De la requête visible à l'inférence invisible

Consommation énergétique de l'IA générative :
modélisation, projections et émergence des usages non perçus

Mars 2026

Sources : Consensus Academic Search, IEA, Goldman Sachs Research,
Google Environmental Report 2024, Microsoft AI Economy Institute, OpenAI

Résumé

La consommation énergétique de l'intelligence artificielle générative fait l'objet d'un discours dominant focalisé sur la requête individuelle. Cet article démontre, par une modélisation physique à partir d'un datacenter de référence de 1 GW, que la valeur médiane publiée par Google (0,24 Wh par requête Gemini) est incompatible avec les volumes d'usage réels et conduit à des absurdités arithmétiques. Une estimation réaliste de 1,8 à 4,2 Wh par requête est établie par calcul inverse. L'application de ce modèle aux projections de capacité 2030 – entre 200 et 400 TWh dédiés à l'inférence IA (IEA-4E) – révèle que les équivalents-requêtes par utilisateur actif dépasseraient 330 par jour, soit un facteur 33 à 66 supérieur à l'usage conversationnel actuel. Cette capacité ne peut être absorbée que par des usages dans lesquels l'humain n'est plus l'initiateur direct : agents autonomes, intégration

systémique, chaînes de traitement automatisées. Nous proposons une taxonomie de ces usages « invisibles », et nous montrons que leur émergence rend structurellement insuffisantes les approches de régulation fondées sur la sobriété individuelle.

Table des matières

1 Introduction

La formule selon laquelle « une requête IA consomme très peu » circule abondamment dans la presse généraliste et dans les communications institutionnelles des grandes entreprises technologiques. Elle s'appuie sur des estimations qui ont une cohérence interne dans leurs hypothèses les plus favorables, mais qui résistent mal à un examen croisé avec les données de capacité d'infrastructure et de volume d'usage réel.

La littérature académique récente permet de structurer une réponse rigoureuse. Samsi et al. (2023) ont conduit les premières évaluations comparatives systématiques de la consommation d'inférence des Large Language Model (grand modèle de langage)s (LLMs) à grande échelle, sur des configurations allant jusqu'à 32 Graphics Processing Unit (processeur graphique) (GPU) en parallèle. Fernandez et al. (2025) ont démontré que les estimations basées sur les Floating Point Operations per Second (opérations en virgule flottante par seconde)s (FLOPss) théoriques ou l'utilisation nominale des GPUs sous-estiment significativement la consommation réelle en production, en raison de la géométrie des charges de travail, de la pile logicielle et de l'architecture matérielle effective. Chien et al. (2023) ont établi que pour des services de type ChatGPT, l'inférence produit en un an 25 fois plus d'émissions carbone que l'entraînement initial du modèle.

Cet article propose un exercice de modélisation à rebours : à partir d'un datacenter IA de 1 GW – ordre de grandeur correspondant aux grandes installations planifiées – nous dérivons les hypothèses physiques nécessaires pour parvenir à une estimation de la consommation par requête cohérente avec les volumes d'usage observés. Nous appliquons ensuite ce modèle aux projections de capacité mondiale à l'horizon 2030 et tirons les conséquences en termes de nature des usages futurs.

2 Cadre méthodologique et hypothèses

2.1 Note sur les sources

Cet article mobilise trois niveaux de sources, dont la fiabilité diffère :

1. **Publications évaluées par les pairs** (IEEE HPEC, AAAI, ACM) – résultats de recherche vérifiés et reproductibles.
2. **Rapports institutionnels** (Agence Internationale de l'Énergie (International Energy Agency) (AIE), IEA-4E, Deloitte, Goldman Sachs) – méthodologie explicitée,

mais non soumise au processus de relecture académique.

3. **Données auto-déclarées par les opérateurs** (Google, OpenAI, Meta) et articles de presse spécialisée – données primaires non vérifiées par un tiers, utilisées faute de meilleures sources publiques.

Les données de capacité et de volume qui fondent le modèle relèvent principalement du troisième niveau. Cette dépendance aux déclarations des opérateurs est elle-même un symptôme du problème d'opacité discuté en section 6.

2.2 Chaîne causale du modèle

La conversion d'une puissance installée en un nombre de requêtes par utilisateur passe par une chaîne de six variables indépendantes. Chacune introduit une incertitude qui se propage multiplicativement. Nous les présentons dans l'ordre logique de la dérivation.

2.3 Puissance installée (H1)

Nous posons comme hypothèse de départ une puissance installée de 1 GW, correspondant aux grandes installations annoncées ou en construction par les hyperscalers en 2024–2025 (McKinsey & Company 2025). Google, Microsoft et Meta ont toutes annoncé des campagnes en Europe et aux États-Unis dépassant ce seuil par campus.

2.4 PUE – Power Usage Effectiveness (H2)

Le Power Usage Effectiveness (efficacité énergétique d'un datacenter) (PUE) exprime le rapport entre l'énergie totale consommée par un datacenter et l'énergie effectivement utilisée par les équipements informatiques. La puissance IT utile est :

$$P_{\text{IT}} = \frac{P_{\text{installée}}}{\text{PUE}} \tag{1}$$

Deux valeurs de référence sont utilisées :

- **PUE = 1,09** : flotte mondiale Google en 2024 (Google 2024), valeur également rapportée pour Meta (Meta Platforms 2024). Représentatif des hyperscalers les plus optimisés.

- **PUE = 1,56** : moyenne mondiale selon l'enquête Uptime Institute 2024 (Uptime Institute 2024), représentatif de l'industrie au sens large.

Ces deux valeurs encadrent le spectre réaliste et conduisent respectivement à $P_{\text{IT}} = 917$ MW et $P_{\text{IT}} = 641$ MW pour 1 GW installé.

2.5 Part de la puissance allouée à l'inférence (H3)

Une fraction de la puissance de calcul totale est allouée à l'entraînement, le reste à l'inférence. Google a publié que l'inférence représentait environ 60 % de sa capacité de calcul IA en 2019–2021 (Institute for Progress 2025). Deloitte Insights (2025) projettent que l'inférence représentera environ deux tiers de la capacité de calcul IA en 2026, contre un tiers en 2023. Nous retenons une fourchette de 50 à 65 %.

