

From computation to environmental cost the resource burden of artificial intelligence

Sophia Falk^{1*}, Nicholas Kluge Corrêa², Sasha Luccioni³, Lisa Biber-Freudenberger⁴, and Aimee van Wynsberghe¹

¹Sustainable AI Lab, Institute for Science and Ethics, Bonn University, Bonn, Germany

*Corresponding author: falk@iwe.uni-bonn.de

²Bonn-Aachen International Center for Information Technology (b-it) / CAISA Lab, Bonn University, Bonn, Germany

³Hugging Face, Paris, France

⁴Center for Development Research, Bonn University, Bonn, Germany

Supplementary Materials

Table S1. AI model Overview.

Overview of large-scale dense Transformer models published between 2022 and 2024 that were trained on Nvidia A100 SXM hardware. The estimated number of FLOPs per model is derived from reported or inferred parameter and token counts. Where official data on parameters or token counts is unavailable, italicized values indicate estimates based on industry conventions and credible leaked data sources.

Model	Parameter (<i>N</i>)	Token (<i>D</i>)	FLOPs Compute Budget
Amazon Titan TG-1	200 B	4 trillion	4.80×10^{24}
Mistral Large 2	123 B	<i>2 trillion</i>	1.48×10^{24}
LLama 2	70 B	2 trillion	8.40×10^{23}
DeepSeekLLM	67 B	2 trillion	8.04×10^{23}
Bloom	176 B	366 billion	3.86×10^{23}
GPT-3.5	175 B	<i>300 billion</i>	3.15×10^{23}
Falcon	40 B	1 trillion	2.40×10^{23}
Phytia	12 B	300 billion	2.16×10^{23}

Table S2. GPT-4 MoE Scenarios.

The estimated GPT-4 MoE configurations are based on publicly discussed architectural hypotheses and disclosed MoE architectures assuming a training dataset of 13 trillion tokens. The computational budget is estimated using parameter activation ratios informed by the MoE architectures of Ling-1T, DeepSeek-V3, and Llama 4 Maverick. For the remainder of the study, we use the scenario laid out by SemiAnalysis as a conservative approach (indicated in bolt).

Scenario	Total Parameter (<i>N</i>)	Active Parameter (<i>N</i> *)	Token (<i>D</i>)	FLOPs Compute Budget
George Hotz	1.76T	440B	13 trillion	3.43×10^{25}
SemiAnalysis	1.76T	222B	13 trillion	1.73×10^{25}
Ling-1T-like	1.76T	50B	13 trillion	3.90×10^{24}
DeepSeek-V3-like	1.76T	37B	13 trillion	2.89×10^{24}
Llama 4 Maverick-like	1.76T	17B	13 trillion	1.33×10^{24}

Table S3. FLOPs processed per A100.

Total amount of maximum theoretical throughput expressed in FLOPs processed per A100 (BF16) over its operational lifespan, assuming three lifespan scenarios.

Hardware lifespan (years)	Theoretical Computational Throughput
1	9.8×10^{21}
2	1.9×10^{22}
3	2.9×10^{22}