Estimer et comprendre les impacts environnementaux d'un service d'IA générative

ACV d'un service d'IA générative

Adrien Berthelot

Eddy Caron
Mathilde Jay
Laurent Lefèvre

27.03.2024 | GreenDays 2024 @ Toulouse

ENS DE LYON

OCTO Technology
Part of Accenture

AVALON

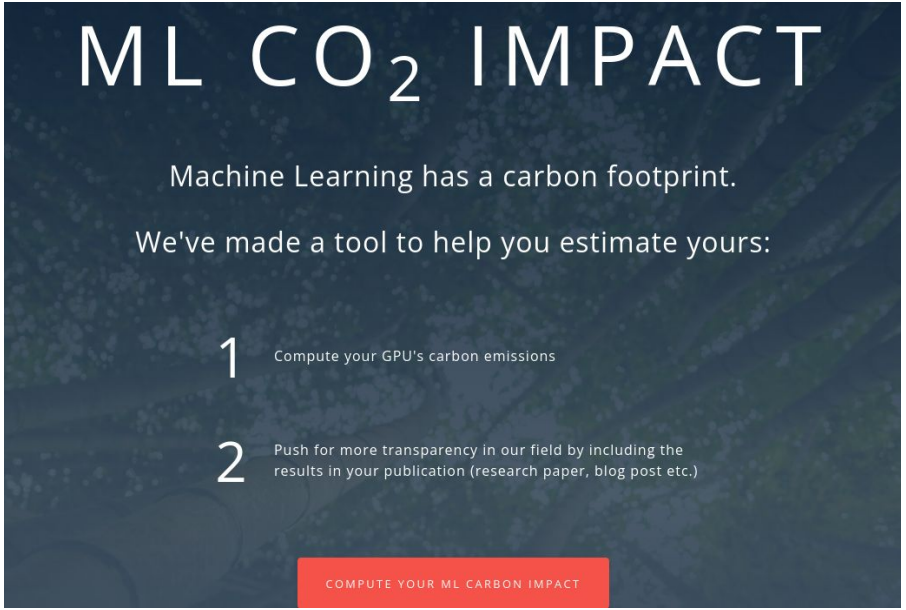
Lyon 1

Inria

Du côté de la recherche académique

Combien coûte l'IA générative ?

