Evaluating Large Language Models

Aurélie Névéol

aurelie.neveol@lisn.fr

March 24, 2025

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
cnrs
université PARIS-SACLAY

Computer Science lab:
Count words in a (small) text collection

Do you:

A Open files sequentially and do a manual count

B Use shell scripting

C Write a Perl/Python/... program

D Prompt chatGPT to write a Python program

Daily life:
Go to Roissy Charles de Gaulle from campus

Do you:

A Walk

B Take RER B all the way

C Drive

D Catch a plane at Orly

Summary

Evaluation should cover
quality + impactS

Evaluation in NLP is designed to foster reproducibility

Shared tasks: task definition, annotated dataset, metrics

NTCIR 2023 task: information extraction from social media

JA

アザチオプリン (イムラン) の副作用で脱毛がひどい。#潰瘍性大腸炎 <url>

EN

Severe hair loss due to azathioprine (Imuran) side effects. #Ulcerative colitis <url>

DE

Azathioprin (Imuran) Nebenwirkungen von schwerem Haarausfall. #Colitis ulcerosa <url>.

FR

Effets secondaires de l'azathioprine (Imuran) sur la perte sévère de cheveux. #Colite ulcéreuse <url>.

Annotated corpus in 4 languages

Table 10: Results of the Exact Match Accuracy for teams each language track.

Team	Japanese	English	German	French
AILABUD	0.75	0.71	0.71	0.67
FRAG	0.86	0.84	0.83	0.83
HPIDHC	0.87	0.85	0.85	0.84
IMNTPU	–	0.82	–	–
SRCB	**0.88**	**0.87**	**0.86**	**0.87**
STIS	–	0.82	–	–
TMUNLP	–	0.83	–	–
VLP	0.85	0.84	0.82	0.83
$\text{Baseline}_{\text{XLM-R}_{\text{ALL}}}$	0.84	0.83	0.80	0.81

Raithel L, Yeh HS, Yada S, Grouin C, Lavergne T, Névéol A, Paroubek P, Thomas P, Nishiyama T, Möller S, Aramaki E, Matsumoto Y, Roller R, Zweigenbaum P. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages. LREC-COLING 2024. 2024:395-414

Sample results on a "language understanding" task

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	**73.9**	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	**77.0**	**55.6**	**81.0**	69.6	**69.3**
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4

Table 9: **Massive Multitask Language Understanding (MMLU).** Five-shot accuracy.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. 2023. LLaMA: Open and Efficient Foundation Language Models https://ar5iv.labs.arxiv.org/html/2302.13971

Sample results on a "language understanding" task

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	**73.9**	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	**77.0**	**55.6**	**81.0**	69.6	**69.3**
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4
Random baseline	-	25.0	25.0	25.0	25.0	25.0

Table 9: **Massive Multitask Language Understanding (MMLU).** Five-shot accuracy.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. 2023. LLaMA: Open and Efficient Foundation Language Models https://ar5iv.labs.arxiv.org/html/2302.13971

Is there a baseline ?

