Scaling Mount AI

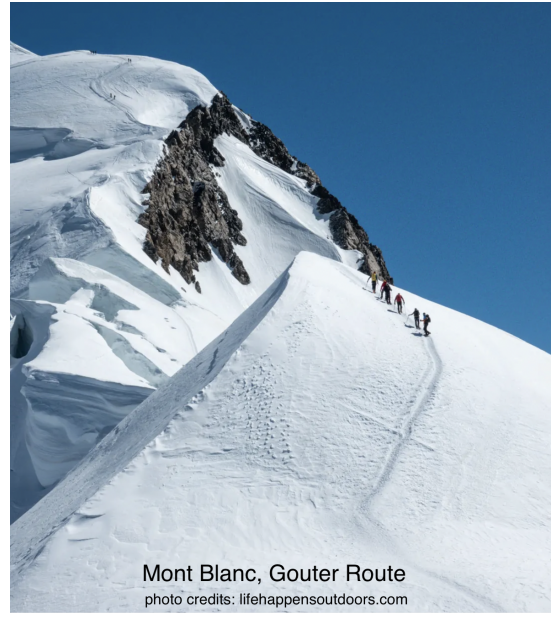
a literature study

Kees van Berkel

MPSoC 2025, Megève

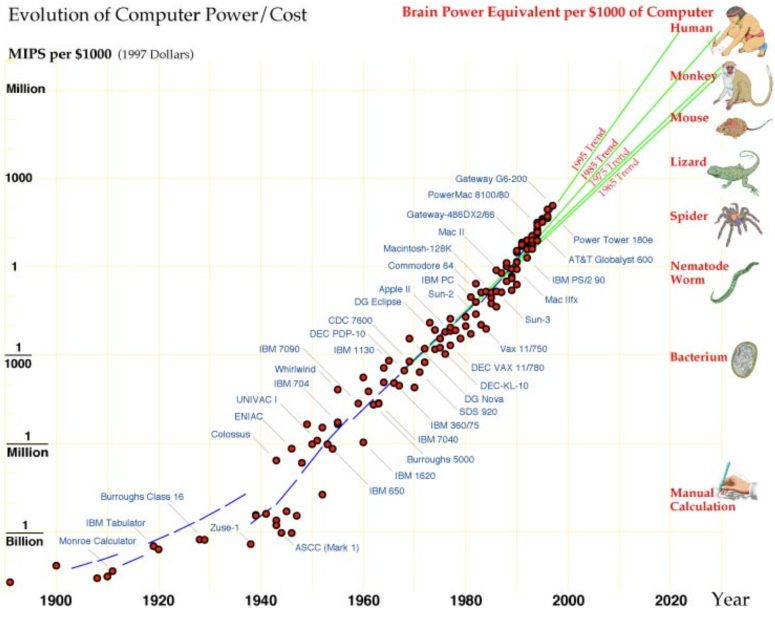
Consultant Snap Inc.
Emeritus professor TU Eindhoven

df
14



Mont Blanc, Gouter Route
photo credits: lifehappensoutdoors.com

When will computer hardware match the human brain?



[Hans Moravec, 1998]

↑
Brain power equivalent per $1000 of computer

How did Moravec estimate "brain-power equivalent"?

