arcep
autorité de régulation
des communications électroniques,
des postes et de la distribution de la presse

RÉPUBLIQUE FRANÇAISE

INTELLIGENCE ARTIFICIELLE GENERATIVE : QUELS DEFIS ENVIRONNEMENTAUX ?

Mai 2026

Avec la collaboration du Pôle d'Expertise de la Régulation Numérique – PEReN

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise
de la Régulation Numérique

ISSN n°2258-3106

Table des matières

Résumé

1. Les défis environnementaux d'un développement accéléré de l'intelligence artificielle générative

L'intelligence artificielle (IA) générative représente une rupture technologique majeure, permettant la création automatique de textes, images, sons ou vidéos à partir de simples instructions en langage naturel. Son adoption massive et rapide par le grand public (48 % des Français l'utilisent en 2025, contre 20 % en 2023) et les entreprises ouvre des perspectives d'importants gains de productivité et d'innovation.

L'IA générative peut également permettre des actions innovantes en faveur de la transition environnementale et de la réduction d'impacts environnementaux de différents secteurs. Elle apparaît toutefois elle-même comme un facteur d'accélération de nombreux impacts environnementaux.

En effet, l'IA générative repose, comme tout service numérique, sur des infrastructures matérielles (centres de données, réseaux et terminaux) et se distingue par d'importants besoins en capacités de calcul. Ces capacités de calcul sont hébergées dans les centres de données et induisent un changement d'ordre de grandeur des puissances électriques nécessaires pour les alimenter.

La croissance de la demande et les perspectives de développement du marché incitent les acteurs à se placer en tant que leaders tout au long de la chaîne de valeur et à consentir des investissements massifs pour développer des modèles de grande taille et accéder aux intrants essentiels comme la puissance de calcul (puces, serveurs, centres de données) et l'énergie.

Cette course à l'innovation peut engendrer des tensions croissantes sur des ressources disponibles en quantité limitée (énergie, eau, métaux, terres rares). En particulier, les dynamiques de marché impliquent un recours accru aux centres de données, dont les capacités doivent continuellement s'adapter à des charges de calcul en forte croissance.

Ces dynamiques de marché invitent à interroger les conditions de soutenabilité du modèle actuel de développement de l'IA et à identifier les leviers susceptibles d'en atténuer les impacts environnementaux. À ce titre, une évaluation rigoureuse et transparente des impacts directs de l'IA générative apparaît comme un préalable indispensable pour appréhender de manière éclairée la soutenabilité d'un tel développement.

2. L'évaluation de l'empreinte environnementale de l'IA

Les données disponibles sur l'impact environnemental de l'IA générative restent à ce jour très limitées, notamment du fait d'un manque de transparence des acteurs : 84 % des modèles d'IA ne font l'objet d'aucune information environnementale. Une meilleure mesure de l'impact environnemental de l'IA sur l'ensemble de la chaîne de valeur nécessite d'utiliser des méthodologies robustes, fiables (multi-critères, multi-composants, multi-étapes) et harmonisées, pour permettre la comparabilité des modèles et des services. Des méthodologies par analyse de cycle de vie (ACV) standardisées et reconnues à l'international existent déjà et doivent être mobilisées et diffusées, pour éviter un foisonnement des méthodes et ainsi s'appuyer sur une base commune

Pour contribuer à l'évaluation de l'empreinte environnementale de l'IA, l'Arcep s'est appuyée sur les résultats issus de la littérature scientifique, complétés par des travaux institutionnels ou de l'industrie.

L'Arcep a également mené des travaux avec le Pôle d'expertise de la régulation numérique (PEReN), qui apportent des résultats inédits sur l'impact environnemental de la phase d'utilisation (inférence).

Les éléments d'appréciation de l'empreinte environnementale de l'IA générative recueillis distinguent les impacts par brique du numérique (centres de données, terminaux, réseaux) puis selon les phases d'entraînement et d'inférence d'un modèle.

L'analyse par brique montre que les centres de données sont la composante structurante des impacts environnementaux de l'IA générative. Leur développement soulève notamment des questions liées à la consommation d'énergie et par extension aux émissions de gaz à effet de serre qui dépendent du mix électrique du pays d'implantation. À titre d'illustration, en moyenne en 2024, le mix polonais possédait un contenu carbone 1,65 fois plus élevé que celui du mix irlandais et 17,6 fois plus élevé que celui du mix français. Il soulève également des questions liées à la consommation d'eau et à l'artificialisation des sols. De fait, le développement de l'IA est considéré comme le principal facteur de croissance de la consommation électrique des centres de données, qui pourrait doubler au niveau mondial d'ici à 2030. Des grands acteurs du numérique ont d'ores et déjà indiqué que l'atteinte de leurs objectifs initiaux de décarbonation était compromise à cause du développement de l'IA générative et ont revu leurs objectifs à la baisse. Au-delà des stratégies des acteurs industriels, la croissance de la consommation électrique des centres de données soulève des enjeux à l'échelle des nations et de leurs trajectoires de décarbonation.

Les effets environnementaux de l'IA générative liés aux terminaux (intensité énergétique d'utilisation des services d'IA générative potentiellement plus élevée et risque de renouvellement des terminaux) et aux réseaux (capacités *a priori* suffisantes à court terme mais des interrogations à moyen terme) sont encore incertains.

L'analyse par phase montre que l'impact environnemental de l'entraînement des modèles dépend principalement de leur taille (nombre de paramètres), du mix électrique et du matériel utilisé (puces, serveurs). Elle relève également que le nombre de modèles et leur taille n'ont cessé de croître.

Les impacts environnementaux de la phase d'inférence sont moins documentés mais pourraient devenir un enjeu significatif en raison de la croissance des usages. Pour enrichir l'analyse, des travaux inédits ont été réalisés par le PEReN sur la consommation électrique des prompts (requêtes). Ils montrent qu'en phase d'inférence, les plus grands modèles sont toujours les plus consommateurs d'énergie mais aussi que des modèles de tailles très différentes peuvent avoir la même consommation énergétique. En outre, certains modèles qui consomment moins sont capables de fournir des réponses aussi pertinentes que des grands modèles. Limiter la consommation énergétique ne revient donc pas nécessairement à faire des compromis sur la performance du modèle. Les tests réalisés mettent également en évidence que des techniques d'optimisation et de compression permettent des gains énergétiques significatifs en phase d'inférence.

3. Nos recommandations afin de rendre le développement de l'IA compatible avec les limites planétaires

Afin de s'assurer que l'IA et ses infrastructures se déploient dans des conditions compatibles avec les limites planétaires et permettant aux générations futures de bénéficier de son potentiel, l'Arcep formule quatre axes d'actions couvrant neuf recommandations. Ces recommandations s'articulent avec le cadre législatif européen, en vigueur ou à venir.

- **Axe 1 : Améliorer la mesure et la connaissance de l'impact environnemental de l'IA**
  - **Mettre en œuvre la collecte et la publication des données environnementales de l'IA par des autorités publiques**
  - **Utiliser des méthodologies internationalement standardisées** d'évaluation de l'impact environnemental, afin de faciliter les comparaisons entre systèmes d'IA
- **Axe 2 : Promouvoir l'écoconception des services d'IA comme levier stratégique de la compétitivité européenne**
  - **Intégrer l'écoconception des services d'IA dans la régulation européenne des fournisseurs**
  - **Renforcer l'écoconditionnalité** dans les modalités de soutien à l'innovation et la commande publique
- **Axe 3 : Donner les moyens aux utilisateurs de choisir leurs services d'IA générative en fonction de leur impact environnemental grâce à une régulation européenne adaptée**
  - **Imposer une plus grande transparence environnementale** aux fabricants de puces et aux grands fournisseurs de modèles et de services d'IA
  - **Garantir l'ouverture des services d'IA**
- **Axe 4 : Construire une stratégie de développement des centres de données en Europe alliant souveraineté et soutenabilité**
  - **Éclairer les choix publics et d'investissement** : faire de la transparence et de la régulation par la donnée des atouts du développement des centres de données en Europe
  - **Renforcer la coordination européenne entre politiques numérique, énergétique et d'infrastructure pour accompagner le développement des centres de données**
  - **Encourager une implantation territoriale concertée des centres de données**

L'Arcep poursuivra ses travaux et les échanges avec l'ensemble des parties prenantes afin d'œuvrer collectivement à des services d'IA générative écoconçus et des infrastructures numériques soutenables.

Introduction

L'intelligence artificielle générative (ci-après « IA générative ») s'est imposée en quelques années seulement comme une innovation majeure et largement adoptée par les utilisateurs particuliers et les entreprises.

Alors que l'intelligence artificielle recouvre un ensemble de technologies développées progressivement depuis plusieurs décennies et qui ont donné lieu à un ensemble d'outils et applications intégrés dans notre quotidien (GPS et navigation, moteurs de recherche, recommandation de contenus, assistants vocaux, détection de fraude…), l'arrivée de l'IA générative constitue un tournant majeur. L'IA générative permet en effet de générer automatiquement de nouveaux contenus sous la forme de texte, mais aussi d'images, de son ou de vidéos. À travers une nouvelle génération d'outils, tels que ChatGPT, Gemini, Claude ou Le Chat, de simples instructions en langage naturel (appelées « prompts »), par écrit ou à l'oral, permettent d'obtenir rapidement la génération de contenus qui pouvaient auparavant nécessiter des tâches humaines complexes.

Les nouvelles capacités d'automatisation, d'assistance et de création ouvertes par ces outils laissent entrevoir la possibilité d'importants gains de productivité et d'innovation, et en font des leviers de croissance potentiels[1]. La maîtrise de ces technologies emporte dès lors des enjeux de compétitivité mais aussi de souveraineté. En effet, le retard en matière d'intelligence artificielle pourrait créer des risques de dépendance vis-à-vis d'autres pays ou régions du monde. Ces enjeux amènent à un soutien massif à cette technologie et aux infrastructures qui la sous-tendent par les puissances publiques à travers le monde. Avec son plan d'action pour le continent de l'IA (« AI Continent Action Plan »), publié en 2025[2], la Commission européenne a notamment affiché sa volonté de devenir un acteur mondial de premier plan dans le domaine de l'IA *via* un ensemble d'initiatives, dont un plan d'investissement de 200 milliards d'euros. En France, le Gouvernement a lancé une stratégie nationale pour l'IA dès 2018, pour structurer l'écosystème IA sur le long terme à tous les stades du développement technologique (recherche, développement et innovation, applicatifs, mise sur le marché et diffusion intersectorielle, soutien et encadrement du déploiement). Près de 2,5 milliards d'euros du plan France 2030 y sont dédiés[3]. De telles initiatives participeront à accroître le développement de capacités et de compétences spécialisées pour l'IA, notamment générative.

Face aux opportunités ouvertes, l'adoption déjà fulgurante des services d'IA générative par le grand public et les entreprises ainsi que les perspectives de développement de cette technologie dans les années à venir soulèvent la question de sa soutenabilité environnementale. Si l'IA, notamment générative, peut permettre des actions en faveur de la transition environnementale et de la réduction d'impacts environnementaux de différents secteurs, elle apparaît elle-même comme un facteur d'accélération potentielle de nombreux impacts environnementaux. Construction de puces et de centres de données, consommation électrique des serveurs pour l'entraînement de modèles, consommation de ressources énergétiques ou en eau à chaque requête : le fonctionnement des services d'IA générative, loin d'être « immatériel », repose sur une matérialité et a des impacts environnementaux

[1] Dans le rapport « IA : notre ambition pour la France » de la Commission de l'intelligence artificielle publié en mars 2024, A. Bouverot et P. Aghion estiment par exemple que l'automatisation de certaines tâches pourrait permettre de doubler la croissance économique annuelle de la France, avec, au bout de dix ans, une hausse de PIB comprise entre 250 et 420 Mds€.

[2] Le plan d'action pour le continent de l'IA | Bâtir l'avenir numérique de l'Europe

[3] La stratégie nationale pour l'intelligence artificielle

Cette situation génère des inquiétudes légitimes qui sont renforcées aujourd'hui par le manque de transparence des acteurs de la chaîne de valeur. Or, l'accessibilité des informations constitue un enjeu central pour s'assurer que l'IA se déploie dans des conditions compatibles avec les limites planétaires, permettant aux générations futures de bénéficier de son potentiel.

Partant de ce constat, l'Arcep s'est engagée, dans le cadre de sa stratégie « Ambition 2030[4] », à « *approfondir et partager la connaissance de l'impact environnemental de l'intelligence artificielle* ». Cette action s'inscrit dans la lignée des travaux de l'Arcep au cours des dernières années sur les enjeux de l'empreinte environnementale du numérique, tels que l'étude ADEME-Arcep sur l'empreinte environnementale du numérique en 2020, 2030 et 2050[5] ou encore la collecte de données environnementales auprès des acteurs du numérique, restituées depuis 2022 dans le cadre de l'enquête annuelle pour un numérique soutenable[6]. En participant à accroître la connaissance des impacts environnementaux du numérique, sur la base de données et méthodologies fiables, l'Arcep entend proposer des leviers d'actions permettant de réduire ces impacts.

Le présent rapport constitue une nouvelle étape dans l'engagement de l'Arcep pour une meilleure évaluation de l'impact environnemental de l'IA, et en particulier de l'IA générative. Ce travail s'appuie sur l'audition de nombreux experts issus des secteurs privé, public, académique et associatif, tout en établissant une revue de littérature scientifique et institutionnelle. Le rapport s'appuie également sur une étude inédite menée en collaboration avec le Pôle d'expertise de la régulation numérique[7] (PEReN), qui a permis la réalisation de tests de consommation électrique en phase d'utilisation de plusieurs outils d'IA générative.

Ce rapport s'organise autour des axes suivants :

- tout d'abord, il analyse les enjeux environnementaux soulevés par la matérialité de l'IA générative et son essor fulgurant du côté de l'offre mais aussi de la demande ;
- ensuite, il propose une synthèse des connaissances sur l'empreinte environnementale de l'IA générative, permettant de fournir de premières évaluations et de partager un socle de connaissances fiables. Les résultats issus de la littérature académique, de sources institutionnelles ou encore de l'industrie ont été complétés par de nouveaux résultats issus des travaux menés en collaboration avec le PEReN, permettant d'analyser les facteurs clés d'impact de la consommation électrique de modèles d'IA générative en phase d'utilisation ;
- enfin, s'appuyant sur ce diagnostic, l'Autorité formule quatre axes de recommandations :
    - Axe 1 : Améliorer la mesure et la connaissance de l'impact environnemental de l'IA
    - Axe 2 : Promouvoir l'écoconception des services d'IA comme levier stratégique de la compétitivité européenne
    - Axe 3 : Donner les moyens aux utilisateurs de choisir leurs services d'IA générative en fonction de leur impact environnemental grâce à une régulation européenne adaptée
    - Axe 4 : Construire une stratégie de développement des centres de données en Europe alliant souveraineté et soutenabilité

[4] Arcep_Ambitions-2030_objectifs_actions_janv2025.pdf

[5] Etude ADEME – Arcep sur l'empreinte environnementale du numérique en 2020, 2030 et 2050 | Arcep

[6] Enquête annuelle "Pour un numérique soutenable" - édition 2025 (données 2023) | Arcep

[7] Le Pôle d'expertise de la régulation numérique (PEReN) est un service public à compétence nationale. Il accompagne les pouvoirs publics dans la régulation des plateformes numériques et de l'IA, en mobilisant une expertise technique souveraine, mutualisée et à l'état de l'art. De l'audit d'algorithmes à l'analyse de données, en passant par le développement d'outils innovants ; lorsque cela est possible, le PEReN partage ses productions en open source pour une régulation transparente et collaborative.

Ces axes, déclinés en 9 recommandations, visent à concilier développement de l'IA, notamment générative, et soutenabilité, afin que cette technologie se déploie dans des conditions favorables pour l'ensemble de la société.

Ce rapport complète ainsi les travaux réalisés par l'Arcep sur les enjeux soulevés par le développement de l'IA générative en matière d'ouverture d'internet, qui ont fait l'objet d'un rapport publié en janvier 2026[8].

[8] IA générative : des défis pour l'avenir de l'internet ouvert

Chapitre 1 Les défis environnementaux d'un développement accéléré de l'intelligence artificielle générative

Comme tout service numérique, l'IA générative repose sur des infrastructures et équipements physiques qui vont des terminaux utilisateurs aux centres de données en passant par les réseaux de communications électroniques. Cette matérialité de l'infrastructure supportant les services d'IA générative, combinée à l'essor rapide de l'usage de ces services et des dynamiques de marché intenses, soulève des enjeux environnementaux croissants. En effet, si l'intelligence artificielle peut offrir des perspectives encourageantes pour la protection de l'environnement, ces opportunités ne doivent pas conduire à négliger les effets adverses liés à ces dynamiques.

Ce chapitre revient d'abord sur la matérialité des services d'IA générative (1.1), en tant que services numériques supportés par une infrastructure et des équipements physiques (1.1.1), puis décrit les particularités techniques et matérielles de cette technologie pouvant renforcer les enjeux environnementaux attachés à ces services (1.1.2). Le chapitre analyse ensuite les enjeux environnementaux découlant de l'essor de l'IA générative (1.2). L'adoption massive de cette technologie et les investissements considérables qu'elle attire (1.2.1) entraînent d'intenses dynamiques concurrentielles aux différents niveaux de la chaîne de valeur (1.2.2). Ces dynamiques de marché peuvent générer des tensions sur certaines ressources et présenter un fort enjeu environnemental (1.2.3). Enfin, une dernière section est consacrée à l'examen des effets environnementaux nets de l'intelligence artificielle, dont l'évaluation demeure, à ce stade, incertaine, en raison notamment de la complexité des effets indirects du numérique et du caractère encore émergent des travaux de recherche en la matière (1.3).

1.1 La matérialité des services d'IA générative

1.1.1 Des services d'IA générative s'appuyant sur des équipements et infrastructures matériels, à l'instar des autres services numériques

À l'instar de tout service numérique, les services d'IA générative reposent sur des infrastructures et équipements matériels qui peuvent être classés en trois grandes catégories ou briques : les terminaux utilisateurs, les réseaux et les centres de données.

Chaque usage numérique – tel que consulter des réseaux sociaux ou un site internet – fait en effet appel à ces trois briques. Pour fonctionner, le service va nécessiter un terminal utilisateur (ex : smartphone, ordinateur…) sur lequel il sera consommé. Si cette brique est généralement la plus tangible, les réseaux (ex : ADSL, fibre optique, box internet…) et les centres de données, qui hébergent les serveurs informatiques, sont également nécessaires pour transmettre les données, les traiter et les stocker, afin de fournir un service numérique. Ces trois briques sont interdépendantes et peuvent influer l'une sur l'autre ; elles ne doivent donc pas être analysées de manière isolée. Par exemple, l'essor des smartphones et d'un écosystème de développeurs d'applications pour ce type de terminaux a rendu accessibles de nombreux services tels que le streaming vidéo, les appels vidéo, les jeux en ligne et des outils numériques comme les services bancaires et les démarches administratives. Partant, le trafic de données a augmenté, influençant à la hausse le dimensionnement des réseaux et des centres de données.

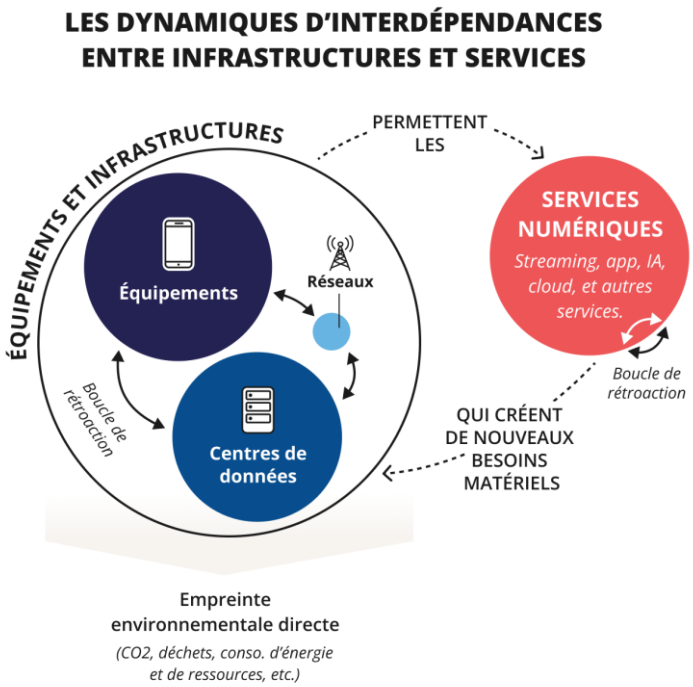


Figure 1 : L'interdépendance entre infrastructures et usages numériques. Source : Arcep.

Ces dernières années, plusieurs études ont été menées sur les impacts environnementaux du numérique par divers acteurs, académiques, privés ou institutionnels. Ces études ont permis de répondre à un déficit de connaissances sur le sujet. Bien que l'IA générative ne soit pas spécifiquement prise en compte dans ces études, du fait de sa dynamique très récente, les enseignements qu'elles apportent apparaissent pertinents pour comprendre et mesurer l'impact environnemental des services numériques dans leur ensemble.

En particulier, l'étude conjointe, menée par l'ADEME et l'Arcep à la demande du gouvernement, sur l'empreinte environnementale du numérique en 2020, 2030 et 2050[9] a permis de mettre en évidence, pour l'analyse des impacts environnementaux du numérique, la pertinence d'une approche qui doit être à la fois :

- **Multi-étapes** : l'évaluation de l'impact de la seule phase d'utilisation n'est pas suffisante, toutes les phases du cycle de vie des équipements doivent être prises en compte (fabrication, distribution, utilisation, fin de vie) ;
- **Multi-composants** : les trois briques constitutives de l'infrastructure (terminaux, réseaux et centres de données) doivent être analysées ;
- **Multi-critères** : les impacts environnementaux du numérique ne se limitent aux émissions de gaz à effet de serre ; la consommation d'eau, d'énergie, de métaux et ressources rares notamment doit également être analysée.

À partir de cette approche, l'étude montre qu'en 2022, le numérique en France représentait environ 4 % de l'empreinte carbone nationale et environ 10 % de la consommation électrique nationale. Elle

[9] Pour aller plus loin sur les impacts environnementaux du numérique : Etude ADEME – Arcep sur l'empreinte environnementale du numérique en 2020, 2030 et 2050 | Arcep

souligne que les services numériques et la « dématérialisation » des échanges qu'ils permettent, dépendent d'infrastructures et de terminaux bel et bien matériels et consommateurs de ressources.

L'analyse prospective met également en évidence que sans action pour limiter la croissance de l'impact environnemental du numérique en France, son empreinte carbone pourrait tripler entre 2020 et 2050 et la consommation d'énergie pourrait doubler sur la même période. Une telle évolution, projetée avant même les dynamiques récentes de développement de l'IA générative, semble donc difficilement compatible avec l'ambition de la France de stabiliser l'empreinte carbone du numérique d'ici 2030-2035, soit l'objectif envisagé dans la « Stratégie nationale bas-carbone[10] ».

La dynamique de croissance des usages numériques soulève également un enjeu central autour de l'épuisement des métaux et autres ressources stratégiques et de leur disponibilité future, pas seulement pour le numérique mais également pour d'autres secteurs, dans un contexte où la demande en métaux pour le numérique pourrait augmenter de +179 % entre 2020 et 2050 (ADEME, 2024b). La consommation de ces ressources pour le numérique s'observe également à travers la génération de déchets. À titre d'exemple, l'étude conjointe menée par l'ADEME et l'Arcep estime qu'une personne vivant en France en 2020 générait, pour ses seuls usages numériques, près de 300 kilos de déchets par an (y compris les déchets électriques et électroniques ainsi que les déchets liés à l'extraction de matières premières).

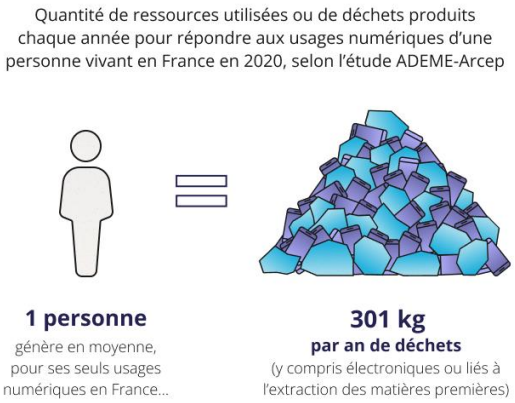


Figure 2 : Les biens et services numériques génèrent des déchets. Source : Arcep.

L'étude ADEME-Arcep a permis d'établir un diagnostic partagé et d'alerter sur la dynamique de l'empreinte environnementale du numérique. Si elle fournit des éléments méthodologiques qui permettent d'identifier les sources de pollution liées à l'entraînement et à l'usage des services d'IA générative, elle ne permet pas d'avoir une vision d'impact par technologie ou par service. Pour isoler spécifiquement la part de l'IA dans l'empreinte environnementale du numérique, d'autres études doivent être réalisées.

Les dynamiques intenses de développement de l'IA générative observées ces dernières années (cf. section 1.2.1) invitent à s'intéresser dès maintenant à la question spécifique des impacts environnementaux de l'IA générative, d'autant que celle-ci présente des particularités techniques et économiques, par rapport aux autres services numériques, qui peuvent avoir une influence sur ses impacts environnementaux.

[10] Stratégie Nationale Bas Carbone (SNBC3), résumé du projet de SNBC3 publié en décembre 2025.

1.1.2 Des particularités liées notamment au besoin en capacités de calcul

L'IA générative se distingue d'un service numérique « classique » notamment par le fait qu'elle nécessite une importante puissance de calcul informatique pour créer, entraîner et affiner les modèles d'IA, mais aussi pour utiliser les services s'appuyant sur l'IA générative.

Les services d'IA générative reposent en effet sur des modèles de fondation qui nécessitent d'importants besoins en puissance de calcul lors de deux phases clés :

- l'entraînement du modèle, qui désigne le processus d'apprentissage initial d'un modèle de fondation, où les données alimentent le modèle, et qui peut éventuellement être complété par une phase de spécialisation, appelée « *fine-tuning* », pendant laquelle le modèle va être adapté à une tâche particulière. Durant cet entraînement, les résultats sont examinés et la sortie du modèle est ajustée pour augmenter la performance (qualité de la réponse) et l'efficacité (temps et quantité de calcul, coût énergétique…) du modèle ;
- l'inférence, qui correspond à l'utilisation du modèle, une fois entraîné, par les utilisateurs finals, pour créer du contenu.

Lors de l'entraînement, un modèle d'IA générative est exposé à un grand ensemble de données (ex : textes, images, sons) afin notamment d'en apprendre les constantes, les structures sous-jacentes qui le caractérisent. Cette phase nécessite d'importantes ressources de calcul pour s'assurer que le modèle pourra générer des contenus cohérents et pertinents.

Une fois le modèle entraîné, il est utilisé, par son intégration au sein de services, pour répondre à des requêtes utilisateurs et générer du contenu (par exemple du texte ou des images, selon le cas d'usage). Cette phase dite d'inférence est, en général, moins consommatrice en puissance de calcul, mais cette consommation est directement liée au volume d'utilisation. Si chaque inférence prise isolément consomme nettement moins de ressources que l'entraînement, elle est en revanche réalisée à très grande échelle (potentiellement des milliards de fois) ce qui nécessite au global une puissance de calcul importante.

La puissance de calcul, nécessaire à l'entraînement du modèle mais également à son inférence, correspond matériellement à des serveurs généralement hébergés dans les centres de données, alimentés en électricité[11]. Or, l'essor de l'IA participe d'un changement d'ordre de grandeur des puissances électriques associées aux équipements informatiques et des infrastructures, tels que les cartes graphiques utilisées dans les serveurs, ou les récents projets de centres de données hébergeant ces serveurs. Les projets actuels de centres de données pour l'IA représentent ainsi, pour certains des plus conséquents, des puissances électriques de l'ordre du gigawatt[12] (GW). À titre de comparaison, les projets qui ont précédé l'essor rapide de l'IA ces dernières années présentent des puissances électriques plutôt de l'ordre de quelques dizaines de mégawatts voire de la centaine pour les plus grands. En parallèle, les nouvelles unités centrales de traitement (CPU) et les processeurs graphiques (GPU), composants centraux des serveurs pour l'IA, ont vu une augmentation importante de leur puissance électrique ces dernières années : certains acteurs prévoient des équipements pouvant atteindre 1 400 W d'ici 2026 et plus au-delà, par rapport à des équipements avoisinant 300 à 400 W en 2017[13].

[11] Pour l'inférence, certains acteurs envisagent d'effectuer cette phase sur le terminal qui fait la requête mais cela nécessite une puissance de calcul dont un grand nombre de terminaux utilisateurs ne disposent pas à date. C'est pourquoi le terme de terminal « compatible IA », notamment pour les smartphones commence à voir le jour.

[12] À titre de comparaison, cela correspond à l'ordre de grandeur de la puissance d'une tranche nucléaire.

[13] Gauthier Roussilhe, Les GPU et IA génératives : une nouvelle phase de l'histoire environnementale de la numérisation, Mars 2025

1.2 Des dynamiques de marché qui nécessitent des ressources

1.2.1 Une adoption massive des services d'IA et des investissements croissants

Outre ses spécificités matérielles, l'IA générative se distingue d'autres services numériques par un essor extrêmement rapide et un développement massif de ses usages. Selon l'édition 2026 du Baromètre du numérique (Arcep, 2026c), les services d'IA générative ont connu une diffusion d'une ampleur inédite en l'espace de seulement deux ans. Alors qu'en 2023, un cinquième de la population déclarait y avoir recours (20 %), ce chiffre atteint déjà près de la moitié en 2025 (48 % exactement), soit une progression de 28 points en deux ans. Ce phénomène s'observe à l'échelle mondiale. La rapidité d'adoption de l'IA générative par la population française est notamment comparable à celle observée aux États-Unis (Bick et *al*., 2024).

La progression des usages s'observe également dans la sphère professionnelle. En l'espace de deux ans, la part des entreprises de l'Union européenne de 10 salariés ou plus déclarant utiliser au moins un service d'IA est ainsi passée de 8 % en 2023 à 20 % en 2025, soit une progression de 12 points. Les entreprises de plus grande taille, avec 250 salariés ou plus, se démarquent par une adoption de l'IA encore plus rapide, passant de 30 % en 2023 à 55 % en 2025[14].

Le potentiel attendu du développement de cette technologie se reflète à travers des investissements très importants et croissants réalisés par les acteurs des services d'IA. Ces derniers doivent faire face à des coûts très élevés notamment pour l'accès à la puissance de calcul nécessaire (*i.e.* aux serveurs dédiés à l'IA[15] mais également aux centres de données pour les héberger). À titre d'illustration, pour l'année 2026, Amazon, Google, Meta et Microsoft ont annoncé à eux seuls des investissements totaux d'environ 650 milliards de dollars dans l'IA, après des investissements de 410 Mds$ en 2025 et 245 Mds$ en 2024[16]. Si ces montants peuvent soulever certaines interrogations quant à la capacité à se traduire rapidement en revenus (les modèles économiques étant encore incertains), les entreprises en tête sur ce secteur[17] disposent de capacités financières et d'endettement leur permettant d'assumer ces investissements, sans nécessairement se préoccuper de la rentabilité de court terme. À date, les modèles économiques de ces acteurs restent encore diversifiés et en cours de structuration (cf. 1.2.2) et témoignent d'une volonté de privilégier une adoption massive de leurs services, en particulier pour certains des services les plus connus tels des *chatbots*, qui proposent généralement une version gratuite.

Outre ces annonces d'investissement par les acteurs privés eux-mêmes, les acteurs publics soutiennent cette dynamique et s'engagent dans une course au développement de l'IA visant à bénéficier du potentiel économique de l'IA en termes de gains de productivité et d'innovation, tout en veillant à préserver leur souveraineté sur les technologies et infrastructures clés. Le sommet mondial pour l'action sur l'IA à Paris, en février 2025, a par exemple été l'occasion d'annonces d'importants investissements, à hauteur de 109 milliards d'euros dans les prochaines années en France. La présidente de la Commission européenne Ursula von der Leyen a également annoncé lors de ce

[14] Eurostat, dernière mise à jour le 27 février 2026

[15] Selon Cyfuture cloud (Cyfuture cloud, site consulté en avril 2026), en 2025, un serveur dédié à l'IA, contenant 8 processeurs « H100 » peut coûter entre 200 000 et 400 000 dollars, sachant qu'il faut plusieurs dizaines voire centaines de serveurs de ce type, ainsi que des capacités de stockage et de réseau, pour avoir un « cluster IA » permettant alors d'entraîner un modèle d'IA générative (selon sa taille).

[16] https://www.lemonde.fr/economie/article/2026/02/07/intelligence-artificielle-la-folle-course-des-geants-de-la-tech_6665758_3234.html

[17] Outre Amazon, Google, Meta et Microsoft, les développeurs de modèles tels qu'OpenAI, Mistral AI ou Anthropic entretiennent des partenariats stratégiques avec ces grands groupes qui leur permettent de bénéficier de capacités de financement.

sommet souhaiter mobiliser 200 milliards d'euros, dont 150 milliards venant d'un consortium d'entités privées regroupées sous la bannière « AI EU Champions Initiative », pour faire de l'Europe un des principaux continents de l'IA, considérant que « *la course à l'IA est loin d'être terminée* »[18]. Ce plan d'investissement constitue le socle du « *AI Continent Action Plan*[19] » publié en avril 2025 et détaillant différentes mesures, allant de la mise en place d'une infrastructure informatique de grande ampleur dédiée à l'IA à la promotion de l'IA dans des secteurs stratégiques, en passant par le renforcement des compétences. Ces annonces européennes sont arrivées peu après celle de Stargate, un programme de 500 milliards de dollars impulsé par l'administration Trump, en lien avec des acteurs privés, pour dynamiser l'investissement privé dans le développement de l'IA aux États-Unis.

