# Rapport d'Analyse : Consommation Énergétique de l'IA

Évaluation structurée du cycle de vie : Fabrication, Entraînement et Inférence

RAPPORT D'ANALYSE TECHNIQUE • BORNÉ EXCLUSIVEMENT À L'ÉNERGIE ÉLECTRIQUE

**Cadrage de l'étude :** L'émergence rapide de l'intelligence artificielle (IA), en particulier de l'IA générative, s'accompagne d'une demande croissante en puissance de calcul et en électricité. Conformément au périmètre strict d'analyse, ce rapport évalue **exclusivement la consommation énergétique** en kilowattheures (kWh) et térawattheures (TWh) à travers les trois phases clés du numérique : la fabrication du matériel (énergie incorporée), l'entraînement des modèles et leur utilisation opérationnelle (inférence).

## 1. Consommation énergétique liée à la fabrication du matériel

La fabrication des composants électroniques de pointe (GPU, puces d'accélération, processeurs, serveurs et infrastructures de centres de données) constitue un processus industriel particulièrement gourmand en électricité (Académie des technologies, 2026 ; AIE, 2025). L'énergie incorporée lors de la fabrication dépend de la complexité de la gravure et de la surface de silicium produite :

- **Volume mondial de la fabrication :** La production globale de semi-conducteurs consomme environ **100 TWh d'électricité par an** (Académie des technologies, 2026).
- **Intensité de gravure des plaquettes (wafers) :** La fabrication d'une seule plaque de silicium gravée en 3 nanomètres (3 nm) nécessite environ **2,3 MWh d'électricité** (Académie des technologies, 2026 ; AIE, 2025).
- **Consommation surfacique :** L'énergie électrique moyenne nécessaire pour fabriquer un centimètre carré de circuit intégré varie de **1 à 2 kWh/cm²** pour les composants standards, et atteint **2,5 à 5 kWh/cm²** pour les puces IA spécialisées les plus complexes. La fabrication d'un processeur haute performance comme le NVIDIA H100 (surface d'environ 8 cm²) requiert environ **20 kWh** (Académie des technologies, 2026).
- **Répartition interne de l'énergie d'usine :** Lors de la fabrication des puces, **60 %** de l'électricité est absorbée directement par les procédés de fabrication du semi-conducteur (dépôt, lithographie, gravure), tandis que **40 %** servent aux équipements auxiliaires de l'usine (refroidissement, salles blanches) (Académie des technologies, 2026).
- **Bilan fabrication vs exploitation :** La fabrication d'un serveur haute performance nécessite plus de **10 MWh**. Cependant, sur une durée de vie opérationnelle de 5 ans, ce même serveur consomme plus de **80 MWh**. Ainsi, les puces d'IA utilisées dans les centres de données consomment annuellement **20 à 100 fois plus d'électricité lors de leur utilisation** que l'énergie requise pour leur fabrication initiale (Académie des technologies, 2026 ; AIE, 2025).

| Indicateur de Fabrication | Valeur Énergétique | Source Référencée |
|---|---|---|
| **Production mondiale de semi-conducteurs** | ~100 TWh / an | Académie des technologies (2026) |
| **Gravure d'un wafer 3 nm** | 2,3 MWh / wafer | AIE (2025) ; Académie des tech. (2026) |
| **Énergie surfacique puces IA de pointe** | 2,5 à 5 kWh / cm² | Académie des technologies (2026) |

| Indicateur de Fabrication | Valeur Énergétique | Source Référencée |
|---|---|---|
| **Fabrication d'une puce NVIDIA H100 (~8 cm²)** | ~20 kWh / puce | Académie des technologies (2026) |
| **Ratio Énergie Exploitation (5 ans) / Fabrication** | 20 à 100 x l'énergie de fabrication | AIE (2025) ; Académie des tech. (2026) |

## 2. Consommation énergétique liée à l'entraînement des IA

L'entraînement (apprentissage) est la phase initiale durant laquelle le modèle d'IA absorbe des masses de données pour ajuster ses paramètres. Cette phase s'effectue sur des clusters de milliers de processeurs graphiques (GPU) fonctionnant à pleine charge pendant plusieurs semaines ou mois (Ifri, 2025 ; AIE, 2025) :