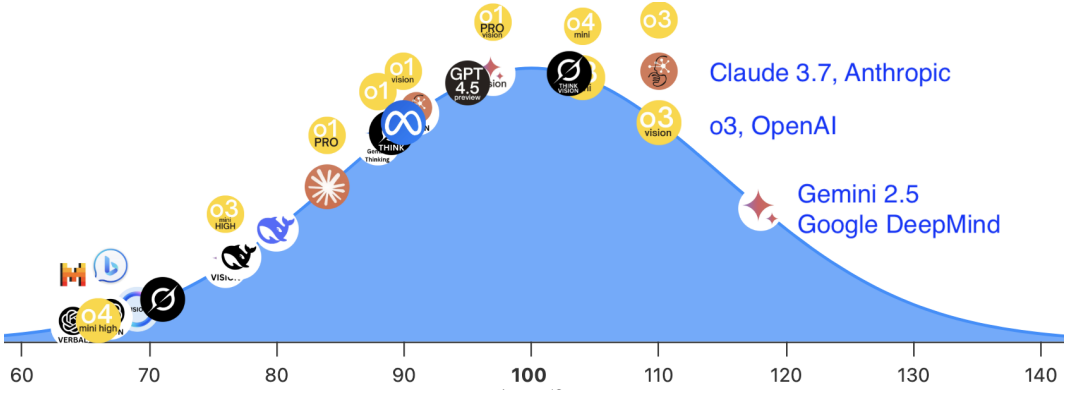
A task-functional model based on audacious "analogy and extrapolation":

MIPS	Neurons	Task
10^0		extract simple features from real-time imagery
10^1		follow complex gray-scale patches
10^2		follow moderately unpredictable features like roads
10^3		robot vision: 100M instructions $\times$ 10Hz
10^3	10^8	$\approx$ retina, $< 1\text{cm}^2 \times 0.5\text{mm}$
10^8	10^{11}	human brain, $\approx 1500\text{cm}^3$ $\approx 10^5\times$ retina

More recent estimates:

- $10^{13} - 10^{17}$ FLOP/sec, based on multiple approaches [Open Philantropy, 2020].
- $10^{16} - 10^{21}$ FLOP/sec: 1 bio neuron $\approx$ 1000s DNN neurons [Beniaguev$^+$, 2021].

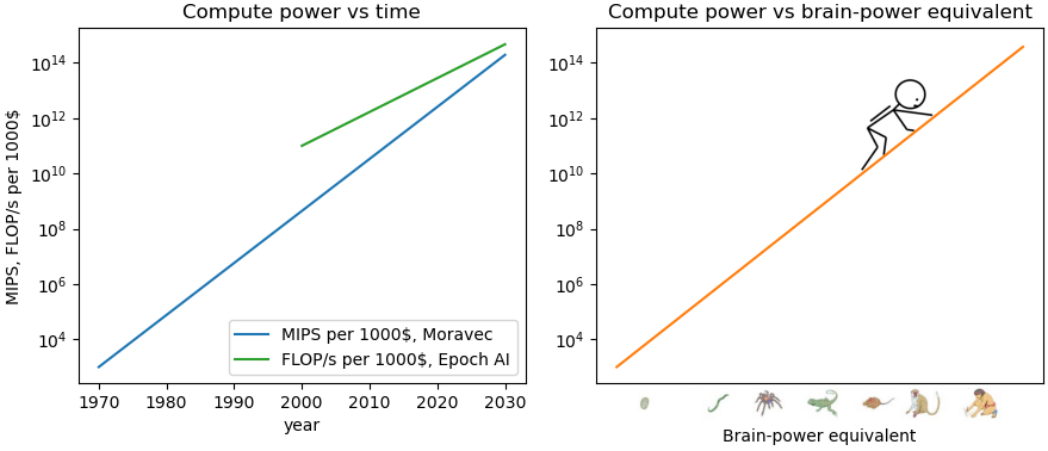
How did Moravec's prediction pan out?



IQ test results, based on Mensa IQ [Tracking AI, 2025].

