Méthodes d'évaluation de l'empreinte de l'IA

GreenDays 2025

Anne-Laure Ligozat (enslIE) &
Aurélie Bugeau (Bordeaux University, IUF)

Context and introduction

- Increasing investments and use of machine learning (ML)
- AI present in most prospective scenarios to achieve carbon neutrality (Bugeau et al., 2024)

- Many estimations of energy consumption of AI in general and of large ML models
  - Few estimations of environmental impacts of AI models (Luccioni et al, 2023; Berthelot et al, 2024; Morand et al, 2024; Desroches et al, 2025; Morrison et al 2025) and sector

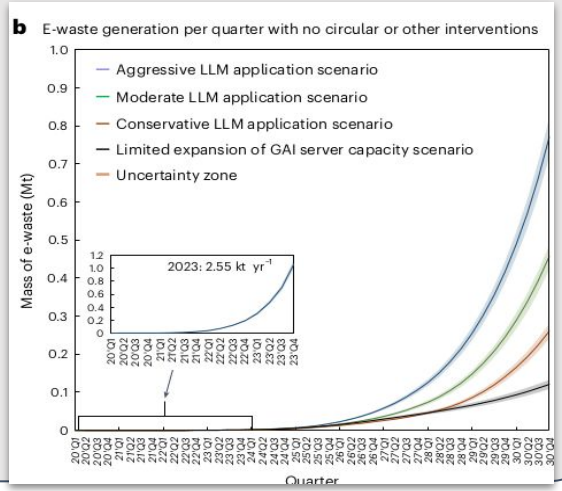
- In this presentation
  - We review large-scale estimations of AI sector
  - We discuss limits of current estimations

Large scale estimations of AI impacts

E-waste challenges of generative artificial intelligence (Wang et al, 2024)

cumulative 2020-2030: 1.2–5.0 Mt

(vs 53Mt of small IT waste)

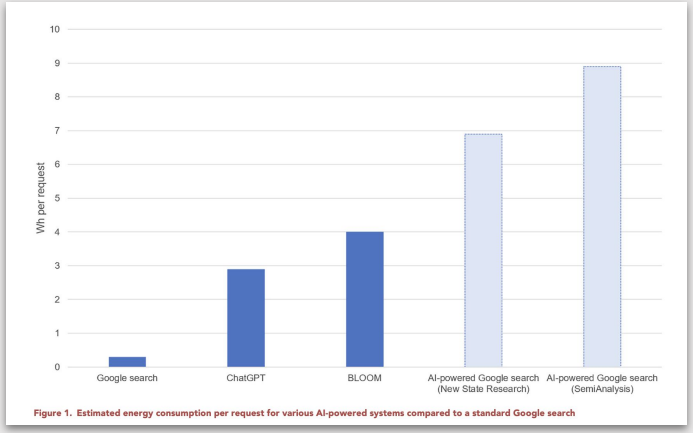


The Inference Cost Of Search Disruption – LLM Cost Analysis (Semianalysis, 2023)

ChatGPT inference: ~3,600 servers
ChatGPT in Google: ~ 513,000 servers
→ annual electricity consumption of 29.2 TWh

The growing energy footprint of artificial intelligence (de Vries, 2023)

