Sustainability concerns with GenAI

Motivation

A few years ago, when people were asked about artificial intelligence, their first thoughts probably turned to smart assistants such as Amazon Alexa or Apple Siri. Now, ChatGPT has become the epitome of AI.

Chatbots can help increase productivity. AI is not only used in research or organizational and creative activities, but is also increasingly being used in business. Among other things, it is used to increase productivity and efficiency and save time.

AI models require an immense amount of data to be able to process inputs correctly and deliver corresponding outputs. Models are trained with human support on the one hand and self-monitored on the other. In addition, the training data must be updated regularly. However, training the models is only part of the resource consumption; all user queries must also be taken into account. Resources refer to the entire infrastructure of data centers, i.e., not only the computers themselves, but also the cooling systems and air conditioning.

The power consumption of data centers worldwide in 2022 was between 240 and 340 TWh of electricity. This corresponds to 1 to 1.3% of global electricity consumption.

Definition:

Definition is the artificial or machine replication of human intelligence. Training gives the system a kind of intelligence. How exactly the decision is made is not comprehensible, similar to the thought process in the brain. (Black box)

Subfields of artificial intelligence:

- Machine learning
    - Supervised learning
    - Unsupervised learning
    - Reinforcement learning
- (Deep learning)
- Natural language processing

LLMs:

AI models that capture and process input in natural language (prompts) and return a generated output in natural language as well. Many LLMs are now capable of performing tasks at a human level. LLMs belong to the group of pretrained foundation models. PFMs specialize either in NLP (LLMs) or in graphical processing. Due to their complexity, the training and use of LLMs consume enormous resources, which results in corresponding environmental and financial costs.

Dimensions of sustainability:

Sustainability means that the needs of the current generation are met in such a way that the needs of future generations do not have to be reduced.

Sustainability encompasses ecological, social, and economic dimensions. Ecological means that natural resources are conserved and protected. Social means fair living conditions worldwide. Economic means long-term prosperity.

Carbon footprint

The carbon footprint describes the amount of carbon dioxide released by humans. Particularly computationally intensive processes such as training large AI models lead to significant CO2 emissions.

Measures at the individual level include switching to public transportation, avoiding domestic flights, using renewable energies, and, in the technology sector, above all, energy-efficient hardware, sustainable training methods, and the use of green data centers.

In addition to emissions during the use of a product, there are also less visible emissions that nevertheless usually have a greater impact.

Embodied emissions (EE): These occur during manufacturing and construction, for example, in the production of hardware or the construction of data centers for AI.

Operational emissions (OE): are released during operation. In the case of AI, for example, these are emissions from cooling data centers or supplying infrastructure.

Power consumption of AI

Global power consumption by data centers in 2022 was around 230-340 TWh, which corresponds to around 1-1.3% of global power demand. In addition, electricity consumption has increased by 20-40% in recent years. Data centers not only consume electricity, but also account for around 1% of global greenhouse gas emissions.

A regionally focused analysis, concentrating on the US, estimates that data centers account for 2-3% of total electricity consumption, with this figure predicted to triple in the coming years. AI applications account for 10-20% of data center electricity consumption. This figure could rise to 70%.

Causes of emissions:

EE:

- Hardware manufacturing
- Model training
- Model development
- Construction of data centers
- Extraction of resources for hardware manufacturing

OE:

- Power consumption of servers
- (Hosting the models)
- (Processing requests)
- Cooling

80-90% of the power consumed by AI is attributable to its operation.

Only 32-40% of the power comes from renewable energy sources. The choice of location for data centers contributes to the level of emissions. The amount of renewable energy available varies from region to region. Regions with plenty of sun but little wind can only provide clean electricity during the day. In the US, 95% of all data centers are located in areas with above-average CO2 intensity. The choice of hardware in data centers, as well as the architecture of AI models and training strategies, also make a difference.

Impact

There are both ecological and financial impacts. The information and communications sector is now responsible for 2% of global CO2 emissions. Another study estimates 3.9% of global greenhouse gas emissions. At Google, greenhouse gas emissions have risen by 48% since 2019, and at Microsoft by 29.1% since 2020. This coincides with the AI boom.

