SPOTLAB / ANALYSE P23

Impact environnemental de l'IA : comment piloter énergie, coût et utilité ?

Comment mesurer l'empreinte et l'utilité d'une IA professionnelle sans extrapoler des chiffres par prompt ?

Analyse / P23

Novembre 2026

Sources consultées au 2 novembre 2026

Use the best of AI.
Build your own intelligence.

ANALYSE P23

Sommaire

01	Trois décisions pour les dirigeants	3
02	Que sait-on réellement de l'empreinte de l'IA ?	4
03	Pourquoi les chiffres « par prompt » sont difficiles à comparer	5
04	Une unité de comparaison plus utile : la tâche acceptée	7
05	La méthode en cinq décisions	8
06	Exemple illustratif : choisir entre deux workflows	9
07	Comment éviter le greenwashing ?	10
08	Le rôle de l'architecture professionnelle	11
09	Périmètre et sources	12

FICHE DE L'ANALYSE

ÉDITEUR
Spotwork.AI — Spotlab, Research & Insights

RÉFÉRENCE
P23

TYPE
Analyse RSE et IA professionnelle

COLLECTION
Économie & impact

SOURCES CONSULTÉES
2 novembre 2026

PUBLICATIONS LIÉES
P05 · P06 · P18

EN BREF

Une requête d'IA n'a pas un impact environnemental universel. Sa consommation dépend du travail demandé, de la configuration, de l'infrastructure et du nombre de tentatives ; son intérêt environnemental net dépend aussi de ce qui aurait été fait sans IA. Pour agir sans greenwashing, les entreprises doivent mesurer l'impact d'une tâche professionnelle effectivement acceptée, puis examiner ses effets à l'échelle des usages et sur le cycle de vie du système.

> *Les comparaisons environnementales doivent partager un périmètre et des hypothèses explicites, et être reliées au travail effectivement réalisé.*
>
> CE QU'IL FAUT RETENIR

01 Trois décisions pour les dirigeants

1 Mesurer plusieurs impacts, pas un seul chiffre de CO_2.

Énergie, émissions, eau et équipements ne suivent pas nécessairement la même évolution. Une consommation électrique plus faible ne signifie pas toujours un bilan environnemental global meilleur.

2 Comparer un travail fini à une alternative crédible.

Le nombre de prompts ne dit rien de la qualité finale, des corrections ni de la tâche qui aurait été réalisée sans IA. Une estimation par requête ne doit pas être présentée comme une analyse de cycle de vie.

3 Réduire les opérations inutiles avant d'optimiser le modèle.

Éviter une exécution sans utilité, limiter les recherches répétées, réutiliser une information validée et adapter les ressources à la tâche peuvent constituer des leviers — à vérifier par la mesure, sans sacrifier la qualité.

DEUX QUESTIONS À NE PAS CONFONDRE

« Quelle est l'empreinte des systèmes d'IA que nous utilisons ? » porte sur les impacts directs. « Quelles activités, économies ou transformations l'IA permet-elle réellement ? » exige de comparer des scénarios d'usage. **Un bénéfice environnemental net ne se déduit pas d'une efficacité informatique ou d'une productivité accrue.**

02 Que sait-on réellement de l'empreinte de l'IA ?

L'Agence internationale de l'énergie (AIE) estime qu'en 2024 les centres de données ont consommé environ 415 TWh d'électricité, soit près de 1,5 % de la consommation mondiale. Dans son scénario central de 2025, cette consommation atteindrait environ 945 TWh en 2030 : il s'agit d'une projection, pour tous les centres de données, pas d'une mesure de la seule IA générative. Son rapport du 14 octobre 2026 approfondit les enjeux de réseaux, d'accessibilité de l'électricité, d'émissions et d'utilisation de l'eau [1] [2].

415 TWh
Consommation des centres de données en 2024 — environ 1,5 % de l'électricité mondiale [1]

945 TWh
Projection 2030 du scénario central de l'AIE (2025) — tous centres de données confondus [1]

34 juridictions
Politiques d'intégration des centres de données analysées par l'AIE en octobre 2026, pour la présidence française du G7 [2]

L'ADEME a publié en juillet 2026 un avis couvrant les impacts directs et indirects des IA génératives et agentiques. Sa recherche sur l'analyse de cycle de vie des GPU rappelle aussi que les équipements matériels ont des impacts de fabrication, indépendants de l'électricité nécessaire pendant l'usage [3] [4].

Ces résultats établissent la matérialité de l'enjeu, pas un facteur d'émission utilisable tel quel pour calculer l'empreinte d'un rapport, d'un dossier de contrôle interne ou d'une proposition commerciale.

