IA générative

Gilles Le Marois

gilles.lemarois@gmail.com

Le 30 Nov. 2022

Sam Altman (OpenAI)
dévoile **ChatGPT**

❖ Convergence

- des **outils de + en + puissants**
  > physiques (ordinateurs, smartphones, IoT, …)
  > numériques (***données*** et traitement, algorithmes…)

- des **moyens** de partage mondialisé et quasi-instantanné
  > internet, réseaux sociaux…
  > réseaux, fibre…

- une **recherche de - en - cloisonnée**
  > sciences du vivant (réseau de neurones, bioinformatique…)

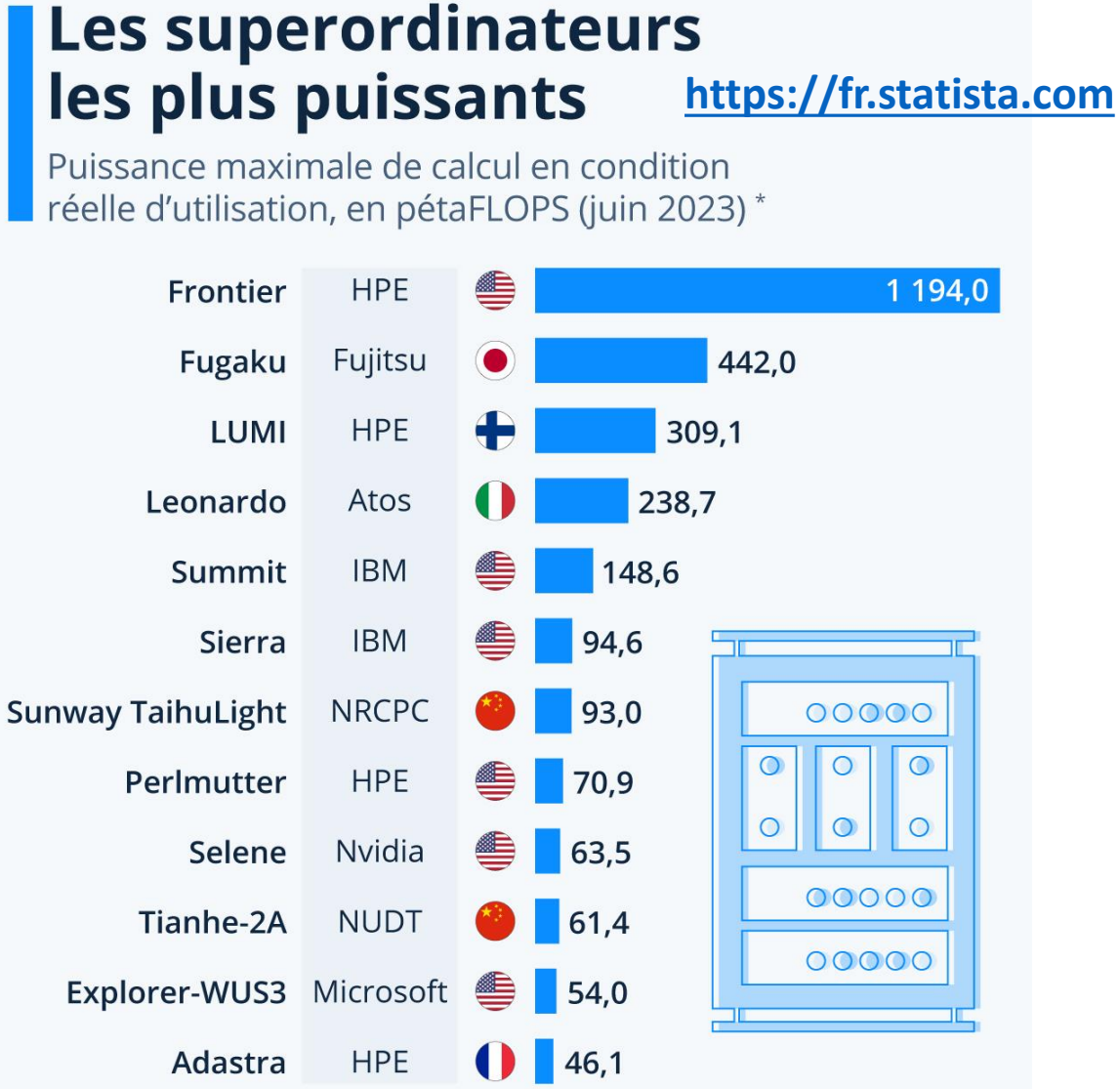
à l'origine de l'IA générative

❖ Puissance des ordinateurs

2000: 3 gigaflops (3. 10^9 op./sec)