2.6 Taux d'utilisation des GPUs (H4)

Les serveurs GPU ne fonctionnent pas à pleine charge en continu. Des études de l'ordonnancement GPU en datacenter (Ye et al. 2024) montrent que des systèmes optimisés peuvent atteindre 94 % d'utilisation dans des conditions idéales. En production réelle, avec la variabilité des charges (pics diurnes, week-ends, maintenance), un taux de 60 à 80 % est retenu comme hypothèse raisonnable.

2.7 Énergie consommée par requête (H5)

C'est la variable centrale du modèle. Trois sous-hypothèses la conditionnent :

1. La **répartition des types de requêtes** : une requête texte courte diffère d'une conversation longue, d'une génération d'image ou d'un appel avec raisonnement. Jegham et al. (2025) montrent que les modèles les plus énergivores dépassent 29 Wh par requête longue, soit un facteur 120 par rapport à la valeur médiane publiée par Google.

2. Le **niveau d'optimisation de l'infrastructure** : le 0,24 Wh mesuré par Google sur Gemini (Elsworth et al. 2025) résulte d'une réduction d'un facteur 33 en un an d'optimisation intensive. Ce chiffre n'est pas représentatif d'un datacenter générique.

3. La **distinction médiane/moyenne** : pour calculer la consommation totale d'un parc, c'est la moyenne qui compte, non la médiane. La distribution des consommations par requête est fortement asymétrique vers le haut.

Fernandez et al. (2025) démontrent que les optimisations d'inférence réduisent la consommation jusqu'à 73 % par rapport aux configurations non optimisées, ce qui signifie qu'un datacenter générique consomme jusqu'à 3,7 fois la valeur d'un système correctement configuré – et *a fortiori* plusieurs fois la valeur médiane Google par requête.

Oviedo et al. (2025) (Microsoft Research) confirment indépendamment une médiane de 0,34 Wh (IQR : 0,18–0,67 Wh) pour des modèles de pointe de plus de 200 milliards de paramètres. Cependant, les requêtes avec raisonnement (*test-time scaling*, 15 fois plus de tokens) atteignent une médiane de 4,32 Wh – un facteur 13 par rapport au texte court. Or, les modèles de raisonnement (o1, o3, R1) représentent une part croissante du trafic, tirant la *moyenne* du parc nettement au-dessus de la médiane publiée.

2.8 Résumé des hypothèses

Le tableau 1 récapitule l'ensemble des paramètres avec leurs sources.

Table 1 : Paramètres du modèle et sources

Variable	**Valeur basse**	**Valeur haute**	**Source**
Puissance installée	1 GW	1 GW	Hypothèse
PUE	1,09	1,56	Google (2024) / Uptime Inst.
Part inférence	50 %	65 %	Institute for Progress (2025) / Deloitte Insights (2025)
Taux utilisation GPU	60 %	80 %	Ye et al. (2024)
Énergie/requête	1,8 Wh	4,2 Wh	Calcul inverse (section 3)
Heures par an	8 760 h	8 760 h	Physique

2.9 Extension à la flotte mondiale (H6)

Le modèle mono-datacenter de 1 GW constitue un exercice pédagogique. Pour confronter la valeur 0,24 Wh à la réalité de l'infrastructure déployée, trois paramètres supplémentaires

sont nécessaires :

1. **Capacité IA mondiale** : l'AIE estime que l'IA représentait 15 % de la consommation totale des datacenters en 2024 (International Energy Agency 2025), soit environ 62 TWh (∼7 GW). Goldman Sachs évalue la capacité totale des datacenters à 55 GW, dont 14 % dédiés à l'IA (∼7,7 GW). Nous retenons **∼7 GW**.

2. **Capacité d'un opérateur de référence** : OpenAI a déclaré une capacité de 0,6 GW en 2024 (SiliconANGLE 2026), pour un volume déclaré de 2,5 milliards de requêtes par jour (TechCrunch 2025). Ce couple (capacité, volume) permet une extrapolation par proportionnalité.

3. **Variation du ratio inférence par opérateur** : la part inférence (H3) varie selon le profil de l'opérateur. Un fournisseur à dominante grand public (OpenAI) consacre une part plus élevée à l'inférence (∼65 %) qu'un opérateur engagé dans l'entraînement de modèles fondamentaux (∼40–50 %). La moyenne pondérée de la flotte est estimée à ∼55 %.

Table 2 : Paramètres additionnels pour l'extrapolation à la flotte mondiale

Variable	**Valeur**	**Source**
Capacité IA mondiale (2024)	∼7 GW	International Energy Agency (2025) / Goldman Sachs Research (2025)
Capacité OpenAI (2024)	0,6 GW	SiliconANGLE (2026)
Volume ChatGPT	2,5 Md req/j	TechCrunch (2025)
Part inférence flotte	∼55 %	Estimation (moyenne pondérée)
Part inférence OpenAI	∼65 %	Estimation (dominante grand public)

3 Modélisation : de 1 GW aux requêtes par utilisateur

Prémisse. Le 0,24 Wh publié par Google est la médiane des requêtes texte sur Gemini Apps, mesurée sur une infrastructure parmi les plus optimisées au monde (PUE 1,09) (Elsworth et al. 2025). Google ne prétend pas que cette valeur s'applique à l'ensemble de l'infrastructure mondiale. Cependant, c'est bien cette valeur qui circule dans la

presse, les rapports institutionnels et les bilans environnementaux comme *la* référence de consommation par requête IA. La question que pose cette section est : *que se passe-t-il si l'on prend cette valeur au mot comme référence système ?*

3.1 Calcul direct

La puissance dédiée à l'inférence est :

$$P_{\text{inf}} = P_{\text{IT}} \times \alpha_{\text{inf}} \times \tau_{\text{GPU}} \tag{2}$$

où α_{inf} est la part allouée à l'inférence et τ_{GPU} le taux d'utilisation effectif des 📖GPUs.

