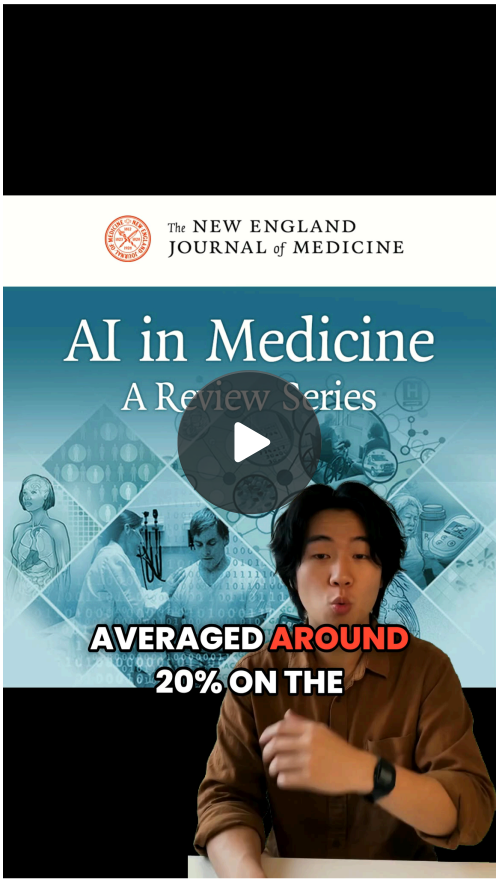
Token Economics Versus the 20 Watt Brain: Why Inference Costs, Not Intelligence, Will Cap Clinical AI, & Whose Job It Is to Decide If Machines Should Diagnose Us When They Reas Better but Cost More

JUN 02, 2026 • PAID

Video Preview



🎧 Podcast episode for paid subscriber only. Also available on Spotify.

Thoughts on Healthcare Markets & Technology

Token Economics Versus the 20 Watt Brain: Why Inference Costs, Not Intelligence, Will Cap Clinical AI, & Whose Job It Is to Decide If Machines Should Diagnose Us When They Reason Better but Cost More

AI models hit 80-85% accuracy on brutal NEJM diagnostic cases. Unaided generalist physicians averaged around 20% on the same cases. So why isn't smarter AI running on every patient? The answer is the economics of thinking...

Listen now

37 minutes ago · Thoughts on Healthcare

To listen to paid episodes in Apple or Spotify, link your Substack subscription v show settings on those platforms (instructions inside the Substack app under Subscriptions → Podcast).

Table of Contents

4. The benchmark that should worry doctors and terrify the people who model margin
5. The agency that signs off on safety never sees the electricity bill
6. Can anyone actually force the smarter, pricier brain into the exam room
7. How this probably ends, and where the money is hiding while it does

Abstract

The bottleneck for clinical AI is no longer capability. It is the marginal cost of e smart answer, and the fact that the smarter the answer, the more it costs to prod

Quick orientation for the skeptics:

1. Training is a one-time capex bonfire. Inference is the opex line that scales w every patient, every note, every prior auth, basically forever.
2. A median frontier query runs about 0.34 Wh, but a long reasoning query wi roughly fifteen times the tokens runs about 4.32 Wh, a thirteen-fold jump fo exact behavior that makes models good at hard diagnosis.
3. The human brain does the same hard reasoning on about 20 watts, which is existence proof that silicon has three or four orders of magnitude of headro has not collected yet.
4. Microsoft's sequential diagnosis work hit roughly 80 to 85.5 percent accurac brutal NEJM cases against about 20 percent for unaided generalist physicia and it did so by spending more compute to spend less on tests.
5. No US regulator owns the question that actually matters, which is whether a better answer is worth the tokens. FDA does safety. CMS does coverage and payment. Nobody does cost-effectiveness with teeth.

Everyone fell in love with training and forgot who pays for thinking

The whole scaling-laws romance was about training. Bigger model, more data, m flops, smarter machine, and the charts went up and to the right in a way that ma venture partners weepy. Fine. But training a frontier model is a capital event. Yo light the money on fire once, you get a set of weights, and then the weights sit th The thing that actually costs money every single day, forever, is inference, which unglamorous act of the model answering a question. Training is the wedding. Inference is the marriage, and it is the marriage that bankrupts people.

Here is the structural problem nobody priced into the 2023 hype. Inference cost with usage, and in a transformer the cost per answer is not flat. Attention is qua in the length of the context, so a long chart or a long reasoning trace does not co little more, it costs a lot more. There is a prefill stage that is compute-bound and decode stage that grinds out one token at a time, and the output tokens dominat energy bill. So the moment you ask a model to actually think, meaning chain its reasoning out across thousands of tokens, the cost curve bends against you in a v never did when the model just autocompleted a sentence.

The numbers are no longer hand-wavy. Sam Altman put a stake in the ground in 2025 with a figure of about 0.34 watt-hours for an average ChatGPT query, roug what a high-efficiency lightbulb sips in a couple of minutes, and an Epoch AI an landed in the same neighborhood even under pessimistic assumptions. So far so cheap. But a 2025 inference-energy paper out of the arXiv crowd put the median frontier-scale query at that same 0.34 Wh and then showed what happens with te time scaling, the reasoning regime where the model generates roughly fifteen tin more tokens. Energy per query rises about thirteen-fold to 4.32 Wh. Serving a bi plain queries a day runs around 0.8 GWh. Shift just ten percent of them to long reasoning and you are at 1.8 GWh a day. The interface is drifting toward tasks th the model to think, plan, search, verify, and act, which is to say toward the exper regime, on purpose, because that is the regime that produces answers worth hav

Zoom out to the grid and the abstraction stops being cute. The IEA's 2025 Energ AI report pegged data centers at about 415 TWh in 2024, roughly 1.5 percent of electricity, growing 12 percent a year, with updated projections of close to 945 te TWh by 2030, somewhere around three percent of world demand and roughly th

entire current consumption of Japan. The US slice was about 183 TWh in 2024, already north of four percent of national electricity, projected to grow about 133 percent to 426 TWh by 2030. The AI-specific portion quadruples or triples depe on the scenario. Yes, efficiency per task is improving at a rate the IEA called unprecedented in energy history. That is exactly the trap. Jevons paradox does n about your good intentions. Make each answer cheaper and people will demand more answers, and the ones they demand will increasingly be the expensive thinl kind. The downfall thesis is not that the models stop getting smart. It is that the marginal cost of the marginal smart answer refuses to trend to zero on the timel the marketing implied, and someone, somewhere, has to keep paying that bill on query for the rest of time.

A three pound lump of fat that out-reasons a server farm on twenty watts

← Previous