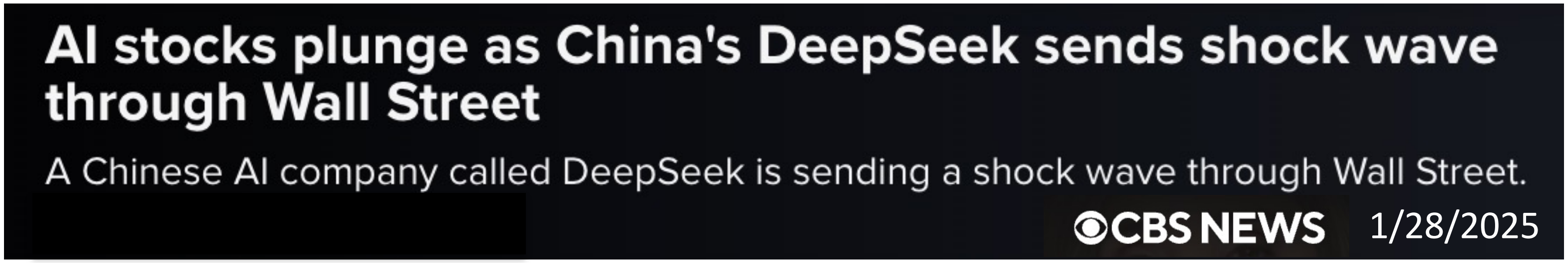
Deep Revolution

- **2024 Nobel prizes in physics (deep learning) & chemistry (Google DeepMind) shook the scientific world, heralding the new era of AI-enabled science**

  https://www.nobelprize.org/prizes/physics/2024

  https://www.nobelprize.org/prizes/chemistry/2024

- **In January 2025, DeepSeek sent a shock wave to Wall Street, White House, & Silicon Valley**

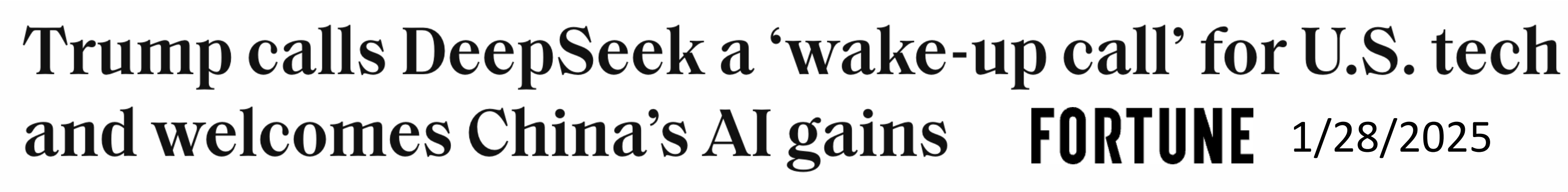


AI stocks plunge as China's DeepSeek sends shock wave through Wall Street

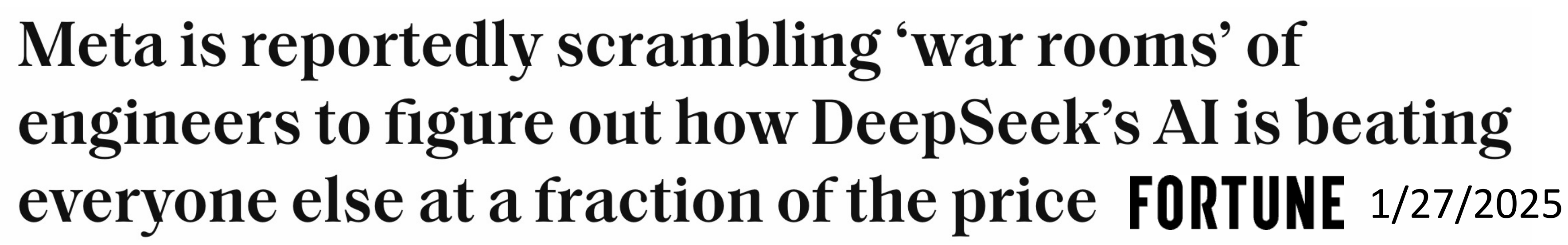
A Chinese AI company called DeepSeek is sending a shock wave through Wall Street.