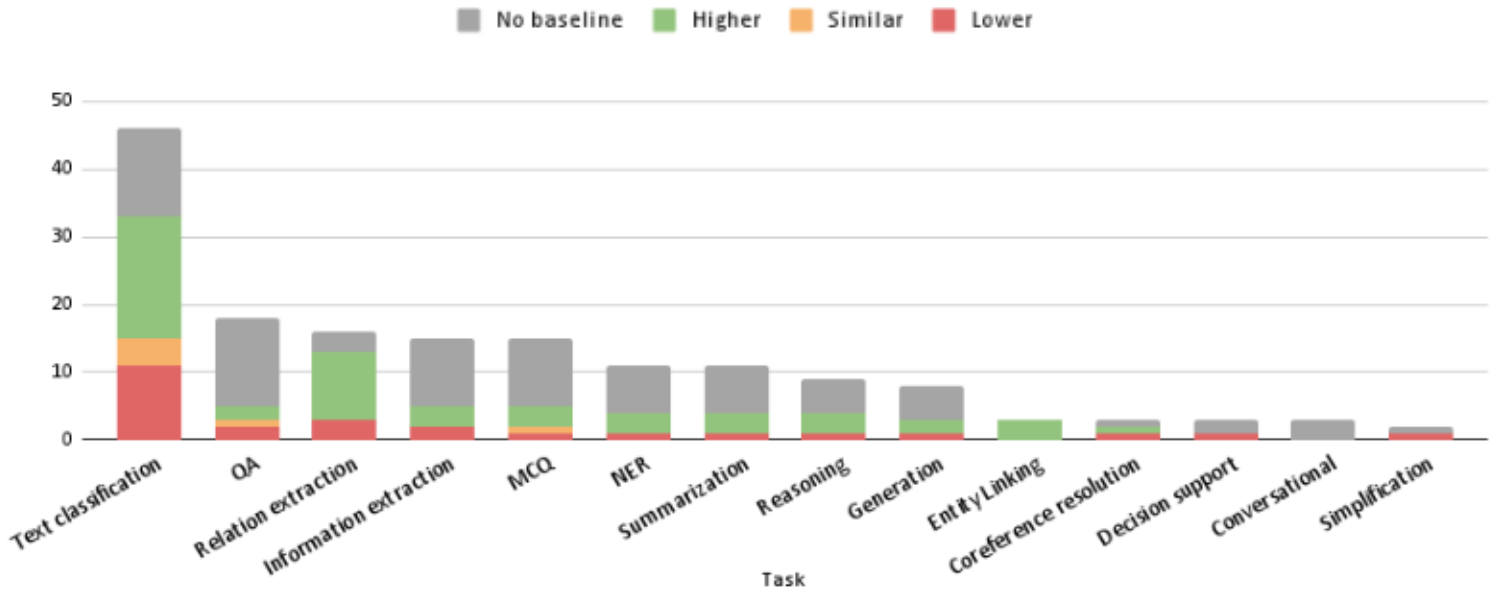


Figure 8: Baseline reports among (a) prompt categories, (b) venues, and (c) addressed NLP tasks. Higher/lower indicates that the performance of the proposed prompt-based approach is higher/lower than the baseline.

Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review and recommendations for better practices JMIR. 2024;26:e60501

Which task(s) are evaluated?

- There are many "benchmarks" out there
  - Are all the tasks relevant in aggregated collections ?

SUPER-NATURALINSTRUCTIONS:
Generalization via Declarative Instructions on 1600+ NLP Tasks

♢Yizhong Wang[2] ♢Swaroop Mishra[3] ♣Pegah Alipoormolabashi[4] ♣Yeganeh Kordi[5]
Amirreza Mirzaei[4] Anjana Arunkumar[3] Arjun Ashok[6] Arut Selvan Dhanasekaran[3]
Atharva Naik[7] David Stap[8] Eshaan Pathak[9] Giannis Karamanolakis[10] Haizhi Gary Lai[11]
Ishan Purohit[12] Ishani Mondal[13] Jacob Anderson[3] Kirby Kuznia[3] Krima Doshi[3] Maitreya Patel[3]
Kuntal Kumar Pal[3] Mehrad Moradshahi[14] Mihir Parmar[3] Mirali Purohit[15] Neeraj Varshney[3]
Phani Rohitha Kaza[3] Pulkit Verma[3] Ravsehaj Singh Puri[3] Rushang Karia[3] Shailaja Keyur Sampat[3]
Savan Doshi[3] Siddhartha Mishra[16] Sujan Reddy[17] Sumanta Patro[18] Tanay Dixit[19] Xudong Shen[20]
Chitta Baral[3] Yejin Choi[1,2] Noah A. Smith[1,2] Hannaneh Hajishirzi[1,2] Daniel Khashabi[21]

[1]Allen Institute for AI [2]Univ. of Washington [3]Arizona State Univ. [4]Sharif Univ. of Tech. [5]Tehran Polytechnic [6]PSG College of Tech. [7]IIT Kharagpur
[8]Univ. of Amsterdam [9]UC Berkeley [10]Columbia Univ. [11]Factored AI [12]Govt. Polytechnic Rajkot [13]Microsoft Research [14]Stanford Univ. [15]Zycus Infotech
[16]Univ. of Massachusetts Amherst [17]National Inst. of Tech. Karnataka [18]TCS Research [19]IIT Madras [20]National Univ. of Singapore [21]Johns Hopkins Univ.

