VU AI in Education Show & Share

Can Generative AI be Sustainable?

Meike Morren [1]

[1]Assistant Professor of Sustainable Consumer Choices, School of Business and Economics, Marketing department, Vrije Universiteit Amsterdam (VU)

December 5, 2024

Large Language Models

More Large Language Models ...

MISTRAL AI_

Claude

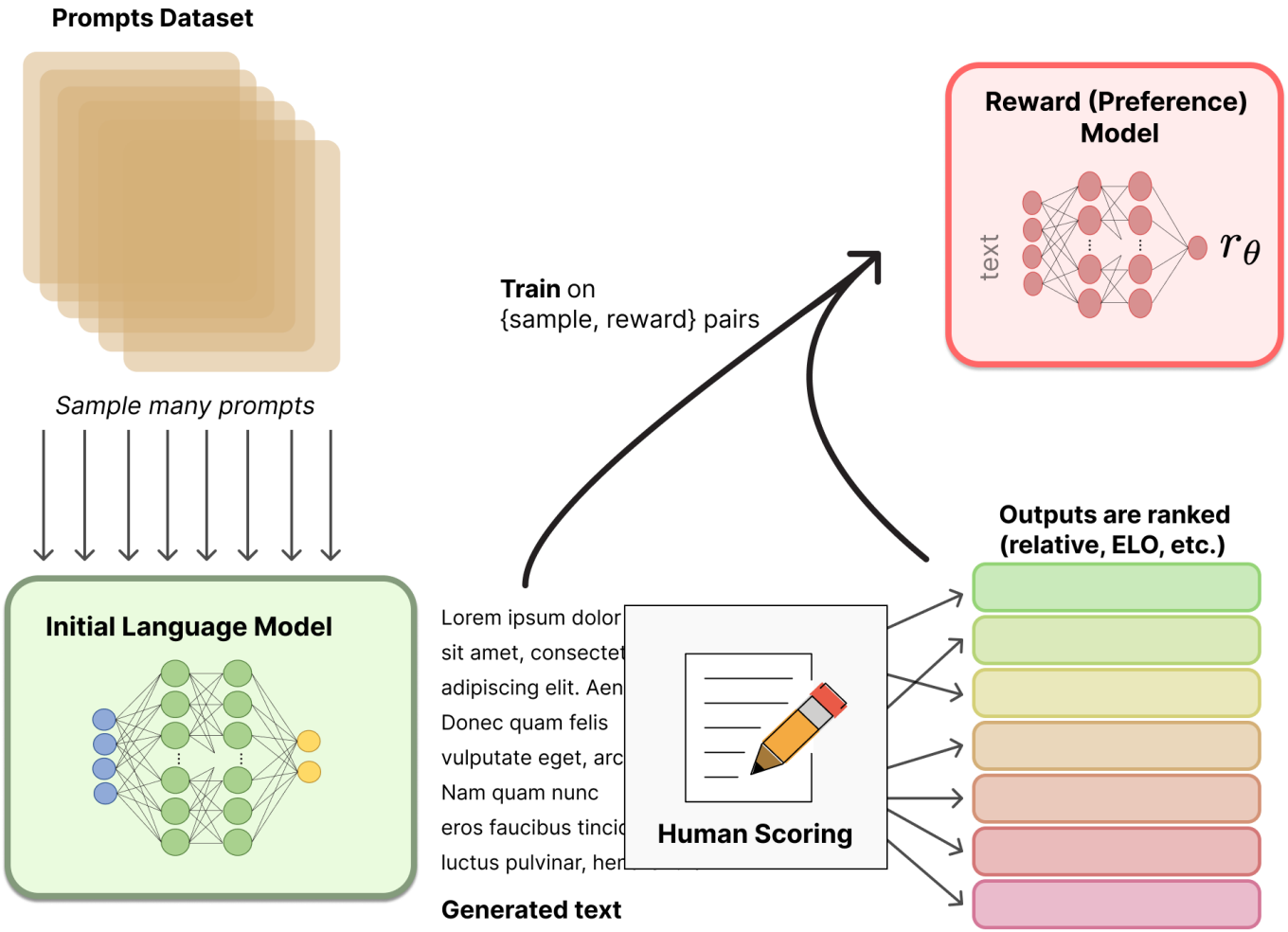
BY ANTHROP\C

Capabilities LLMs

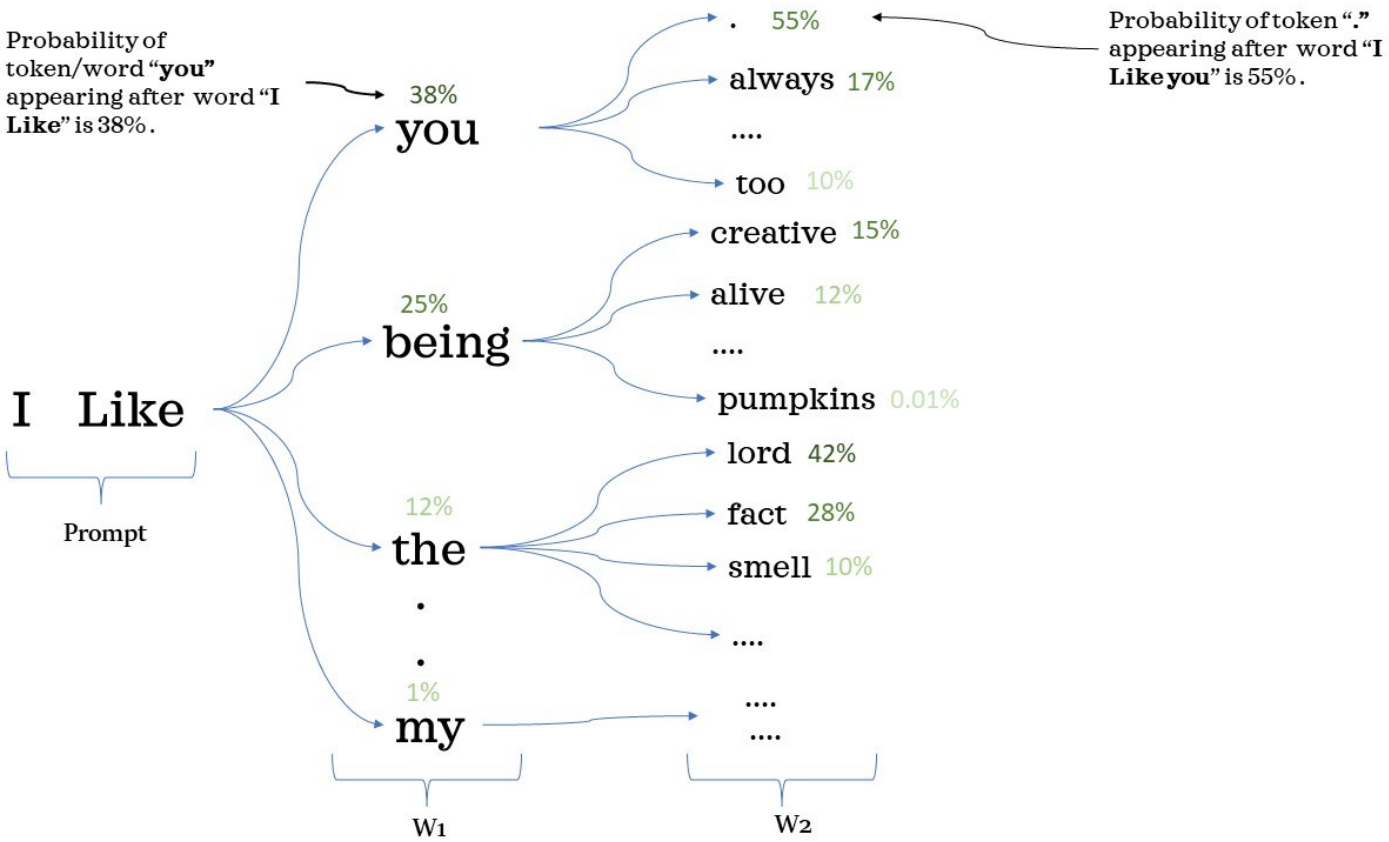
- Perform professional and academic exams at normal intelligence OpenAI's technical report
- Help programmers by increasing efficiency using the copilot
- Can describe, analyze and create images and videos
- Win chess, Jeopardy, Go and argumentation debates from world champions (IBM debater)
- ...

Under the hood

Reinforcement Learning Human Feedback



Under the hood



Want to know more? See this gentle introduction.

Limitations I: errors

- Hallucination: made-up facts
- Developed and screened by mainly white (high educated) males
- Some easy tasks are hard for GPT4o: select words with third letter being k
- Need for computational resources increases exponentially

Bias, disinformation, over-reliance, privacy, cybersecurity

More on what GPT4 can and cannot do, is described by Bubeck et al 2023

Hallucination

Lawyers have real bad day in court after citing fake cases made up by ChatGPT

Lawyers fined $5K and lose case after using AI chatbot "gibberish" in filings.