Specific emission values

In 2018, US data centers still produced 31.5 million tons of CO2, but from September 2023 to August 2024, this figure rose to 105 million tons of CO2.

The training of BLOOM, an LLM with 176 billion parameters, generated 50 tons of CO2. The training of GPT-3 generated more than ten times that amount, 552 tons of CO2.

Image generation consumed as much electricity as a smartphone charge.

Data centers will emit as many greenhouse gases as 16 million cars in the next few years.

Challenges

While energy consumption can be easily measured, emissions from resource extraction and hardware manufacturing are difficult to quantify.

Another challenge will be to develop more efficient hardware and algorithms without running into the rebound effect, i.e., negating the savings through more intensive use.

It is hardly possible to operate data centers completely emission-free, as renewable energies are not constantly available.

Goals of companies and governments

Meta, Microsoft, and Google are failing to meet their own sustainability goals as a result of the rapid advancement of AI. In principle, they are striving for climate neutrality, but due to the increasing energy requirements of AI applications, this goal is being postponed or abandoned altogether. In addition, 64% of companies consider the actual energy consumption of AI to be too complex to measure.

Google continues to strive for climate neutrality by 2030, but its CO2 emissions have risen by 50% between 2019 and 2023. Google will not be able to meet its climate targets for 2024 due to the development of AI.

At the global level, the United States, China, and supranational actors such as the EU face similar challenges.

Reducing emissions

To reduce emissions caused by AI, resource consumption must be reduced. On the one hand, the use of AI can be reduced; on the other hand, a more efficient and targeted use of resources would be conceivable. The first approach is not considered feasible given the potential of AI.

Development approaches:

Sustainability by design:

- Develop products and services in such a way that they already meet sustainability standards
- Can be done both technically and organizationally
- Durability, ease of repair, reusability, efficient energy use

The three Rs of the circular economy:

- Reduction: same computing power with fewer hardware resources
- Reuse: component-based systems allow individual parts of the entire unit to be replaced in the event of defects or upgrades
- Recycling: giving hardware a second life

Early Stopping mechanism:

- Stops training prematurely to make training more energy-efficient
- Measures performance metrics and detects when no further improvements are being made or accuracy is actually deteriorating (overfitting)
- Reduces training time
- No relevant differences in accuracy

AI against emissions

AI can be used to reduce or even avoid emissions. There are numerous use cases in which AI can have a greater effect. When used correctly, AI actually improves the state of ecosystems.

Examples:

- Integration in the electricity market
- Construction or further development of smart buildings in conjunction with the Internet of Things to increase energy efficiency
- Urban development, intelligent transport networks, and mobility
- Tracking emissions and other environmental problems

The problem with these solutions is that they are expensive and complex to purchase and are therefore not yet widely used.

Strange study

A study investigated whether AI causes less CO2 than humans for certain tasks. Four AI models were used for this purpose, which were tasked with generating both text and images. CO2 emissions from training and queries were taken into account. Average values were used for human emissions, showing how much CO2 is generated per hour per citizen. The results showed 130 to 1500 times less CO2 for text generation and 310 to 2900 times less CO2 for image generation. The large intervals are due to the different consumption of the models, but also because a US citizen has a larger carbon footprint on average than an Indian citizen.

This finding cannot be generalized, as the use case is very limited.

Conclusion

Server hardware accounts for the largest share of AI's electricity consumption. The operation of cooling/air conditioning systems also contributes significantly.

Data centers currently account for 1.0 to 1.3% of global electricity consumption, and a significant increase is expected in the coming years.

A distinction is made between embodied and operational emissions. EE includes emissions from the extraction of resources for the hardware, its manufacture, and the training of the AI. OE consists of emissions from the electricity used to operate the AI servers.

Companies and governments are aware of the high emissions caused by AI. However, there is still a lack of implementation.

Translated with DeepL.com (free version)