[redacted]

CBS NEWS 1/28/2025



Trump calls DeepSeek a 'wake-up call' for U.S. tech and welcomes China's AI gains FORTUNE 1/28/2025



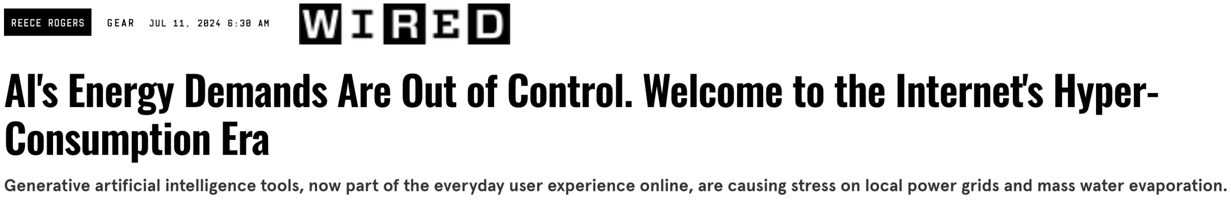
Meta is reportedly scrambling 'war rooms' of engineers to figure out how DeepSeek's AI is beating everyone else at a fraction of the price FORTUNE 1/27/2025

Key Computational Enablers of DeepSeek?

- **DeepSeek is a large language mode (LLM) that outperforms OpenAI's ChatGPT with less computing**
- **Multi-head Latent Attention guarantees efficient inference through significantly compressing the Key-Value cache into a latent vector, while DeepSeekMoE (Mixture-of-Experts) enables training strong models at an economical cost through sparse computation** [https://arxiv.org/abs/2405.04434]
- **DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance** [https://arxiv.org/html/2412.19437v1]
- **Reasoning: DeepSeek-R1 directly applies reinforcement learning to the base model, thereby generating a long chain-of-thoughts** [https://arxiv.org/abs/2501.12948]

My expert friend thinks it's their ingenious engineering, not these known & some new methods

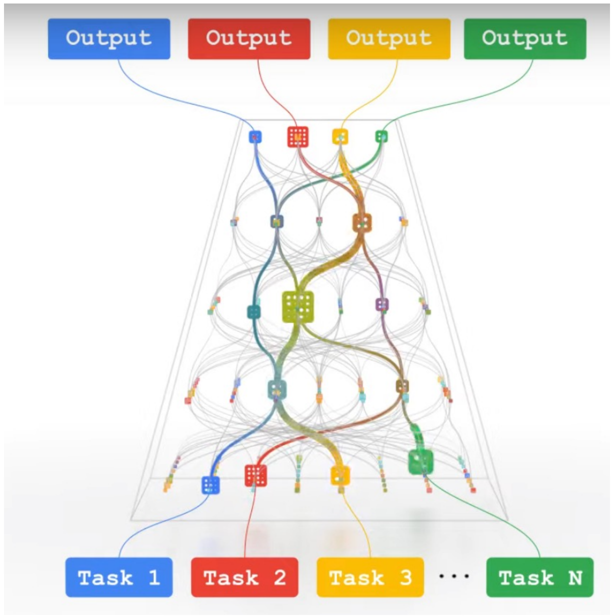
- **Will brain-like sparse spiking of neurons solve the AI power catastrophe (*cf*. Google's Pathways)?**



REECE ROGERS GEAR JUL 11, 2024 6:30 AM WIRED

AI's Energy Demands Are Out of Control. Welcome to the Internet's Hyper-Consumption Era

Generative artificial intelligence tools, now part of the everyday user experience online, are causing stress on local power grids and mass water evaporation.