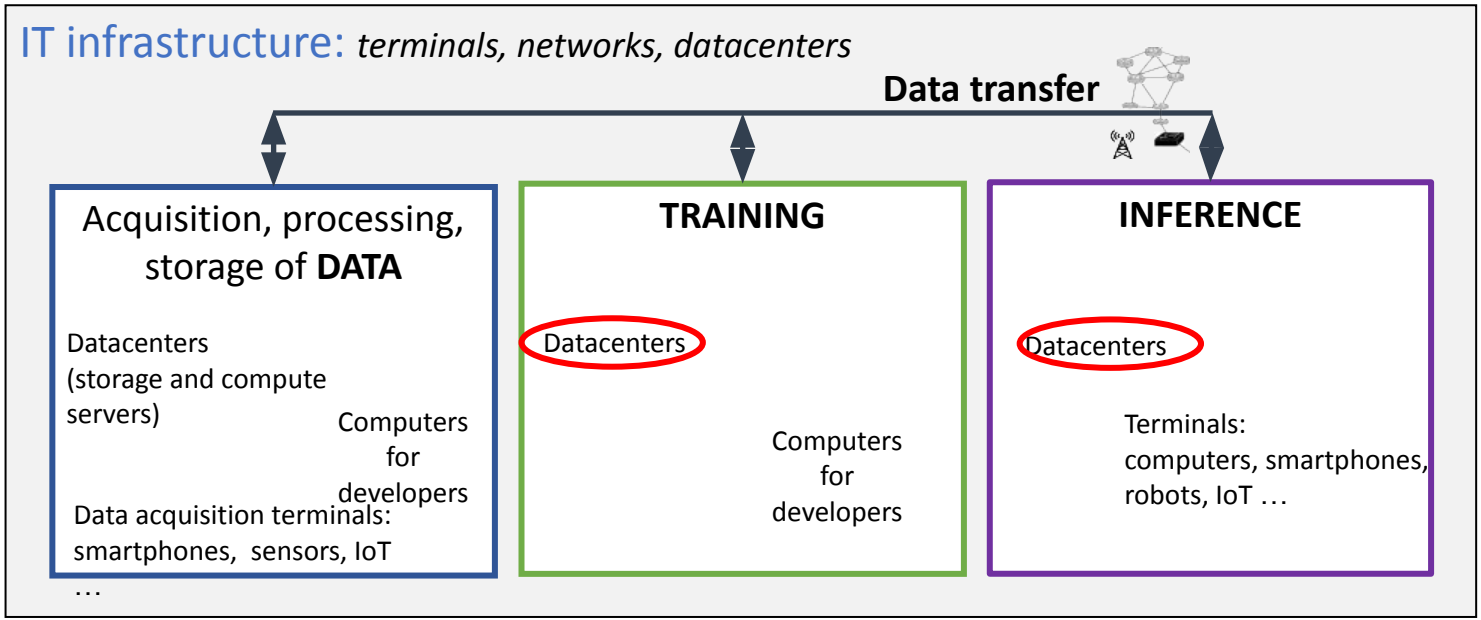
3Wh per LLM interaction



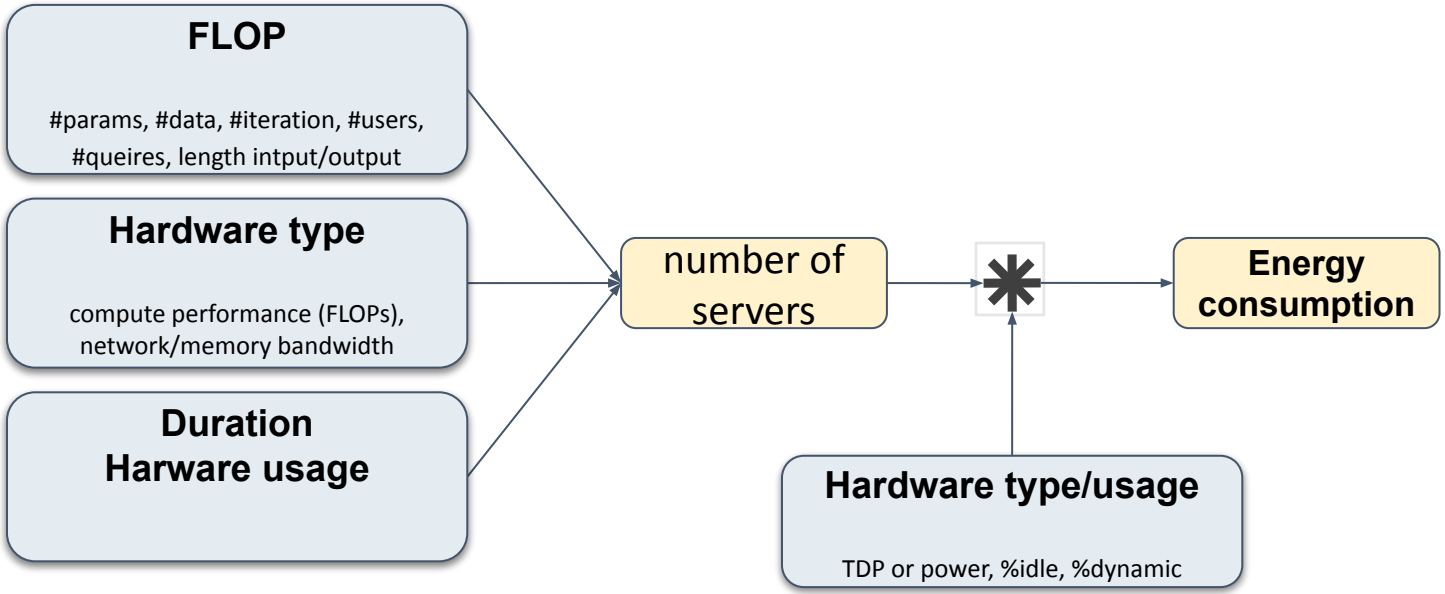
nvidia sales: in 2027, AI ≈ ⅔ of present data center electricity use