L'énergie annuelle dédiée à l'inférence est :

$$E_{\text{inf}} = P_{\text{inf}} \times 8\,760 \text{ (Wh)} \tag{3}$$

Le nombre de requêtes par jour, pour une consommation unitaire e_{req}, est :

$$N_{\text{req/j}} = \frac{E_{\text{inf}}}{365 \times e_{\text{req}}} \tag{4}$$

3.2 Application numérique – Scénario A (📖hyperscaler optimisé)

Avec PUE = 1,09, $\alpha_{\text{inf}} = 0{,}60$, $\tau_{\text{GPU}} = 0{,}80$:

$$P_{\text{IT}} = 917 \text{ MW} \quad \Rightarrow \quad P_{\text{inf}} = 440 \text{ MW} \quad \Rightarrow \quad E_{\text{inf}} = 3{,}86 \text{ TWh/an}$$

En appliquant $e_{\text{req}} = 0{,}24$ Wh (valeur médiane Google) :

$$N_{\text{req/j}} = \frac{3{,}86 \times 10^{12}}{365 \times 0{,}24} \approx 44 \times 10^{9} \text{ requêtes/jour}$$

> **Résultat absurde (Scénario A, 0,24 Wh) :** un seul datacenter de 1 GW, dans les meilleures conditions d'efficacité, produirait 44 milliards de requêtes par jour, soit **17,6 fois** le volume réel déclaré par OpenAI pour ChatGPT en juillet 2025 (2,5 milliards de requêtes/jour – TechCrunch, 2025).

Test de robustesse avec la médiane Oviedo. Oviedo et al. (2025) rapportent une médiane de 0,34 Wh (IQR : 0,18–0,67 Wh) pour des modèles de pointe. Avec $e_{\text{req}} = 0{,}34$ Wh :

$$N_{\text{req/j}} = \frac{3{,}86 \times 10^{12}}{365 \times 0{,}34} \approx 31 \times 10^{9} \text{ requêtes/jour} \quad (12{,}4\times \text{ le volume ChatGPT})$$

Même avec la borne basse de l'IQR ($e_{\text{req}} = 0{,}18$ Wh) : $N \approx 59 \times 10^{9}$, soit 23,5× – l'absurdité *augmente*. Ce résultat n'est pas contre-intuitif : une énergie par requête *plus basse* que 0,24 Wh implique un nombre de requêtes *encore plus élevé* pour absorber la même capacité, et donc un écart encore plus grand avec le volume réel. Toute valeur inférieure au 0,24 Wh de Google aggrave l'absurdité au lieu de la résoudre.

3.3 Application numérique – Scénario B (industrie moyenne)

Avec PUE = 1,56, $\alpha_{\text{inf}} = 0{,}50$, $\tau_{\text{GPU}} = 0{,}60$:

$$P_{\text{IT}} = 641 \text{ MW} \quad \Rightarrow \quad P_{\text{inf}} = 192 \text{ MW} \quad \Rightarrow \quad E_{\text{inf}} = 1{,}68 \text{ TWh/an}$$

Avec $e_{\text{req}} = 0{,}24$ Wh :

$$N_{\text{req/j}} \approx 19 \times 10^{9} \text{ requêtes/jour} \quad \text{soit encore} \quad 7{,}7\times \text{ le volume réel ChatGPT}$$

3.4 Calcul inverse : estimation réaliste de l'énergie par requête

Si un datacenter de 1 GW servait l'intégralité du trafic ChatGPT (2,5 milliards de requêtes/jour), l'énergie implicite par requête serait :

$$e_{\text{req}} = \frac{E_{\text{inf}}}{N_{\text{req/j}} \times 365} \tag{5}$$

Scénario	E_{inf}	e_{req} implicite
A (📖 hyperscaler)	3,86 TWh	**4,2 Wh** (17,5× valeur Google)
B (industrie)	1,68 TWh	**1,8 Wh** (7,7× valeur Google)

Ces valeurs de 1,8 à 4,2 Wh par requête sont cohérentes avec la littérature : Jegham et al. (2025) estiment, par triangulation indirecte sur 30 modèles commerciaux, des consommations de 0,42 Wh pour les requêtes courtes et jusqu'à 29 Wh pour les invites longues. Sidorkin (2025) citent 4,3 grammes de CO_2 par requête, ce qui correspond, avec une intensité carbone électrique médiane américaine, à une consommation de l'ordre de 4–6 Wh.

> **Conclusion méthodologique :** La valeur 0,24 Wh de Google est une médiane obtenue dans des conditions d'optimisation extrême, chez l'opérateur le mieux classé au monde (PUE 1,09). La prendre comme référence pour des projections de politique énergétique ou des bilans environnementaux revient à évaluer la consommation d'un parc automobile en retenant le record du monde de consommation comme valeur de référence.

3.5 Normalisation par utilisateur

Choix du dénominateur. La normalisation par « internautes mondiaux » (5,4 milliards, UIT 2024) produit des résultats trompeurs car la majorité de ces internautes n'utilise pas l'📖 Intelligence Artificielle (IA) générative. Les sources disponibles permettent d'affiner le dénominateur :

- **Utilisateurs gen AI MAU** (∼1 milliard) : le Microsoft AI Economy Institute (Microsoft AI Economy Institute 2026) mesure qu'environ une personne sur six dans le monde utilise des outils d'IA générative en H2 2025. Eurostat rapporte 32,7 % de la population européenne 16–74 ans en 2025 (Omniflow 2025).
- **Utilisateurs actifs quotidiens** (DAU, ∼250 millions pour ChatGPT seul) : OpenAI rapporte 122 millions de DAU en février 2025 (NerdyNav 2025).

Requêtes par utilisateur : 0,24 Wh vs 1,8 Wh.

Dénominateur	**0,24 Wh** (scénario A)	**1,8 Wh** (scénario B)
Requêtes/jour (1 GW)	44 Md	2,6 Md
Internautes mondiaux (5,4 Md)	8,1	0,5
Utilisateurs gen AI MAU (∼1 Md)	44	2,6
DAU ChatGPT (∼250 M)	176	10,4

À 0,24 Wh, un *seul* datacenter de 1 GW produirait 44 requêtes par jour et par utilisateur MAU – soit plus de quatre fois l'usage observé chez ChatGPT. À 1,8 Wh, la cohérence interne est satisfaisante : 1 GW produit ∼10 requêtes/jour pour chaque utilisateur actif quotidien, ce qui correspond exactement au ratio observé chez ChatGPT ($2{,}5 \times 10^9$ req/j ÷ 250×10^6 DAU = 10 req/j/DAU).

3.6 De un datacenter à la flotte mondiale

Les sections précédentes modélisent un datacenter unique de 1 GW. Or, l'AIE rapporte que l'IA représentait 15 % de la consommation totale des datacenters en 2024 (International Energy Agency 2025), soit environ 62 TWh – l'équivalent de ∼7 GW de capacité IA mondiale. Goldman Sachs estime la capacité totale des datacenters à 55 GW, dont 14 % dédiés à l'IA (∼7,7 GW) (Goldman Sachs Research 2025). En appliquant la part inférence de 50 à 65 % (section 2), l'énergie effectivement dédiée à l'inférence IA mondiale est de 31 à 40 TWh/an.