Quatre niveaux, quatre frontières de calcul

Niveau	Ce qui est évalué	Ce qui ne peut pas être déduit automatiquement
Centre de données	Charge électrique, infrastructure, réseaux, eau	Énergie ou CO_2 d'un prompt donné
Système d'IA	Modèle, calcul, stockage, réseaux, matériel, cycle de vie	Bilan environnemental global d'une entreprise
Service applicatif	Recherche, appels aux modèles, outils, reprises, validation	Résultat métier effectivement accepté
Usage professionnel	Tâche réalisée, alternative sans IA, effets rebond	Bénéfice net sans scénario de référence

03 Pourquoi les chiffres « par prompt » sont difficiles à comparer

Dans une étude de 2025, Google estime que le prompt texte médian de Gemini Apps représente environ 0,24 Wh dans le périmètre opérationnel mesuré par son infrastructure. Cette mesure est documentée pour un produit, une période, un profil de requêtes et une méthode ; elle n'est pas transférable à n'importe quel modèle, requête longue, agent ou architecture. Elle ne remplace pas non plus un bilan complet du cycle de vie [7].

Comparer deux chiffres demande donc de connaître au minimum : le type de tâche, le modèle et sa version, le volume de tokens et les outils, les serveurs et leur taux d'utilisation, les charges de refroidissement et les facteurs d'émission retenus. Il faut préciser ce qui est mesuré directement, fourni par un opérateur ou estimé. Une étude locale ne doit jamais être présentée comme une moyenne mondiale.

Même opérateur, deux périmètres ; deux opérateurs, deux méthodes

Publication	Périmètre	Valeur publiée	Limite déclarée
Google, 21 août 2025 — méthode complète [7]	Puce active, machines en réserve, CPU et mémoire, surcoût du centre de données (PUE), eau de refroidissement	0,24 Wh · 0,03 gCO_2e · 0,26 mL par prompt texte médian	Données de mai 2025, non vérifiées par un tiers
Google — accélérateur actif seul [7]	Consommation des seuls TPU et GPU actifs	0,10 Wh · 0,02 gCO_2e · 0,12 mL	Jugé « optimiste » par Google lui-même
Mistral AI, avec Carbone 4 et l'ADEME, 22 juillet 2025 [8]	Analyse de cycle de vie : entraînement de Mistral Large 2 et inférence, hors terminaux des utilisateurs	1,14 gCO_2e · 45 mL d'eau par réponse de 400 tokens	Méthodologie Frugal AI de l'AFNOR ; revue par Resilio et Hubblo

Ces valeurs ne se comparent pas entre elles : unité fonctionnelle, phases du cycle de vie et frontières diffèrent. L'écart entre les deux estimations de Google — d'un facteur 2,4 sur l'énergie — tient au seul choix du périmètre.

> **ESTIMATIONS DE RECHERCHE : MESURÉ N'EST PAS INFÉRÉ**
>
> L'étude How Hungry is AI? explore la variabilité entre une trentaine de modèles, à partir de données de performance publiques des API et de configurations matérielles inférées statistiquement, non de mesures d'opérateurs [9]. Dans une publication institutionnelle, mieux vaut ne pas reprendre ces valeurs comme facteurs universels et n'utiliser les chiffres des opérateurs que dans leurs limites déclarées. Faute d'unité fonctionnelle, de frontières et de facteurs communs, on parle de comparaison exploratoire, pas de réduction de CO_2 démontrée.

03 / TROIS RÈGLES DE LECTURE

1

Le cycle de vie ne se limite pas au GPU actif

Machines hôtes, capacité inactive allouée, refroidissement, stockage et réseaux contribuent aux consommations ; le matériel a des impacts de fabrication et de fin de vie. La fabrication est de loin la phase la plus contributrice aux impacts embarqués des cartes graphiques, tandis que l'usage pour l'IA reste la phase la plus impactante au total [4]. Il faut documenter les frontières et distinguer production, exploitation et fin de vie [4] [5].

2

Plusieurs indicateurs, plusieurs méthodes comptables

Le choix d'un serveur, d'une région ou d'un mode de refroidissement peut modifier différemment énergie, empreinte carbone et eau. Les méthodes de comptabilité électrique fondées sur la localisation (location-based) et sur le marché (market-based) reposent sur des hypothèses distinctes ; des achats contractuels d'électricité n'annulent ni les contraintes locales de réseau ni les impacts matériels. Les déclarations doivent préciser la méthode utilisée [10].

