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 Berthelot1,2, Eddy Caron1, Mathilde Jay1,3 and Laurent Lefèvre1

1Univ. Lyon 1 UCBL, EnsL, CNRS, Inria, LIP, Lyon, France
2OCTO Technology. Paris, France
3Univ. 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.

1. Average utilization rate (AUR) of data center equipment
2. Scope of training phases taken into account.

Sensitivity to the utilization rate and the scope of training

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 GtCO2 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 2nd and 3rd 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.

Functional Unit (FU)

<table>
<thead>
<tr>
<th>FU 1 - Average single use of service</th>
<th>FU 2 - A year of service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abiotic Depletion Potential (kg.Sb.eq)</td>
<td>6.72e-6</td>
</tr>
<tr>
<td>Warming Potential (kg.CO2.eq)</td>
<td>7.84e-3</td>
</tr>
<tr>
<td>Primary Energy (MJ)</td>
<td>2.02e-1</td>
</tr>
<tr>
<td>4.64</td>
<td>3.60e-5</td>
</tr>
<tr>
<td>8.93e-6</td>
<td></td>
</tr>
</tbody>
</table>

Environmental impact of Stable Diffusion for FU1 and FU2

Impact distributions for FU1 and FU2

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