Towards a multi-criteria evaluation of the environmental footprint of generative ai services

Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefèvre

TOWARDS A MULTI-CRITERIA EVALUATION OF THE ENVIRONMENTAL FOOTPRINT OF GENERATIVE AI SERVICES

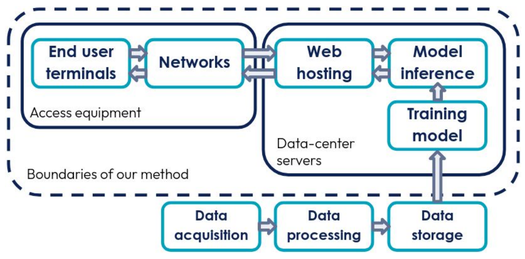
Adrien Berthelot[1,2], Eddy Caron[1], Mathilde Jay[1,3] and Laurent Lefèvre[1]

[1]Univ. Lyon 1 UCBL, EnsL, CNRS, Inria, LIP. Lyon, France
[2]OCTO Technology. Paris, France
[3]Univ. Grenoble Alpes, Grenoble INP, LIG. Grenoble, France

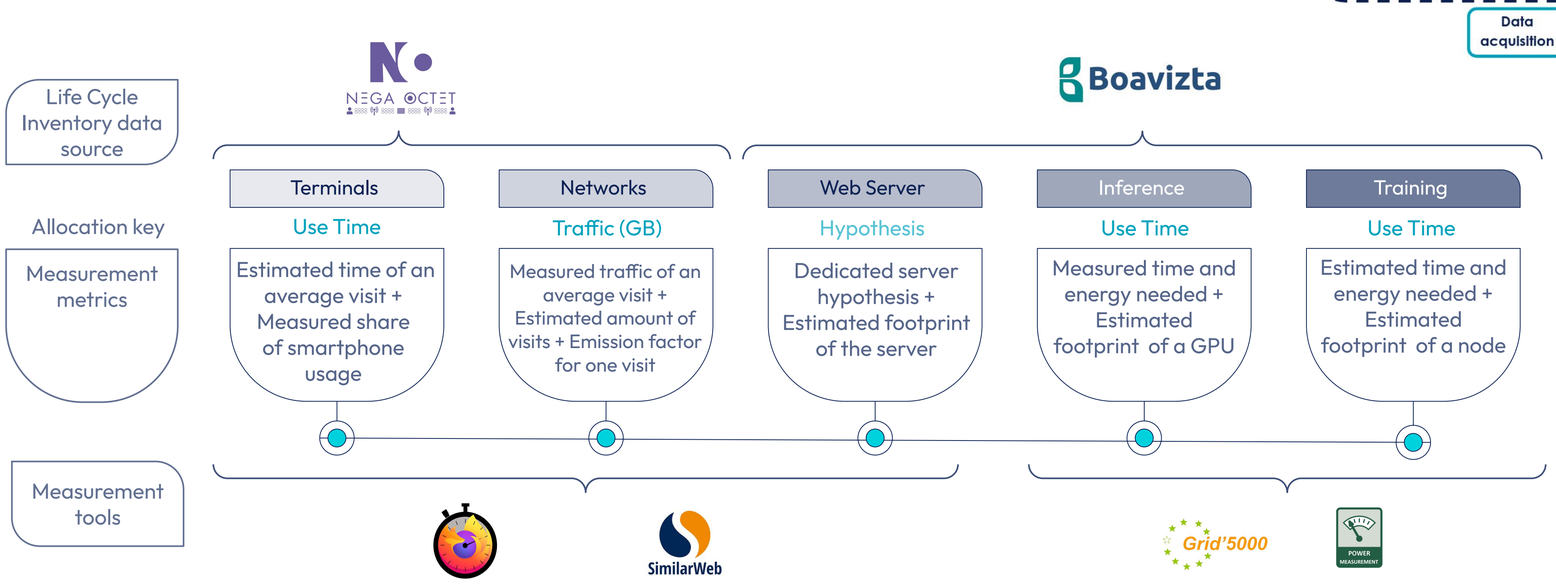
Introduction

Generative AI represents a new stage in digital transformation through its many applications. Unfortunately, by accelerating the growth of digital technology, Gen-AI is contributing to the **multiple environmental damages** caused by its sector.