3

Des référentiels non interchangeables

UIT-T L.1801 concerne l'évaluation par analyse de cycle de vie des systèmes d'IA et leur comparaison à d'autres solutions ; ISO/IEC 21031:2024 spécifie un score d'intensité carbone logicielle (SCI), qui n'est pas une ACV multicritère. Aucun de ces textes n'est une « certification de sobriété » applicable de plein droit à un produit [5] [6].

Ne jamais transférer 0,24 Wh à une tâche agentique multi-étapes, à une autre région, à un autre modèle ou à un autre opérateur. **Une enquête publique ne remplace pas une instrumentation technique spécifique.**

04 Une unité de comparaison plus utile : la tâche acceptée

L'optimisation d'un système professionnel ne devrait pas commencer par « quel modèle consomme le moins par appel ? », mais par : combien de ressources sont utilisées pour obtenir un résultat métier exploitable ?

Considérons une analyse documentaire. La première configuration mobilise un grand modèle, une recherche documentaire et une validation humaine. La seconde utilise un modèle plus léger, mais nécessite davantage de reprises. Le calcul pertinent inclut toutes les tentatives, les appels d'outils, les vérifications et les échecs. Si la seconde produit davantage de conclusions erronées, elle peut déplacer le travail vers les collaborateurs, sans améliorer le bilan de bout en bout.

Nous proposons de suivre séparément :

Indicateur	Objet	Précaution de lecture
Travaux acceptés	Nombre de livrables conformes aux critères métier	Ne pas compter les réponses simplement générées
Énergie par travail accepté	Énergie de tous les essais / livrables acceptés	Dépend de données d'infrastructure disponibles
Émissions estimées	Énergie et équipements avec facteurs documentés	Publier périmètre, zone, période et incertitude
Eau / matériel	Consommation et impacts alloués selon méthode	Éviter d'assimiler eau prélevée et eau consommée
Efficacité professionnelle	Temps, coûts, qualité, corrections et utilité	Un coût inférieur ne prouve pas un impact moindre

Cette grille est une proposition opérationnelle Spotwork.AI, non une méthode normalisée de calcul carbone en elle-même.

APPELS PAR RÉSULTAT ACCEPTÉ

Tous les appels, échecs inclus, divisés par les réussites. Comparer des tâches et des permissions similaires.

IMPACT NET DE SUBSTITUTION

Variante IA comparée au scénario sans IA ; documenter le scénario évité et les éventuels effets rebond.

Ce contrat d'évaluation n'est ni une norme d'ACV ni une méthode de comptabilité GES certifiée : il prépare une étude conforme et vérifiable. Il rejoint l'unité de coût proposée dans notre analyse « Combien coûte réellement l'IA générative en entreprise ? Du token au livrable accepté » : le livrable accepté.

05 La méthode en cinq décisions

DÉCISION 1

Définir l'utilité

Identifier les tâches réellement nécessaires, leur destinataire, les critères d'acceptation et l'alternative sans IA. Éliminer les générations sans usage identifié avant toute optimisation technique.

Le résultat est accepté sur des critères de qualité, de droits et de conformité prédéfinis.

DÉCISION 2

Définir le périmètre

Séparer entraînement et inférence, équipements et énergie d'exploitation, impacts directs du service et effets indirects de son adoption. Décrire la méthode d'allocation lorsque le même modèle sert plusieurs usages.

Décrire modèles, versions, outils, recherche et ingestion, stockage, infrastructures, régions et périodes ; énoncer ce qui est exclu, et pourquoi.

DÉCISION 3

Collecter les preuves

Demander au fournisseur ses données d'énergie, d'eau, d'équipements, de localisation et de méthode ; instrumenter les coûts, appels, retries et latences côté application. Les données manquantes doivent rester indiquées comme telles.

Classer chaque valeur : mesure directe, mesure tierce, estimation explicite, donnée absente.

DÉCISION 4

Comparer des configurations à qualité équivalente

Tester les mêmes scénarios, droits et critères de réussite, y compris cas difficiles et refus corrects. Un incident critique de sécurité ou une dégradation métier substantielle peut rendre une optimisation inacceptable.

Compter tous les essais et les échecs ; conserver l'écart de réussite et les incidents critiques.

DÉCISION 5

Décider et surveiller

Mettre en production la variante retenue dans un périmètre défini ; mesurer l'évolution des volumes, la consommation, les reprises et les effets rebond. Une meilleure efficacité unitaire ne garantit pas une baisse de consommation totale si le nombre d'usages augmente.

GO LIMITÉ | CORRECTION PUIS RETEST | NO-GO

Aucun succès ponctuel ne prouve une performance environnementale à grande échelle. Une véritable ACV suit les référentiels applicables et s'appuie sur des données suffisantes.

