Degenerative AI?

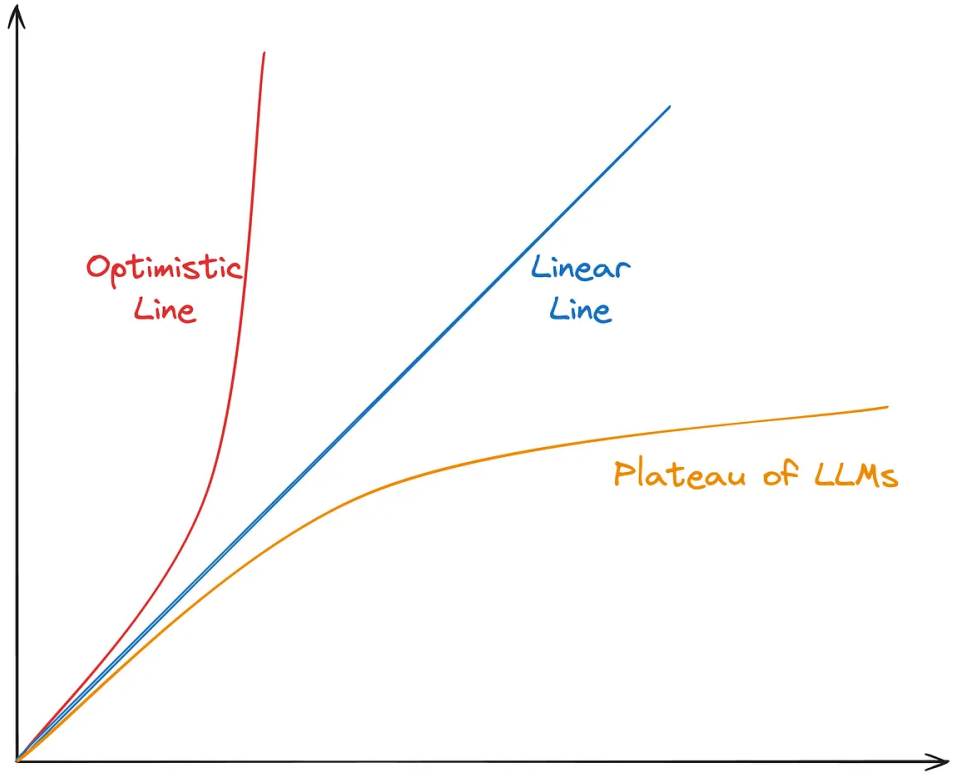
If GenAI Has Already Hit its Limit,

What is its Next Evolution?

