20 CARTES POUR ABORDER L'IMPACT ÉNERGÉTIQUE DE L'IA GÉNÉRATIVE

conseil national du numérique

TOKEN:

Les requêtes (prompts) formulées par les utilisateurs sur des outils d'IA conversationnelles sont découpées en jetons (dits token) pour être traitées par le modèle d'IA.

Un token représente une séquence de caractères considérée comme commune. Un token peut alors correspondre à un ou plusieurs caractères.

Bonjour
1 TOKEN

Bonjour, un café svp
5 TOKENS

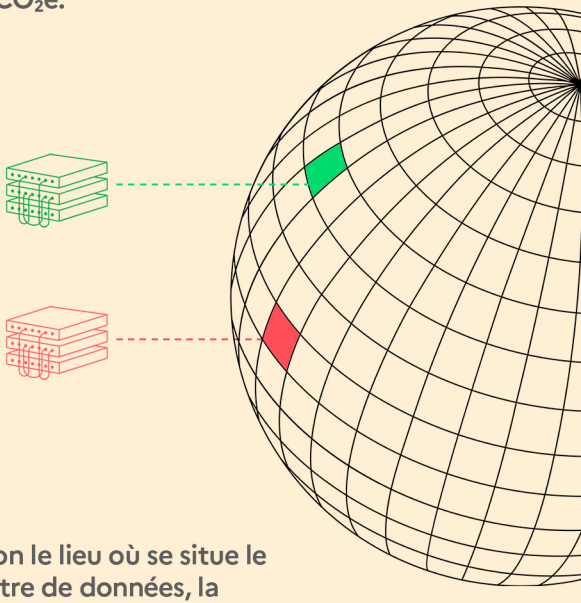
1 COURT PROMPT C'EST ENVIRON 50 TOKENS (250 CARACTÈRES)

EXEMPLE DE PROMPT (7 tokens) :
Quelle est la définition du silicone ?

EXEMPLE DE RÉPONSE (58 tokens) :
Les silicones, ou polysiloxanes, sont des composés inorganiques formés d'une chaîne silicium-oxygène dans laquelle des groupes se fixent sur les atomes de silicium. Certains groupes organiques peuvent être utilisés pour relier plusieurs de ces chaînes.

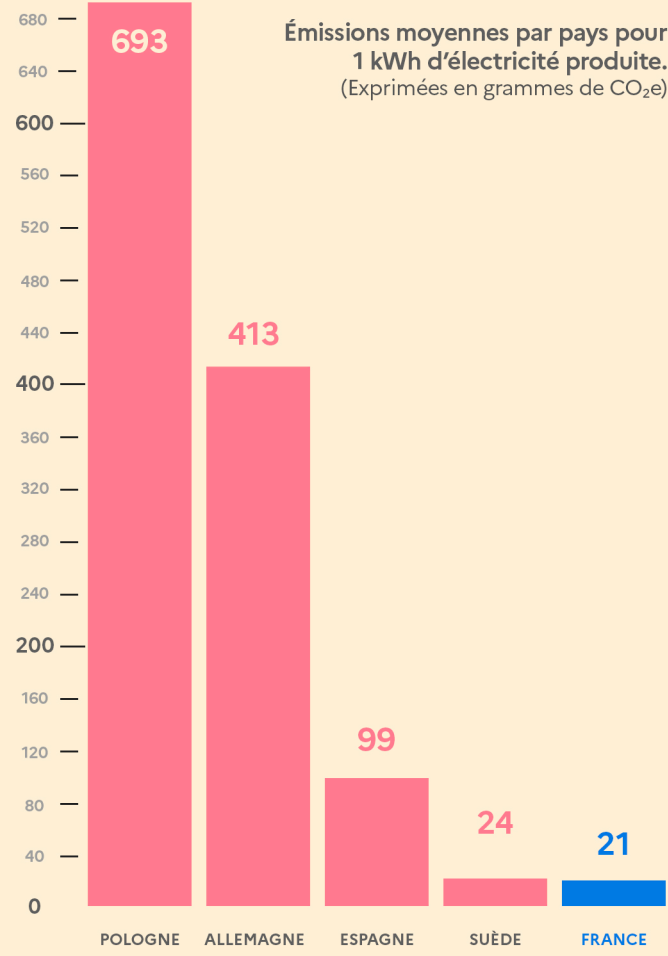
Les tokens n'ont pas le même impact à l'entrée et à la sortie. **Taper le prompt consomme très peu d'énergie** : Les quelques secondes qu'il faut pour l'écrire et l'envoyer utilisent juste un fragment de l'énergie de votre appareil (smartphone ou ordinateur).