06 Exemple illustratif : choisir entre deux workflows

SIMULATION Simulation sans mesures réelles de Spotwork.AI. Une équipe traite un lot de 1 000 demandes. Le workflow A exécute 1 000 appels (1 par demande) et produit 900 livrables acceptés. Le workflow B exécute 1 300 appels (1,3 par demande en moyenne) et produit 950 livrables acceptés. Les acceptations portent sur les demandes traitées, non sur les appels individuels. On observe donc environ 1,11 appel par résultat accepté pour A et 1,37 pour B.

Sur 1 000 demandes	Workflow A	Workflow B
Appels exécutés	1 000	1 300
Résultats acceptés	900	950
Cas non acceptés à traiter	100	50
Appels par résultat accepté	**1 000 / 900 ≈ 1,11**	**1 300 / 950 ≈ 1,37**

Cette illustration ne permet pas de conclure que A consomme moins d'énergie : les modèles, longueurs des requêtes et infrastructures peuvent différer. Elle montre seulement que le nombre d'appels par tâche acceptée apporte davantage d'information que le coût d'un appel isolé. La meilleure configuration dépend encore des impacts mesurés et du niveau de qualité exigé.

La comparaison exige les données d'énergie et de matériel, la qualification des travaux acceptés, les coûts de validation et le traitement des 100 ou 50 cas non acceptés. Les pistes d'optimisation — réduire les reprises et recherches, alléger le contexte inutile, router vers des modèles spécialisés — n'ont de valeur que si critères de qualité, droits, volume et limites de latence restent constants.

Protocole de test et règles d'arrêt

La recette comporte au minimum : un cas simple, un cas documentaire complexe, une source absente, une réponse nécessitant abstention, une récupération répétée, une panne de connecteur, un changement de modèle et un pic de volume. Nombre de cas, répétitions, budget et durée sont fixés avant l'essai ; variance et incertitude environnementale sont rapportées, non masquées par un score moyen.

ARRÊT SÉCURITÉ OU MÉTIER

Donnée non autorisée, validation contournée ou erreur critique.

ARRÊT TECHNIQUE

Budget, délai ou nombre d'appels dépassé.

ARRÊT MÉTHODOLOGIQUE

Données insuffisantes pour revendiquer une réduction carbone : les gains de coût ou de latence restent publiables, sans être renommés « réduction d'empreinte ».

07 Comment éviter le greenwashing ?

L'UIT a approuvé en février 2026 la recommandation UIT-T L.1801, fondée sur l'analyse du cycle de vie (ACV) pour évaluer les systèmes d'IA. Elle prévoit aussi l'examen des effets liés à l'utilisation de l'IA par rapport à des solutions alternatives ; elle ne détermine pas à l'avance si l'IA est bénéfique à l'environnement. L'ISO/IEC 21031:2024 traite pour sa part de l'intensité carbone d'un système logiciel. Ces deux référentiels ne doivent pas être décrits comme équivalents [5] [6].

Une communication solide distingue mesure, estimation, hypothèse et bénéfice futur, et ne promet pas de réduction de CO_2 à partir de seules économies financières, d'un nombre de prompts réduit ou d'un hébergement présenté comme « vert » sans comptabilité explicite.

Quatre niveaux d'affirmation, quatre niveaux de preuve

01 **« Nous avons réduit le nombre d'appels »**

Possible si seuls les journaux applicatifs sont disponibles.

02 **« Nous avons mesuré moins d'énergie pour le service étudié »**

Seulement si le périmètre, les mesures et le protocole le prouvent.

03 **« Nous avons réduit les émissions »**

Uniquement si des facteurs vérifiés et des frontières appropriées permettent le calcul.

04 **« L'usage de l'IA améliore l'environnement »**

Seulement si la comparaison avec l'alternative sans IA et ses effets indirects a été examinée.

CE QU'UNE POLITIQUE D'ACHAT PEUT DEMANDER AU FOURNISSEUR

Périmètre des données énergétiques, région et mix électrique, méthodes d'attribution, matériel et taux d'utilisation, eau, période et possibilités de vérification. Les réponses partielles sont utiles si elles sont étiquetées ; elles ne deviennent pas une certitude par la seule existence d'un tableau de bord. Une formule du type « jusqu'à X % plus vert » n'est pas justifiée par une comparaison de coûts ou de tokens isolée.

08 Le rôle de l'architecture professionnelle

Une orchestration raisonnée peut être une piste de sobriété : affecter les modèles aux tâches adaptées, borner les retries, éviter les récupérations inutiles, conserver les résultats validés et suivre qualité et coût. Ce potentiel doit être testé dans l'environnement réel ; aucune économie d'énergie ou de carbone de Spotwork.AI n'est mesurée ni revendiquée dans cet article.

