Settling the Frontier

Shifting priorities in Artificial Intelligence

Noah Dockery

November 20, 2026

Contents

Introduction

I. The Frontier Will Get Harder to Scale

II. The Best Model for the Job Will Often Be Local

III. Open Models Improving Themselves

IV. Mispricing the Frontier

V. Compute Will Trade Like a Commodity

VI. AI for the American People

Sources

Introduction

A fundamental shift in what we as a society are pushing artificial intelligence towards is occurring.

As talks arise of 'Pacing the Frontier', as frontier labs/data centers are becoming incredibly unpopular to the American voter, and as politicians debate the fate of American AI, I believe that we will shift priorities from developing larger LLMs towards applying AI more efficiently in everyday cases.

In other words, I think AI is going local. The best general models may still come from a few labs. But the more we rely on AI for sensitive industries pivotal to the future of the national image like space and robotics, the less sense it makes to work every important job through a single private corporation, a corporation that can raise prices, change terms, or restrict access at will. Any company -or government agency- hung up on a model that is subject to pressure from outside investors and regulation will soon realize the value of local AI.

I. The Frontier Will Get Harder to Scale

The biggest labs are making many bets at once. They have to keep finding enough power, hardware and data to make better models. They have to find customers willing to pay enough to finance the next round. They have to subsidize their services with twenty- and hundred-dollar plans so the everyday user finds them appealing. And on top of that, they're praying (and paying) for good fortune in regulation and government. It is easy to see how one bet works in isolation, but I have a harder time seeing every bet working at today's scale in the future.

Start with power. Global data centers used 415 TWh of electricity in 2024 and 485 TWh in 2025, and the IEA projects 950 TWh by 2030. Going from 485 to 950 TWh a year would mean about 53 GW of additional average load worldwide. The chart covers all data-center workloads. What matters for a new project is how much electricity the local grid can deliver: a grid connection, transformers and cooling at one site, on a deadline.

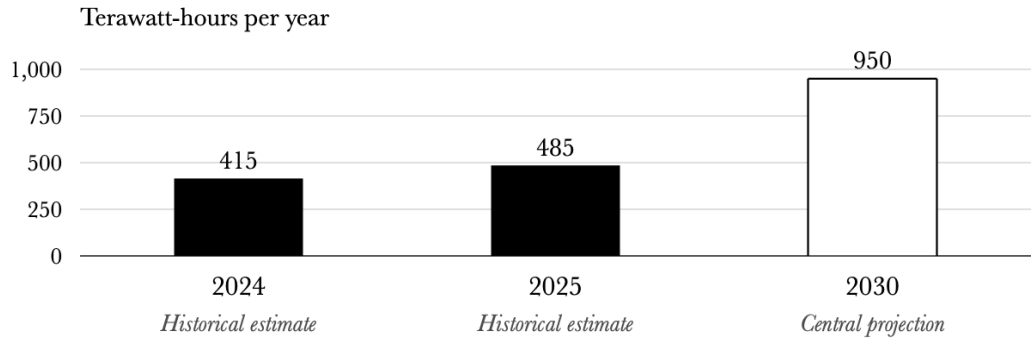

Figure 1. Global data-center electricity use · *International Energy Agency [2]*

Labs are also running low on text. One 2024 estimate has training using up the entire collection of human-written text between 2026 and 2032. Labs have resorted to collecting millions of books, scanning them, and destroying the originals. I find it unsettling, even dystopian, that a private company is the one turning shared literature into a training advantage. And even when there is more text, it helps less each time. Prediction error falls as a power law as training data grows, so each extra batch buys less than the last. Repeating the same text works for a few passes and then stops helping.

After the November midterms, Democrats preparing to take control of Congress said they would push for independent safety testing of advanced AI models and clearer rules for when companies can be held liable for harm. [7] These proposals make it clear that Washington is headed towards regulation. Regulation and liability will only take away from the frontier's rapidly shrinking moat and their ability to subsidize the cost of their models, their ability to provide 20 or 100 dollar plans.