Générer la réponse nécessite par contre de faire fonctionner les calculateurs d'un centre de données. C'est ce qu'on appelle la phase d'inférence. C'est cette phase qui nous intéresse pour estimer ou mesurer la consommation électrique et les émissions de CO_2e.

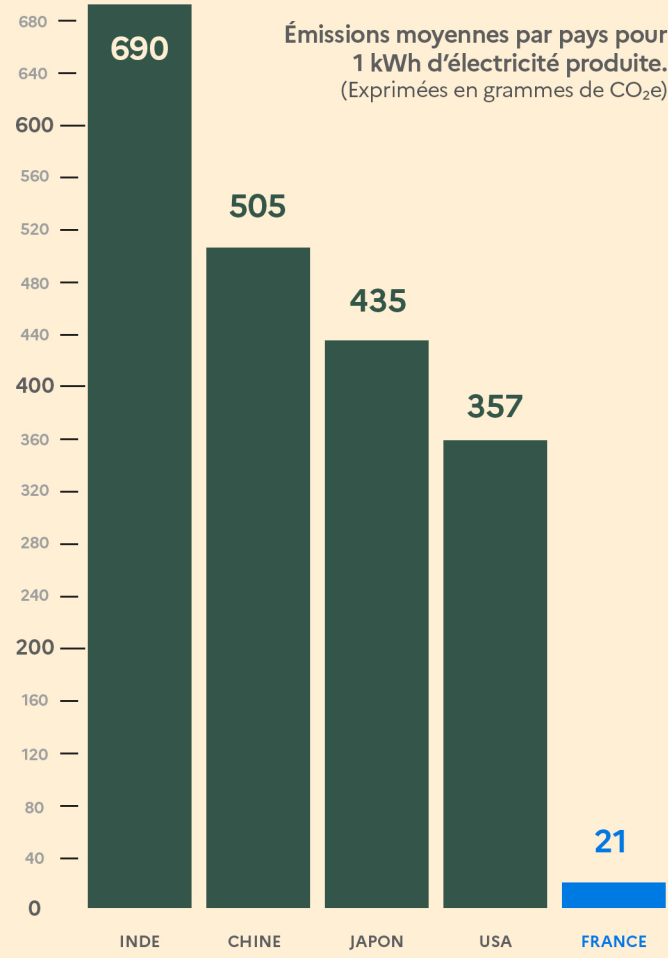


Selon le lieu où se situe le centre de données, la consommation et les émissions sont plus ou moins importantes.

