280x Cheaper: The Real AI Revolution Is Accessibility

Published May 19, 2025

Christopher Gannatti, CFA
Global Head of Research

Key Takeaways

- The cost of generating GPT-3.5-level AI outputs dropped 280-fold in less than two years, transforming advanced AI from a luxury tool into a mass-market utility.
- As inference costs plummet, a deflationary wave is reshaping the AI ecosystem, favoring widespread deployment over elite performance and enabling AI-native applications to flourish in the margins.
- For investors, the collapsing cost of intelligence signals a pivotal shift: the next big returns may come not from AI builders, but from those embedding cheap, ubiquitous AI into everyday products.

In a year dominated by multimodal marvels and reasoning breakthroughs, perhaps the most economically significant shift in AI went largely underplayed: cost collapse. The 2025 AI Index Report delivers an extraordinary statistic that should be front-page news for any enterprise strategist, systems architect or national policy maker:

The cost to generate one million tokens at GPT-3.5 quality fell from $20.00 in November 2022 to just $0.07 by October 2024. That's a 280-fold reduction in less than two years.

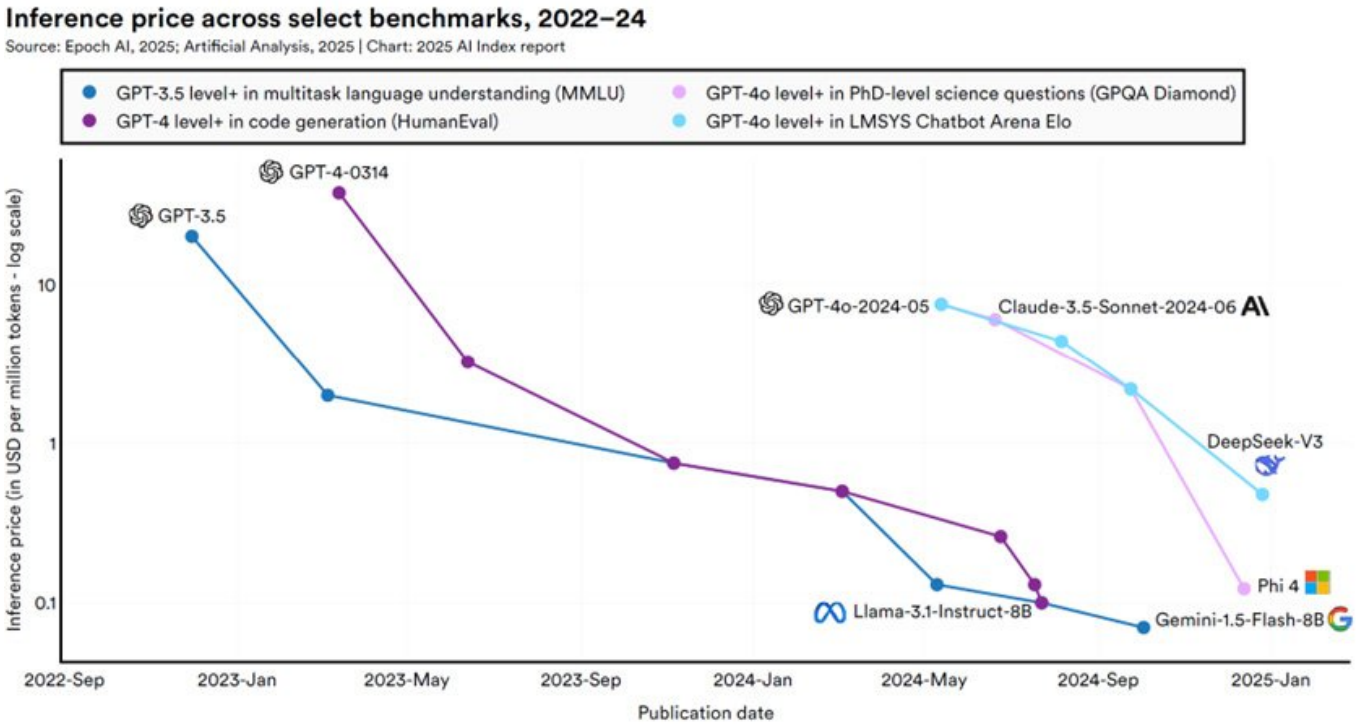
In less than two years, the cost to run advanced AI models has collapsed by more than 90%. At first glance, this might seem like a footnote in a broader AI narrative dominated by model breakthroughs and benchmark scores. But this is the economic fulcrum on which the real AI revolution turns. The chart (figure 1) tracks models that meet or exceed performance thresholds in language understanding, code generation, scientific reasoning and chatbot ability—each line representing not just a drop in price, but a breaking of economic barriers. Models that once cost $10–$20 per million tokens to run now deliver similar or even superior performance for pennies. This isn't just about cost efficiency; it's about unlocking entire classes of applications that were previously too expensive to deploy at scale.

