Sputnik Moment, Democratization of AI, or Media Hype?

2025.02.25 - DigHum Lecture
Johannes Oster · johannes.oster@tuwien.ac.at

DeepSeek - Sputnik Moment, Democratization of AI, or Media Hype?

~800 lines of code

671B Floating-point numbers

Alex Cheema – e/acc

The Washington Post

DeepSeek has created a 21st-century Sputnik moment

Export controls may have helped turbocharge Chinese innovation.

January 31, 2025

App Store

Top Free Apps >

1 DeepSeek - AI Assistant — Update
Intelligent AI Assistant

2 ChatGPT — Open
The official app by OpenAI

3 Threads
Connect and share ideas

Google Finanzen

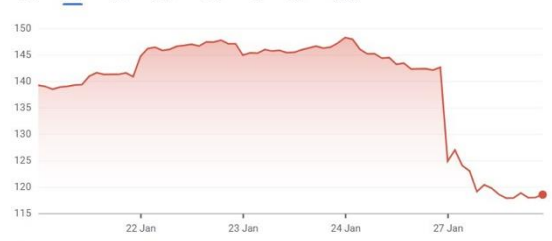
NVIDIA Corp

$118.58 ↓14.83% -20.64 5 D

After hours: $121.27 (↑2.27%) +2.69

Closed: 27 Jan, 19:30:04 UTC-5 · USD · NASDAQ · Disclaimer

1 D 5 D 1 M 6 M YTD 1 Y 5 Y MAX

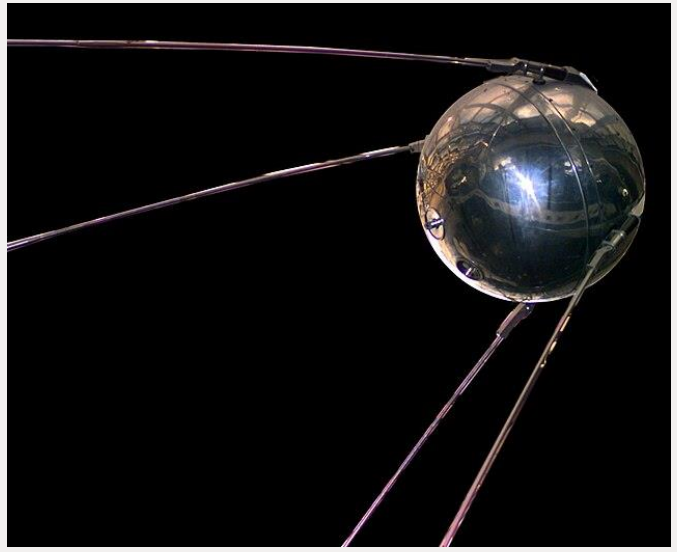


Bloomberg

IN. Capital's Scott: DeepSeek Heralds AI Democratization

January 28th, 2025, 6:38 AM GMT+0100

3 Chosen Characteristics of a Sputnik Moment



I. High Stake Field & Arms Race Dynamics

II. Debunking of Perceived Technological Supremacy

III. Strong Response from the West

Img src: defense360.csis.org

3 Chosen Characteristics of a DeepSeek Moment?

?

I. High Stake Field & Arms Race Dynamics

II. Debunking of Perceived Technological Supremacy

III. Strong Response from the West

I. High Stake Field & Arms Race Dynamics

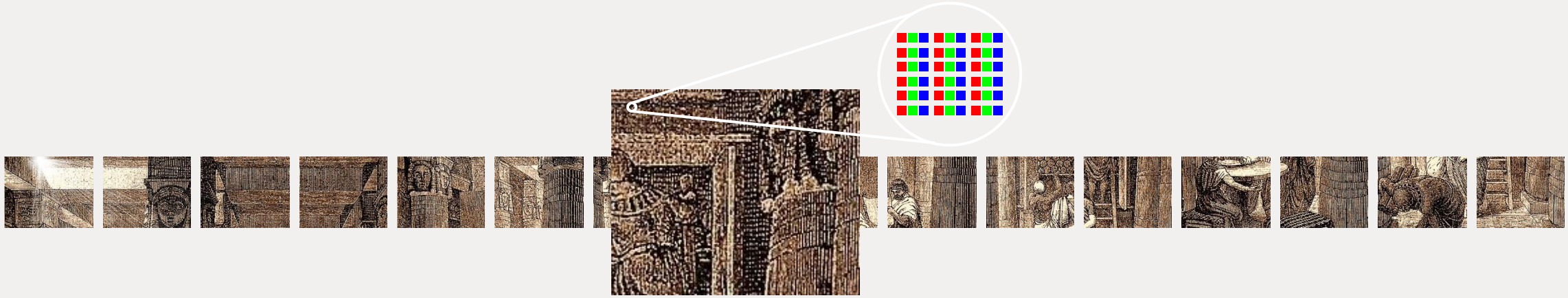
1. Language is a powerful vehicle and vital component of cross-domain intelligence.

