) Intrinsic hallucination

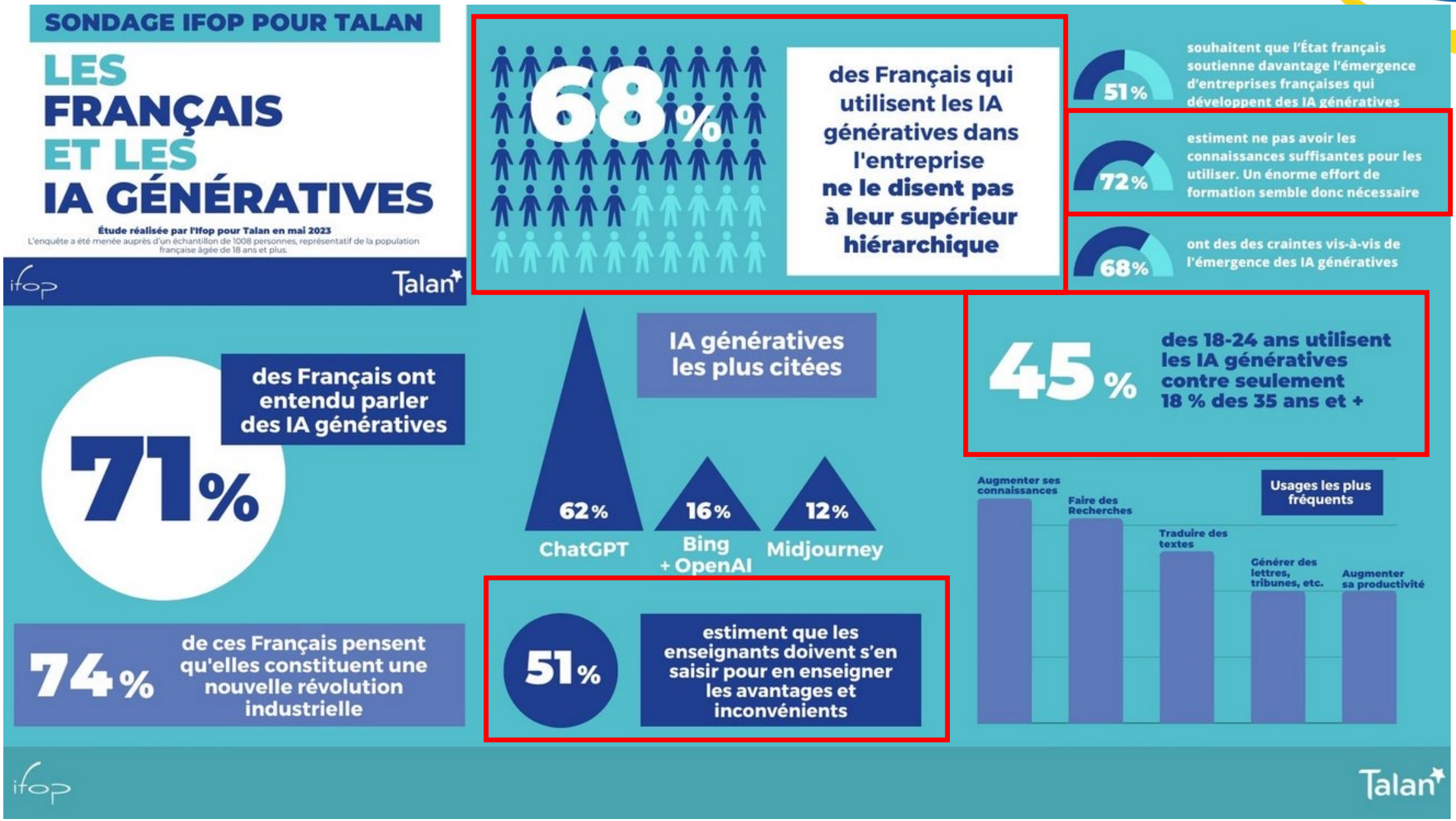


(b) Extrinsic hallucination

Model name	Number of parameters	Datacenter PUE	Carbon intensity of grid used	Power consumption	CO_2eq emissions	CO_2eq emissions × PUE
GPT-3	175B	1.1	429 gCO_2eq/kWh	1,287 MWh	*502 tonnes*	552 tonnes
Gopher	280B	1.08	330 gCO_2eq/kWh	*1,066 MWh*	*352 tonnes*	380 tonnes
OPT	175B	*1.09* [2]	*231 gCO_2eq/kWh*	*324 MWh*	70 tonnes	*76.3 tonnes* [3]
BLOOM	176B	1.2	57 gCO_2eq/kWh	433 MWh	25 tonnes	30 tonnes