Task114: the given word

Definition: In this task, you need to answer 'Yes' if the given word is the longest word (in terms of number of letters) in the given sentence, else answer 'No'. Note that there could be multiple longest words in a sentence as they can have the same length that is the largest across all words in that sentence.

Input: Sentence: 'a man is surfing on a crashing wave.'. Is **'a'** the longest word in the sentence? **Output**: No

Input: Sentence: 'a man is riding on the back of an elephant'. Is **'is'** the longest word in the sentence? **Output**: No

Illustration: P. Langlais

Sociology research shows that users are actors

example of transfer (*déplacement*)

source

source

$\Rightarrow$ we cannot predict all of the uses of a tool

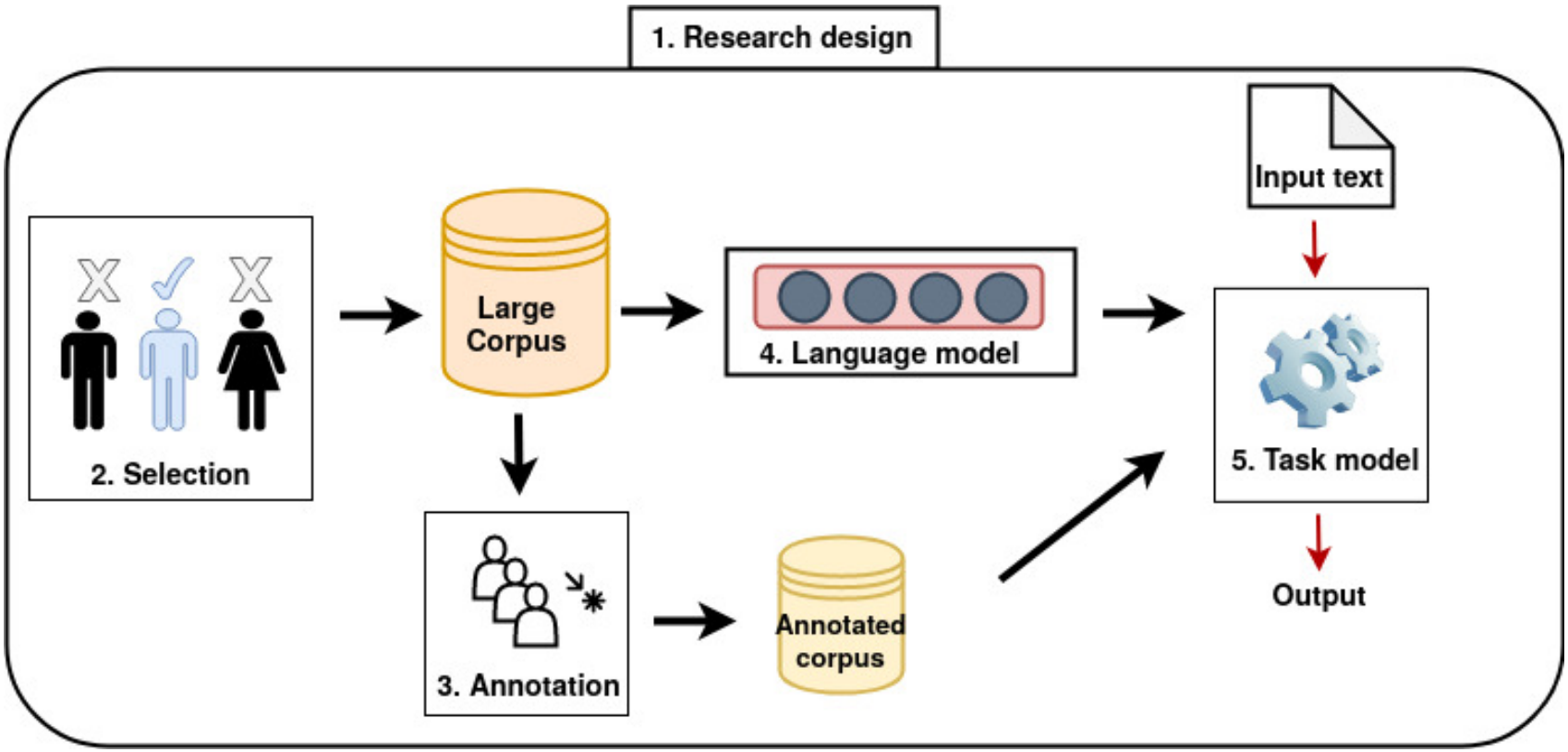
Slide credit: K. Fort

Akrich, M. (2006). Sociologie de la traduction : Textes fondateurs, chapitre - Les utilisateurs, acteurs de l'innovation. Presses des Mines.

Evaluation is more than a measure of task performance

Illustration: adapted from F. Ducel