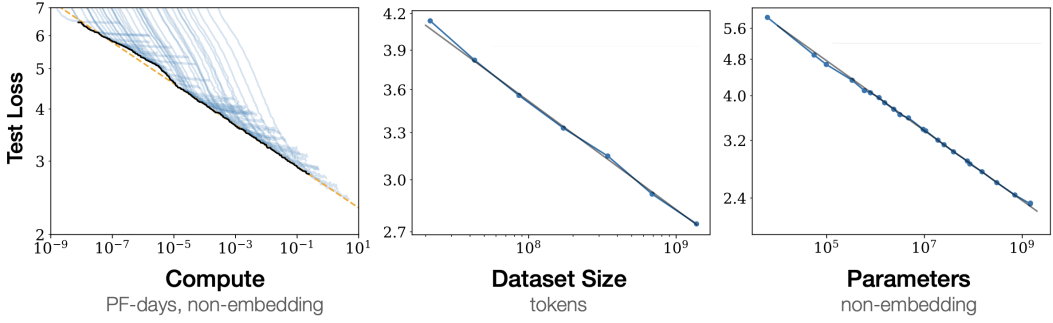
AI = Large Language Models.

Moravec's prediction = 2 scaling laws



"performance of AI machines tends to improve at the same pace that AI researchers get access to faster hardware" [Hans Moravec, 1998].

Scaling Laws for Neural Language Models [OpenAI, 2020]



$$L(C) = \left(\frac{C_c}{C}\right)^{\alpha_C}$$

$\alpha_C \approx 0.076$
$C_c \approx 3.1 \times 10^8$ [PF days]

$$L(D) = \left(\frac{D_c}{D}\right)^{\alpha_D}$$

$\alpha_D \approx 0.095$
$D_c \approx 5.4 \times 10^{13}$ [tokens]

$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}$$

$\alpha_N \approx 0.050$
$N_c \approx 8.8 \times 10^{13}$ [params]

"For optimal performance all three factors must be scaled up in tandem".

Chinchilla scaling [DeepMind, 2022]

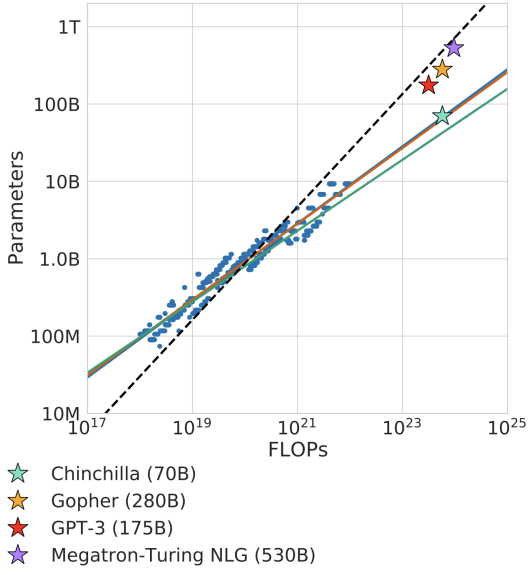
Large models need to be trained over more data to reach their best performance.

$\Rightarrow$ Current large models should be substantially smaller and therefore trained much longer than is currently done.

$\Rightarrow$ Chinchilla uses substantially less compute for fine-tuning and inference.

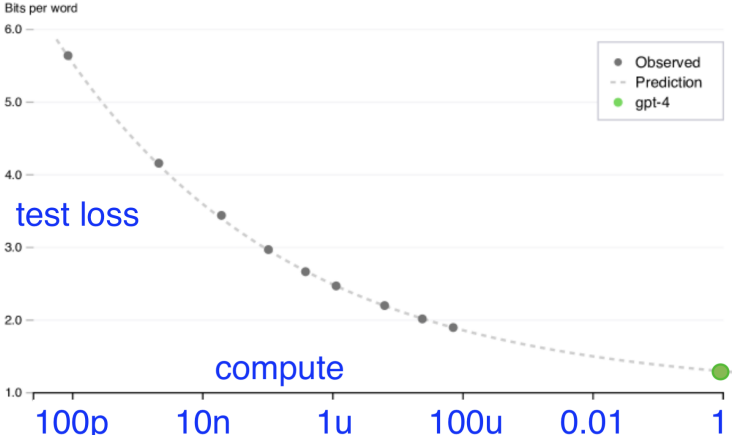
$$L(N, D) = E + \frac{A}{N^{0.34}} + \frac{B}{D^{0.28}} \ ,$$