Face à cet engouement, certaines institutions évoquent aujourd'hui quelques réserves et l'intérêt d'ajuster la stratégie de soutien à l'IA. Ainsi, en France, le comité de surveillance des investissements d'avenir (CSIA), observe une accélération des investissements pour l'IA dans le monde, ce qui pose des questions « *sur une possible bulle, les modèles économiques des investissements dans l'IA n'étant pas encore établis* »[20].

1.2.2 Une dynamique d'innovation aux équilibres encore incertains

À l'heure actuelle, le marché des services d'IA est en pleine expansion : de 189 milliards de dollars de chiffre d'affaires en 2023, il est estimé à 4 800 milliards de dollars en 2033 selon l'ONU Commerce et Développement. L'IA pourrait ainsi, sur cette période, quadrupler sa part dans le marché mondial des technologies de pointe, en passant de 7 % à 29 %[21]. Face à ces perspectives, la course actuelle à l'innovation et à l'adoption des services incite très fortement les acteurs à se placer en tant que *leaders* tout au long de la chaîne de valeur et à consentir des investissements conséquents dans des intrants essentiels comme l'accès à l'énergie, la puissance de calcul, les données et l'expertise technique.

La chaîne de valeur de l'IA générative peut être découpée en trois couches : l'infrastructure, la modélisation et le déploiement. Elle se compose actuellement d'un nombre important d'acteurs, qui se répartissent différemment selon les services proposés. En fonction des couches, les intrants et les investissements nécessaires diffèrent et conduisent à des concentrations de marché variables. Certains segments de marché, notamment au niveau de l'infrastructure, tels que la fourniture de puces spécialisées (avec un quasi-monopole de Nvidia) ou de services *cloud* (avec Amazon, Microsoft et Google représentant l'essentiel du marché), apparaissent davantage concentrés que d'autres, notamment pour la fourniture de services intégrant de l'IA.

Dans son avis sur le fonctionnement concurrentiel du secteur de l'IA générative[22], l'Autorité de la concurrence met en évidence que la chaîne de valeur est occupée par de grands acteurs du numérique verticalement intégrés et capables ainsi de contrôler la conception, l'entraînement et l'exploitation des modèles, accompagnés en aval, d'une myriade de fournisseurs de services, entreprises intégrant l'IA à leurs offres et produits. Au sommet de cette pyramide se trouvent les géants du numérique : Alphabet (Google), Microsoft, Amazon, Meta ou Nvidia. Certains, comme Microsoft ou Alphabet, sont intégrés verticalement sur l'ensemble de la chaîne de valeur ; d'autres se concentrent sur des segments clés (GPU, *cloud*, collecte de données). En outre, les développeurs de modèles tels qu'OpenAI, Mistral AI ou Anthropic entretiennent des partenariats stratégiques avec ces grands groupes, renforçant ainsi la concentration sectorielle. Enfin, des acteurs positionnés sur des marchés

[18] Intelligence artificielle : Ursula von der Leyen annonce 200 milliards d'euros d'investissements en Europe

[19] The AI Continent Action Plan | Shaping Europe's digital future

[20] rapport-2025-du-csia-france-2030.pdf

[21] ONU Commerce et développement (CNUCED), 2025. L'IA pourrait atteindre 4 800 milliards de dollars d'ici 2033 et s'imposer comme la principale technologie d'avant-garde.

[22] Avis 24-A-05, 28 juin 2024, Autorité de la concurrence

connexes comme la vente de terminaux (ex : Samsung, Apple), intègrent de l'IA à leurs produits ou services déjà existants afin de maintenir leur attractivité.

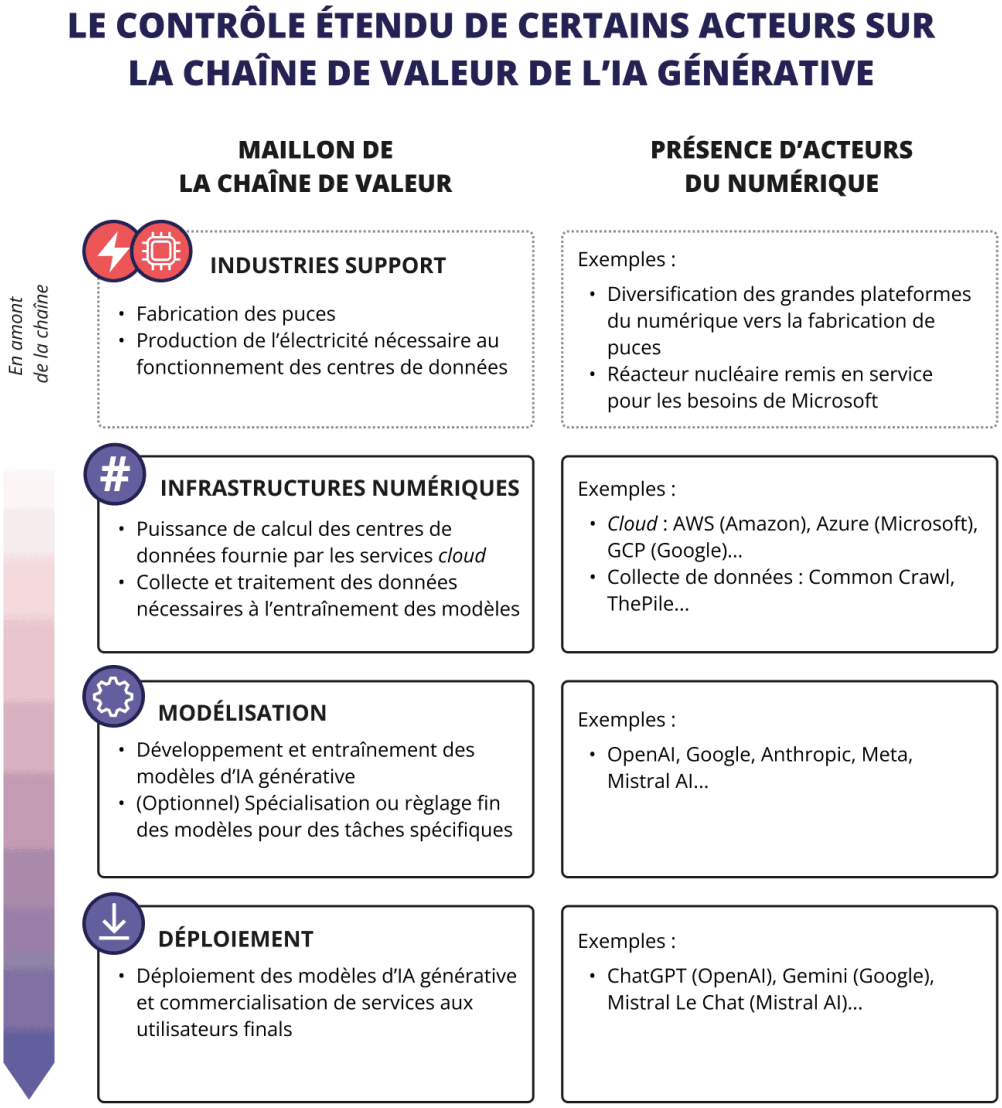


Figure 3 : Chaîne de valeur de l'IA générative. Source : Arcep.

Si certains acteurs sont ainsi présents et bien établis sur l'ensemble de la chaîne de valeur, les dynamiques de marché sont encore soumises à différentes incertitudes. Par exemple, la modélisation et les services d'IA pourraient se concentrer autour d'un faible nombre de modèles généralistes ou tendre vers une plus grande diversité de modèles spécialisés. Dans son avis relatif au fonctionnement concurrentiel du secteur de l'IA générative[23], l'Autorité de la concurrence estime que la course à l'innovation et au développement de nouveaux modèles d'IA générative devrait se poursuivre sur deux axes : la taille des modèles et l'optimisation des modèles à taille constante. Empiriquement, le nombre de modèles et leur taille n'ont cessé de croître au cours des dernières années. À titre d'illustration, dans le cas de ChatGPT, si le premier modèle publié en 2018 contenait 120 millions de paramètres,

[23] Avis 24-A-05, 28 juin 2024, Autorité de la concurrence

GPT-3 sorti en 2020 comprenait 175 milliards de paramètres et, selon certaines estimations[24], le modèle GPT-5, déployé en mars 2026, comprendrait de l'ordre de 1 700 milliards de paramètres.

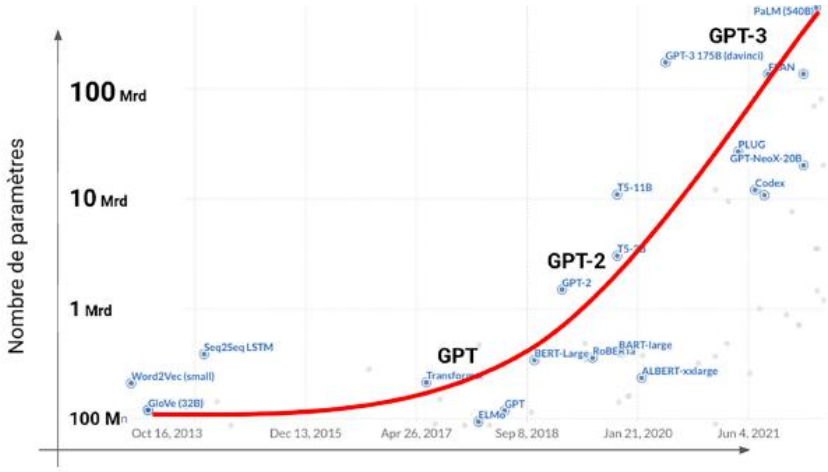


Figure 4 : Évolution de la taille des modèles dans le domaine du traitement du langage naturel[25]

Dans le même temps, la disponibilité de nombreux jeux de données et modèles ouverts permet à un plus grand nombre d'acteurs de déployer des services d'IA générative, notamment pour ceux (entreprises, administrations ou chercheurs par exemple) qui n'ont pas les moyens de développer et entraîner leurs propres modèles, et pour lesquels il est plus aisé de procéder à la spécialisation de modèles existants. Toutefois, si l'*open source* favorise l'ouverture du marché à de nouveaux entrants, l'intégration verticale de certains acteurs tout au long de la chaîne de valeur (de l'infrastructure au déploiement) renforce la position de ces derniers, en raison de la forte dépendance des développeurs de services d'IA aux ressources de calcul et aux données de haute qualité auxquelles ont accès les grands acteurs intégrés.

Par ailleurs, les acteurs qui développent des modèles et services d'IA générative sont en cours de structuration de leur modèle économique. Quatre modèles commerciaux se distinguent à date : (i) un accès au service d'IA soumis à un paiement *via* un abonnement, une licence ou par appel d'une interface applicative (API) ; (ii) l'intégration des outils d'IA dans d'autres produits ou services[26] ; (iii) la vente et/ou le conseil aux entreprises ; (iv) la publicité. Ainsi, bien que les acteurs ne semblent pas prioriser la rentabilité économique de court terme, la volonté d'assurer une adoption et une diffusion massives de leurs services d'IA générative disponibles gratuitement s'inscrit dans une logique de monétisation et de retour sur investissements à terme.

[24] GPT-5: Just How Big Is This Model, Really? - A2E

[25] Source : Data For Good à partir de l'article de Sevilla *et al.* (2022)

[26] Par exemple : l'intégration d'IA dans les messageries instantanées, les moteurs de recherche ou encore les systèmes d'exploitation des terminaux.

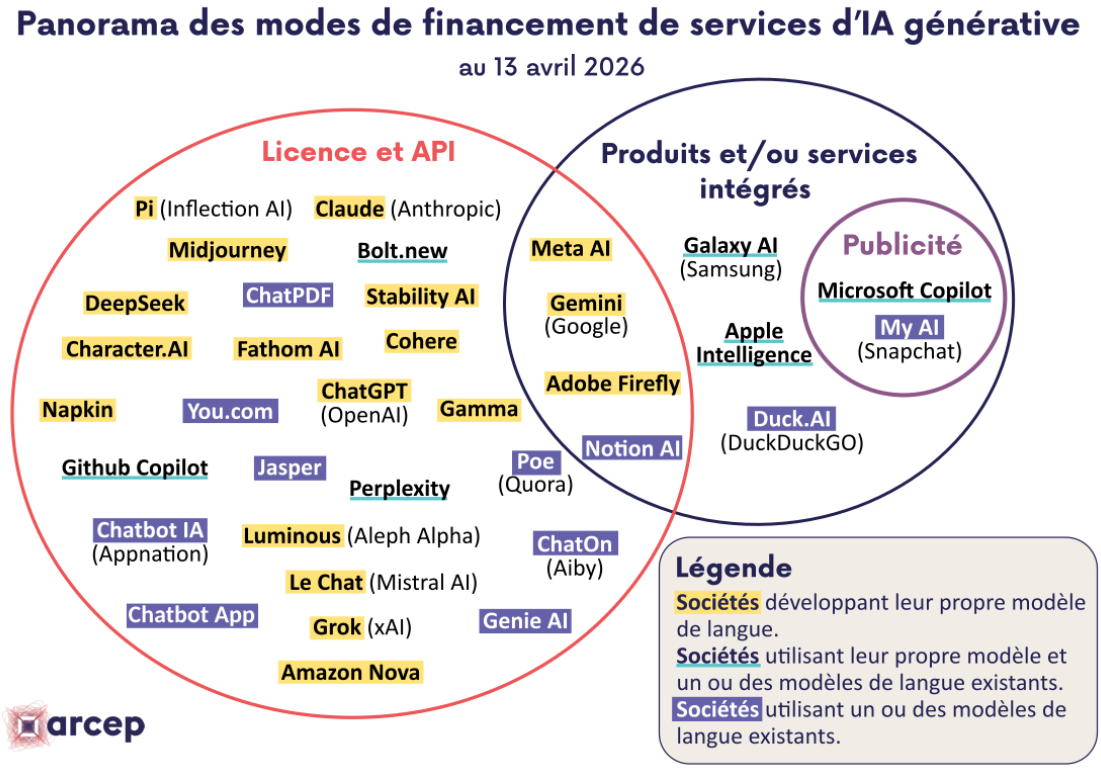


Figure 5 : Panorama des modes de financement de services d'IA générative. Source : Arcep

Ainsi, les dynamiques de marché découlant de la course à l'innovation présentent encore de nombreuses incertitudes sur la structuration du marché à terme. Ces dynamiques se traduisent actuellement et pour les prochaines années par une intensité concurrentielle stimulant fortement les investissements et la consommation de ressources. L'effet de ces dynamiques ne se limite ainsi pas au champ économique mais entraîne également des effets sur l'environnement.

1.2.3 Des risques de tension sur les ressources qui présentent des enjeux environnementaux et économiques

Comme rappelé en section 1.1, loin d'être immatérielle, l'intelligence artificielle repose sur des infrastructures physiques et la consommation de ressources naturelles. Dès lors, les dynamiques concurrentielles entourant le développement de l'IA générative, marquées par une course à l'innovation, un besoin croissant en capacités de calcul et une hausse soutenue des investissements, ne sont pas sans conséquence sur les impacts environnementaux de l'IA générative et peuvent engendrer des tensions croissantes sur certaines ressources stratégiques.

Les dynamiques de marché présentées précédemment impliquent un recours accru aux centres de données dont les capacités doivent continuellement s'adapter à des charges de calcul en forte croissance, ainsi qu'une demande soutenue en serveurs spécialisés, notamment pour l'entraînement ou l'inférence des modèles d'IA générative de grande taille. Elles peuvent également se traduire par un renforcement des équipements de réseau, selon l'architecture retenue (calcul centralisé ou en périphérie) ou un renouvellement anticipé des terminaux utilisateurs pour des équipements compatibles avec l'IA.

Ces évolutions induites par les dynamiques concurrentielles du secteur de l'IA générative reposent sur des chaînes d'approvisionnement particulièrement intensives en ressources. En effet, la fabrication des semi-conducteurs omniprésents dans l'ensemble des trois briques de l'infrastructure numérique (terminaux, réseaux, centres de données) mobilise des quantités importantes d'énergie, d'eau

hautement purifiée et de produits chimiques, générant des impacts environnementaux significatifs en amont de l'usage, en particulier pour les semi-conducteurs les plus performants (Roussilhe, 2021). De même, la demande croissante en équipements numériques entraîne une tension accrue sur les matières premières critiques. L'extraction et le raffinage de ces ressources sont souvent associés à des dégradations environnementales importantes, notamment en termes de pollution des sols, de consommation d'eau et de perte de biodiversité. Dans ce contexte, la dynamique d'investissements massifs et de montée en puissance des capacités de calcul tend à renforcer les tensions sur ces ressources stratégiques, tout en accentuant les impacts environnementaux à l'échelle locale.

Par ailleurs, la concentration géographique des infrastructures de centres de données, motivée par des logiques d'économies d'échelle et d'accès à des ressources énergétiques et de connectivité abondantes, est susceptible d'exacerber la concurrence pour l'accès aux ressources locales (notamment l'électricité, l'eau et le foncier), avec à la clé des tensions accrues sur les infrastructures comme le réseau électrique (e.g. réservation de capacité), voire des conflits d'usage avec d'autres secteurs. Ces évolutions s'inscrivent ainsi dans un contexte plus large d'arbitrages énergétiques, où la croissance de la demande liée à l'IA générative est susceptible d'entrer en concurrence avec d'autres usages énergétiques, posant la question de sa compatibilité avec les trajectoires de transition énergétique et avec l'ambition politique d'électrification de l'économie.

Dans ce contexte, les acteurs les plus puissants du secteur de l'IA générative peuvent chercher à préempter les ressources stratégiques, qu'il s'agisse de l'accès à l'électricité nécessaire au fonctionnement des centres de données, ou des semi-conducteurs de dernière génération indispensables au développement des modèles d'IA générative, accentuant ainsi les tensions environnementales. Dans ce contexte, la frugalité des modèles d'IA générative pourrait devenir, d'une part, un facteur différenciant majeur dans la compétition entre acteurs et, d'autre part, un atout pour favoriser un développement soutenable de l'IA générative. Cette incitation à la frugalité pourrait voir ses effets réduits en cas de concentration accrue du marché des services d'IA entre les mains de quelques grands acteurs qui capteraient l'accès aux intrants (calcul, données, talents, accès à l'énergie).

Enfin, au-delà des infrastructures et des ressources, le développement de l'IA engendre des effets systémiques susceptibles d'amplifier cette tension sur les ressources. L'un des principaux mécanismes à l'œuvre est celui de l'effet rebond (cf. encadré en section 1.3) : les gains d'efficacité et la baisse des coûts d'accès à ces technologies, dont certaines sont aujourd'hui proposées gratuitement, favorisent une diffusion rapide des usages, qui conduit à une augmentation globale de la demande en calcul et en données. À titre d'exemple, du fait d'une importante popularité pour générer des images s'inspirant du style graphique du studio Ghibli, OpenAI a été contraint de limiter l'accès à son outil de génération d'image basé sur l'IA générative car la libre utilisation entraînait une consommation électrique trop importante des processeurs graphiques de l'entreprise. Cet exemple illustre que les ressources mobilisées pour fournir des services d'IA générative, qu'il s'agisse de puissance de calcul, d'électricité ou d'infrastructures, ne sauraient être considérées comme illimitées. L'intégration croissante de l'IA dans de multiples services du quotidien — des moteurs de recherche aux outils de productivité — participe ainsi à une expansion continue des besoins en ressources.

Ainsi, l'analyse environnementale des effets des mécanismes économiques décrits précédemment met en évidence un ensemble d'impacts directs et indirects qui s'articulent étroitement avec les dynamiques économiques de l'intelligence artificielle. Loin d'être marginales, ces implications, s'illustrant notamment par une tension sur des ressources en quantité limitée, invitent à interroger les conditions de soutenabilité du modèle actuel de développement de l'IA, ainsi que les leviers susceptibles d'en atténuer les impacts environnementaux. À ce titre, une évaluation rigoureuse et transparente de ces impacts apparaît comme un préalable indispensable pour appréhender de manière éclairée la soutenabilité d'un tel développement.

1.3 Des effets environnementaux nets attendus qui restent à évaluer

Si le numérique génère des effets directs, qui correspondent à l'empreinte environnementale liée aux différentes phases de cycle de vie du numérique (cf. 1.1.1), il peut également avoir des effets indirects. À l'inverse des effets directs, les effets indirects peuvent avoir un impact environnemental positif ou négatif. Il s'agit concrètement des effets que va produire un usage numérique notamment sur les comportements humains ou les processus de production (cf. encadré « Comprendre les effets environnementaux directs et indirects du numérique » pour plus de détails).

Un exemple d'effet indirect positif du numérique est l'efficacité, c'est-à-dire l'optimisation d'un processus en le numérisant. À titre d'illustration, un cas d'usage connu du grand public est la domotique, avec l'ajout d'équipements connectés à la maison. Cela permet d'optimiser le fonctionnement des équipements de chauffage et réduire la dépense énergétique du foyer et les émissions de gaz à effet de serre (GES) associées.

Plusieurs cas d'usage de solutions numériques, en particulier dans des domaines industriels et techniques, peuvent engendrer des gains environnementaux, d'après une récente étude de l'ADEME (2024a ; 2025b). L'étude montre par exemple que le *Dynamic Line Rating (DLR),* pour optimiser la gestion du système électrique, entraîne quasiment systématiquement une réduction des émissions de GES dans l'ensemble des configurations testées [27].

Cependant, dans certains cas, les économies d'énergie ou de ressources permises par l'utilisation de nouvelles technologies numériques peuvent avoir aussi pour effet une augmentation de la consommation. En effet, cet effet « rebond » décrit une situation où une amélioration de l'efficacité permettant de réduire un coût (économies d'énergie ou de ressources) pour un usage donné, entraîne de ce fait une augmentation de l'usage qui peut neutraliser les gains attendus, voire générer un impact négatif si l'accroissement des usages fait plus que compenser les économies permises[28]. Pour reprendre l'exemple de la domotique précédemment cité, l'optimisation du chauffage permet de réduire la facture énergétique (financière et environnementale) mais les gains économiques engendrés pour le foyer peuvent inciter à chauffer à un niveau plus élevé (température de consigne plus élevée), ce qui peut contrebalancer les gains initiaux.

Ainsi, l'ADEME rappelle que, pour certains cas, les impacts nets de la solution numérique considérée vont dépendre de la configuration et du contexte de mise en œuvre de la solution. Par exemple, pour le cas de l'éclairage public connecté, les impacts nets sont négatifs lorsque le périmètre de comptabilité inclut les risques environnementaux liés à des effets de rebond et d'induction (cf. encadré « Comprendre les effets environnementaux directs et indirects du numérique »). Dès lors, l'ADEME souligne la difficulté d'extrapoler les résultats d'un cas d'usage à une échelle nationale et donc l'impossibilité de conclure sur les effets de la numérisation à l'échelle globale.

À ce jour, aucun consensus scientifique n'existe sur l'impact environnemental net du numérique c'est-à-dire sur son empreinte environnementale nette de ses effets indirects (positifs ou négatifs). *A*

[27] En effet, en France, les distances à respecter entre un câble électrique et les différents obstacles qui pourraient se trouver en dessous sont réglementées pour des raisons de sécurité. Les conditions météorologiques et le courant électrique qui le parcourt modifient la température d'un câble, ce qui peut le dilater ou le contracter, modifiant alors sa distance au sol. Les systèmes de *DLR* permettent d'optimiser en temps réel la capacité de transit maximale du courant dans des lignes électriques (par opposition au *Static Line Rating* qui fixe une valeur en prenant en compte des conditions météorologiques fixes et conservatrices). Concrètement, le DLR permet ainsi, par exemple lorsqu'il y a beaucoup de vent qui refroidit le câble, de ne pas inutilement limiter la production de parcs éoliens (« l'écrêtement ») et de favoriser l'intégration des énergies renouvelables (donc décarbonées) dans le réseau.

[28] On parle de « paradoxe de Jevons » lorsque les gains d'efficacité sont quantitativement plus que compensés par l'augmentation des usages (c'est ce qu'on appelle un effet « rebond »), au point que la consommation totale de ressources dépasse le niveau qu'elle avait avant l'apparition des gains d'efficacité permis par une innovation technologique.

fortiori, s'agissant de l'IA générative, qui connait des évolutions technologiques très rapides et a récemment émergé dans les usages, aucune étude à date ne permet de conclure sur son impact environnemental net, en particulier car de telles évaluations doivent être faites en fonction d'un cas d'usage et d'un contexte précis de déploiement d'une solution IA.

Évaluer les effets environnementaux nets du numérique est donc une tâche complexe. La diffusion massive des technologies numériques limite en pratique la réalisation de telles évaluations, qui exigent un volume important de données, à la fois sur le numérique, mais aussi sur de nombreux secteurs dans lesquels des solutions numériques sont déployées. Cette tâche est d'autant plus ardue dans un contexte où la transparence reste insuffisante. Une évaluation au cas par cas apparaît ainsi comme plus pertinente car elle seule permet d'intégrer pleinement les effets indirects d'une solution numérique, permettant alors d'éviter des extrapolations à plusieurs cas d'usages qui seraient peu robustes. Cette analyse doit prendre en compte les effets indirects et systémiques de la solution numérique[29]. Pour ce faire, il est préférable de réaliser une analyse de cycle de vie conséquentielle (ACV-C), qui vise précisément à comprendre les conséquences globales d'un changement (apporté par une solution numérique) plutôt qu'une ACV dite attributionnelle (ACV-A), qui fournit une simple image statique des impacts.

[29] G.Roussilhe explique ainsi que le problème de certaines études sur les émissions évitées du numérique est l'extrapolation à partir d'études de cas, la représentativité des études de cas sélectionnées et l'absence d'intégration des impacts du cycle de vie et des impacts structurels (Gauthier Roussilhe, Les émissions évitées de la numérisation, septembre 2023).

ACV attributionnelle vs conséquentielle : quels objectifs et quelles différences ?

ACV attributionnelle (ACV-A)

L'analyse de cycle de vie attributionnelle (ACV-A) est une méthode qui vise à évaluer les impacts environnementaux d'un système tel qu'il existe à un instant donné. Elle repose sur une approche descriptive et statique : le but est d'estimer l'empreinte environnementale associée à un produit ou service - par exemple numérique - sur son cycle de vie (fabrication, distribution, utilisation, fin de vie). Les impacts sont généralement répartis selon des règles d'allocation.

Exemples : Estimer l'empreinte environnementale d'un smartphone en intégrant les impacts liés aux différentes étapes de son cycle de vie ou estimer l'empreinte environnementale d'une heure de streaming en allouant les impacts de l'infrastructure numérique nécessaire (le terminal sur lequel est regardé le contenu, les réseaux qui transmettent les données, les centres de données qui hébergent le contenu) *via* des règles d'allocation.

Cette approche est notamment utilisée pour comparer des produits ou établir des bilans d'empreinte, mais une des limites principales est qu'elle ne prend pas en compte les effets de changements ou les conséquences indirectes d'une modification des usages.

ACV conséquentielle (ACV-C)

L'analyse de cycle de vie conséquentielle (ACV-C) vise à évaluer les impacts environnementaux induits par une décision ou un changement, plutôt que de décrire un système existant. Elle adopte une approche dynamique et comparative, en s'intéressant aux conséquences d'un scénario par rapport à un autre (par exemple le déploiement d'une solution numérique par rapport à une situation sans cette solution). Cette méthode permet d'intégrer les effets du système étudié sur son environnement extérieur, au-delà de ses effets directs.

Exemple : Que se passe-t-il si une partie des achats en magasin est transférée vers l'e-commerce ? Si cela permet *a priori* moins de trajets individuels (en mutualisant les trajets avec les livraisons), cela peut par exemple engendrer une hausse de la consommation par des achats facilités. Il faut donc raisonner avec une logique conséquentielle et ne pas se limiter à une photographie du système.

L'ACV-C est donc particulièrement adaptée à l'aide à la décision et à l'évaluation des politiques publiques ou des innovations, mais ses résultats peuvent être plus incertains car fortement dépendants des hypothèses retenues, notamment sur les scénarios choisis.

Comprendre les effets environnementaux directs et indirects du numérique

Les effets environnementaux du numérique, positifs ou négatifs, peuvent être directs ou indirects.

Les effets environnementaux directs d'une solution numérique correspondent aux impacts environnementaux induits par le cycle de vie des équipements et infrastructures numériques qui assurent le fonctionnement de cette solution. Il s'agit donc des impacts des terminaux utilisateurs, réseaux et centres de données mobilisés pour fournir le service, le long du cycle de vie de ces trois briques (extraction des matières premières, fabrication, distribution, usage et fin de vie).

Les effets environnementaux indirects d'une solution numérique correspondent aux externalités environnementales que génère le déploiement de technologies numériques. Elles recouvrent des modifications notamment techniques (e.g. utiliser l'IA pour faciliter la détection d'incendies) ou sociétales (e.g. utiliser l'IA pour de la publicité ciblée modifie les comportements de consommation), et peuvent être positives ou négatives.

Les effets indirects peuvent en effet entrainer une réduction de l'impact environnemental par rapport à une situation de référence où la solution numérique en question n'est pas déployée. Parmi les effets indirects positifs, on peut par exemple citer :

- **l'effet de substitution** : un effet de substitution ayant un impact environnemental positif se produit quand une technologie est remplacée par une technologie plus efficace d'un point de vue environnemental ;
- **l'effet d'optimisation** : les effets d'optimisation ou d'efficacité correspondent à des modifications d'une solution lui permettant d'utiliser moins de ressources (ex : matières, énergie...) à quantité produite ou consommée constante.

Les effets indirects peuvent également se traduire par une augmentation de l'impact environnemental par rapport à la situation de référence. Parmi les effets indirects négatifs, on peut par exemple citer :

- **l'effet d'induction** : les effets d'induction correspondent aux situations où la solution numérique induit de nouveaux usages ou l'utilisation de nouvelles ressources (e.g. : l'innovation permise par l'imprimante induit l'usage de papier) ;
- **l'effet rebond** : les effets de rebond correspondent aux situations où le gain d'efficacité permis par une solution numérique (ex : chauffage connecté) amène à une augmentation de la consommation de l'activité ayant bénéficié de la solution numérique ou bien d'une autre activité (ex : réallouer le budget économisé sur le chauffage dans une autre activité).

Les impacts nets, c'est-à-dire en considérant les effets directs et indirects, sont donc relatifs à une situation de référence, ce qui implique de comparer des scenarii où la solution numérique est déployée avec ce qui se serait passé si un tel déploiement n'avait pas eu lieu. **Comme l'indique l'ADEME dans une récente étude (ADEME, 2024a) sur l'évaluation environnementale des effets directs et indirects du numérique pour des cas d'usages, la définition du contexte de mise en œuvre de chaque solution est donc cruciale pour réaliser une analyse de cycle de vie conséquentielle (ACV-C). Cette contrainte rend donc toute extrapolation générale des résultats sur l'impact environnemental net du numérique très délicate et incomplète.**

Dès lors, si l'évaluation des effets nets est déjà complexe pour des innovations numériques moins récentes, elle le devient encore plus lorsqu'il s'agit de l'IA générative du fait de ses évolutions très rapides et d'un manque de données nécessaires à de telles évaluations. Il est ainsi délicat, tout comme

pour le numérique, de trancher sur les effets environnementaux nets de l'IA générative, qu'ils soient positifs ou négatifs, en adoptant une conclusion générale.

En effet, le progrès technique des centres de données[30], infrastructures essentielles au développement de l'IA générative, est aujourd'hui observable avec des données factuelles (cf. encadré « Zoom sur les travaux sur les centres de données dans l'Enquête annuelle « Pour un numérique soutenable » en section 2.1.1), et permet de réduire l'impact environnemental de l'IA générative. Cependant, les effets indirects de l'IA générative peuvent être de nature très variable en fonction des secteurs dans lesquels elle est déployée et même au sein d'un unique secteur. Par exemple, dans l'énergie, les effets nets peuvent être négatifs lorsque l'IA générative facilite l'extraction d'énergies fossiles, mais ils sont susceptibles d'être positifs lorsqu'elle est déployée pour soutenir le développement des énergies renouvelables et leur intégration au système électrique. De fait, dans certains secteurs, l'IA peut contribuer à réduire les émissions de GES (ex : électricité, filière viande et lait, véhicules légers) mais Stern et *al*. (2025) avancent que laisser le marché déterminer les applicatifs de l'IA et sa gouvernance peut se révéler risqué. Selon eux, avoir des États interventionnistes est un facteur central de la bonne utilisation de l'IA dans ces secteurs car les seules forces du marché pourraient ne pas suffire à engendrer les mutations requises pour exploiter pleinement le potentiel de l'IA, tout en atteignant les objectifs climatiques.

En définitive, analyser les effets environnementaux nets du numérique (et plus encore de l'IA générative) se heurte aujourd'hui à des limites méthodologiques importantes. La diversité des usages, la forte dépendance aux contextes sectoriels et les effets indirects, tels que les rebonds, rendent toute évaluation globale particulièrement incertaine et, à ce stade, largement exploratoire.