2023: [Frontier](), 1er ordi exaflops (10^{18} op./sec)

Au-delà,
[l'ordinateur quantique]()



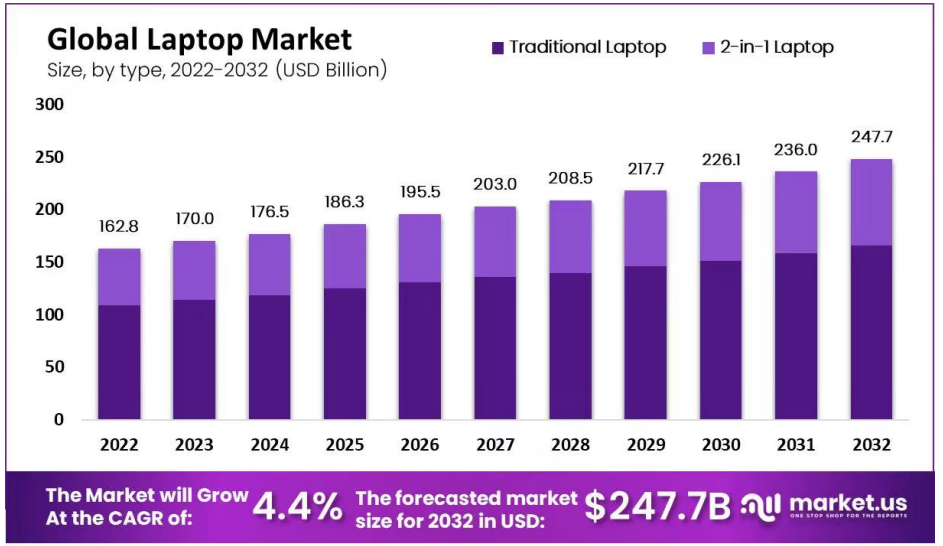
❖ Des secteurs en croissance rapide

Classement des ventes de smartphones en 2024

	MONDE	FRANCE
Apple	27,7%	24,8%
Samsung	23,6%	34,0%
Xiaomi	11,8%	15,2%
Oppo	5,9%	3,9%
Vivo	5,4%	3,9%
Autres,	7,2%	4,8%

En 2023:

- 289 M de **smartphones** ont été vendus
- **4.3 Md** de personnes possèdent un smartphones (oct 2023), soit 54% de la population mondiale



En 2023:

- 188 M de **laptop** ont été vendus (248 Md$)
- Apple (17%), Lenovo (15%), HP (12%)
- Le nombre de laptop en service est de **2Md** (équivaut à 25% de la population mondiale)

IoT: le marché devrait atteindre **40 Mds** d'appareils connectés d'ici **2025**

❖ Les Données: l'or noir de l'IA

L'IA générative est dépendante des données:

IA gen utilise des modèles qui nécessitent une **« formation »** pour aboutir à la solution du problème.

Cette formation se fait grâce aux **données** que l'on lui injecte.

En 2023, le volume de données mondiales échangées est estimé à **64.2 zettaoctets (*)**, avec une progression qui double tous les 5 ans.

*(*) soit **30Md de Md de pages** échangées*

*A noter: **90% du volume des données en circulation ont moins de 2 ans d'existence.***

Sources: tout ce que nous consultons et échangeons via des terminaux digitaux:

- sites web et cookies, réseaux sociaux via smartphones, portables,

Les objets connectés (écouteurs, montres, domotique, capteurs, satellites,…) génèrent 10% des données