JON BRODKIN - 6/23/2023, 7:32 PM

For more on the hallucinations among several LLMs, see Leaderboard

Limitations II: Not OpenSource

"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."

The architecture of GPT4 is not shared with public by OpenAI

Some opensource models perform equally well on some tasks (e.g. LLama3.1).

Limitations III: Computation Resources

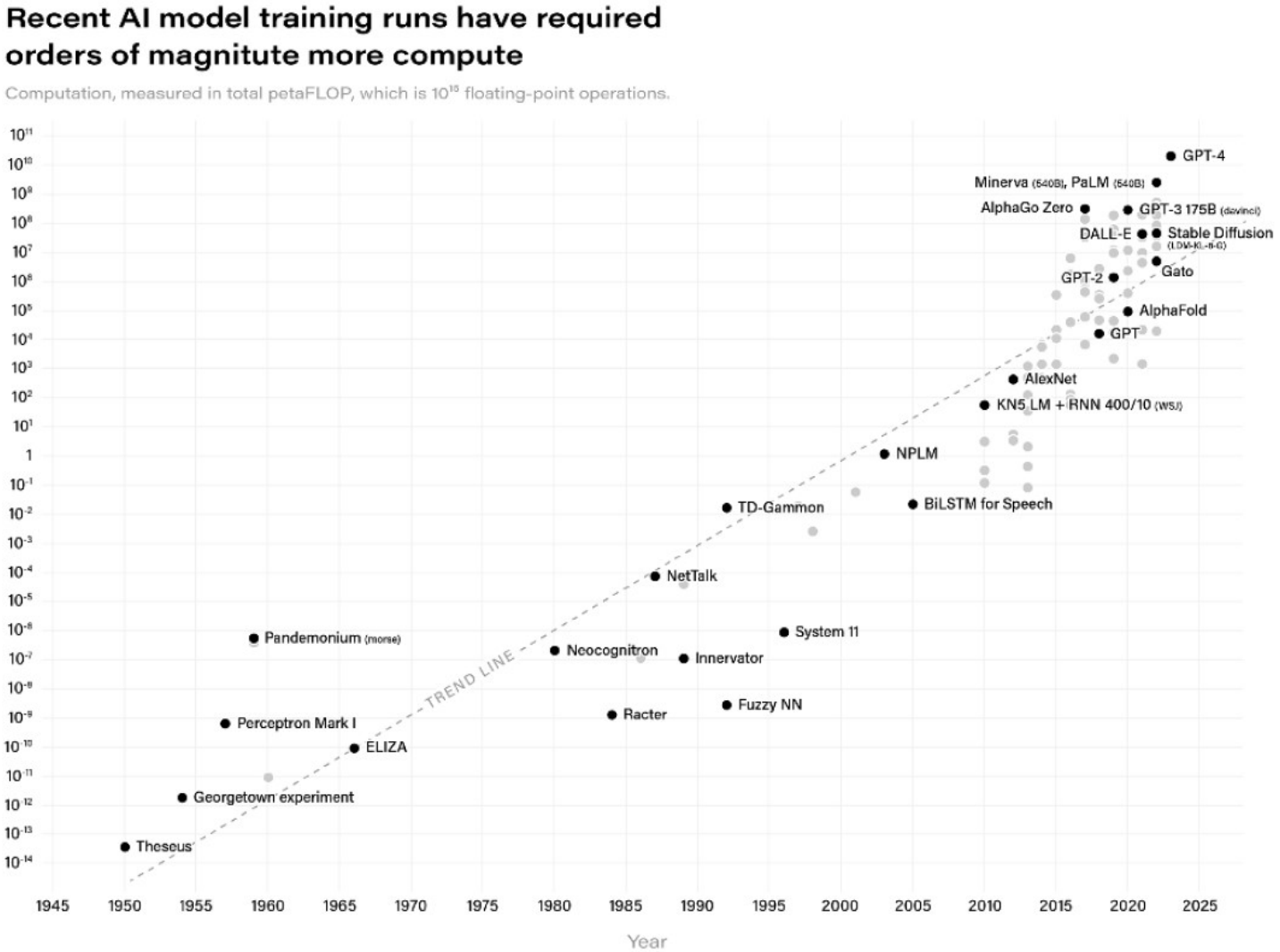


Figure 1. OpenAI is estimated to have used approximately 700% more compute to train GPT-4 than the next closest model (Minerva, DeepMind), and 7,000% more compute than to train GPT-3 (Davinci). Depicted above is an estimate of compute used to train GPT-4 calculated by Ben Cottier at Epoch, as official training compute details for GPT-4 have not been released. Data from: Sevilla et al., 'Parameter, Compute and Data Trends in Machine Learning,' 2021 [upd. Apr. 1, 2023].

