

AI Sustainability From Standards to Science to Action



July 7th, 2026



Boris Gamazaychikov
Co-Founder, Chief Executive Officer

Agenda

01

AI as a Field of Research

02

Connecting Research with Practice

03

Case Study

04

Scaling This Approach

DSIR ACE PILOT MODEL 1950

NPL

VOL. LIX. No. 236.]

[October, 1950

MIND

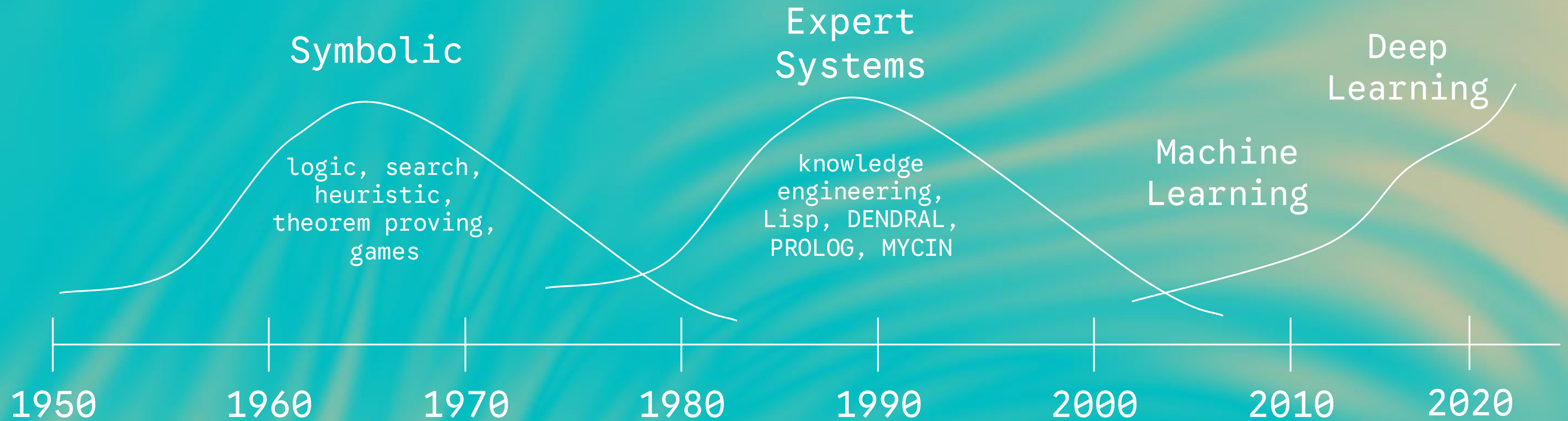
A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning





Deep
Learning

Machine
Learning

2010

2020

tokens_embed.weight

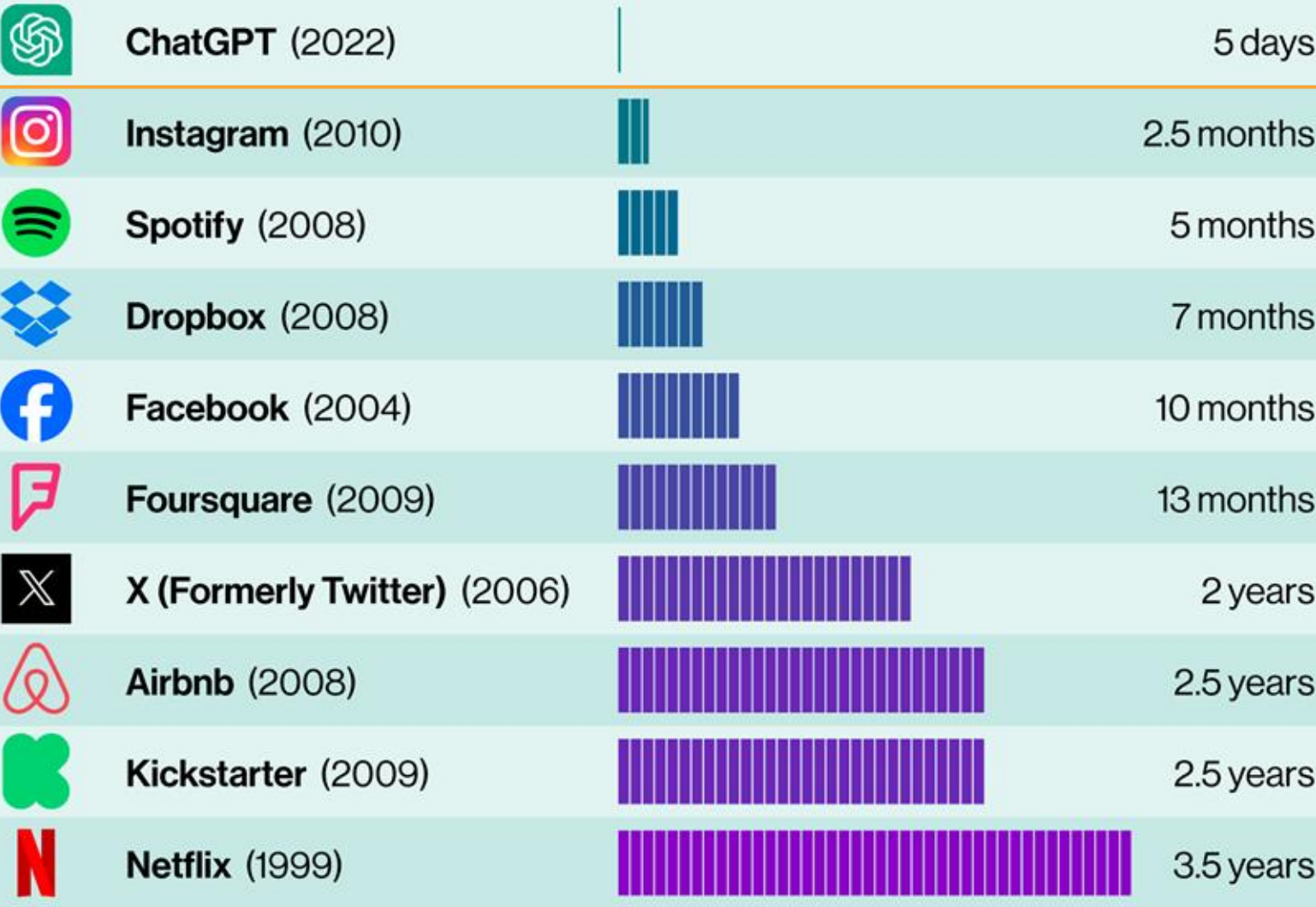
Vol 6 · p.1

offset 7,300,000

0.047256	0.143392	0.091252	-0.003474	0.006906	-0.005400	0.028910	0.048116	-0.057612	-0.003166
-0.052038	0.032670	-0.054014	-0.021879	0.021770	0.126981	0.097208	-0.050762	0.042653	-0.063476
-0.036264	0.059425	0.063276	-0.147230	0.152061	0.023890	0.007424	0.029486	0.012734	-0.075877
0.054146	-0.042039	0.060790	0.124063	-0.102504	0.065056	-0.030073	-0.037027	0.010899	-0.142807
-0.007973	-0.083414	-0.025297	-0.208081	-0.094468	0.149324	-0.007554	0.197698	0.019672	-0.037999
0.072393	-0.268340	0.133260	0.129678	-0.016137	-0.015233	-0.009790	-0.059881	-0.015974	-0.023074
0.005078	-0.022209	0.501190	0.131886	-0.005228	-0.069399	-0.116387	0.191962	-0.064171	0.003595
-0.036219	0.008179	0.038161	-0.064962	0.083053	-0.000786	-0.024285	-0.095446	-0.053381	0.018605
0.009730	-0.072944	0.132345	0.004272	0.001375	-0.006742	0.135036	-0.003096	0.005343	-0.025908
-0.082640	0.016974	0.087535	-0.018612	-0.063020	0.076850	-0.020838	-0.064494	-0.054081	-0.002033
0.047256	0.143392	0.091252	-0.003474	0.006906	-0.005400	0.028910	0.048116	-0.057612	-0.003166
-0.052038	0.032670	-0.054014	-0.021879	0.021770	0.126981	0.097208	-0.050762	0.042653	-0.063476
-0.036264	0.059425	0.063276	-0.147230	0.152061	0.023890	0.007424	0.029486	0.012734	-0.075877
0.054146	-0.042039	0.060790	0.124063	-0.102504	0.065056	-0.030073	-0.037027	0.010899	-0.142807
-0.007973	-0.083414	-0.025297	-0.208081	-0.094468	0.149324	-0.007554	0.197698	0.019672	-0.037999
0.072393	-0.268340	0.133260	0.129678	-0.016137	-0.015233	-0.009790	-0.059881	-0.015974	-0.023074
0.005078	-0.022209	0.501190	0.131886	-0.005228	-0.069399	-0.116387	0.191962	-0.064171	0.003595
-0.036219	0.008179	0.038161	-0.064962	0.083053	-0.000786	-0.024285	-0.095446	-0.053381	0.018605
0.009730	-0.072944	0.132345	0.004272	0.001375	-0.006742	0.135036	-0.003096	0.005343	-0.025908
-0.082640	0.016974	0.087535	-0.018612	-0.063020	0.076850	-0.020838	-0.064494	-0.054081	-0.002033