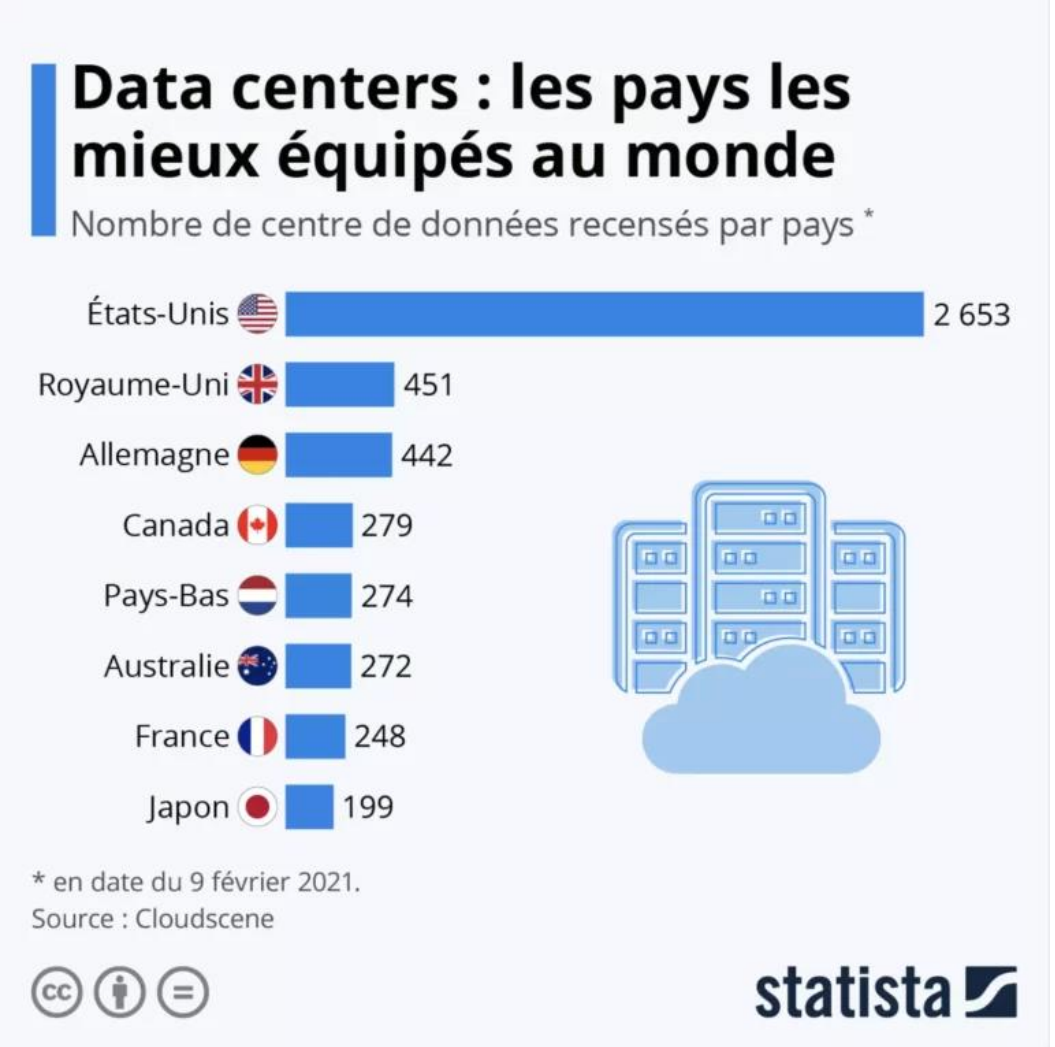
❖ Les données: **le rôle clé des datacenters**
Ils offrent les infrastructures de stockage et traitement des données

AWS - Amazon (18%)

RACKSPACE (US, 5%)

HETZNER (D, 5%)

OVH (FR, 4%)