Final project?

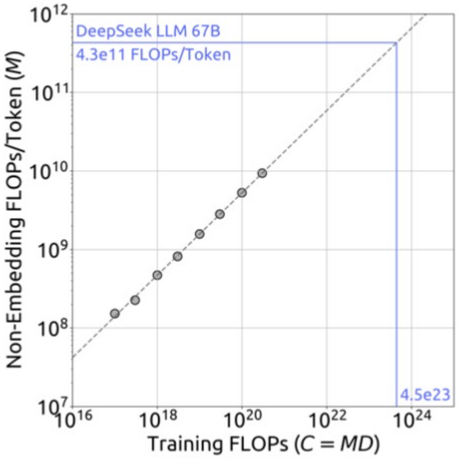
https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture

Scaling Analysis Is Important

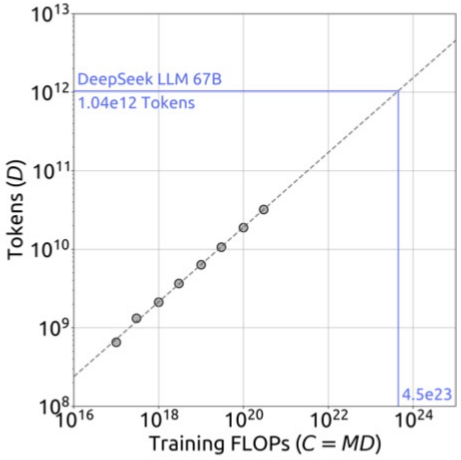
- **Understanding scaling laws of LLMs is essential for long-term projection**

https://arxiv.org/abs/2401.02954

- **Use the same scaling exponent analysis (log-log plot & linear fit) as in assignment 2, Part I-2!**



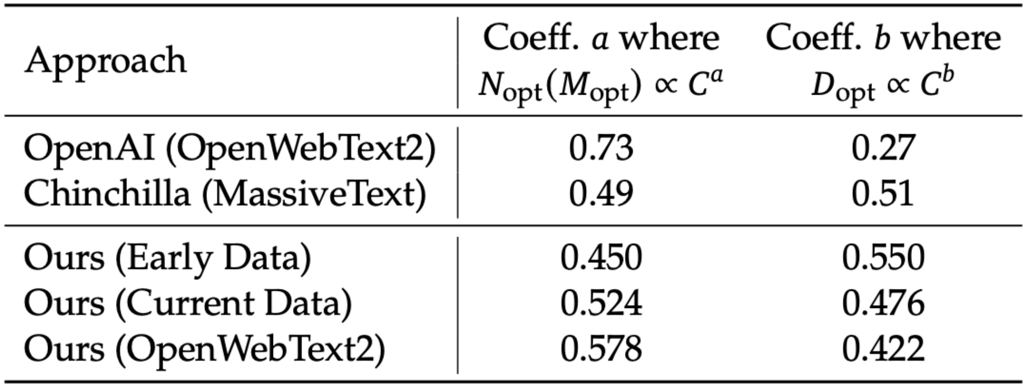
(b) Optimal model scaling



(c) Optimal data scaling

$$M_{\text{opt}} = M_{\text{base}} \cdot C^{a}, \quad M_{\text{base}} = 0.1715, \quad a = 0.5243$$

$$D_{\text{opt}} = D_{\text{base}} \cdot C^{b}, \quad D_{\text{base}} = 5.8316, \quad b = 0.4757$$



Approach	Coeff. a where $N_{\text{opt}}(M_{\text{opt}}) \propto C^{a}$	Coeff. b where $D_{\text{opt}} \propto C^{b}$
OpenAI (OpenWebText2)	0.73	0.27
Chinchilla (MassiveText)	0.49	0.51
Ours (Early Data)	0.450	0.550
Ours (Current Data)	0.524	0.476
Ours (OpenWebText2)	0.578	0.422