Dr. Sabri Boughorbel

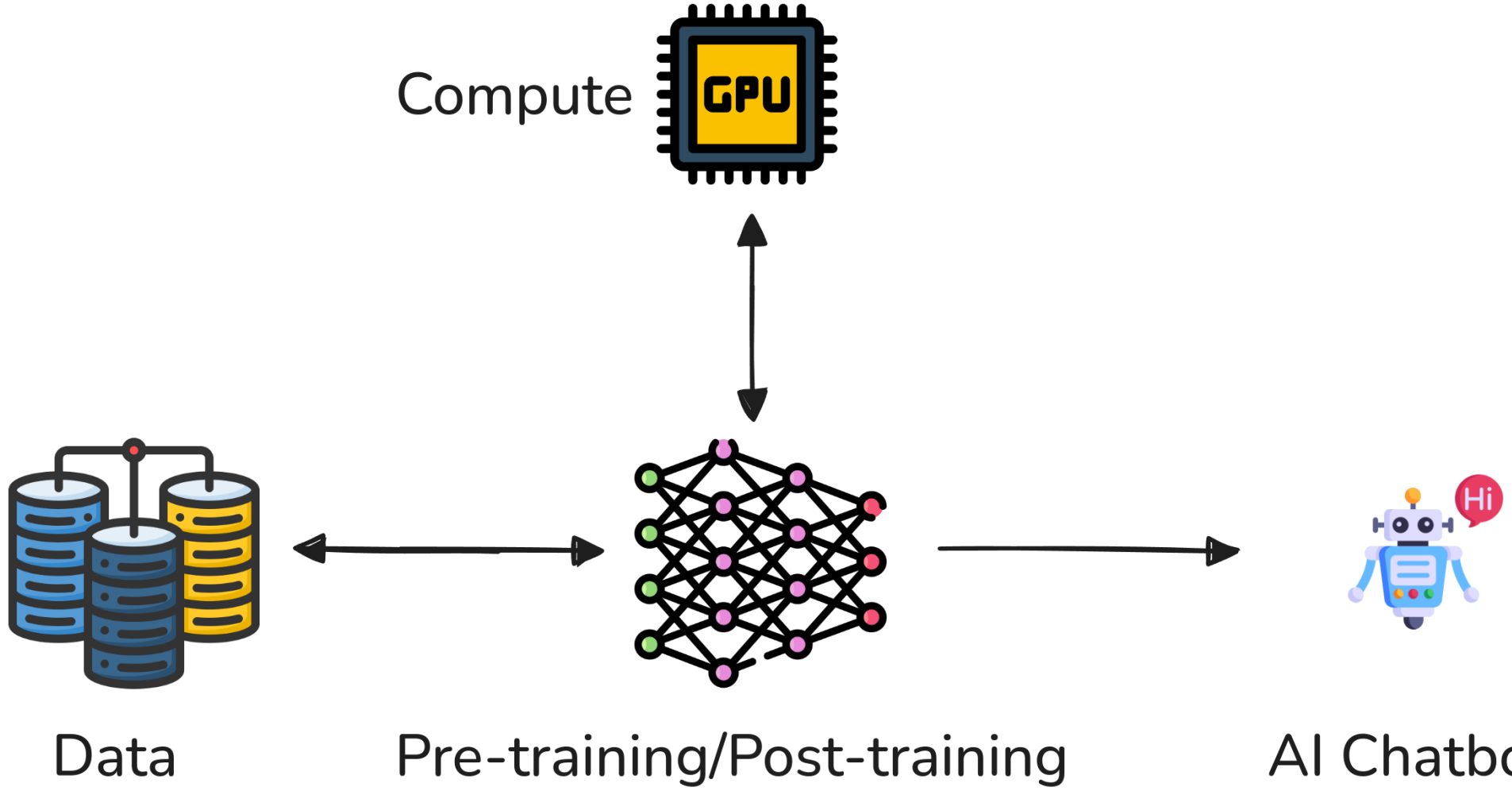
Outline

- Review of AI progress
- Challenges on the road of AI
- Promising directions



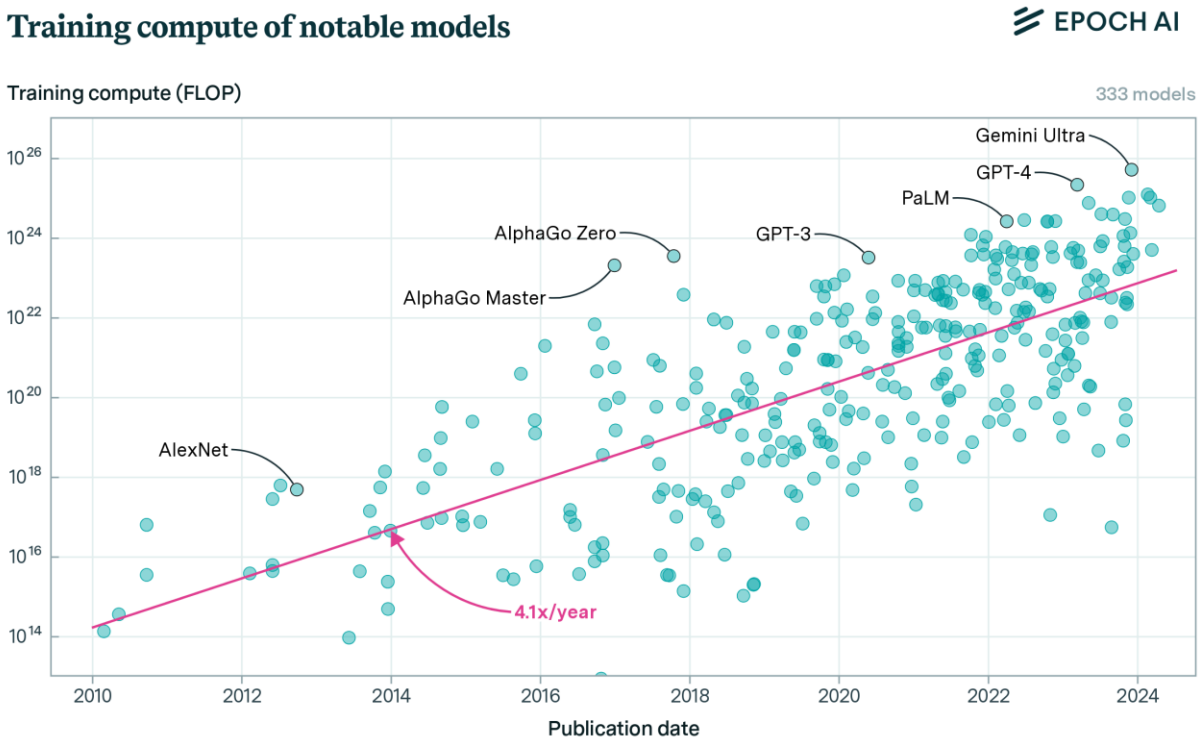
https://nishu-jain.medium.com/are-llms-hitting-a-plateau-c8e185d0992e

GenAI Under the Hood



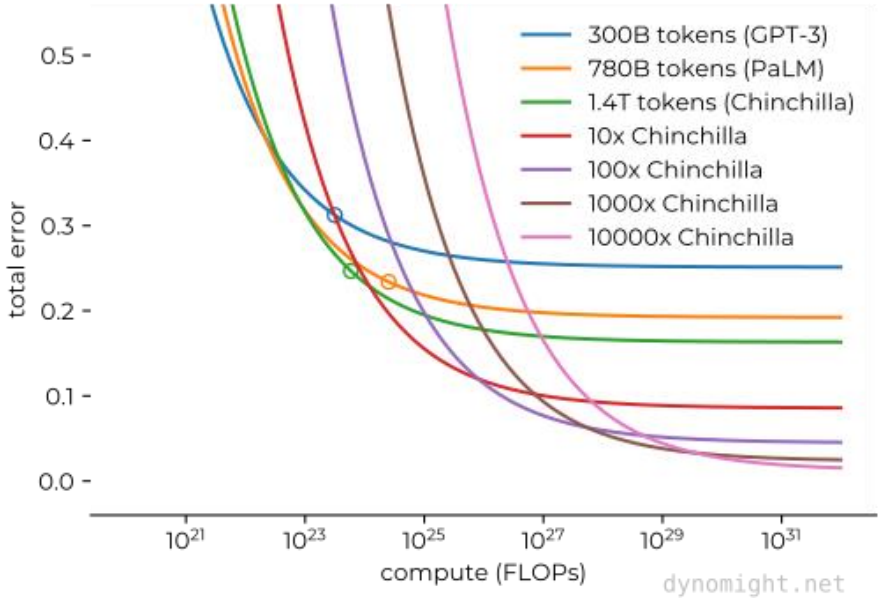
AI Progress

- Compute expanding at 4x per year
- Fastest technological expansions in recent history
  - Mobile adoption (2x/year)
  - Solar energy capacity (1.5x/year)
  - Human genome sequencing (3.3x/year)
  - Moore's Law (1.3x/year)