Library of Alexandria – wikimedia commons

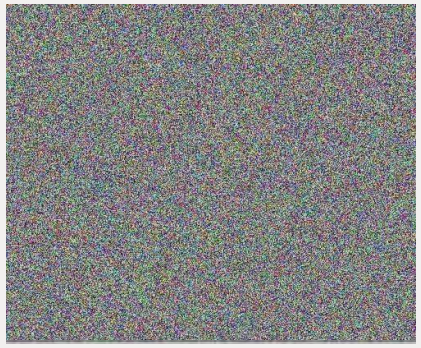
1. Computer Vision is a powerful vehicle and vital component of cross-domain intelligence.

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

1. Computer Vision is a powerful vehicle and vital component of cross-domain intelligence.



1. Image Generation is a powerful vehicle and vital component of cross-domain intelligence.

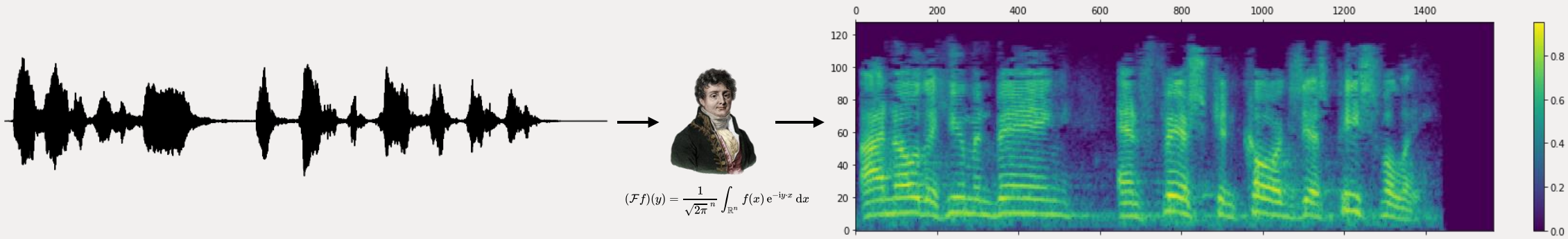


Scalable Diffusion Models with Transformers

1. Image Generation is a powerful vehicle and vital component of cross-domain intelligence.

Short Sidenote

1. Audio Processing is a powerful vehicle and vital component of cross-domain intelligence.



AST: Audio Spectrogram Transformer

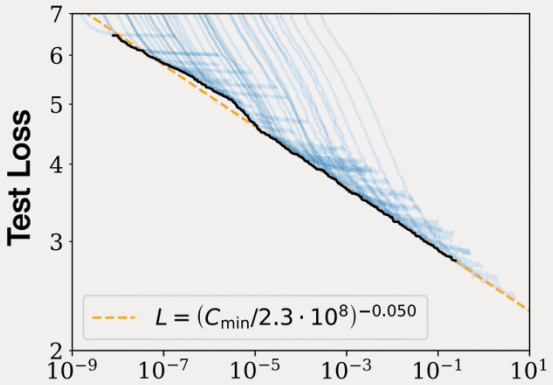
Img src: GitHub - Tim Sainburg, Wikipedia

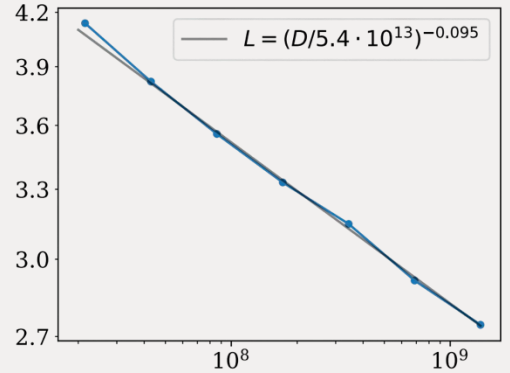
I. High Stake Field & Arms Race Dynamics

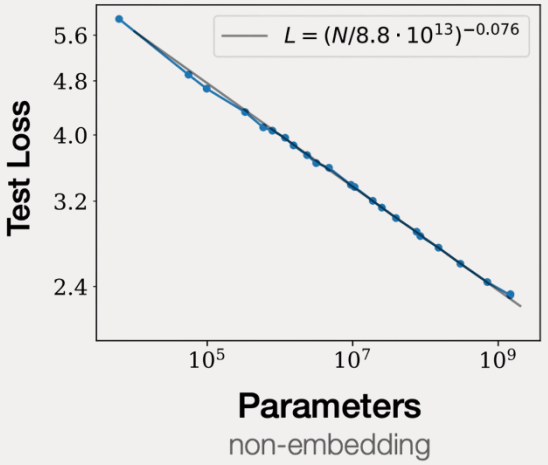
1. Language is a powerful vehicle and vital component of cross-domain intelligence.