Five sources of bias in Natural Language Processing



Hovy D, Prabhumoye S. (2021). Five sources of bias in natural language processing. Language and Linguistics Compass, e12432. https://doi.org/10.1111/lnc3.12432

Bias in research design

Problem statement:
How to best use an LLM for my problem?

AFNOR Specification for "Frugal AI"

31 recommendations including

- AI solutions should be as efficient as possible
- Benefits of using an AI system rather than another solution are shown
- Uses and needs are intended to remain within planetary boundaries

source

Measuring the environnemental impact of a language model

Need to account for:

- life-cycle of models: training, fine-tuning, distillation, inference, ...
- hardware equipment
- life-cyle of hardware

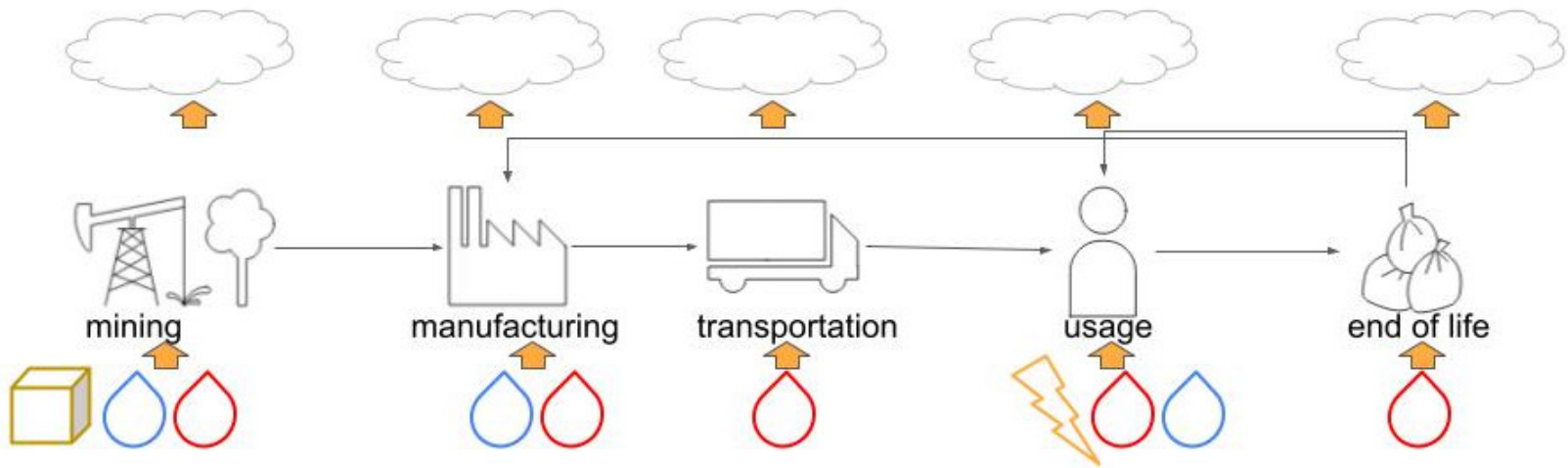


figure adapted from J Combaz and A-L Ligozat

Review of 85 publications to identify 6 tools for CO2 impact measurement

- Online tools
  1. Green Algorithms
  2. ML CO2 Impact
- Python toolkits
  2'. Code Carbon
  3. Energy Usage
  4. Experiment Impact Tracker
  5. Carbon Tracker
  6. Cumulator