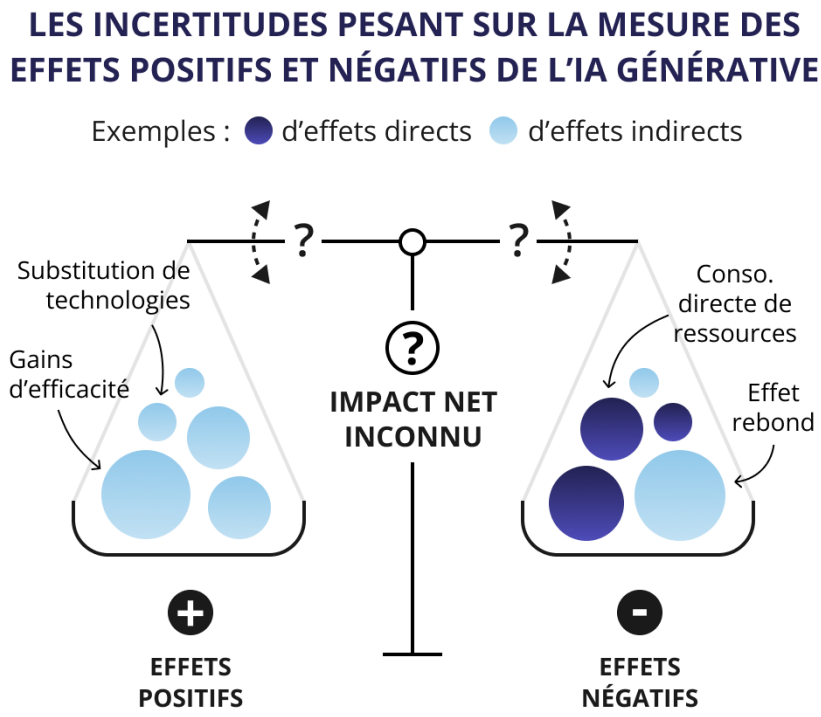


Figure 6 : Illustration des effets directs et indirects de l'IA générative. Source : Arcep.

[30] Par exemple, au cours des 15 dernières années, l'urbanisation des salles serveurs dans les centres de données, c'est-à-dire l'organisation physique et logique des infrastructures IT, notamment avec des allées froides et des allées chaudes, a permis de les rendre plus faciles à refroidir et donc d'améliorer l'efficacité énergétique globale des centres de données.

Cette difficulté ne saurait toutefois justifier l'inaction. En effet, dans le même temps, l'ampleur des dynamiques en cours, notamment la croissance rapide des besoins énergétiques associés à l'IA générative (AIE, 2025), confère à ces enjeux un caractère à la fois immédiat et structurant.

Sans préjuger des effets environnementaux nets de l'IA générative à long terme, il apparaît d'ores et déjà essentiel de renforcer les connaissances sur ses impacts environnementaux directs, qui demeurent insuffisamment documentés, afin d'éclairer au mieux le débat public.

⁂

L'IA générative met ainsi en évidence, de manière particulièrement marquée, la matérialité du numérique : son développement repose sur des infrastructures physiques, des capacités de calcul élevées et des investissements massifs dans les centres de données et les composants informatiques. La rapidité de son adoption et l'intensité des dynamiques concurrentielles qui accompagnent le développement de l'IA accélèrent ses besoins, au risque de renforcer les tensions sur certaines ressources et d'amplifier les enjeux environnementaux associés.

Dans ce contexte, si l'évaluation des effets environnementaux nets de l'IA demeure très délicate et complexe à traiter, il apparaît dès l'amont essentiel de s'interroger sur les effets environnementaux directs de l'IA qui sont abordés dans le prochain chapitre.

Chapitre 2 L'évaluation de l'empreinte environnementale de l'IA

Ce chapitre propose une synthèse des connaissances sur l'empreinte environnementale de l'IA générative. Elle permet de fournir de premières évaluations, et de partager un premier socle de connaissances fiables, sur la base duquel des leviers d'actions pourront être discutés. Ce travail apparaît d'autant plus important que les données disponibles restent à ce jour très limitées, notamment du fait d'un manque de transparence des acteurs de la chaîne de valeur de l'IA générative, en particulier des fournisseurs de modèles (cf. encadré *infra*). Pour contribuer à l'évaluation de l'empreinte environnementale de l'IA, l'Arcep s'est donc appuyée autant que possible sur les résultats issus de la littérature scientifique, complétés par des travaux institutionnels ou de l'industrie. Constatant un manque d'informations fiables en particulier sur la phase d'inférence, l'Arcep a également mené des travaux avec le Pôle d'expertise de la régulation numérique (PEReN), qui apportent des résultats inédits sur l'impact environnemental de la phase d'inférence.

Un manque de transparence environnementale des modèles d'IA

L'évaluation de l'impact environnemental de l'IA nécessite de s'appuyer sur un ensemble d'informations fiables, accessibles principalement auprès des fournisseurs eux-mêmes, qui sont les seuls à disposer d'une vision complète sur leur production.

Or, actuellement, il n'existe que peu de chiffres précis et fiables sur les différents indicateurs de l'empreinte environnementale des modèles et de leur utilisation sur l'ensemble du cycle de vie. Les principaux acteurs du numérique notamment ne partagent pas d'informations sur l'impact environnemental de leurs activités liées au développement de leurs modèles, leur entraînement et leur usage. Luccioni *et al*. (2025) ont ainsi montré que, sur 754 modèles analysés, déployés entre 2010 et le premier trimestre 2025, 84 % ne font l'objet d'aucune information environnementale, 14 % des modèles font indirectement l'objet d'information environnementale (*e.g.* données sur l'entraînement ou la taille des modèles qui constituent des informations techniques et non environnementales, mais qui sont utiles pour une évaluation environnementale) et seulement 2 % des modèles font l'objet d'information de manière directe (*e.g.* mesures de consommation énergétique ou intensité carbone) à date de mai 2025. Les informations sont donc très parcellaires.

Faute de transparence des acteurs, les impacts environnementaux de l'IA sont donc aujourd'hui encore peu évalués sur l'ensemble de la chaîne de valeur (Ligozat *et al*., 2022), ce qui limite la capacité à anticiper et maîtriser ces impacts. Les acteurs pourraient d'ores et déjà fournir davantage d'informations environnementales sur l'ensemble du cycle de vie des modèles, en particulier sur la phase d'entraînement et la phase d'inférence mais le manque actuel de transparence environnementale tend à montrer que les incitations « naturelles » à la transparence sont insuffisantes. Plusieurs éléments peuvent contribuer à cette situation, notamment une situation de déséquilibre concurrentiel, l'absence de méthodologie partagée ou une mauvaise prise en compte des enjeux du côté de la demande.

Malgré le manque de transparence des modèles d'IA générative, des travaux académiques ont réalisé des évaluations quantitatives de l'impact environnemental de cette technologie. Ces dernières s'appuient généralement sur les déclarations environnementales volontaires mais partielles des fournisseurs de services d'IA générative ou sur des modèles ouverts moins connus. Des hypothèses moyennes (ex : intensité carbone du mix électrique moyen mondial ou puissance moyenne des puces) sont considérées en l'absence de données industrielles dans les travaux étudiés.

Les éléments d'appréciation de l'empreinte environnementale de l'IA générative recueillis sont d'abord présentés par brique du numérique (centres de données, terminaux, réseaux) (2.1) puis en suivant une analyse par cycle de vie d'un service d'IA, en particulier les phases d'entraînement et d'inférence du modèle (2.2)[31]. L'analyse proposée étudie, dans la mesure des données disponibles, différents critères environnementaux (notamment énergie, carbone, ressources abiotiques ou eau).

Si l'analyse environnementale qui suit se centre autant que possible sur les services d'IA générative, les méthodologies utilisées s'appliquent généralement à tout système d'IA[32] puisque l'IA générative n'est qu'une sous-catégorie de l'IA comme l'illustre la figure ci-dessous, avec toutefois des consommations de données et d'énergie qui sont relativement moindres pour ces autres systèmes selon l'Union internationale des télécommunications (UIT)[33].

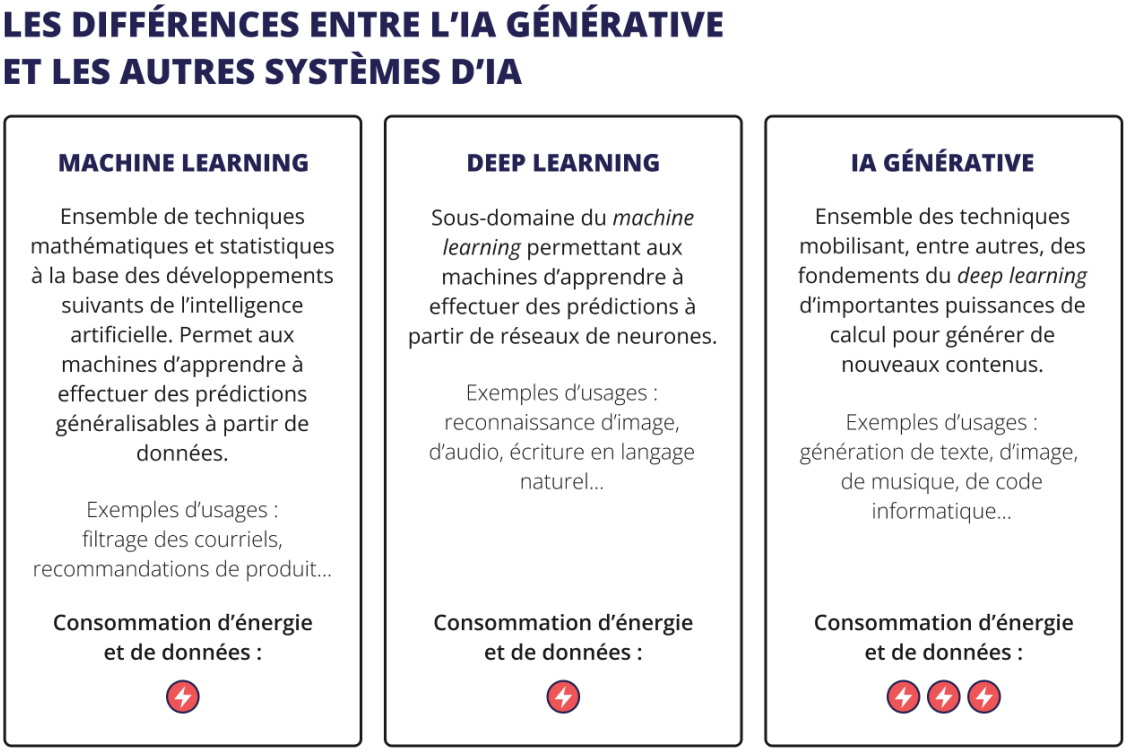


Figure 7 - Types de systèmes d'IA. Source : Arcep.

[31] Il convient de distinguer la phase d'entrainement et la phase d'inférence du modèle, parmi d'autres phases (collecte, réentraînement, *fine-tuning*…) nécessaires au développement de services d'IA générative tels que ChatGPT. Ces deux phases sont étudiées en détail en raison de leur impact environnemental significatif par rapport aux autres phases. Les étapes du cycle de vie d'un système d'IA sont détaillées dans la recommandation UIT-T L.1801 en page 5. Des paramètres-clés spécifiques doivent être étudiés en fonction de ces différents systèmes d'IA (IA générative, *deep-learning, machine-learning* ou systèmes experts).

[32] Selon la norme ISO/IEC 22989, un système d'IA est un système conçu pour générer des résultats tels que du contenu, des prévisions, des recommandations ou des décisions pour un ensemble donné d'objectifs définis par l'humain.

[33] Voir cette synthèse pédagogique de la recommandation UIT-T L.1801.

2.1 Analyse par brique du numérique : les centres de données comme composante structurante des impacts environnementaux de l'IA générative, des effets encore incertains sur les réseaux et les terminaux

Comme précisé en partie 1.1, l'évaluation de l'empreinte environnementale d'un service numérique, tel que l'IA générative, passe par une analyse de chacune des briques du numérique qui sous-tendent ces services : les centres de données, les réseaux et les terminaux. Cette partie revient ainsi sur les études existantes permettant d'apporter des éclairages sur l'impact de l'IA, tout d'abord sur la brique des centres de données (2.1.1), qui concentre à ce jour les principaux enjeux environnementaux identifiés, puis sur les briques des terminaux et des réseaux (2.1.2).

2.1.1 Les centres de données : la composante structurante des impacts environnementaux de l'IA générative

D'après l'étude ADEME-Arcep, mise à jour en 2025, les centres de données représentent 46 % de l'empreinte carbone du numérique en France en 2022, soit près de 13,5 millions de tonnes de $CO2_e$[34]. Cette estimation a été réalisée avant l'essor de l'IA générative ; elle ne permet pas d'isoler la part de l'impact environnemental du numérique, ni même des centres de données qui lui est attribuable. Il est toutefois probable que l'essor de l'IA générative et de ses usages (cf. 1.2.1) entraîne une augmentation des impacts associés aux centres de données, du fait notamment des besoins en capacités de calcul spécifiques à l'IA générative (cf. section 1.1.2).

Les impacts liés à l'IA générative sur la brique centres de données peuvent être appréhendés en utilisant l'ACV, à travers l'analyse des impacts liés aux serveurs et à leur environnement technique (a). Si plusieurs enjeux environnementaux à l'échelle locale sont associés à l'implantation des centres de données (b), leur consommation d'énergie soulève en particulier des préoccupations importantes à l'échelle globale (c).

a) Les impacts liés aux serveurs et à leur environnement technique

La brique centres de données comprend notamment les serveurs informatiques qui y sont hébergés. Or, les usages liés à l'IA générative nécessitent généralement des capacités de calcul plus élevés ainsi que des serveurs spécifiques avec des puces spécialisées, notamment des cartes graphiques, disposant de capacités de calcul beaucoup plus importantes que des serveurs traditionnels. Ces serveurs, ainsi que leurs composants, génèrent plusieurs impacts environnementaux, aux différentes étapes de leur cycle de vie.

- **La fabrication des serveurs et des puces**

La phase de fabrication des serveurs et des puces adaptées à l'utilisation de l'IA[35] nécessite tout d'abord la consommation de métaux contribuant à l'épuisement des ressources abiotiques, dans un contexte où la demande en métaux pour le numérique pourrait augmenter de +179 % entre 2020 et 2050, principalement en raison d'une très forte croissance du nombre d'équipements totaux sur la période[36] (ADEME, 2024b).

---

[34] Par comparaison, selon le CITEPA, la France a émis 369 $MtCO2_e$ en 2024.

[35] Comme les puces GPU de NVIDIA ou Google qui assemble ses propres puces TPU.

[36] Dans le scénario tendanciel, le nombre d'équipements augmente de +522 % (+53 % hors équipements de l'internet des objets) entre 2020 et 2050. Il est à noter que l'étude (ADEME, 2024b) a été produite sur la base de données ne prenant pas

La fabrication de ces puces est également énergivore et émettrice de gaz à effet de serre (Boavizta, 2021) et nécessite des quantités importantes d'eau, exerçant des pressions locales non négligeables notamment en période de sécheresse[37]. S'appuyant sur le démontage d'une carte Nvidia A100 SXM 40GBG GPU, Falk *et al.* (2025a) estiment par exemple l'empreinte carbone de la fabrication d'une telle carte graphique à 150 $kgCO_2e$ et soulignent la nécessité de compléter leur analyse en prenant en compte, en plus de l'impact carbone en phase d'utilisation, déjà bien identifié, l'épuisement des ressources abiotiques, l'impact sur la santé humaine et l'écotoxicité pour la phase de fin de vie.

- **L'utilisation des serveurs**

En phase d'utilisation, les serveurs ont besoin d'être alimentés en électricité. Même si les impacts embarqués, c'est-à-dire les impacts liés à la fabrication des cartes graphiques, restent importants sur les critères environnementaux liés à l'extraction des matières premières, une étude récente de l'ADEME, qui a analysé la composition de certaines cartes graphiques[38], a montré que la phase d'utilisation de ces cartes graphiques pour l'IA[39] a de loin le plus d'impact pour une majorité des critères d'impacts environnementaux analysés, que ce soit pour l'entraînement ou pour l'inférence (ADEME, 2026). Ces résultats corroborent ceux de Schneider *et al.* (2025), qui ont réalisé pour Google une analyse de cycle de vie (ACV) du matériel utilisé pour l'IA générative (le *hardware*, comme les serveurs). Les impacts carbone de la phase d'utilisation des puces représenteraient ainsi 70 à 90 % de l'impact carbone total du cycle de vie d'un entraînement ou d'une inférence (ADEME, 2026).

Par ailleurs, les travaux de Zhang *et al*. (2025) ont mis en évidence que la localisation des serveurs utilisés pour l'entraînement ou de l'inférence est un paramètre déterminant de leur impact carbone, du fait de différences de mix électrique selon le pays où ils sont implantés. À titre d'illustration, en moyenne en 2024, le mix polonais possédait un contenu carbone 1,65 fois plus élevé que celui du mix irlandais et 7,2 fois plus élevé que celui du mix finlandais[40].

À l'exception des résultats de ces travaux, très peu de données environnementales récentes relatives aux semi-conducteurs (GPU, TPU) sont disponibles. En particulier, s'agissant de leur utilisation dans des serveurs adaptés à des usages *cloud* également mobilisés pour l'IA générative, l'absence de transparence sur les *workloads*[41] (*i.e.* sur la quantité de CPU/GPU mobilisés pour remplir une tâche ou sur le taux d'utilisation) ne permet pas de mesurer l'efficacité environnementale de ces serveurs. Or, de tels indicateurs sont clés pour comprendre si les serveurs (et par association les semi-conducteurs) sont sous-utilisés ou utilisés à 100 % de leur capacité maximale et pour évaluer le profil énergétique de ces puces en conditions réelles (Schneider *et al.*, 2025).

---

en compte le développement important de l'IA générative mais sur la base de l'étude ADEME-Arcep sur l'empreinte environnementale du numérique publiée en 2022 et 2023.

[37] Gauthier Rousilhe, Eau et puces électroniques : l'avenir climatique et industriel de Taïwan, 2021.

[38] Cette étude a rassemblé des données issues de la littérature scientifique et a analysé la composition de certaines cartes graphiques qui ont été récupérées, en les démontant et en pesant chacun de ses composants. L'objectif était de construire ce qui est communément appelé un « Inventaire de Cycle de Vie » en ACV afin de modéliser l'impact environnemental d'une carte graphique à partir de sa composition.

[39] Une modélisation a été réalisée sur un logiciel spécialisé en ACV à partir des données de puissance électrique et de la composition des cartes graphiques.

[40] Selon Electricity Maps, en 2024 (site consulté le 28 avril 2026), l'intensité carbone du mix électrique de la Pologne était de 581 gCO_2e/kWh tandis qu'en Irlande (resp. en Finlande), elle était de 352 gCO_2e/kWh (resp. 80 gCO_2e/kWh). Selon l'Agence internationale de l'énergie (AIE), l'intensité carbone moyenne du mix électrique à l'échelle mondiale sur l'année 2024 était de 445 gCO_2e/kWh.

[41] Un *workload* est la quantité de temps et de ressources informatiques nécessaires à un système ou à un réseau pour accomplir une tâche ou générer un résultat particulier.

- **L'environnement technique des serveurs**

Les impacts liés aux centres de données proviennent également des infrastructures qui fournissent un environnement technique permettant le fonctionnement des serveurs dans de bonnes conditions, notamment grâce au refroidissement de ces équipements.

Le choix du système de refroidissement des serveurs a une influence forte sur la consommation d'électricité associée au refroidissement, avec des gains pouvant aller jusqu'à 50 % pour des systèmes de refroidissement avancés comme le refroidissement par liquide ou par immersion, par rapport à des systèmes traditionnels par évaporation ou par air (Patel *et al.*, 2025), mais souvent au détriment de la consommation d'eau. L'efficacité environnementale des techniques de refroidissement s'apprécie notamment à travers les indicateurs d'efficacité d'utilisation de l'énergie (respectivement de l'eau) appelés « *Power Usage Effectiveness* » ou PUE (resp. « *Water Usage Effectiveness* » ou WUE)[42]. L'amélioration de ces indicateurs d'efficacité s'observe notamment à travers la diminution du PUE moyen que ce soit au niveau international, dans les rapports environnementaux des acteurs industriels[43], ou en France où cette tendance est visible dans les chiffres fournis par l'Enquête annuelle « Pour un numérique soutenable » de l'Arcep avec un PUE moyen de 1,42 en 2024 contre 1,46 en 2023 pour les centres de données analysés (cf. encadré « Zoom sur les travaux sur les centres de données dans l'Enquête annuelle Pour un numérique soutenable »).

Au-delà des enjeux liés au refroidissement des serveurs, il existe des enjeux relatifs à la récupération de la chaleur générée par les serveurs. En effet, les serveurs et les autres équipements informatiques produisent de la chaleur qui est rejetée dans l'environnement si elle n'est pas utilisée ; elle est appelée chaleur fatale. En France, il est à noter que tout projet de création ou de modification d'ampleur, pour les centres de données dont la puissance est supérieure à 1 MW, a l'obligation de réaliser au préalable une analyse coûts-avantage de la « faisabilité économique d'améliorer l'efficacité énergétique de l'approvisionnement en chaleur et en froid »[44]. L'ADEME estime que le potentiel de récupération de la chaleur fatale des centres de données pourrait atteindre entre 4 et 13 TWh en 2035 selon les scénarios considérés de développement des centres de données (ADEME, 2026).

b) Les enjeux locaux liés à l'implantation des centres de données

Comme évoqué au chapitre 1, l'essor extrêmement rapide de l'intelligence artificielle conduit à un développement des centres de données, à la fois en nombre et en taille, en attestent les annonces récentes d'investissement en la matière. Au-delà de leur empreinte énergétique globale et des questionnements prospectifs qui en découlent (cf. *infra*), le développement des centres de données emporte des enjeux environnementaux plus locaux, en particulier en termes d'accès à l'électricité, de consommation d'eau et d'emprise foncière. En effet l'implantation d'un centre de données est conditionnée par trois contraintes principales : accès à la connectivité, accès à l'énergie et accès au foncier. Cela a naturellement conduit l'industrie à être géographiquement concentrée (par exemple dans des *hubs* européens comme Paris, Dublin, Londres, Amsterdam et plus récemment Marseille où arrivent de nombreux câbles sous-marins pour les communications électroniques).

---

[42] Le PUE est un indicateur qui mesure l'efficacité de l'utilisation de l'énergie, défini par la norme ISO 30134-2 : il s'agit du ratio entre l'énergie totale consommée par le centre de données et l'énergie consommée par les équipements informatiques qu'il héberge. Ainsi, plus le PUE est élevé, moins le centre de données est réputé efficace. Similairement le WUE correspond au ratio entre le volume l'eau consommée et la consommation d'énergie des équipements informatiques. Un centre de données qui n'utilise pas d'eau pour le refroidissement a donc un WUE nul.

[43] Uptime Institute estime que la valeur moyenne mondiale du PUE était de 1,58 en 2023, contre 2,5 en 2007.

[44] Conformément à la loi n°2025-391 du 30 avril 2025, dite loi DDADUE, transposant la directive (UE) 2023/1791 relative à l'efficacité énergétique en droit français.

- **L'accès à l'électricité**

Du fait d'une concentration géographique des centres de données, les gestionnaires de réseau électrique font face à des problématiques locales de demandes importantes de raccordement électrique. Cela a abouti par exemple à l'annonce d'un moratoire, désormais levé, par l'opérateur national du réseau électrique EirGrid en 2021 pour la région de Dublin[45]. En outre, les demandes de raccordement des centres de données, déjà fortement électro-intensifs par nature, peuvent entrer en concurrence avec celles d'autres secteurs industriels engagés dans une trajectoire de décarbonation, pour lesquels l'accès à l'électricité constitue une condition indispensable à leur électrification (comme le secteur des transports avec les flottes de véhicules individuels et les flottes de bus qui s'électrifient) et, à terme, à la réduction de leurs émissions de GES. À cet égard, en France, le gestionnaire du réseau de transport d'électricité (RTE) a mis en consultation publique un document sur la refonte du raccordement des clients[46], notamment pour faire évoluer la logique de « premier arrivé, premier servi », afin de nourrir sa réflexion sur cette question, en concertation avec les parties prenantes.

- **La consommation d'eau**

Il existe actuellement peu de données portant spécifiquement sur la consommation d'eau de l'IA générative dans les centres de données mais certains articles ont réalisé des estimations de la quantité d'eau consommée par l'entraînement de modèles d'IA générative, comme GPT-3 dans des centres de données Microsoft (voir section 2.2.1) et fournissent des approches méthodologiques permettant de réaliser ces estimations et ainsi de discuter de l'efficacité hydrique des systèmes de refroidissement (Li *et al*., 2023).

Sans distinguer la part liée à l'utilisation de l'IA, les données collectées par l'Arcep dans son enquête annuelle « Pour un numérique soutenable » montrent qu'en 2024, le volume d'eau prélevé ou consommé par les centres de données progresse d'environ 8 %, en raison de la hausse significative de la consommation totale d'électricité[47]. Ce volume est estimé à 6,5 millions de m^3 en 2024, soit la consommation annuelle moyenne d'eau en France de plus de 100 000 personnes (Arcep, 2026d).

Toutefois, les prélèvements d'eau par les centres de données peuvent générer des conflits d'usage dans les localités où les centres sont implantés. L'enjeu de la consommation d'eau des centres de données alimente ainsi des polémiques et oppositions citoyennes comme cela a pu être le cas aux Pays-Bas[48].

Des travaux récents montrent que le cycle de l'eau sera aussi de plus en plus affecté par le changement climatique et qu'il s'agit un enjeu important à l'échelle territoriale (France Stratégie, 2025).

- **L'emprise foncière**

Enfin, si l'artificialisation des sols et l'emprise foncière sont observables à la maille locale avec la naissance de contestations citoyennes sur certains projets[49], la littérature disponible à date reste moins fournie mais relève l'importance de l'enjeu d'artificialisation des sols (Lopez, Diguet, 2023 ;

[45] Ce moratoire s'est terminé fin 2025.

[46] RTE, Raccordement des clients au réseau de transport d'électricité, consultation publique sur la refonte du raccordement des clients.

[47] Au volume d'eau prélevé directement par les centres de données s'ajoute le volume d'eau consommé indirectement, c'est-à-dire le volume d'eau utilisé pour la production d'électricité nécessaire à l'activité des centres de données concernés.

[48] En 2022, selon le journal Noordhollands Dagblad, les données de consommation d'eau d'un centre de données de Microsoft dépassaient fortement les estimations initiales de l'entreprise (84 millions de litres contre un intervalle de 12 à 20 millions de litres évoqué par Microsoft). Cette nouvelle arrivait également dans le contexte d'une sécheresse importante à l'été 2022.

[49] Quadrature du Net, https://www.laquadrature.net/moratoire-data-centers/

Diguet, Lopez, ADEME, 2019). L'enjeu foncier peut générer des polémiques locales, y compris dans un contexte où il existe également des programmes d'attractivité établis et soutenus par les acteurs économiques. Dans certaines villes, comme à Marseille, le développement des centres de données peut entrer en concurrence avec d'autres projets dans des zones déjà très urbanisées (Arcep, 2025c). C'est également l'une des causes du moratoire sur l'implantation de nouveaux centres de données dans des villes comme Amsterdam ou Singapour, où la ressource foncière devient rare.

Zoom sur les travaux sur les centres de données dans l'Enquête annuelle « Pour un numérique soutenable »

Chaque année, l'Arcep rend compte de son travail de collecte de données environnementales effectué en collaboration avec les acteurs, dans le cadre de son **enquête annuelle « Pour un numérique soutenable »**. Pour la troisième année consécutive, en 2026, l'Arcep a restitué les résultats issus de la collecte de données qu'elle réalise auprès des opérateurs de centres de données dont le chiffre d'affaires annuel est égal ou supérieur à 10 millions d'euros hors taxes ou dont la demande de puissance des technologies de l'information installées est supérieure ou égale à 100 kW, soit 23 opérateurs de centres de données. En 2024, ces opérateurs exploitent près de 160 centres de données en France, ce qui représente environ la moitié des centres de données de colocation en service en 2020, selon l'étude ADEME-Arcep. L'étude des impacts de ces acteurs est essentielle dans le contexte d'une croissance des usages numériques qu'ils hébergent.

Les principaux résultats issus de cette édition sont les suivants :

- La consommation des centres de données ne cesse de progresser ces trois dernières années alors que la consommation électrique du secteur tertiaire a significativement diminué en 2023, en raison des hausses de prix de l'électricité, et conserve en 2024, malgré une légère hausse, un niveau inférieur à celui de 2020, année marquée par la crise sanitaire. La consommation électrique des centres de données étudiés continue ainsi de progresser à un rythme soutenu en 2024 pour atteindre 2,7 TWh. Après un ralentissement en 2023 (+ 8 % en un an), la croissance de la consommation électrique des centres de données s'accélère à nouveau en 2024, pour s'établir à + 12 %, revenant à un niveau proche de celui de 2022 (+ 14 %).
- Les opérateurs de centres de données interrogés ont émis au total 178 000 tonnes de GES en 2024[50]. Ces émissions progressent pour la troisième année consécutive à un rythme de plus en plus soutenu. Après avoir augmenté de huit points en 2023 pour atteindre + 13 %, le rythme de croissance de dix points en 2024 s'élève à + 23 %.
- Après deux années de hausse en 2022 (+ 17 %) et 2023 (+16%), dont une hausse exceptionnelle en 2023 liée à des facteurs externes à l'activité de centres de données, tels que des travaux d'aménagement de sites anciens, le volume d'eau prélevé par les centres de données diminue en 2024 (- 15 %) et retrouve un niveau similaire à celui observé en 2022. Cependant, le volume d'eau prélevé ou consommé, progresse d'environ 8 %, en raison de la hausse significative de la consommation totale d'électricité (cf. section précédente 2.1.1b) « La consommation d'eau »). Ce volume est estimé à 6,5 millions de m^3 en 2024, soit la consommation annuelle moyenne d'eau en France de plus de 100 000 personnes.

Ces hausses s'expliquent en partie par la croissance du nombre de centres de données et de leur utilisation. Depuis 2020, leur implantation s'intensifie, se concentre en Île-de-France et concerne des

[50] Sont comptabilisés ici les scopes 1 et 2 des émissions de gaz à effet de serre. S'agissant du scope 2, la méthodologie retenue pour le calcul de ces émissions est la méthodologie *location-based.*

centres avec des capacités informatiques de plus en plus importantes. En effet, comme l'indique RTE[51], l'implantation des centres de données en France connaît un véritable essor ces dernières années. Les projets se multiplient et portent sur des centres de plus en plus importants, avec par exemple des demandes de raccordement de 100 à 200 MW, contre quelques mégawatts auparavant[52].

S'appuyant sur la décision de collecte de données environnementales homologuée le 21 janvier 2026, l'Arcep étend la collecte de données aux fournisseurs de services cloud pour l'édition de 2027. Elle prévoit de collecter auprès de ces acteurs des indicateurs pour suivre l'évolution du nombre de serveurs et processeurs dans les centres de données et évaluer leur capacité à exécuter des tâches d'IA. Ces informations permettront notamment d'apprécier l'impact de l'IA générative sur la consommation électrique des centres de données.

Les travaux se poursuivront pour enrichir la collecte de données en co-construction avec les parties prenantes.

⁂

Dès 2019, l'Arcep s'est emparée de l'enjeu de l'impact environnemental du numérique, et s'est vue ensuite progressivement confier de nouveaux pouvoirs par le législateur, pour collecter auprès des acteurs du numérique des données sur leur empreinte environnementale. Cette collecte de données, entamée dès 2020 auprès des quatre principaux opérateurs de communications électroniques, a été élargie grâce à la loi « REEN 2 » du 23 décembre 2021 aux opérateurs de centres de données notamment. En 2024, la loi visant à sécuriser et réguler l'espace numérique (loi « SREN ») a élargi les pouvoirs de collecte de l'Arcep aux fournisseurs de services d'informatique en nuage.

La démarche de l'Arcep a d'ailleurs été saluée par la Banque mondiale et l'Union internationale des télécommunications (UIT). En 2025, ces deux institutions ont publié un rapport conjoint intitulé « Mesurer l'impact environnemental du numérique – étude du cas Arcep » (Banque Mondiale et UIT, 2025), consacré à l'enquête annuelle de l'Arcep, « *premier et unique régulateur sectoriel à publier régulièrement des indicateurs fondés sur les données collectées auprès des acteurs du numérique afin d'évaluer et de suivre dans le temps leur impact environnemental* ». Ces institutions invitent à reproduire les conditions de cette collecte de données environnementales à l'international.

c) Les enjeux liés à la consommation d'énergie et aux émissions de gaz à effet de serre

Une empreinte énergétique en forte accélération

Il existe un consensus sur la hausse anticipée de la consommation d'énergie des centres de données liée à la croissance de l'IA générative. L'Agence internationale de l'énergie (AIE) a par exemple estimé dans son dernier rapport que la consommation d'électricité mondiale des centres de données était de 415 TWh en 2024, soit 1,5 % de la consommation d'électricité mondiale, et qu'elle pourrait doubler d'ici à 2030 notamment du fait de l'essor de l'IA, pour atteindre potentiellement près de 1 000 TWh, soit l'équivalent de la consommation électrique totale du Japon en 2024[53] (AIE, 2025). D'après l'Observatoire de l'énergie et de l'IA de l'AIE, en 2024, la consommation électrique des centres de

[51] RTE, sigle du Réseau de transport d'électricité, désigne le gestionnaire du réseau public de transport d'électricité haute tension en France métropolitaine

[52] Source : RTE, Data centers : 11 chiffres sur leur essor en France et leurs besoins en électricité

[53] Japan Ministry of Economy, Trade and Industry, 2025 (lien).

données représentait 4,4 % de la consommation électrique totale aux États-Unis et 2,3 % dans l'Union européenne[54].

En France, l'étude ADEME-Arcep (2023) a permis d'estimer que la consommation électrique du parc global des centres de données en France était d'environ 12 TWh en 2020 et pourrait atteindre près de 16 TWh en 2030, sans tenir compte des centres de données à l'étranger servant un usage français et sans une prise en compte spécifique des dynamiques récentes sur la croissance de l'IA générative.