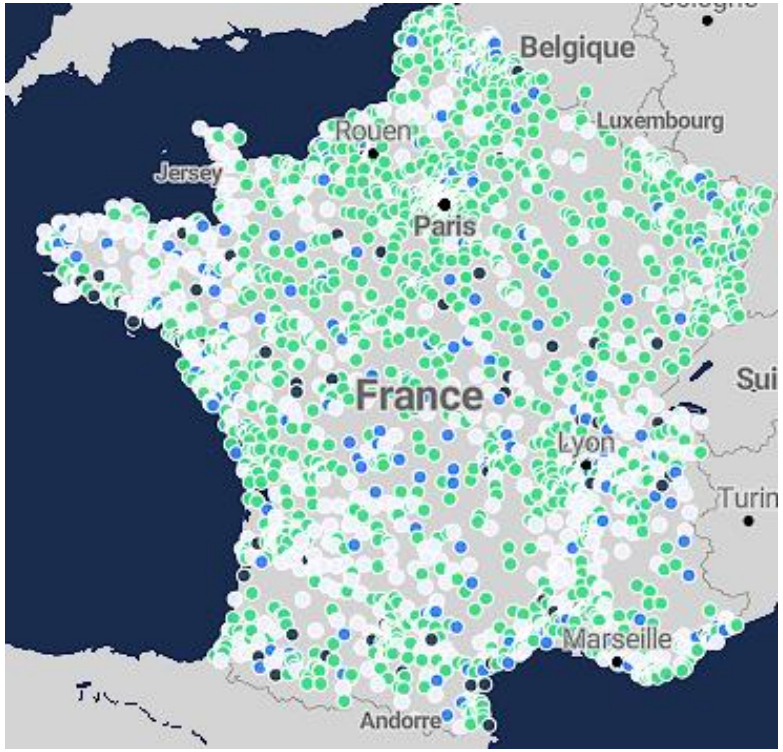


❖ **Internet et les réseaux** jouent un rôle clé dans la circulation des données et l'accessibilité à l'IA gen

Déploiement en France

- réseaux fixes: fibres
- réseaux mobiles : 4G, 5G (2020-2025)
- réseaux IoT : LTE-M (Orange, Bouygues), NB-IoT (SFR), wifi, bluetooth…

en France métrpolitaine (10/2024)
83.1% des logements
sont raccordés à la fibre



Au niveau mondial, on estime à **6 Md**
Le nombre des personnes **connectées**

La recherche

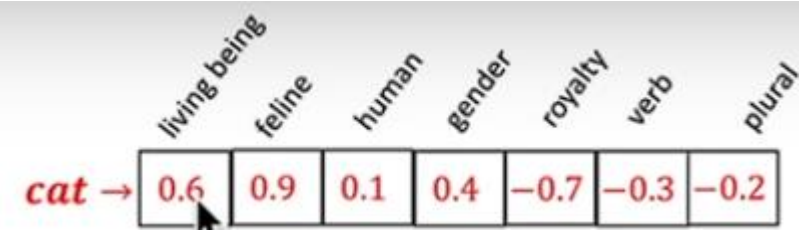
Elle contribue aussi à cette convergence

> algorithme et math
> physique quantique,
> neuroscience, biologie, génétique…
> robotique…

Comment une machine

peut-elle comprendre votre requête?

❖ Comprendre une requête



1ère étape « Comprendre » les mots (2013)

> ***représentation numérique*** des mots

(un vecteur de nombres, qui capture les informations sémantiques et syntaxiques du mot)

> on ajuste cette représentation par ***apprentissage*** grâce aux données disponibles

> les **mots « reliés »** entre eux ont des représentations similaires

2ème étape Positionnement des mots dans une phrase (2013)

« ***La voiture dépasse le camion***» est différent de « ***Le camion dépasse la voiture***»

3ème étape « Comprendre » les mots dans leur contexte (2017)

> mesure de la ***relation*** entre les mots pour capturer le sens de la phrase

> relation : similarité, ressemblance, association

Ex: « Je me promène le long de la côte.»
« côte » peut avoir plusieurs sens, on différencie grâce au contexte

Comment une machine

peut-elle répondre à votre requête?

❖ Répondre à une requête

L'opération prend en entrée **la requête « comprise »**

> il **génère la réponse** à la requête, a priori la plus **probable**

Cette génération se fait selon un **processus itératif :**

> à chaque itération, le système prédit le mot (ou groupe de mots) suivant,

> il se base pour cela sur la requête et les données disponibles

Exemple: à un stade de la réponse, il a écrit: « **Paris est la...** »

Son corpus de données lui propose:

- "Paris est la capitale de la France." (très probable, compte tenu de la requête)
- "Paris est la ville de l'amour." (probable, cliché associé à Paris)
- "Paris est la ville lumière." (probable, autre cliché)

En fonction de la requête, il peut choisir différentes stratégies:

> choix du mot avec la probabilité la plus élevée

> choix aléatoire, pour obtenir des résultats plus créatifs et variés…

Les outils

❖ Accéder aux principaux outils (ordinateur)

Outils	Sites *(officiels)*	Procédure d'accès
ChatGPT-4o	https://chatgpt.com/ *ou sous windows à partir du Microsoft Store*	Sign up: pour s'inscrire et créer un compte openai Log in : pour se connecter
Gemini 1.5 Flash	https://gemini.google.com	Il faut disposer d'un compte google
Claude 3.5 Sonnet	https://claude.ai	Il faut ouvrir un compte en renseignant les rubriques

Accès élargi via d'autres sites (US)
POE (https://poe.com) qui propose un large éventail de produits.