$E = 1.69, \ A = 406.4, \ B = 410.7.$



Scaling and the age of pretraining

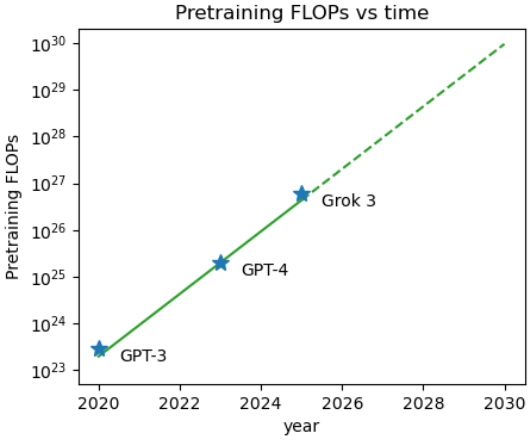
GPT4 performance was predicted using scaling laws based on 0.01% of the training FLOPs. [OpenAI, 2023]

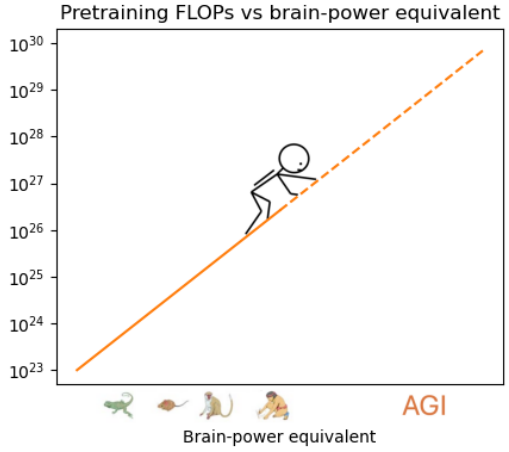


- What are the other 99.99% of the FLOPs good for?
- What is the value of reducing the test loss from 1.9 to 1.3?

"This is what has been the driver of all progress we see today - extraordinarily large neural networks trained on huge datasets." - Ilya Sutskever, co-founder OpenAI.

Scaling another 10,000× ?





Scaling and the age of pretraining

The sceptics:

- "scaling has hit a plateau";
  "scaling has hit a wall".
- Ilya Sutskever:
  "we have achieved peak data",
  "there is only one internet",
  "pretraining as we know it will end".

The believers:

- Sam Altman (CEO OpenAI):
  "there is no wall".
- Dario Amodei (CEO Anthropic):
  "scaling is probably... going to continue".

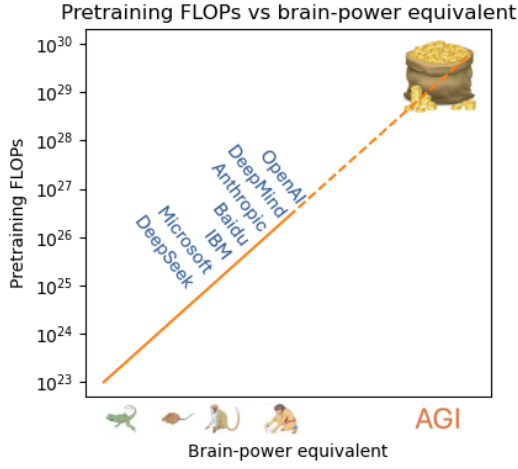
Beyond pretraining:

- Task-specific fine-tuning, domain adaptation, personalization, ...
- Reinforcement learning from human feedback (RLHF).
- Mixture of Experts (MoE).
- Chain of thought (CoT) prompting.
- Rejection sampling Fine-Tuning (RFT).
- ...