Les données sur la consommation électrique des centres de données collectées par l'Arcep sur un périmètre restreint (cf. encadré « Zoom sur les travaux sur les centres de données dans l'Enquête annuelle « Pour un numérique soutenable » » *supra*) révèlent également une hausse de la consommation électrique sur les dernières années (+14 % entre 2021 et 2022 et +8 % entre 2022 et 2023 et +12 % entre 2023 et 2024). Si cette tendance à la croissance de la consommation électrique des centres de données est observée à différentes échelles, il est difficile d'estimer la part attribuable à l'IA. Des estimations existent mais sont actuellement difficiles à contre-expertiser, en raison d'un manque de données. L'ADEME estime par exemple que la part de l'IA dans la consommation électrique des centres de données en France est de 13 % pour l'année 2025 en tenant compte des centres de données à l'étranger servant des usages français (ADEME, 2026) et le Shift Project estime que cette part est de l'ordre de 15 % dans le monde (The Shift Project, 2025).

De fait, le développement de l'IA est considéré comme le principal facteur de croissance de la consommation électrique des centres de données au cours des prochaines années. Des études prospectives de la consommation d'électricité des centres de données, notamment en lien avec l'IA, ont été réalisées ces dernières années par différents types d'acteurs (institutionnels, industriels, analystes financiers) et étayent ce postulat. Au-delà de la nécessaire prudence quant aux résultats[55], il est intéressant de noter que ces différentes études dessinent un consensus sur l'augmentation significative à venir, de la consommation d'électricité liée aux centres de données, en raison d'un accroissement des usages. Comme le montre le graphique ci-dessous, les études convergent pour mettre en évidence que, d'ici à 2030, la consommation d'électricité des centres de données devrait augmenter considérablement, notamment du fait du développement de l'IA générative.

[54] AIE, 2025 Energy and AI Observatory.

[55] Ces analyses prospectives doivent être prises avec prudence car elles reposent sur des données partielles et des hypothèses qui influencent fortement les résultats. Elles constituent donc des trajectoires possibles, et non des prévisions.

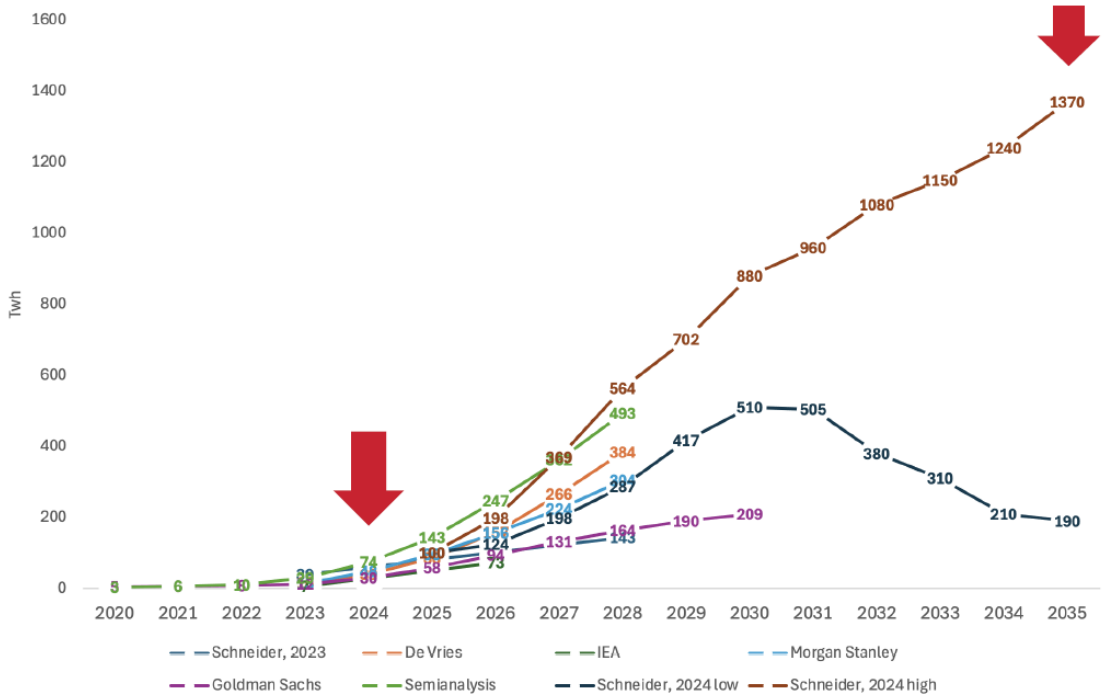


Figure 8 : Consommation d'électricité liée à l'IA selon différentes sources[56]

Au-delà de 2030, l'évolution de la consommation énergétique dépendra particulièrement du rythme de déploiement des centres de données et des gains d'efficacité réalisés. À cet égard, dans tous les scénarios considérés par l'AIE, l'agence a estimé (cf. graphique ci-dessous) que la consommation globale des centres de données connaîtrait une hausse marquée d'ici 2030 (entre 700 et 1250 TWh, comparé à 415 TWh en 2024), et de manière plus incertaine entre 2030 et 2035 (entre environ 750 TWh et 1750 TWh). Cette évolution dépend notamment des hypothèses de gains d'efficacité énergétique, de l'évolution de la demande des services d'IA générative et des contraintes locales précédemment évoquées comme l'accès à l'électricité, au foncier ou l'acceptabilité sociale. Ces hypothèses se traduisent par les scénarios suivants : un scénario *Lift-Off* (« Décollage ») qui considère un taux élevé d'adoption, un scénario *Base* (« Base ») tendanciel, un scénario *High Efficiency* (« Haute efficacité ») qui considère des progrès technologiques importants en matière d'efficacité énergétique et un scénario *Headwinds* (« Vents contraires ») qui considère un taux faible d'adoption.

[56] Source : Lynn Kaack, graphique publié en 2025 à l'occasion du Sommet pour l'action sur l'IA à Paris, à partir des données issues du Roegen Centre for Sustainability et d'INFRAS à partir des données de Schneider Electric, 2023 ; Schneider Electric, 2024 ; De Vries, 2023 ; AIE, 2024 ; Morgan Stanley, 2024 ; Goldman Sachs, 2024 ;Semianalysis, 2024.

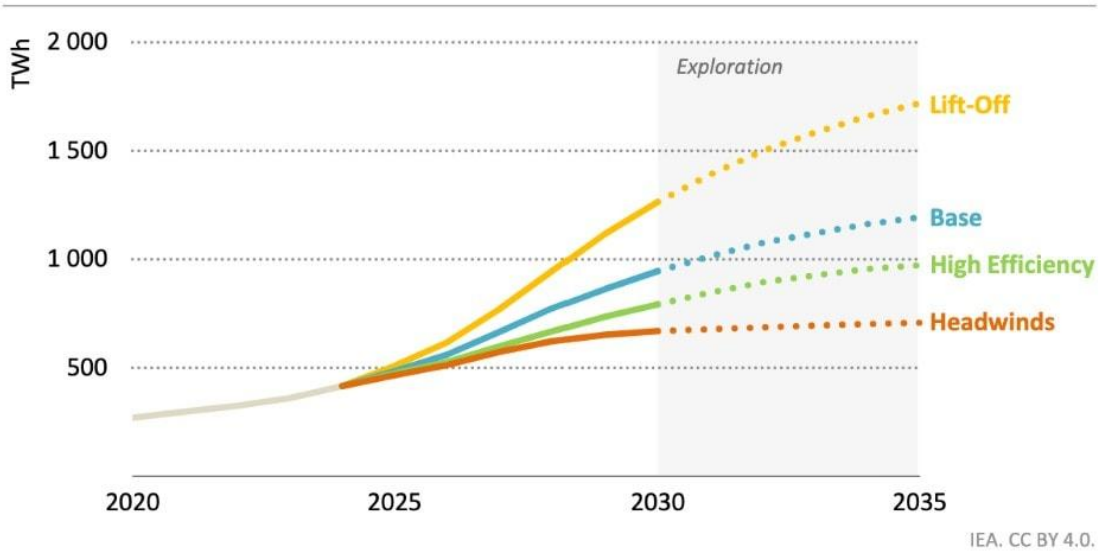


Figure 9 : Projections de la consommation d'électricité mondiale des centres de données selon différents scénarios. Source : AIE, 2025

Une pression croissante sur les stratégies et trajectoires de décarbonation

Cette hausse rapide de la consommation électrique des centres de données due à la croissance des usages de l'IA générative conduit certains des principaux acteurs de l'IA générative, notamment aux États-Unis, à faire évoluer leurs stratégies. Ces grands acteurs tendent à élargir leur positionnement pour se placer en tant que producteurs d'électricité pour garantir la couverture de leurs besoins, comme l'illustre l'intérêt manifesté pour la relance de certaines capacités existantes, à l'image de la centrale nucléaire de *Three Mile Island*[57] ou de centrales à gaz. Le développement de l'IA fait également émerger des interrogations plus larges quant à la capacité des plus grands acteurs de l'IA générative et du *cloud* à concilier croissance de leurs activités et respect de leurs engagements climatiques. Malgré le peu d'informations environnementales communiquées par les acteurs de l'IA générative, ceux-ci reconnaissent que les impacts environnementaux des modèles d'IA, en particulier d'IA générative sont loin d'être négligeables et remettent en cause leur stratégie bas carbone compte tenu du développement des centres de données. En 2024, Google et Microsoft ont par exemple reconnu que l'atteinte de leurs objectifs initiaux de décarbonation à horizon 2030 étaient compromis à cause du développement des IA génératives et ont revu leurs objectifs à la baisse[58]. Melanie Nakagawa, la responsable environnement de Microsoft a notamment déclaré en 2025 : « *En 2020, les dirigeants de Microsoft avaient qualifié nos objectifs de durabilité d' « objectif Lune » et près de cinq ans après, nous avons dû reconnaître que la Lune s'est éloignée* »[59].

Au-delà des stratégies des acteurs industriels, la croissance de consommation électrique des centres de données soulève des enjeux à l'échelle des nations et de leurs trajectoires de décarbonation. Cela

[57] AP News, Le département américain de l'énergie accorde un prêt d'un milliard de dollars pour contribuer au financement du redémarrage d'un réacteur nucléaire de Three Mile Island, 19 novembre 2025.

[58] Technology Magazine, 2025 ; Les Echos, 2024; Novethic, 2024. Voir également les rapports environnementaux 2024 de Google et de Microsoft.

[59] Microsoft, 2025, Progress on the road to 2030. Traduction de courtoisie de : "*In 2020, Microsoft leaders referred to our sustainability goals as a "moonshot," and nearly five years later, we have had to acknowledge that the moon has gotten further away.*"

a par exemple abouti à une relance massive d'infrastructures fonctionnant aux énergies fossiles (notamment au gaz naturel aux États-Unis) en raison de capacités de production bas carbone limitées (The Shift Project, 2025). Face à ce développement et à un manque de données pour en mesurer toute l'intensité, l'agence fédérale américaine de l'énergie (EIA), à l'initiative des sénateurs Warren et Hawley, prévoit ainsi de lancer une enquête nationale obligatoire pour mesurer la consommation électrique des centres de données aux États-Unis[60], après le lancement de projets pilotes dans les États du Texas et de Washington. L'évaluation couvrira non seulement la consommation d'électricité annuelle, mais aussi la production d'énergie locale, souvent issue du gaz, et l'efficacité des systèmes de refroidissement de ces infrastructures[61].

Pour certains pays comme l'Irlande, l'accueil massif de centres de données remet d'ores et déjà en cause des objectifs de décarbonation. En effet, cette dynamique s'accompagne d'une pression croissante sur le système électrique (dont une partie importante est aujourd'hui basée sur le gaz) : la consommation des centres de données y a augmenté de 24,7 % entre 2012 et 2022, portée notamment par le développement du *cloud*[62]. Aujourd'hui, ces infrastructures représenteraient déjà près de 20 % de la consommation électrique nationale irlandaise, soit environ 18 points de plus que la moyenne mondiale, une proportion qui pourrait atteindre 30 % dès 2028 selon certaines projections (AIE, 2024). Dans ce contexte, la question de l'accueil et de l'encadrement de ces installations constitue un véritable enjeu pour le pays, tant sur le plan énergétique qu'environnemental.

2.1.2 Des interrogations sur les conséquences à moyen terme sur les terminaux et les réseaux

a) Terminaux : une intensité énergétique d'utilisation des services d'IA générative potentiellement plus élevée et un risque de renouvellement des terminaux

Comme précisé en section 1.1.1, la brique terminaux regroupe les équipements (ex : téléviseurs, smartphones, ordinateurs...) qui sont utilisés pour consommer des services numériques, dont des usages d'IA générative. Pour rappel, en France en 2024, les foyers détiennent en moyenne 9,6 appareils numériques avec écran, dont 1,8 sont inutilisés et pourraient être recyclés ou vendus[63].

D'après l'étude ADEME-Arcep, les terminaux représentent 50 % de l'empreinte carbone du numérique en France en 2022, soit près de 14,7 millions de tonnes de $CO2_e$[64]. À l'échelle des services numériques, il s'agit de la brique ayant le plus fort impact sur la plupart des critères.

Le développement des services d'IA générative est susceptible d'avoir des impacts sur cette brique de différentes façons. Tout d'abord, ce développement conduit à un accroissement de la consommation énergétique lié à l'utilisation des terminaux car ces services sont gourmands en énergie. La consommation des services d'IA générative peut en effet conduire au traitement de requêtes dans le *cloud* (augmentant l'échange et le traitement de données avec le terminal) ou directement sur le

[60] Wired, 2026, The US Government Will Ask Data Centers How Much Power They Use

[61] EIA, 2026, EIA launches pilot survey on energy use at data centers

[62] Comme mentionné par l'Autorité de la Concurrence dans son avis sur l'intelligence artificielle publié en 2024, « *le* cloud *apparait comme un passage obligé pour accéder à la puissance de calcul nécessaire à l'entrainement de modèle. Il est également un vecteur de diffusion des modèles en aval sur les places de marchés. Ces places de marché permettent aux développeurs de rendre leurs modèles facilement accessibles aux entreprises utilisatrices de services* cloud*, ce qui encourage les développeurs à rendre leurs modèles disponibles sur chacun des fournisseurs de* cloud*.* »

[63] Arcep, Arcom, CGE, ANCT, Baromètre du numérique, édition 2025.

[64] D'après la mise à jour de l'étude ADEME-Arcep réalisée par l'ADEME en novembre 2024.

terminal, accélérant la décharge de la batterie notamment pour les smartphones[65] (The Shift Project, 2025). En outre, ces services intensifs en énergie peuvent être imposés aux utilisateurs finals par les fournisseurs de services (Arcep, 2026a ; Beignon *et al*., 2025 ; Limites Numériques, 2025), le consommateur ayant une capacité limitée à choisir que ces fonctionnalités soient ajoutées aux services utilisés (ex : intégration de fonctionnalités d'IA générative aux services de messageries instantanées). Certains terminaux pourraient alors subir une obsolescence accélérée s'ils ne sont pas assez performants pour utiliser de façon satisfaisante ces services, en dépit des gains d'efficacité énergétique des terminaux recherchés par les industriels[66] (The Shift Project, 2025).

Le développement de services d'IA générative peut également conduire à la production de nouveaux terminaux spécifiquement adaptés, avec l'implantation de processeurs graphiques (GPUs) ou d'autres types de puces (NPUs). Des terminaux comportant une IA générative embarquée[67] ont d'ores et déjà été mis sur le marché. Ces nouveaux terminaux pourraient avoir un impact environnemental plus élevé, lors de la phase de fabrication et d'utilisation et surtout conduire au renouvellement anticipé des anciens terminaux qui ne disposeraient pas de capacités de calculs compatibles avec des services d'IA générative, accélérant ainsi le phénomène d'obsolescence.

Actuellement, des effets massifs ne sont pas constatés sur le marché, les services actuels fonctionnant *a priori* à matériel constant car les requêtes sont principalement traitées *via* le *cloud* et transmises par le réseau. Ces tendances pourraient toutefois évoluer rapidement, notamment si les stratégies des acteurs les incitent à davantage embarquer de la puissance de calcul dans le terminal.

[65] Greenspector, Quels impacts environnementaux pour les IA locales sur nos smartphones ?, 2025

[66] GSMA liste des actions que mettent en place les opérateurs pour améliorer l'efficacité énergétique des terminaux, avec l'existence d'une initiative dédiée appelée EcoRating.

[67] Les services d'IA générative embarqués sont par exemple Apple Intelligence (Apple), Galaxy AI (Samsung), Gemini Nano (Google), AI Engine (Qualcomm), HyperOS (Xiaomi) ou Harmony IA (Huawei). Les smartphones iPhone 16 Pro Max, Pixel 9 Pro XL et Galaxy S25 permettent d'embarquer ces services d'IA générative.

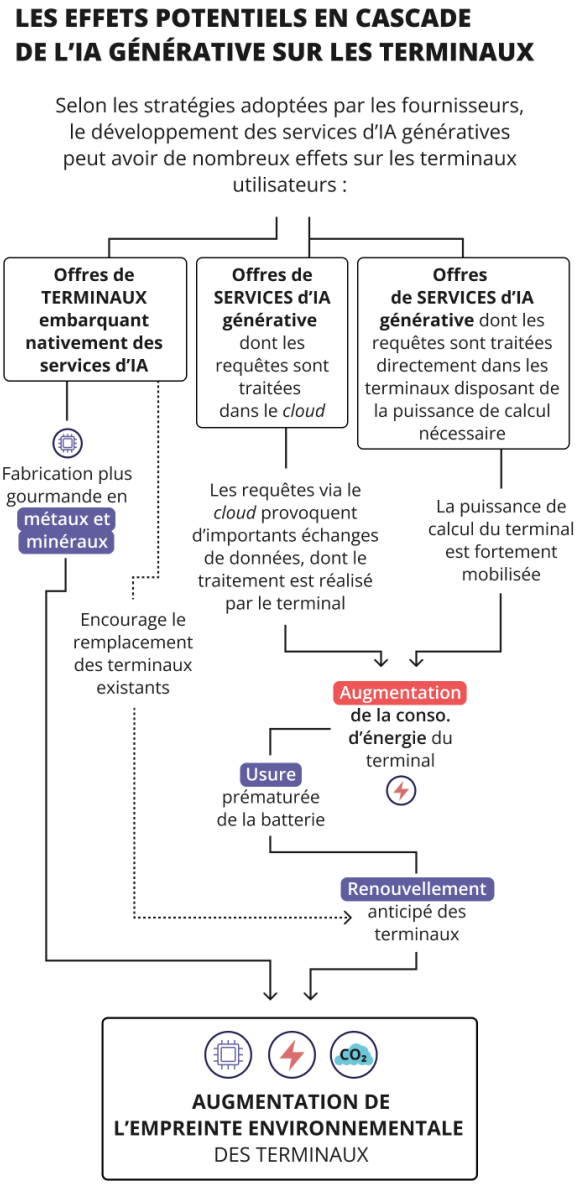


Figure 10 : Représentation simplifiée des différents effets de l'IA sur les terminaux. Source : Arcep.

b) Réseaux : des capacités *a priori* suffisantes à court-terme mais des interrogations à moyen-terme

D'après l'étude ADEME-Arcep, la brique réseaux représente 4 % de l'empreinte carbone du numérique en France en 2022 soit près de 1,2 millions de tonnes de $CO2_e$[68]. Il s'agit de la brique la moins influente des trois sur l'essentiel des critères d'impact environnemental étudiés.

[68] D'après la mise à jour de l'étude ADEME-Arcep réalisée par l'ADEME en novembre 2024.

L'impact des services d'IA générative sur les réseaux est susceptible de se matérialiser de plusieurs façons[69]. Tout d'abord, plusieurs travaux montrent que la phase d'entraînement des modèles d'IA repose sur une puissance de calcul d'ampleur et une communication entre les nœuds d'apprentissage qui sollicite davantage les réseaux que les usages moyens du numérique[70] (Arcep, 2025c).

Ensuite, en phase d'inférence, les applications d'IA, en particulier la génération de contenus intensifs en consommation de données (notamment la génération de vidéos ou la génération de contenus interactifs nécessitant une latence réduite et une synchronisation accrue avec l'agent IA sur le serveur), pourraient occuper une part croissante du trafic Internet, en fonction de l'augmentation du nombre de requêtes. En particulier, selon les estimations d'équipementiers télécoms[71], le développement du « trafic IA indirect », c'est-à-dire du trafic associé à la création de contenu hyper-personnalisé adapté aux utilisateurs (à partir de leurs données de navigation, de localisation, la météo ou l'heure) par des algorithmes de recommandation pouvant s'appuyer sur l'IA générative, pourrait générer un effet « boule de neige » où chaque prompt générerait à son tour une quantité importante de requêtes ou d'appels à d'autres systèmes IA, amplifiant ainsi la quantité de données en circulation. Le trafic IA indirect pourrait alors dépasser le trafic IA direct (The Shift Project, 2025).

Enfin, de manière prospective, de nombreuses questions se posent sur la distribution[72] des calculs réalisés : ils peuvent avoir lieu de manière centralisée ou décentralisée ce qui n'impliquerait pas les mêmes niveaux de dimensionnement des réseaux selon les cas (Arcep, 2025c). Il existe également une incertitude sur la capacité des terminaux à avoir des capacités IA en local (cf. *supra*), ce qui permettrait de soulager la charge sur les réseaux mais qui nécessiterait toujours de se synchroniser avec le *cloud*. Il faut également noter que ces incertitudes sur le développement, le dimensionnement des réseaux et les impacts associés sont induites par des modèles d'affaires et des stratégies différents d'un acteur à l'autre (cf. 1.2.2).

De manière générale, même avec un déploiement massif de l'IA générative, les capacités réseaux pourraient être suffisantes à court terme pour répondre à cette demande : l'IA générative ne semble pas actuellement être le facteur de saturation des réseaux. Sur la base de la littérature étudiée et des entretiens menés dans le cadre de la démarche Réseaux du Futur, il ne semble actuellement pas que l'IA générative remette en cause le dimensionnement des réseaux fixes (Arcep, 2025c). À moyen terme en revanche, des questions se posent pour les réseaux mobiles. En effet, la croissance des services utilisateurs reposant sur l'IA pourrait entraîner une augmentation des capacités de calcul à l'*edge* qui concernerait certains lieux spécifiques du réseau**.** Elle pourrait par ailleurs amplifier le volume du trafic ascendant de données (de l'utilisateur vers le réseau), entraînant des besoins d'évolution du dimensionnement des réseaux habituellement conçus avec le sens descendant (du réseau vers l'utilisateur) comme sens dominant. Des travaux de standardisation sont aussi actuellement en cours sur la 6G qui a vocation à intégrer l'IA de façon native dans les réseaux. Enfin, l'intégration de l'IA dans les réseaux pourrait améliorer les performances du réseau et offrir un levier d'optimisation de sa capacité.

⁂

[69] Cette analyse a été menée de manière spécifique pour les réseaux mobile et fixe dans le cadre des travaux prospectifs de l'Arcep (2025), Réseaux du futur, dans la note « Intelligence artificielle et les réseaux télécom ».

[70] Le site Statista estime qu'Azure offre en 2023 une bande passante de 1 600 Gbit/s pour les besoins d'instances de calcul pour de l'IA.

[71] Nokia. (2024). Global network traffic report. https://onestore.nokia.com/asset/213660

[72] Autrement dit, la répartition des calculs selon leur localisation.

Il ressort de cette analyse par brique que le principal enjeu environnemental associé au développement rapide de l'IA et à ses perspectives de déploiement massif concerne les centres de données. L'essor de l'IA devrait en effet s'accompagner d'une croissance des besoins en alimentation électrique des centres de données et de leurs infrastructures de calcul intensif. Il devrait également se traduire par une croissance de la consommation de ressources métalliques et minérales, nécessaires notamment à la fabrication des puces utilisées dans les serveurs. Localement, l'accès à l'électricité et les impacts sur la consommation d'eau peuvent générer des conflits d'usage tandis que l'accès au foncier peut générer des contestations citoyennes. Si les enjeux liés aux centres de données sont *a priori* les plus importants et les moins incertains, les impacts environnementaux de l'IA liés au renouvellement des terminaux utilisateurs et au redimensionnement des réseaux mobiles, notamment les réseaux mobiles sont à considérer à moyen terme.

2.2 Analyse par phase : un accroissement des usages qui augmente l'impact de la phase d'inférence par rapport à la phase d'entraînement des modèles

Des travaux récents sur l'IA générative permettent d'établir un premier socle de connaissances sur les impacts environnementaux des phases d'entraînement et d'inférence des services d'IA, malgré l'opacité des fournisseurs de services. Cette partie se concentre sur les travaux apportant des éclairages sur ces deux phases. Il convient toutefois de noter que les étapes du cycle de vie d'un modèle d'IA ne se limitent pas à l'entraînement et à l'inférence. En effet, les entraînements qui n'aboutissent pas à la commercialisation d'un modèle, de même que les réentraînements (voire à l'apprentissage continu) ou le *fine-tuning* sont consommateurs d'énergie et génèrent d'autres impacts environnementaux.

2.2.1 Phase d'entraînement : un impact dépendant du nombre de paramètres des modèles, du mix électrique et du matériel utilisé

Plusieurs travaux existent sur l'impact environnemental de la phase d'entraînement, provenant de recherches académiques ou d'études du secteur lui-même. Alors que les travaux académiques s'appuient sur des méthodologies et des hypothèses transparentes, les études venant d'acteurs privés s'appuient en grande majorité sur des méthodologies opaques et présentent donc des résultats non comparables et non opposables.

a) Nombre de paramètres et durée d'entraînement : deux facteurs clés de la consommation électrique de l'entraînement d'un modèle

Comme développé au chapitre 1, le nombre de paramètres des modèles a considérablement augmenté (cf. Figure 4) au cours des dernières années (Varoquaux *et al.*, 2025)[73], ce qui interroge sur la croissance de la puissance de calcul, et donc de la consommation électrique, nécessaire à leur entrainement.

Il ressort des travaux académiques que la durée d'entraînement et les ressources de calculs mobilisées constituent des facteurs déterminants de la consommation électrique de la phase d'entraînement de ces modèles (Strubell *et al*., 2019 ; Patterson, 2021). Or, le nombre de paramètres a un effet direct sur la durée d'entraînement : plus il y a de paramètres, plus la durée d'entraînement est allongée et plus la quantité de ressources à mobiliser est importante.

[73] 120 millions de paramètres pour le premier modèle de ChatGPT sorti en 2018 contre 1 000 milliards pour le modèle GPT-4 sorti en 2024.

Comment mesurer la consommation électrique de la phase d'entraînement ?

De nombreux travaux académiques ont fourni des chiffres sur l'impact environnemental de la phase d'entraînement en s'appuyant sur des outils transparents et librement accessibles ou en s'appuyant sur des estimations à partir de quelques données industrielles publiques. Kim *et al*. (2025), dans leur synthèse des travaux académiques sur les impacts en matière de consommation d'électricité et d'émissions de gaz à effet de serre de la phase d'entraînement, font le constat qu'il existe de nombreux outils de mesure déjà existants qui permettent de quantifier et de suivre les émissions de gaz à effet de serre liées à la phase d'entraînement des modèles. Il peut s'agir d'estimations (modèles empiriques ou modèles statistiques) comme ML Emissions Calculator (Lacoste *et al.*, 2019) ou Green Algorithms (Lannelongue *et al*., 2021), de mesures par capteurs comme Carbontracker (Anthony *et al.*, 2020), ou d'outils hybrides comme Code Carbon (Schmidt *et al*., 2021) ou Eco2AI (Budennyy *et al.*, 2020). L'évaluation de la phase d'entraînement permise par ces outils constitue un point de départ pour discuter de la durabilité des services d'IA générative.

Strubell *et al*. (2019) ont réalisé la première étude quantitative évaluant l'impact carbone de l'entraînement des modèles de langage naturel (Transformer, ELMo, BERT et GPT-2). Cette étude, qui intègre une mesure de la consommation d'énergie des GPU et des CPU, a mis en évidence que la taille des modèles constitue bien un facteur clé de leur consommation énergétique : les modèles légers émettent bien moins de carbone que les modèles les plus grands. À titre d'exemple, les émissions de gaz à effet de serre de l'entraînement d'un grand modèle Transformer (avec un réseau de neurones dit *neural architectural search*) étaient estimées à 284 tCO_2e contre 0,087 tCO_2e pour un modèle plus léger de type Transformer big sans réseau de neurones.

Budennyy *et al*. (2022) ont quant à eux estimé les émissions de gaz à effet de serre des modèles de *machine-learning* générant des images Malevich (1,3 milliard de paramètres) et Kandinsky (12 milliards de paramètres) et estimé à 17 le rapport entre les émissions de gaz à effet de serre de ces deux modèles. D'après leurs estimations, en tenant compte du mix électrique russe, le premier modèle émet 330 gCO_2e[74] (correspondant à une consommation d'énergie de 1,37 kWh) pour une durée d'entraînement de 4h19, et le second 5,89 $kgCO_2e$[75] (soit une consommation d'énergie de 24,5 kWh) pour une durée d'entraînement de 9h45. Bien que ces modèles ne soient pas comparables aux grands modèles d'IA générative, cette estimation de l'empreinte carbone de leur entraînement met en évidence l'impact du nombre de paramètres des modèles.

Dans une évaluation d'un modèle ouvert, le modèle généraliste BLOOM[76], entraîné sur 176 milliards de paramètres, Luccioni *et al*. (2023) ont aussi montré que la phase d'entraînement est particulièrement consommatrice en ressources (serveurs, GPUs) et en électricité : 348 GPUs (NVIDIA A100) ont été mobilisées pendant 118 jours, pour une consommation de 433 MWh, et une empreinte carbone liée à la consommation électrique des serveurs de 30 tCO2e. En y ajoutant l'empreinte carbone de la fabrication des équipements, Luccioni *et al*. (2023) estiment que l'empreinte carbone totale de l'entraînement de ce modèle était d'environ 124 tCO2e, un ordre de grandeur relativement faible en comparaison par exemple avec l'entraînement de GPT-3, en raison d'un matériel utilisé

[74] Soit 1,27 gCO2e/min d'entraînement (ou 0,005 kWh/min d'entraînement).

[75] Soit 10 gCO2e/min d'entraînement (ou 0,04 kWh/min d'entraînement).

[76] Il s'agit d'un modèle initié dans le cadre du projet *BigScience* principalement par l'entreprise franco-américaine HuggingFace, le CNRS, GENCI ou le Ministère de l'Enseignement Supérieur et de la Recherche. Le modèle a été entraîné en France sur le supercalculateur Jean Zay.

considéré comme efficace et du mix électrique français bas carbone (voir Annexe 5.5 qui compare les ordres de grandeur de l'entraînement des GPT-3 et de BLOOM). Il est toutefois à noter que ce modèle est relativement modeste puisqu'il n'est entraîné que sur un ensemble de données de 1,6 To (contre 45 To estimés pour entraîner le modèle GPT-3[77]).

Des travaux ont également cherché à identifier des pistes pour réduire la consommation électrique de la phase d'entraînement. Si la taille des modèles constitue un facteur clé, Patterson *et al*. (2021) ont montré, dans des travaux réalisés pour Google, que les progrès réalisés en matière d'efficacité énergétique dans les centres de données Google et dans la conception des modèles (architecture notamment) auraient permis de réduire les émissions de gaz à effet de serre liées à l'entraînement des modèles étudiés par Strubell *et al*. (2019) d'un facteur 80. Si cette étude peut être discutable d'un point de vue méthodologique[78], elle permet de souligner l'importance de tenir compte des gains d'efficacité liées aux évolutions technologiques sur les dernières années. Le tableau ci-dessous récapitule les ordres de grandeurs des différents travaux académiques mis en avant dans cette section : ces chiffres doivent être pris avec précaution sachant que toutes les informations ne sont pas toujours communiquées et qu'ils peuvent interroger quant à la fiabilité de certaines données d'entrée.

Modèle entraîné	Nombre de paramètres	Durée d'entraînement	Ressources mobilisées	Consommation d'électricité des serveurs liée à l'entraînement	Émissions de gaz à effet de serre associées à l'utilisation des serveurs
GPT-3	175 milliards	15 jours	10 000 GPU	1 287 MWh	552 tCO_2e
Transformer big avec réseau de neurones	213 millions (avec 979 millions d'étapes d'entraînement)	10 heures (équivalent de 274 120 heures sur 8 GPU P100)	1 TPUv2 core	656 MWh	284 tCO_2e
BLOOM	176 milliards	118 jours	384 GPU	433 MWh	30 tCO_2e
Transformer big sans réseau de neurones	213 millions (avec 300 000 étapes d'entraînement)	3,5 jours	8 GPU	0,201 MWh	0,087 tCO_2e
Kandinsky	12 milliards	9h45	Non communiqué	0,0245 MWh	0,00589 tCO_2e
Malevich	1,3 milliards	4h19	Non communiqué	0,00137 MWh	0,00033 tCO_2e

Tableau 1 – Récapitulatif des ordres de grandeur sur la consommation d'électricité et les émissions de gaz à effet de serre en phase d'entraînement des modèles analysés dans le cadre de la revue de littérature scientifique

Des acteurs du numérique tels que Meta ou Mistral AI ont également réalisé leurs propres études d'impact mais elles sont, soit incomplètes car ne remplissant pas les exigences d'une ACV, soit très peu transparentes sur les hypothèses considérées et donc difficilement contre-expertisables. Meta a par

[77] Sénat, 2024, ChatGPT, et après ? Bilan et perspectives de l'intelligence artificielle

[78] En particulier, le choix de communiquer uniquement les résultats en approche *market-based* est discutable, les résultats des analyses de cycle de vie devant également être communiquées suivant une approche *location-based*. Cf. section 2.1.1b)).

exemple communiqué en juillet 2024 sur l'empreinte carbone de la phase d'entraînement de LLaMa3-405B[79] : celle-ci aurait mobilisé 16 000 GPU H100 et émis environ 9 000 tonnes de CO_2e sur un périmètre restreint puisque seule l'électricité consommée par les puces GPU a été comptabilisé, sans prise en compte de la consommation électrique des CPU ou des équipements réseaux, des équipements non-informatiques servant à refroidir les serveurs ou de la fabrication des serveurs[80]. En juillet 2025, Mistral a communiqué quelques résultats d'une ACV de son modèle Large 2[81] : il aurait généré 20 400 tonnes de CO_2e en phase d'entraînement sur une durée de 18 mois (Mistral, 2025)[82].

