Towards a multi-criteria evaluation of the environmental footprint of generative ai services
Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefèvre

To cite this version:
Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefèvre. Towards a multi-criteria evaluation of the environmental footprint of generative ai services. ICT4S 2024 - International Conference on Information and Communications Technology for Sustainability, Jun 2024, Stockholm, Sweden. pp.1-1, 2024. hal-04586653

HAL Id: hal-04586653
https://inria.hal.science/hal-04586653
Submitted on 24 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License
TOWARDS A MULTI-CRITERIA EVALUATION OF THE ENVIRONMENTAL FOOTPRINT OF GENERATIVE AI SERVICES

Adrien Berthelot\textsuperscript{1,2}, Eddy Caron\textsuperscript{1}, Mathilde Jay\textsuperscript{3,5} and Laurent Lefèvre\textsuperscript{1}

\textsuperscript{1}Univ. Lyon 1 UCL, EnsL, CNRS, Inria, LIP, Lyon, France
\textsuperscript{2}OCTO Technology, Paris, France
\textsuperscript{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.

Environmental footprint of the service

<table>
<thead>
<tr>
<th>Functional Unit (FU)</th>
<th>Abiotic Depletion Potential (kg.Sb.eq)</th>
<th>Warming Potential (kg.CO$_2$.eq)</th>
<th>Primary Energy (MJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FU1 - Average single use of service</td>
<td>6.72e$^6$</td>
<td>7.84e$^3$</td>
<td>2.02e$^1$</td>
</tr>
<tr>
<td>FU2 - A year of service</td>
<td>4.64</td>
<td>3.60e$^5$</td>
<td>8.93e$^6$</td>
</tr>
</tbody>
</table>

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.
   - 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.

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 GtG 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$^\text{nd}$ and 3$^\text{rd}$ 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.