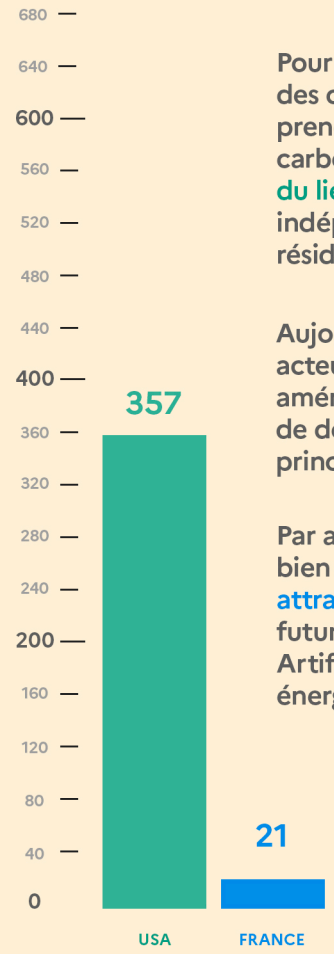
ÉMISSIONS DE CO_2e ET LOCALISATION DES CENTRES DE DONNÉES



ÉMISSIONS DE CO_2e ET LOCALISATION DES CENTRES DE DONNÉES



ÉMISSIONS DE CO_2e ET LOCALISATION DES CENTRES DE DONNÉES



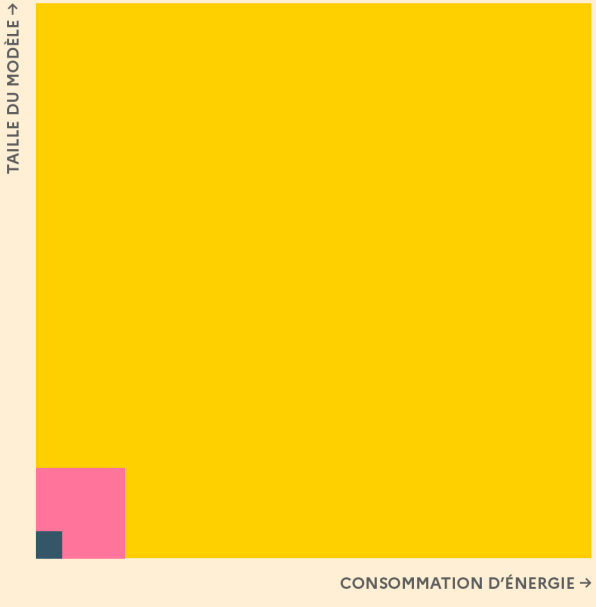
Pour mesurer l'impact direct des centres de données il faut prendre en compte l'intensité carbone du **mix électrique du lieu où ils sont situés,** indépendamment du lieu de résidence de l'utilisateur.

Aujourd'hui la majorité des acteurs sont des entreprises américaines. Leurs centres de données sont localisés principalement aux USA.

Par ailleurs, on comprend bien que **la France semble attractive** pour accueillir de futures usines à Intelligence Artificielle en raison de son énergie nucléaire décarbonée.

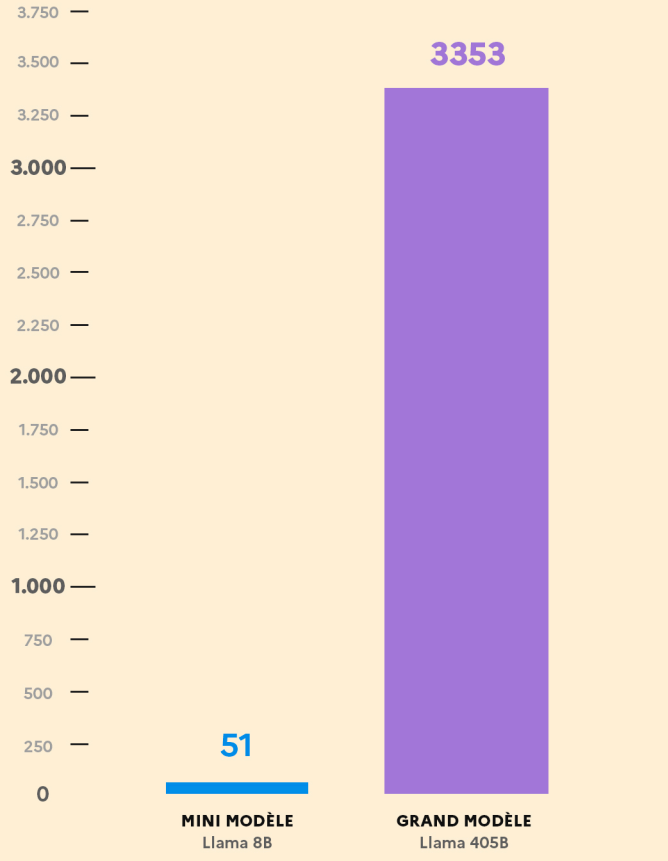
(Émissions exprimées en grammes de CO_2e par kWh d'éléctricité produite)