Uber used discounts and driver incentives to get people into the habit of using its app. [8, 9] Frontier labs are making a similar bet. They hope customers will have built enough of their work around the service that they stay when prices go up. Labs that can't bring costs down may have to raise prices, cut what their plans include, or go out of business. The bet only works if switching costs more than staying. The better local models get, the harder that bet is to win.

II. The Best Model for the Job Will Often Be Local

By local I mean the model runs somewhere the operator controls and can keep working without calling the provider. That could be a personal laptop or a university server room. Open weights can make independent deployment possible under the license. Fully open-source AI, as the Open Source Initiative defines it, sets a higher bar for what people can inspect and modify.

The relevant gap is how long you have to wait for a model you can run yourself to do the job you need. The best open-weight models trail the closed frontier by about four months, and models that run on a single high-end consumer GPU trail by around six. Most of the best open models today are Chinese, though, and that matters for government and defense. The best American open models are further behind, which is why I argue that we should build them. Plenty of hard work will still justify paying for the best model available. Plenty of regular work is already within reach of a model you can own.

- **Government.** A town can keep its records where it has authority over access, use them to produce an answer with sources, and keep an approved model running through a vendor outage. A regional service can make that possible for towns too small to staff their own systems.
- **Defense.** Maintenance records, inventories, and technical manuals are work a unit should be able to keep doing if its provider has an outage or a breach. Running an approved model on infrastructure the department controls gives it another way to keep working.
- **Robotics.** A warehouse robot has to process what it sees while the network is congested. You can train in a big cluster and still want the decision made on the machine. What counts is what the robot can do in the building, on time, under load. [15]

Defense, government, robotics and space are the industries that carry America's reputation. We want to lead in them, and I think we should. Leading means being able to keep working and keep innovating, whatever a vendor decides about its price, its terms or its uptime.

The same logic holds for the cost. Suppose it costs $15,000 a year to operate a local system, plus two cents per accepted job. A hosted equivalent costs twelve cents per job. If they do the work equally well, local breaks even at **150,000 jobs a year**. Cut the hosted price in half and the break-even goes to **375,000**. Those are hypothetical prices. They show why you have to count the actual workload,

including retries and human review. An extra minute for a person can cost more than the inference.

A smaller model fine-tuned on good examples can beat a general model on narrow, well-defined tasks. On general-domain relation extraction, a fine-tuned 500-million-parameter Qwen scored 0.83 F1, against 0.69 for GPT-5.4 and 0.66 for Claude Sonnet 4.6 when those models were prompted without examples. The smaller model had task-specific training.

If I were choosing a model for a town, I'd give both systems the same real tickets and count correct answers with current sources, time spent fixing mistakes, and jobs that still work through an outage. If the local model does the work as well or better on those terms, I would run it locally.

III. Open Models Improving Themselves

I think an open-weight model will repeatedly improve its own ability to do AI research and engineering before 2030. Open models are about four months behind the closed frontier, and I expect that gap to keep shrinking. As open models get smarter, I expect them to be able to do the work of improving models, including themselves.

The early signs are real. This month, MIT researchers used OpenAI's GPT-6 Astra to rewrite the weights of a small open model, Qwen, directly, with no training run. Weights are the numbers inside a model that decide how it behaves. Its score on a math test went from 7.9% to 48.1% after Astra changed about 0.016% of them, and the improvement came mainly from fixing repetitive reasoning. That makes it a repair, not a leap, and it was a closed frontier model improving an open one, not an open model improving itself. But a frontier model improving a model anyone can download is the kind of thing I expect open models to do for each other soon. DeepMind's AlphaEvolve used Gemini and automated testing to improve algorithms, including one used in data-center scheduling. The authors of the Darwin Gödel Machine reported that a system rewriting its own coding agent went from 20% to 50% on a software-engineering test.

The big implication is that whatever governments decide about the leading labs, more of the ability to improve AI could sit outside them. That is good for competition and independent research. It also means we have to pay attention to what these systems can actually do when they spread.