+ MLCA https://github.com/blubrom/MLCA
+ Ecologits https://ecologits.ai

Bannour N, Ghannay S, Névéol A, Ligozat AL. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. ACL Workshop SustainNLP 2021:11-21
Morand C, Névéol A, Ligozat AL. MLCA: a tool for Machine Learning Life Cycle Assessment. ICT4S 2024

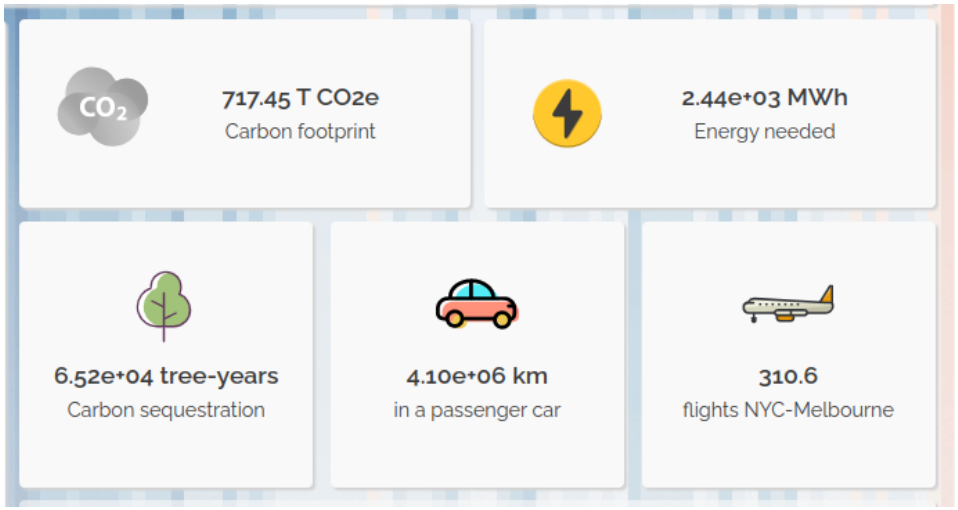
Features of measurement tools

Feature	**online**	**toolkit**
direct measure	X	✓∼
estimation	✓	X
asynchronous	✓	X
comparison on same hardware	∼	✓
easy to install	✓	∼

What is the environemental impact of chatGPT?

Training

- Data is hard to find!
  - OpenAI is estimated to have used 3,617 NVIDIA A100 HGX GPUs for [90-100] days on Azure cloud for training chatGPT



http://calculator.green-algorithms.org/
https://semianalysis.com/2023/02/09/the-inference-cost-of-search-disruption/

What is the environemental impact of chatGPT?