L'objectif est double : employer l'IA lorsqu'elle crée une utilité démontrable, puis rendre l'exécution de cette IA aussi efficiente et contrôlée que possible. Le premier volet relève de la pertinence métier ; le second, de l'ingénierie et de l'évaluation environnementale. Les confondre fragilise les décisions RSE.

Qui fait quoi dans une expérimentation

Rôle	Responsabilité
Propriétaire de l'usage	Définit la valeur professionnelle et les critères d'acceptation
Responsable technique	Décrit le workflow complet et ses dépendances
Équipe RSE ou ACV	Définit périmètres, indicateurs environnementaux et limites d'allocation
Achats ou DSI	Sollicitent les fournisseurs pour obtenir des informations contextualisées
Responsable sécurité	Contrôle les frontières d'accès et les contraintes de traitement

Aucun de ces rôles ne doit être remplacé par un score d'IA auto-attribué. Un registre de décisions conserve la version des configurations comparées, la distribution des tâches, les résultats acceptés, les données du prestataire, les hypothèses de calcul, les intervalles d'incertitude et les arbitrages refusés. Un même cas n'est pas comptabilisé comme gain à plusieurs endroits ; les évaluations sont reconduites lorsque changent modèles, régions, volumes, matériel ou qualité des résultats.

L'IA comme outil de pilotage RSE — analyse de données, extraction de justificatifs, consolidation du reporting — est un sujet distinct : ses résultats et sa propre empreinte s'examinent séparément. **Un premier pilote vise à rendre visibles les inconnues, pas à publier une réduction de CO_2.**

09 Périmètre et sources

Cette publication synthétise des travaux de l'AIE, de l'ADEME, de l'UIT, de l'ISO et de Google. Elle propose un cadre décisionnel, sans bilan carbone propriétaire, audit d'un fournisseur ou certification Spotwork.AI. Le cas numérique est fictif et limité à la comparaison des volumes d'appels.

Les chiffres de l'AIE concernent les centres de données ; ceux de Google, l'inférence d'un service déterminé. Les normes de l'UIT et de l'ISO proposent des méthodes sans certifier de technologie. L'ADEME décrit les enjeux et les frontières pertinentes ; ses conclusions ne servent pas de coefficient d'émission par prompt. Aucune mesure énergétique, hydrique ou carbone de Spotwork.AI n'a été réalisée pour cette publication.

[1] AIE — Energy and AI, Executive Summary (avril 2025).
iea.org/reports/energy-and-ai/executive-summary

[2] AIE — Powering Artificial Intelligence: Strategies for integrating data centres into electricity systems (14 octobre 2026).
iea.org/reports/powering-artificial-intelligence

[3] ADEME — Avis : l'intelligence artificielle générative, des impacts environnementaux importants (juillet 2026).
librairie.ademe.fr/economie-circulaire-et-dechets/9495-avis-de-l-ademe-l-intelligence-artificielle-generative-des-impacts-environnementaux-importants.html

[4] ADEME — Analyse de cycle de vie de GPU (cartes graphiques) pour l'intelligence artificielle (mars 2026).
librairie.ademe.fr/economie-circulaire-et-dechets/9103-analyse-de-cycle-de-vie-de-gpu-cartes-graphiques-pour-l-intelligence-artificielle.html

[5] UIT-T L.1801 (02/2026) — Guidelines for assessing the environmental impact of artificial intelligence systems.
itu.int/epublications/fr/publication/itu-t-l-1801-2026-02-guidelines-for-assessing-the-environmental-impact-of-artificial-intelligence-systems

[6] ISO/IEC 21031:2024 — Software Carbon Intensity (SCI) specification.
iso.org/standard/86612.html

[7] Google — Measuring the environmental impact of AI inference (21 août 2025).
cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference

[8] Mistral AI, avec Carbone 4 et l'ADEME — Our contribution to a global environmental standard for AI (22 juillet 2025).
mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai

[9] Jegham et al. — How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference (arXiv, 2025).
arxiv.org/abs/2505.09598

[10] GHG Protocol — Scope 2 Guidance (2015).
ghgprotocol.org/scope-2-guidance

Spotwork.AI

L'IA professionnelle de confiance

Mieux travailler aujourd'hui,
construire votre intelligence de demain.

SPOTLAB · RESEARCH & INSIGHTS · ANALYSE P23

Impact environnemental de l'IA : comment piloter énergie, coût et utilité ?

Mesurer les impacts, comparer les usages et éviter les faux gains

© 2026 SPOTWORK.AI