The environmental impact of Generative AI

Adrien Berthelot, **Mathilde JAY**,
Laurent Lefevre, Eddy Caron

Context: Generative AI

« An artificial intelligence capable of generating text, images, or other media »

- **1966:** ELIZA
  - Simulate a conversation
- Chat bots
- **2018:** Interest regain
  - Google, Open AI
- Summer **2022:** Explosion!
  - Text-to-Image: Stable Diffusion
  - Text-to-Text: ChatGPT

A photo of a lake on a sunny day, blue sky with clouds, beautiful, small reeds behind lake, bushes in the foreground, varied trees in the back, summer, 4k, Kieran Stone, Mandy Lea, Sapna Reddy, muted colors, nature photography

Context: AI & Energy globally

From the International Energy Agency (IEA)

- Alphabet (Google)
  - **10-15%** of it total energy consumption are related to AI workload (2019 - 2021)
  - In 2021: Total consumption of 12k GWh (From Statistica)
  - Growth: **20-25%** per year

- Meta & Google
  - **60-70%** Inference
  - **20-40%** Training

- Generative AI will accelerate this growth

State of the art: Impact of AI

- Focus on the **electricity consumption** of AI **training**
  - Eventually adding the carbon emissions
  - Eventually including the life cycle of equipment used
- **Inference** is starting to gain interest
- No study on the deployment of AI
- No other environment indicator than carbon and energy

Evaluation methodology adapted to Generative AI

At the service level

3 impact indicators
- Primary Energy
- Global Warming Potential
- Abiotic Depletion Potential

Plan

1. Estimation of the electricity consumption (Mathilde)

2. Life Cycle Assessment (LCA) based methodology for the estimation of the environmental impact of Generative AI (Adrien)

Use case : Stable Diffusion

- Text-to-Image
- Open-source

The electricity consumption of training a Generative AI model

Electricity consumption

Estimation

Based on Thermal Design Power (TDP)

Estimation requiring information on the training

+ Easy/accessible

- Not reliable
- Don’t take into account the whole server

Complete measure

Need to be done during the training or replicated

+ Accurate

- Not accessible

Estimation from measures

Replicate and monitor only part of the training

+ Accurate
+ Reproducible

- Not as accessible as TDP

Electricity consumption

- Definitions
  - A batch = Quantity of data that the algorithm processes at the same time
  - One step = Execution of the algorithm on a batch
  - Training = number of steps to achieve the desired quality of result

- Replicate a training step
  - Same settings/material as initial training
  - If multiple versions of the model, steps may be different

- Observations on different numbers of steps
  - Proof that the cost per step is identical

- Linear regression: $Energy = x * number\ of\ steps + y$

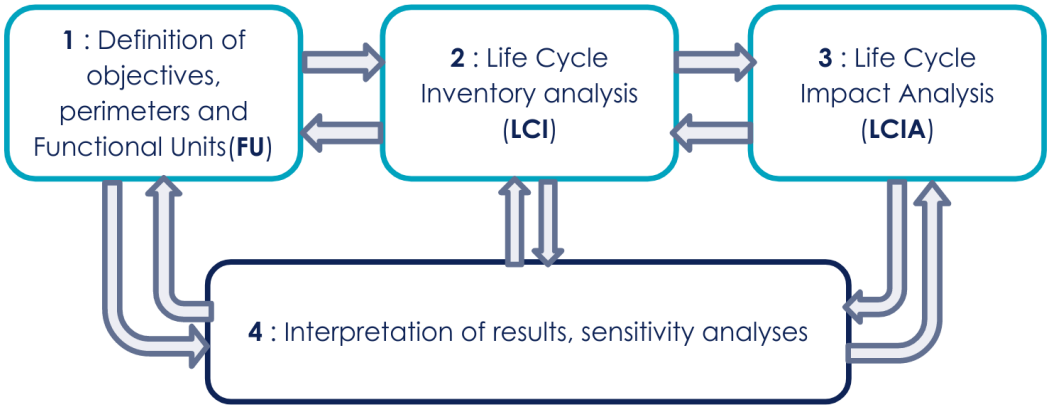
Electricity consumption

- Benefits
  - No need to replicate the entire training
  - More accurate estimation than TDP
  - Estimation of the different possible versions
  - Reproducible

- Drawbacks
  - Need to have access to a cluster identical to the initial training
  - Need to have access to open source code and model
  - Need information on the parameters used and the number of steps performed