Different perimeters

Training and/or inference
AI and/or large scale models



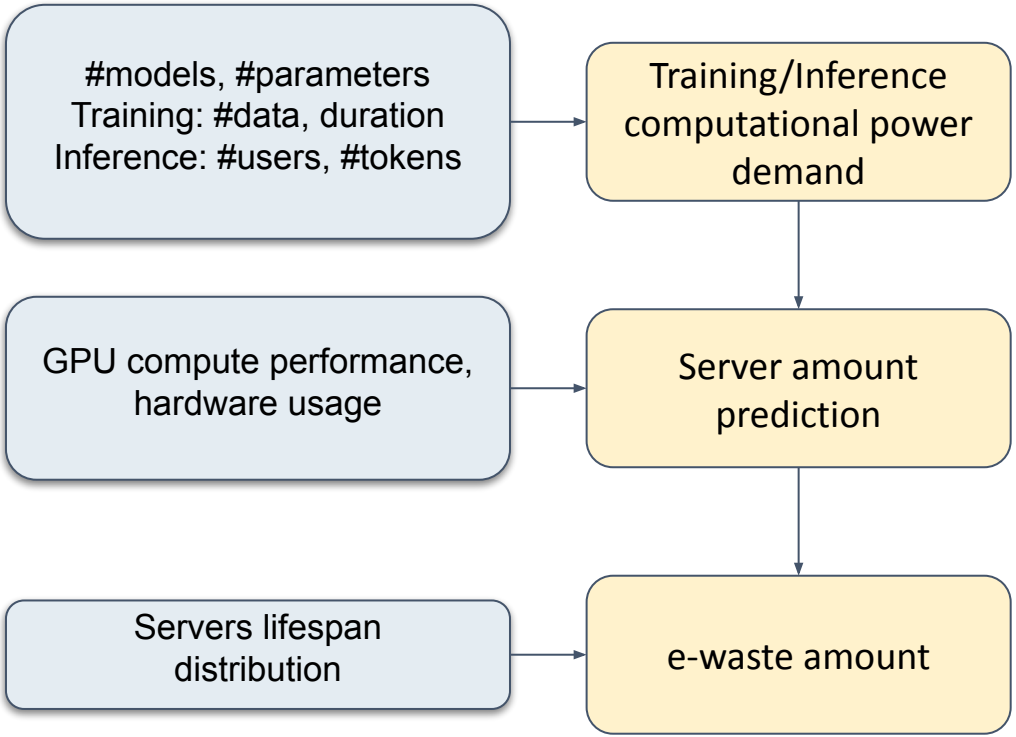
General methodology for estimating the amount of hardware and energy consumption of AI models



Remark: Various application methodologies with different hypotheses, mainly due to data scarcity

E-waste challenges of generative artificial intelligence (Wang et al, 2024)

Restricted to LLM



Training compute: $\sim 6NT_t/d$ FLOPs

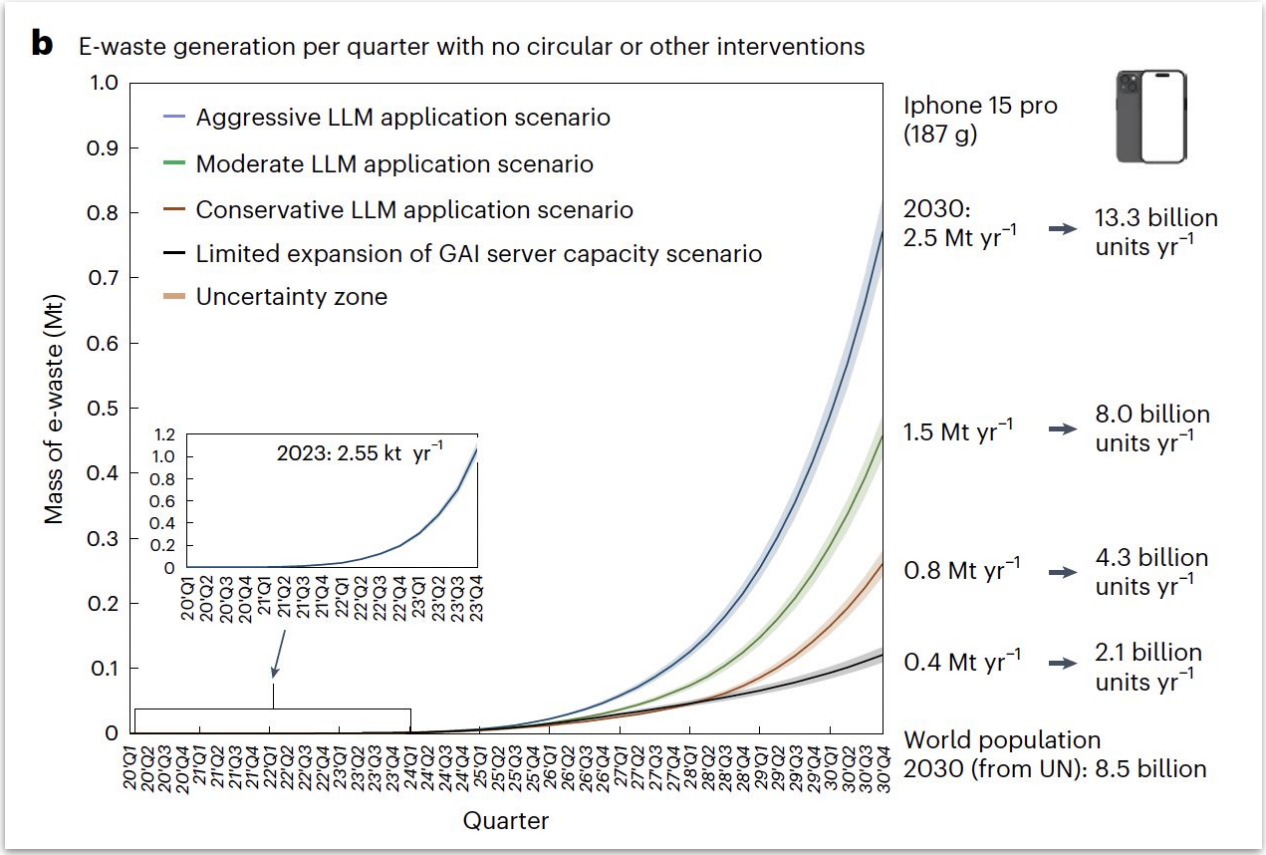
- *N*: number of non-embedding parameters
- T_t: total number of tokens used for training
- *d:* training duration in seconds