IV. Mispricing the Frontier

I worry that there may be serious repricing by the end of 2028. I expect people to use more AI through it, but some labs, hosts and infrastructure projects tied to the most aggressive growth assumptions will have to change their plans. Anthropic just went public, and it's spending far more than it earns. In 2025 it brought in nearly $4.6 billion and lost more than $8 billion on operations. In the first half of 2026 it reported $16.3 billion in revenue, and a net loss of $80.7 billion, though most of that is accounting for the rising value of its investors' notes. Walmart made $22.3 billion in net income last fiscal year. Eventually those numbers have to line up.

What also worries me is how many of the companies that make up such a big part of the economy are in circular deals. Nvidia agreed to invest in OpenAI. OpenAI agreed to pay Oracle $300 billion for computing power, and Oracle agreed to buy $40 billion of chips from Nvidia. If I give you $5 and you give my friend $5 and my friend gives me that $5 back, none of us made $5. It's the same money going around. The real deals aren't that simple, but when so much of a company's revenue comes from a customer it has funded, growth can look stronger than the demand behind it.

I think frontier access is going to get much more expensive, and I think that happens after the bubble pops, not before. Once labs have to earn what they spend, the discounts that bring in new users go away. I can't say what will set it off. I'd count the bubble call as a hit if a sustained repricing comes with real project deferrals, major contract renegotiations or provider restructurings. If the major labs fill the capacity they signed for, generate durable cash and keep frontier access affordable, I got this part wrong.

I don't think Anthropic goes bankrupt. My point is about anyone who puts all of their trust in a company that spends this far ahead of what it earns. Prices and terms change when the money tightens. Do you really want your data and your budget depending on one company's finances? Open models look much more attractive next to that, because you can run them yourself and nobody can change the terms on you.

V. Compute Will Trade Like a Commodity

When you pay for frontier access, part of what you're buying is time on computers.

Cloud providers already rent GPUs by the hour. Say a university reserves eight H100s for a week and finishes its experiment two days early. Could another lab buy those last two days? It would need to know where the GPUs are, whether they're connected to each other fast enough for its work, and whether the contract lets the university transfer the reservation. If buyers can compare offers on those terms, unused capacity can change hands.

There's already a resale market for the hardware. From late September to mid-November 2026, 94 tracked U.S. eBay sales of used RTX 3090s had a median price of $1,650. [24] At a different scale, CoreWeave announced an $8.5 billion investment-grade financing facility in March 2026. The customer contract gives lenders a source of repayment, and the equipment is part of what they can sell if the borrower defaults. How much they recover depends on what someone will pay for those servers at the time. [25]

CME has proposed futures tied to H100 and B200 rental prices. [26] The contracts would let buyers hedge a price increase and would settle in cash. A buyer would still have to reserve the machines separately. Regulators extended their review in November, delaying the proposed launch. [27]

By the end of the decade, I expect organizations to be able to buy and resell at least one standardized type of GPU reservation and hedge its rental price. If providers still set the terms one contract at a time and buyers can rarely resell unused hours, I got it wrong. A town deciding what to do when its contract comes up for renewal could compare offers for the extra capacity it needs. If it owns enough hardware to handle routine work, it can keep that work running while it shops around.

VI. AI for the American People

Most people rent their AI. They type into someone else's servers, on someone else's terms, at someone else's price, and the lab can change any of them tomorrow or use this data to train new models.

Bitcoin is the closest comparison I know. Set aside whether it is a good investment. The idea it proved is custody: you can hold your own money without a bank's permission. Open-weight models do the same for intelligence. Download one, run it on your own hardware, and no lab can raise its price, rewrite its terms or switch it off.

The usual story of AI is about pushing the frontier outward: bigger models, harder benchmarks. But exploring land and settling it are different jobs. Value arrives when the territory is put to work, in a hospital, a factory, a city agency. I call that settling the frontier, and it is how a country earns its reputation. Any American who wants us to lead should want the frontier settled here, by our own towns and factories, not leased from a few firms.