Summit fever

Laying siege to mount AGI:

- Microsoft, Alphabet, Amazon, and Meta plan to invest $320B in data centres [2025] [...] to remain at the forefront of AI ... [Financial Times, 2025]
- The Stargate Project is a new company which intends to invest $500B over the next four years in AI ... [OpenAI, 2025]



Thus the first ultraintelligent machine is the last invention that man need ever make. [I.J. Good, 1965]

Artificial General Intelligence (AGI)

Levels of AGI for operationalizing progress on the path to AGI [Google Deepmind, 2024].

	performance	skilled adults	narrow AI	general AI
0	no AI		calculator SW	Amazon Mechanical Turk
1	emerging		simple rule-based systems	ChatGPT, Llama, Gemini
2	competent	$\geq$50%	Siri, Alexa, Watson, ...	-
3	expert	$\geq$90%	Grammarly, Dall-E	-
4	virtuoso	$\geq$99%	Deep Blue, AlphaGo	-
5	superhuman	$\geq$100%	AlphaFold, AlphaZero	(artificial superintelligence)

An Approach to Technical AGI Safety and Security [Google Deepmind, 2025].

Can AI scaling continue through 2030?

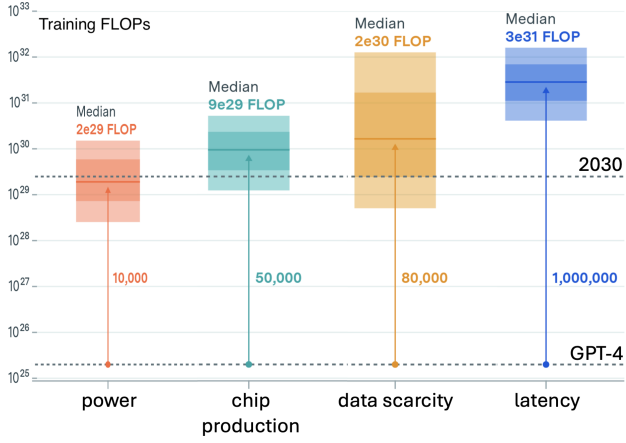
Stargate [AI data center, 2025]
- Microsoft, OpenAI, ..
- $100bn, 5GW.

Amazon investing in SMRs to deploy 5GW by 2039.

Meta seeks nuclear power for AI, data center support: 4GW by 2030.

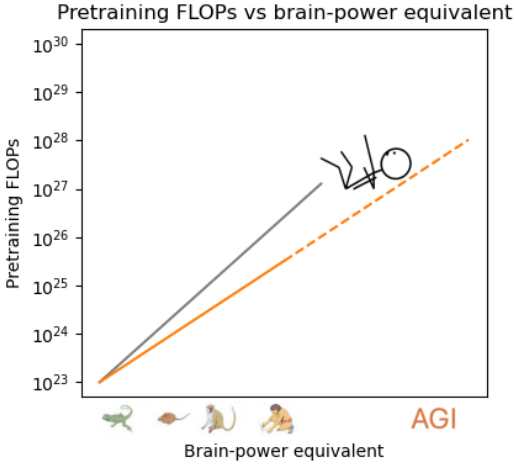
...

Money is not seen as a constraint. "AI may deliver explosive growth", [Open Philantropy, 2021].

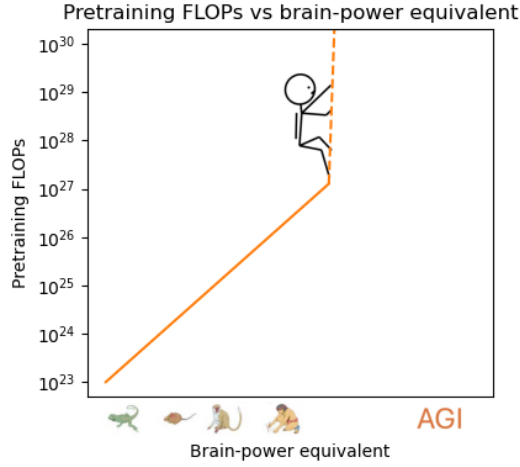


Constraints to scaling training runs by 2030. [EPOCH AI, 2024]

Scaling disruptions?