(and other models) - Usage

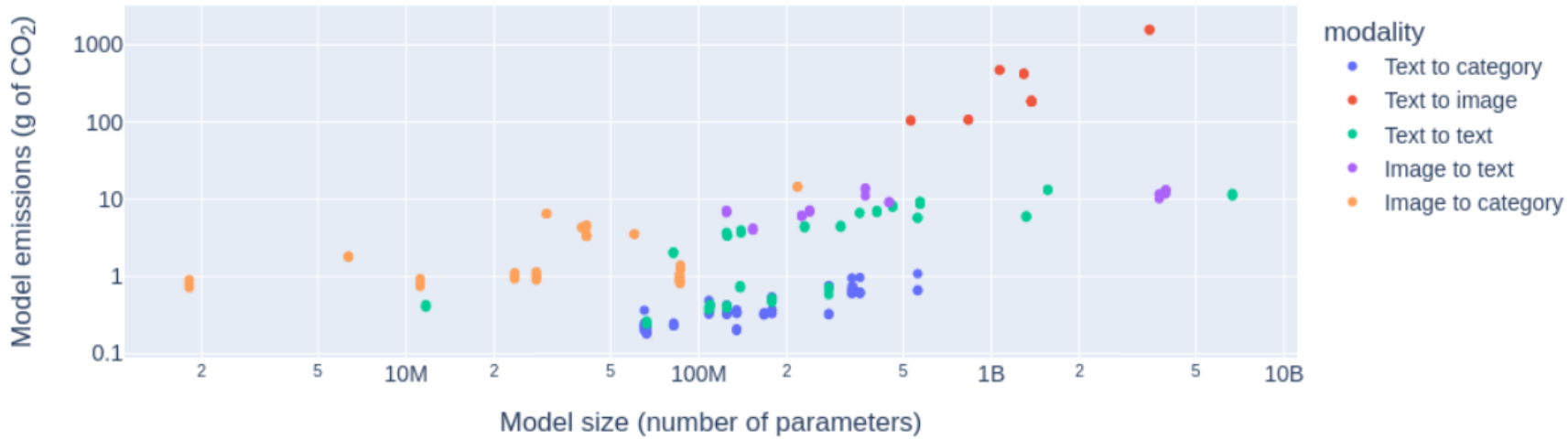


Figure 2: The 5 modalities examined in our study, with the number of parameters of each model on the x axis and the average amount of carbon emitted for 1000 inferences on the y axis. NB: Both axes are in logarithmic scale.

Luccioni S, Jernite Y, Strubell E. 2024. Power hungry processing: Watts driving the cost of ai deployment?. Proc. ACM conference on fairness, accountability, and transparency (pp. 85-99).

What does this impact mean?

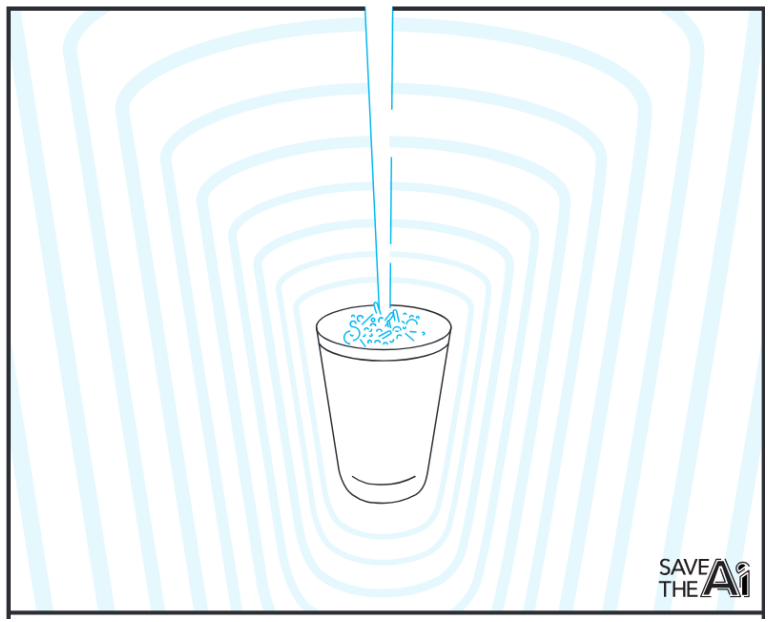
- 718 T CO2 = yearly target impact for 359 people according to Paris agreement

OpenAI report the use of **30 000 A100 GPUs** for keeping its AI up and running
(**Jean Zay** boasts 1,456 H100 GPUs and 416 A100 GPUs)

What does this impact mean?