Inference compute per second: $\sim 2NT_i$ FLOPs

- T_i: number of tokens demand per second

(Kaplan, 2020)

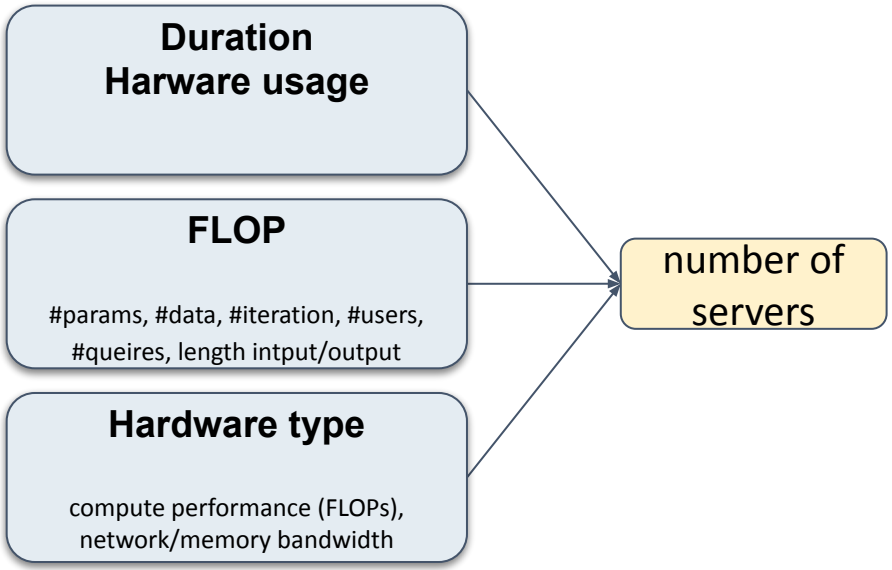
E-waste challenges of generative artificial intelligence (Wang et al, 2024)