Modèle entraîné	Nombre de paramètres	Durée d'entraînement	Ressources mobilisées	Consommation d'électricité des serveurs liée à l'entraînement	Émissions de gaz à effet de serre associées à l'utilisation des serveurs
Mistral Large 2	123 milliards	18 mois	Non communiqué	Non communiqué	20 400 tCO_2e
Meta LLaMa3-405 B	405 milliards	Non communiqué	16 000 GPU H100	Non communiqué	9 000 tCO_2e

Tableau 2 – Récapitulatif des ordres de grandeur sur la consommation d'électricité et les émissions de gaz à effet de serre en phase d'entraînement communiqués par des fournisseurs de modèles

Au-delà de la consommation d'électricité et des émissions de gaz à effet de serre associées, d'autres travaux ont analysé l'impact de l'entraînement d'un modèle sur la consommation d'eau. Li *et al.* (2023) ont ainsi estimé que l'entraînement de GPT-3 avait nécessité 700 000 litres d'eau, en 15 jours, pour le refroidissement des centres de données, pour une empreinte en eau totale (*i.e.* y compris prélèvements indirects liés à la production d'électricité[83]) de 3,5 millions de litres [84].

b) Intensité carbone du mix électrique du pays d'entraînement : un facteur-clé à anticiper dès la conception du modèle

Comme vu dans la section précédente, la phase d'entraînement est particulièrement électro-intensive car elle nécessite d'importantes puissances de calcul. Or, les émissions de gaz à effet de serre liées à l'utilisation des serveurs dépendent notamment de l'intensité carbone du mix électrique utilisé (cf. section 2.1.1). Dans la mesure où un modèle est généralement entraîné une seule fois sans enjeu

[79] Voir l'analyse sur le site de Hugging Face des données qui sont disponibles sur le site de NVIDIA, utilisant cette méthodologie.

[80] Carbone 4 estime qu'en considérant le cycle de vie complet, l'impact carbone serait de 19 600 tonnes de CO_2e.

[81] Cette ACV, réalisée par Carbone 4 en partenariat avec l'ADEME, s'est appuyée sur le Référentiel général pour l'IA frugale (AFNOR, 2024), initiative de référence développée en partenariat avec le Ministère de la Transition Écologique. Toutefois, seul un communiqué a été mis à disposition sur le site internet de Mistral et aucun rapport détaillé n'a été publié, ce qui ne permet pas de vérifier les informations transmises et notamment les paramètres clés de la modélisation.

[82] À titre de comparaison, l'empreinte carbone annuelle moyenne d'un Français s'établit à 8,5 tonnes de CO_2 équivalent par adulte et par an pour l'année 2023, selon le Citepa et l'ABC. Pour rappel, le numérique représente 4,4 % de l'empreinte carbone du pays en 2022, soit 29,5 MtCO2e de GES émises en 2022 (ADEME, 2025c).

[83] L'empreinte eau se base notamment sur l'indicateur EWIF (Energy Water Intensify Factor) qui quantifie la quantité d'eau nécessaire pour la production d'électricité. Selon le rapport annuel de Microsoft pour l'année 2022, l'empreinte eau était de 6,4 milliards de litres.

[84] Par ordre de comparaison, en 2021, selon l'INSEE, le volume d'eau prélevé était de 29 543 millions de m^3 en France. Selon Eurostat, ce chiffre s'élève à environ 218 000 millions de m^3 dans l'Union européenne.

important lié à la latence, la localisation de l'entraînement est donc un paramètre important à anticiper pour les fournisseurs de modèles dès la conception du modèle.

Le niveau de transparence sur la localisation de l'entraînement est faible. La localisation de l'entraînement n'est généralement pas une information communiquée et les informations transmises par les fournisseurs de services d'IA générative sur les émissions de gaz à effet de serre ne permettent pas d'en déduire le lieu où a été réalisé l'entraînement et peuvent s'avérer trompeuses. En effet, les émissions de gaz à effet de serre liées à l'entraînement des modèles (pour les émissions indirectes liées à la consommation d'énergie, représentant « le scope 2 ») peuvent être communiquées de différentes manières et les acteurs privilégient l'approche dite *market-based* (approche économique) qui minimise leur impact au détriment d'une approche dite *location-based* (approche physique) qui reflète les flux physiques réels[85]. L'approche *market-based* permet de rendre compte des efforts réalisés par les entreprises lorsqu'elles achètent des contrats d'énergie renouvelable[86]. Cette méthodologie permet de déclarer des émissions de gaz à effet de serre nulles pour une partie de leur consommation d'électricité équivalente à la quantité d'énergie renouvelable couverte par ces contrats. Le GHG Protocol recommande d'utiliser ces deux méthodologies en parallèle afin d'établir une documentation et une évaluation les plus globales et précises possibles.

L'évaluation des émissions de gaz à effet de serre de l'entraînement reste ainsi conditionnée au niveau de transparence des acteurs et aux modalités de *reporting*, les approches de type *market-based* utilisées pouvant ne pas refléter les émissions effectivement générées par l'entraînement.

c) Matériel utilisé : la consommation électrique de l'entraînement très dépendante des caractéristiques des puces

L'entraînement des modèles d'IA générative peut être réalisé à partir de différents matériels, et en particulier de modèles de puces, qui vont avoir un impact sur la consommation électrique nécessaire à l'entraînement. Les progrès technologiques ont considérablement fait évoluer les caractéristiques des puces sur les dernières années. Ces puces ont vu leur puissance électrique augmenter[87] mais ont enregistré dans le même temps d'importants gains d'efficacité en matière de capacité de calcul. Selon une analyse d'Epoch, la puissance de calcul totale installée de NVIDIA double chaque année, comme l'illustre le graphique ci-dessous. Face à de tels gains, les puces utilisées ont donc une influence forte sur la consommation d'électricité requise pour entraîner le modèle, toutes choses égales par ailleurs.

[85] The Guardian, Data center emissions probably 662% higher than big tech claims. Can it keep up the ruse?, 2024

[86] Ces contrats peuvent désigner : 1/ les certificats d'énergie renouvelables, qui sont des options d'achat d'électricité bas carbone pouvant être utilisées jusqu'à un an après la date d'achat, partout dans le monde ; 2/ les contrats d'achat direct d'énergies renouvelables (dits PPA pour Power Purchase Agreements) qui peuvent être physiques impliquant la livraison physique d'une quantité d'électricité (passant par le réseau d'électricité public ou pas) ou des PPA virtuels, qui sont des contrats financiers sans livraison physique d'électricité.

[87] La puissance électrique d'une puce H100 de NVIDIA est de 700 W (soit la puissance d'un micro-onde) contre 400 W pour une puce A100 de NVIDIA. Les nouvelles puces peuvent aller jusqu'à 1 400 W en 2026 comme mentionné dans un article de Gauthier Roussilhe. Par comparaison, la puissance électrique moyenne en charge d'un CPU grand public varie entre 75 et 150 W mais reste relativement stable selon le chercheur. Un graphique retrace l'historique et l'évolution potentielle de cette puissance pour les puces.

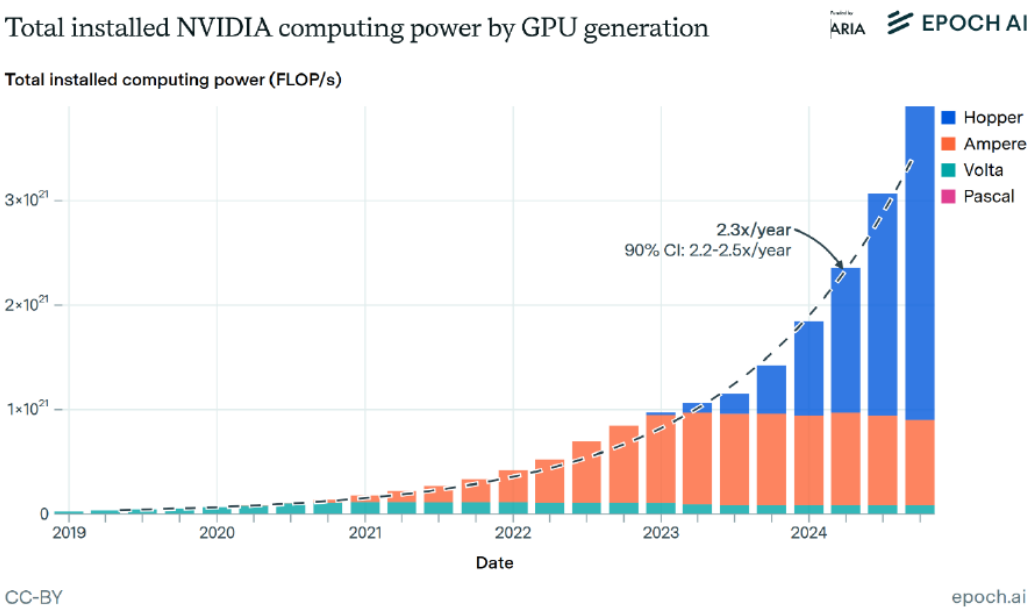


Figure 11 - Capacité installée de calcul NVIDIA par génération de processeurs graphiques NVIDIA. Source : Luke Emberson et David Owen (2025). Publié sur le site de epoch.ai

Il convient toutefois de noter que les progrès technologiques réalisés, en accroissant considérablement les capacités de calcul, permettent d'entraîner des modèles de plus en plus grands et alimentent l'inflation du nombre de paramètres. Ainsi, la puissance requise pour entraîner les modèles d'IA continue de doubler chaque année, selon une autre étude d'Epoch AI, comme le montre la figure ci-dessous. Cette dynamique illustre l'effet rebond : les gains d'efficacité des puces se traduisent pour le moment par une consommation énergétique globale croissante.

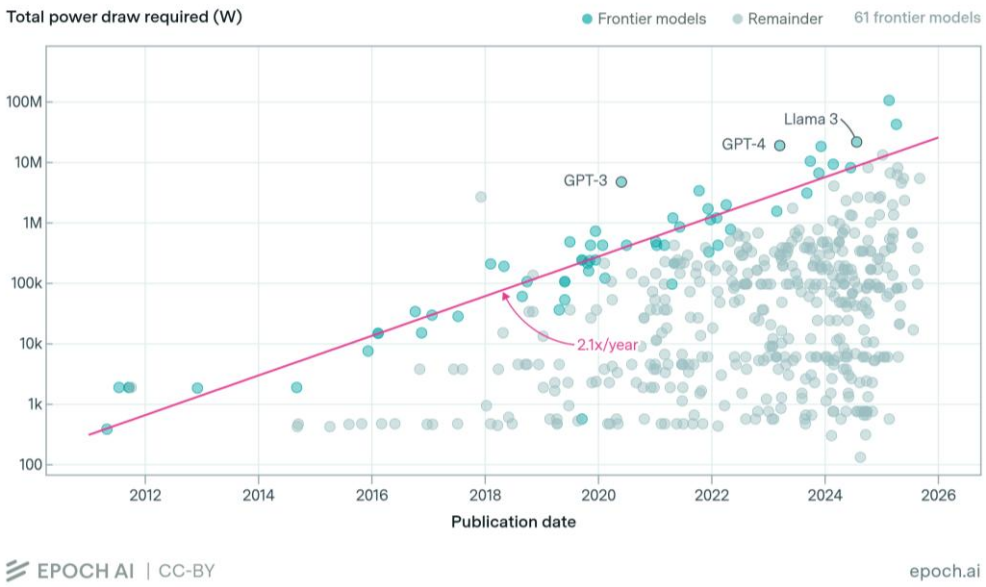


Figure 12 - Puissance de calcul requise estimée pour entraîner les modèles d'IA. Source : Epoch AI

Enfin, le nombre de puces utilisées pour l'entraînement des modèles d'IA croît. Or, au-delà de leur utilisation, la fabrication des puces a un impact carbone embarqué et nécessite une consommation de métaux non négligeable. Falk *et al*. (2025b) ont par exemple estimé que l'entraînement de GPT-4 aurait

nécessité entre 1 174 et 8 800 cartes graphiques GPU A100, soit entre 176 tCO_2e et 1 320 tCO_2e. L'étude estime que l'impact carbone embarqué représenterait près de 13 % de l'impact carbone total associé à l'entraînement.

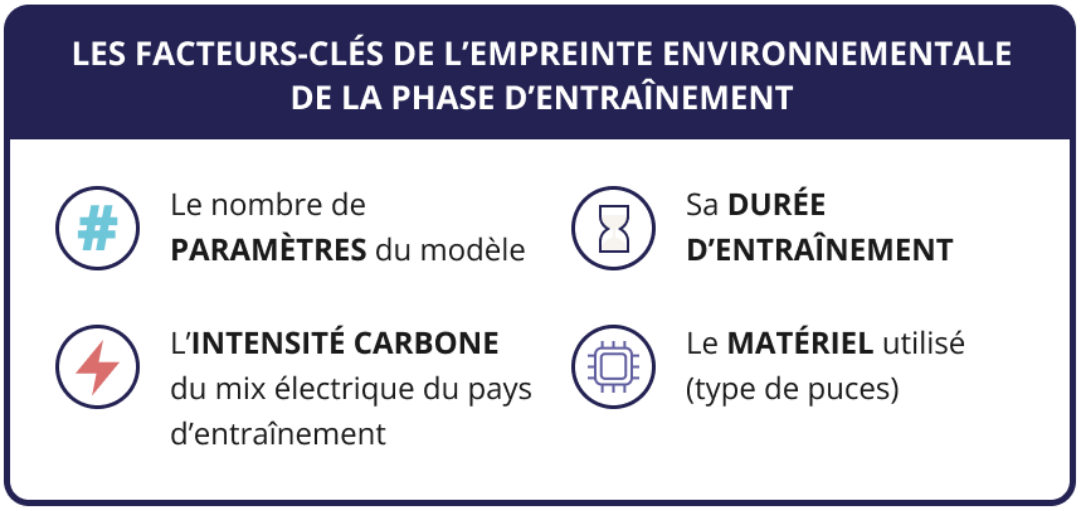


Figure 13 : Les facteurs-clés de l'empreinte environnementale de la phase d'entraînement d'un modèle d'IA générative. Source : Arcep.

2.2.2 Phase d'inférence : au-delà de l'impact par prompt dépendant du nombre de paramètres, l'impact global en forte croissance du fait de la généralisation rapide de l'IA générative

Concernant les impacts environnementaux de la phase d'usage dite phase d'inférence, les estimations sont moins nombreuses que pour la phase d'entraînement. Il existe toutefois des travaux qui permettent d'identifier les facteurs-clés de l'impact environnemental de la phase d'inférence. Ces travaux mettent en particulier en évidence un impact de l'inférence qui reste limité lorsqu'il est considéré par prompt, mais qui pourrait au global devenir un enjeu significatif face à la croissance des usages (a), ainsi que des impacts très variables selon le type de contenus générés (b). Afin d'étudier l'impact d'autres facteurs peu observés jusqu'à présent dans la littérature, des travaux complémentaires ont été réalisés par le PEReN pour l'Arcep et permettent d'enrichir l'analyse (c).

a) Un impact par prompt limité mais un enjeu relatif au volume de prompts

Plusieurs travaux académiques fournissent des estimations de la consommation énergétique d'un prompt (Elsworth *et al.*, 2025). De Vries (2023) estime qu'un prompt sur ChatGPT-3.5 consommerait environ 3 Wh et You (2025) qu'un prompt sur ChatGPT-4o consommerait environ 0,3 Wh. Rincé et Banse (2025) estiment quant à eux qu'une réponse de 50 *tokens*[88] consommerait entre 1,83 Wh et 6,95 Wh selon le modèle testé. Ces chiffres ne proviennent que d'estimations et non de mesures *in situ.*

[88] Un *token* (ou jeton en français) est une unité de données traitées par les modèles d'IA générative afin de permettre la prédiction, la génération et le raisonnement le cas échéant. Pour plus de précision, voir la définition en glossaire.

D'autres travaux scientifiques ont permis de mesurer la consommation d'énergie d'un prompt à partir de tests *in situ.* Luccioni *et al.* (2022), en utilisant l'outil CodeCarbon pour le modèle BLOOM ont évalué que la consommation d'énergie s'élève à 4 Wh/prompt (sur près de 230 000 prompts évalués) et les émissions de gaz à effet de serre sont de 1,5 gCO2e/prompt. Samsi *et al*. (2023) ont quant à eux réalisé des tests des modèles LLaMA (7B, 13B et 65B) sur un supercalculateur du MIT utilisant des GPU NVIDIA V100 et A100 et estimé que le modèle LLaMA-65B consommait 0,3 Wh/prompt.

Si les impacts unitaires (par exemple, par prompt) estimés apparaissent ainsi relativement faibles, ils doivent être mis en regard des impacts environnementaux absolus (comme les émissions de gaz à effet de serre directes et indirectes sur une période d'un an par exemple) et des volumes de prompts. En juillet 2025, OpenAI revendiquait 912,5 milliards de requêtes chaque année, soit une moyenne estimée à 2,5 milliards d'échanges journaliers[89].

D'autres estimations et ordres de grandeur sur l'impact de requêtes sont souvent relayés dans les médias mais la validité des données utilisées pour réaliser ces estimations est contestée (Vanderbauwhede *et al.*, 2024) et ces résultats sont difficilement contre-expertisables en raison de méthodologies parfois peu transparentes ou du manque de comparabilité des données[90]. Par exemple, Open AI avait affirmé en 2020 qu'une requête GPT-3 équivalait à 4 Wh (OpenAI, 2020), et à l'été 2025 qu'un prompt moyen sur ChatGPT qui équivaut à une requête GPT-4 consommait 0,34 Wh[91]. Mistral et Google ont quant à eux publié les résultats de leur propre évaluation de l'impact d'une requête en phase d'inférence. Mistral (2025) a ainsi évalué l'impact de l'utilisation de l'agent conversationnel Le Chat pour une réponse de 400 tokens[92] (environ une page de texte) à 1,14 gCO2e émis, 45 mL d'eau consommés et 0,16 mg de ressources abiotiques consommées. Google (2025) a évalué l'impact d'une requête médiane textuelle Gemini à 0,24 Wh d'énergie, émet 0,03 gCO2e et consomme 0,26 mL d'eau. Si ces dernières estimations sont davantage documentées, le manque de transparence (absence de méthodologie détaillée ou des données primaires utilisées) ou l'emploi de certaines hypothèses (comme l'emploi de l'approche *market-based* pour les émissions de gaz à effet de serre liées à la consommation d'électricité en phase d'utilisation) en limitent la portée.

De manière générale, l'absence de référentiel et de méthodologie harmonisée pose un problème de comparabilité au regard des communications des entreprises qui ne s'appuient pas sur les mêmes méthodes d'évaluation. Cela entraîne également une difficulté à évaluer l'impact environnemental lié à la substitution d'un usage numérique non-IA (par exemple une requête sur un moteur de recherche) par un usage numérique d'IA générative (cf. encadré).

[89] Les Numériques, Juillet 2025, « ChatGPT explose absolument tous les compteurs : voici combien de requêtes sont traitées chaque jour par l'IA ». En considérant qu'un prompt consommerait entre 0,34 Wh (selon l'hypothèse d'OpenAI pour GPT-4) et 4 Wh (selon l'hypothèse d'OpenAI pour GPT-3 et des ordres de grandeur fournis par d'autres articles scientifiques), la consommation énergétique totale des requêtes en 2025 se situerait entre 310 GWh et 3650 GWh, soit des émissions de GES comprises entre 138 000 tCO2e et 1 600 000 tCO2e (en considérant une intensité carbone moyenne du mix électrique mondiale de 445 gCO2e/kWh qui est celle de 2024 renseignée par l'AIE).

[90] Pour plus de transparence et de recul sur la méthodologie utilisée et les données mobilisées, il serait nécessaire d'avoir une contre-expertise ou de mener une revue critique menée par un panel expert comme pour toute ACV.

[91] Sam Altman, 2025, https://blog.samaltman.com/the-gentle-singularity

[92] Les tokens (jetons) sont des unités de données traitées par les modèles d'IA pendant l'entraînement et l'inférence afin de permettre la prédiction, la génération et le raisonnement.

Peut-on comparer l'impact environnemental d'une requête sur un moteur de recherche et d'une requête sur un service d'IA générative ?

Plusieurs études sur les usages numériques montrent une tendance à la substitution des moteurs de recherche par l'IA générative (*via* des *chatbots*) : selon une étude Ipsos[93] de 2024, 48 % du panel de Français interrogés indiquent utiliser des services d'IA générative pour effectuer des recherches sur le web ; selon le Baromètre du numérique publié en 2026[94], la recherche d'informations constitue l'usage le plus répandu des utilisateurs de l'IA générative : 73 % des utilisateurs effectuent ce type de requêtes au moins une fois par mois, 21 % quotidiennement. L'Arcep a bien identifié cette tendance dans son rapport de janvier 2026 sur les enjeux entre IA et internet ouvert[95].

Face à cette tendance, la question de l'impact environnemental de cette substitution des usages se pose et, partant, celle de la comparaison des impacts entre recours à un moteur de recherche ou un *chatbot* d'IA générative. Toutefois, le manque de transparence, à la fois des moteurs de recherche et des modèles d'IA générative rend difficile une comparaison de l'impact environnemental de ces deux usages. Vanderbauwhede *et al*. (2024), s'appuyant sur diverses estimations, concluent que, selon les hypothèses, une requête ChatGPT serait entre 50 et 90 fois plus énergivore qu'une requête sur le moteur de recherche Google. Il pourrait ainsi exister un risque d'augmentation de la consommation énergétique si l'IA générative est généralisée pour des tâches courantes, comme en substitution des moteurs de recherche.

Au contraire, d'autres spécialistes affirment que les *chatbots* pourraient être plus précis dès la première requête par rapport aux moteurs de recherche. Duprat (2026)[96] avance de son côté que la comparaison peut être favorable aux IA génératives si l'on ne compare plus uniquement une requête *via* un moteur de recherche, dont le modèle économique s'appuie sur la publicité (consommant à cause du système d'enchères publicitaires entre 15 et 44% de l'énergie totale associée à la requête sans apporter aucune information pour l'utilisateur), à une requête *via* une IA générative, dans le cas où on s'intéresse au service rendu. En effet, pour l'auteur, il est nécessaire de prendre en compte l'énergie totale permettant de fournir une information, au lieu de comparer une seule requête *via* un moteur de recherche à une seule requête *via* une IA générative. Plusieurs requêtes *via* un moteur de recherche peuvent être nécessaires pour obtenir le résultat que peut obtenir un service d'IA générative en une seule fois.

En outre, plutôt que de constater une substitution, il est davantage constaté que les deux fonctionnalités (IA générative et moteur de recherche) ont tendance à s'additionner. En effet, les principaux moteurs de recherche intègrent maintenant une synthèse des réponses fournies par une IA générative en complément d'une liste de sites internet, complexifiant la capacité à séparer et comparer l'impact environnemental associé au total des usages d'IA générative (*via* des *chatbots*) et à celui associé au total des usages de moteurs de recherche.

Ainsi, il n'existe pas à l'heure actuelle de consensus sur l'impact environnemental comparé des deux services, face aux difficultés d'évaluation et à la diversité des cas d'usage. Toutefois, ces difficultés d'évaluation n'empêchent pas d'identifier certaines bonnes pratiques qui peuvent guider les usages des utilisateurs pour limiter leurs impacts, tels que limiter le recours à des IA génératives pour des recherches sur des questions simples ou privilégier des IA frugales pour des questions plus

[93] Ipsos & CESI, 2025. Intelligence artificielle : quels sont les usages des Français ?

[94] Baromètre du numérique, 2026 (lien).

[95] Arcep, 2026, « IA générative : des défis pour l'avenir de l'internet ouvert ».

[96] Charles Duprat, 2026, The thermodynamic Efficiency Inversion

sophistiquées[97]. Avec des prompts plus précis, qui réduisent la durée d'inférence, les utilisateurs pourraient également réaliser jusqu'à 50 % de gains sur la consommation énergétique de leur prompt (Unesco, 2025).

b) Un impact qui dépend fortement du type de tâches réalisées

En dépit des difficultés d'évaluation et de comparaison, les travaux scientifiques s'accordent sur le fait que l'impact des usages en phase d'inférence dépend fortement du type de tâches.

D'après les estimations disponibles, la génération d'image serait en moyenne 60 fois plus consommatrice en énergie que la génération de texte (2,9 kWh pour la génération d'image contre 0,047 kWh pour la génération de texte). En outre, les modèles généralistes seraient en moyenne plus consommateurs que les modèles spécifiques (Luccioni *et al.*, 2023). Par ailleurs, générer du texte serait en moyenne 25 fois plus énergivore que classifier des textes[98] (Luccioni *et al.*, 2023). D'après les premières estimations de l'Agence internationale de l'énergie, générer une vidéo de 6 secondes nécessiterait environ 115 Wh (AIE, 2025).

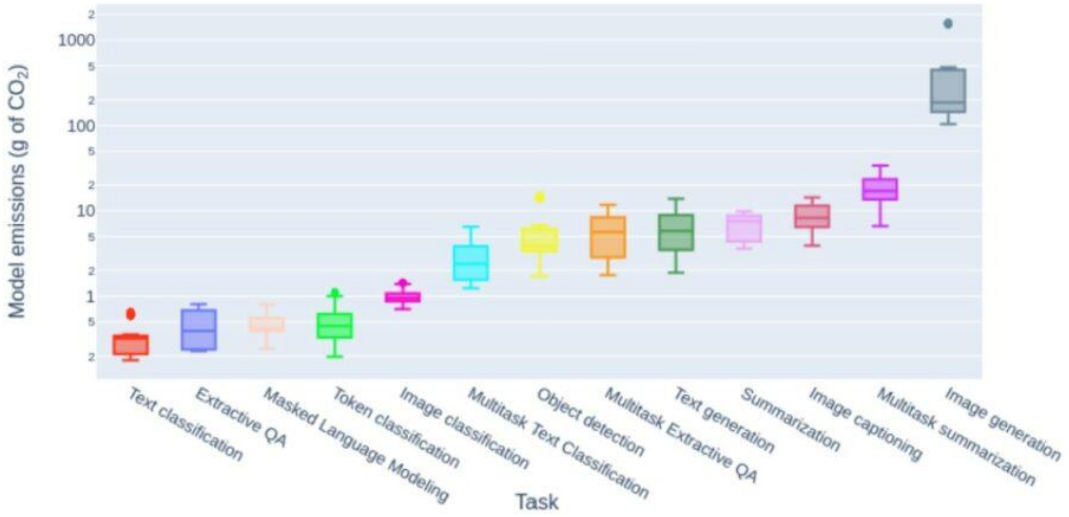


Figure 14 : Émissions de CO2e en fonction des différentes tâches réalisées. Source : Luccioni et al., 2023

c) L'étude inédite réalisée par le PEReN : un approfondissement des facteurs-clés de la consommation énergétique de modèles d'IA générative en phase d'inférence

Pour approfondir la connaissance sur l'impact environnemental de la phase d'inférence et disposer de davantage de données fiables, l'Arcep a proposé au PEReN de mener une étude *in situ* dont l'objectif est de mesurer la consommation d'énergie de modèles d'IA générative en phase d'inférence et d'étudier le rapport entre consommation d'énergie et performance des modèles (au sens de la capacité du modèle à fournir une réponse attendue) pour des cas d'usage donnés. À travers

[97] Dans un article récent, Le Monde listait notamment sept bonnes pratiques pour limiter l'empreinte carbone liées aux usages d'IA (lien).

[98] La classification de texte est une tâche de *machine-learning* qui consiste à attribuer des étiquettes prédéfinies à des données textuelles afin de les classer automatiquement en groupes.

l'évaluation d'une vingtaine de modèles dans un environnement contrôlé, cette étude inédite visait ainsi à identifier les facteurs clés explicatifs de leur consommation énergétique en phase d'inférence.

Le protocole expérimental de l'étude réalisée avec le PEReN

Les travaux du PEReN ont été menés sur le supercalculateur Jean Zay[99] en suivant un protocole détaillé dans le rapport complet de l'étude publié séparément en annexe[100]. Pour cette étude, des cas d'usage ont été déterminés à partir de l'état de l'art scientifique sur les thématiques d'usage grand public (usages du quotidien, aide au code, usages professionnels et usages créatifs) et à partir des données de Compar:IA[101]. Une fois les thématiques identifiées, des jeux d'évaluation[102] (*benchmarks)* ont été sélectionnés pour tester la performance des modèles sur ces thématiques, c'est-à-dire leur capacité à fournir une réponse attendue.

Les 22 modèles d'IA générative sélectionnés sont des modèles à poids ouverts, dits « *open-weights »,* d'une taille comprise entre 3 et 123 milliards de paramètres, disponibles sur la plateforme de référence Hugging Face pour l'hébergement des modèles d'IA et publiés entre juin 2024 et septembre 2025. Ce choix pour des modèles à poids ouverts a permis au PEReN de mesurer directement la consommation énergétique des modèles en phase d'inférence sur une même infrastructure connue, et ainsi d'obtenir des résultats exploitables pour comprendre les facteurs-clés de la consommation énergétique en phase d'inférence. Ce choix méthodologique ne permet toutefois pas d'intégrer dans les tests ni les très gros modèles (plus de 125 milliards de paramètres) ni les modèles les plus utilisés du grand public (ex : ChatGPT), et a également limité le nombre de modèles spécialisés pouvant être étudiés.

La consommation énergétique de l'inférence des modèles a été mesurée avec l'outil CEEMS[103] intégré au sein du supercalculateur Jean Zay. Seule la consommation énergétique des GPU est présentée dans les résultats, en raison d'une forte variabilité de la consommation électrique du CPU pour une même tâche réalisée[104].

Il ressort de cette étude que, à usage et matériel constant, le nombre de paramètres d'un modèle est le premier facteur explicatif de sa consommation électrique en phase d'inférence, mais il n'est pas le seul. L'architecture du modèle et l'activation ou non d'un mode « raisonnement » ont également un impact significatif sur la consommation électrique de la phase d'inférence.

[99] Ces travaux ont bénéficié de ressources de calcul en IA et de stockage au IDRIS au travers de l'allocation de ressources 2025-AD011016936 attribuée par GENCI sur les partitions A100 /H100 du calculateur Jean Zay.

[100] Ce test se déroule à utilisation du matériel constante, sur un seul type de machine. Le choix du GPU a une influence significative sur la consommation énergétique, pour un modèle donné. Chaque GPU possède son propre profil de consommation énergétique pour un même modèle.

[101] Compar:IA est un comparateur de modèles d'IA générative permettant de créer des jeux de données de préférence centrés sur des usages réels exprimés dans les langues européennes à partir de thématiques comme les biais culturels et linguistiques, l'impact environnemental, le pluralisme des modèles ou l'esprit critique et les questions sociétales.

[102] Un jeu d'évaluation (aussi appelés benchmarks) est (i) un ensemble de questions sur la thématique donnée (ii) avec les réponses attendues et (iii) une méthode pour évaluer le score de qualité (qu'on appellera performance dans la suite de la présentation) entre la réponse donnée par le modèle testé et la réponse attendue.

[103] L'outil est librement accessible sur Github en cliquant ici.

[104] Il est à noter qu'exclure la consommation énergétique des CPU du périmètre réduit d'environ un tiers la consommation d'énergie totale du modèle lors de l'inférence. Le PEReN a considéré que la variabilité de la consommation énergétique des CPU pour une même tâche était trop importante (pour deux mesures, l'écart entre la consommation d'énergie des CPU est très variable alors que les conditions sont les mêmes) pour prendre en compte les CPU pour cette expérience.

Les tests réalisés révèlent ainsi que les plus grands modèles sont toujours les plus consommateurs d'énergie mais il ressort également des mesures effectuées que des modèles de tailles très différentes peuvent générer la même consommation énergétique ; certains gros modèles consomment en effet autant voire moins d'énergie que des modèles ayant un nombre de paramètres total plus faible, en raison de l'influence d'autres déterminants explicités ci-après.

De plus, limiter la consommation énergétique ne revient pas nécessairement à faire des compromis sur la performance. En effet, des modèles à faible consommation énergétique peuvent avoir des performances équivalentes aux modèles qui ont une plus forte consommation énergétique.

Les résultats des travaux du PEReN montrent que certaines architectures de modèles développées ces dernières années[105], comme le « mélange d'experts »[106], permettent de réduire la consommation énergétique des grands modèles. Les modèles présentant cette architecture consomment en moyenne 45 % de moins que les modèles denses ayant un nombre de paramètres équivalent. Il ressort également des tests que la mise en œuvre de techniques de compressions des modèles (quantification)[107] permet de réduire significativement la consommation électrique en phase de d'inférence. L'effet de la quantification (8-bit ou 4-bit) sur la consommation énergétique est significatif, avec un gain moyen de 39 %. L'impact du raisonnement sur la consommation peut énormément varier selon le type de tâche demandée, allant jusqu'à +849% sur une tâche de génération de code.