Water consumption on the rise

Using chatGPT to write a 100 word email or answer 10 queries requires 500 ml water

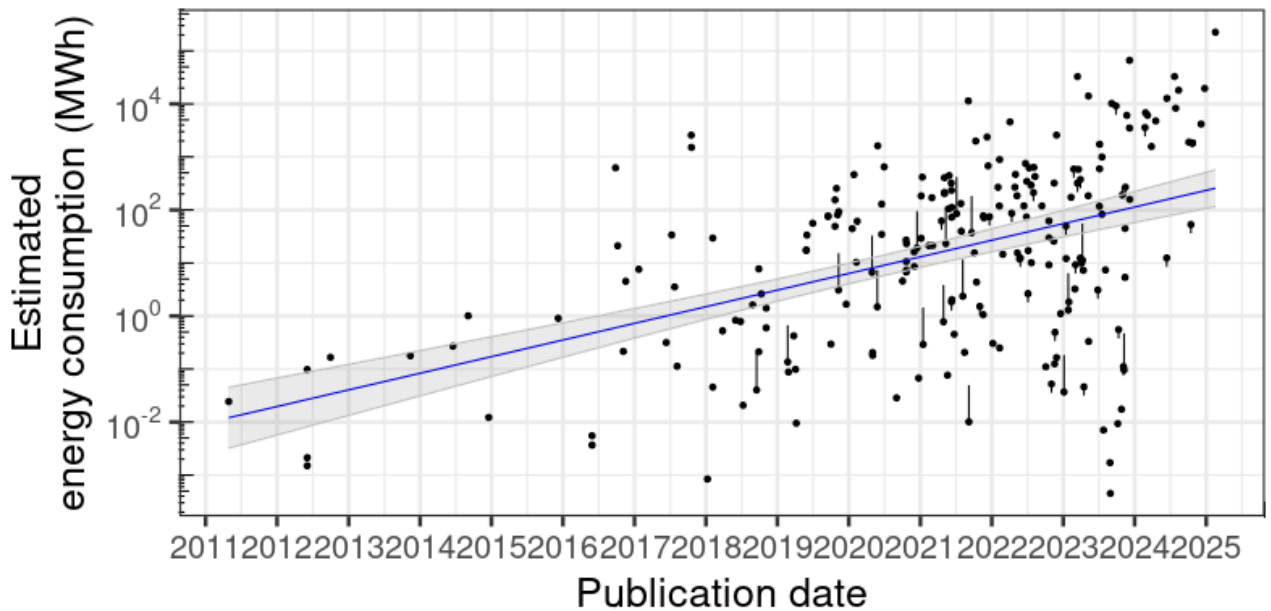


source

Li P, Yang J, Islam MA, Ren S. (2023). Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models arXiv.2304.03271

What does this impact mean?