A deeper-seek disruption?



A steeper-seek disruption?

Scaling Mount AI: summary

AI history:

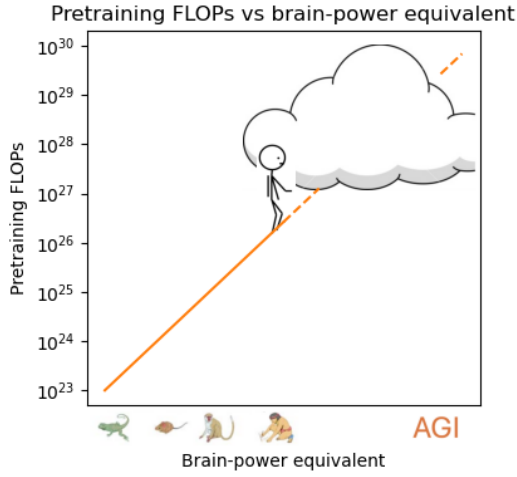
- Moravec: "brain-power equivalent" scales with compute resources.
- Scaling and "the age of pretraining" delivered todays AI state of the art.

AI today:

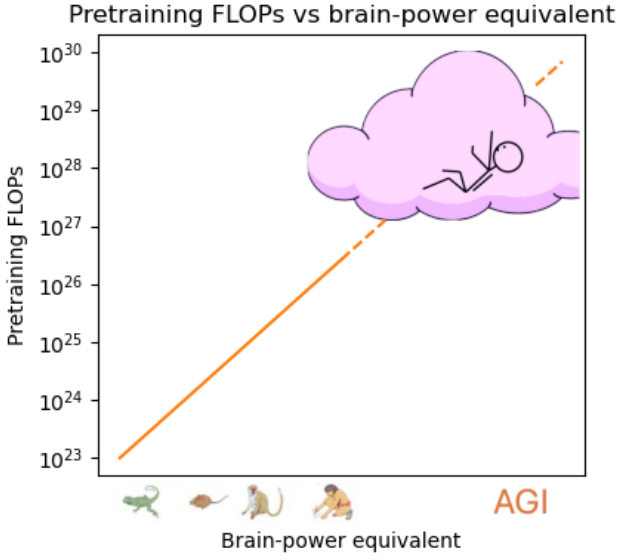
- Doubts about the future of AI scaling.
- AI $\gg$ pretraining.

AI future (industry consensus):

- AI scaling likely to continue to 2030.
- AGI is expected to occur before 2030.



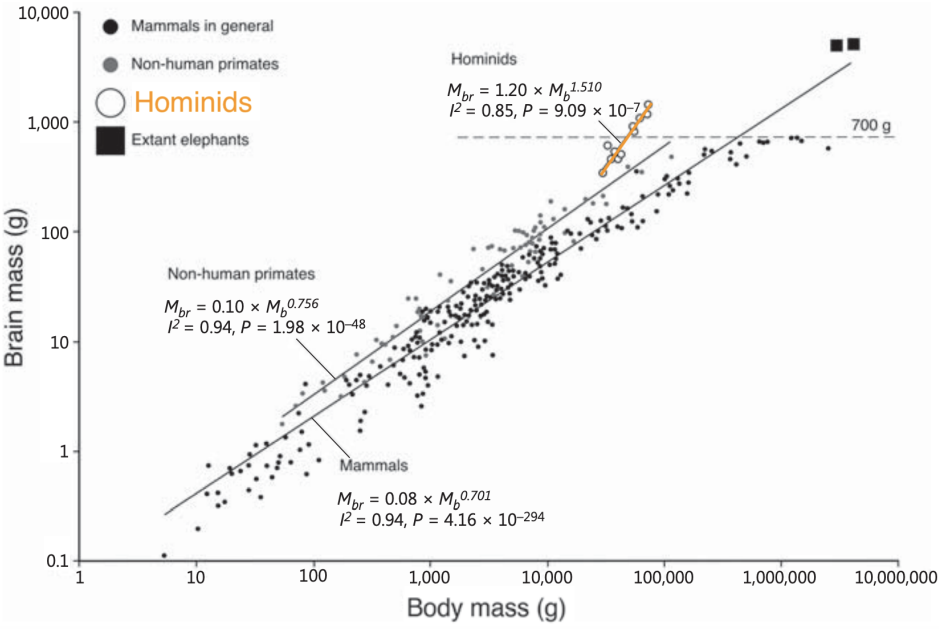
Summit fever?