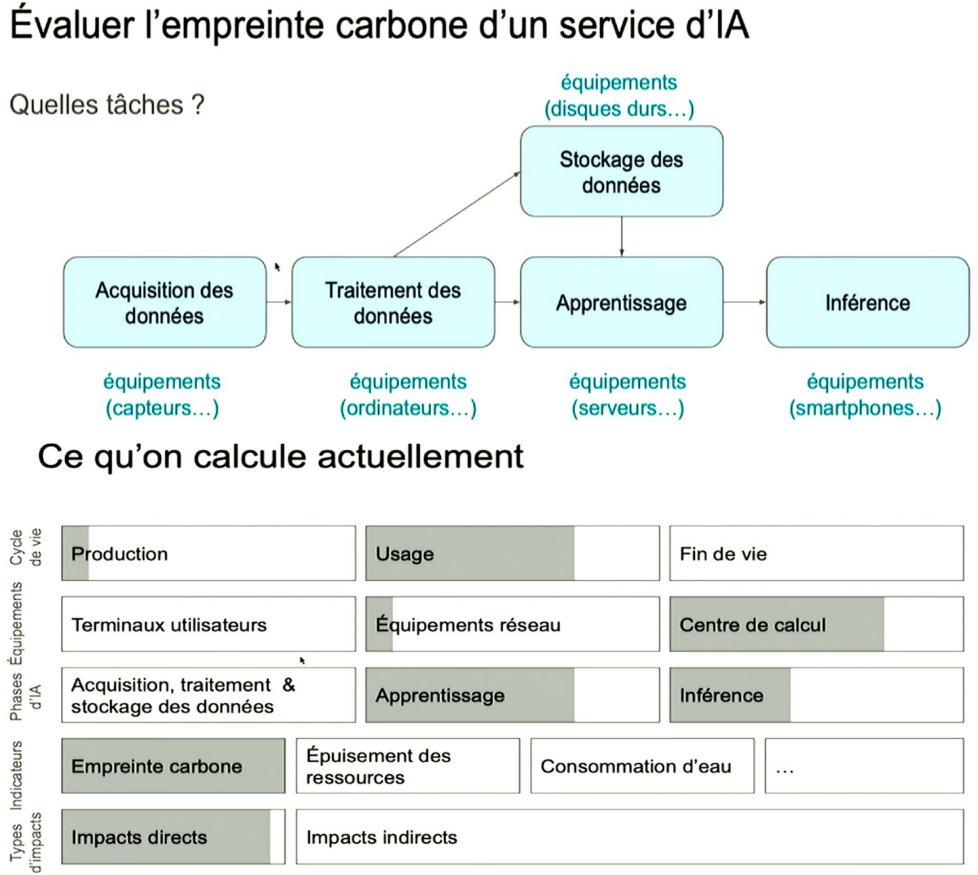
- Des premiers travaux dès 2019 dans la continuité des recherches sur la mesure de consommation électrique du numérique et notamment du *machine learning* comme le travail de *Lacoste et al. 2019* qui donne cet outil
- Des papiers très aboutis comme celui sur BLOOM de *Luccioni et al. 2023*
- Des articles issus de chercheurs de l'industrie qui se veulent rassurant avec des noms évocateurs comme "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink" - *Patterson et al. 2022*



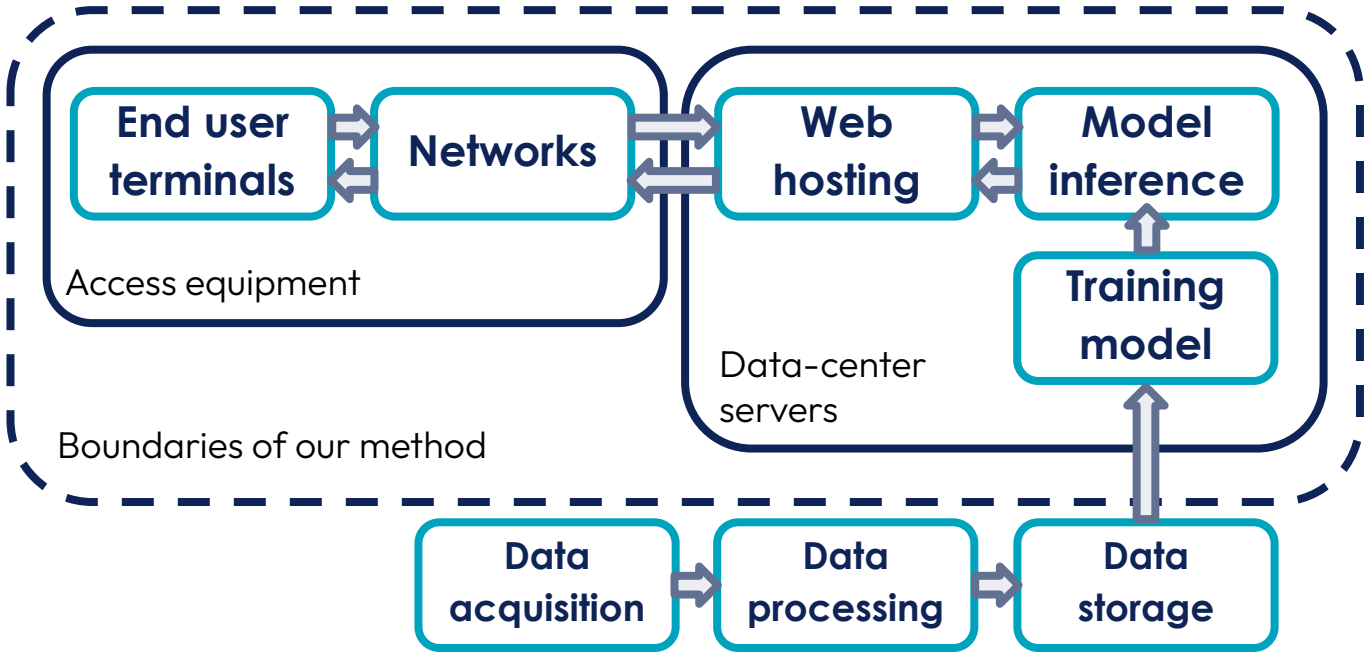
Au final, une vision du problème encore restreinte