- **Modèle GPT-3 (175 milliards de paramètres) :** L'entraînement a consommé plus de **1 287 MWh** (1,287 GWh) d'électricité (Patterson et al., 2021 ; AIE, 2025 ; Wikipédia).
- **Modèle GPT-4 :** L'entraînement a mobilisé environ 25 000 GPU avec une puissance nominale de 10 MW, portée à environ **22 MW au total** avec les CPU auxiliaires, les équipements réseau et le refroidissement. Sur 14 semaines de calcul (facteur de charge de 84 %), la consommation électrique totale est estimée entre **42,4 GWh** (~0,43 GWh/jour) et **50 à 60 GWh** (Ifri, 2025 ; Académie des technologies, 2026 ; AIE, 2025).
- **Volume agrégé mondial de l'apprentissage :** La consommation cumulée attribuée à l'entraînement de l'ensemble des grands modèles développés depuis 2020 s'élève à environ **1 700 GWh (1,7 TWh)**, soit 0,001 % de la consommation électrique mondiale sur la période. Les pics de puissance pour les très grands clusters d'entraînement atteignent jusqu'à **154 MW** (AIE, 2025).
- **Sautes de charge et stabilité du réseau :** Lors des phases d'apprentissage, l'allumage ou l'extinction simultanés de dizaines de milliers de GPU (sauvegardes d'étapes, interruptions de calcul) génèrent de fortes sautes de consommation de l'ordre de **plusieurs dizaines de mégawatts** (Ifri, 2025 ; Académie des technologies, 2026).
- **Optimisations algorithmiques :** Les architectures frugales et le pré-entraînement permettent de réduire la charge électrique. Par exemple, le modèle **DeepSeek v3** a été pré-entraîné sur des clusters de 2 000 puces (contre 16 000 puces pour des modèles équivalents), réduisant fortement l'énergie consommée par rapport aux approches de bout en bout (Ifri, 2025 ; Académie des technologies, 2026).

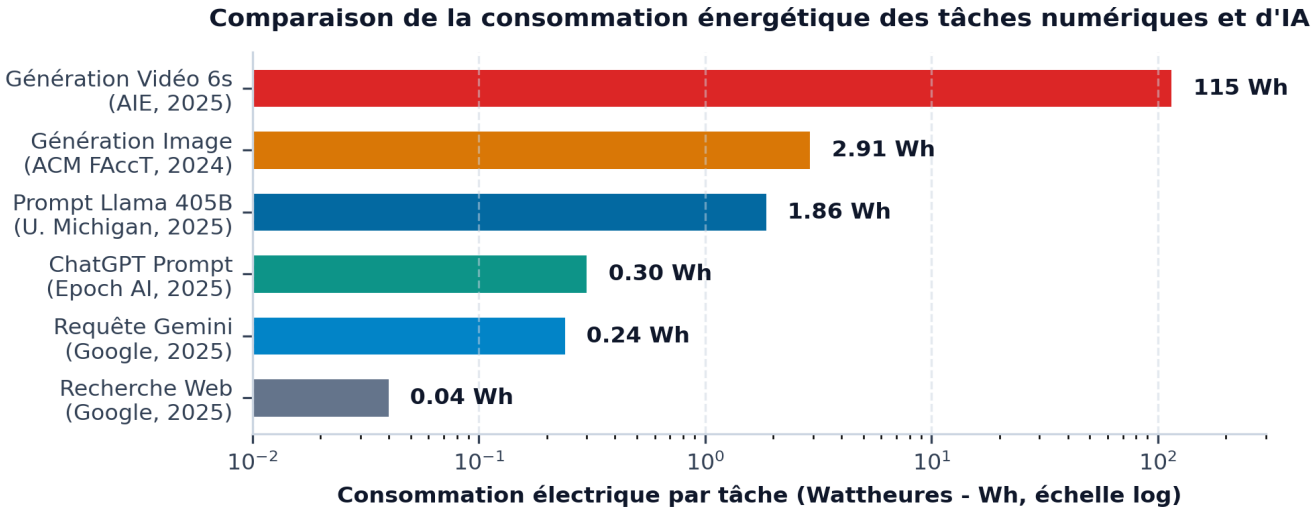
## 3. Consommation énergétique liée à l'utilisation des IA (Inférence)

Une fois déployé, le modèle intervient lors de chaque requête utilisateur (phase d'inférence). Si la consommation unitaire d'une requête est modeste, la répétition à grande échelle (milliards de requêtes quotidiennes) fait de l'inférence le facteur prépondérant de la consommation électrique de l'IA (Polytechnique Insights, 2024 ; Ifri, 2025) :

- **Basculement du poids énergétique :** L'inférence représente désormais entre **60 % et plus de 90 %** de la consommation électrique totale liée à l'exploitation des grands modèles de langage, contre 10 à 40 % pour l'apprentissage (Polytechnique Insights, 2024 ; Ifri, 2025 ; Académie des technologies, 2026 ; Wikipédia).
- **Requête textuelle vs Recherche Web :** Une requête texte courte sur ChatGPT consomme ~**0,3 Wh** (Epoch AI / You et al., 2025) à **0,34 Wh** (Sam Altman, 2025). La requête médiane sur Google Gemini consomme **0,24 Wh** (Elsworth et al., 2025). En comparaison, une recherche classique sur Google consomme entre 0,04 Wh (estimations académiques récentes) et 0,3 Wh (chiffre officiel 2009). Le ratio réel varie ainsi entre **1,5x et 3x**, et non 10x (Projet Celsius, 2026 ; Epoch AI, 2025).