High-altitude oxygen deprivation is a known cause of hallucination and delirium.

Thank you.

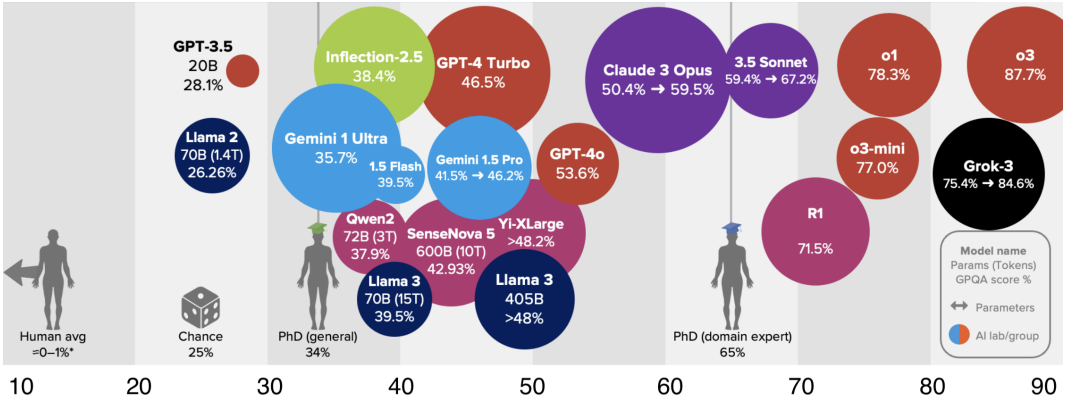
Backup 1: The Evolutions of Large Brain Size in Mammals



[Manger+, 2013]

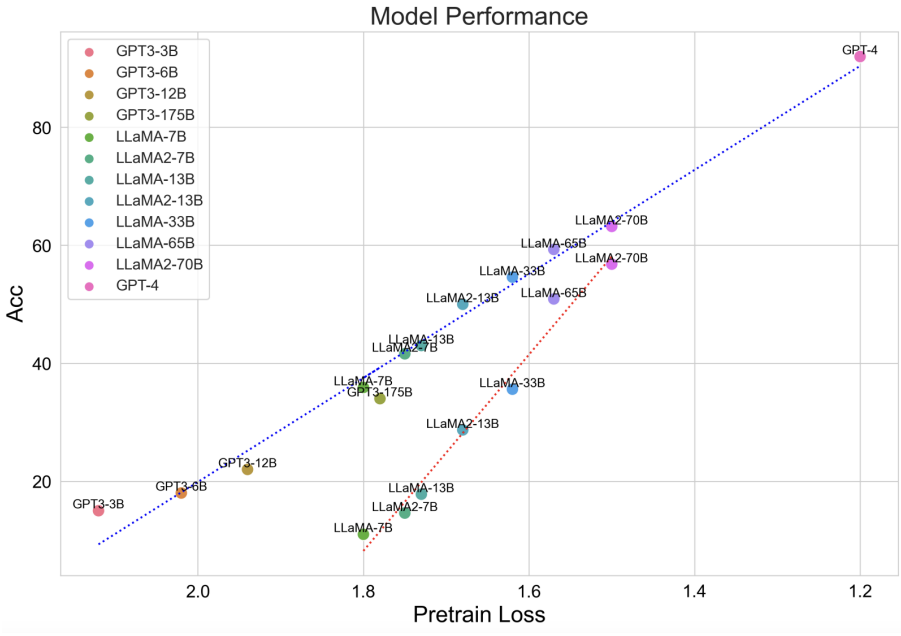
Backup 2: Google-Proof Q&A Benchmark

GPQA is a challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry. The questions are high-quality, extremely difficult, and "Google-proof" [paper, 2023].