- En réalité, la majorité des travaux se limitent au coût d'entraînement du modèle
  - > Éventuellement son coût carbone
  - > Éventuellement le cycle de vie des machines d'entraînement
- Pourtant l'inférence représenterait aujourd'hui de la consommation de ressources
- Une meilleure représentation du coût environnemental de l'IA générative nécessite la considérer intégrant un **service** aux **impacts divers**

Anne-Laure Ligozat (CNRS, Université Paris-Saclay) : "Côté obscur de l'IA : quels bénéfices réels de l'IA pour faire face aux crises environnementales ?" - Keynote au GreenDays 2023

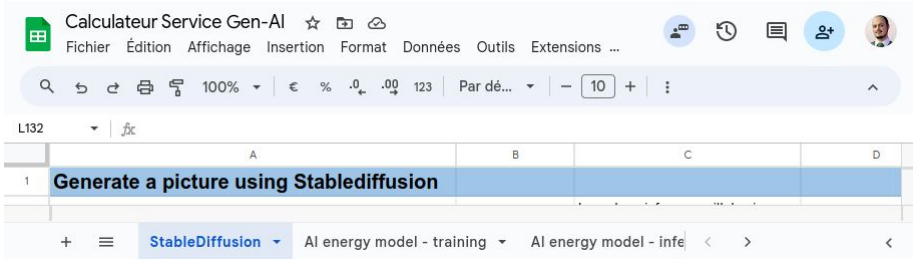


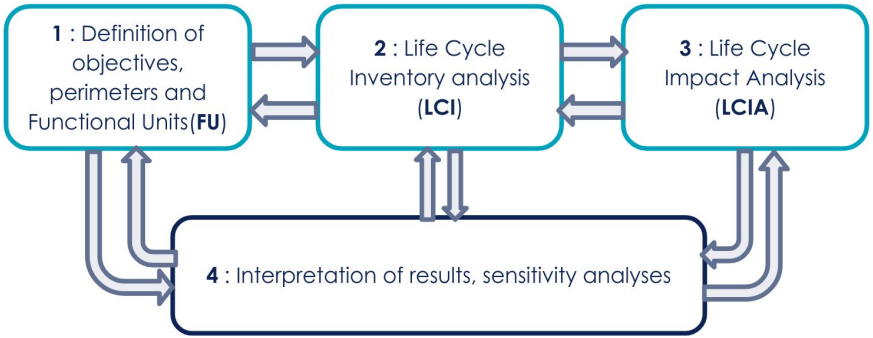
Le périmètre de notre étude



Outils

- Un GoogleSheet fait maison pour l'**ACV**
- Des données environnementales issues de NegaOctet/ADEME ou du projet Boavizta
- Des infrastructures de recherches (Grid'5000) équipées de Wattmètres physiques et logiciels
- Des outils de mesure web (Similarweb, HypeStats ou les outils dev firefox)





Méthode :