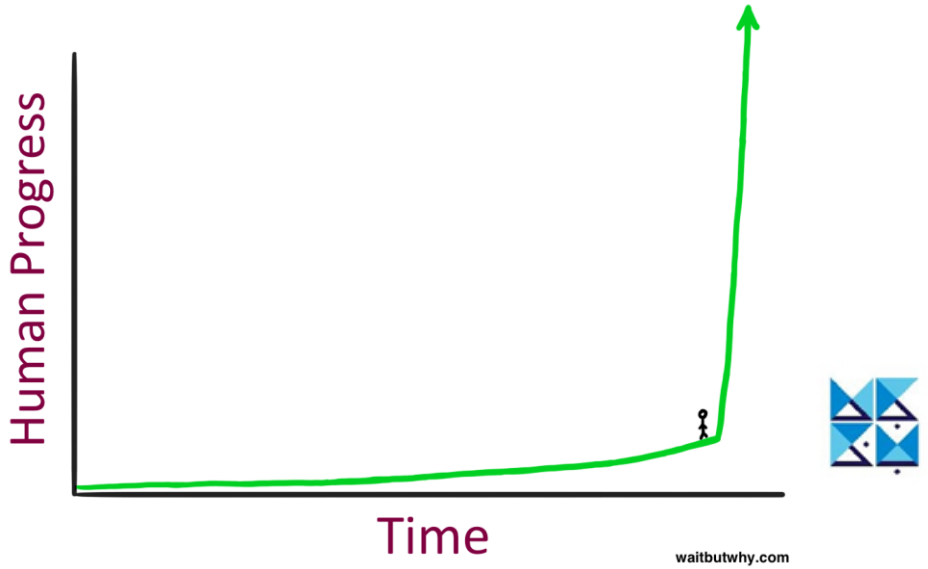


AI Progress

- AI Performance is predictable
  - Scaling of model, data and compute

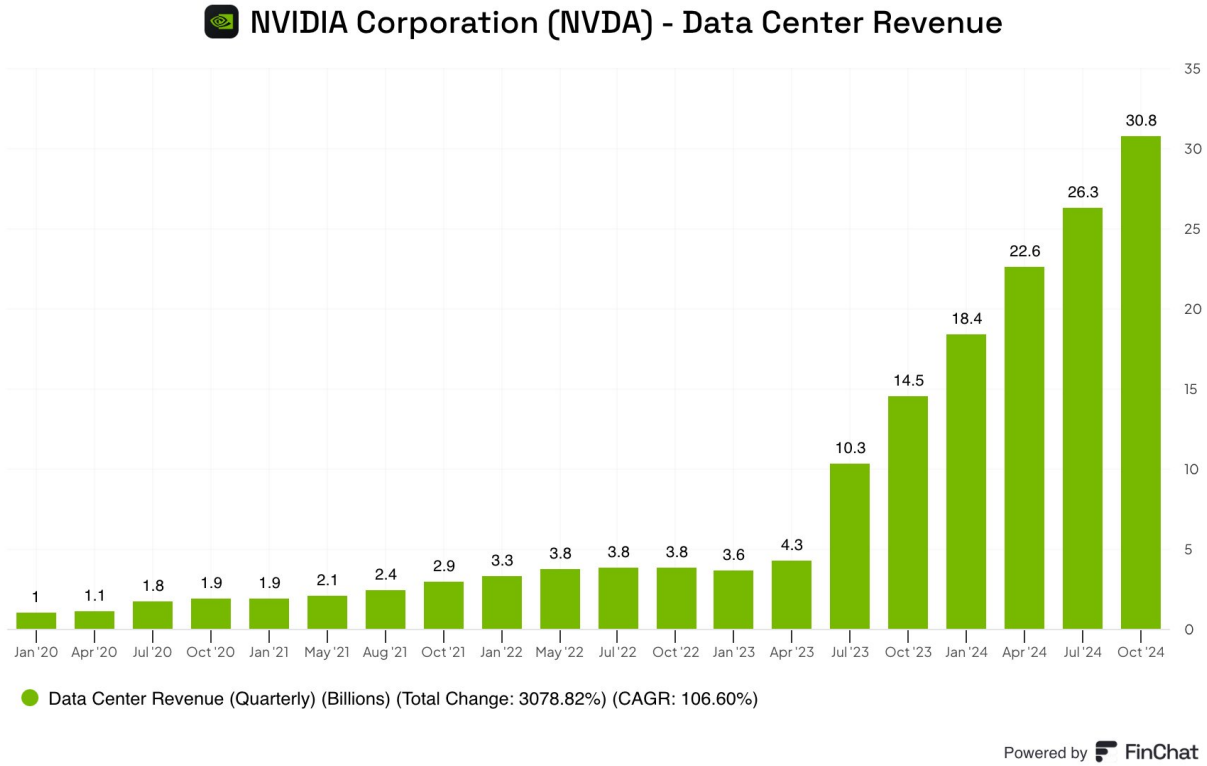


- Would this lead to capability explosion ?
  - Task automation
  - Fast economy growth



Monetization of AI

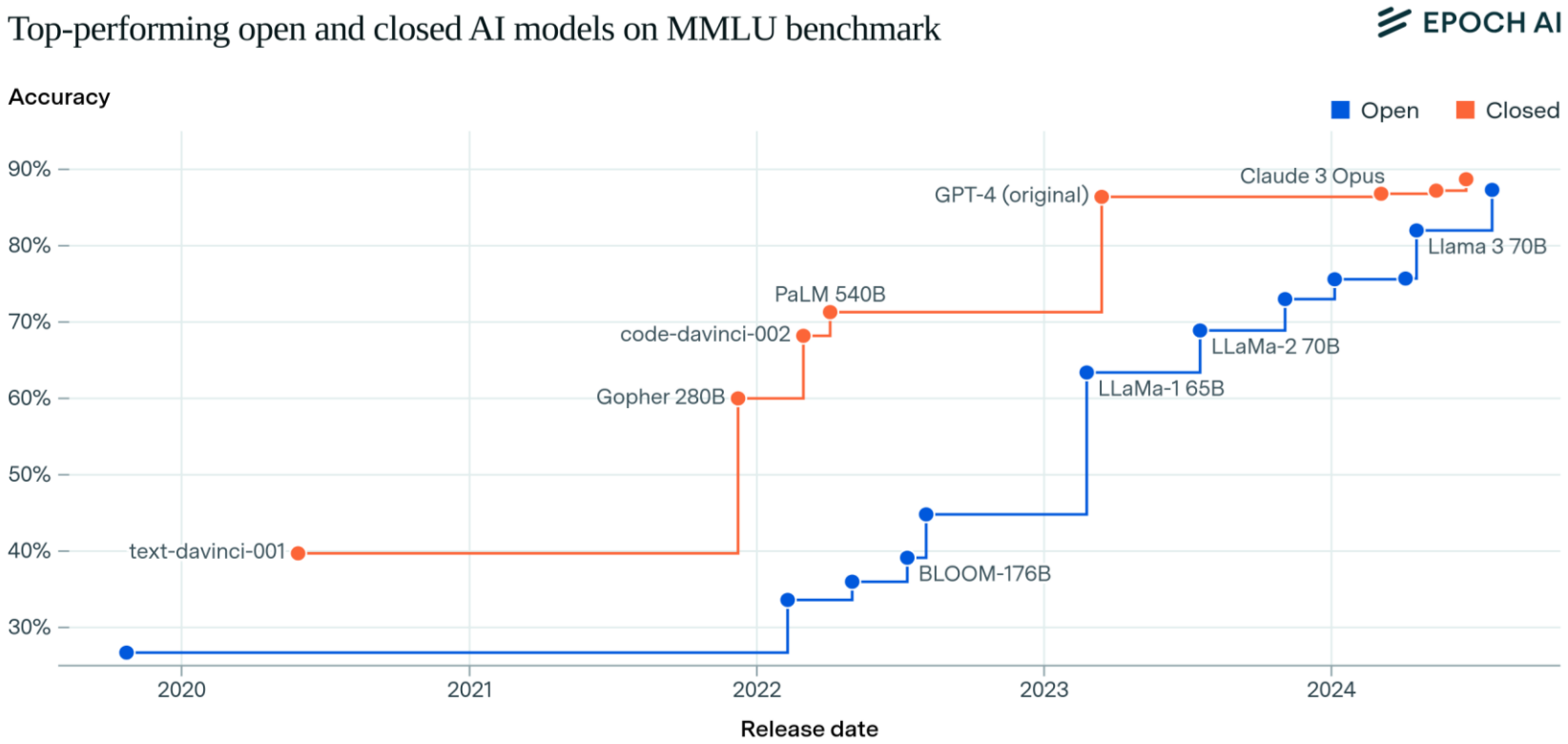
- **Nvidia**:
  - ~100B annual revenue projected for Nvidia from data center
- **OpenAI**:
  - $1B revenue in Aug 2023
  - $2B revenue in Feb 2024
  - doubling every ~6 months
- **Microsoft**:
  - ~$5B estimation of incremental AI revenue



Models are Reaching a "ceiling" ?