Online platform (year)

Time taken to reach one million users



Source: Statista

Proportion of S&P 500 mentioning "AI" during quarterly earnings







Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell, Ananya Ganesh, Andrew McCallum

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware. In this paper we bring this issue to the attention of NLP researchers by quantifying the approximate financial and environmental costs of training a variety of recently successful neural network models for NLP. Based on these findings, we propose actionable recommendations to reduce costs and improve equity in NLP research and practice.

Anthology ID: P19-1355

Volume: [Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics](#)

Month: July

Year: 2019

Address: Florence, Italy

Editors: [Anna Korhonen](#), [David Traum](#), [Lluís Màrquez](#)

Venue: [ACL](#)

SIG: –

Publisher: Association for Computational Linguistics

The inherent gap

Research

GOAL

Answers

TIMELINE

Years

MEASURED BY

Peer review

VALUES

Rigor, Openness

Business

GOAL

Value

TIMELINE

Quarters

MEASURED BY

Revenue

VALUES

Speed, Scale

The inherent gap **must be closed**

Research + Business

STANDARDS

Create shared measurement approaches

DISCLOSURE

Establish decision-useful metrics

TOOLS

Make knowledge usable

PROCUREMENT

Translate to contract-ready language

to drive AI sustainability progress

Connecting Research with Practice

RESEARCH

ACTION

AI Inference Measurement

*Power Hungry Processing:
Watts Driving the Cost of AI
Deployment?*

AI Energy Score:
standardized, labeled,
created procurement
language

Model Substitution

*Balancing Sustainability and
Performance: The Role Of
Small-Scale LLMs In Agentic
AI*

Integration of decision
framework into corporate
AI model selection process

Proprietary AI Video Model Estimation

*Lights, Camera, Carbon:
Architectural Scaling Laws
for Video Generation Energy
Consumption*

Integration of equations
into AI carbon accounting
software

NEW

Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?

ALEXANDRA SASHA LUCCIONI and YACINE JERNITE, Hugging Face, Canada/USA
EMMA STRUBELL, Carnegie Mellon University, Allen Institute for AI, USA