NEGA OCTET

Boavizta

Terminaux

Allocation temporelle :

$$\frac{\text{Temps d'usage device}}{\text{Temps d'usage total}}$$

- Estimation durée visite moyenne
- Mesure part smartphone

Réseaux

Allocation au Go

- Mesure du trafic moyen lié à une visite
- Estimation nombre de visites
- Facteur d'émission pour 1 Go

Serveur Web

Hypothèse de serveur dédié, allocation à l'usage

- Estimation nombre d'utilateur par durée et nombre visites
- Estimation empreinte du serveur

Inférence

Allocation temporelle :

$$\frac{\text{Temps d'usage GPU}}{\text{Temps d'usage total}}$$

- Mesure temps et conso inférence
- Estimation empreinte GPU

Entraînement

Allocation temporelle :

$$\frac{\text{Temps d'usage GPU}}{\text{Temps d'usage total}}$$

- Estimation temps et conso entraînement
- Estimation empreinte nœud

SimilarWeb

Grid'5000

Résultats

- FU 1 : Une visite et soumission d'un prompt sur stablediffusion.com générant en retour 4 petites images
- FU 2 : Une année d'hébergement de Stable Diffusion

Table 2. Environmental impact of Stable Diffusion for FU1 and FU2

FU	Abiotic Depletion Potential (kgSb eq)	Warming Potential (kgCO² eq)	Primary energy (MJ)
FU1 - Single use of service	$6.72e^{-08}$	$7.84e^{-03}$	$2.02e^{-01}$
FU2 - A year of service	$4.64e^{+00}$	$3.60e^{+05}$	$8.93e^{+06}$