REUTERS

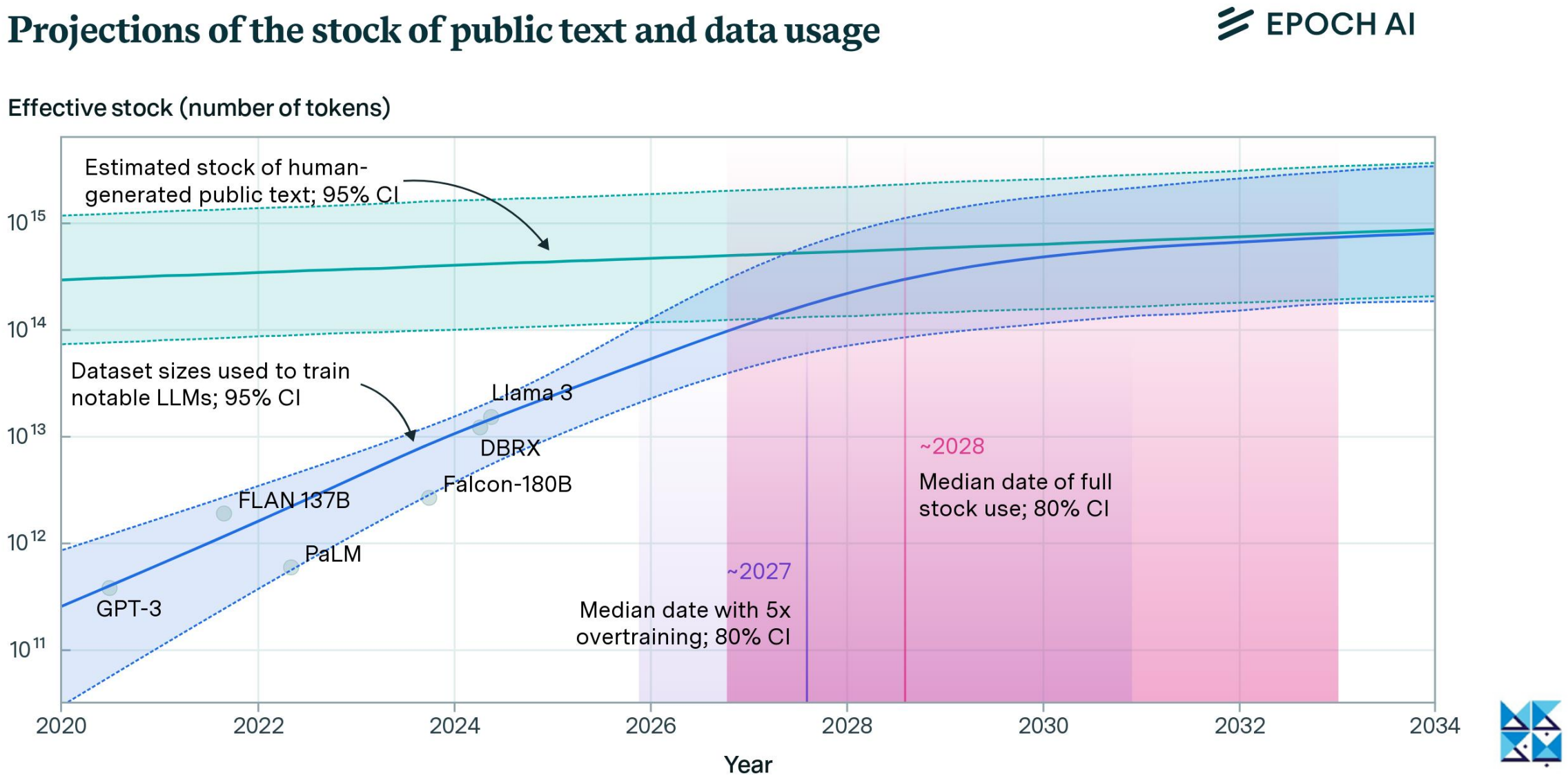
*"...**Ilya Sutskever**, co-founder of AI labs Safe Superintelligence (SSI) and OpenAI, told Reuters recently that results from scaling up pre-training - the phase of training an AI model that uses a vast amount of unlabeled data to understand language patterns and structures* ***have plateaued****."* Nov 11, 2024