That is what American AI should be, widespread AI, intelligence for the people: who gets to settle the frontier, and whether they need permission. Right now a few unelected companies and investors decide who gets the most capable AI, at what price and under what rules. This techno-oligarchy, a term for wealthy, influential, and most importantly unelected people and entities who control key parts of technological development, should not be able to determine the future of American intelligence. You don't have to accept the label to see the structure. America's tradition is to spread power around so that no one can hold everyone else hostage.

The next fight is over open models. After China's Moonshot released its open-weight Kimi K3 in July, Axios reported that parts of the administration had considered ways to make it harder for American companies to use Chinese open models. [37] More than two dozen companies, including Nvidia, Microsoft and Meta, signed a letter asking Washington not to impose premature restrictions on open weights. [38]

The danger is losing this AI race, the open race. Chinese labs have led in open-weight models since about April 2025, [42] and Chinese models passed American ones in downloads by August 2025. [41] On the Artificial Analysis index, the top three Chinese open models score 42 to 45, while the best American ones score 23 and 26, and American companies including Harvey, Cursor, DoorDash, Airbnb and Perplexity are already building on Chinese open weights. [42] Banning Chinese models without an American alternative leaves American users with worse tools. Restricting American open development would finish the job. The answer is to build better American open models. Settling the frontier with our own towns and factories requires American open models to run it.

I believe that rules should follow capability and use, not openness. Test the most dangerous systems, open or closed. Hold people responsible for what they deploy. But a rule that lets only the largest firms run capable models hands the few a permanent

monopoly and calls it safety. If we restrict who may own and run this technology, elected officials should decide it in public. Anything else is undemocratic and un-American.

Open models are about four months behind the frontier, and the gap is closing. So ask your representatives where they stand on open weights. Back the people who build them. Run one yourself.

No matter how this new great frontier will be built, settling it should be our priority.

Sources

[1] AP–NORC Center for Public Affairs Research, Most Adults Think AI Is Developing Too Fast. October 8, 2026.

[2] International Energy Agency, Key Questions on Energy and AI. 2026. Executive summary and Figure 1.4.

[3] Pablo Villalobos and colleagues, Position: Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data. ICML, 2024.

[4] U.S. District Court for the Northern District of California, Bartz v. Anthropic PBC, Order on Fair Use. June 23, 2025.

[5] Jared Kaplan and colleagues, Scaling Laws for Neural Language Models. 2020.

[6] Niklas Muennighoff and colleagues, Scaling Data-Constrained Language Models. Journal of Machine Learning Research, 2025.

[7] Reuters, Democrats Planning Flood of AI Bills When They Take Control of Congress. November 6, 2026. Reuters wire syndicated by The Spokesman-Review.

[8] Uber Technologies, Form S-1/A. 2019. See consumer discounts and driver incentives.

[9] Uber Technologies, Form 10-K. Fiscal year ended December 31, 2025. See incentives and promotions.

[10] TechCrunch, OpenAI Is Losing Money on Its Pricey ChatGPT Pro Plan, CEO Sam Altman Says. January 5, 2025.

[11] Open Source Initiative, The Open Source AI Definition 1.0. 2024.

[12] Jack Edwards and Luke Emberson, Epoch AI, Open Models Lag State-of-the-Art Closed Models by 4 Months. May 29, 2026. Based on the Epoch Capabilities Index.

[13] Epoch AI, Frontier AI Capabilities Can Be Run at Home Within a Year or Less. 2025. Hardware-constrained benchmark analysis, separate from [12].

[14] Cyber Safety Review Board, Review of the Summer 2023 Microsoft Exchange Online Intrusion. 2024. Executive summary.

[15] NVIDIA, Getting Started with Edge AI on NVIDIA Jetson. 2025.

[16] Predibase, LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4. Technical report, 2024.

[17] Cheng-Yu Hsieh and colleagues, Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. 2023.

[18] Despina Christou and Grigorios Tsoumakas, Sub-Billion, Super-Frontier: Small Language Models Rival Zero-Shot Frontier LLMs on General and Literary Relation Extraction. Preprint, June 2026.

[19] Jenny Zhang and colleagues, Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. 2025.

[20] Google DeepMind, AlphaEvolve: A Gemini-Powered Coding Agent for Designing Advanced Algorithms. 2025.