Les requêtes

❖ Echanger avec un ChatGPT: **pourquoi faire?**

> pour rechercher des **informations**

(c'est la base)

> pour réaliser des **tâches**: articles, traduction, résumé, codage, création (jeux, site web, graphisme, sons…),…

(c'est le plus)

> pour effectuer des **analyses** sur des documents, traiter des données, gagner du temps, améliorer la productivité…

(cadre entreprise)

…..

❖ Echanger avec ChatGPT: **comment?**

> Généralement en tapant une **requête**(*) écrite, en attendant une réponse

> Cette requête peut s'appuyer sur un fichier (**image, tableau…**) que l'on **télécharge**.

à partir d'un ordinateur ou d'un smartphone,
on peut aussi utiliser le **micro** pour exprimer **oralement** sa demande.

(*): on parle aussi d'invite ou de prompt

❖ Echanger avec ChatGPT: **comment?**

Pour **améliorer** la réponse souhaitée, pour **compléter** votre demande

> on peut l'utiliser en mode **« conversation »**, via des requêtes successives

Exemple:

1/ Prépare un mail à …pour…

2/ Sois plus concis dans ta réponse,…

❖ Chat: **limites d’utilisation**

Nb de requêtes / jour, selon le type de requête….(version gratuite)

Requêtes concernant la sécurité, l’éthique…

Requêtes proscrites

> pour générer du contenu illégal, haineux ou offensant.
> pour diffuser de fausses informations.
> à des fins commerciales ou professionnelles.

Requêtes trop complexes.

Les fournisseurs se réservent le droit de modifier les limites d'utilisation à tout moment.

❖ Optimiser la requête

Pour **améliorer la réponse, Il faut guider l'IA**

L'importance des mots
L'IA s'appuie sur les mots importants, ne pas « diluer » la demande

Orienter vers un type de **modèle**
Utiliser des termes comme prédire, classer…

Orienter le choix des **données**
Préciser le domaine, le cadre…

L'importance du contexte
L'IA recherche le contexte pour « comprendre » les mots, Il faut le guider

Gérer la complexité
L'IA sait bien gérer une tâche, moins bien plusieurs tâches

Améliorer le rendu
en précisant le ton, le style et le format de sortie

❖ Requêtes: **l'importance des mots** (1/6)

Allez droit au but

> Évitez les fioritures, les formules de politesse… qui génère du « bruit »

Vous seriez bien aimable d'avoir l'obligeance, éventuellement…

> Soyez clair, concis, précis et directif *(utiliser des **verbes d'action** : expliquer, analyser)* éviter les généralités, bien délimitée les tâches à accomplir.

Parlez-moi des voitures

Décrivez les caractéristiques des voitures électriques, par rapport aux voitures à essence classiques

> Employez des directives affirmatives, éviter les termes négatifs

Comment les bâtiments restent-ils stables lors des tremblements de terre ?

> **Répétez** le mot (important) ou une phrase spécifique plusieurs fois

L'évolution, en tant que concept, a façonné le développement des espèces. Quels sont les principaux moteurs de l'évolution et comment l'évolution a-t-elle affecté les humains modernes ?

❖ Requêtes: **le choix du modèle** (2/6)

Préciser le type de tâche: *Cette tâche concerne….ou Il s'agit de …*

Pour orienter l'IA vers le choix de l'algorithme le plus approprié

> **Prédire** une valeur, une donnée, par corrélation, relation

Prix recherché par l'agent immobilier

> **Classer** des images, des mots ou phrases…

Analyse de texte, classement de vins selon leur qualité

> **Regrouper**, rechercher des affinités

Très utilisé en marketing, pour définir le comportement d'un client

> **Aider** à la décision

Arbre de décision pour l'analyse de données

> **Générer** de nouvelles entités

Pour la production audio-visuelle, pour générer de nouveaux matériaux

> **Atteindre** un objectif

Système de recommandation (Amazon)