Library of Alexandria – wikimedia commons

2. More parameters, data, and compute promise better performance.



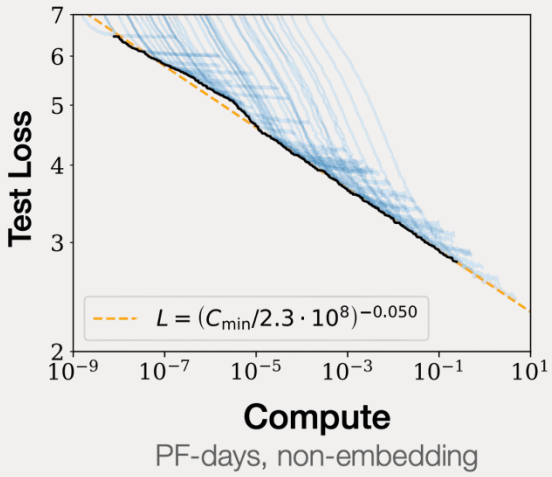


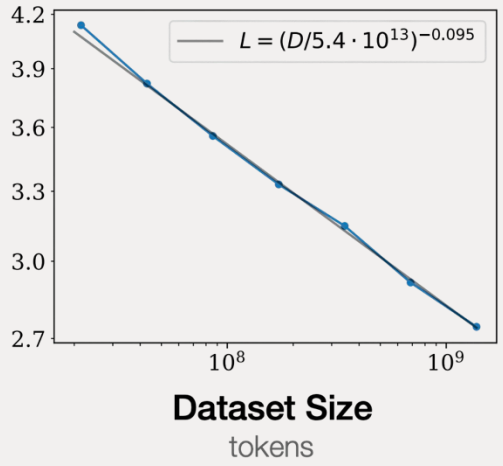


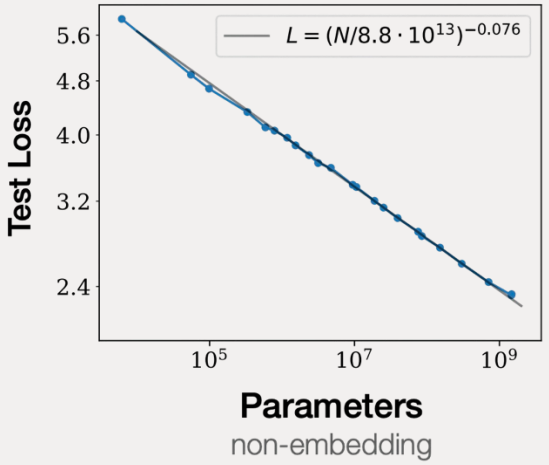
Kaplan et al. (2020), **Scaling Laws for Neural Language Models**.

https://arxiv.org/abs/2001.08361

2. More parameters, data, and compute promise better performance.

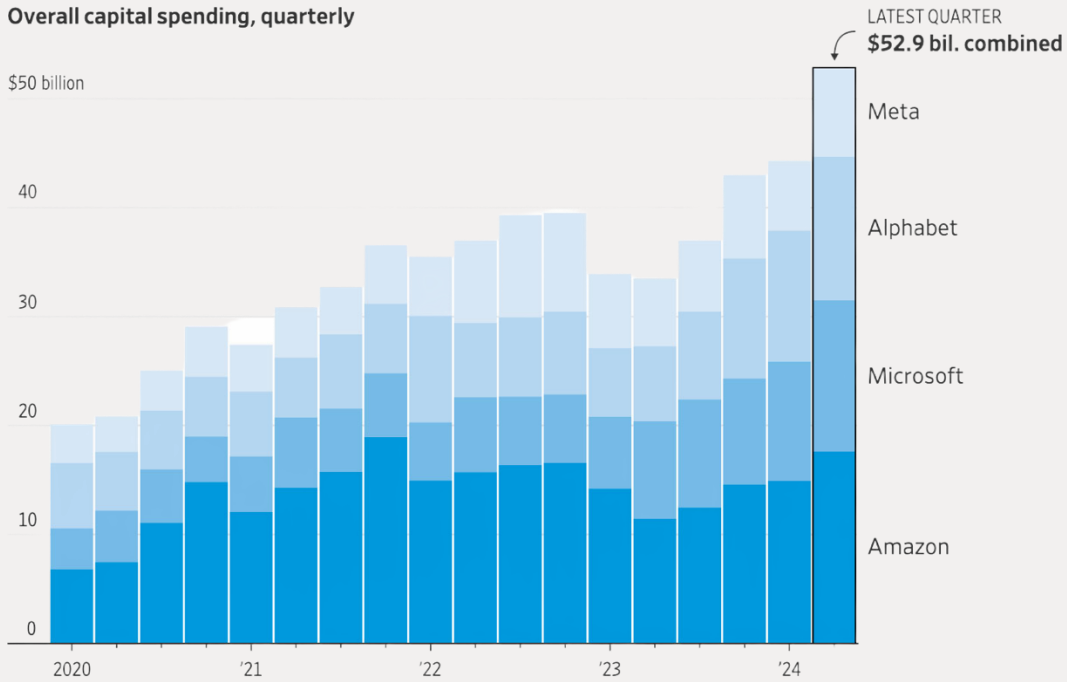






Kaplan et al. (2020),
Scaling Laws for Neural Language Models.
https://arxiv.org/abs/2001.08361