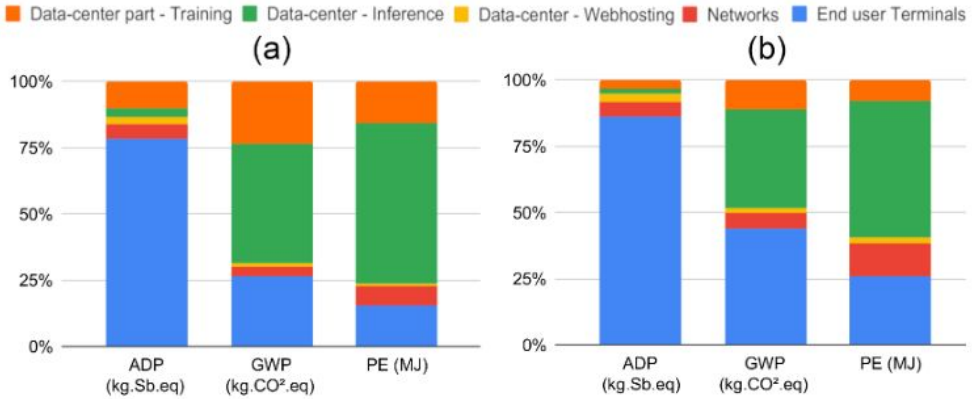


Fig. 2. Impact distributions for (a) FU1 and (b) FU2

Analyse de sensibilité

La notion critique de taux d'utilisation actif

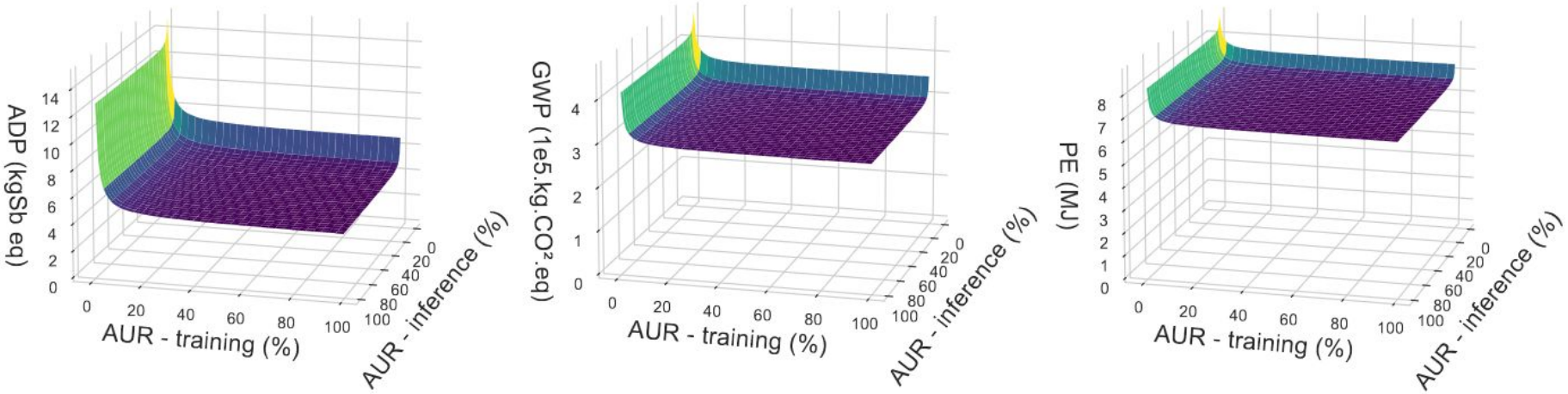


Fig. 3. Impact of the average active utilization rate (AUR) of data-center equipment

Analyse de sensibilité

Jusqu'où inclure le coût de l'entraînement ?

- Nous avons inclus dans l'étude de base uniquement le coût d'entraînement des versions 1.4 et 1.5 du modèle, celles utilisées - Scénario **S** standard
- Mais pour obtenir ces versions, il a fallu l'entraînement d'au moins les versions 1, 1.1, 1.2 et 1.3, ce coût est pris en compte dans le scénario **L** legacy