Energy consumption on the rise

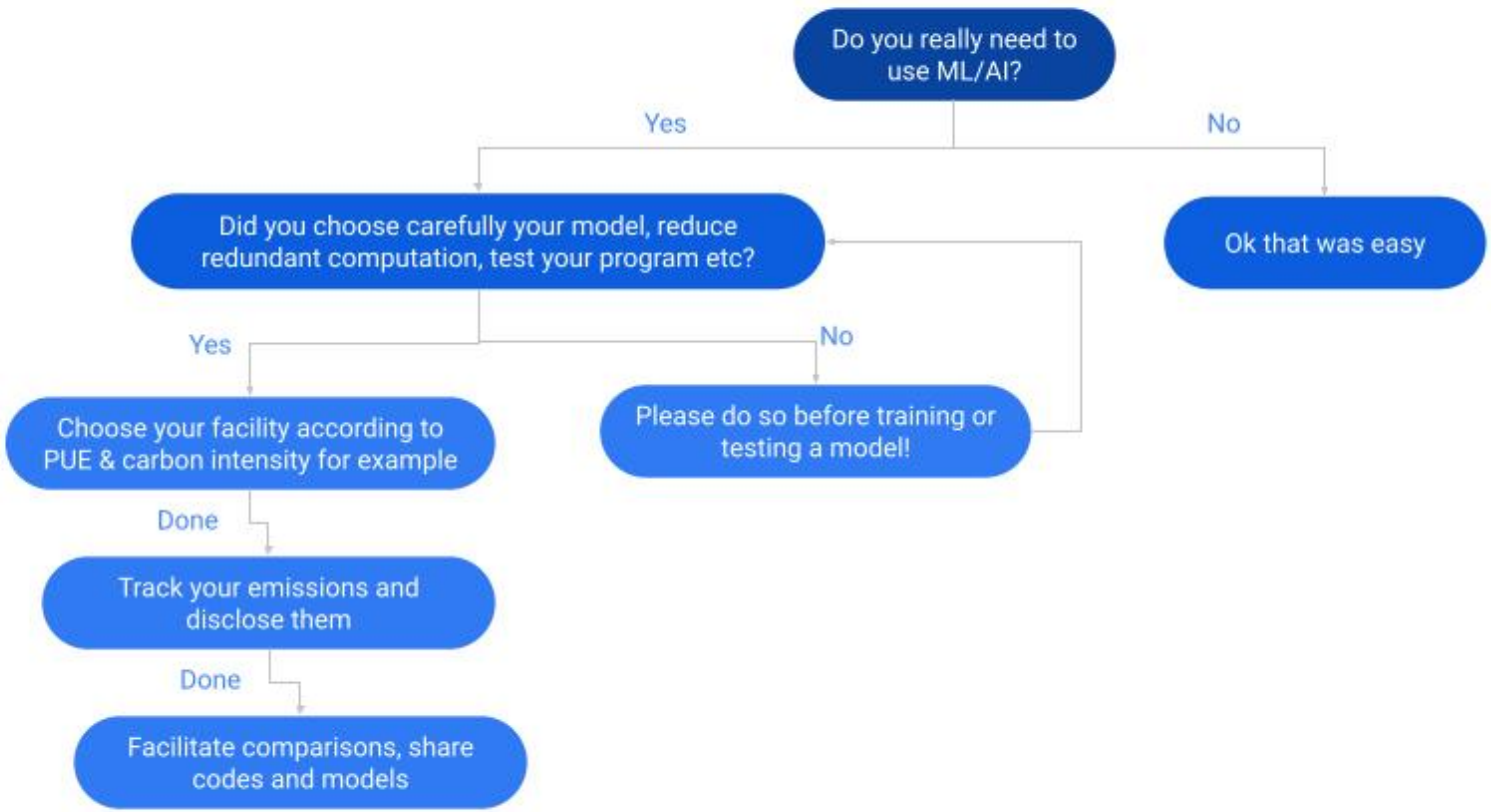


Morand C, Névéol A, Ligozat AL. How Green Can AI Be? A Study of Trends in Machine Learning Environmental Impacts. arXiv:2412.17376

Can we trust these numbers?

- Hypothesis and approximations are needed
- However, looking at the big picture:
  - Relative differences in impacts stand
  - Impacts are high overall

What can we do about it?



Ligozat AL. Luccioni S. 2021. A practical guide to quantifying carbon emissions for machine learning researchers and practitioners

Ten simple rules to make your computing more environmentally sustainable

- Rule 1: Calculate the carbon footprint of your work
- Rule 2: Include the carbon footprint in your cost–benefit analysis
- Rule 3: Keep, repair, and reuse devices to minimise electronic waste
- Rule 9: Be aware of unanticipated consequences of improved software efficiency

Lannelongue L, Grealey J, Bateman A, Inouye M (2021) Ten simple rules to make your computing more environmentally sustainable. PLoS Comput Biol 17(9): e1009324.

Acknowledgements

- Colleagues
  - LISN/STL - especially: Nesrine Bannour, Fanny Ducel, Clément Morand, Marco Naguib
  - Aurélie Bugeau, Karën Fort, Loïc Lannelongue, Anne-Laure Ligozat, Xavier Tannier
- Funding
  - ANR-23-IAS1-0004 InExtenso
  - ANR-20-CE23-0026-01 CODEINE
  - ITMO Cancer
  - ED STIC

Take home messages

Evaluation should cover quality + impactS

$\Rightarrow$ Measure impacts

don't fly from ORY to CDG