Estimation du volume réel par extrapolation. OpenAI a déclaré une capacité de 0,6 GW en 2024 (SiliconANGLE 2026), pour un volume de 2,5 milliards de requêtes par jour (TechCrunch 2025). En pondérant par la part inférence respective – estimée à ∼65 % chez OpenAI (produit grand public dominant) contre ∼55 % pour la flotte mondiale (mix incluant davantage d'entraînement, cf. tableau 2) – le ratio corrigé est :

$$\frac{7 \text{ GW} \times 0{,}55}{0{,}6 \text{ GW} \times 0{,}65} \approx 10$$

Le volume mondial de requêtes d'inférence IA est ainsi estimé à ∼25 milliards par jour, tous services confondus. Cet ordre de grandeur constitue une borne haute, les profils de charge variant significativement entre opérateurs.

Limites de cette extrapolation. Ce calcul repose sur les données auto-déclarées d'OpenAI (capacité de 0,6 GW, volume de 2,5 milliards de requêtes/jour), qui ne sont

pas vérifiées par un tiers. Il suppose aussi que le ratio capacité/volume d'OpenAI est transposable à l'ensemble de la flotte, ce qui n'est vrai qu'en ordre de grandeur : OpenAI est un opérateur orienté grand public, là où d'autres consacrent une part plus importante à l'entraînement ou à des charges non conversationnelles.

Les données de Google offrent cependant un point de contrôle indépendant. Elsworth et al. (2025) rapportent que le 0,24 Wh est une médiane obtenue après une réduction d'un facteur 33 en un an d'optimisation, sur une infrastructure à PUE 1,09. Ce chiffre reflète donc le *plancher* atteignable dans les conditions les plus favorables, pas la consommation moyenne du parc Google lui-même – qui inclut des requêtes longues, du raisonnement, de la génération d'images et des charges non conversationnelles. Le rapport Google le confirme implicitement en ne publiant qu'une médiane (insensible aux valeurs extrêmes) et non une moyenne. Le 0,24 Wh est donc cohérent avec les requêtes texte courtes sur Gemini Apps, mais *non généralisable* comme référence système, y compris chez son propre opérateur.

Confrontation avec le 0,24 Wh à l'échelle de la flotte.

Table 3 : Flotte IA mondiale 2024 : requêtes impliquées selon l'énergie par requête

	0,24 Wh (valeur Google)	**1,8 Wh** (réaliste)	**Volume réel** (extrapolation)
Requêtes/jour (flotte IA)	354–457 Md	47–61 Md	∼25 Md
Ratio vs réel estimé	**14–18**×	1,9–2,4×	—
Par MAU gen AI (∼1 Md)	354–457	47–61	∼25
Par DAU (∼250 M)	1 416–1 828	188–244	∼100

Résultat (flotte mondiale) : À 0,24 Wh par requête, la flotte IA mondiale actuelle impliquerait 354 à 457 milliards de requêtes quotidiennes, soit **14 à 18 fois** le volume réel estimé par extrapolation. Même en supposant que l'ensemble des services IA traite collectivement 5 fois le volume de ChatGPT, cela ne représenterait que 12,5 milliards de requêtes – moins de 4 % de la capacité impliquée par le 0,24 Wh. À 1,8 Wh, la capacité dépasse le volume visible d'un facteur ∼2. Cet écart quantifie l'ampleur des usages non directement initiés par des utilisateurs humains – les usages invisibles analysés en section 5 – déjà présents dans l'infrastructure en 2024.

3.7 Limites du modèle et portée de la démonstration

Notre modèle comporte deux hypothèses structurantes qu'il convient d'expliciter.

Hypothèse 1 : toute la capacité sert des requêtes. Le modèle convertit la totalité de la capacité d'inférence en requêtes, comme si les datacenters fonctionnaient à charge nominale. Or, les installations sont dimensionnées pour absorber la croissance projetée, pas la demande instantanée. Une fraction de la capacité peut être structurellement inoccupée pendant la montée en charge.

Cette remarque affecte la *validité arithmétique* de la conversion capacité→requêtes : si les serveurs sont à 40 % d'utilisation effective au lieu de 70 %, le nombre de requêtes impliquées est divisé par ∼1,75 – ce qui réduit l'écart avec le volume réel sans l'éliminer (8–10× au lieu de 14–18× à 0,24 Wh).

Mais elle ne résout pas le *problème énergétique* : un datacenter surdimensionné consomme de l'énergie réelle (refroidissement, réseau, consommation au repos des GPUs à 30–50 % de la puissance de crête) sans la convertir en requêtes utiles. L'énergie par requête *effectivement servie* est alors *supérieure* à la valeur calculée à pleine charge, et non inférieure.

Hypothèse 2 : toutes les requêtes sont visibles. Le modèle attribue toute la capacité à des requêtes initiées par des utilisateurs humains. C'est une hypothèse *maximisante* : elle produit le plus grand nombre de requêtes par utilisateur et donc l'absurdité la plus manifeste. En l'absence de données publiques sur la répartition entre usages visibles et invisibles, trois cas se présentent :

1. **L'usage visible est prédominant.** Alors l'essentiel de la capacité est bien consacré à répondre aux requêtes humaines, et le 0,24 Wh est réfuté par la démonstration par l'absurde exposée ci-dessus.

2. **L'usage invisible est prédominant.** Alors la capacité dédiée aux requêtes humaines est moindre, mais la consommation énergétique globale de l'infrastructure reste identique. Ce sont nos usages quotidiens – envoyer un message, ouvrir une application, effectuer une recherche – qui *déclenchent* ces chaînes de traitement automatisées en arrière-plan. L'empreinte réelle par geste humain est alors *supérieure* à la consommation directe d'une requête, et non inférieure.

3. **La capacité est en surcapacité structurelle.** Ni les usages visibles ni les usages invisibles ne saturent l'infrastructure en 2024 : les 🕮 datacenters sont pré-positionnés pour la croissance attendue. Ce cas réduit le facteur d'absurdité arithmétique (voir hypothèse 1 ci-dessus), mais l'énergie est consommée quand même, et soit elle sera absorbée par la montée en charge – confirmant les projections des sections suivantes – soit elle représente un gaspillage énergétique structurel.