[21] METR, AI R&D-Relevant Model Capabilities Are Improving Rapidly and Are Getting Harder to Measure. November 20, 2026.

[22] Anthropic, Announcement of IPO Closing. November 20, 2026.

[23] Anthropic, Form S-1/A. November 17, 2026. See commitment disclosures and cash requirements.

[24] SoldComps, NVIDIA RTX 3090 Sold Price on eBay. November 15, 2026 snapshot of completed U.S. sales.

[25] CoreWeave, Announcement of $8.5 Billion GPU-Backed Financing, March 31, 2026. Transaction overview, p. 4 (collateral).

[26] CME Group, Compute Futures. Proposed H100 and B200 rental-index contract specifications.

[27] Financial Times, US Regulator Postpones Wall Street Push to Trade AI Computing Futures. November 10, 2026.

[28] Satoshi Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System. 2008.

[29] Pew Research Center, Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. June 17, 2026. See the 2024 and 2026 figures on confidence in U.S. companies.

[30] U.S. National Telecommunications and Information Administration, Dual-Use Foundation Models with Widely Available Model Weights. 2024.

[31] U.S. National Science Foundation, NSF Establishes Operations Center for National Artificial Intelligence Research Resource. September 1, 2026.

[32] Rikiya Takehi, Jayden Teoh, Alex Zhang and Omar Khattab, MIT CSAIL, LLMs Can Natively Read & Write Neural Network Weights. November 2, 2026.

[33] Wall Street Journal, via SiliconANGLE, OpenAI and Oracle Strike $300B Cloud Computing Deal to Power AI. September 10, 2025. https://siliconangle.com/2025/09/10/openai-oracle-strike-300b-cloud-computing-deal-power-ai/

[34] Reuters, reporting the Financial Times, Oracle to Buy $40 Billion of Nvidia Chips for OpenAI's US Data Center, FT Reports. May 23, 2025. https://www.aol.com/news/oracle-buy-40-billion-nvidia-181443605.html

[35] NVIDIA, OpenAI and NVIDIA Announce Strategic Partnership to Deploy 10 Gigawatts of NVIDIA Systems. September 22, 2025. https://nvidianews.nvidia.com/news/openai-and-nvidia-announce-strategic-partnership-to-deploy-10gw-of-nvidia-systems

[36] Retail Insight Network, Walmart Revenue Rises by 4.7% in Fiscal 2026. https://www.retail-insight-network.com/news/walmart-fy26-revenue-rises-4-7/

[37] Axios, The Secret Trump Administration Battle to Fight Chinese AI. July 20, 2026. https://www.axios.com/2026/07/20/ai-us-china-open-source-kimi

[38] Open Weights and American AI Leadership, industry coalition letter. July 24, 2026.

[39] Dario Amodei, Anthropic, Our Position on Open-Weights Models. July 27, 2026. https://www.anthropic.com/news/position-open-weights-models

[40] RAND Corporation, Open-Weight AI Models May Increase Biological Misuse Risks: Assessing Anti-Refusal Tampering, Capability Enhancement, and Publicly Available Uncensored Models. https://www.rand.org/pubs/research_reports/RRA5112-1.html

[41] Nathan Lambert and Florian Brand, The ATOM Report: Measuring the Open Language Model Ecosystem. April 2026. https://arxiv.org/abs/2604.07190

[42] Nathan Lambert, The Current Balance of Power in Open. Interconnects. https://www.interconnects.ai/p/the-current-balance-of-power-in-open

[43] Fortune, Anthropic, DeepMind, OpenAI and Meta Employees Back Washington AI Slowdown Plan (Pacing the Frontier statement). July 29, 2026. https://fortune.com/2026/07/29/anthropic-deepmind-openai-meta-washington-ai-slowdown-plan

[44] Michael Scherer and Ashley Parker, The Atlantic, The Tech Oligarchy Arrives. January 20, 2025. https://www.theatlantic.com/politics/archive/2025/01/tech-zuckerberg-trump-inauguration-oligarchy/681381/