Forum TERATEC

IA génératives : leur impact dans la société



Forum TERATEC

Atelier

« 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 »

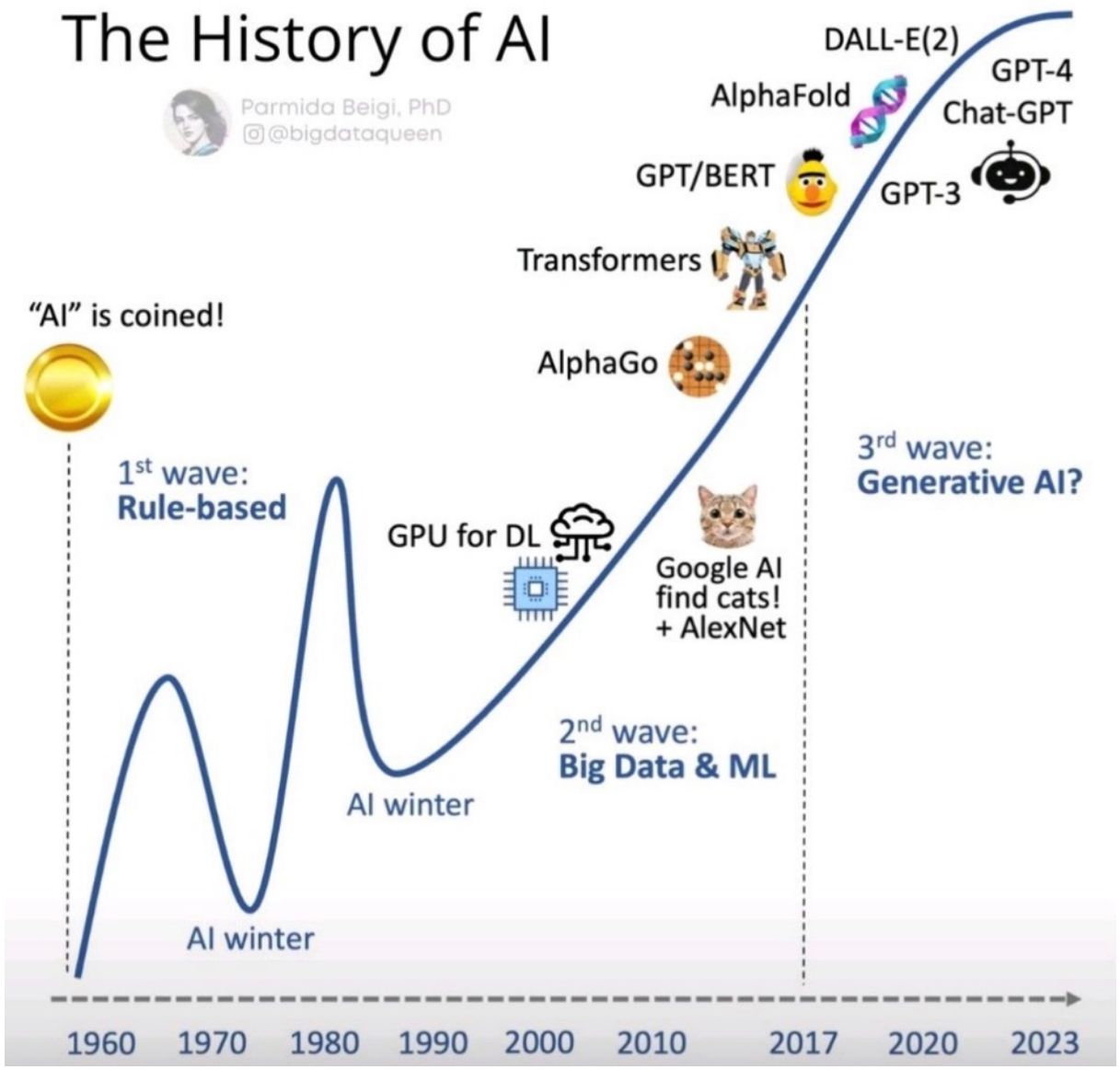
…… Pause ……..

→ 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

Forum TERATEC

IA Génératives : un peu d'histoire



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 1er chatbot
1982 : réseaux neuronaux récurrents (RNN)
1997 : introduction des LSTM (long-short term memory)
2003 : Y. Bengio introduit notion feed-foward 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