Dans les trois cas, le 0,24 Wh ne rend pas compte de la consommation énergétique réelle par geste humain. Et indépendamment de la répartition visible/invisible, cette valeur ne correspond qu'à des requêtes très spécifiques : un texte court, sans raisonnement, sur une infrastructure optimisée à l'extrême. Une requête courante – rédiger un article pour un réseau social, synthétiser un document de plusieurs pages, générer du code – mobilise des contextes longs, souvent du raisonnement, et se situe bien au-delà de cette médiane communiquée.

La flotte actuelle invalide donc le 0,24 Wh comme référence de calcul. Les sections suivantes projettent ces résultats aux horizons 2026 et 2030.

4 Projections

4.1 Horizon 2026 – trajectoire à court terme

L'🕮 AIE établit que les serveurs accélérés (IA) croissent à un rythme de 30 %/an (International Energy Agency 2025). En appliquant ce taux aux 62 TWh de 2024 :

$$E_{\text{IA}}^{2026} = 62 \times 1{,}3^2 \approx 105 \text{ TWh}$$

Epoch AI estime la capacité IA installée à ∼30 GW fin 2025 (Epoch AI 2025), corroborant cette trajectoire. OpenAI illustre l'accélération : sa capacité a été multipliée par 3,2 en un an (0,6 GW fin 2024 à 1,9 GW fin 2025 – SiliconANGLE, 2026).

Deloitte Insights (2025) projettent que l'inférence représentera ∼2/3 de la capacité de calcul IA en 2026, soit environ **68 TWh** d'inférence.

	0,24 Wh (Google)	**1,8 Wh** (réaliste)
Requêtes/jour (flotte 2026)	∼776 Md	∼104 Md
Par MAU gen AI (∼1,5 Md)	517	69
Par DAU (∼400 M)	1 940	259

Entre 2024 et 2026, la capacité d'inférence IA croît de ∼70 %, mais le nombre d'utilisateurs humains croît plus modestement (∼50 % pour les MAU). L'écart se creuse : à 0,24 Wh, un utilisateur actif quotidien en 2026 déclencherait 1 940 requêtes par jour – un volume incompatible avec tout comportement individuel humain.

4.2 Capacité mondiale planifiée à l'horizon 2030

Les estimations convergentes des principales institutions permettent de borner la capacité 2030 :

- **122 GW** de capacité totale datacenters à fin 2030 selon Goldman Sachs Research (Goldman Sachs Research 2025).
- **945 TWh/an** de consommation totale des datacenters en 2030 dans le scénario central de l'AIE (International Energy Agency 2025), soit un doublement par rapport aux 415 TWh de 2024.
- **200 à 400 TWh** de consommation IA spécifique en 2030, selon la revue critique des modèles publiée par l'IEA-4E (Coroamă 2025), représentant 35 à 50 % de la consommation totale des datacenters projetée.

L'AIE précise que les serveurs accélérés (IA) représentaient 15 % de la consommation totale des datacenters en 2024, et croissent à un rythme de 30 %/an (International Energy Agency 2025).

4.3 Calcul des équivalents-requêtes à l'horizon 2030

En appliquant le même modèle (part inférence 65 %, valeur réaliste 1,8 Wh) :

Scénario	Énergie inférence	Requêtes/jour
Bas (IEA, 200 TWh IA)	130 TWh	198 milliards
Haut (IEA-4E, 400 TWh IA)	260 TWh	395 milliards

En projetant 2,5 milliards d'utilisateurs gen AI MAU et 600 millions de DAU à l'horizon 2030 (extrapolation prudente à partir des trajectoires actuelles (NerdyNav 2025)) :

Dénominateur 2030	Scénario bas	Scénario haut
Par MAU gen AI (2,5 Md)	79 req/j	158 req/j
Par DAU gen AI (600 M)	330 req/j	659 req/j

> **Résultat central :** La capacité IA planifiée pour 2030 représente, pour un utilisateur actif quotidien, un facteur **33 à 66** supérieur à l'usage conversationnel actuel (∼10 req/j/DAU). Aucun comportement individuel humain ne peut expliquer cet écart. La seule interprétation cohérente est que la majorité de la puissance de calcul sera déclenchée par des processus automatisés, sans intervention humaine directe.

5 Taxonomie des usages invisibles

5.1 Définition

Nous définissons un **usage invisible** de l'IA comme toute requête d'inférence générée sans que l'utilisateur final n'ait formulé explicitement une demande à ce moment précis. Le geste conscient – ouvrir une interface, taper une question – est absent. Le calcul est déclenché par un événement applicatif, temporel ou systémique.

Cette définition exclut les interfaces conversationnelles directes (ChatGPT, Claude, Gemini en mode conversationnel), qui constituent la fraction la mieux documentée et la plus visible de l'usage.

5.2 Classe 1 – Agents autonomes et agentic AI

Sapkota et al. (2025) proposent une taxonomie distinguant les AI Agents (systèmes

modulaires pour l'automatisation de tâches spécifiques) des Agentic AI (systèmes multi-agents à collaboration dynamique, décomposition de tâches et mémoire persistante). Dans les deux cas, le modèle génère des requêtes d'inférence en boucle, sans attendre d'intervention humaine entre les étapes.

Des exemples opérationnels documentés incluent les chaînes de traitement automatisées de surveillance de prix, de classification de documents contractuels, de génération automatique de rapports financiers et de réponses de premier niveau en service client. La caractéristique commune est la dissociation entre l'instruction initiale (humaine, ponctuelle) et les centaines à milliers d'appels d'inférence générés en conséquence.

Deloitte Insights (2025) projettent qu'en 2026, les charges de travail d'inférence représenteront environ deux tiers de la capacité de calcul d'inférence totale, contre un tiers en 2023, précisément sous l'effet de cette montée en charge des agents.

5.3 Classe 2 – Intégration dans les outils logiciels de masse

Greshake et al. (2023) ont étudié les vecteurs d'attaque des applications LLM intégrées et décrivent exhaustivement le mécanisme d'intégration : les LLMs sont désormais intégrés comme moteur de fonctionnalités dans des applications qui n'ont pas « intelligence artificielle » dans leur intitulé. Chaque fois qu'un utilisateur ouvre un document, envoie un courriel ou effectue une recherche, une requête d'inférence peut être déclenchée en arrière-plan.