Enfin, l'activation ou non d'un mode « raisonnement »[108] influe également significativement sur la consommation énergétique des modèles en phase d'inférence en raison de l'incidence de cette fonction sur la génération de *tokens*. Les modèles consomment davantage lorsqu'ils raisonnent (jusqu'à +92 % en moyenne pour les modèles testés quand le mode « raisonnement » est activé), pour un gain de performance inégal selon les tâches.

En conclusion des travaux menés avec le PEReN, les modèles les plus performants ne sont pas nécessairement ceux qui sont les plus énergivores, et les modèles les plus énergivores ne sont pas nécessairement les plus performants : limiter l'empreinte énergétique n'implique pas toujours des compromis sur la performance.

Enfin, si les tests réalisés par le PEReN ont été réalisés à matériel constant pour permettre de mettre en évidence l'impact de différents facteurs, le choix de l'infrastructure de calcul pour réaliser l'inférence constitue également un paramètre important souligné par le PEReN. Il ressort de la revue de littérature et des travaux empiriques que le type de carte graphique GPU est un facteur explicatif de la consommation énergétique (Chen *et al*., 2024) et que des innovations technologiques sur les puces peuvent avoir un impact significatif sur la réduction de la consommation (Kachris, 2025).

[105] En 2024, peu de modèles denses (non-MoE) ont été publiés. Au contraire, les modèles MoE ont pris leur essor à partir de septembre 2024.

[106] Le mélange d'experts (MoE, ou Mixture of Experts) est une approche qui consiste à diviser un modèle en sous-réseaux distincts (les experts), chacun spécialisé dans un sous-ensemble de données d'entrée, afin d'effectuer conjointement une tâche. Les modèles n'utilisant pas cette approche sont à distinguo appelés « denses ». Lors de l'inférence, un MoE n'utilise que la partie du réseau dont il a besoin, le nombre de ces paramètres dits « activés » est ainsi inférieur au nombre de paramètres totaux du modèle.

[107] Ces techniques consistent à réduire la taille des paramètres (et donc, à réduire la précision de ces paramètres) du modèle sans réduire la performance. Par exemple, il s'agit de compresser le modèle pour passer de 16 bits à 8 bits. Par analogie, il s'agirait en mathématique de réduire le nombre de décimales d'un nombre.

[108] Les modèles de raisonnement sont apparus plus fréquemment sur le marché à partir de septembre 2024.

Des techniques d'optimisation permettant de réduire la consommation énergétique des modèles

Plusieurs progrès en termes de consommation énergétique des modèles ont déjà été permis par des techniques d'optimisation. Parmi celles-ci, le mélange d'experts (MoE pour Mixture of Experts), consiste à diviser un modèle en plusieurs sous-modèles spécialisés qui sont ensuite mobilisés en fonction des requêtes. Il existe des pratiques de compression des modèles comme la quantification (*quantisation*), mise en avant dans un article récent (Delavande *et al*., 2026) ou par des travaux expérimentaux menés par la *University College London* (UCL) pour l'Unesco[109] (2025) qui montrent que cette dernière technique pourrait permettre d'obtenir jusqu'à 44 % de gains en consommation d'énergie et réduire ainsi le coût de fonctionnement des LLM (*Large Language Models).* D'autres techniques comme l'adaptation de rang faible (Low-Rank Adaptation ou LoRA), la distillation ou l'élagage (*pruning*), visent à réduire le nombre de poids ou à les simplifier afin de diminuer les ressources matérielles nécessaires au déploiement du modèle. En outre, en permettant aux modèles de mobiliser des données absentes lors de l'entraînement, le RAG (*Retrieval-Augmented Generation*) a aussi permis d'éviter des entraînements de bout en bout. Il est à noter que si les approches par optimisation sont nécessaires, elles pourraient s'avérer insuffisantes en cas d'effet rebond dû à la croissance des usages.

Il est à noter que l'intensité carbone du mix électrique des serveurs utilisés pour l'inférence peut constituer un paramètre à considérer. Toutefois, contrairement à la phase d'entraînement, la phase d'inférence doit intégrer un paramètre de latence : les marges de manœuvre s'agissant de la localisation des serveurs dédiés à l'inférence sont donc moins importantes que pour la phase d'entraînement.

[109] Les prompts ont été expérimentés sur des GPU NVIDIA GeForce RTX 3090 Ti, en combinaison avec un CPU1 12th Gen Intel Core i9-12900K et 128GB of RAM. Le protocole est détaillé en annexe du rapport.

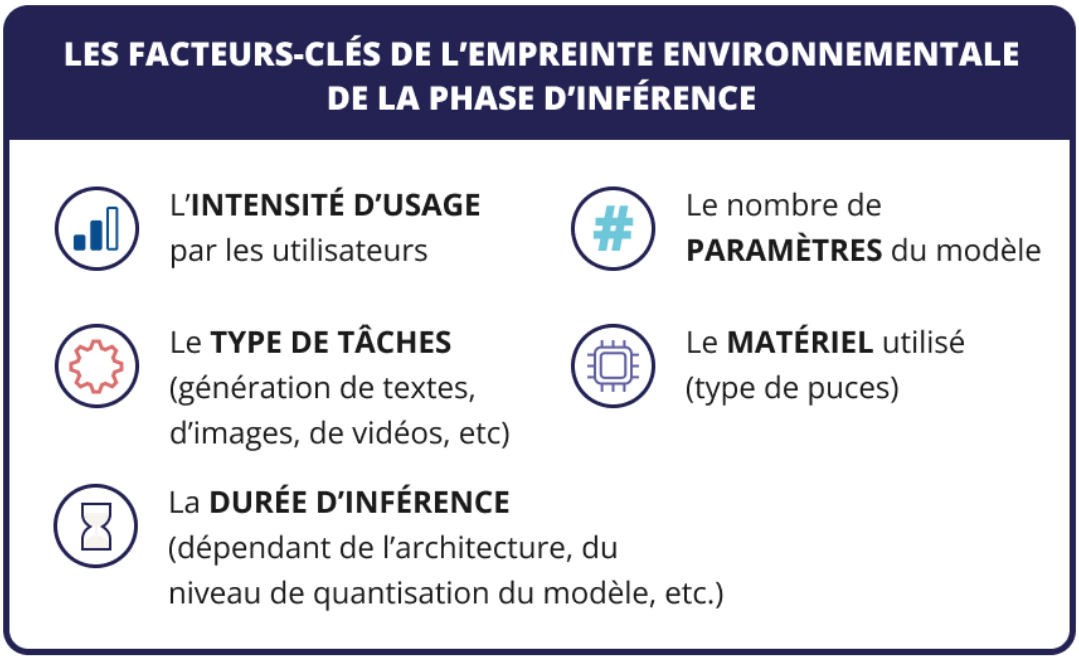


Figure 15 : Les facteurs-clés de l'empreinte environnementale de la phase d'inférence d'un modèle d'IA générative. Source : Arcep.

⁂

En conclusion de l'analyse par phase, il ressort que si l'entrainement des modèles d'IA générative est source d'impacts très significatifs, avec le développement des usages, la phase d'inférence est *a priori* celle qui se déroulera le plus fréquemment, notamment pour les services les plus employés, comme ChatGPT. Il est ainsi possible que la part des émissions de gaz à effet de serre (GES) liées à l'usage puisse finalement dépasser celle liée à l'entraînement au-delà d'un certain nombre de requêtes, en fonction du nombre d'utilisateurs et de la fréquence d'utilisation du service (Data For Good, 2023). Aujourd'hui, il est ainsi difficile d'affirmer qu'une des deux phases, d'entraînement ou d'inférence, aurait un impact plus important que l'autre, car cela dépend du modèle analysé (et du nombre de réentraînements[110]) et du nombre d'utilisateurs (De Vries, 2023).

Au niveau global, pour l'ensemble des services d'IA, les impacts environnementaux liés aux phases d'entraînement et d'inférence dépendront des dynamiques de marché dans les prochaines années (cf. section 1.2). Les dynamiques d'innovation pourraient donner lieu à la poursuite d'une course au développement de nouveaux modèles comprenant toujours plus de paramètres, comme la tendance actuelle, ou au contraire au développement de modèles plus petits (cf. encadré).

D'un point de vue prospectif, de nombreuses questions se posent également sur l'impact environnemental d'innovations telles que le développement des services d'IA agentique[111] qui pourrait démultiplier l'impact environnemental des services d'IA générative en phase d'inférence. Cette

[110] Afin d'obtenir une meilleure précision, un modèle peut être ré-entraîne plusieurs fois (ou *fine-tuned).*

[111] Il s'agit de systèmes d'IA générative intégrant des capacités d'interaction avec d'autres services numériques (moteurs de recherche, calendriers, messageries, outils bureautiques, plateformes sociales ou applications tierces.

nouvelle génération de services combine les capacités conversationnelles des *LLM* avec des fonctions d'automatisation (Arcep, 2026a). De premières études montrent que les IA agentiques auraient un impact 60 fois plus important en matière d'émissions de gaz à effet de serre et de consommation d'eau qu'un « simple » service d'IA générative (Green IT, 2025). Ces résultats sont toutefois à prendre avec précaution compte tenu de la littérature très récente à ce sujet et du manque de transparence de ces systèmes.

« Small is beautiful » : Faut-il privilégier les petits modèles au regard de leur performance et de leur impact environnemental ?

En amont du Sommet pour l'Action sur l'IA à Paris de février 2025, l'Inria a publié un livre blanc dont l'une des recommandations principales est de privilégier des modèles spécialisés aux modèles généralistes, qui possèdent moins de paramètres. Selon l'organisme de recherche, quand une technologie émerge, il y a une tendance naturelle à développer des outils généralistes. Or, les modèles spécifiques pourraient être capables de réconcilier protection de l'environnement et innovation (Inria, 2025 ; Varoquaux, 2025).

Des travaux montrent que le volume de données utilisées pour entraîner un modèle, et donc le nombre de paramètres, n'est pas nécessairement corrélé à la performance d'un modèle (au sens de la capacité du modèle à fournir une réponse attendue). Des solutions d'IA sophistiquées conçues pour résoudre des problèmes spécifiques peuvent parfois être préférables du point de vue de la performance et des moindres ressources mobilisées, à des modèles à usage général qui soient *fine-tunés*. Varoquaux *et al*. (2025) ont mis en évidence que dans différents cas d'usage, augmenter le nombre de paramètres du modèle, la durée d'entraînement ou les ressources GPU mobilisées pour l'entraînement ne revient pas systématiquement à améliorer la performance (cf. annexe 7).

Les petits modèles, appelés SLM *(Small Language Models),* peuvent être également très performants : l'adoption de ces modèles pour des applications spécifiques pourrait permettre de réaliser jusqu'à 90 % d'économies d'énergie par rapport à des LLM tout en conservant une grande précision (Unesco, 2025). Outre leur efficacité énergétique, ils sont plus accessibles dans les environnements à faibles ressources et à connectivité limitée et avec des temps de réponses plus rapides.

À titre d'illustration, des chercheurs de Nvidia et du Georgia Institute of Technology considèrent que les SLM sont suffisamment puissants pour des systèmes d'IA agentique de sorte qu'il y aurait un intérêt environnemental (économie de ressources) à passer des LLM aux SLM pour ce type de systèmes (Belcak *et al*., 2025). Les investissements massifs dans des centres de données puissants, où sont réalisés les calculs pour l'entraînement et l'inférence des LLM, constitueraient toutefois les principales barrières à l'adoption de SLM par la création d'un phénomène de verrouillage technico-économique, incitant à privilégier les LLM pour rentabiliser ces investissements. Belcak et *al*. (2025) pointent également le manque de publicité faite aux SLM comparativement aux LLM.

Il n'en demeure pas moins que le déploiement des petits modèles peut présenter des limites. D'une part, de nombreux modèles dérivent des LLM[112], ce qui en réduit l'intérêt environnemental par rapport à des gammes de modèles de plus petite taille qui ne dérivent pas des LLM[113] et qui ont donc été entraînés sur un plus petit nombre de paramètres. D'autre part, selon des experts interrogés, le déploiement plus large de SLM conduirait à une superposition des usages de SLM aux usages des LLM existants, risquant de minimiser les gains environnementaux permis par les SLM.

[112] C'est le cas pour les modèles comme BERT Mini, Llama3.2-1B, Qwen2.5-1.5B ou DeepSeek-R1-1.5B.

[113] C'est le cas pour les gammes Gemma de Google, Ministral de Mistral AI ou Phi de Microsoft.

Chapitre 3 Nos recommandations afin de rendre le développement de l'IA compatible avec les limites planétaires

À mesure que le numérique et ses infrastructures deviennent toujours plus essentiels à l'économie et à la vie quotidienne, il apparaît nécessaire d'en garantir la soutenabilité. En ce qui concerne l'IA générative, si les données sur son impact environnemental restent parcellaires, le constat actuel et les tendances anticipées appellent d'ores et déjà à agir afin de s'assurer que cette technologie se développe et se déploie dans des conditions soutenables pour les générations futures.

Une meilleure compréhension des enjeux autour des dynamiques de marché de l'IA générative ainsi que de son empreinte environnementale permet d'identifier des leviers d'action pertinents. L'Arcep propose dans ce dernier chapitre de développer quatre axes de travail afin de rendre le développement de l'IA compatible avec les limites planétaires.

Ces propositions intègrent la mobilisation d'outils de droit souple et règlementaires, se fondant sur le cadre législatif européen en vigueur ou à venir. Ces propositions visent à alimenter le débat public et les futurs travaux de régulateurs ou de parties prenantes sur l'IA générative.

Axe 1 : Améliorer la mesure et la connaissance de l'impact environnemental de l'IA

Comme souligné à plusieurs reprises dans ce rapport, il existe aujourd'hui un déficit évident d'informations permettant d'évaluer pleinement l'impact environnemental de l'IA, en particulier faute de transparence des acteurs à tous les niveaux de la chaîne de valeur. Si des travaux académiques ou institutionnels permettent d'apporter des éclairages précieux, l'identification, la priorisation et le pilotage des actions à mener passent par une meilleure mesure des impacts. Pour établir un diagnostic robuste et évaluer les impacts à partir de méthodologies standardisées, il est ainsi nécessaire de disposer de données fiables, obtenues auprès des acteurs de toute la chaîne de valeur.

Recommandation 1. Mettre en œuvre la collecte et la publication des données environnementales de l'IA par des autorités publiques

Face à l'accélération rapide des usages de l'intelligence artificielle, l'Union européenne s'est engagée dans une dynamique d'investissement ambitieuse en faveur du développement de l'IA. Toutefois, l'impact environnemental de ces technologies demeure encore insuffisamment documenté (cf. Chapitre 2) alors même que cette connaissance apparaît indispensable pour concilier les objectifs de compétitivité et de soutenabilité d'une intelligence artificielle dont les générations futures pourront bénéficier.

Dans ce contexte, la capacité des pouvoirs publics à adopter des décisions éclairées dépend directement de la disponibilité de données fiables et harmonisées. A cet égard, la régulation par la donnée apparaît comme un levier essentiel aux ambitions européennes : sans dispositif de mesure robuste, il devient difficile de traduire concrètement et d'évaluer les ambitions liées au développement d'une IA soutenable.

Dès lors, l'Union européenne doit se doter des moyens de produire une connaissance objectivée de l'empreinte environnementale de l'IA, en confiant un mandat de collecte de données environnementales aux autorités de régulations nationales en charge des communications électroniques sur l'ensemble de la chaîne de valeur de l'IA. À cet égard, certaines dispositions envisagées dans le cadre du *Digital Networks Act (DNA)*, notamment l'article 115 (1) (c) qui prévoit une mission de collecte de données auprès des acteurs du numérique, y compris en matière de durabilité,

constituent une évolution intéressante et ouvrent des perspectives pour commencer à mettre en place une régulation par la donnée, efficace à l'échelle européenne.

Pour répondre à ce besoin, il est en effet nécessaire de dépasser les dispositifs actuels, qui restent insuffisants pour structurer une collecte de données exploitable. Certaines dispositions de l'*AI Act* imposent des obligations de documentation pour les fournisseurs d'IA et de transparence pour les acteurs (cf. encadré sur l'*AI Act* ci-dessous, notamment ses articles 53 (1) (a) et 95). Toutefois, ces dispositions ne constituent pas un cadre de collecte de données environnementales à proprement parler : elles reposent largement sur la capacité et la volonté des acteurs à produire ces données, sans garantir leur disponibilité, leur fiabilité et leur homogénéité à l'échelle européenne. En particulier, elles laissent en suspens la question de l'accès à des données clés, telles que la consommation énergétique liée à l'entraînement des modèles d'IA ou à l'inférence qui constitue un angle mort du règlement, lorsque ces informations ne sont pas systématiquement mesurées ou estimées par les fournisseurs eux-mêmes.

L'expérience récente de la régulation des centres de données illustre clairement les limites d'une telle approche. Face à un déficit de données fiables, sur leur consommation énergétique et leur empreinte environnementale, l'Union européenne a introduit, dans la révision de la directive efficacité énergétique (DEE), une obligation explicite de *reporting* harmonisé au niveau européen (cf. encadré sur la DEE dans l'Axe 4). Ce dispositif impose la collecte de données précises selon des formats standardisés, afin d'alimenter une base de connaissances robustes au service de l'action publique. Ce cadre juridique clair permettra de réduire les asymétries informationnelles mais il ne concerne que les centres de données, qui ne constituent qu'un maillon de la chaîne de valeur de l'IA. D'autres segments essentiels restent aujourd'hui largement hors du champ de toute collecte structurée, dont les fabricants de semi-conducteurs et les fournisseurs de modèles d'IA. L'absence de données harmonisées sur ces segments empêche toute appréhension globale de l'empreinte environnementale de l'IA. Dans ce contexte, l'attribution d'un mandat explicite de collecte de données environnementales pour les autorités de régulation apparaît comme une solution pertinente pour améliorer la mesure des impacts environnementaux de l'IA le long de toute la chaîne de valeur.

La mise en œuvre d'un tel mandat doit cependant s'inscrire dans une logique de proportionnalité et d'efficacité administrative. Conformément au principe du « dites-le nous une fois » (« *once-only* »), il convient d'éviter la multiplication des obligations déclaratives pesant sur les acteurs économiques et de privilégier l'exploitation des cadres existants. À cet égard, certaines évolutions envisagées dans le *DNA* peuvent constituer des points d'appui intéressants. Outre la mention d'une mission de collecte pour les autorités de régulation, le DNA prévoit l'élaboration de lignes directrices au niveau européen par le BEREC (ORECE en français) sur les procédures et demandes d'informations. Aller au-delà de ces mécanismes et confier aux autorités de régulation un rôle clair et un périmètre explicite dans la collecte de données sur toute la chaîne de valeur de l'IA (incluant les fournisseurs de modèles) leur permettrait d'agir comme des tiers de confiance, à la fois indépendants et transparents. Une telle évolution permettrait d'améliorer significativement la mesure de l'impact environnemental de l'IA sur l'ensemble de sa chaîne de valeur, répondant ainsi à l'ambition européenne d'un développement compétitif et durable de l'IA.

La Directive Efficacité Énergétique

Cette directive (UE) 2023/1791 sur l'efficacité énergétique vise à réduire la consommation d'énergie et les émissions de gaz à effet de serre des entreprises. Elle fixe un objectif de réduction de la consommation finale d'énergie d'au moins 40 % d'ici à 2030 (par rapport à 2007) en Europe.

La directive a été mise à jour en 2023, et prévoit désormais une obligation pour les centres de données dont la puissance informatique est supérieure à 500 kW de transmettre les données relatives à leur consommation énergétique et à des indicateurs environnementaux requis conformément à l'Article 12 de cette directive. La Commission européenne doit mettre en place une base de données européenne sur les centres de données qui regroupe les informations communiquées par les centres de données.

Les États membres doivent encourager les propriétaires et exploitants de centres de données situés sur leur territoire dont la demande de puissance est égale ou supérieure à 1 MW à tenir compte des bonnes pratiques figurant dans la version la plus récente du code de conduite européen sur l'efficacité des centres de données.

Enfin, la Commission européenne, à partir de ces données disponibles, doit formuler des propositions législatives visant à améliorer l'efficacité énergétique et doit évaluer la faisabilité de la transition vers la « neutralité carbone du secteur des centres de données », pouvant ainsi définir un calendrier dans lequel les centres de données existants seront tenus de « satisfaire à une performance minimale ».

Recommandation 2. Utiliser des méthodologies internationalement standardisées d'évaluation de l'impact environnemental, afin de faciliter les comparaisons entre systèmes d'IA

Une meilleure mesure de l'impact environnemental de l'IA sur l'ensemble de la chaîne de valeur nécessite d'utiliser des méthodologies robustes, fiables (multicritère, multi-composants, multi-étapes) et harmonisées, pour permettre la comparabilité des modèles et des services.

Des méthodologies ACV standardisées reconnues existent déjà et doivent être mobilisées et diffusées, pour éviter un foisonnement des méthodes et ainsi s'appuyer sur une base commune. Elles sont précisées et détaillées dans le Référentiel général pour l'IA frugale, développé par l'AFNOR en partenariat avec le Ministère de la Transition Écologique, avec la participation de l'Arcep, et directement mobilisables par les acteurs pour une évaluation quantitative et qualitative globale des services d'IA**.** Ce référentiel s'appuie notamment sur des articles académiques de référence (AFNOR, 2024 ; Ligozat *et al.*, 2022) ou des recommandations existantes, comme la recommandation UIT-T L.1410.

Au niveau européen, l'article 40 du Règlement IA (voir encadré ci-dessous) pourrait être mobilisé pour aboutir à une méthodologie européenne commune. Dans sa stratégie « Appliquer l'IA » publiée en octobre 2025, la Commission européenne a d'ailleurs annoncé l'adoption prochaine d'une demande de normalisation adressée aux organismes de normalisation concernant les processus communs de déclaration et de documentation relatifs à l'impact des systèmes d'IA et des modèles à usage général sur la consommation d'énergie[114]. C'est une initiative positive à poursuivre qui permettrait d'établir un cadre harmonisé pour la mesure de l'impact environnemental des systèmes d'IA.

Au niveau international, des travaux ont déjà été publiés comme la recommandation UIT-T L.1801 visant à évaluer l'impact environnemental des systèmes d'IA. Cette recommandation permettra d'avoir un cadre cohérent et international pour évaluer l'impact environnemental de l'IA, avec des

[114] Commission européenne, « Apply AI Strategy », COM(2025) 723 final, 08/10/2025, Bruxelles, p.9, Appliquer la stratégie en matière d'IA | Bâtir l'avenir numérique de l'Europe

orientations pour comparer deux systèmes d'IA ou un système d'IA avec un système non-IA et alimentera les réflexions pour l'élaboration d'un score environnemental (UIT, 2025).

En particulier, il apparaît nécessaire que soient articulées les différentes initiatives de normalisation des méthodologies d'évaluation de l'impact environnemental de l'IA, notamment celles de l'UIT, de l'Institut européen des normes de télécommunications (ETSI) en lien avec le Comité européen de normalisation (CEN) et le Comité européen de normalisation électrotechnique (CENELEC), de l'Institut des ingénieurs électriciens et électroniciens (IEEE), de l'Organisation internationale de normalisation (ISO)[115]. À ce sujet, une feuille de route sur la normalisation a été publiée à l'occasion du Sommet mondial pour l'action sur l'IA à Paris. Cette feuille de route sur la normalisation de l'impact environnemental de l'IA a été établie par plus de 30 partenaires privés et publics définissant des lignes directrices pour l'évaluation environnementale de l'IA et identifiant des domaines de coopération entre les principales organisations internationales de normalisation (ISO, UIT, UNESCO et IEEE). La cartographie des différentes initiatives de normalisation des méthodes d'évaluation de l'impact environnemental de l'IA est détaillée en Annexe 6. Plus largement, la coopération internationale demeure indispensable afin de coordonner une action pour une IA plus durable, notamment entre les acteurs historiques du secteur de l'énergie et de l'écosystème du numérique[116].

[115] L'Autorité de la concurrence a également souligné l'importance de la standardisation pour garantir une concurrence entre acteurs sur la base de leurs mérites respectifs. Pour limiter les risques concurrentiels, l'Autorité rappelle cependant l'importance de garantir, dans les instances de normalisation, un équilibre entre les représentants des intérêts privés et des pouvoirs publics. Elle insiste également sur l'importance d'assurer la transparence vis-à-vis des utilisateurs sur le fonctionnement d'un standard et sur l'importance d'une vérification indépendante des informations fournies par les parties prenantes lors de l'élaboration du standard. En effet, certains standards peuvent avoir des limites quand le standard est biaisé au profit de certains acteurs du marché. À l'inverse, la standardisation peut être entravée par le comportement d'acteurs qui ne communiquent pas les informations indispensables à l'élaboration ou la mise en œuvre du standard.

[116] À titre d'illustration, l'Agence internationale de l'énergie (AIE) a également lancé le premier observatoire mondial dédié à l'énergie et à l'IA qui permet, en centralisant et en analysant les données : (i) de mieux anticiper les besoins énergétiques des centres de données et des modèles d'IA, sur la base d'une méthodologie transparente, (ii) d'optimiser les systèmes énergétiques et de réduire les émissions de carbone grâce à l'IA, et (iii) de promouvoir des innovations révolutionnaires pour l'IA appliquée à l'énergie.

Le Règlement IA

Ce règlement (UE) 2024-1689 établit un cadre harmonisé pour le développement, la mise sur le marché et l'utilisation des services d'intelligence artificielle. Il adopte une approche fondée sur les risques et prévoit des obligations notamment en matière de sécurité, transparence, gouvernance des données, documentation technique et supervision humaine. Le texte inclut aussi des dispositions spécifiques pour les modèles dits à usage général (MIAUG), notamment dans le domaine de l'IA générative.

L'article 40, paragraphe 2 prévoit la possibilité pour la Commission européenne d'émettre une demande de normalisation aux organismes dédiés (CEN, CENELEC et/ou ETSI) visant à réduire la consommation d'énergie et d'autres ressources tout au long du cycle de vie des services IA et à améliorer l'efficacité énergétique des modèles à usage général.

L'article 95 stipule que la Commission européenne (*via* le Bureau IA) et les États Membres doivent encourager les acteurs à élaborer des codes de conduite volontaires pour minimiser l'impact environnemental des services IA.

L'article 112 prévoit une évaluation des deux dispositions précédentes : la publication d'un rapport sur le progrès en matière de normalisation d'ici le 2 août 2028 puis tous les quatre ans et la publication d'un rapport sur l'impact et l'efficacité des codes de conduite volontaires d'ici le 2 août 2028 puis tous les trois ans.

L'article 53(1)(a) prévoit que les fournisseurs de modèles d'IA à usage général doivent fournir une documentation technique incluant notamment la consommation d'énergie mesurée ou estimée de leurs modèles, a minima les ressources de calcul mobilisées en cas d'incapacité à fournir ces informations.

Axe 2 : Promouvoir l'écoconception des services d'IA comme levier stratégique de la compétitivité européenne

Les principes d'écoconception numérique sont un levier d'action permettant d'assurer l'intégration des enjeux environnementaux dès la conception de services d'IA et tout au long de leur cycle de vie. S'emparer des meilleures pratiques existantes à ce sujet contribue à minimiser l'impact environnemental des services d'IA, et pourrait devenir un axe de différenciation pour bénéficier d'un avantage compétitif dans la dynamique concurrentielle actuelle. En effet, l'écoconception peut être mise au service d'une plus grande compétitivité européenne, par une meilleure efficacité des ressources utilisées et l'intégration de pratiques de sobriété numérique permettant d'adapter les services aux besoins des utilisateurs finals.

Recommandation 3. Intégrer l'écoconception des services d'IA dans la régulation européenne des fournisseurs de services

L'Arcep a pour ambition d'œuvrer à l'émergence d'initiatives normatives, incitatives ou règlementaires principalement au niveau européen mais aussi au niveau international sur l'écoconception des services numériques.

Au sein de l'Union européenne, si les produits matériels font l'objet d'une règlementation étendue, les services, y compris les services d'IA, n'y sont pas soumis. De premières initiatives de droit souple ont émergé au niveau des États membres pour favoriser l'écoconception des services numériques et d'IA, en particulier le RGESN (Référentiel Général de l'Écoconception des Services Numériques, Arcep & Arcom,2024), le référentiel général pour l'IA frugale (Afnor, 2024) ou encore la récente norme ISO *Digital Services Ecodesign*. Face à l'évolution rapide des services d'IA générative, la mise à jour

régulière de ces référentiels apparaît nécessaire pour tenir compte des évolutions technologiques, en s'appuyant notamment sur les experts du secteur.

Toutefois, les outils non-contraignants comme les référentiels ou les codes de conduite peuvent s'avérer insuffisants. En effet, si les pouvoirs publics peuvent favoriser l'écoconception des services d'IA par ces mécanismes, ces outils peuvent ne pas être suffisamment incitatifs lorsque ce critère de différenciation est peu ou mal pris en compte dans la dynamique concurrentielle, notamment lorsque les acteurs ne sont pas directement concernés par les externalités négatives liées à l'utilisation de leurs produits ou services. Par exemple, la hausse de consommation électrique des terminaux utilisateurs ne constitue pas une source de coûts directs pour le fournisseur de services d'IA qui pourrait l'inciter à une plus grande efficacité de ses services. En revanche, elle peut accélérer la fréquence des cycles de charge et décharge de la batterie du terminal, réduisant ainsi la durée de vie de cette composante.

Il est donc nécessaire d'aller plus loin pour accompagner le développement rapide des services d'IA et ses impacts environnementaux associés. L'Arcep défend ainsi la nécessité de mettre en place au niveau européen un cadre permettant d'intégrer l'écoconception des services d'IA dans la régulation des fournisseurs.

Différentes initiatives en cours de discussion au niveau européen (*Digital Networks Act* et *Digital Fairness Act)* pourraient permettre des avancées en ce sens. L'Arcep a par exemple répondu à plusieurs consultations publiques de la Commission européenne sur l'IA, afin de défendre la prise en compte des enjeux environnementaux associés (Arcep, 2024 ; Arcep 2025d ; Arcep ; 2025e ; Arcep 2026b). Dans le cadre de sa réponse à la consultation publique sur la future législation sur l'équité numérique (le *Digital Fairness Act - DFA)*, dont le projet devrait être publié en 2026, l'Arcep soulignait que ce texte pourrait constituer un véhicule intéressant pour favoriser la liberté de choix des consommateurs et une meilleure efficacité environnementale de leurs produits et services numériques. En particulier, l'Arcep invitait à la mise en place de mesures d'écoconception pour les services d'IA en faveur de l'usage de modèles plus frugaux, en s'inspirant des mesures déjà identifiées dans des référentiels tels que le RGSEN (cf. encadré *infra*) ou le Référentiel général pour une IA frugale. En outre, le *Digital Networks Act (DNA)* comporte dans son projet publié en janvier 2026 un élargissement des objectifs de régulation des réseaux numériques aux questions d'efficacité de l'acheminement de trafic et de soutenabilité des services numériques. Dans la pratique, les fournisseurs de services numériques, y compris de services d'IA, devraient être tenus de mettre en place des mesures d'écoconception visant à réduire le trafic de données superflu et la consommation énergétique associée.

À travers une ou plusieurs initiatives législatives à venir, l'Union européenne pourrait ainsi se doter d'instruments œuvrant au développement de services d'IA alignés avec ses objectifs environnementaux, et dont l'efficacité pourrait être, à terme, un avantage compétitif compte tenu des contraintes croissantes sur les ressources.

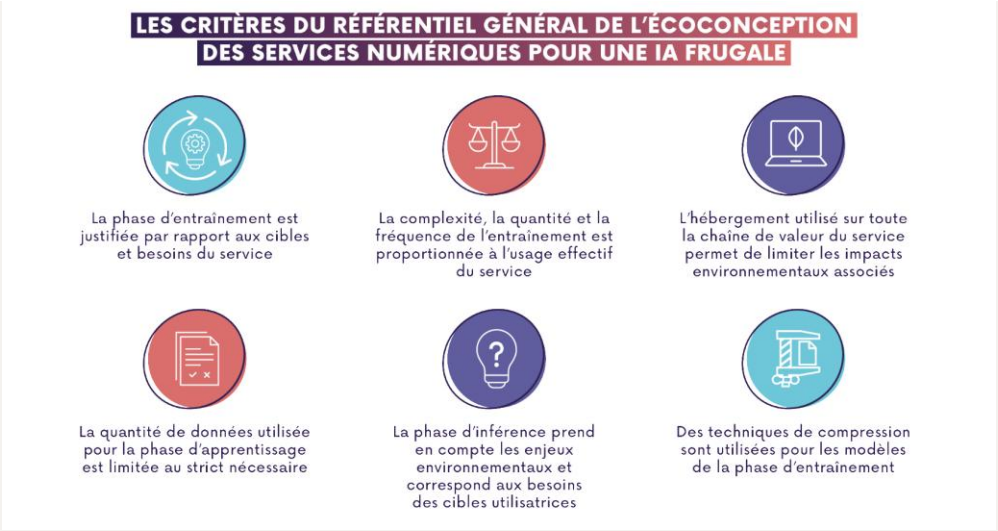
Zoom sur la partie « Algorithmie » du Référentiel général de l'écoconception des services numériques (RGESN)

Des bonnes pratiques et recommandations sont d'ores et déjà disponibles en matière d'écoconception des services d'IA, d'IA générative en particulier. À titre d'exemple, le référentiel général de l'écoconception des services numériques[117] (RGESN, Arcep et Arcom, 2024) intègre une série de bonnes pratiques que les concepteurs peuvent mettre en œuvre pour construire une démarche volontaire d'écoconception pour leurs services intégrant des systèmes d'IA. Le RGESN souligne en particulier la nécessité de définir le besoin du service et des utilisateurs cibles afin d'adapter la conception du modèle d'IA envisagé. Après avoir défini la pertinence du modèle, il est aussi nécessaire de choisir les options les plus sobres, par exemple en utilisant des modèles pré-entraînés à chaque fois que possible. Il s'agit également de prévoir des modalités d'entraînement des algorithmes qui sont proportionnées aux besoins essentiels du service en termes de quantité, de fréquence, de compression ou de données collectées. Outre les pratiques spécifiques au modèle algorithmique, les bonnes pratiques relatives à l'adoption de solutions d'hébergement soutenables ou à l'optimisation de l'interface utilisateurs sont essentielles à considérer pour limiter l'empreinte environnementale associée aux services fondés sur un système IA, de l'entraînement du modèle à son inférence.