❖ Requêtes: **le choix des données** (3/6)

> Préciser le domaine, le cadre

Pour sélectionner un corpus de données spécifique et guider l'IA

> **Plus explicitement**, lui demander d'utiliser **un corpus de données spécifiques, à jour…**

Utilise des articles publiées récemment par CNN…

> Donner à l'IA un rôle

*Vous êtes un **économiste expert**, répondez à cette question: quelles sont les principales différences entre un système économique capitaliste et socialiste ?*

*En tant qu'**astrologue**, dites-moi quel est mon signe astrologique et mon ascendant : Je suis né le* [date à heure], à [lieu]

[À partir d'une image d'un plat} *En tant que **chef cuisinier**, dites-moi les étapes de création de ce plat*

*Vous êtes **Jeff Bezos**, suggérez 3 stratégies pour développer une entreprise de commerce électronique.*

❖ Requêtes: **l'importance du contexte** (4/6)

Préciser le contexte

> Rédiger une invite **détaillée**, en rajoutant des informations claires et spécifiques sur **la situation**, pour la tâche qui lui est demandée

Rédigez un avis sur ce produit.
Rédigez une critique de ce produit en mettant l'accent sur ses performances pour les activités de plein air.

> **Ajuster** le niveau de clarté, de compréhension recherché

Explique -moi comme si j'avais 11 ans, en termes simples, comme si j'étais un débutant/expert en [domaine], rédige [un texte] en utilisant un français simple…

> **Préciser** quel est le public visé

Décrire le fonctionnement des smartphones, à destination des seniors qui n'en ont jamais utilisé.

❖ Requêtes: **l'importance du contexte** (4/6)

Le **guider**, pour une réponse plus précise et plus pertinente:

> Fournir des **exemples** (factuels, documentés) , ex en téléchargeant un article sur le sujet

> Fournir une expérience, une opinion personnelle, des instructions.

L'IA apprend de ces exemples et intègre ce qu'il a appris au contexte

[Décrire vos symptômes]: fournir un diagnostic possible et expliquez pourquoi, s'inspirer des ex. suivants

Exemple 1: Symptômes : fièvre, toux, fatigue
Diagnostic : rhume
Explication : la combinaison de fièvre, de toux et de fatigue est typique d'un rhume. Aucun symptôme grave n'est présent, ce qui suggère une infection virale légère.

Exemple 2: Symptômes : douleur thoracique, essoufflement, étourdissements
Diagnostic : crise cardiaque possible
Explication : la combinaison de douleur thoracique, d'essoufflement et d'étourdissements sont des signes avant-coureurs d'une possible crise cardiaque. Une attention médicale immédiate est requise.

Exemple 3: Symptômes : maux de tête, sensibilité à la lumière, nausées
Diagnostic: migraine….

❖ Requêtes: **gérer la complexité** (**5**/6)

Plutôt que d'écrire des requêtes complexes, avec trop de paramètres, trop d'attendus, utiliser la capacité de ChatGPT au **mode « conversation »**, ex:

> **Approche progressive**: décomposer en étapes successives simples, pour réaliser une tâche. Utiliser la réponse comme entrée pour l'invite suivante

Ex: préparer un voyage

1. *définir des destinations pour une plage de dates*
2. *En fonction du résultat, sélectionner les modes d'accès appropriés*
3. *Idem, choisir un type de résidence*
4. *Idem, choisir des prestations, des excursions…*

> **Interagir pour clarifier et affiner votre requête** -> guider l'IA et améliorer sa réponse

Après une 1ère réponse > identifier et éliminer les ambiguïtés potentielles dans votre invite
> revoir ce que vous avez écrit, peaufiner votre requête, demander des précisions

❖ Requêtes: **gérer la complexité** (5/6)

> **Réfléchir avant d'agir**: on ajoute des questions intermédiaires, nécessitant des réponses qui seront rajoutées à son contexte, ce qui l'aide à mieux « comprendre » la demande et à obtenir un résultat plus satisfaisant.

Peux-tu développer une intrigue d'histoire d'amour

Peux-tu développer une intrigue d'histoire d'amour en décrivant les événements clés, les motivations des personnages et les relations de cause à effet.

> **Améliorer sa mémoire**: ChatGPT apprend des interactions que vous avez avec lui, ce qui lui permettra d'utiliser ces informations dans les discussions ultérieures, et vous répondre plus personnellement.