REPÈRE : DES DIFFÉRENCES GIGANTESQUES D'UN MODÈLE À L'AUTRE



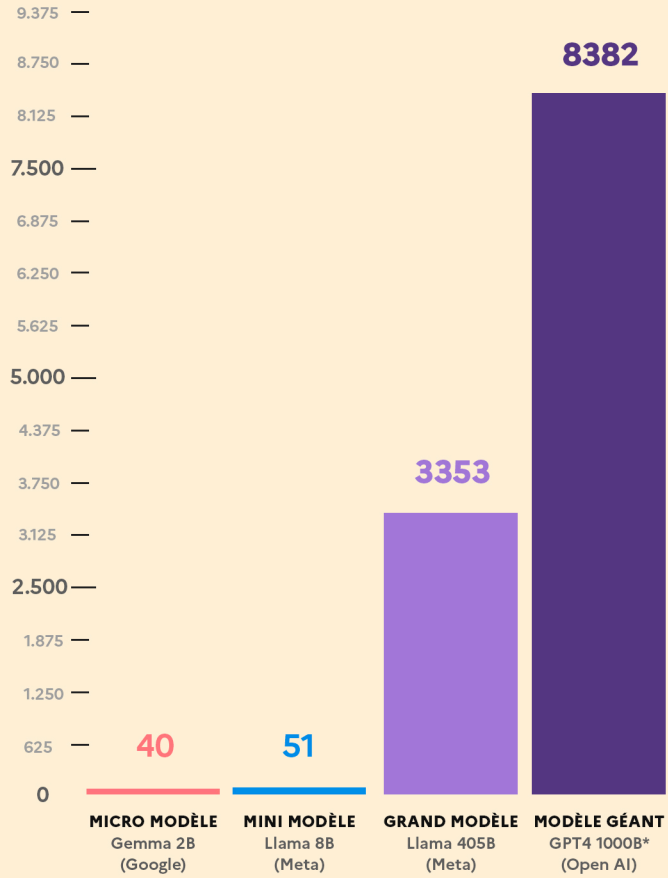
Au delà du lieu où sont situés les centres de données, la consommation d'énergie dépend aussi fortement du modèle utilisé. La différence peut-être vraiment immense d'un outil à l'autre.

Consommation énergétique moyenne pour un court prompt avec les modèles Llama de la société Meta.
(Exprimée en joules)

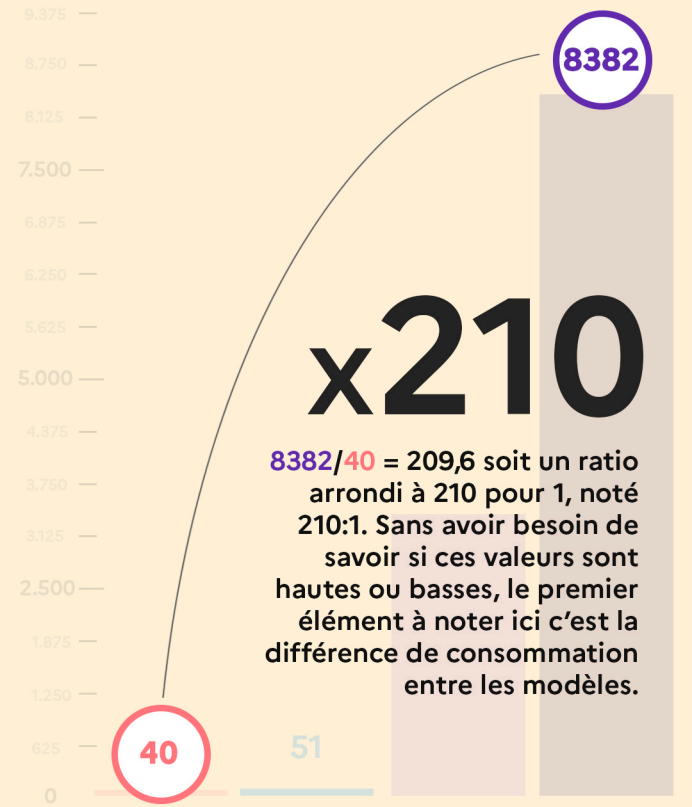


Consommation énergétique moyenne pour un court prompt, du modèle le plus frugal au plus énergivore.