Le RGESN inclut dans la partie « Algorithmie » sept critères relatifs à la mise en place de principes d'écoconception et de frugalité quant à l'entraînement et l'inférence des modèles algorithmiques utilisés par l'IA :

- 9.1 - Le service numérique a-t-il interrogé la nécessité d'une phase d'entraînement pour éviter un usage non justifié et déraisonné ?
- 9.2 - Le service numérique utilise-t-il une phase d'apprentissage avec un niveau de complexité minimisé et proportionné à l'usage effectif du service ?
- 9.3 - Le service numérique a-t-il mis en place des mécanismes visant à limiter la quantité d'entraînement nécessaire à son fonctionnement ?
- 9.4 - Le service numérique limite-il la quantité de données utilisées pour la phase d'apprentissage au strict nécessaire ?
- 9.5 - Le service numérique optimise-t-il l'occurrence de mise à jour et de réentraînement des modèles en fonction de ses besoins et des cibles utilisatrices ?
- 9.6 - Le service numérique utilise-t-il des techniques de compression pour les modèles utilisés lors de la phase d'entraînement ?
- 9.7 - Le service numérique utilise-t-il une stratégie d'inférence optimisée en termes de consommation de ressources et des cibles utilisatrices ?

[117] Le référentiel général de l'écoconception des services numériques est un document technique destiné aux experts et métiers du numérique souhaitant mettre en œuvre une démarche d'écoconception pour un service (sites, applications, IA, logiciels, API). Il a été élaboré par l'Arcep et l'Arcom, en collaboration avec l'ADEME, la DINUM, la CNIL et l'Inria. L'Inria a été impliquée tout particulièrement sur ce volet Algorithmie.



Au-delà de la partie « Algorithmie », d'autres critères du RGESN pourraient être mobilisés par les développeurs de services d'IA générative pour mettre en place de bonnes pratiques telles que :

- Interroger la nécessité du déploiement de l'IA générative pour la déployer là où il y en a le plus besoin[118],
- Rationaliser et optimiser l'entraînement des modèles pour limiter sa fréquence et adapter leur consommation aux contraintes du réseau électrique,
- Utiliser des modèles pré-entraînés, des bases existantes ou issues de la recherche pour éviter des entraînements systématiques aux coûts énergétiques élevés,
- Systématiser le *reporting* d'indicateurs de suivi de l'empreinte environnementale, pour l'entraînement et pour l'inférence des modèles d'IA,
- Choisir des solutions d'hébergement soutenable sur toute la chaîne de valeur du service d'IA[119],
- Assurer un parcours de navigation optimisée par les agents conversationnels fondés sur l'IA, y compris en limitant les pratiques de captation de l'attention ou des éléments trompeurs dans leur interface[120],
- Informer l'utilisateur des requêtes qui ont des impacts environnementaux importants[121],
- Inciter l'utilisateur à limiter son impact environnemental et permettre à l'utilisateur de supprimer les questions et propositions ajoutées automatiquement par l'IA à la suite d'une réponse à une question.

[118] En lien avec le critère 9.1 « Le service numérique a-t-il interrogé la nécessité d'une phase d'entraînement pour éviter un usage non justifié et déraisonné ? » et le critère 1.1 « Le service numérique a-t-il été évalué favorablement en termes d'utilité en tenant compte de ses impacts environnementaux ? ».

[119] En lien avec le critère 8.1 « Le service numérique utilise-t-il un hébergement ayant une démarche de réduction de son empreinte environnementale ? » et le critère 8.3 « Le service numérique utilise-t-il un hébergement dont le PUE (*Power Usage Effectiveness*) est minimisé ».

[120] En lien avec le critère 4.3 du RGESN (« Le service numérique optimise-t-il le parcours de navigation pour chaque fonctionnalité principale ? ») et le critère 1.6 (« Le service numérique collecte-t-il la donnée de façon responsable et raisonnée ? ») ;

[121] En lien avec le critère 4.12 du RGESN (« Le service numérique indique-t-il à l'utilisateur que l'utilisation d'une fonctionnalité a des impacts environnementaux importants ? » et le critère 4.15 du RGESN « Le service numérique fournit-il à l'utilisateur un moyen de contrôle sur ses usages afin de suivre et de réduire les impacts environnementaux associés ? » ;

Recommandation 4. Renforcer l'écoconditionnalité dans les modalités de soutien à l'innovation et la commande publique

Les pouvoirs publics, comme l'État ou les collectivités territoriales, peuvent jouer un rôle pour favoriser des modèles et des services plus vertueux, notamment en renforçant l'écoconditionnalité[122] dans les modalités de soutien à l'innovation[123] et la commande publique[124]. De telles mesures s'inscriraient en cohérence avec les objectifs de politique publique en permettant de concilier développement de l'IA et maîtrise des impacts environnementaux[125].

Les enjeux environnementaux pourraient être intégrés aux politiques de soutien à l'innovation de l'IA, que ce soit à l'échelle nationale ou à l'échelle européenne. À titre d'exemple, certaines incitations publiques existent, notamment en matière d'écoconditionnalité, dans le cadre du dispositif de soutien à l'innovation France 2030. Les projets soumis à ce dispositif ont l'obligation d'utiliser l'outil Green Algorithms[126] (Lannelongue, 2021) pour mener une évaluation de l'impact environnemental du modèle candidat au financement. Cette quantification des impacts environnementaux sur la base d'un outil commun d'évaluation permet de tenir compte de ces indicateurs dans l'évaluation de la qualité des dossiers soumis et le choix des financements. En Europe, le conditionnement des aides publiques à une évaluation systématique à partir d'une méthodologie commune et harmonisée au sein de l'UE pourrait être envisagé. Ce mécanisme d'écoconditionnalité pourrait par exemple s'appliquer aux financements de l'initiative européenne GenAI4EU pour les prochaines vagues de financement[127].

Outre les appels à projets, la prise en compte de la frugalité des services d'IA dans les conditions d'obtention de marchés *via* la commande publique participerait d'une logique d'exemplarité et permettrait de stimuler la demande, compte tenu de la taille potentielle des marchés cumulés (besoins en services d'IA à différentes échelles de la puissance publique). L'introduction d'un critère environnemental caractérisant la frugalité des services d'IA suffisamment élevé dans la notation finale dans les marchés publics d'État ou des collectivités territoriales pourrait être une mesure concrètement applicable. Afin d'évaluer ce critère environnemental, les donneurs d'ordre publics pourraient s'appuyer sur les référentiels précédemment cités comme le RGESN ou le Référentiel général pour l'IA frugale qui proposent des bonnes pratiques objectives. Ce critère environnemental pourrait être rendu obligatoire dans la loi, à l'instar de l'Article 55 de la loi Réduire l'empreinte environnementale du numérique (REEN) qui dispose que *« l'administration publique, lors de ses achats, favorise le recours à des logiciels dont la conception permet de limiter la consommation énergétique associée à leur utilisation* ».

---

[122] L'écoconditionnalité consiste à conditionner le versement des aides publiques ou à subordonner les achats au respect de principes et de critères environnementaux.

[123] Le soutien à l'innovation passe notamment par des aides de type « guichet », des concours d'innovation ou des prises de participation.

[124] La commande publique correspond à l'ensemble des contrats conclus à titre onéreux par un acheteur public ou une autorité concédante ayant une mission de service public. En France, l'ensemble de la réglementation est codifié dans le Code de la commande publique.

[125] À titre d'exemple, le projet de 3e édition de la Stratégie Nationale Bas Carbone (SNBC) en France, qui intègre pour la première fois un volet numérique en visant à « *stabiliser les émissions de gaz à effet de serre d'ici 2030-2035 tout en préservant la souveraineté numérique* », recommande d'« *encourager le recours à l'IA frugale* » pour contribuer à « *la réduction de l'empreinte des centres de données et une meilleure gestion du cycle de vie des données* » (Ministère de la Transition Écologique, 2025).

[126] Cette suite d'outils intègre en particulier un calculateur en ligne fournissant des estimations d'utilisation énergétique et d'empreinte carbone à partir de détails sur l'algorithme (ex : temps de calcul, nombre de processeurs utilisés, localisation...).

[127] Commission européenne, *GenAI4EU*, Consulté en 2026

Concrètement, que ce soit dans les politiques de soutien à l'innovation ou la commande publique, les pouvoirs publics pourraient favoriser des modèles d'IA générative dont la taille est adaptée aux besoins. Des modèles de plus petite taille peuvent être adaptés pour certains cas d'usage spécifiques[128] (ex : robot conversationnel, génération de code, traduction, contraction de texte ou applications médicales), et présenter de plus faibles impacts, en particulier en phase d'entraînement. Les pouvoirs publics pourraient privilégier les acteurs qui respectent des codes de conduite reconnus pour réduire l'impact environnemental de leurs services par la mise en place d'une démarche d'écoconception. Une piste à privilégier pour avoir un code de conduite européen reconnu serait de rendre effective la disposition prévue dans le Règlement IA (article 95, cf. encadré en fin d'Axe 1) visant à minimiser l'impact environnemental des services d'IA par des codes de conduite volontaires.

[128] Hugging Face, John Johnson, Small Language Models (SLM): A Comprehensive Overview, 2025

Axe 3 : Donner les moyens aux utilisateurs de choisir leurs services d'IA générative en fonction de leur impact environnemental grâce à une régulation européenne adaptée

Le développement des services d'IA générative doit être compatible avec la préservation de la liberté de choix des utilisateurs finals (Arcep, 2026a). Ainsi, ces derniers doivent pouvoir avoir le choix de leurs services et de leurs fonctionnalités, et plus largement des technologies qui seront utilisées au quotidien. Pour cela, il est important que les fournisseurs donnent aux utilisateurs les moyens de maîtriser leurs usages et donc leur impact environnemental.

Recommandation 5. Imposer une plus grande transparence environnementale aux fabricants de puces et aux grands fournisseurs de modèles et de services d'IA

Une plus grande transparence environnementale de la part des acteurs de la chaîne de valeur (fabricants de puces, fournisseurs de modèles, fournisseurs de services…) permettrait de fournir aux acteurs économiques et aux utilisateurs les informations nécessaires pour permettre des choix éclairés. La communication de l'empreinte environnementale, sur la base d'informations fiables et de méthodologies harmonisées, est notamment essentielle pour permettre la comparabilité des modèles et des services d'IA générative et par conséquent le bon fonctionnement concurrentiel de leur marché.

Les informations sur l'impact environnemental des services d'IA et des usages associés sont aujourd'hui très peu accessibles pour les utilisateurs, qu'il s'agisse du grand public, d'entreprises ou de développeurs de services d'IA Une étude réalisée en 2025 montre d'ailleurs que les personnes qui utilisent le plus les services d'IA générative sont généralement peu sensibilisées aux enjeux environnementaux associés (Moravec *et al.*, 2025). Les développeurs de services d'IA, s'appuyant sur des modèles fournis par des tiers, pourraient également bénéficier d'une information environnementale claire et transparente, pour utiliser les modèles les plus adaptés à leurs besoins et avec le moins d'impact. Une plus grande transparence environnementale permettrait une meilleure appréciation par l'ensemble des entreprises des impacts relatifs à l'IA générative, qui peuvent limiter leur capacité à atteindre leurs objectifs climatiques (Desroches et al., 2025).

Face à cette asymétrie d'information, les pouvoirs publics doivent imposer une plus grande transparence environnementale aux fabricants de puces et aux fournisseurs de modèles et de services d'IA. Un certain nombre d'entreprises sont d'ailleurs déjà soumises[129] à des obligations de *reporting* environnemental, comme celles fixées par la directive (UE) 2022/2464 sur la publication d'informations en matière de durabilité par les entreprises, dite CSRD, dans l'Union européenne, ou les Bilans d'Émissions de Gaz à Effet de Serre, dit BEGES, en France. Dans ce cadre, elles collectent des données environnementales, notamment sur leurs émissions directes et indirectes de gaz à effet de serre ou leur consommation d'énergie. Certaines entreprises réalisent même des ACV sans que cela ne soit obligatoire et possèdent donc déjà un certain nombre d'informations environnementales utiles à communiquer.

[129] Avant *la loi Omnibus I*, les entreprises concernées par ces obligations étaient celles remplissant deux des trois conditions suivantes : un bilan total de 25 millions d'euros, un chiffres d'affaires net de 50 millions d'euros ou un nombre moyen de salariés employés au cours de l'exercice de 250. Étaient également concernées, les entreprises cotées sur le marché réglementé européen et les entreprises non européennes ayant à la date de clôture des deux derniers exercices consécutifs un chiffre d'affaires net européen supérieur à 150 millions d'euros et disposant d'une succursale en France dont le chiffre d'affaires net excède 40 millions d'euros. La loi Omnibus I restreint le champ d'application aux entreprises de plus de 1 000 salariés et dont le chiffre d'affaires annuel net est supérieur à 450 millions d'euros. En ce qui concerne les entreprises de pays tiers, les exigences actualisées ne s'appliqueront qu'aux entreprises dont la société mère réalise dans l'UE un chiffre d'affaires annuel net de plus de 450 millions d'euros, et de plus de 200 millions d'euros pour la filiale ou la succursale.

Les pouvoirs publics pourraient renforcer les obligations de transparence par différentes mesures. La mise en place d'étiquettes obligatoires (à l'instar des étiquettes énergétiques)[130] ou de labels pour les services transparents sur leur impact environnemental (ou les services écoconçus dans un deuxième temps), s'appuyant sur une approche multicritère et sur l'ensemble du cycle de vie des systèmes d'IA, pourrait par exemple être envisagée à l'échelle européenne. La Commission européenne a d'ailleurs lancé une étude visant à développer un cadre de mesure et une étiquette européenne portant sur la consommation d'énergie et les émissions de gaz à effet de serre des services d'IA[131]. Une consultation publique a été lancée le 7 avril dernier pour recueillir des informations auprès des acteurs[132]. L'aboutissement de cette initiative permettrait d'améliorer la transparence environnementale des services d'IA.

Du côté des fournisseurs de modèles d'IA, en plus des informations environnementales évoquées précédemment, il serait utile de pouvoir bénéficier d'informations techniques comme la taille du modèle, la durée et la localisation de l'entraînement ou le matériel utilisé (notamment les puces) pour l'entraînement ou l'inférence. Ces informations pourraient figurer dans la documentation technique prévue par la disposition de l'article 53(1)(a) du Règlement IA (voir encadré en fin d'Axe 1), permettant de comprendre les facteurs-clés de la consommation énergétique des modèles.

Du côté des fabricants de puces, des informations environnementales, telles que les fiches *Product Carbon Footprint* et des indicateurs environnementaux, devraient être mises à disposition, à l'instar des dispositifs contraignants de transparence environnementale existants dans des secteurs comme celui du bâtiment ou des transports, compte tenu de l'influence significative des puces sur la consommation énergétique, que ce soit en phase d'entraînement ou d'inférence (cf. Chapitre 2). Améliorer la transparence environnementale sur les puces récentes permettrait d'avoir une vision granulaire fiable et fine de la consommation énergétique de l'IA et plus largement des centres de données.

Recommandation 6. Garantir l'ouverture des services d'IA

Comme souligné dans un rapport sur les défis de l'internet ouvert publié par l'Arcep[133], la liberté de choix des utilisateurs finals doit également s'accompagner de moyens d'actions dans leur usage des services d'IA générative. Or, face à l'hybridation croissante des services numériques et d'IA générative et à l'inclusion de fonctionnalités d'IA générative dans les interfaces des services, il peut s'avérer difficile, voire impossible, pour les utilisateurs de désactiver les services d'IA générative proposés par défaut, notamment au sein de certaines applications.

Parallèlement, l'intégration par défaut de l'IA générative au sein de services numériques déjà largement utilisés renforce le pouvoir de marché des acteurs verticalement intégrés, et accroît le risque de pratiques préjudiciables aux bonnes dynamiques de marché. Elle représente ainsi une barrière supplémentaire à la fourniture de services concurrents par des acteurs proposant des solutions plus respectueuses de l'environnement.

---

[130] Une étiquette énergétique pourrait donner des informations sur la consommation énergétique des modèles et des services d'IA générative, intégrant les phases d'entraînement et d'inférence, par exemple la consommation énergétique totale de l'entraînement et la consommation énergétique par prompt du service.

[131] Commission européenne, "*Artificial Intelligence Act: Call for tenders to measure and foster energy efficient and low emission artificial intelligence in the EU"*,2024

[132] European Commission, *Targeted consultation on measuring energy consumption and emissions of AI models and systems*, 2026

[133] Arcep, 2026, IA générative : des défis pour l'avenir de l'internet ouvert

Dans cette optique, l'Arcep défend notamment la mobilisation du règlement sur les marchés numériques, aussi appelé *Digital Markets Act* ou DMA, pour garantir l'ouverture des services d'IA générative. Plusieurs acteurs et services intégrant des fonctionnalités d'IA, notamment certains moteurs de recherche, systèmes d'exploitation et navigateurs[134], sont déjà concernés par ce règlement. Cependant, certaines de ses obligations nécessitent d'être adaptées[135], afin de permettre, d'une part, aux utilisateurs de choisir librement des services d'IA distincts de celui du contrôleur d'accès dans les cas d'une fourniture liée[136] et, d'autre part, aux agents IA fournis par des tiers d'accéder aux outils et services numériques du contrôleur d'accès (e.g. moteur de recherche), afin de répondre à la requête de l'utilisateur.

Enfin, le règlement sur les données (Data Act) pourrait également s'appliquer à certains fournisseurs d'IA générative dont les services sont susceptibles d'être identifiés à des services *cloud*, notamment les services de type « *AI as a service » (AIaaS)*. L'application du règlement sur les données pourra ainsi participer à l'ouverture des services d'IA générative en facilitant la portabilité des données des utilisateurs de ces services et l'interopérabilité des systèmes.

La disponibilité de solutions de paramétrage, de désinstallation des services d'IA générative intégrés et d'interopérabilité renforcerait la liberté de choix des utilisateurs, permettrait une maîtrise accrue de l'impact environnemental des usages, et pourrait contribuer à l'émergence de solutions alternatives, potentiellement plus vertueuses.

[134] Commission européenne, 2024. High-Level Group for the Digital Markets Act Public Statement on Artificial Intelligence.

[135] Notamment *via* l'adoption d'actes délégués.

[136] L'article 6(3) du DMA, qui permet par exemple aux utilisateurs finals de modifier les paramètres par défaut de certains services de plateforme essentiels, pourraient être élargis aux services cloud, aux moteurs de recherche et aux messageries instantanées afin d'éviter l'intégration par défaut, voire exclusive, des services IA, comme c'est le cas aujourd'hui avec Microsoft Copilot et Microsoft 365, ou avec Meta AI et WhatsApp, entre autres.

Axe 4 : Construire une stratégie de développement des centres de données en Europe alliant souveraineté et soutenabilité

Les centres de données sont des infrastructures numériques stratégiques pour la souveraineté européenne et nationale. Leur développement joue également un rôle croissant dans la compétitivité économique et l'essor des usages numériques. Dans ce contexte, en tant que régulateur des infrastructures numériques, l'Arcep propose des pistes de réflexion afin de contribuer à concilier les enjeux de soutenabilité, de compétitivité et de souveraineté dans la stratégie de développement des centres de données en Europe.

Recommandation 7. Éclairer les choix publics et d'investissement : faire de la transparence et de la régulation par la donnée des atouts du développement des centres de données en Europe

Dans le cadre de ses travaux prospectifs[137], l'Arcep notait la nécessité de trouver des moyens de prendre en compte les préoccupations sociales dans la façon dont se déploient les nouvelles technologies et les infrastructures numériques sous-jacentes. À cet égard, elle relevait l'importance que les utilisateurs finals puissent avoir accès à une information fiable et transparente pour éclairer, de manière neutre et objective, une réflexion individuelle et partant, nourrir utilement le débat public.

Comme évoqué précédemment dans ce chapitre, pour éclairer les choix, publics comme privés, professionnels comme personnels, la transparence doit être effective à tous les niveaux de la chaîne de valeur de l'IA et ce dès l'amont de la fourniture des services d'IA, c'est-à-dire en ce qui concerne les infrastructures de centres de données.

Pour ce faire, le suivi d'indicateurs pertinents pour les centres de données, tels que l'utilisation des ressources (notamment électricité et eau) ou le dimensionnement de la puissance électrique associée à ces infrastructures, constitue un prérequis. En ce sens, la révision de la directive efficacité énergétique et le règlement délégué (UE) 2024/1364 de la Commission européenne, qui ont permis d'instaurer un mécanisme de collecte de données environnementales sur les centres de données au niveau européen, constituent un premier pas. Les analyses qui pourront être proposées à partir des données recueillies sur les centres de données contribueront à nourrir la réflexion de l'Union européenne en matière de planification énergétique.

Ce type de mécanisme constitue ainsi une première avancée pour mener des analyses utiles à l'élaboration de futurs standards minimaux de performance pour les centres de données. Pour donner toute la portée de cet instrument de politique publique, il convient ainsi de veiller à sa mise en œuvre effective, à travers un niveau élevé de complétude et une fiabilité des données collectées. Selon l'Union européenne[138], les premiers retours indiquent en effet une participation encore partielle des centres de données, ce qui limite à ce stade la représentativité des analyses pouvant en être tirées. Dans ce contexte, il apparaît souhaitable d'accompagner la montée en charge de ce mécanisme par une amélioration des pratiques de *reporting* et de leur suivi, afin de garantir, à terme, une base d'information suffisamment exhaustive pour orienter efficacement les futures décisions réglementaires de l'Union européenne.

[137] Arcep, « Choisir son numérique : les réseaux télécoms au regard des usages numériques », Réseaux du futur, septembre 2025.

Recommandation 8. Renforcer la coordination européenne entre politiques numérique, énergétique et d'infrastructure pour accompagner le développement des centres de données

La planification électrique est un des facteurs-clés du succès des politiques européennes visant la décarbonation, la réindustrialisation et la souveraineté énergétique du continent. Il apparaît donc essentiel d'anticiper les futurs et nécessaires investissements sur le réseau électrique ainsi que les conflits d'accès et d'utilisation de la ressource en électricité, en évitant d'analyser le développement de l'IA, et des centres de données qui la supporteront en Europe, de manière isolée. En effet, le développement de l'IA et des centres de données peut entrer en concurrence avec l'électrification d'autres industries comme le transport qui, vont aussi nécessiter d'importants investissements pour développer le réseau électrique. Dans ce contexte, il est donc important d'avoir une vision à l'échelle nationale et européenne, pour coordonner l'implantation des centres de données, industrie géographiquement concentrée en *hub* par nature. En outre, l'industrie des centres de données étant électro-intensive, il est aussi pertinent de veiller à garantir de telles implantations là où le mix électrique est fortement décarboné comme en France.

À cet égard, des travaux tels que ceux menés par RTE dans le cadre du schéma décennal de développement du réseau (SDDR) ou sur les évolutions possibles des conditions de raccordement au réseau de transport d'électricité, permettent d'éclairer les décideurs publics. À l'échelle européenne, la publication à venir de la Feuille de route stratégique pour la numérisation et l'intelligence artificielle dans le secteur de l'énergie par la Commission européenne pourrait tenir compte de l'articulation entre développement des centres de données et développement du système électrique. Les centres de données doivent être envisagés comme des composantes à part entière du système électrique en tenant compte des services qu'ils pourraient fournir au réseau, notamment en matière de flexibilité. L'Europe peut utilement s'inspirer de l'Irlande qui a été confrontée plus tôt aux conséquences d'un développement très rapide des centres de données sur le système électrique (participant par exemple à rendre plus difficile l'atteinte de ses objectifs de décarbonation à 2030), et demande aujourd'hui des conditions ambitieuses pour le raccordement au réseau des nouveaux projets (ex : critère de développement de capacité d'énergies renouvelables).

Recommandation 9. Encourager une implantation territoriale concertée des centres de données

Au-delà des orientations stratégiques définies à l'échelle européenne, le succès de l'implantation des centres de données repose également sur une réflexion concertée avec les territoires, car elle soulève aussi des enjeux locaux.

Le stress hydrique, la valorisation effective de la chaleur fatale et la consommation foncière des centres de données sont intimement liés avec l'acceptabilité sociale de ces projets. En particulier, certaines contestations locales face à des projets de centres de données commencent à voir le jour, par exemple en France. À cet égard, le Haut Conseil pour le Climat appelle, en France, à réaliser un suivi des projets à la maille locale, indiquant notamment que les conflits d'usage ne se limitent pas aux questions électriques, mais concernent également les ressources en eau et le foncier disponible, qui ont une dimension principalement locale.

Dans un contexte de forte ambition de l'Europe sur le développement des centres de données, à travers l'*AI Continent Action Plan*, ces enjeux plus locaux doivent être pris en compte en amont, lors de l'élaboration des politiques locales de développement du tissu industriel pour garantir le succès et la bonne intégration dans les territoires de ces projets. L'ambition ne doit pas se limiter à un niveau élevé de capacités de calcul disponibles sur le sol européen, mais aussi sur la capacité des futurs projets à donner du sens à leur insertion dans les territoires, par leur soutenabilité effective, tant sur le plan économique qu'environnemental (ex : emploi et retombées fiscales au niveau local, anticipation en

amont pour réellement[139] valoriser la chaleur fatale, considération prospective sur le stress hydrique de la localité en question).

Pour ce faire, la gestion des enjeux liés à l'eau, au foncier et à la valorisation de la chaleur fatale des centres de données doit s'appuyer en priorité sur les dispositifs juridiques et réglementaires existants. Ceux-ci prévoient déjà des mécanismes de participation et de concertation du public[140]. Signe d'un intérêt grandissant pour la question et les réflexions autour d'une implantation durable et concertée des centres de données, certaines initiatives collaboratives commencent par ailleurs à voir le jour et proposent des outils pour accompagner cette démarche, comme le guide du datacenter durable et acceptable[141]. Ces initiatives innovantes peuvent aussi nourrir la déclinaison opérationnelle à venir des orientations proposées par la Commission européenne dans « l'*AI Continent Action Plan ».*

[139] Si la chaleur fatale ne dépasse pas une certaine température, elle est en général plutôt considérée comme un déchet thermique car elle n'est pas utilisable pour des usages comme approvisionner un réseau de chaleur urbain.

[140] Par exemple, en France, ces mécanismes reposent notamment sur l'enquête publique, intégrée à la procédure d'instruction d'un projet (pas uniquement pour l'industrie des centres de données) susceptible de porter atteinte à l'environnement, ainsi que sur des dispositifs de débat public conduits par la Commission nationale du débat public (CNDP) en amont pour des projets de grande envergure. À titre d'illustration, un projet de centre de données à Châteauroux a récemment fait l'objet d'une saisine conjointe de la CNDP à l'initiative des maîtres d'ouvrage, en l'occurrence Google et le gestionnaire du réseau de transport d'électricité (RTE).

[141] https://www.ville-demain.com/

Annexes

1. Bibliographie

Abitbol M., Aghion P., Antonin C., Barrage L. (2026), « Comment l'IA peut-elle contribuer aux objectifs d'efficacité énergétique dans l'industrie ? », L'institut Veolia et Microsoft.

ADEME (2026), « Prospective d'évolution des consommations des data centers à court, moyen et long terme de 2024 à 2060 ».

ADEME (2025a), « ACV de GPU pour l'intelligence artificielle ».

ADEME (2025b), « Évaluation environnementale des effets directs et indirects du numérique pour des cas d'usage », Étape 2 : numérisation de 10 cas d'usage.

ADEME (2025c), Évaluation de l'impact environnemental du numérique en France, Mise à jour de l'étude ADEME-Arcep.

ADEME (2024a), « Évaluation environnementale des effets directs et indirects du numérique pour des cas d'usage », Étape 0 : état de l'art des méthodologies existantes, Étape 1 : analyse de 30 cas d'usages.

ADEME (2024b), « Étude numérique et métaux ».

ADEME & Arcep (2022 et 2023), « Étude sur l'empreinte environnementale du numérique en France ».

AFNOR, Ministère de la Transition Écologique et de la Cohésion des Territoires (2024), « Référentiel général pour l'IA frugale ».

AIE (2025), "AI & Energy".

AIE (2024), "Electricity 2024".

Arcep (2026a), « IA générative : des défis pour l'avenir de l'internet ouvert ».

Arcep (2026b), « Contribution à la consultation publique sur le réexamen du programme d'action pour la décennie numérique (PADN) ».

Arcep (2026c), « Baromètre du numérique – édition 2026 ».

Arcep (2026d), « Enquête annuelle "Pour un numérique soutenable" - édition 2026 (données 2024) ».

Arcep (2025a), « Baromètre du numérique - édition 2025 ».

Arcep (2025b), « L'intelligence artificielle et les réseaux télécoms ».

Arcep (2025c), « Choisir son numérique : les réseaux télécoms au regard des usages du numérique ».

Arcep (2025d), « Contribution à la consultation publique sur les futures politiques publiques en matière d'IA et de cloud dans l'UE ».

Arcep (2025e), « Contribution à la consultation publique sur la feuille de route stratégique pour la numérisation et l'IA dans le domaine de l'énergie ».

Arcep (2025f), « Enquête annuelle "Pour un numérique soutenable" - édition 2025 (données 2023) ».

Arcep (2024), "Arcep's contribution to the call for contributions on competition in generative AI".

Arcep & Arcom (2024), « Référentiel général de l'écoconception des services numériques ».

Autorité de la Concurrence (2025), « Les questions concurrentielles relatives à l'impact énergétique et environnementale de l'Intelligence Artificielle ».

Banque Mondiale, Union internationale des télécommunications (2025), "Measuring National ICT Sector Environmental Impact, Arcep Case Study – France".

Beignon, A., Thibault, T., & Maudet, N. (2025). Imposing AI: Deceptive design patterns against sustainability. arXiv preprint arXiv:2508.08672.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., ... & Molchanov, P. (2025). Small language models are the future of agentic ai. arXiv preprint arXiv:2506.02153.

Bick, A., Blandin, A., & Deming, D. J. (2024). The Rapid Adoption of Generative AI (NBER Working Paper No. 32966). National Bureau of Economic Research.

Boavizta (2021), « Numérique et environnement : Comment évaluer l'empreinte de la fabrication d'un serveur, au-delà des émissions de gaz à effet de serre ? ».

Budennyy, S.A., Lazarev, V.D., Zakharenko, N.N. *et al*. (2022), eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. Dokl. Math. 106 (Suppl 1), S118–S128 (2022). https://doi.org/10.1134/S1064562422060230.

Carbone 4, Vasselin Z., Maquet P., Benedini C., Paulmier P. (2024), « Net Zero Initiative for IT: Understanding the role of digital solutions in the global net zero effort (NZI for IT)".

Chen, Z., Lin, W., Xie, X., Hu, Y., Li, C., Tong, Q., ... & Li, S. (2024, December). An empirical study on the power consumption of LLMs with different gpu platforms. In 2024 IEEE International Conference on Big Data (BigData) (pp. 8640-8642). IEEE.

Coalition for Sustainable AI (2025), "Standardization for AI Environmental Sustainability".

Commission de l'intelligence artificielle, Aghion P., Bouverot A. (2025) « IA : Notre ambition pour la France ».

Data for Good (2023), « Les grands défis de l'IA générative ».

Delavande, J., Pierrard, R., & Luccioni, S. (2026). Understanding Efficiency: Quantization, Batching, and Serving Strategies in LLM Energy Use. arXiv preprint arXiv:2601.22362.

Deloitte (2024), "Powering artificial intelligence".

Desroches, C., Chauvin, M., Ladan, L., Vateau, C., Gosset, S., & Cordier, P. (2025). Exploring the sustainable scaling of AI dilemma: A projective study of corporations' AI environmental impacts. arXiv preprint arXiv:2501.14334.

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

Diguet, C., Lopez, F., & Lefèvre, L. (2019). L'impact spatial et énergétique des data centers sur les territoires (Doctoral dissertation, ADEME, Direction Villes et territoires durables).

European Green Digital Coalition (2024), "Net Carbon Impact Assessment Methodology for ICT Solutions".

Falk, S., Ekchajzer, D., Pirson, T., Lees-Perasso, E., Wattiez, A., Biber-Freudenberger, L., ... & van Wynsberghe, A. (2025). "More than Carbon: Cradle-to-Grave environmental impacts of GenAI training on the Nvidia A100 GPU". arXiv preprint arXiv:2509.00093.

Falk, S., Corrêa, N. K., Luccioni, S., Biber-Freudenberger, L., & van Wynsberghe, A. (2025). From FLOPs to Footprints: The Resource Cost of Artificial Intelligence. arXiv preprint arXiv:2512.04142.

France Stratégie (2025), « La demande en eau, Prospective territorialisée à l'horizon 2050 ».

Google (2025), "Measuring the environmental impact of delivering AI at Google Scale".