If you were building a product in 2022, using GPT-4-level reasoning in real time was either a luxury or a loss leader. Today, that same capability is suddenly viable for mass-market software, IoT devices and even embedded AI agents that work without cloud connectivity. The convergence of affordability and capability means we're transitioning from a world where access to intelligence was scarce and gated to one where it becomes ambient and assumed. And while headlines focus on model capabilities, this steep drop in

inference cost is arguably the more important story for long-term investors and operators. Why? Because infrastructure-level cost collapses tend to precede explosive waves of application-layer innovation.

There's a lesson here we've seen before: When the cost of computation fell, we got the personal computer. When bandwidth costs fell, we got YouTube and Zoom. What happens when the cost of deploying PhD-level reasoning or high-quality code generation falls to near zero? That's possibly a trillion-dollar question. This isn't just about OpenAI, Google or Anthropic. The real story is in what thousands of smaller companies, solo developers and emerging market entrepreneurs will do when this kind of power is no longer a capital expense but a creative tool. The frontier models may get the glory, but the real magic is happening in the margins, where costs are low, constraints breed ingenuity, and the next AI-native products are already being built.

Figure 1: Evolution of Inference Price over Time for Select Models

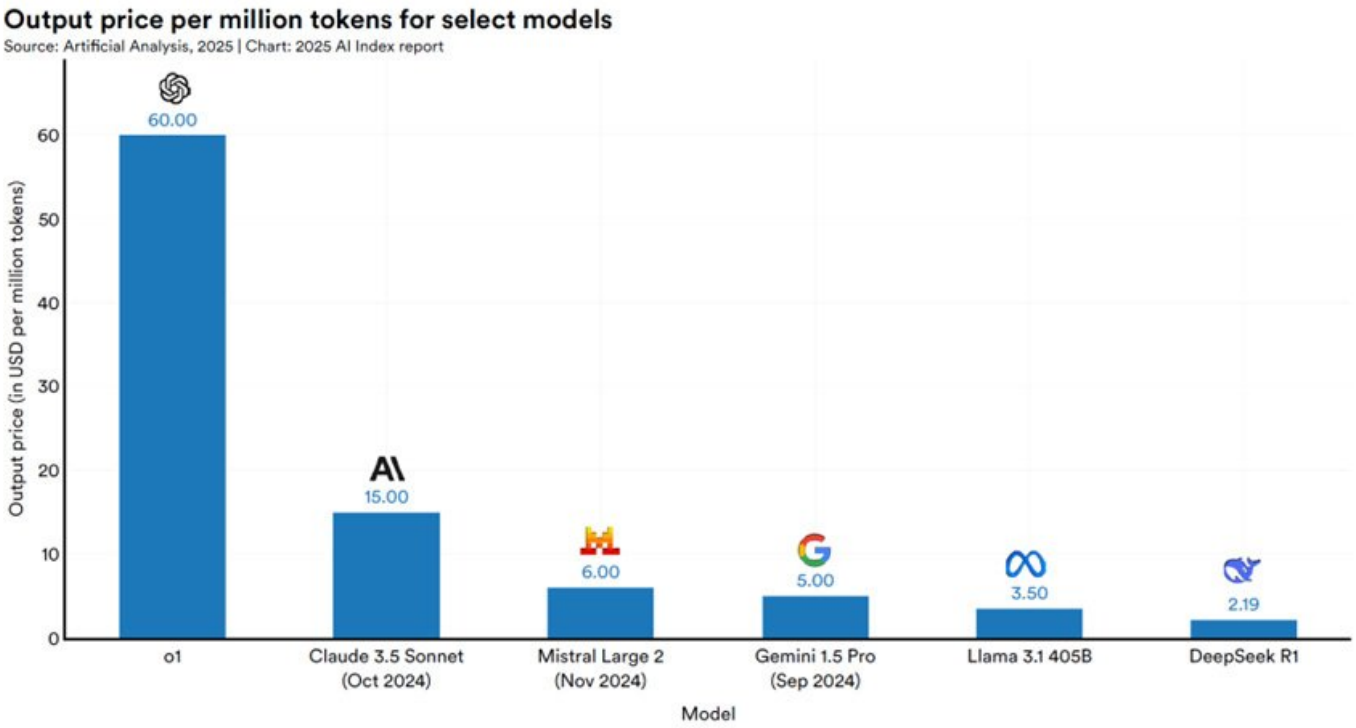


In 2025, the economics of intelligence are fracturing in real time. OpenAI's "o1" model charges $60 per million tokens—a stark contrast to DeepSeek R1 at just $2.19, as we see in figure 2. That 27-fold price gap isn't just a curiosity; it's a signal. We're watching the early innings of a massive deflationary wave in AI utility, where the cost of intelligence could soon approach zero for many applications. The analog here isn't just computing hardware or cloud storage; it's electricity. Once scarce and expensive, now ubiquitous and assumed. The companies priced like luxuries today may not be the ones that scale to ambient, everyday intelligence tomorrow.

Just as index funds reshaped finance by lowering the cost of ownership to near-zero, we may now be witnessing the index-fund moment for AI. A lower output price per token doesn't guarantee a better product, but it fundamentally rewires the incentive to build with it. Meta, Google, Mistral and DeepSeek seem to

be betting that AI isn't just a race for power; it's a race for ubiquity. Claude and OpenAI may still lead in quality for high-stakes enterprise use cases, but the terrain is shifting underneath them. The models that win in the long run might not be the smartest or flashiest, but the ones cheap enough to disappear into the background, everywhere, all the time.