Challenges

- Data Wall
- Scaling Training Runs

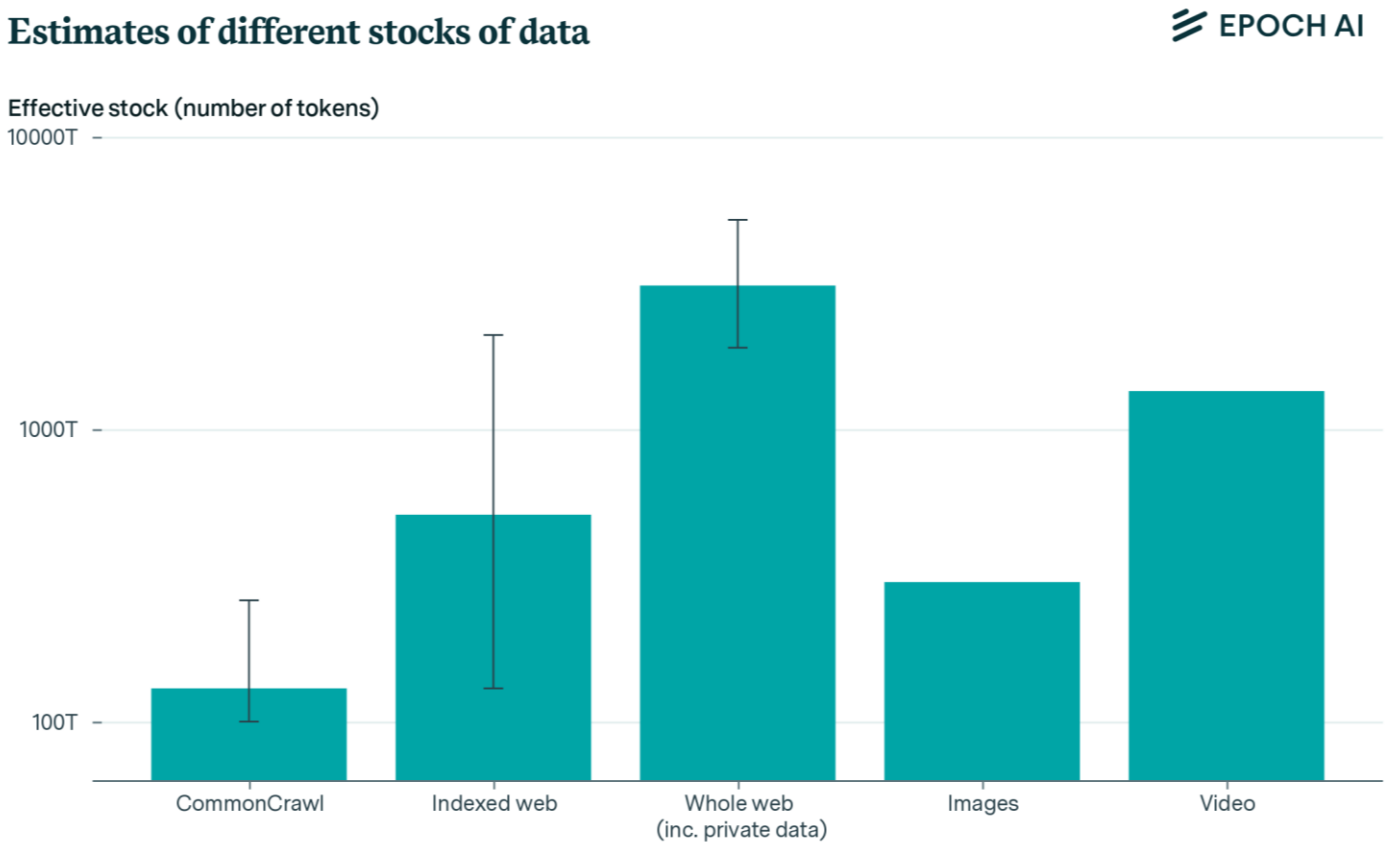
Will We Run Out of Data ?



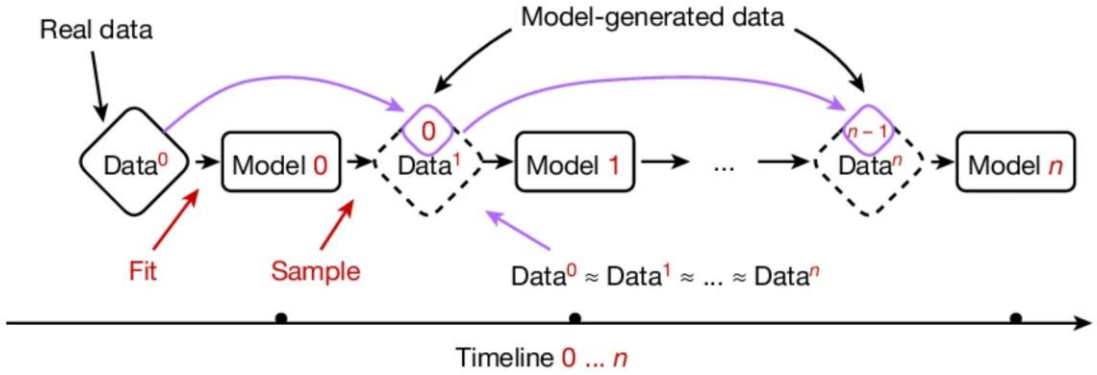
https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Will we run out of data ?

- Data from other modalities
- Enterprise data
- Synthetic data



Synthetic Data Could Lead to Degenerative AI

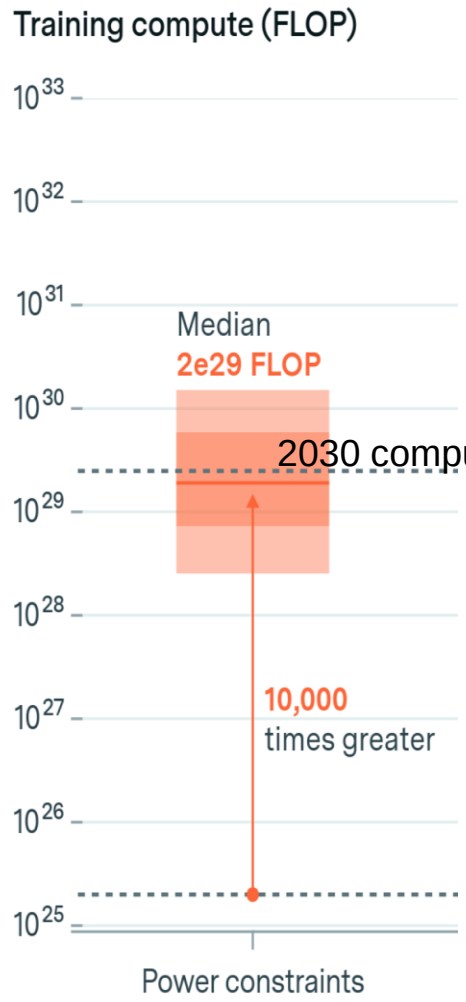


Shumailov, Ilia, et al. "AI models collapse when trained on recursively generated data." *Nature* 631.8022 (2024): 755-759.

Gibney, Elizabeth. "AI models fed AI-generated data quickly spew nonsense." *Nature* 632.8023 (2024): 18-19.