QPGA score [LifeArchitect, 2025]

Backup 3: Mathematical reasoning performance vs pretraining loss



[Alibaba, 2023]

Backup 4: Ultraintelligent machines

Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make. [I.J. Good, 1965]

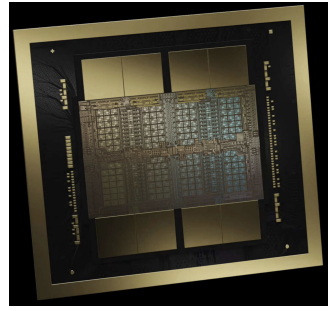
"Extrapolating the spectacular performance of GPT-3 into the future suggests that the answer to life, the universe and everything is just 4.398 trillion parameters" [Geoffrey Hinton, 2020].

Backup 5: AI factory *Stargate UAE* (Abu Dhabi)

A 1GW Stargate UAE cluster in Abu Dhabi with 200MW expected to go live in 2026 [OpenAI, 2025].

NVIDIA GB300 NVL72 = 72x Blackwell B300 GPU:
Designed for AI reasoning performance .. in AI factories.

The estimates below are from [NextPlatform, 2025].



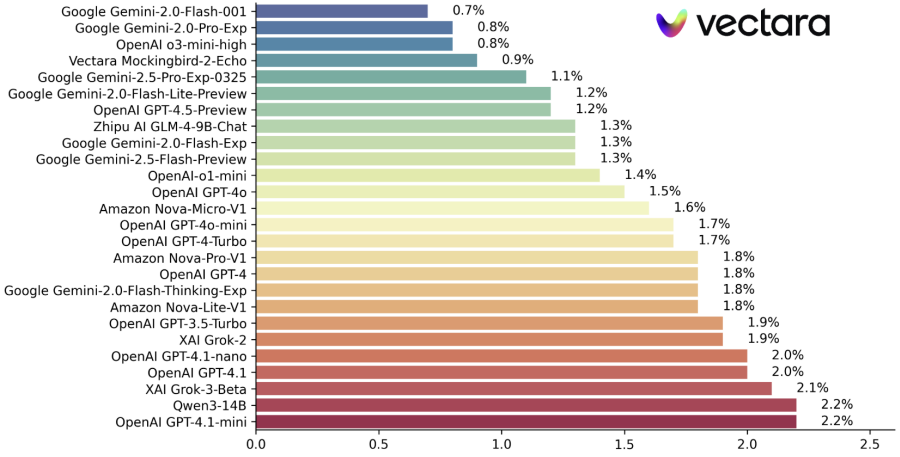
Blackwell B200 GPU

	GPU	HBM3	FLOPs/s	power	costs
Blackwell B300 GPU	1	288 GB	10^{15}		
NVIDIA GB300 NVL72 (rack)	72	20 TB	10^{17}	120kW	
20% Stargate UAE	100k	28 PB	10^{21}	200MW	7^{+} B$

Four months of pretraining at 20% capacity: $\approx 10^{28}$ FLOPs.

Backup 6: AI hallucination

"AI hallucinations are getting worse – and they're here to stay".
New Scientist, 2025 May.



Grounded hallucination rates for top-25 LLMs, 2025 April [github].
AI models collapse when trained on recursively generated data [paper, 2024].

Backup 7: my favorite books on deep learning and on AI