Les cas documentés incluent :

- **Assistance à la rédaction** (correcteurs, suggestions, résumés automatiques) dans les suites bureautiques – Microsoft 365 compte 345 millions d'abonnés, dont la fonctionnalité Copilot est déployée sur environ 15 millions de sièges (Guadamuz 2025).
- **Complétion de code** (GitHub Copilot, 1,3 million d'utilisateurs déclarés (Guadamuz 2025)) : chaque frappe de clavier peut déclencher une requête d'inférence.
- **Moteurs de recherche augmentés** : Google AI Overviews, Bing Copilot, Perplexity – toute recherche web génère désormais potentiellement une requête LLM supplémentaire.
- **Systèmes de recommandation** pilotés par LLM dans les plateformes de contenu (Ma et Wu 2025).

5.4 Classe 3 – Chaînes de traitement automatisées en entreprise

À l'échelle des organisations, les LLMs sont déployés comme composants de chaînes de traitement entièrement automatisées : classification de demandes d'assistance, extraction d'information de documents réglementaires, génération de métadonnées, modération de contenu. Ces processus s'exécutent en continu, 24 h/24, sur des volumes de documents ou de flux de données qui échappent à toute estimation à partir des statistiques d'usage individuel.

Le rapport OpenAI (*State of Enterprise AI*) cité par DataReportal (DataReportal 2026) indique que les messages professionnels ont été multipliés par 8 entre novembre 2024 et début 2026, avec 20 % des messages professionnels transitant via des Custom GPT ou des Projects configurés en automatisation.

5.5 Classe 4 – Inférence embarquée dans les appareils

Deloitte Insights (2025) rapportent que des centaines de millions de PC et smartphones équipés de Neural Processing Unit (processeur neuronal)s (NPUs) ont été vendus en 2025. Ces appareils exécutent des inférences locales – traitement d'image, transcription, assistance contextuelle – en arrière-plan, sans connexion visible à un service cloud. L'inférence embarquée échappe aux statistiques de trafic serveur mais consomme de l'énergie sur le parc matériel distribué.

5.6 Classe 5 – IA systémique dans les infrastructures critiques

L'UNEP et l'AIE documentent des déploiements IA pour l'optimisation de réseaux électriques, la détection de fraudes financières en temps réel, la surveillance environnementale par analyse satellite et la gestion du trafic (International Energy Agency 2025). Ces applications génèrent un flux permanent de requêtes d'inférence déclenchées par des événements physiques (variations de charge réseau, transactions, images satellite) et non par des utilisateurs humains.

5.7 Récapitulatif de la taxonomie

TABLE 4 : Taxonomie des usages invisibles de l'IA – caractéristiques principales

Classe	**Mécanisme déclencheur**	**Exemples documentés**	**Volume relatif**
Agents autonomes	Événement applicatif ou temporal	Surveillance prix, génération rapports	Très élevé
Intégration logicielle	Action utilisateur indirecte	Copilot, correcteurs, search augmenté	Élevé
Traitement automatisé	Flux documentaire continu	Modération, classification demandes	Élevé
Inférence embarquée	Capteurs, contexte appareil	NPU smartphones, PC IA	Croissant
Infrastructures critiques	Événements physiques	Réseaux, fraude, satellite	Émergent

Portée de la taxonomie. Cette classification est *existentielle* : elle établit que ces usages existent, croissent et sont documentés dans la littérature. Elle n'est pas *dimensionnante* : aucune donnée publique ne permet aujourd'hui de quantifier la part que chaque classe représente dans la capacité de calcul totale. L'opacité des opérateurs sur la répartition de leurs charges de travail (cf. section 6) est précisément ce qui empêche ce dimensionnement. La taxonomie n'est donc pas la preuve que les usages invisibles sont majoritaires – elle montre qu'ils sont nombreux, croissants, et structurellement non comptabilisés dans les estimations fondées sur les usages individuels.

6 Pistes pour la gouvernance et la régulation

6.1 L'insuffisance structurelle de la sobriété individuelle

Si la majorité de la puissance de calcul est déclenchée par des processus automatisés, les approches de régulation fondées sur la sensibilisation des utilisateurs individuels perdent leur pertinence. C'est le même glissement qu'on a observé avec la consommation électrique des datacenters depuis 2010 : pendant des années, le discours portait sur « éteindre

son ordinateur en veille », alors que la vraie croissance se produisait dans les serveurs que personne ne voyait.

Hacker (2024) argumente dans la *Common Market Law Review* que les simples mécanismes de transparence sont insuffisants et propose un cadre de régulation plus contraignant : *sustainability by design*, plafonds de consommation et éventuelle intégration de l'IA dans le marché européen du carbone.

6.2 L'opacité comme problème de gouvernance

Desroches et al. (2025) soulignent que l'opacité des grands fournisseurs empêche les entreprises d'évaluer leur propre empreinte IA. Les seules données de production à grande échelle proviennent des hyperscalers eux-mêmes, créant un problème structurel d'auto-évaluation (Elsworth et al. 2025).

La revue IEA-4E (Coroamă 2025) recommande explicitement aux journalistes et décideurs de « critically assess the quality of data centre energy estimates », constatant que les projections divergent d'un facteur 40 selon les hypothèses. Cette divergence n'est pas épistémique : elle reflète des choix de périmètre et des intérêts divergents dans la communication des acteurs.

6.3 La régulation doit cibler les opérateurs, pas les utilisateurs

Si les usages invisibles représentent la majorité de la capacité de calcul future, la seule régulation efficace se situe au niveau des opérateurs d'infrastructure. Les obligations de reporting environnemental (PUE, WUE, intensité carbone) en cours d'élaboration dans le cadre européen vont dans ce sens, mais leur granularité actuelle ne permet pas de distinguer les charges de travail par type d'usage.

7 Conclusion

Cet article a démontré, par un exercice de modélisation physique, que la valeur médiane de consommation par requête publiée par Google (0,24 Wh) est incompatible avec les volumes d'usage réels, et ce à trois échelles successives :

- **Un datacenter de 1 GW** (section 3) : le 0,24 Wh implique 44 milliards de requêtes par jour, soit 176 par utilisateur actif quotidien de ChatGPT – un volume

sans rapport avec l'usage observé.