Scaling Training Runs

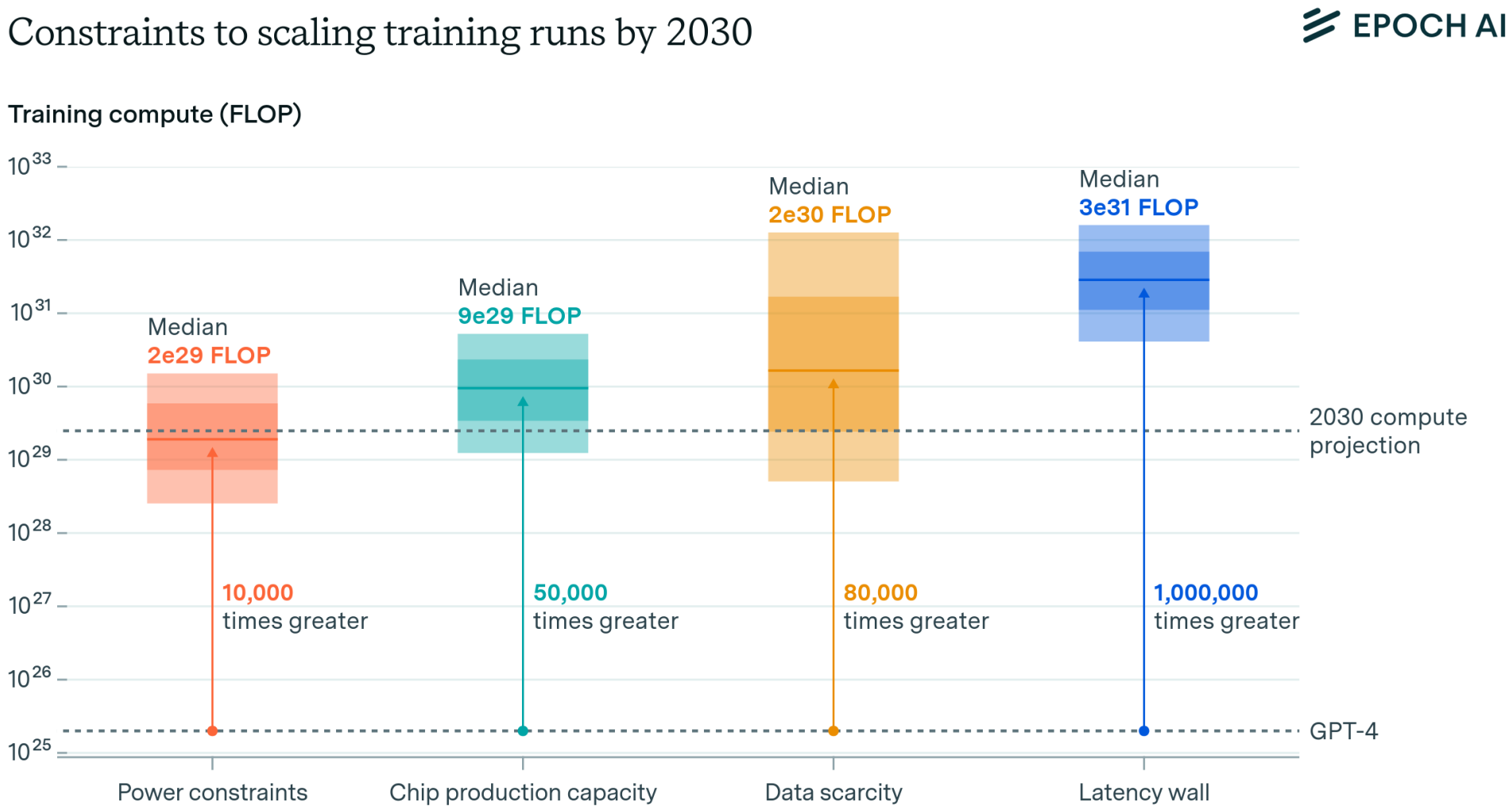
Year	OOMs	H100s-equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	$10s of billions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	$1T+	~100GW	>20% of US electricity production