Pour l'aider à créer cette base de connaissances plus rapidement :

As-tu des questions pour mieux me connaître et approfondir tes connaissances à mon sujet ? Pose-moi tout ce que tu trouves important, une seule question à la fois et attends ma réponse avant de continuer.

❖ Requêtes: **indiquer le rendu souhaité** (6/6)

Soigner le ton, style et format de la réponse

> **Préciser vos exigences :**

- longueur, structuration (liste avec puces, sur une page, sous forme de tableau…)
- restrictions (*contenu sûr, approprié et non nuisible …)*
- directives spécifiques (nb d' options à proposer), mots-clés,

** Créez une todo liste pour des vacances à la plage, comprenant les mots-clés suivants" crème solaire ", " maillot de bain " et " serviette de plage " comme éléments essentiels.*

> **Donner** un style de contrôle *(ie convivial, professionnel, humoristique…)*

** Écrivez un paragraphe sur une alimentation saine. Répondez d'une manière naturelle et humaine.*

** Comment les origines culturelles influencent-elles la perception de la santé mentale ? Assurez-vous que votre réponse est impartiale et évite les stéréotypes*

❖ Requêtes: **l'ordre est important**

1. Commencer par le **rôle** et le **contexte**
2. Décrire la **tâche** (commencer par un verbe d'action)
3. Fournir des **exemples** (améliore la qualité du résultat de l'IA)
4. Préciser les **contraintes**, les limites à respecter (évite que l'IA ne s'écarte du sujet)
5. Indiquer le **format** de présentation du résultat
6. Préciser le **ton** que vous souhaitez

Perspectives

pour l’IA générative

❖ Perspectives sur les outils

Les recherches visent à **répondre aux contraintes** liées au développement d'outils de + en + puissants et à **améliorer leur performance, leur fiabilité:**

- **Rendre le modèle plus frugal en énergie** et en **consommation de données**
- **Réduire les coûts**, en particulier d'apprentissage
- **Permettre à l'IA de s'adapter plus facilement** à de nouvelles situations (capacité à généraliser)
- **Explicabilité** (confiance, fiabilité) : rendre la réponse explicable
- **Fiabilité** : le modèle doit faire preuve d'humilité s'il ne sait pas

❖ **Consommation énergétique:** ex ChatGPT-4

> pré-formation du modèle: **50 GWh**

> réponse aux requêtes des utilisateurs : **500 MWh par jour**
Soit l'empreinte énergétique d'une ville de taille moyenne.

Prédiction à l'horizon 2026 pour l'IA gen : **1000 TWH par an**
Soit la **consommation annuelle d'énergie du Japon**

Cela n'est pas viable: nécessité de développer des solutions plus frugales en énergie:

> CRAM (Computational Random-Access Memory)
> Combine mémoire et puissance de traitement
> On divise par **1000** la consommation
> Encore au stade de développement

❖ Perspectives sur les outils: **nature de coûts des LLMs**

Chaque étape a un coût:

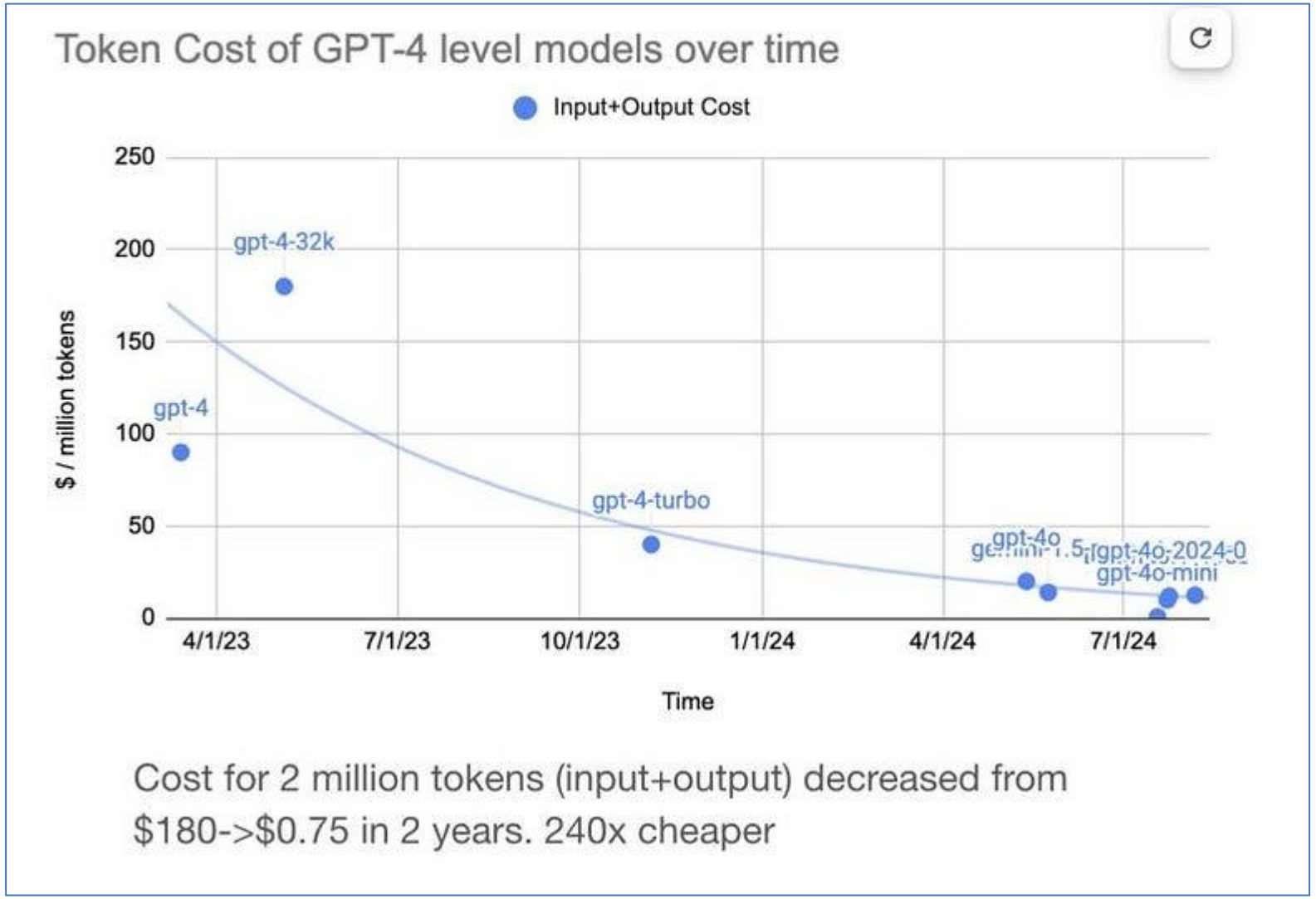
> préapprentissage du modèle > **coût fixe initial**

> compréhension de la requête et réponse > **coût à chaque invite**

L’objectif des acteurs est bien sur de réduire ces coûts

❖ Coût

Traitement
de la requête



L’innovation a permis de créer des modèles plus petits, moins chers et tout aussi puissants

❖ Coût

Cependant, les coûts de formation et d'exploitation restent élevés

D'après **The Information**,

> OpenAI a dépensé plus de 7 Md$ en formation à l'IA et 1,5 Md$ en personnel,

> L'exploitation de ChatGPT, coûte plus de 700 000 dollars par jour.

OpenAI pourrait afficher 5 Md$ de déficit fin 2024

Implications

liées à son déploiement

❖ Implications

L'émergence de l'IA gen fait craindre des bouleversements qui toucheraient plusieurs secteurs clés.

- **L'emploi, l'économie et le sociétal**

  > *évolution des rôles professionnels vers les services*

  > la (perte de la) prise de décisions

https://ai.gopubby.com/the-decision-dilemma-embracing-ai-in-an-imperfect-world-b3b0728a244e

- **L'éducation, pour une adaptation face à une évolution rapide**

  « *En tant que responsable national de l'Éducation des Jeunes, pour mieux éduquer et préparer les jeunes à leur entrée dans la vie active, dans un monde où l'IA générative aura pris une place prédominante, propose des stratégies d'adaptation concernant les méthodes et les contenus scolaires ».*

  > *formation continue*

- **Les considérations éthiques, la confidentialité et la protection des données**

  > *garantir un traitement juste, transparent et responsable > pb de réponse avec contenu préjudiciable*

  > *utilisation des données de manière responsable et respectueuse > pb des droits d'auteurs*