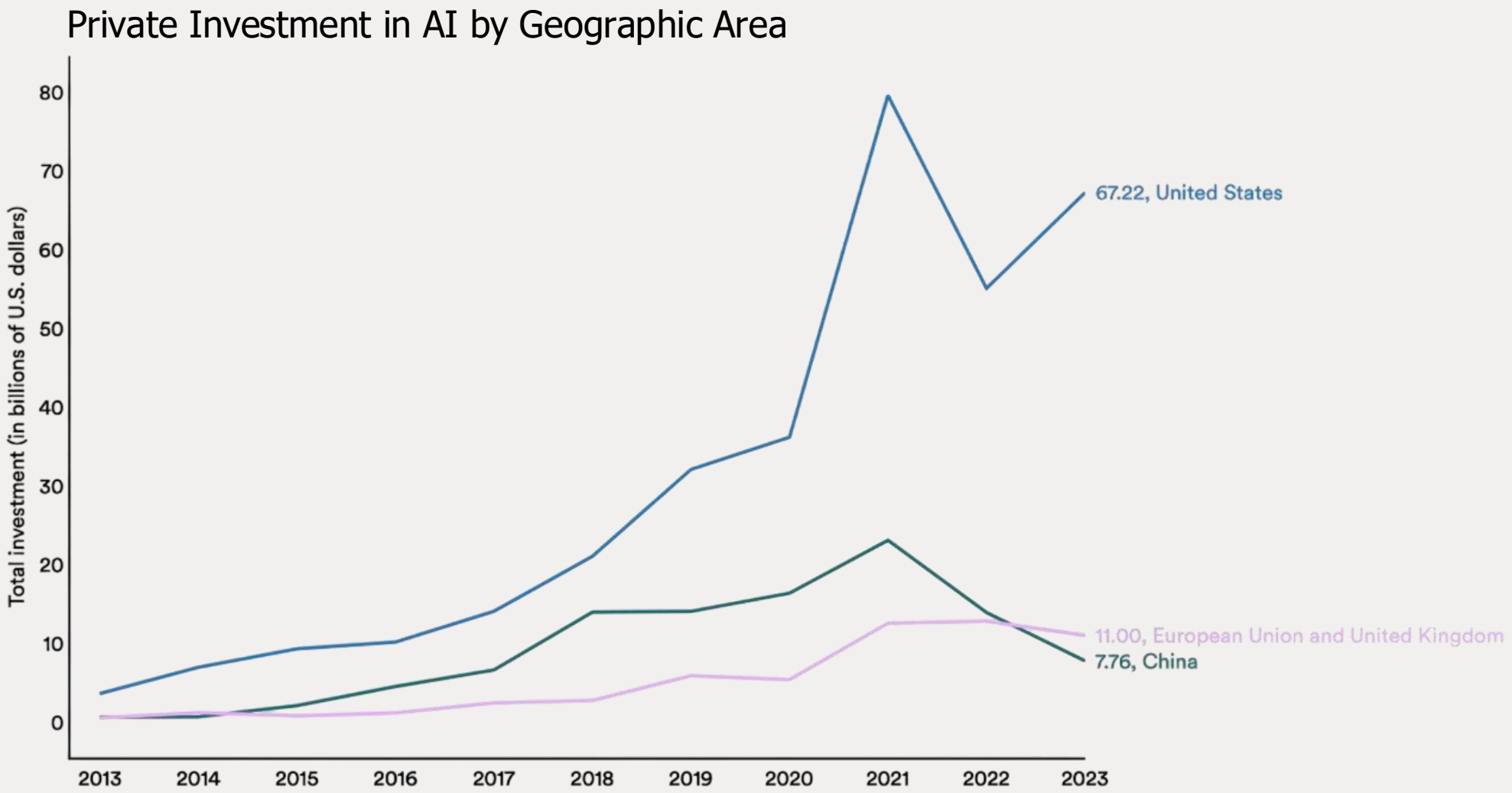
DERIVED HYPOTHESIS:
Linear Correlation between Money and Intelligence*



The AI Spending Spree, in Charts; The Wall Street Journal

* With GPUs acting as the translator

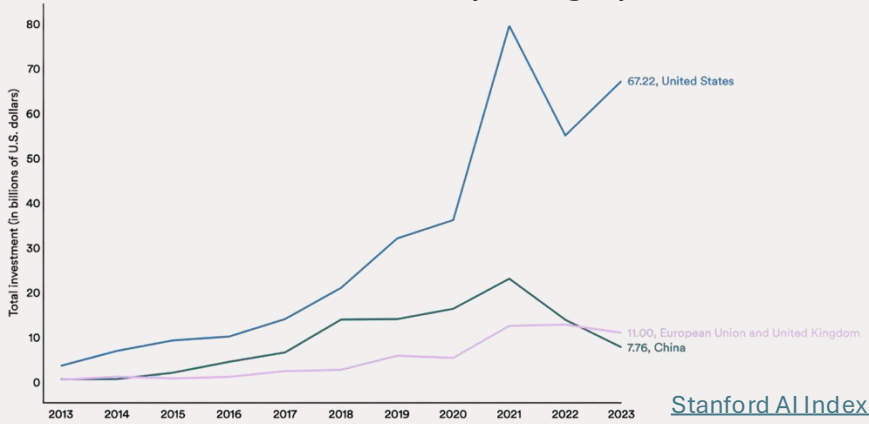
II. Debunking of Perceived Technological Supremacy