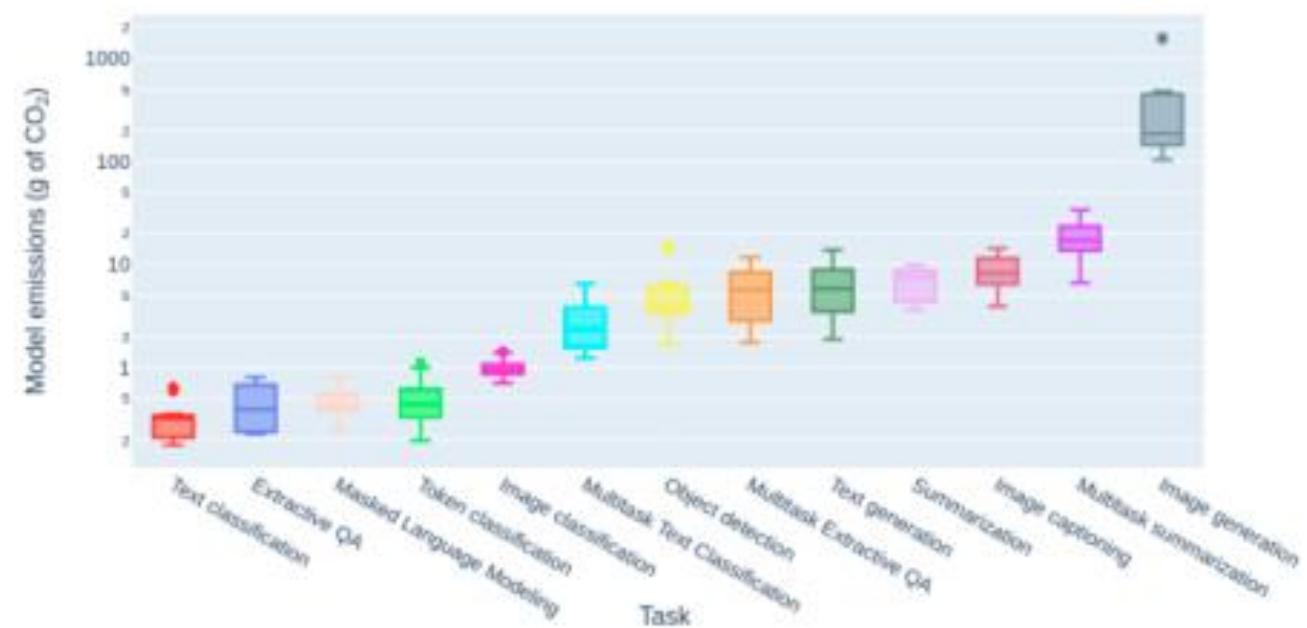


Fig. 1. The tasks examined in our study and the average quantity of carbon emissions they produced (in g of CO₂eq) for 1,000 queries. N.B. The y axis is in logarithmic scale.

Recent years have seen a surge in the popularity of commercial AI products based on generative, multi-purpose AI systems promising a unified approach to building machine learning (ML) models into technology. However, this ambition of “generality” comes at a steep cost to the environment, given the amount of energy these systems require and the amount of carbon that they emit. In this work, we propose the first systematic comparison of the ongoing inference cost of various categories of ML systems, covering both task-specific (i.e. finetuned models that carry out a single task) and ‘general-purpose’ models, (i.e. those trained for multiple tasks). We measure deployment cost as the amount of energy and carbon required to perform 1,000 inferences on representative benchmark dataset using these models. We find that multi-purpose, generative architectures are orders of magnitude more expensive than task-specific systems for a variety of tasks, even when controlling for the number of model parameters. We conclude with a discussion around the current trend of deploying multi-purpose generative ML systems, and caution that their utility should be more intentionally weighed against increased costs in terms of energy and emissions. All the data from our study can be accessed via an interactive demo to carry out further exploration and analysis.

CCS Concepts: • **Computing methodologies** → **Machine learning; Neural networks**; • **Hardware** → **Impact on the environment; Power estimation and optimization**.

ACM Reference Format:

Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24)*, June 3–6, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3630106.3658542>

★ AI Energy Score

Model
berkeley-nest
Starling-LM-7B-alpha

📅 Scored	February 2025
✅ Task	Text Generation
🖨 Hardware	NVIDIA H100-80GB

⚡ Inference Energy (Wh) per 1k tasks **19.01**

Energy Score ★★☆☆☆

Learn More: tiny.cc/AIEnergy



AI Energy Score

Feb'25: Leaderboard, label, submission toolkit, methodology

[Submission Portal](#) [Label Generator](#) [FAQ](#) [Documentation](#) [Download Data](#) [Community](#)

 **AI Energy Score**

Text Generation

Reasoning

Image Generation

Text Classification

Image Classification

Image Captioning

Summarization

...

Select Model Class

A (Single Consumer GPU) <20B parameters

Sort

Low to High

There's a 342,822x difference between the highest and lowest energy use in this leaderboard.

There's a 23x difference between the highest and lowest energy use in this class.

Model	Provider	GPU Energy (Wh) per 1k Queri...	Score	Test Date
Qwen3-0.6B	Qwen	546.99	★★★★★	Dec 25
Hunyuan-1.8B-Instruct	tencent	1,478.96	★★★★★	Dec 25
Phi-4-reasoning-plus	microsoft	9,461.61	★★	Dec 25
gemma-3-4b-pt	google	11,811.18	★	Dec 25
SmolLM3-3B	HuggingFaceTB	12,791.22	★	Dec 25

Outcomes to Date

- ✓ Influencing policy frameworks (EU AI Act Code of Practice)
- ✓ Informing emerging standards (ITU, IEEE, European Commission)
- ✓ Used in enterprise AI procurement vendor evaluation
- ✓ Integrated into enterprise AI governance workflows (Salesforce model cards)
- ✓ Enabling further research

53k+ page views ★ 150+ models ★ 11 tasks

🏆 Recognized by:



📰 Featured in:



Moving Forward, Together



Researchers

AI Labs

Policy
Makers

Customers

Infrastructure
Providers

Everyone has a role to play

Thank You