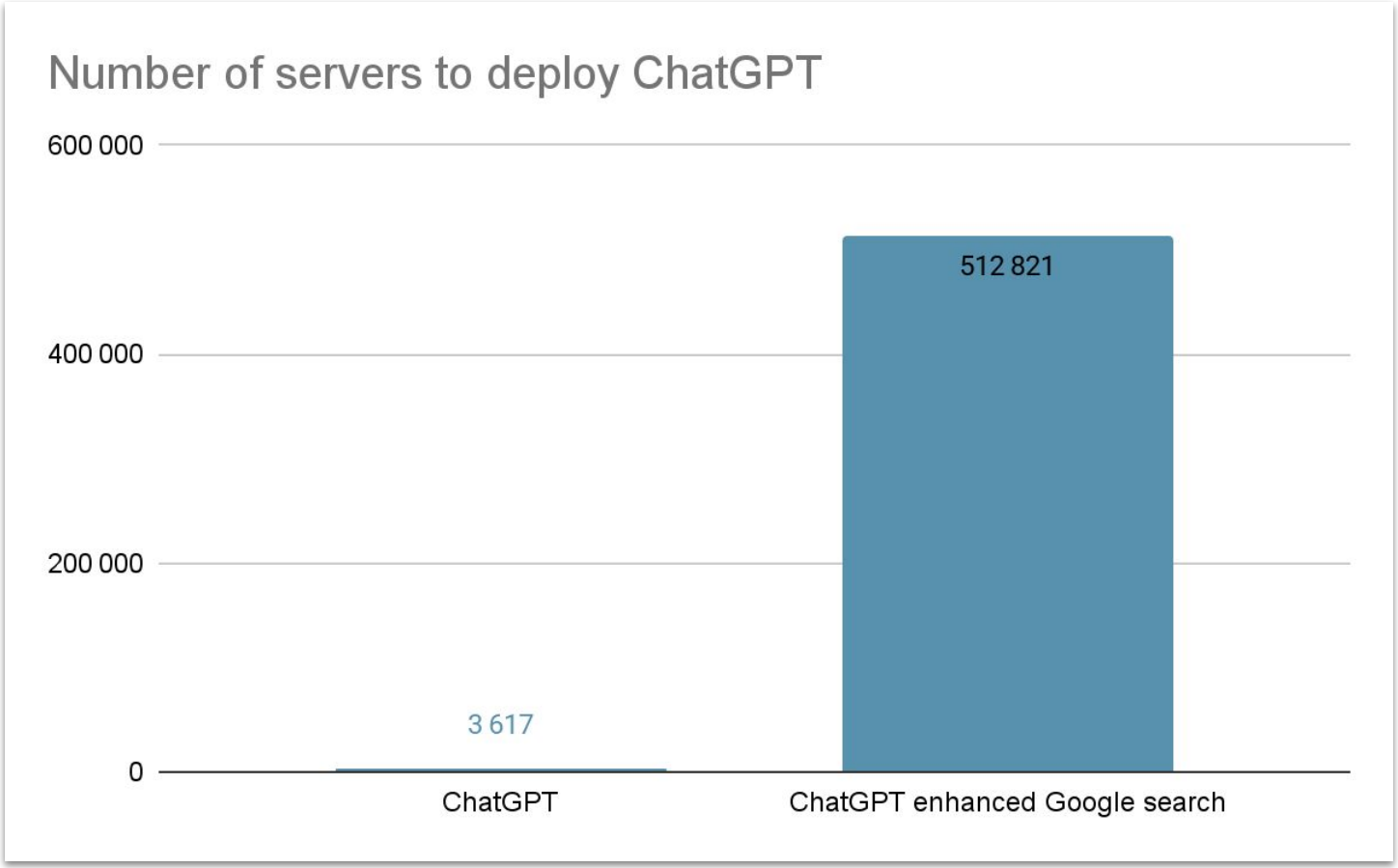
The Inference Cost Of Search Disruption – Large Language Model Cost Analysis (Semianalysis, 2023)

OUR MODEL IS BUILT FROM THE GROUND UP ON A PER-INFERENCE BASIS, BUT IT LINES UP WITH SAM ALTMAN'S TWEET AND AN INTERVIEW HE DID RECENTLY. WE ASSUME THAT OPENAI USED A GPT-3 DENSE MODEL ARCHITECTURE WITH A SIZE OF 175 BILLION PARAMETERS, HIDDEN DIMENSION OF 16K, SEQUENCE LENGTH OF 4K, AVERAGE TOKENS PER RESPONSE OF 2K, 15 RESPONSES PER USER, 13 MILLION DAILY ACTIVE USERS, FLOPS UTILIZATION RATES 2X HIGHER THAN FASTERTRANSFORMER AT <2000MS LATENCY, INT8 QUANTIZATION, 50% HARDWARE UTILIZATION RATES DUE TO PURELY IDLE TIME, AND $1 COST PER GPU HOUR.

- an independent research and analysis company
- restricted access to model
- many hypotheses based on GPT-3



The Inference Cost Of Search Disruption – Large Language Model Cost Analysis (Semianalysis, 2023)



The growing energy footprint of artificial intelligence (de Vries, 2023)