(Exprimée en joules)



Consommation énergétique moyenne pour un court prompt, du modèle le plus frugal au plus énergivore.
(Exprimé en joules)

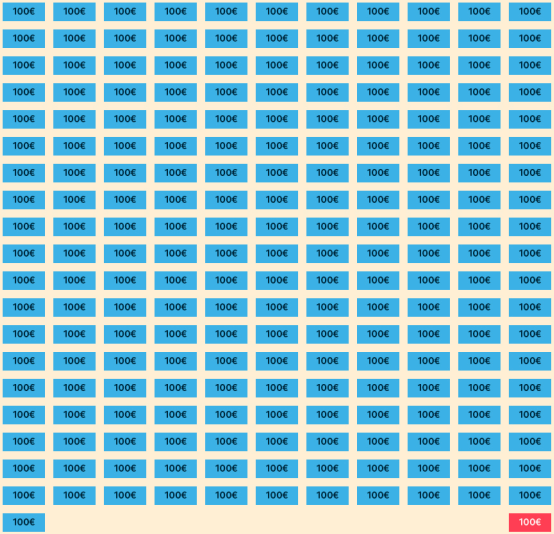


x210

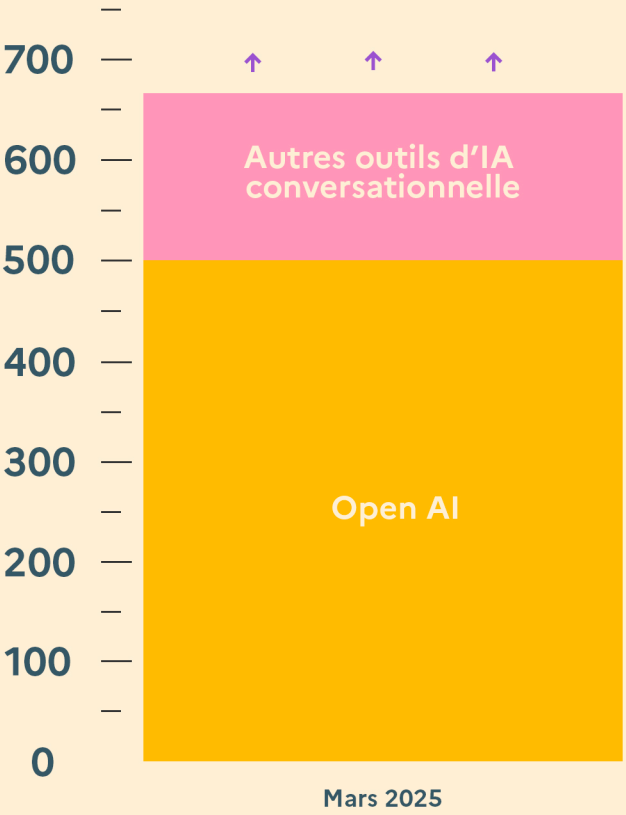
8382/40 = 209,6 soit un ratio arrondi à 210 pour 1, noté 210:1. Sans avoir besoin de savoir si ces valeurs sont hautes ou basses, le premier élément à noter ici c'est la différence de consommation entre les modèles.

210:1

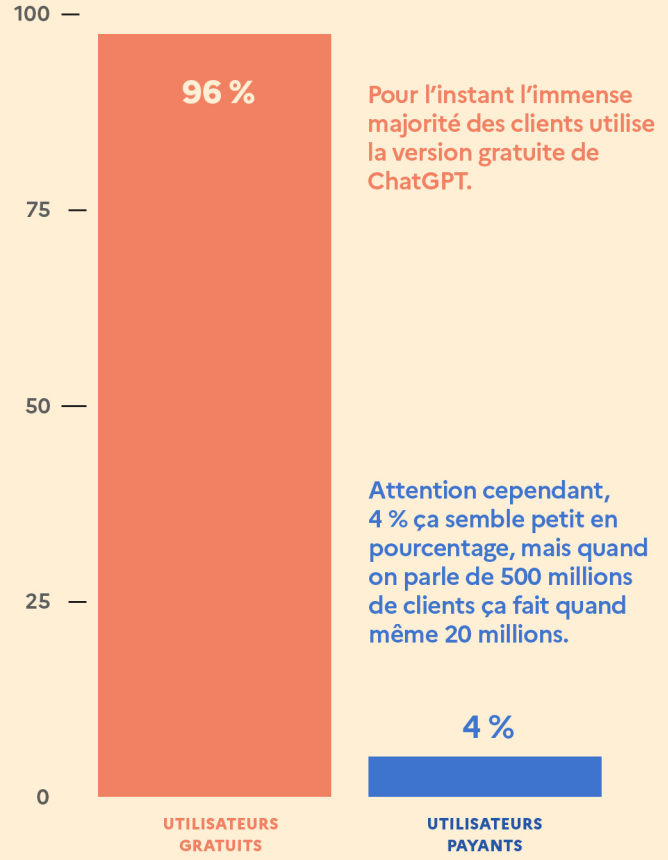
Un ratio de 210 pour 1 c'est gigantesque. C'est comparable à la différence entre 100 € et 21 000 €.