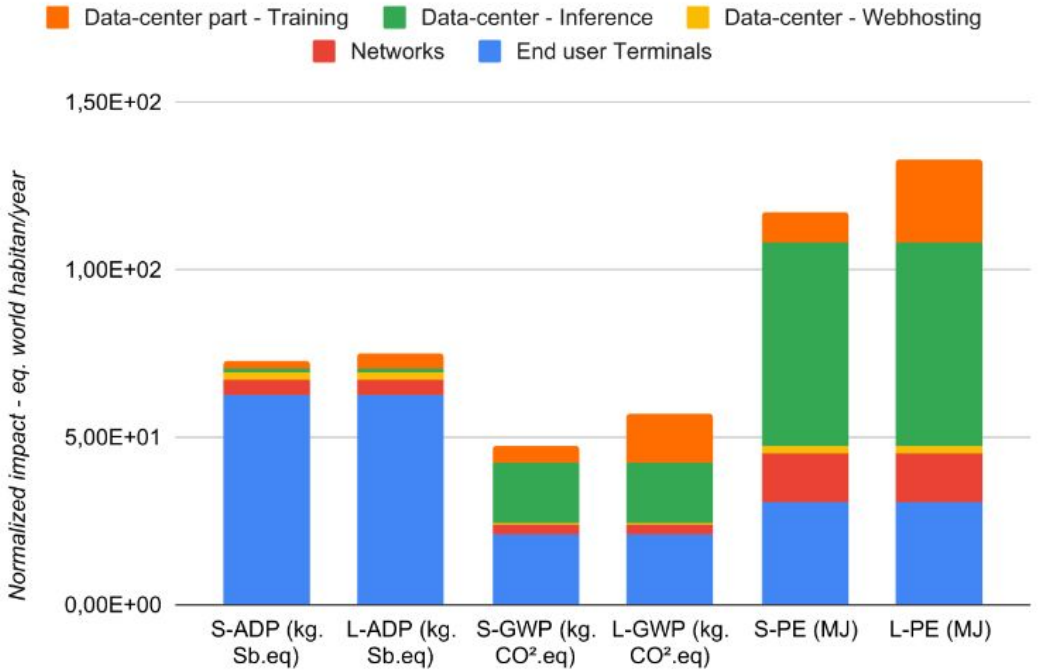
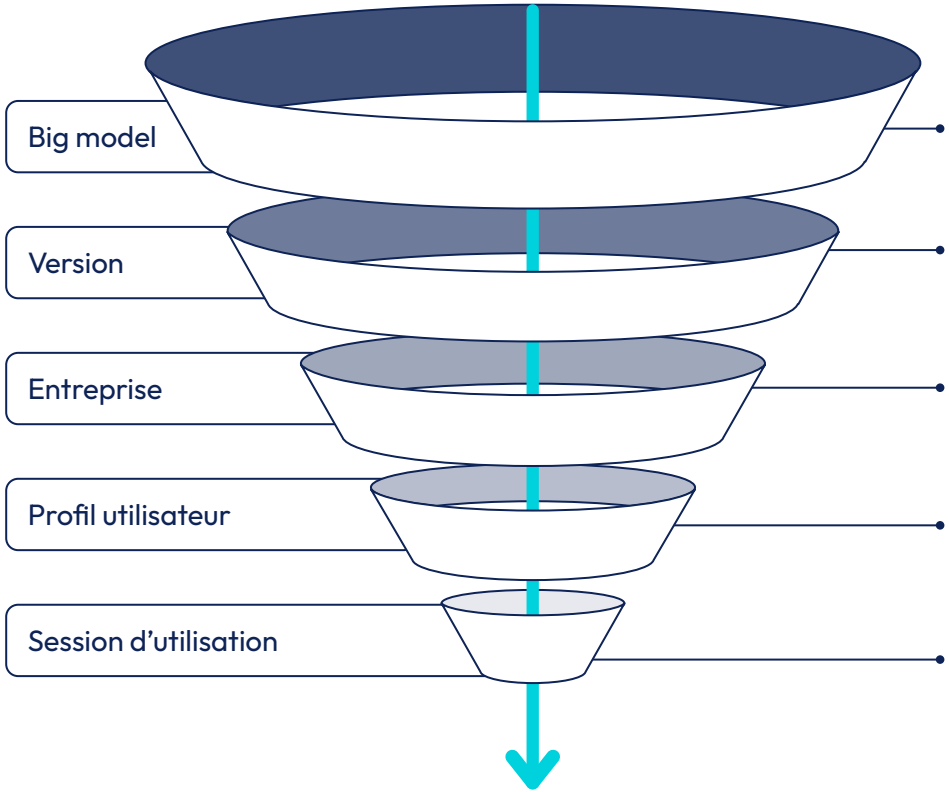


Fig. 4. Comparison between the standard scenario (S) and the full legacy scenario (L) on equivalent world habitant normalized impact for FU2

Principal limite actuelle de l'approche service

Les différentes couches d'entraînement jusqu'à l'utilisateur final



Notre périmètre actuel :

- Version initiale
- Sous-version du modèle

Déploiement du service avec entraînement sur données d'entreprises

AWS Bedrock

Entraînement supplémentaire dit "fine tuning", à l'échelle :

- Utilisateur
- Session d'utilisateur

Conclusions

- L'IA générative propose des services numériques particulièrement coûteux environnementalement.
- L'impact environnemental n'est pas concentré dans une unique partie et un unique impact.
- Une grande part des émissions de GES pourra être évitée mais cela sera insuffisant.
- La transformation des datacenter induite par la multiplication de ces services va engendrer de nombreux impacts de 2e et 3e ordre.
- Plus que l'IA générative en tant que technologie, c'est le déploiement rapide, croissant, incontrôlé de celle-ci comme service qui représente un problème pour l'environnement.
- Présentation basée sur l'article : *Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. CIRP LCE 2024 - 31st Conference on Life Cycle Engineering, Jun 2024, Turin, Italy. ⟨hal-04346102v2⟩*