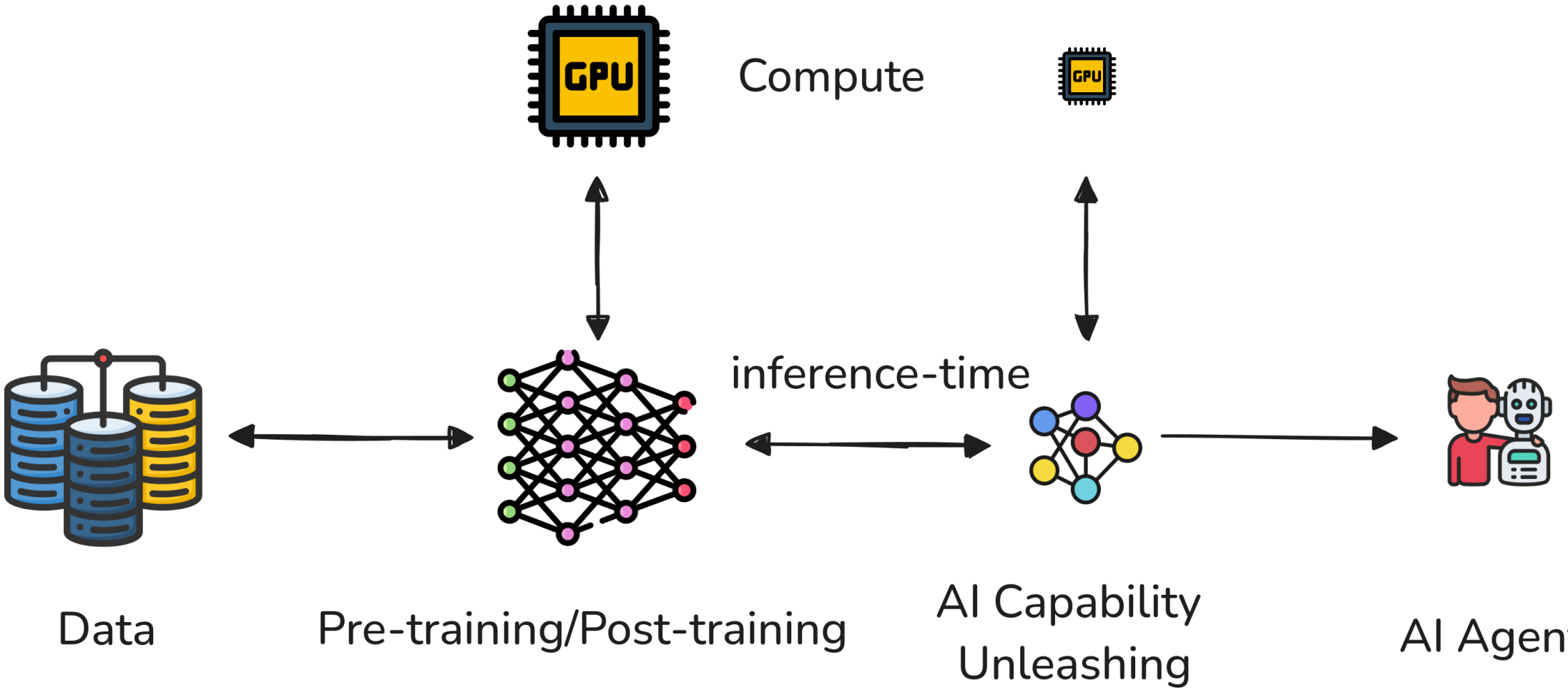
https://situational-awareness.ai/

https://epoch.ai/blog/can-ai-scaling-continue-through-2030

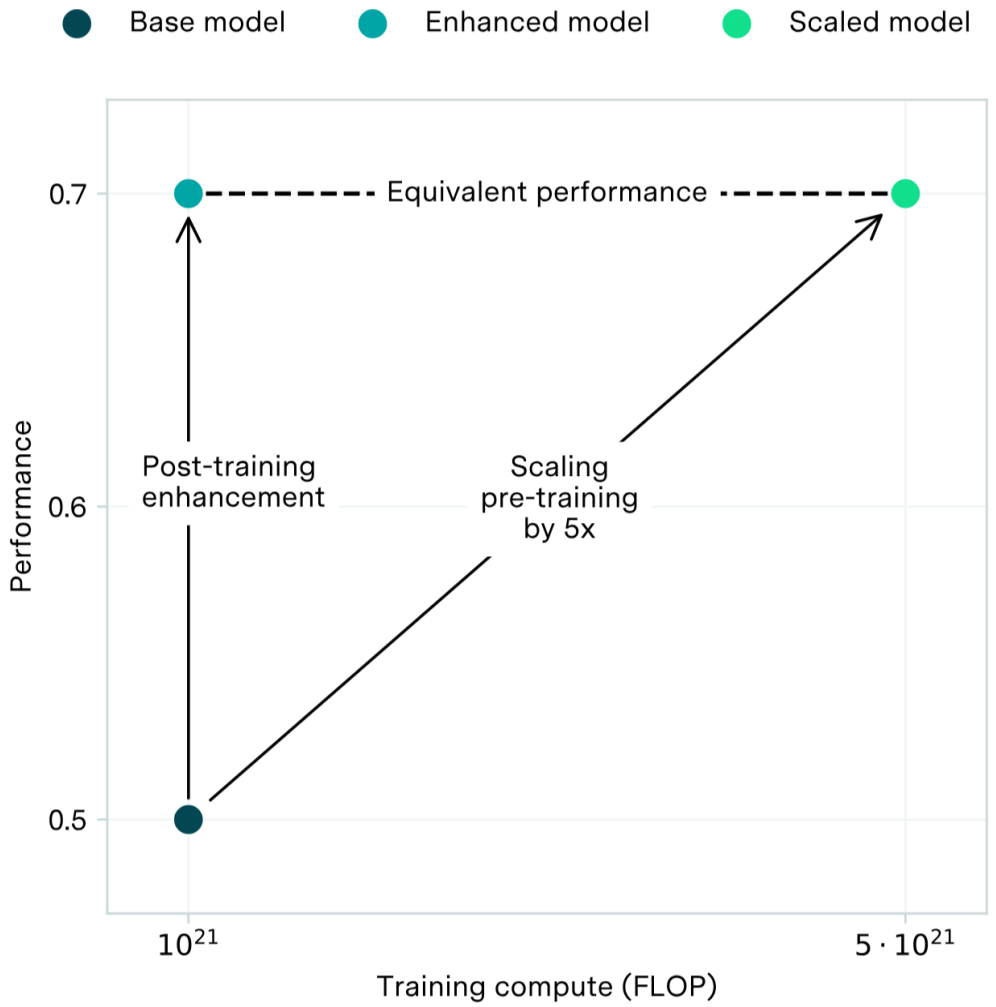
Scaling Training Runs



AI Capabilities Can be Improved Without Expensive Retraining



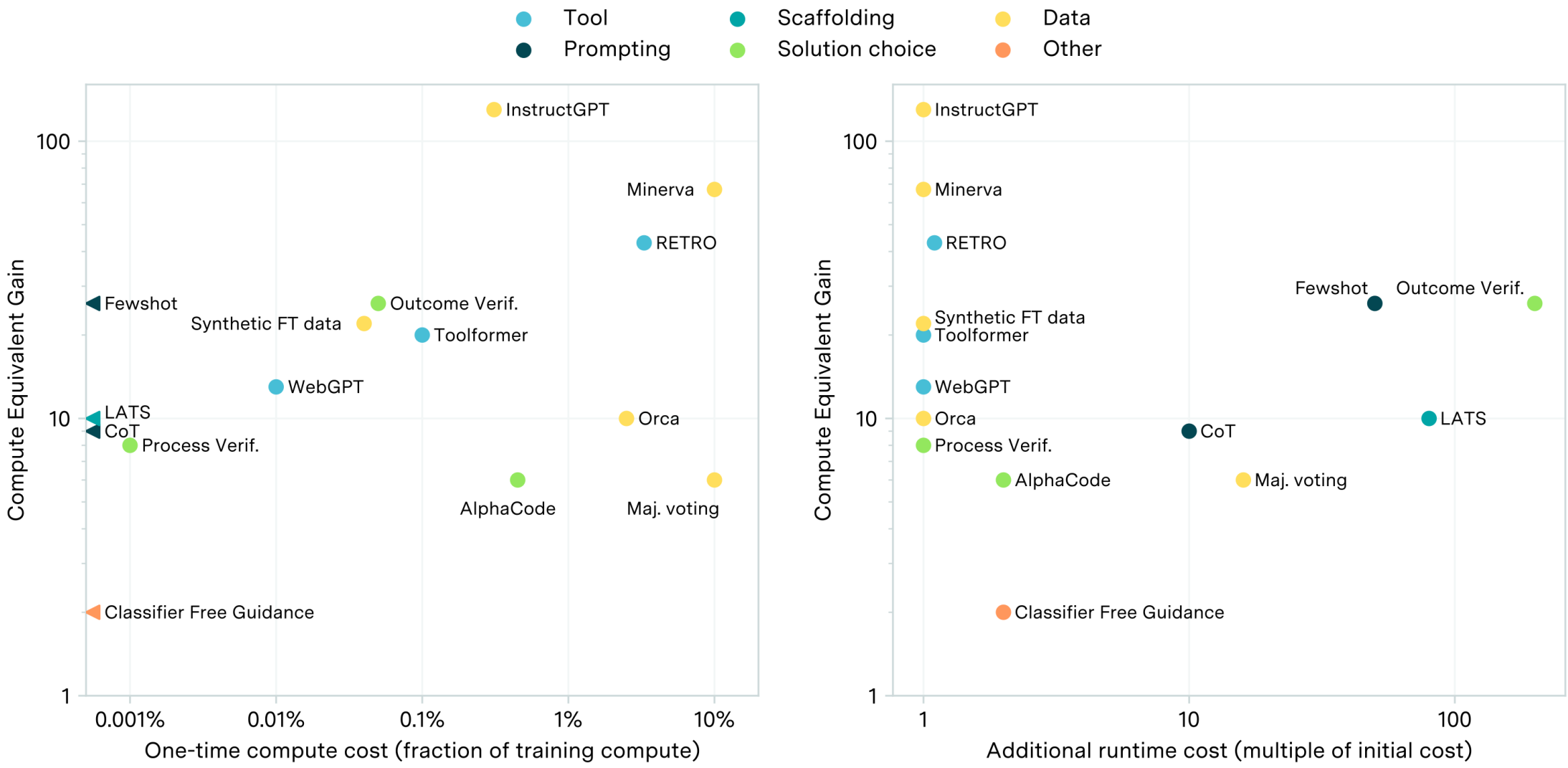
AI Capability Unleashing



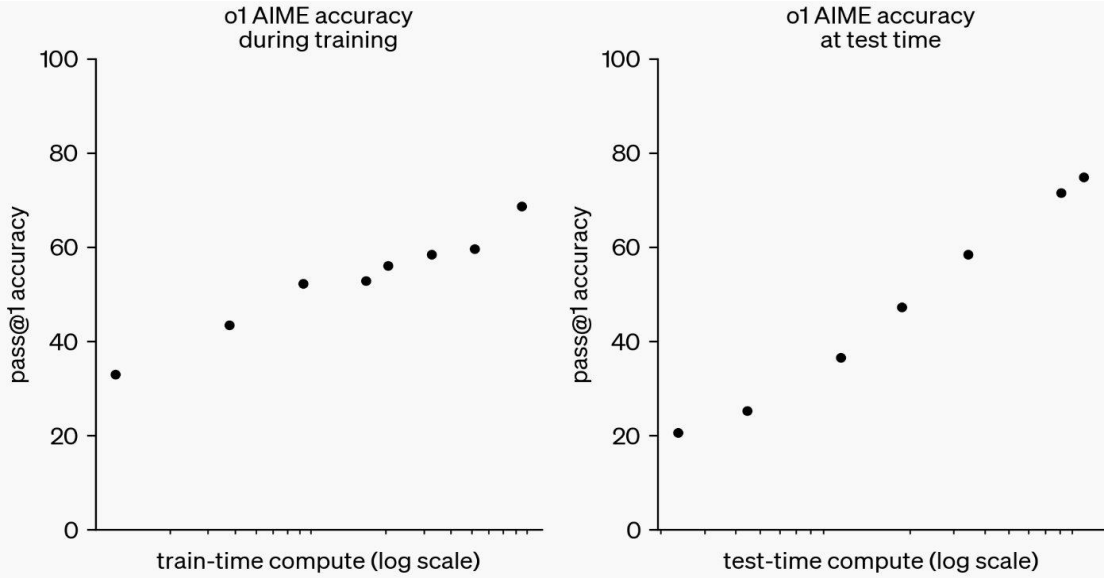
Compute Equivalent Gain

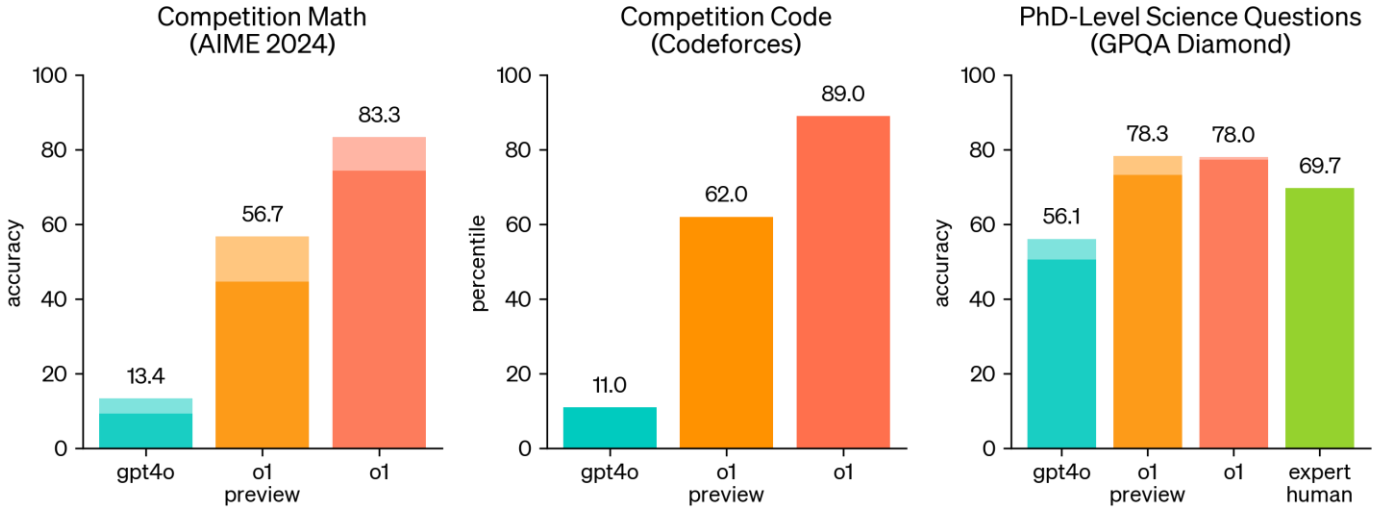
https://epoch.ai/blog/ai-capabilities-can-be-significantly-improved-without-expensive-retraining

AI Capability Unleashing



AI Capability Unleashing

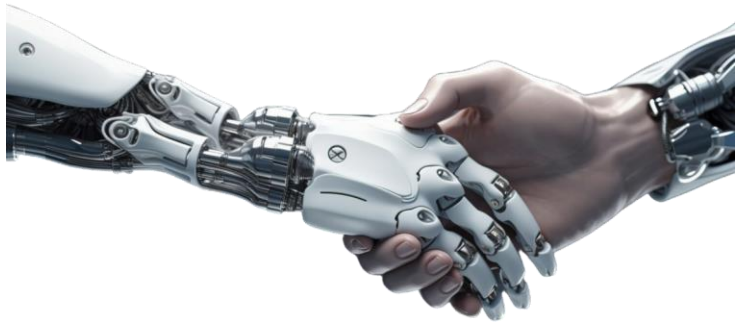




Test-time compute

https://openai.com/index/learning-to-reason-with-llms/

Conclusion



- Current AI Progress is dependent on scaling of data and compute
- Signs of saturation in pretraining performance
- New algorithmic techniques in inference time leading to significant improvement