Stanford AI Index 2024

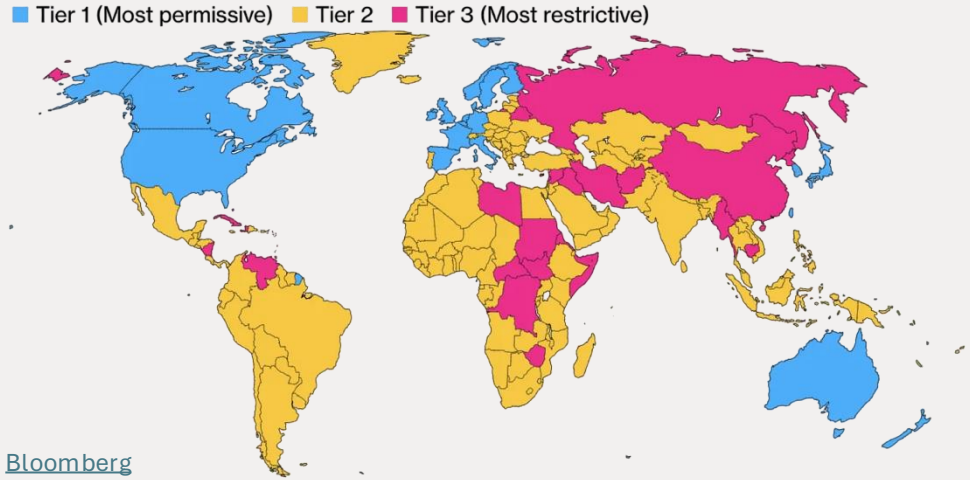
3. The "West" has the *capital*, *the knowledge* and the *GPUs*.

Private Investment in AI by Geographic Area



Stanford AI Index 2024

Level of Chip Shipment Restrictions



Bloomberg

Google

Meta

ANTHROP\C

OpenAI

I. High Stake Field & Arms Race Dynamics

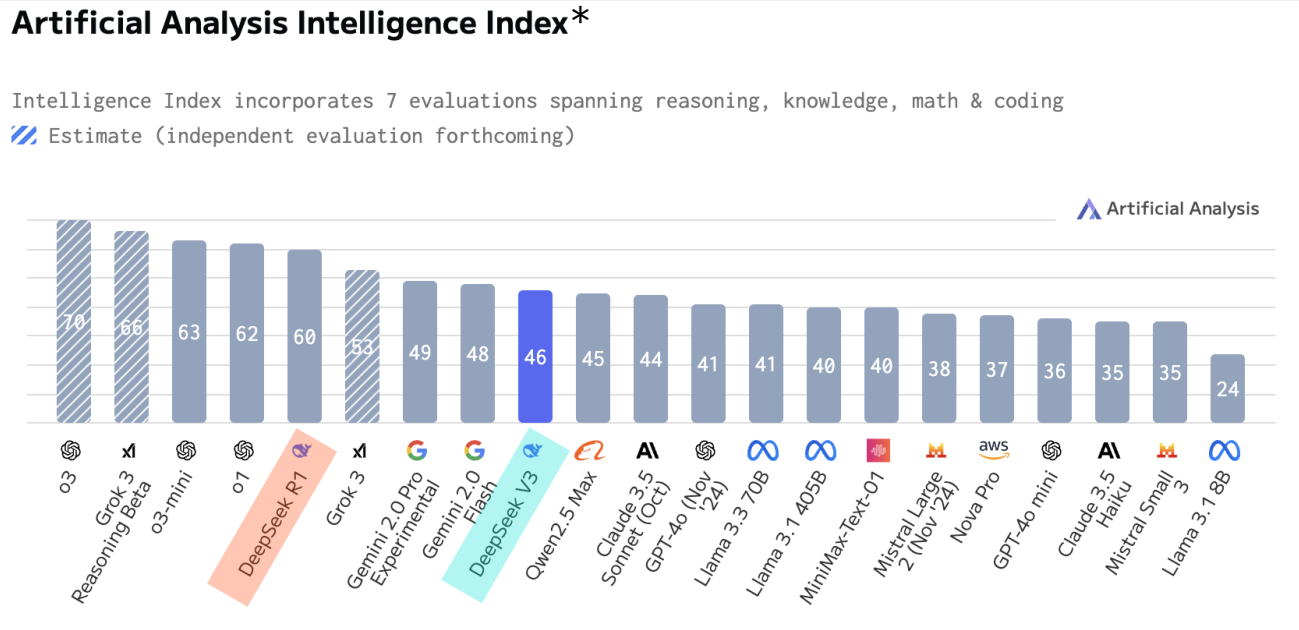
- 1. **Language** is a powerful vehicle and vital component of cross-domain intelligence.
- 2. **More** parameters, data, and compute promise better performance.

II. Perceived Techno-logical Supremacy

- 3. **The "West" has** the *capital*, the *knowledge* and the *GPUs*.

deepseek

1	DeepSeek-V3-Base & DeepSeek-V3	*2024.12.26*
2	DeepSeek-R1-Zero & DeepSeek-R1	*2025.01.20*
3	DeepSeek-R1 Distilled Versions	*2025.01.20*



https://artificialanalysis.ai/models/deepseek-v3

Demonstration *that* and *how* ...