- **La flotte mondiale de ∼7 GW** (section 3.6) : étendu à l'ensemble des datacenters IA, le 0,24 Wh implique 354 à 457 milliards de requêtes par jour, soit 14 à 18 fois le volume réel estimé à ∼25 milliards. L'absurdité n'est plus locale, elle est systémique.
- **La trajectoire 2026–2030** (section 4) : avec une croissance de 30 %/an des serveurs accélérés, la capacité d'inférence atteint ∼68 TWh dès 2026 et 130 à 260 TWh en 2030, creusant encore l'écart entre la capacité construite et tout usage humain plausible.

Cette démonstration repose sur deux hypothèses maximisantes : nous attribuons toute la capacité à des requêtes, et toutes les requêtes à des usages visibles. Comme discuté en section 3.7, relâcher ces hypothèses ne sauve pas le 0,24 Wh : soit l'usage visible domine, et le chiffre est réfuté ; soit l'usage invisible domine, et l'empreinte réelle par geste humain est *supérieure* à la consommation directe d'une requête ; soit la capacité est en surcapacité structurelle, et l'énergie par requête effectivement servie est encore plus élevée que dans notre modèle à pleine charge.

Les valeurs réalistes se situent entre 1,8 et 4,2 Wh par requête, soit un facteur 7 à 17 au-dessus du chiffre communiqué. De plus, les travaux de Oviedo et al. (2025) montrent que les requêtes avec raisonnement (*test-time scaling*) atteignent 4,32 Wh – un facteur 13 par rapport au texte court – tirant la moyenne du parc nettement au-dessus de la médiane publiée, à mesure que les modèles de raisonnement se généralisent. Et indépendamment de ces moyennes, le 0,24 Wh ne correspond qu'à des requêtes très spécifiques : un texte court, sans raisonnement, sur une infrastructure optimisée à l'extrême. Les usages courants – rédiger un article, synthétiser un document, générer du code – se situent bien au-delà.

La capacité construite ne peut être absorbée par des usages conversationnels humains. Elle implique nécessairement l'émergence massive d'usages invisibles – agents autonomes, intégration logicielle, chaînes de traitement automatisées, inférence embarquée, infrastructures critiques – dans lesquels l'humain n'est plus l'initiateur direct des requêtes d'inférence (section 5).

Cette montée en charge rend structurellement insuffisantes les approches de régulation fondées sur la sobriété individuelle (section 6). La vraie question que soulève cette transition n'est pas « combien consomme une requête ? » mais « qui décide du nombre de requêtes générées, au nom de qui, et avec quelle transparence ? ». Cette question est fondamentalement politique avant d'être technique.

Acronymes

AIE Agence Internationale de l'Énergie (International Energy Agency). 3, 7, 11, 14, 15, 18

DAU Daily Active Users (utilisateurs actifs quotidiens). 10, 11, 12, 15, 16

FLOPs Floating Point Operations per Second (opérations en virgule flottante par seconde). 3

GPU Graphics Processing Unit (processeur graphique). 1, 3, 5, 6, 8, 13

IA Intelligence Artificielle. 10

LLM Large Language Model (grand modèle de langage). 3, 17, 18

MAU Monthly Active Users (utilisateurs actifs mensuels). 10, 11, 12, 15, 16

NPU Neural Processing Unit (processeur neuronal). 18, 19

PUE Power Usage Effectiveness (efficacité énergétique d'un datacenter). 1, 4, 5, 6, 10, 20

WUE Water Usage Effectiveness (efficacité hydrique d'un datacenter). 20

Glossaire

agentic AI Système d'IA multi-agents à collaboration dynamique, décomposition de tâches et mémoire persistante, générant des requêtes d'inférence en boucle sans intervention humaine entre les étapes. 1, 16

datacenter Installation physique hébergeant des serveurs informatiques, leur alimentation électrique, leur refroidissement et leur connectivité réseau. 3, 4, 5, 6, 7, 9, 11, 13, 14, 15, 19, 20, 21

entraînement Phase initiale de construction d'un modèle d'IA, consistant à ajuster ses paramètres à partir d'un corpus de données. Opération ponctuelle et coûteuse, mais amortie sur la durée de vie du modèle. 5

hyperscaler Opérateur de datacenters à très grande échelle (Google, Microsoft, Meta, Amazon) disposant de capacités dépassant le gigawatt et d'infrastructures optimisées (PUE $<$ 1,2). 1, 4, 8, 10, 20

inférence Phase d'exploitation d'un modèle d'IA entraîné, consistant à produire une réponse à partir d'une entrée (requête). Par opposition à l'entraînement, l'inférence s'exécute en continu en production et constitue la majorité de la consommation énergétique à l'échelle. 5, 11, 16

usage invisible Requête d'inférence IA générée sans que l'utilisateur final n'ait formulé explicitement une demande à ce moment précis. Le calcul est déclenché par un événement applicatif, temporel ou systémique. 13, 16, 19, 20, 21

Bibliographie

Chien, A. et al. (2023). « Reducing the Carbon Impact of Generative AI Inference (today and in 2035) ». In : *Proceedings of the 2nd Workshop on Sustainable Computer Systems (HotCarbon)*. 135 citations, p. 1-7. URL : https://consensus.app/papers/reducing-the-carbon-impact-of-generative-ai-inference-chien-lin/7bb9cf02cc345f6dacd7d57183dd6920/.

Coroamă, V. C. (2025). *Data Centre Energy Use : Critical Review of Models and Results*. Plage IA spécifique 2030 : 200–400 TWh, soit 35–50 % du total DC. URL : https://www.iea-4e.org/wp-content/uploads/2025/05/Data-Centre-Energy-Use-Critical-Review-of-Models-and-Results.pdf.

DataReportal (2026). *Digital 2026 : One Billion People Using AI*. URL : https://datareportal.com/reports/digital-2026-one-billion-people-using-ai.

Deloitte Insights (2025). *More compute for AI, not less*. L'inférence représentera ∼2/3 du compute IA en 2026. Centaines de millions de PC/smartphones avec NPU vendus en 2025. URL : https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/compute-power-ai.html.

Desroches, C. et al. (2025). « Exploring the sustainable scaling of AI dilemma : A projective study of corporations' AI environmental impacts ». In : *ArXiv*. 1 citation. url : https://consensus.app/papers/exploring-the-sustainable-scaling-of-ai-dilemma-a-desroches-chauvin/c2a3d1f120ab59ac930ef7b855b0d561/.