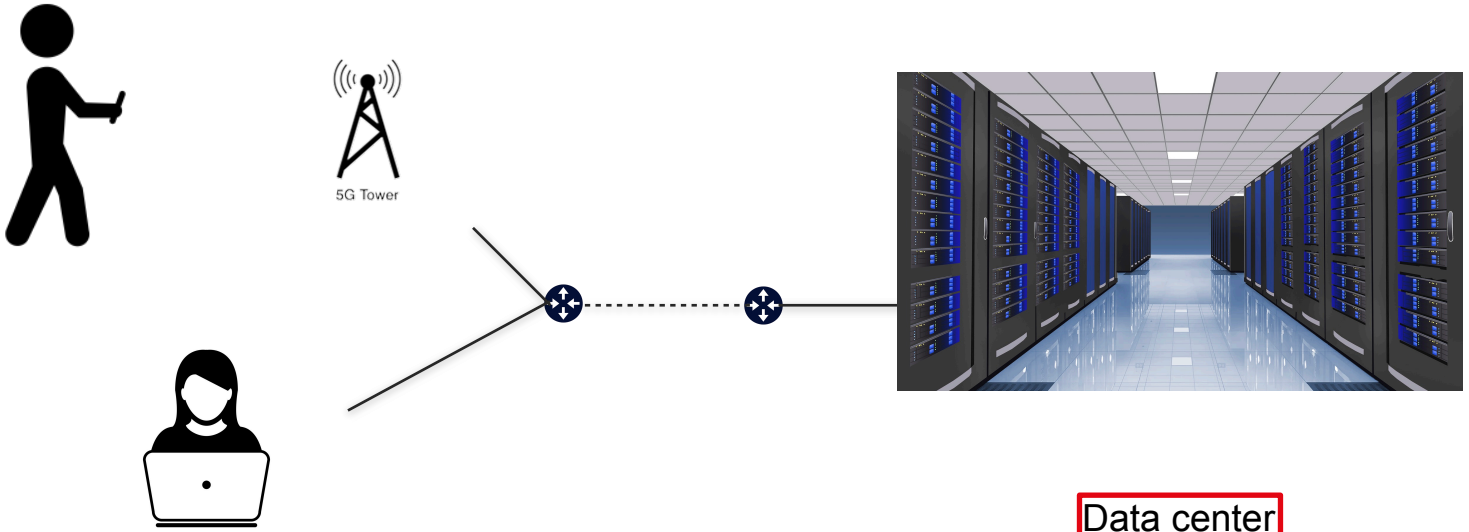
LCA-based methodology for the estimation of the environmental impact of a Generative AI service

Life Cycle Assessment methodology



- FU 1
  - A single use of the service
- FU 2
  - One year of hosting the service
- Tools
  - ADEME / NegaOctet
  - Boavista
  - Statistic tools

Gen AI services



End user terminal

Networks

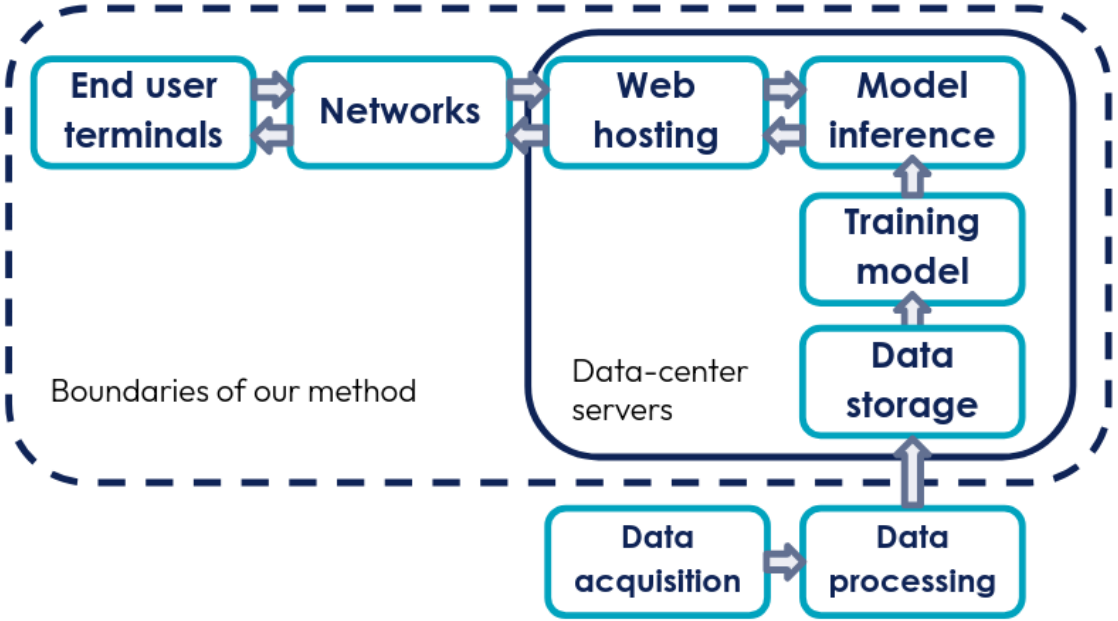
Data center

Web Hosting

Training

Inference

Our method boundaries



End-user terminals

Average electricity consumption

Impact for one type of terminal →

$$I_{term} = I_{term,footprint} + C_{term} \times I_{elec}$$

3 minutes

Allocation for the service →

$$a_{term} = \frac{\Delta t}{AUR \times T_{term}}$$

Total Lifetime

Active Utilisation Rate

Impact of terminals for the service →

$$I_{Terminals} = \sum_{term} s_{term} \times a_{term} \times I_{term}$$

Share

Networks

Quantity of data transferred (Go)