- **Impact de la taille du modèle et des prompts :** Une classification simple consomme 0,0005 à 0,007 Wh (ACM FAccT, 2024 ; Ifri, 2025). Un petit modèle (Meta Llama 3.1 8B) consomme ~0,03 Wh par réponse, tandis qu'un grand modèle (Llama 3.1 405B) requiert 1,86 Wh (Université du Michigan, 2024). L'analyse de documents longs (~7 500 mots / 200 pages) par des modèles de raisonnement peut atteindre **2,5 Wh à 40 Wh** par requête (Projet Celsius, 2026 ; Wikipédia).

- **Variations selon la modalité :** La réponse textuelle simple consomme ~0,05 à 0,3 Wh ; la génération d'image consomme **0,32 Wh à 2,91 Wh** par image (jusqu'à 11,49 Wh pour les modèles les plus lourds) (ACM FAccT, 2024 ; AIE, 2025). La génération d'une vidéo courte de 6 secondes requiert environ **115 Wh**, soit **700 fois plus qu'une image** et l'équivalent de près de deux recharges d'ordinateur portable (AIE, 2025).




***Graphique 1 :*** *Consommation électrique unitaire par type de tâche numérique et d'IA (Données issues d'Epoch AI, Google arXiv:2508.15734, ACM FAccT 2024 et AIE 2025).*

## 4. Projections de la demande globale et leviers d'efficience

À l'échelle mondiale, la consommation électrique globale des centres de données (cloud et IA) devrait passer d'environ **415 à 460 TWh** (en 2022/2024) à une fourchette comprise entre **800 TWh et 1 500 TWh d'ici 2030** (AIE, 2025 ; Ifri, 2025 ; The Shift Project, 2025 ; Académie des technologies, 2026). La part attribuable à l'IA au sein des centres de données, estimée à 10-15 % en 2024 (<0,2 % de l'électricité mondiale), devrait atteindre **19 % à 35 %** d'ici 2028-2030 (Ifri, 2025 ; The Shift Project, 2025).

**Leviers de modération énergétique :** L'augmentation globale de la consommation est contenue par les progrès d'efficacité du matériel spécialisé (l'énergie par token ayant historiquement été divisée par 4 chaque année), par l'optimisation des traitements en lots (*batching*, réduisant l'énergie par tâche de 50 %), ainsi que par le déploiement du calcul en périphérie (*edge computing* / NPU sur terminal), capable de réduire la consommation d'inférence de **plus de 90 %** par rapport au cloud (Ifri, 2025 ; AIE, 2025 ; Académie des technologies, 2026).

## 5. Références bibliographiques citées


1. **Académie des technologies** (Mars 2026). *Les consommations du numérique en 2030 : électricité, eau et empreinte carbone*. Rapport rédigé sous la direction de J.-C. André et Y. Caseau, Paris.

2. **De Vries, A.** (2023). *The growing energy footprint of artificial intelligence*. Joule, 7(10), 2191–2194.

3. **Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., et al.** (2025). *Measuring the environmental impact of delivering AI at Google Scale*. arXiv:2508.15734.

4. **IEA - International Energy Agency** (2025). *Energy and AI*. Special Report, Paris.

5. **Ligozat, A.-L., & De Vries, A.** (13 novembre 2024). *IA générative : la consommation énergétique explose*. Polytechnique Insights, Revue de l'Institut Polytechnique de Paris.

6. **Luccioni, S., Jernite, Y., & Strubell, E.** (2024). *Power Hungry Processing: Watts Driving the Cost of AI Deployment?*. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24), pp. 85–99.

7. **Pakula, G. / Projet Celsius** (Juin 2026). *ChatGPT consomme-t-il vraiment plus qu'une recherche Google ?*. Analyse technique Celsius.

8. **Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., et al.** (2022). *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*. Computer, 55(7), 18–28.

9. **Roucy-Rochegonde, L. de, & Buffard, A. / Ifri** (Février 2025). *IA, centres de données et énergie : Quelles tendances ?*. Notes de l'Ifri, Institut français des relations internationales, Paris.

10. **Strubell, E., Ganesh, A., & McCallum, A.** (2019). *Energy and Policy Considerations for Deep Learning in NLP*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3645–3650.

11. **The Shift Project** (Octobre 2025). *Intelligence artificielle, données, calcul : quelles infrastructures dans un monde décarboné ?*. Rapport final, Paris.

12. **You, J. / Epoch AI** (2025). *How much energy does ChatGPT use?*. Epoch AI Gradient Updates.