We propose to study not only the impact of developing a model but also that of its deployment and use as a **service**. The figure at the left shows what we consider the standard structure of a Gen-AI service.
Our methodology (presented in the diagram below) is applied and validated on an AI service based on **Stable Diffusion**, an open-source text-to-image generative deep-learning model.



Boundaries of our methodology

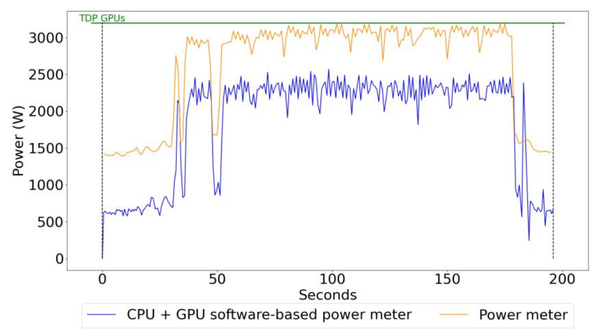


Training electricity cost

Replication is the most reliable method to estimate the electricity consumption. But for training, it would too expensive.

Our solution: replicate a fraction of it and use **linear regression to extrapolate** the results.

The figure below shows the difference between the methods used to estimate the electricity consumption. It leads to a **better understanding of electricity consumption.**



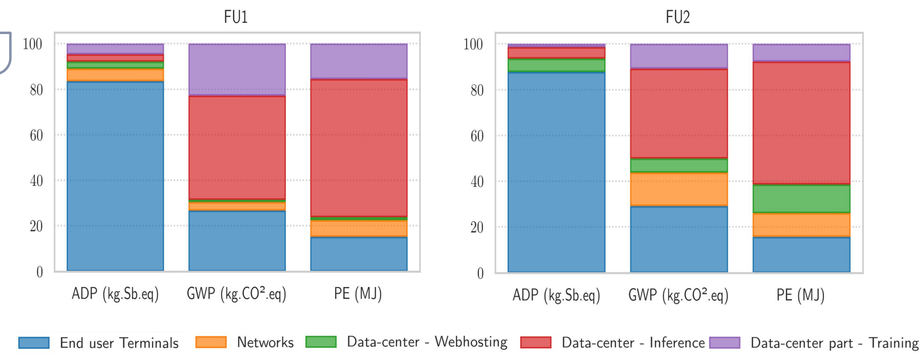
Electricity consumption measurement method comparison

Environmental footprint of the service

Functional Unit (FU)	Abiotic Depletion Potential (kg.Sb.eq)	Warming Potential (kg.CO2.eq)	Primary Energy (MJ)
FU1 - Average single use of service	$6.72e^{-8}$	$7.84e^{-3}$	$2.02e^{-1}$
FU2 - A year of service	4.64	$3.60e^{+5}$	$8.93e^{+6}$