6 Mo

$$I_{Networks} = \sum_{network} s_{network} \times V_{data} \times I_{network}$$

40% mobile (smartphone, laptop)
60% fixed

Web Hosting

Average electricity consumption for 1 year

$$I_{server} = I_{server,footprint} + C_{server} \times I_{elec} \times PUE$$

With Boavista LCA
1 GPU A100

$$a_{server} = \frac{V_{visit}}{V_{total}}$$

Power Usage Effectiveness

$$I_{WebHosting} = \sum_{server} s_{server} \times a_{server} \times I_{server}$$

1 web server (AWS instance)

Inference

Experimentally measured electricity consumption

One GPU dedicated to the inference

$$I_{proc,elec} = C_{proc,infer} \times I_{elec} \times PUE$$

$$a_{proc} = \frac{\Delta t}{AUR \times T_{proc}}$$

Experimentally measured duration

$$I_{Inference} = V_{infer} \times (I_{proc,elec} + a_{proc} \times I_{proc,footprint})$$

Training

Estimated electricity consumption for 1 GPU

$$I_{proc,elec} = C_{proc,training} \times I_{elec} \times PUE$$

$$a_{proc} = \frac{\Delta t}{AUR \times T_{proc}}$$

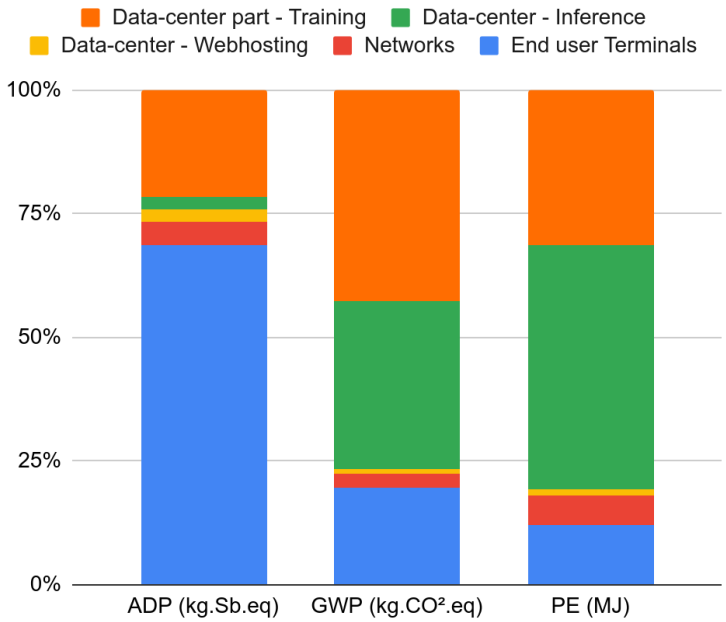
← Estimated duration

$$I_{Training} = \sum_{proc} I_{proc,elec} + a_{proc} \times I_{proc,footprint}$$

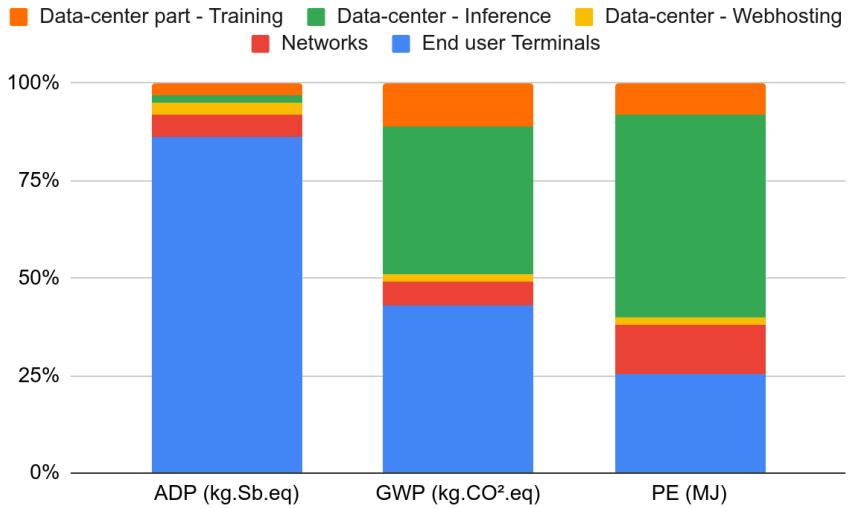
32 nodes of 8 GPUs (Nvidia A100)

Results for Stable Diffusion

For a single visit



For a year of service



- Around **5700 smartphones** in terms of rare metals
- **356 Paris-NewYork** person travelling by plane in CO2
- More than **1000 meters of wood energy**: Approximately what a wooden Eiffel Tower is and burn it

Conclusion

- Methodology with focus on completeness and accuracy

- Applied on Stable Diffusion
  - The impact of training is not negligible
  - Inference the most significant impact
  - Carbon is not the only impact

Thank you for your attention