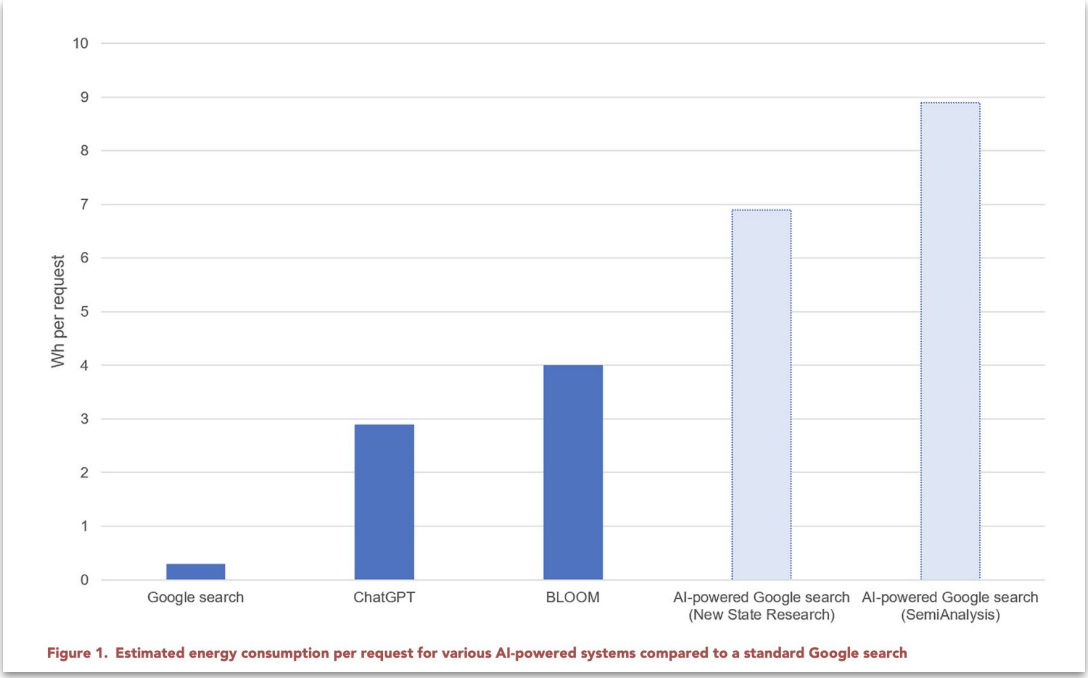
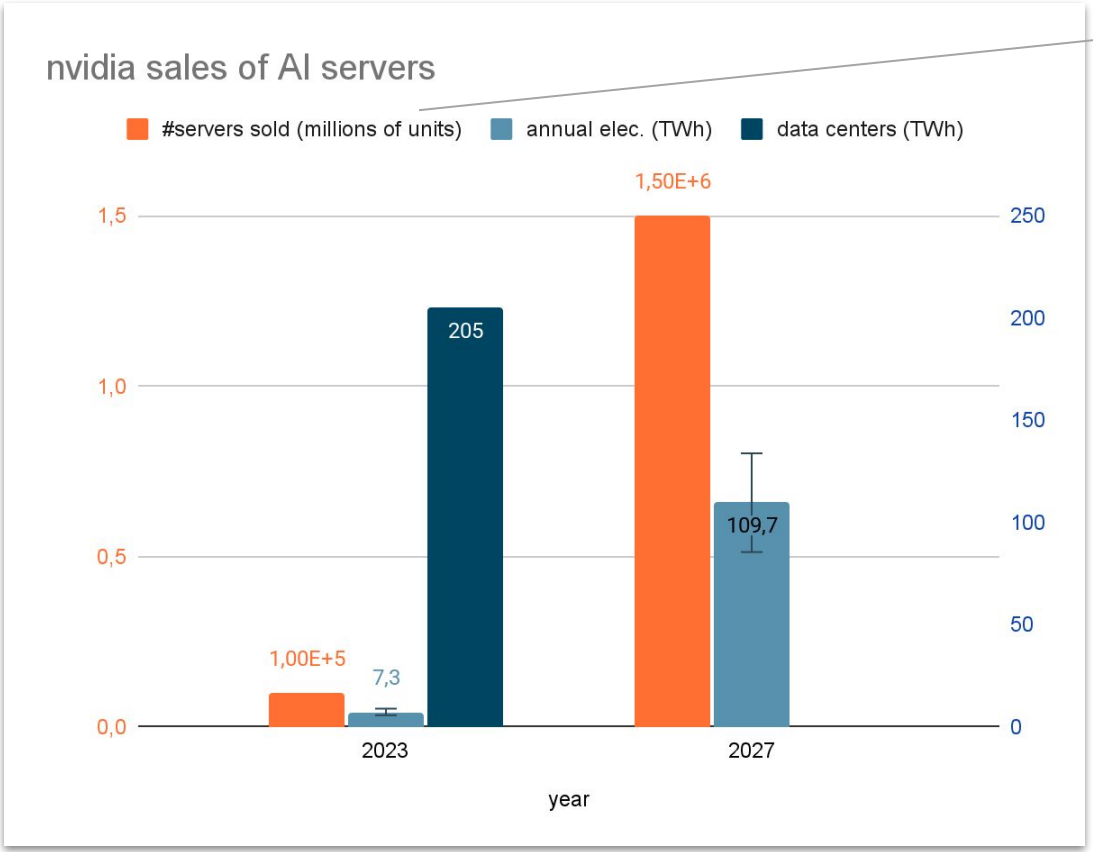


Figure 1. Estimated energy consumption per request for various AI-powered systems compared to a standard Google search

“Alphabet’s chairman indicated in February 2023 that interacting with an LLM could “likely cost 10 times more than a standard keyword search.[6]"

- a standard Google search reportedly uses 0.3 Wh of electricity (Remark: data from 2009)
- → 3 Wh per LLM interaction

The growing energy footprint of artificial intelligence (de Vries, 2023)



#servers sold from MarketWatch

Discussion

- Several estimations of large-scale AI impacts
  - different functional units, perimeters, hypotheses and types of impacts
  - but common restriction on compute servers (ignoring other IT and non-IT equipment)
  - no consideration of indirect impacts
- Large variability of results and lack of uncertainties analysis

- Consensus on growth of environmental impacts

- Both academic papers and media coverage should be more careful on data
  - when were data produced and by whom