Nombre total d'utilisateurs d'outils d'IA générative, Open AI + reste du marché, mars 2025.
(Exprimés en millions)



Répartition entre utilisateurs gratuits et payants des outils Open AI au mois de mars 2025.
(Exprimés en pourcentage)



Pour l'instant l'immense majorité des clients utilise la version gratuite de ChatGPT.

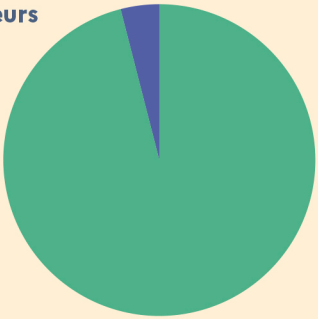
Attention cependant, 4 % ça semble petit en pourcentage, mais quand on parle de 500 millions de clients ça fait quand même 20 millions.

CONSOMMATION ÉLECTRIQUE LIÉE À L'USAGE D'OPEN AI PAR 500 MILLIONS D'UTILISATEURS

Répartition des utilisateurs

UTILISATEURS GRATUITS
480 Millions
(96%)

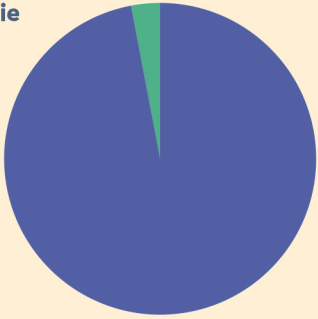
UTILISATEURS PAYANTS
20 Millions
(4%)



Consommation d'Énergie

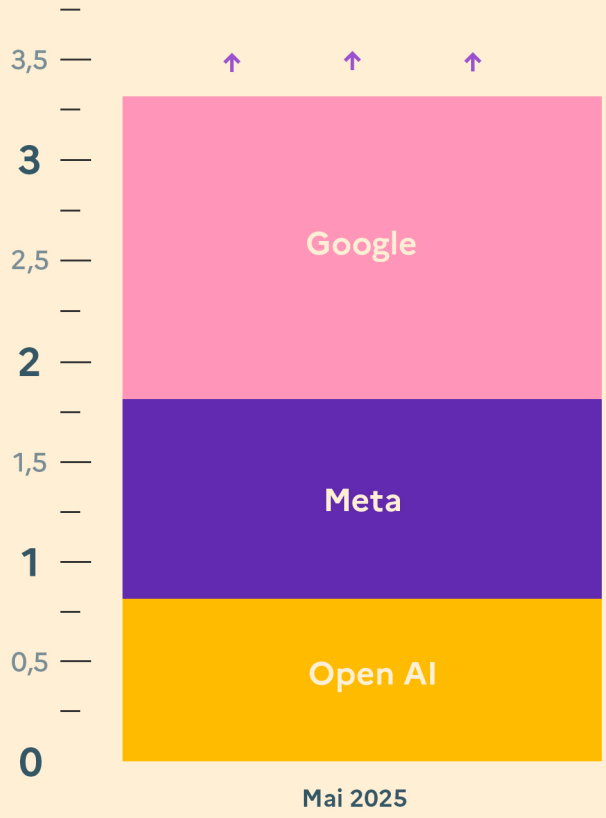
UTILISATEURS GRATUITS
20 GWh
(3%)

UTILISATEURS PAYANTS
680 GWh
(97%)