4. Efficiency can be improved

5. RL on LLMs works

6. Reasoning is distillable

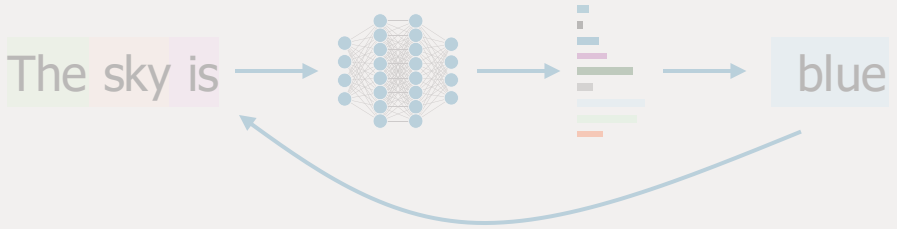
... on a large scale, frontier model

* Combination of multiple established test sets (e.g. MMLU-Pro, Math-500, etc.): https://artificialanalysis.ai/methodology/intelligence-benchmarking

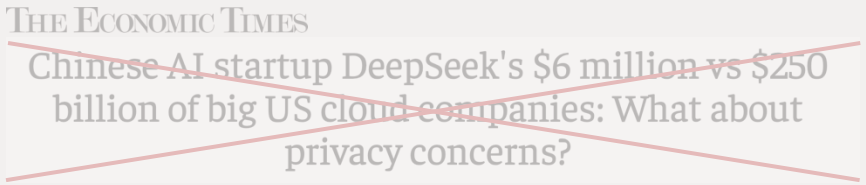
1 DeepSeek-V3-Base & DeepSeek-V3

2024.12.26

Generative Pre-training



- Llama 3.1 - 405B took **30.84M** H100 GPU hours to train[1]
- DeepSeek's V3 final training run took **2.79M** H800 GPU hours (~9% of Llama)[2]



Demonstration *that* and *how* …

4. Efficiency can be improved

How?

- Mixture of Experts (Sparse Model)
- Multi-Token-Prediction
- Dual Pipe Training
- Mixed Precision Training
- Multi-head Latent Attention
- …

1. Llama 3 Model Card by Meta
2. DeepSeek V3 Technical Report

1 DeepSeek-V3-Base & DeepSeek-V3

2024.12.26

- **Demonstration** and **thorough explanation** of efficient training process through excellent engineering efforts
- Clear frontier open weights model with very permissive licence and without major downstream use case restrictions
- No unexpected shift of the performance trajectory of LLMs

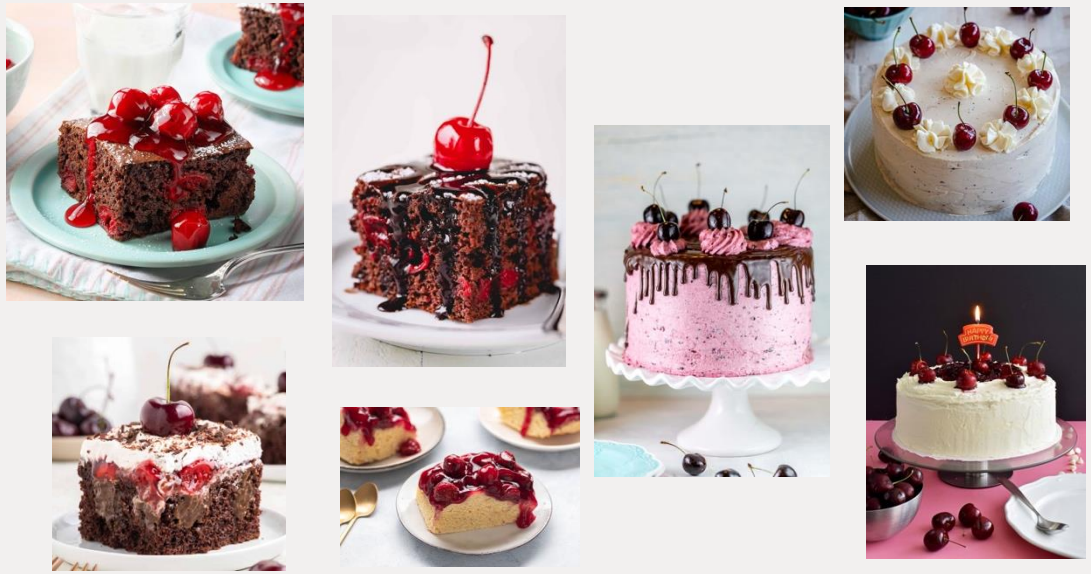
2 DeepSeek-R1-Zero & DeepSeek-R1 *2025.01.20*

Inspired by ISSCC 2019: Deep Learning Hardware: Past, Present, and Future Yann LeCun

Unsupervised Pre-training

Inspired by [ISSCC 2019: Deep Learning Hardware: Past, Present, and Future Yann LeCun]

II. Debunking of Perceived Technological Supremacy | *An Analogy To Yann LeCun's Cake*



Supervised Fine-Tuning

Preference Learning

Inspired by [ISSCC 2019: Deep Learning Hardware: Past, Present, and Future Yann LeCun]

II. Debunking of Perceived Technological Supremacy | *An Analogy To Yann LeCun's Cake*

Inspired by: ISSCC 2019: Deep Learning Hardware: Past, Present, and Future Yann LeCun