Références

- Berthelot, Caron, Jay, Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. Conf.on Life Cycle Engineering, 2024 <hal-04346102v2 >
- Desroches, Chauvin, Ladan, Vateau, Gosset, Cordier.. Exploring the sustainable scaling of AI dilemma: A projective study of corporations' AI environmental impacts. 2025. <hal-04908002>
- de Vries, The growing energy footprint of artificial intelligence, Joule, 2023, https://doi.org/10.1016/j.joule.2023.09.004
- Kaplan, McCandlish, Henighan, Brown, Chess, Child, Gray, Radford, Wu, Amodei, Scaling laws for neural lan-guage models, 2020, <arXiv:2001.08361>.
- Luccioni, Viguier, Ligozat (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, JMLR, https://jmlr.org/papers/v24/23-0069.html
- Morand, Ligozat, Névéol. How Green Can AI Be? A Study of Trends in Machine Learning Environmental Impacts. 2024. <hal-04839926v2>
- Morrison, Na, Fernandez, Dettmers, Strubell, Dodge(2025). Holistically Evaluating the Environmental Impact of Creating Language Models, 2025. <arXiv:2503.05804
- SemiAnalysis, The Inference Cost Of Search Disruption – Large Language Model Cost Analysis, 2023 (webpage)
- Wang, Zhang, Tzachor, Chen, E-waste challenges of generative artificial intelligence, Nat Comput Sci 4, 818–823, 2024. <https://www.nature.com/articles/s43588-024-00712-6 >

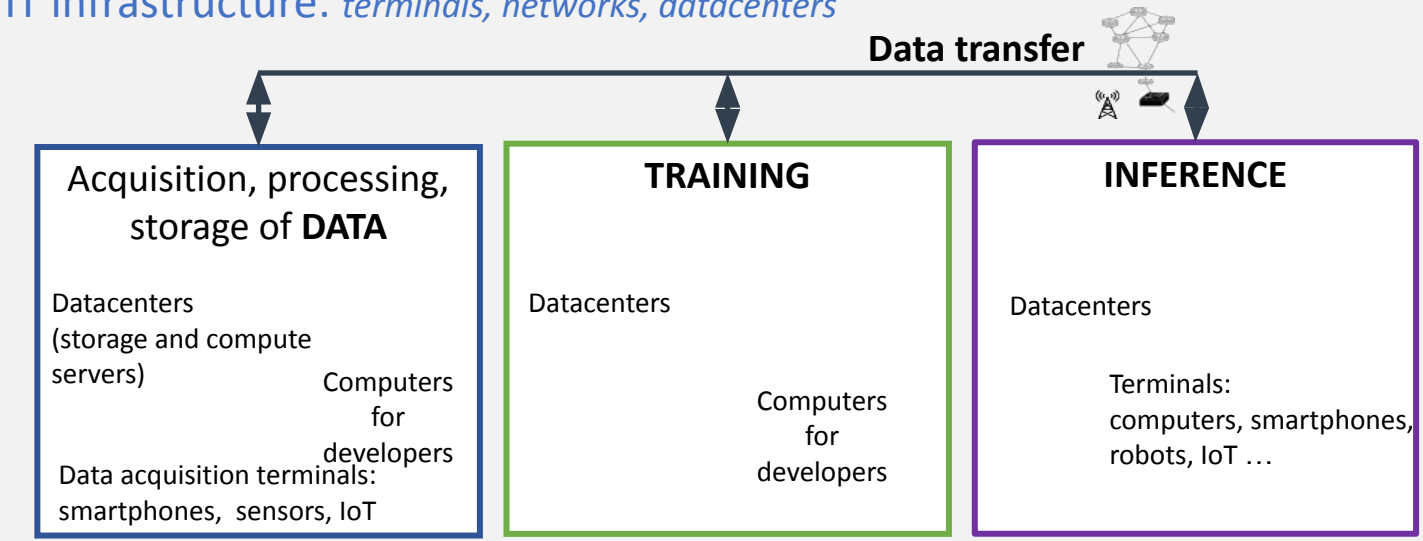
No Effects from interaction with society

Structural and systemic transformation: *changes in consumption, production and socio-economic structures, accelerating flows (logistics, people, financial, high-frequency trading, etc.)*