Emissions

- BERT: 340 million parameters, 1.536 Mwh, 0.7 tCO2
- GPT2: 1.5 billion parameters, 1.7 Mwh, 0.7 tCO2
- GPT3: 175 billion parameters, 1.287 Mwh, 552 tCO2
- GTP4: 1.8 trillion parameters, 7200 Mwh
- PALM (google): 540 billion parameters, 3436 Mwh, 271 tCO2
- GLAM (google): 1.2 trillion parameters, 456 Mwh, 40 tCO2
- LLama (meta): 70 billion parameters, 688 Mwh, 291 tCO2
- OPT (meta): 175 billion parameters,

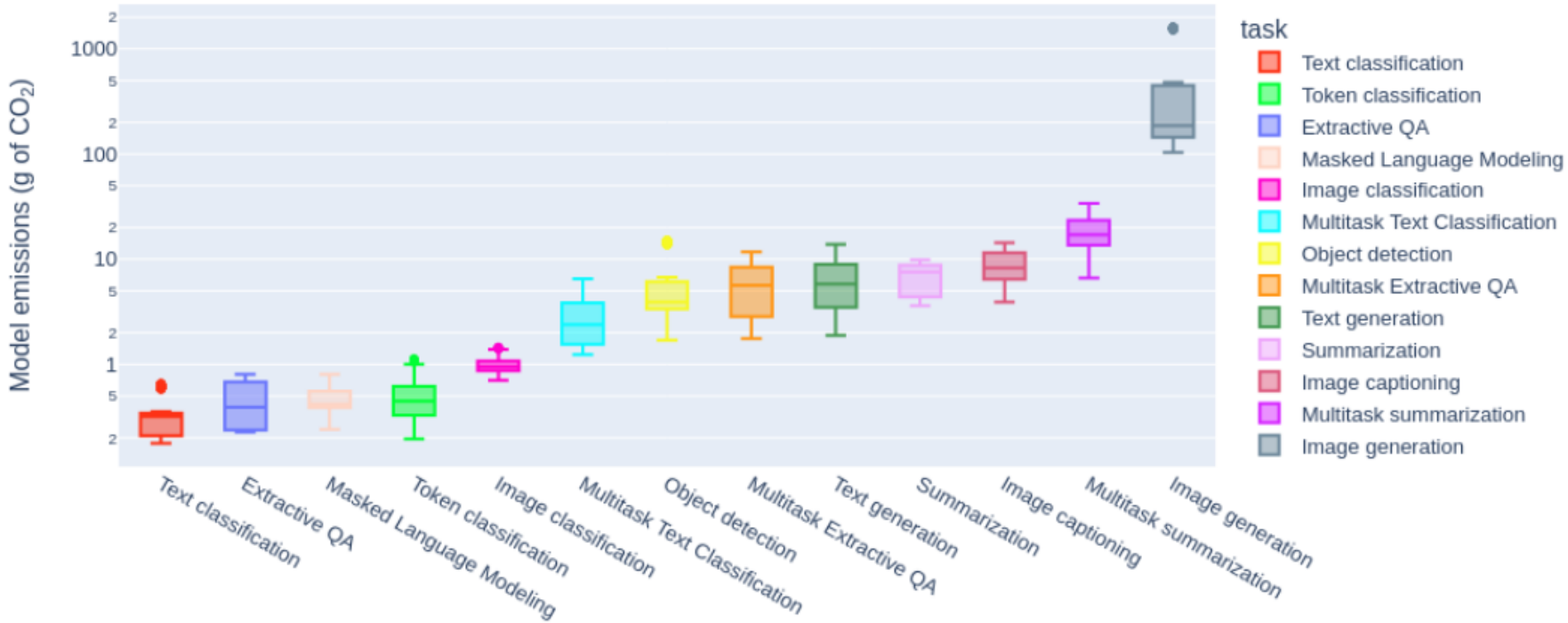
For more information about environmental impact, see Center for Data Innovation

Most emissions are from using the model (about 80% of AI workload in data centers is from inference (i.e. usage) and the rest for training). Meta says that this varies across use cases and estimate 65% of carbon footprint is associated with inference.

Multimodal LLMs - example problems

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



To summarize...

1. LLMs are not a panacea to all problems
2. LLMs have their limitations (e.g. hallucination)
3. LLMs generate pollution

1. LLMs are very powerful
2. Used wisely, LLMs can help people a lot
3. LLMs should not replace search engines