Elsworth, C. et al. (2025). *Measuring the environmental impact of delivering AI at Google Scale*. 9 citations. Requête médiane Gemini Apps : 0,24 Wh. url : https://consensus.app/papers/measuring-the-environmental-impact-of-delivering-ai-at-elsworth-huang/af825dad75c65294a026c1e326581460/.

Epoch AI (2025). *Global AI power capacity is now comparable to peak power usage of New York State*. Capacité IA totale ∼30 GW fin 2025. url : https://epoch.ai/data-insights/ai-datacenter-power.

Fernandez, J. et al. (2025). « Energy Considerations of Large Language Model Inference and Efficiency Optimizations ». In : *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 14 citations, p. 32556-32569. url : https://consensus.app/papers/energy-considerations-of-large-language-model-inference-fernandez-na/0336437d038550999374e842e921ee80/.

Goldman Sachs Research (2025). *AI to Drive 165% Increase in Data Center Power Demand by 2030*. Capacité totale datacenters 2024 : ∼55 GW (14 % IA). Projection fin 2030 : ∼122 GW. url : https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030.

Google (2024). *Power Usage Effectiveness – Google Data Centers*. PUE flotte mondiale 2024 = 1,09. url : https://datacenters.google/efficiency/.

Greshake, K. et al. (2023). « Not What You've Signed Up For : Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection ». In : *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 700 citations. url : https://consensus.app/papers/not-what-youve-signed-up-for-compromising-realworld-greshake-abdelnabi/667ace18801a59ce89d56d0e636d376c/.

Guadamuz, A. (2025). *How Many People Are Using Generative AI on a Daily Basis ?* url : https://www.technollama.co.uk/a-gemini-report-how-many-people-are-using-generative-ai-on-a-daily-basis-a-gemini-report.

Hacker, P. (2024). « Sustainable AI Regulation ». In : *Common Market Law Review*. 9 citations. url : https://consensus.app/papers/article-sustainable-ai-regulation-hacker/bb2a56535e6354e996b118f16805adf6/.

Institute for Progress (2025). *How to Build the Future of AI in the United States*. Google : l'inférence = 60 % du compute IA en 2019–2021. url : https://ifp.org/future-of-ai-compute/.

International Energy Agency (2025). *Energy and AI – Energy demand from AI*. Scénario central : 945 TWh datacenters en 2030, croissance serveurs IA +30 %/an. URL : https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai.

Jegham, N. et al. (2025). « How Hungry is AI ? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference ». In : *ArXiv*. 25 citations. URL : https://consensus.app/papers/how-hungry-is-ai-benchmarking-energy-water-and-carbon-jegham-abdelatti/2aed36133ef8566592bf8c9730f45c76/.

Ma, Z. et J. Wu (2025). « Design of an LLM-Driven Personalized Learning Resource Recommendation System ». In : *Frontiers in Computing and Intelligent Systems*. URL : https://consensus.app/papers/design-of-an-llmdriven-personalized-learning-resource-ma-wu/eb809637c0f15f2a9e5c737c62dcb0fd/.

McKinsey & Company (2025). *The next big shifts in AI workloads and hyperscaler strategies*. URL : https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-next-big-shifts-in-ai-workloads-and-hyperscaler-strategies.

Meta Platforms (2024). *Sustainability Report 2024*. PUE flotte mondiale Meta = 1,08. URL : https://sustainability.meta.com/.

Microsoft AI Economy Institute (2026). *Global AI Adoption in 2025*. ∼1 personne sur 6 dans le monde utilise des outils IA générative en H2 2025. URL : https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2025/.

NerdyNav (2025). *Latest ChatGPT Statistics : 800M+ Users, Revenue*. 122,58 millions de DAU ChatGPT (fév. 2025) ; 1 milliard de requêtes/j déclarées par Sam Altman (déc. 2024). URL : https://nerdynav.com/chatgpt-statistics/.

Omniflow (2025). *AI Usage Statistics : How People Are Using AI in 2026*. Eurostat 2025 : 32,7 % des Européens 16–74 ans utilisateurs gen AI. URL : https://www.omniflow.team/blog/ai-usage-statistics.

Oviedo, F. et al. (2025). « Energy Use of AI Inference : Efficiency Pathways and Test-Time Compute ». In : *ArXiv*. 3 citations. Médiane 0,34 Wh (IQR 0,18–0,67) pour modèles frontier >200B paramètres ; 4,32 Wh pour raisonnement (test-time scaling). URL : https://arxiv.org/abs/2509.20241.

Samsi, S. et al. (2023). « From Words to Watts : Benchmarking the Energy Costs of Large Language Model Inference ». In : *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. 221 citations, p. 1-9. URL : https://consensus.app/papers/from-words-to-watts-benchmarking-the-energy-costs-of-large-samsi-zhao/8ae46eeb1b9954e5bf31a8b552c69fec/.

Sapkota, R., K. I. Roumeliotis et M. Karkee (2025). « AI Agents vs. Agentic AI : A Conceptual Taxonomy, Applications and Challenges ». In : *Information Fusion* 126. 103 citations, p. 103599. URL : https://consensus.app/papers/ai-agents-vs-agentic-ai-a-conceptual-taxonomy-applications-sapkota-roumeliotis/f2dae358a9155cbc9eb0188c44ecdfbd/.

Sidorkin, A. M. (2025). « Environmental Impact of Generative AI : Carbon and Water Footprint ». In : *AI-EDU ArXiv*. URL : https://consensus.app/papers/environmental-impact-of-generative-ai-carbon-and-water-sidorkin/221db0b0a6cf5b22a041e

SiliconANGLE (jan. 2026). *OpenAI reveals its data center capacity tripled to 1.9GW in 2025*. Capacité OpenAI : 0,2 GW (2023), 0,6 GW (2024), 1,9 GW (2025). Déclaration de Sarah Friar, CFO. URL : https://siliconangle.com/2026/01/19/openai-reveals-data-center-capacity-tripled-1-9gw-2025/.

TechCrunch (2025). *ChatGPT users send 2.5 billion prompts a day*. URL : https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/.

Uptime Institute (2024). *Global Data Center Survey 2024*. PUE moyen mondial = 1,56. URL : https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2024.

Ye, Z. et al. (2024). « Deep Learning Workload Scheduling in GPU Datacenters : A Survey ». In : *ACM Computing Surveys* 56. 71 citations, p. 1-38. URL : https://consensus.app/papers/deep-learning-workload-scheduling-in-gpu-datacenters-a-ye-gao/059bef8608b1553988b670478d1a19a7/.