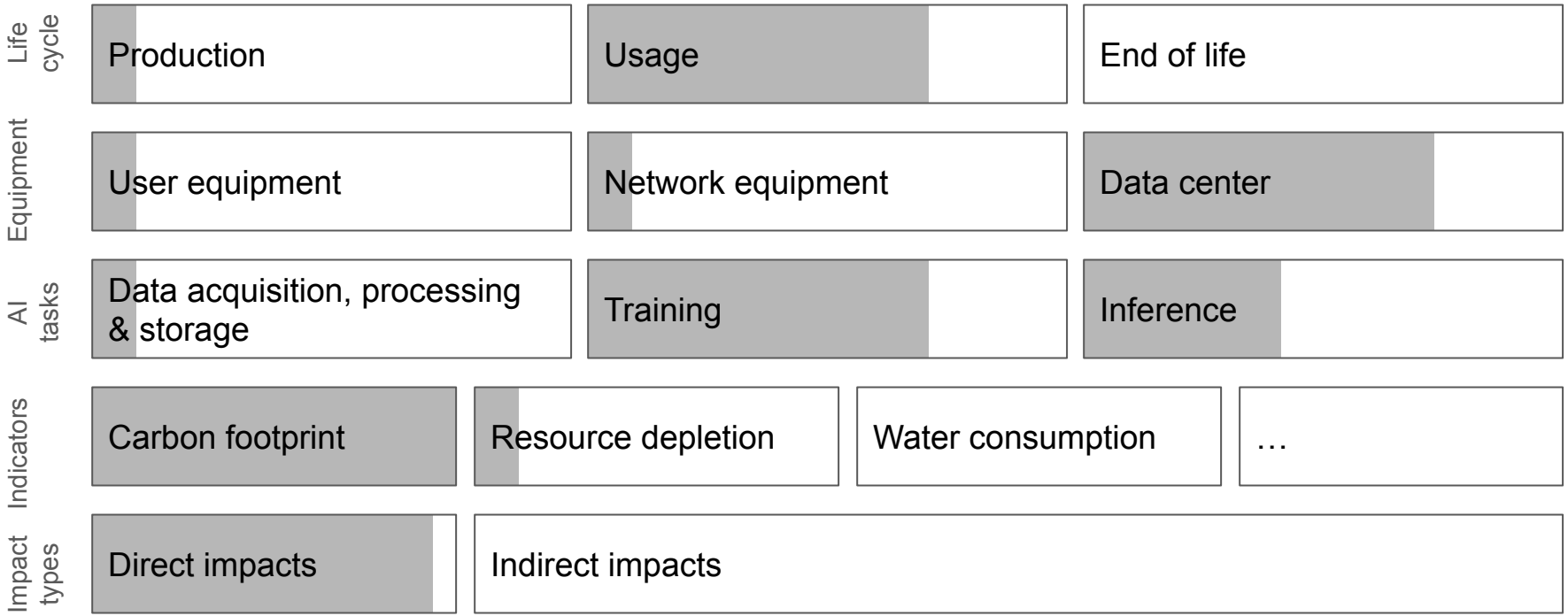
Effects from usage and interactions with the society: *optimization, substitution, induced/rebound effects, obsolescence, stacking*

Non-IT infrastructure: *factories, power plants, other networks, R&D, marketing, training, etc.*

IT infrastructure: *terminals, networks, datacenters*



What is presently assessed



Limits

- No strong correlation between parameters/flops and energy consumption (Henderson et al., 2020)

- Additional information necessary
  - latency, % usage of hardware/peak-to-average ratio, #users, #queries, length input/output
  - precision in floating-point arithmetic

- No standard methodology → differences between tools
  - Only GPUs/CPUs/RAMs or Full Servers
  - Static or Dynamic consumptions,
  - Only IT equipment
    - use of PUE (with its limits), or infrastructure consumption (Luccioni et al., 2023)

Estimating trends on global environmental impacts of AI

Methodologies

- Present estimations
  - Bottom-up approaches
    - Aggregation of data on individual models or HPC nodes (e.g. Desroches et al. 2025)
  - Top-down approaches
    - Number of hardware from shipments, market shares or total number of computation needed (e.g. IEA 2024, de Vries 2024)
  - A mix of bottom-up and top-down (ex: Wijnhoven et al., Schneider Electric, 2024)
- Future scenarios by varying several factors
  - datacenters planned, hardware efficiency, increases in model size and complexity adoption of generative AI, market demand, etc.

Biases of impact studies (Rasoldier et al., 2022)

Perimeter

- life cycle not taken into account: (Ligozat et al., 2021) for AI
- indirect (2nd and 3rd order) not taken into account: 5G

Uncertainties

- model choices, access and quality of data

Hypotheses

- comparison to what reference scenario?

Disconnection from global scenarios

- minimal benefits + poorly managed uncertainties
- incompatibility between measures

Conclusion

- More and and more evaluations of AI environmental impacts; fast evolving domain
  - But incomplete and not always fully transparent evaluations
  - Difficult access to data

- All trends: Important growth of every impact

(de Vries, 2023) The growing energy footprint of artificial intelligence

2023

- 100,000 nvidia AI servers (A100 and H100) sold
- full capacity: 650–1,020 MW or 5.7–8.9 TWh annually (vs 205 TWh for data centers)

2027

- 1.5 million servers
- 9.75–15.3 GW power demand or 85.4–134.0 TWh annually

“Alphabet’s chairman indicated in February 2023 that interacting with an LLM could “likely cost 10 times more than a standard keyword search.[6]"

- a standard Google search reportedly uses 0.3 Wh of electricity
- → 3 Wh per LLM interaction