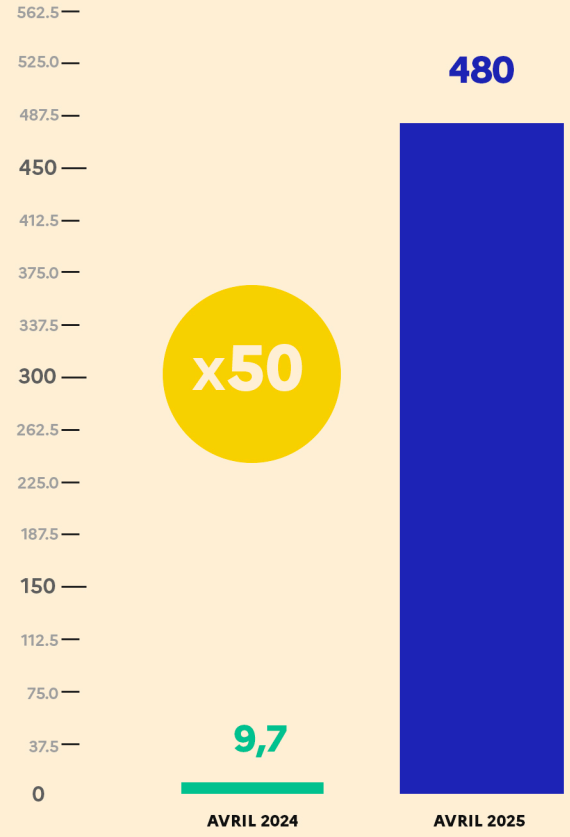
En projection annuelle, pour 500 millions de clients, le total de la consommation directe liée à l'usage des outils conversationnels d'Open AI avoisinerait déjà 700 GWh (estimation). 97% de cette energie est consommée par seulement 4% des utilisateurs.

Nombre total d'utilisateurs des outils d'IA générative de Google + Meta + Open AI, mai 2025.
(Exprimés en milliards)



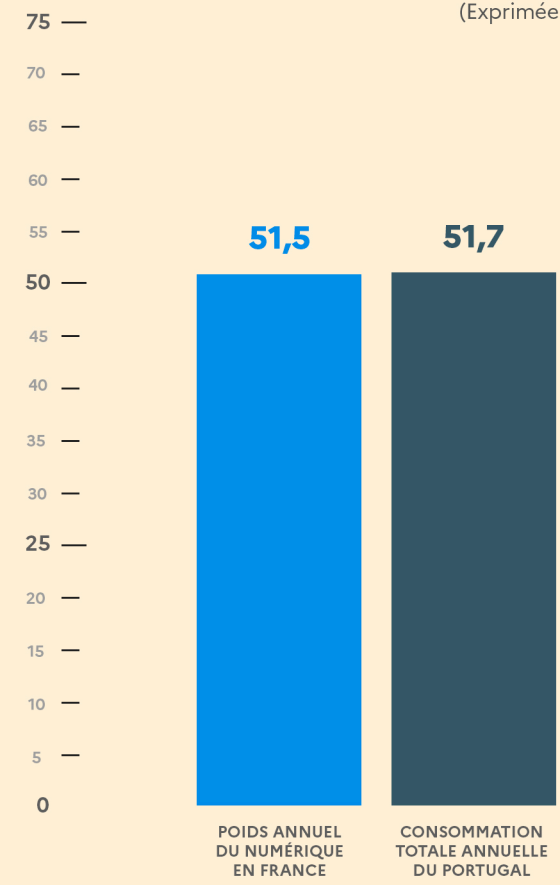
Evolution du volume de tokens générés par mois via Gemini, l'IA de Google.