Green IT (2025), « Baromètre de l'Écoconception Digitale ».

Green IT (2025) « Impacts environnementaux et sanitaires de l'intelligence artificielle ».

Gupta, U., Kim, Y. G., Lee, S., Tse, J., Lee, H. H. S., Wei, G. Y., ... & Wu, C. J. (2021, February). Chasing carbon: The elusive environmental footprint of computing. In 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA) (pp. 854-867). IEEE.

Hintemann, R., & Hinterholzer, S. (2019). Energy consumption of data centers worldwide. Business, Computer Science (ICT4S).

I Care (2025), "Intégrer la responsabilité environnementale dans le contexte de déploiement rapide de l'IA en entreprise ».

Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., & Hendawi, A. (2025). How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference. arXiv preprint arXiv:2505.09598.

You J., How much energy does ChatGPT use? (2025). URL https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use.

Kaack, L.H., Donti, P.L., Strubell, E. *et al*. Aligning artificial intelligence with climate change mitigation. Nat. Clim. Chang. 12, 518–527 (2022).

Kachris C. (2025), A Survey on Hardware Accelerators for Large Language Models, Appl. Sci. 15 (2), 586.

Kim M., Y. T-A., Chung J. (2025), Toward Sustainable Generative AI: A Scoping Review of Carbon Footprint and Environmental Impacts Across Training and Inference Stages, Pre-Print, DOI: 10.48550/arXiv.2511.17179.

Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: quantifying the carbon footprint of computation. Advanced science, 8(12), 2100707.

Ligozat, A. L., Lefevre, J., Bugeau, A., & Combaz, J. (2022). Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. Sustainability, 14(9), 5172.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making ai less' thirsty'. Communications of the ACM, 68(7), 54-61.

Limites Numériques (2025), « Comment les entreprises de la tech nous forcent à utiliser l'IA », 2025.

Lopez, F., & Diguet, C. (2023). Sous le feu numérique Spatialités et énergies des data centers.

Luccioni, A. S., Strubell, E., & Crawford, K. (2025, June). From efficiency gains to rebound effects: The problem of jevons' paradox in AI's polarized environmental debate. In Proceedings of the 2025 ACM conference on fairness, accountability, and transparency (pp. 76-88).

Luccioni, S., Gamazaychikov, B., Da Costa, T. A., & Strubell, E. (2025). Misinformation by omission: The need for more environmental transparency in AI. arXiv preprint arXiv:2506.15572.

Luccioni, S., Jernite, Y., & Strubell, E. (2024, June). Power hungry processing: Watts driving the cost of AI deployment?. In Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (pp. 85-99).

Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. Journal of machine learning research, 24(253), 1-15.

Ministère de la Transition Écologique et de la Cohésion des Territoires (2025), « Projet de Stratégie nationale bas-carbone n°3 ».

Ministère fédéral de l'Environnement, du Climat, de la Protection de la nature et de la Sécurité nucléaire, allemand (2026), “Towards a Sustainable and Competitive AI Economy, Recommendations to the German Federal Environment Ministry”.

Mistral (2025), “Our contribution to a global environmental standard for AI”.

Moravec, V., Gavurova, B., & Kovac, V. (2025). Environmental footprint of GenAI–Changing technological future or planet climate? Journal of Innovation & Knowledge, 10(3), 100691.

Patel, A., Mahalingam, N., & Patel, R. (2025). The Environmental Impact of AI Servers and Sustainable Solutions. arXiv preprint arXiv:2601.06063.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

Qiu, X., Parcollet, T., Fernandez-Marques, J., Gusmao, P. P., Gao, Y., Beutel, D. J., Topal, T., Mathur, A., & Lane, N. D. (2023). A first look into the carbon footprint of federated learning. Journal of machine learning research, 24(129), 1-23.

Rincé S., and Banse A., Ecologits: Evaluating the environmental impacts of generative AI. Journal of Open Source Software, 10(111):7471, 2025.

Roussilhe G. (2022), « Les effets indirects de la numérisation ».

Roussilhe G. (2025), “La phase G : les GPU et les IA génératives comme nouvelle phase de l’histoire environnementale de la numérisation ? (Partie 1).

Roussilhe G. (2021), « Eau et puces électroniques : l’avenir climatique et industriel de Taïwan ».

Samsi S., Zhao D., McDonald J., Baolin Li, Michaleas A., Jones M., Bergeron W., Kepner J., Tiwari D., and Gadepally V. (2023), From words to watts: Benchmarking the energy costs of large language model inference. In 2023 IEEE High Performance Extreme Computing Conference (HPEC), pages 1–9. IEEE, 2023.

Schneider, I., Xu, H., Benecke, S., Patterson, D., Huang, K., Ranganathan, P., & Elsworth, C. (2025). Life-cycle emissions of ai hardware: A cradle-to-grave approach and generational trends. arXiv preprint arXiv:2502.01671.

The Shift Project (2025), « Intelligence artificielle, données, calculs. Quelles infrastructures dans un monde décarboné ? ».

Sopra Steria (2025), « IA & environnement : sortir du brouillard informationnel ».

Stern, N., Romani, M., Pierfederici, R., Braun, M., Barraclough, D., Lingeswaran, S., ... & Niemann, N. (2025). Green and intelligent: the role of AI in the climate transition. npj Climate Action, 4(1), 56.

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 3645-3650).

UIT (2025), “Measuring what matters: How to assess AI’s environmental impact”.

Unesco (2025), “Smarter, Smaller, Stronger: Resource-Efficient Generative AI & the Future of Digital Transformation”.

Vanderbauwhede, W. (2024). Estimating the Increase in Emissions caused by AI-augmented Search. arXiv preprint arXiv:2407.16894.

Varoquaux, G., Luccioni, S., & Whittaker, M. (2025, June). Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 61-75).

Zhang, M., Carbajales-Dale, M., Ma, X., Guo, L., & Fan, C. (2025). Cleaner Grid or Smarter Cooling? Environmental Impact Trade-offs of a Data Center Using the Life Cycle Assessment Method. Cleaner Energy Systems, 100223.

2. Entretiens réalisés

Dans le cadre de ce rapport, des auditions ont été menées avec :

- Théo Alves Da Costa (Data for Good)
- Cyriaque Dubois (Mistral AI)
- Juliette Fropier (Ecolab)
- Marie-Liane Lekpeli, Véronique Bouvelle, Emma Le Boulicaut (Ministère de l'économie des finances – Direction générale des entreprises)
- François Philippe (RTE)
- Gaël Varoquaux (Inria)
- Eloïse Villani (Ministère de l'économie et des finances – Direction générale du Trésor)

Ces auditions ont servi à alimenter les réflexions de l'Arcep, toutefois le rapport ne reflète pas l'opinion des personnes auditionnées.

L'Arcep remercie également les spécialistes qui ont été consultés à d'autres occasions :

- Deux séminaires académiques internes ont été organisés à ce sujet, en invitant Anne-Laure Ligozat du LISN et Loïc Lannelongue de l'Université de Cambridge.
- Lors de la rédaction du Référentiel général pour l'écoconception des services numériques (RGESN), l'Inria a été interrogé, notamment Denis Trystram et Laurent Lefèvre (Inria).
- Lors de la rédaction de la note « Choisir son numérique : les réseaux télécoms au regard des usages du numérique », des auditons ont été organisées avec des acteurs économiques (Ericsson, Télécoop, Google, Meta et Orange), des académiques (Christian Licoppe, Clément Marquet, Jean-Samuel Beuscart, Nicolas Jullien & Soazig Lalancette et Laura Draetta), des collectivités territoriales (Nantes Métropole et Rennes Métropole) et des associations de collectivités (Association des Maires Ruraux de France et Les Inteconnectés).

L'Autorité remercie également les spécialistes suivants qui ont été rencontrés lors de réunions multilatérales ou bilatérales comme Sasha Luccioni (Hugging Face), Aurélie Bugeau (LaBRI) ou Marlène de Bank (The Shift Project). Elle remercie également ses partenaires avec qui elle a pu échanger sur le sujet sur la durée comme l'ADEME.

L'Arcep remercie les étudiants (Maela Guillaume-Le Gall, Rafaël Mourouvin et Julian Rojas) et professeurs (Franck Malherbet et Pierre Boyer) du master *Data and Economics for Public Policy* (DEPP) de l'École Polytechnique, l'ENSAE et de Telecom Paris pour leur projet sur les impacts environnementaux de l'IA.

Enfin, l'Autorité remercie tout particulièrement le PEReN pour sa collaboration dans le cadre du projet de mesure de la consommation énergétique des modèles d'IA en phase d'inférence.

3. Glossaire

Agent conversationnel : Système informatique conçu pour interagir avec un utilisateur humain en langage naturel. Il utilise des technologies d'intelligence artificielle, telles que le traitement du langage naturel et l'apprentissage automatique, pour comprendre des requêtes, générer des réponses pertinentes et simuler une conversation fluide. Les agents conversationnels peuvent être déployés sous forme de chatbots, d'assistants virtuels ou d'interfaces vocales.

Agent d'IA : Système logiciel qui emploie des techniques d'intelligence artificielle (notamment des LLM) pour percevoir son environnement, prendre des décisions et accomplir des tâches de manière autonome pour le compte d'utilisateurs humains ou d'organisations.

Ajustement (*fine tuning*) : Technique consistant à spécialiser un modèle d'IA pré-entraîné à l'accomplissement d'une tâche spécifique. Cela consiste généralement à entraîner le modèle dans son ensemble, ou seulement certaines couches d'un réseau de neurones, pour un faible nombre d'itérations sur un ensemble de données spécifiques correspondant à la tâche visée. Cette pratique est parfois traduite par « affinage », « réglage fin », « peaufinage » ou encore « spécialisation »[142].

Apprentissage automatique (*machine learning*) : Domaine de recherche relatif aux techniques mathématiques et statistiques qui visent à apprendre à des machines à effectuer des prédictions généralisables à partir de données (dites « d'entraînement »).

Apprentissage profond (*deep learning*) : Sous-domaine de l'apprentissage automatique (voir ci-dessus) qui vise à apprendre à des machines à effectuer des prédictions à l'aide de techniques fondées sur les réseaux de neurones (voir ci-dessous).

Assistant virtuel : L'article 2(12) du règlement sur les marchés numériques (voir ci-dessous) définit un assistant virtuel comme « *un logiciel qui peut traiter des demandes, des tâches ou des questions, notamment celles fondées sur des données d'entrée sonores, visuelles ou écrites, de gestes ou de mouvements, et qui, sur la base de ces demandes, tâches ou questions, donne accès à d'autres services ou contrôle des appareils connectés physiques* ».

Assistant vocal : Assistant virtuel (voir ci-dessus) qui fonctionne principalement à l'aide de données d'entrée sonores.

Cloud : L'article 2(8) du règlement sur les données définit un service cloud comme « *un service numérique* […] *qui permet un accès par réseau en tout lieu et à la demande à un ensemble partagé de ressources informatiques configurables, modulables et variables de nature centralisée, distribuée ou fortement distribuée, qui peuvent être rapidement mobilisées et libérées avec un minimum d'efforts* »

Données d'entraînement : Ensemble de données (ex : texte, sons, images...) utilisé pour entrainer un modèle d'apprentissage automatique ou d'intelligence artificielle, c'est-à-dire pour en ajuster les paramètres afin qu'il apprenne à accomplir une tâche.

Éditeurs ou fournisseurs de contenu et d'applications : Acteurs impliqués dans la fourniture de contenu (pages web, blogs, vidéos) et/ou des applications (moteurs de recherche, applications mobiles, service d'itinéraire) sur internet.

Entraînement/Apprentissage : Désigne le processus par lequel un système réalise, à partir de données et *via* des modèles algorithmiques, des calculs afin de proposer des fonctionnalités, d'améliorer ses performances ou d'acquérir une capacité à réaliser certaines tâches de manière autonome.

[142] CNIL. *Glossaire : définition d'ajustement (fine tuning).*

Grand modèle de langue ou LLM (*Large Language Model*) : Modèle d'IA de traitement du langage naturel comprenant un grand nombre de paramètres.

Hyperscaler : Fournisseur important de services *cloud*, capable de proposer des services de calcul et de stockage à grande échelle.

IA agentique : Champ des techniques liées aux agents d'IA (voir ci-dessus).

IA générative : Champ des techniques permettant d'utiliser des modèles d'intelligence artificielle pour générer de nouveaux contenus comme du texte, du code informatique, des images ou de la musique.

Inférence (phase) : Phase durant laquelle un modèle d'intelligence artificielle est utilisé pour effectuer des prédictions, par opposition à la phase d'entraînement. Un service d'IAG repose sur un (ou plusieurs) LLM en phase d'inférence.

Intelligence artificielle : Domaine de recherche qui vise à permettre à des systèmes artificiels, le plus souvent informatiques, d'effectuer des tâches associées à l'intelligence humaine ou naturelle. L'apprentissage automatique en est aujourd'hui la principale approche.

Instruction (*prompt*) : Requête textuelle adressée à un modèle par un utilisateur.

ISO : organisation internationale indépendante et non gouvernementale de normalisation. Les normes ISO visent à garantir la fiabilité, la sûreté et la qualité des produits et des services[143].

Jeton (ou *token* en anglais) : selon la norme ISO/IEC 22989/DAmd 1, unité de contenu qu'un modèle IA considère comme ayant une signification sémantique.

Modèle d'IA à usage général (MIAUG) : l'article 3(63) du règlement sur l'intelligence artificielle définit un modèle d'IA à usage général comme « *un modèle d'IA, y compris lorsque ce modèle d'IA est entraîné à l'aide d'un grand nombre de données utilisant l'auto-supervision à grande échelle, qui présente une généralité significative et est capable d'exécuter de manière compétente un large éventail de tâches distinctes, indépendamment de la manière dont le modèle est mis sur le marché, et qui peut être intégré dans une variété de systèmes ou d'applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant leur mise sur le marché* »

Modèle d'IA : Un modèle d'IA est un algorithme, dont le fonctionnement est déterminé par un ensemble d'attributs, et qui est conçu pour opérer, selon les cas, différentes tâches, telles que la prédiction, la classification, l'inférence ou la génération (définition issue du glossaire de la CNIL).

Modèle d'IA générative : Modèle d'IA caractérisé par sa capacité à produire, en réponse à une requête, un résultat *ad hoc* : texte, image, vidéo.

Modèle de fondation : Modèle d'IA entraîné sur de grands volumes de données d'un ou de plusieurs types afin de pouvoir être ensuite adapté à de nombreuses tâches. À titre d'exemple, dans le domaine du traitement du langage naturel, il s'agit des LLM.

Plateforme numérique : Le terme de plateforme numérique est généralement utilisé pour désigner un ensemble varié de services en ligne et d'acteurs offrant des services d'intermédiation, comme des places de marché, des plateformes communautaires ou encore des magasins d'applications. Ces acteurs peuvent présenter des caractéristiques très variables (en taille, revenus…) et évoluent dans de nombreux secteurs d'activité. Certaines grandes plateformes concentrent à elles seules de nombreux

Règlement IA : Règlement européen établissant un cadre harmonisé pour le développement, la mise sur le marché et l'utilisation des systèmes d'intelligence artificielle. Il adopte une approche fondée sur

[143] ISO. Site web de l'ISO.

les risques et prévoit des obligations notamment en matière de sécurité, transparence, gouvernance des données, documentation technique et supervision humaine. Le texte inclut aussi des dispositions spécifiques pour les modèles dits à usage général (MIAUG), notamment dans le domaine de l'IA générative[144].

Réseau de neurones : dans le domaine de l'intelligence artificielle, un réseau de neurones est un ensemble organisé de neurones artificiels interconnectés qui apprennent, grâce à des techniques d'optimisation issues de l'apprentissage automatique, à détecter des motifs (*patterns*) dans les données d'entraînement afin de résoudre des problèmes dans des domaines complexes tels que la vision par ordinateur ou le traitement du langage naturel (définition reprise à partir du glossaire de la CNIL).

Service d'IA générative : Service numérique reposant sur un modèle d'IA générative.

Système d'IA : selon la norme ISO/IEC 22989, système conçu pour générer des résultats tels que du contenu, des prévisions, des recommandations ou des décisions pour un ensemble donné d'objectifs définis par l'humain.

Système expert : selon la norme ISO/IEC 22989 système IA qui accumule, combine ou encapsule de la connaissance fournie par une ou des expertises humaines dans un domaine spécifique pour fournir une solution à des problèmes.

Trafic : quantité de données transitant à un instant T entre deux machines, par exemple le cas d'une interconnexion de données. Lorsque le trafic dépasse la capacité du lien, la liaison sature.

[144] CNIL, 2024. *Entrée en vigueur du règlement européen sur l'IA : les premières questions-réponses de la CNIL*.

4. Liste des acronymes

ACV : Analyse de Cycle de Vie

ACV-A : Analyse de Cycle de Vie – Attributionnelle

ACV-C : Analyse de Cycle de Vie – Conséquentielle

ADEME : Agence de la transition écologique (anciennement Agence de l'environnement et de la maîtrise de l'énergie)

ADSL : *Asymmetric Digital Subscriber Line*

AFNOR : Association française de normalisation

API : *Application Programming Interface*

Arcep : Autorité de Régulation des Communications Electroniques, des Postes et de la distribution de la Presse

ARN : Autorité de Régulation Nationale des communications électroniques

BEGES : Bilan des émissions gaz à effet de serre

BLOOM : *BigScience Large Open-science Open-access Multilingual Language Mode*

CEEMS : *Compute Energy & Emissions Monitoring Stack*

CEN/CENELEC : Comité européen de normalisation en électronique et en électrotechnique

CGDD : Commissariat général au développement durable

CNRS : Centre national de la recherche scientifique

CPU : *Central Processing Unit* (Unité centrale de calcul)

CSIA : Comité de surveillance des investissements d'avenir

CSRD : *Corporate Sustainability Reporting Directive* (Directive relative à la publication d'informations en matière de durabilité par les entreprises)

DLR : *Dynamic Line Rating*

DFA : *Digital Fairness Act*

DGE : Direction générale des entreprises

DNA : *Digital Networks Act*

ETSI : *European Telecommunications Standards Institute* (Institut européen des normes de télécommunications)

GES : Gaz à Effet de Serre

GPT : *Generative Pre-trained Transformer*

GPU : *Graphics Processing Unit* (Processeur graphique)

IA : Intelligence Artificielle

IEEE : *Institute of Electrical and Electronics Engineers* (Institut des ingénieurs électriciens et électronicien)

Inria : Institut national de recherche en sciences et technologies du numérique

ISO : *International Organization for Standardization* (Organisation internationale de normalisation)

LLM : *Large Language Model* (Grand modèle de langue)

LoRA : *Low-Rank Adaptation*

MoE : *Mixture of Experts* (Mélange d'experts)

ONU : Organisation des Nations Unies

PEReN : Pôle d'Expertise de la Régulation Numérique

PUE : *Power Usage Effectiveness*

RAG : Retrieval Augmented Generation (Génération augmentée de récupération)

REEN (loi) : Réduire l'Empreinte Environnementale du Numérique

RGESN : Référentiel Général de l'Ecoconception des Services Numériques

RTE : Réseau de Transport d'Electricité

SLM : *Small Language Model* (Petit modèle de langue)

TPU : *Tensor Processing Unit* (Unité de traitement de tenseur)

UE : Union européenne

UIT : Union internationale des télécommunications

UNESCO : *United Nations Educational, Scientific and Cultural Organization* (Organisation des Nations unies pour l'éducation, la science et la culture)

WUE : *Water Usage Effectiveness*

5. Ordres de grandeur et équivalences relatives à la consommation d'énergie et aux émissions de gaz à effet de serre

Périmètre	Émissions de gaz à effet de serre	Source
Union européenne en 2023	3 222 000 000 tCO_2e	UE, Edgar, 2025
France en 2024	369 000 000 tCO_2e	MTE, SDES, 2025
Numérique – France - 2022	29 500 000 tCO_2e	ADEME, 2025c
Terminaux utilisateurs - France - 2022	14 700 000 tCO_2e	ADEME, 2025c
Centres de données – France (dont usages centres de données à l'étranger) - 2022	13 500 000 tCO_2e	ADEME, 2025c
Réseaux - France - 2022	1 200 000 tCO_2e	ADEME, 2025c
Centres de données – France (périmètre collecte de données Arcep) – 2023	137 000 tCO_2e	Arcep, 2025
Entraînement du modèle Large 2 (18 mois)	20 400 tCO_2e	Mistral, 2025
Entraînement du modèle LLaMA3-405B	9 000 tCO_2e	Meta, 2025
Entraînement du modèle BLOOM (118 jours) – Centres de données	124 tCO_2e	Luccioni *et al.*, 2023
Empreinte carbone annuelle moyenne d'un français – 2023	8,5 tCO_2e	Citepa, ABC, 2023
Carte graphique GPU Nvidia A100 SXM 40 GBG - Fabrication	0,150 tCO_2e	Falk *et al.*, 2025

Tableau 3 – Ordres de grandeur et équivalences relatives aux émissions de gaz à effet de serre

Périmètre	Consommation d'énergie finale	Source
Union européenne – Électricité – 2023	2 326 TWh	UE, Eurostat, 2025
France - Électricité - 2024	449 TWh	RTE, 2025
Centres de données – Monde - 2024	415 TWh	AIE, 2025
Centres de données – France - Périmètre collecte de données Arcep – 2023	2,4 TWh	Arcep, 2025
Ville d'environ 100 000 habitants – Électricité	200 000 MWh	CRE, 2025
Entraînement de GPT-3 – Serveurs	1 287 MWh	Luccioni *et al.*, 2023
Entraînement du modèle BLOOM - Serveurs	433 MWh	Luccioni *et al.*, 2023
Foyer moyen – France – Estimation réalisée en 2025 pour 2023-2024	4,3 MWh	CRE, 2025

Tableau 4 - Ordres de grandeur et équivalences relatives à la consommation d'électricité

6. Comparaison de l'entraînement des modèles GPT-3 et BLOOM à partir de la revue de la littérature

Le tableau ci-dessous compare des données caractérisant les ressources nécessaires à l'entraînement des deux modèles d'IA « GPT-3 » et « BLOOM » (Patterson *et al.*, 2022 ; Luccioni *et al.*, 2022 ; Data for Good, 2023).

Dégroupage	**GPT-3 (OpenAI)**	**BLOOM (BigScience)**
Nombre de paramètres	175 milliards	176 milliards
Volume de données sur lesquelles s'entraîne le modèle	45 To	1,6 To
Architecture	Transformer (decoder-based)	Transformer (decoder-based)
Durée d'entraînement	15 jours	118 jours
Ressources mobilisées	10 000 GPUs (NVIDIA V100)	384 GPUs (NVIDIA A100)
Consommation électrique des serveurs pour entraînement	1 287 MWh	433 MWh
Intensité carbone de l'électricité consommée pour entraînement	427 gCO2e/kWh	57 gCO2e/kWh
Empreinte carbone lié à la consommation électrique des serveurs	**552 tCO2e**	**30 tCO2e**
Consommation électrique des serveurs et des équipements annexes (refroidissement, switchs, routeurs, …)	N/A	690 MWh
Empreinte carbone lié aux centres de données mobilisés (hors fabrication équipements)	**N/A**	**39,3 tCO2e**
Empreinte carbone lié aux centres de données mobilisés (comprenant fabrication équipements)	**N/A**	**50,5 tCO2e**
Empreinte carbone lié aux centres de données mobilisés (comprenant fabrication équipements) et avec tests, évaluations et corrections	**N/A (estimation à ~2200 tCO2e à partir du ratio constaté pour BLOOM)**	**123,8 tCO2e**

7. Performance de modèles d'IA en fonction du temps d'entraînement ou de l'empreinte mémoire pour différentes tâches

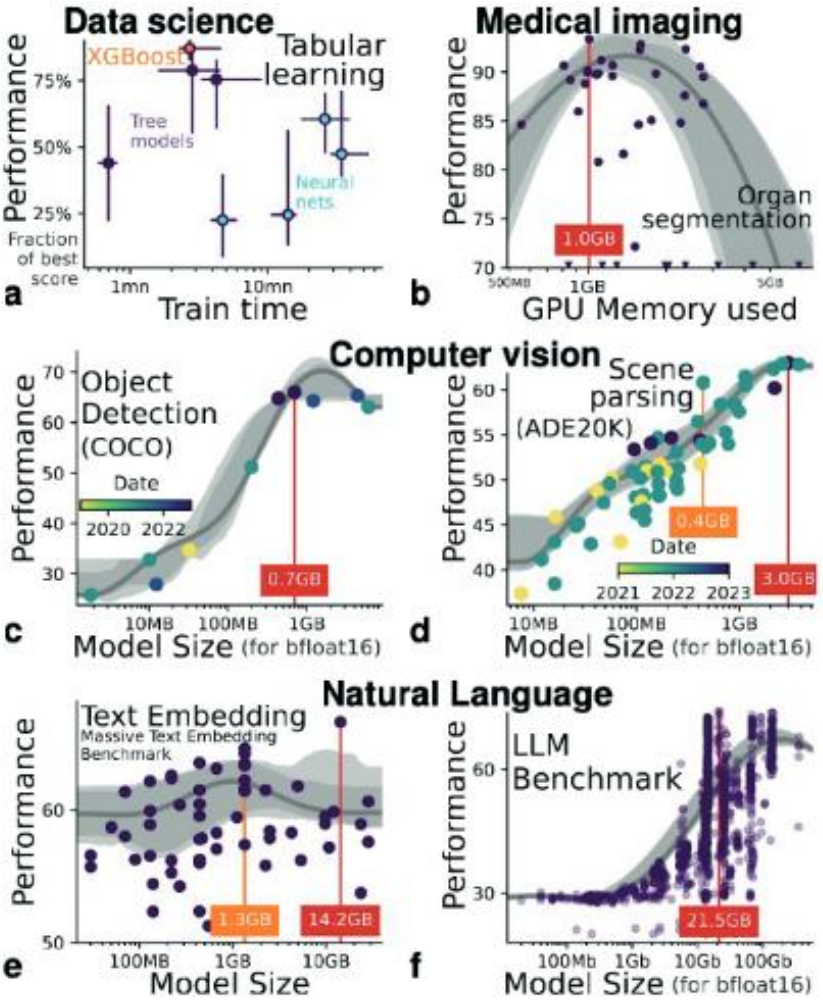


Figure 16 : Performance de modèles d'IA en fonction du temps d'entraînement ou de l'empreinte mémoire pour différentes tâches[145]

[145] Source : Varoquaux et al., 2025. Liste des tâches : apprentissage tabulaire ; segmentation d'images médicales ; détection d'objets par vision à l'ordinateur ; analyse syntaxique de scènes ; intégration de texte ; compréhension de texte.

8. Cartographie des différentes initiatives de normalisation des méthodes d'évaluation de l'impact environnemental de l'IA

La Coalition pour une IA durable lancée dans le cadre du Sommet pour l'Action sur l'IA de février 2025 a initié une feuille de route commune entre les différentes initiatives de normalisation, réunissant des experts de l'ISO, de l'UIT et de l'IEEE, en partenariat avec l'OCDE et l'UNESCO.

Cette feuille de route a permis d'initier une coordination entre les différents organismes de normalisation et d'optimiser les ressources dédiées.

Le schéma ci-dessous présente ainsi les initiatives de normalisation existantes et les liens avec d'autres standards en vigueur sur le sujet plus large de l'évaluation de l'impact environnemental du numérique ou des systèmes d'IA.

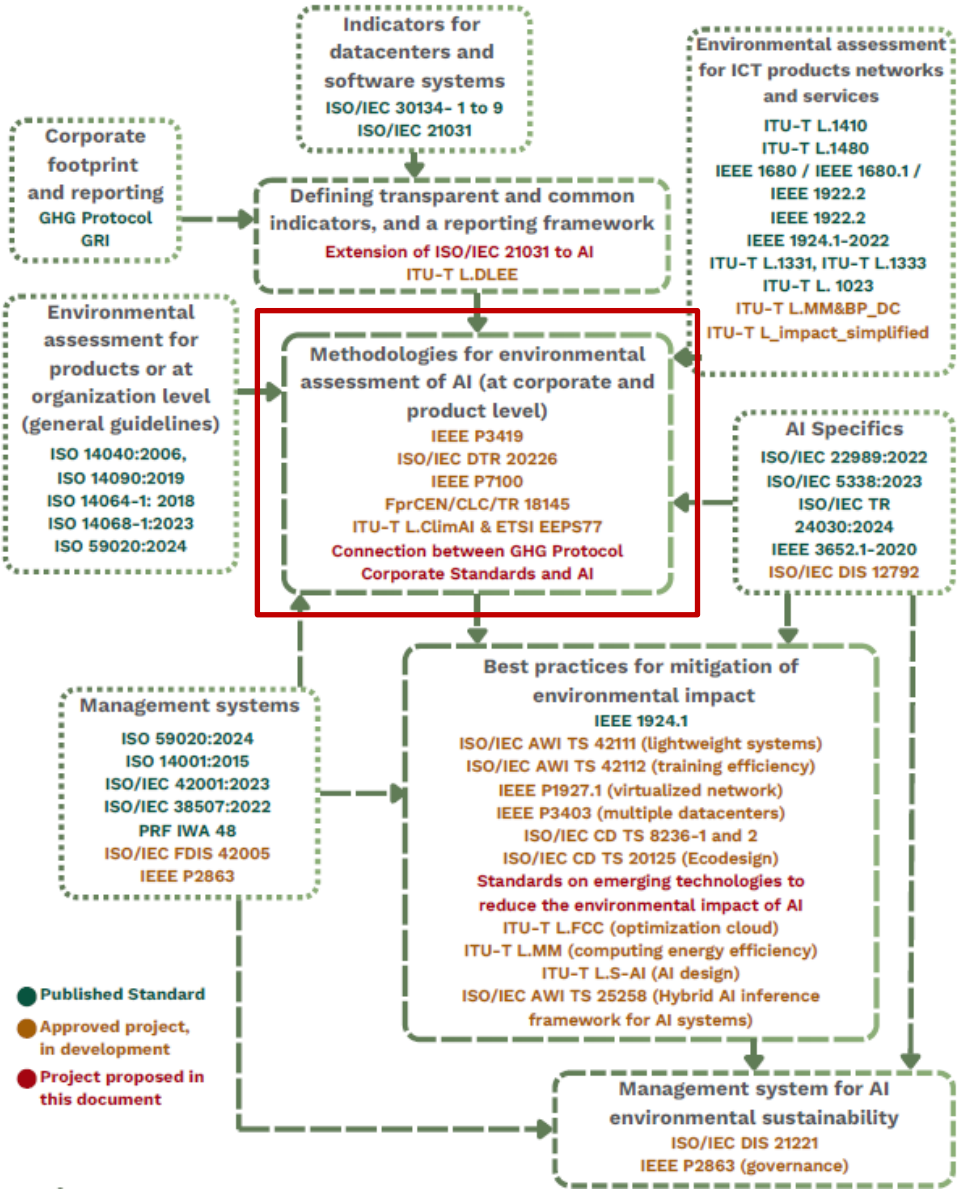


Figure 17 - Standards sur l'évaluation de l'impact environnemental des systèmes d'IA en cours ou publiés. Source : Coalition pour une IA durable, février 2026

Ce document a été réalisé par l'Arcep

Olivier Corolleur, directeur général
Rodolphe Le Ruyet, conseiller de la Présidente

DIRECTION « ÉCONOMIE, MARCHÉS ET NUMÉRIQUE »

Anne Yvrande-Billon, directrice

Blaise Soury-Lavergne, adjoint à la directrice

Unité « Analyse économique et Intelligence numérique »

Marion Panfili, cheffe d'unité

Chiara Caccinelli, adjointe à la cheffe d'unité

Charles Joudon-Watteau, chargé de mission

Tom Nico, chargé de mission

DIRECTION « INTERNET, DONNEES, PRESSE, POSTES ET UTILISATEURS »

Olivier Delclos, directeur

Unité « Internet ouvert »

Sandrine Elmi Hersi, cheffe de l'unité

Oriane Piquer-Louis, chargée de mission

Vivien Guéant, expert

Unité « Services de données et Cloud »

Léo Quentin, chef d'unité

Philéas Samir, chargé de mission

DIRECTION « MOBILE ET INNOVATION »

Franck Tarrier, directeur

Ahmed Haddad, conseiller technique

DIRECTION « EUROPE ET INTERNATIONAL »

Anne Lenfant, directrice

Unité « Europe »

Emmanuel Leroux, chef d'unité

Malo Le Cointe, chargé de mission

DIRECTION « COMMUNICATION ET PARTENARIATS »

Clémentine Beaumont, directrice

Victor Schmitt, chargé de mission

DIRECTION « AFFAIRES JURIDIQUES »

Élisabeth Suel, directrice

Unité « Marché mobile et Ressources rares »

Léa Ployaert, cheffe d'unité

Annabel Gandar, adjointe à la cheffe d'unité

Un grand merci à l'équipe du PEReN pour sa contribution !

Le Pôle d'expertise de la régulation numérique (PEReN) accompagne les services de l'Etat et les administrations indépendantes intervenant dans le champ de la régulation des plateformes numériques et de l'intelligence artificielle. Centre d'expertise technique, il mutualise des compétences à l'état de l'art en sciences des données, traitement et audit algorithmiques, évaluation de l'IA et développement logiciel pour produire des outils, des études et du conseil numérique. Il conduit également des projets de recherche publique exploratoire ou scientifique. Pour en savoir plus : www.peren.gouv.fr

Publication

Arcep
14, rue Gerty-Archimède
75012 Paris

Direction de la Communication et Partenariats : comm@arcep.fr

21 mai 2026

ISSN n° 2258-3106