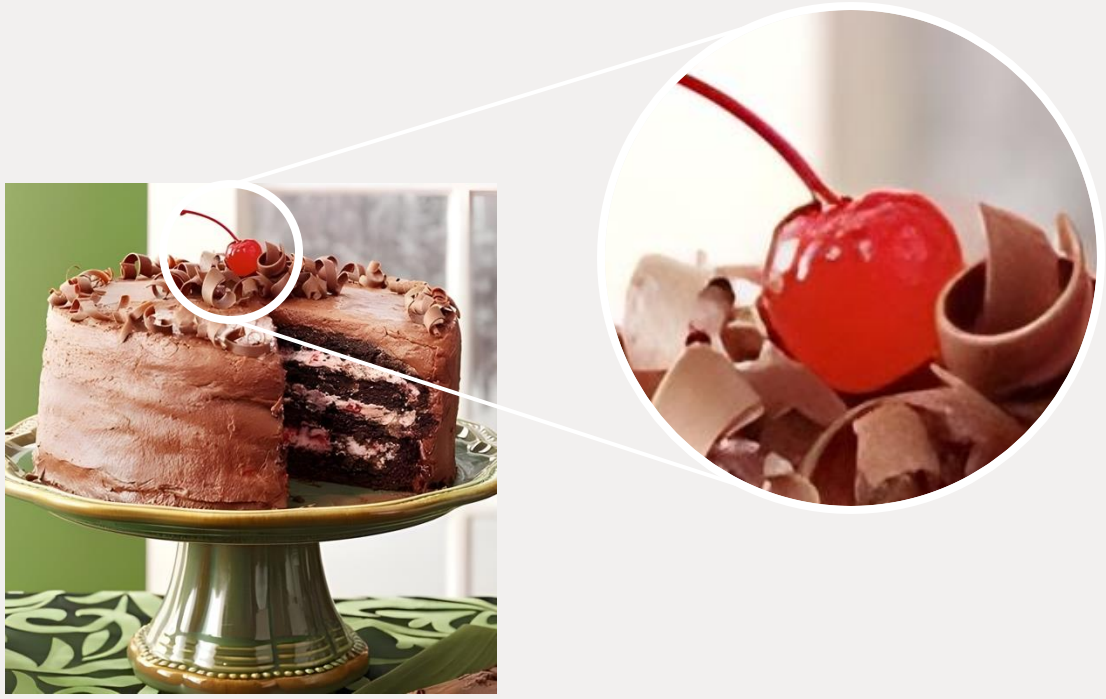
A. Can we trade test time compute* for better performance?

B. How can we overcome the potential bottleneck of **Humans** in RL**HF.**

* test time compute = computational resources required to run inference = inference compute

Reinforcement Learning from Verifiable Rewards

Inspired by ISSCC 2019: Deep Learning Hardware: Past, Present, and Future Yann LeCun



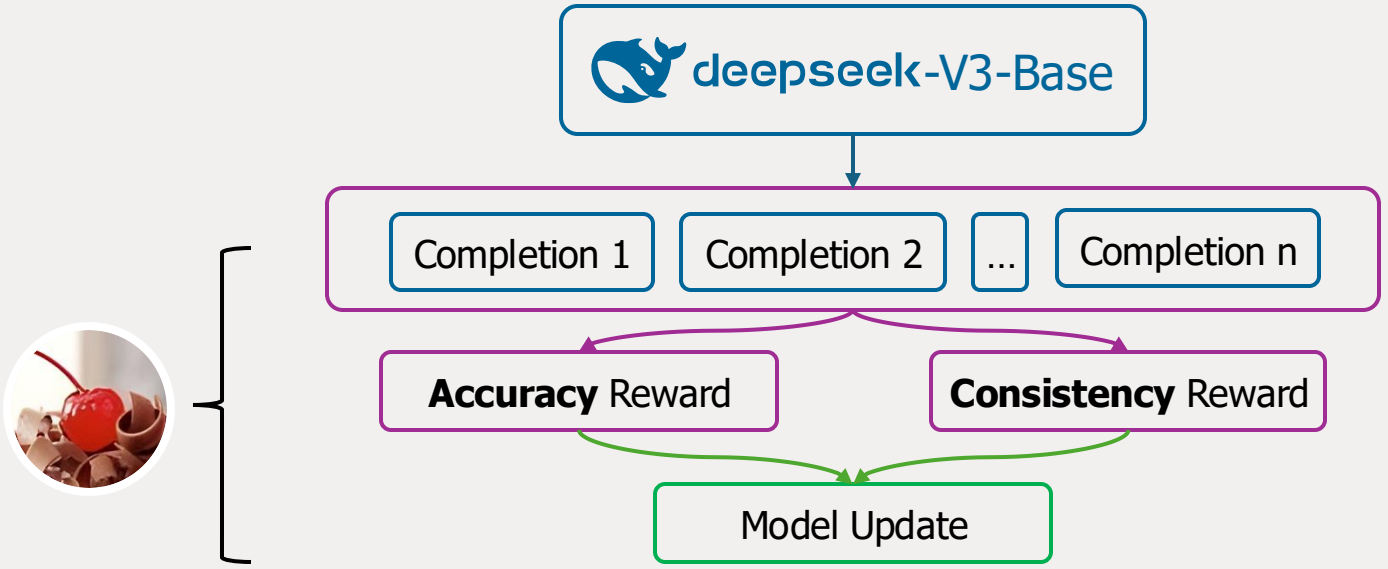
Reinforcement Learning from Verifiable Rewards

[…] Building how we think we think does not work in the long run. […]