Environmental impact of Stable Diffusion for FU1 and FU2

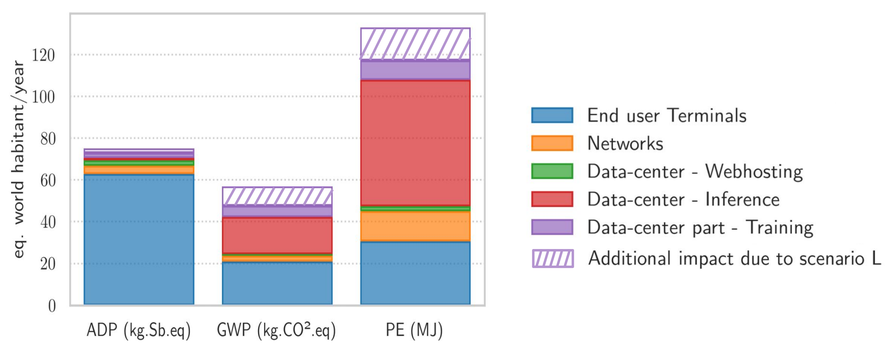
Impact distributions for FU1 and FU2

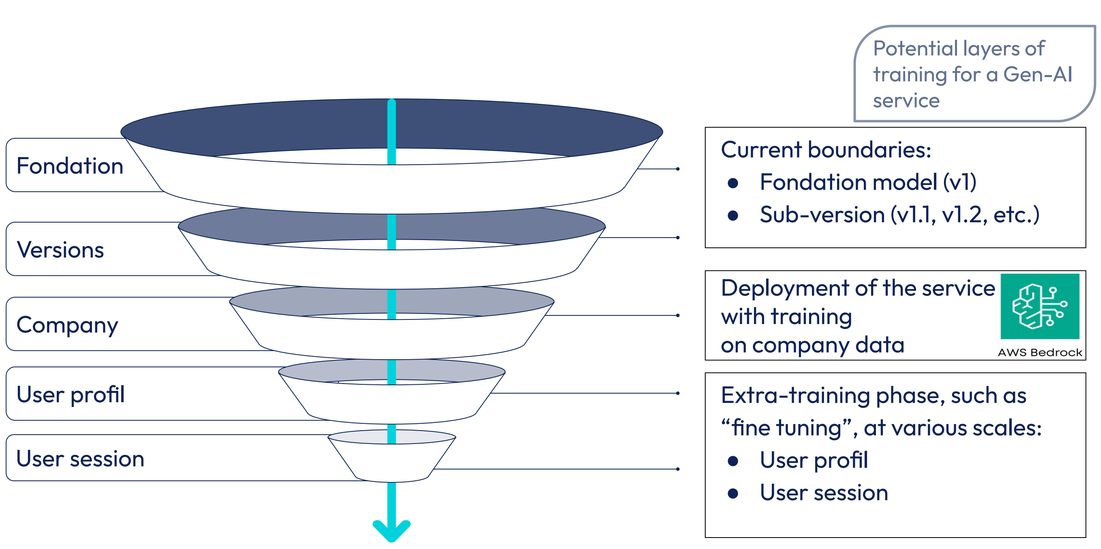


Sensitivity to the utilization rate and the scope of training

1. Average utilization rate (**AUR**) of data center equipment
   - AUR = percentage of time in the equipment lifespan during which it is actively used, as opposed to when it is either idle or on standby.
   - Crucial to the allocation for cost-heavy data center equipment.
   - Impact remains small if the AUR is higher than **20%** (is that reasonable?)
2. Scope of training phases taken into account.
   - Diagram below shows hierarchical **fine-tuning layers** in training. Our standard scenario (S) only includes versions v1.1 and v1.2.
   - We consider versions up to v1.5 in a full legacy scenario (L). Figure on the right shows the additional impact due to these additional versions.

Comparison between the standard scenario (S) and the full legacy scenario (L) for FU2





Potential layers of training for a Gen-AI service

Current boundaries:
- Fondation model (v1)
- Sub-version (v1.1, v1.2, etc.)

Deployment of the service with training on company data

AWS Bedrock

Extra-training phase, such as "fine tuning", at various scales:
- User profil
- User session

Conclusions

- Generative AI offers digital services that are particularly costly in environmental terms.
- The total footprint is not concentrated in a single part and a single impact.
- A large proportion of GHG emissions can be avoided, but this will **not be enough**.
- The transformation of data centers induced by the multiplication of these services will generate numerous impacts of 2^e and 3^e order.
- More than generative AI as a technology, it is the rapid, growing, uncontrolled **deployment** of it as a service that represents a problem for our environment.

Poster based on: *Adrien Berthelot, Eddy Caron, Mathilde Jay et Laurent Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology.* Procedia CIRP, vol. 122, pages 707–712, May 2024.
https://www.sciencedirect.com/science/article/pii/S2212827124001173

Contacts: adrien.berthelot@octo.com, mathilde.jay@univ-grenoble-alpes.fr, laurent.lefevre@ens-lyon.fr, eddy.caron@ens-lyon.fr