(Exprimé en billions)

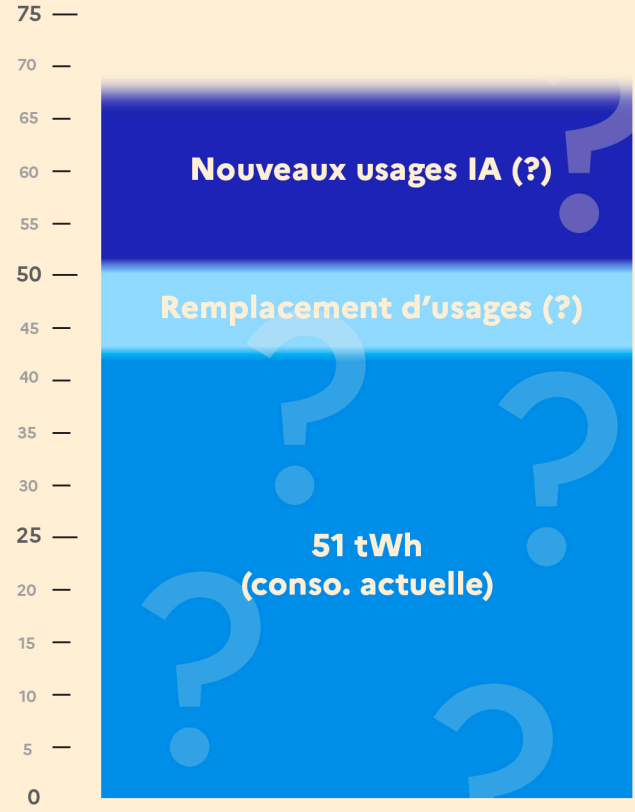


Consommation annuelle électrique du numérique en France comparée à la consommation annuelle totale d'électricité du Portugal

(Exprimée en tWh)



Quelle sera l'évolution de la consommation annuelle électrique du numérique en France en incluant l'IA ?
(Exprimée en tWh)



IL N'EXISTE PAS DE MODÈLE FIABLE À L'HEURE ACTUELLE CAPABLE DE SIMULER LA PROGRESSION À VENIR DES USAGES LIÉS À L'IA, NI DE SA PERFORMANCE ÉNERGÉTIQUE (ÉVOLUTION RAPIDE).

CARTE 1, 2 : Le comptage des tokens présentés dans les cartes 1 et 2 est réalisé avec le simulateur « Tokenizer » d'Open AI.

CARTE 5, 6, 7 : Ces moyennes sont établies à partir des données d'Electricity Maps qui mesure l'intensité carbone en équivalent CO_2 (CO_2e). Elles couvrent la période [mai 2024 - avril 2025].

CARTE 9, 10, 11 : Les données chiffrées s'appuient sur ce dossier de la MIT Technology Review publié en mai 2025. La MIT Tech. Rev. a collaboré avec une équipe de l'Université du Michigan pour réaliser leurs mesures en testant plusieurs modèles sur 2 types de GPU. Les résultats sont accessibles ici.

Par ailleurs, la MIT Tech. Rev. s'appuie également sur ces travaux d'une équipe de recherche de Microsoft pour appliquer un facteur multiplicateur (x2) aux résultats de leurs mesures (sans entrer dans les détails, il s'agit de prendre en compte la chaîne des autres équipements qui tournent dans le processus complet).

Sur la carte 10, le score énergétique attribué à Open AI est une estimation sur l'hypothèse que GPT4 compte 1 000 milliards de paramètres (Open AI ne publie pas les informations). Cependant c'est un chiffre qui semble faire consensus parmi les experts. C'est l'ordre de grandeur que donne la MIT Tech. Rev. dans son dossier.

CARTES 13, 14, 15 ET 16 : Les données concernant les utilisateurs sont extraites des Keynote Meta 2025 et Google I/O '25 Keynote ainsi que de données communiquées directement par Open AI.

Les données de consommation d'énergie de la carte 15 proviennent d'un tableau réalisé par la Conseil national du numérique, (disponible en libre téléchargement ici) basé sur les travaux de la MIT Tech. Rev. précités et proposant différentes hypothèses d'usages. C'est le scénario moyen qui est utilisé.

CARTES 17 : Les données Google sont extraites de la « Google I/O '25 Keynote » de mai 2025. Ce sont les chiffres présentés et commentés directement par Sundar Pichai (CEO Google).

CARTE 18 : Les données proviennent d'ici pour la France (2022) et du Figaro pour le Portugal (chiffres 2024)

20 CARTES POUR ABORDER L'IMPACT ÉNERGÉTIQUE DE L'IA GÉNÉRATIVE

Éditeur : Conseil national du numérique, juin 2025.
© Licence libre Creative Commons CCBYSA.