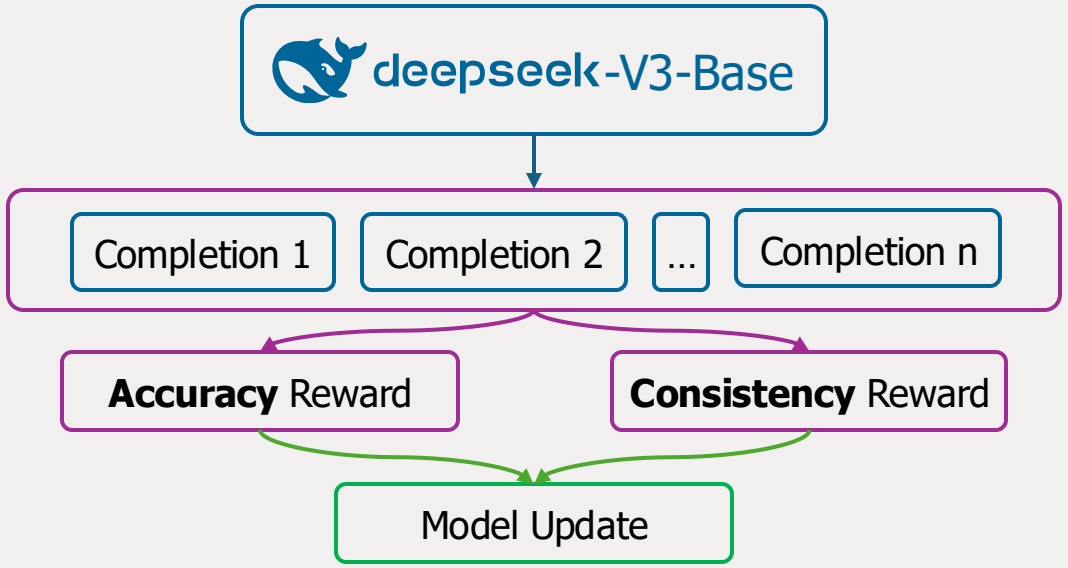
The Bitter Lesson by Rich Sutton

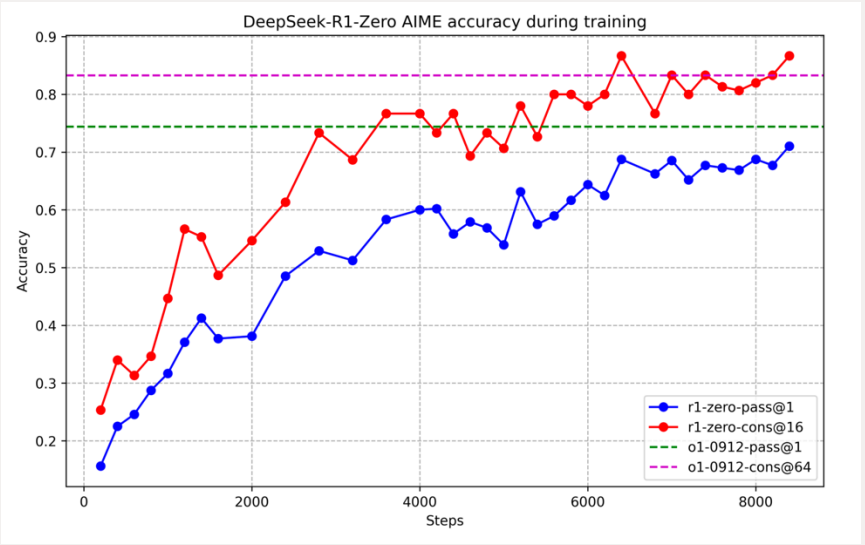
Inspired by ISSCC 2019: Deep Learning Hardware: Past, Present, and Future Yann LeCun

2 DeepSeek-R1-Zero & DeepSeek-R1 *2025.01.20*



2 DeepSeek-R1-Zero & DeepSeek-R1 *2025.01.20*





Demonstration *that* and *how* …

5. RL on LLMs works

⊗ No unexpected shift of the performance trajectory of LLMs

3 DeepSeek-R1 Distilled Versions

2025.01.20

Model	**AIME 2024** *pass@1*	**AIME 2024** *cons@64*
GPT-4o	9.3	13.4
Claude-3.5-Sonnet	16.0	26.7
o1-mini	63.6	80.0
QwQ-32B	**44.0**	**60.0**
R1-Distill-Qwen-1.5B	28.9	52.7
R1-Distill-Qwen-7B	**55.5**	**83.3**
R1-Distill-Qwen-14B	69.7	80.0
R1-Distill-Qwen-32B	72.6	83.3
R1-Distill-LLama-8B	50.4	80.0
R1-Distill-LLama-70B	70.0	86.7

Demonstration *that* and *how* …

6. Reasoning is distillable

https://api-docs.deepseek.com/news/news250120

Democratization of AI?

~800 lines of code

671B Floating-point numbers

Alex Cheema – e/acc

- ✓ Open Weight & Inference Code for all models
- ✓ Comprehensive Technical Report
- ✗ Open Data
- ✗ Open Training / Evaluation Code
- ✗ Open Checkpoints / Logs

Great step in the right direction but far from democratization

FINAL REMARKS

Is deepseek a Sputnik Moment, Democratization of AI, or Media Hype?

⊗ Neither

DeepSeek…

✓ … should be a **wake up** call / reminder

✓ … **reduced** the **entry barrier**

? III. Strong Response from the West

deepseek