Figure 2: Output Price per Million Tokens for Select Models

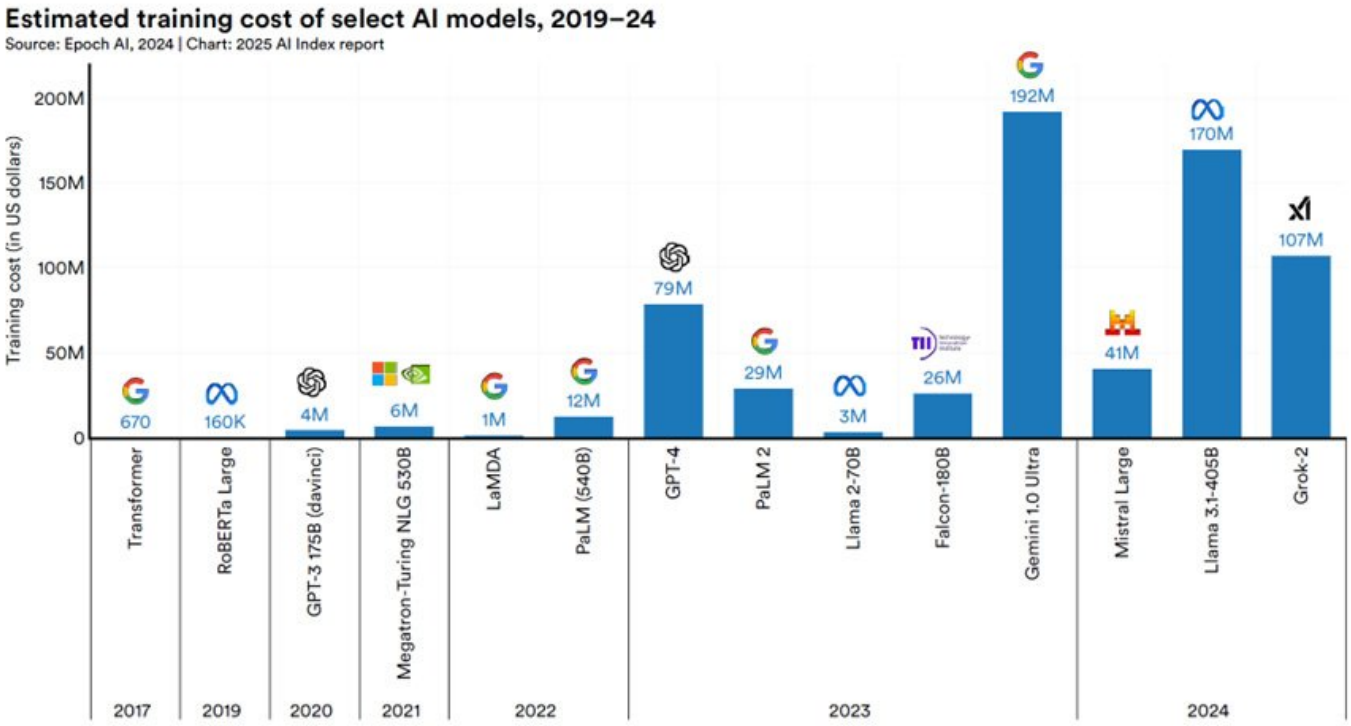


Over the past seven years, the cost of training leading AI models has gone from less than a used laptop to more than a fleet of private jets, as we can see in figure 3. In 2017, Google's seminal Transformer architecture was trained for just $670, a number that feels almost quaint in hindsight. But as the promise of large language models became apparent, capital flooded in. GPT-3 cost $4 million to train in 2020. By 2023, GPT-4 pushed that number to $79 million, and in 2024, Google's Gemini 1.0 Ultra topped the chart at an estimated $192 million. What began as a pure research endeavor has rapidly become one of the most capital-intensive frontiers in modern technology—where the barrier to entry is no longer talent or ideas but how much you're willing to spend on compute. Companies like Meta, OpenAI and xAI are now operating like sovereign research labs, channeling budgets more commonly associated with aerospace defense into the race for model supremacy.

But hidden behind these eye-popping figures is a more nuanced story. Not everyone is playing the same game. While some players chase scale at all costs, others are proving that clever architectures and efficient training pipelines can produce competitive models at a fraction of the budget. Mistral and Falcon, for example, reveal that you don't always need nine figures to make something powerful. It's a divergence in philosophy: one path driven by brute-force scaling, the other by optimization and openness. As this plays out, we're likely witnessing the early formation of distinct AI ecosystems—with different training styles,

deployment philosophies and even geopolitical alignments. The most important takeaway might be that the future of AI won't be written just by the biggest wallets but by those who understand where scale actually delivers value, and where it's just noise.

Figure 3: Evolution of Estimated Training Costs of Select Models



U.S. vs. China

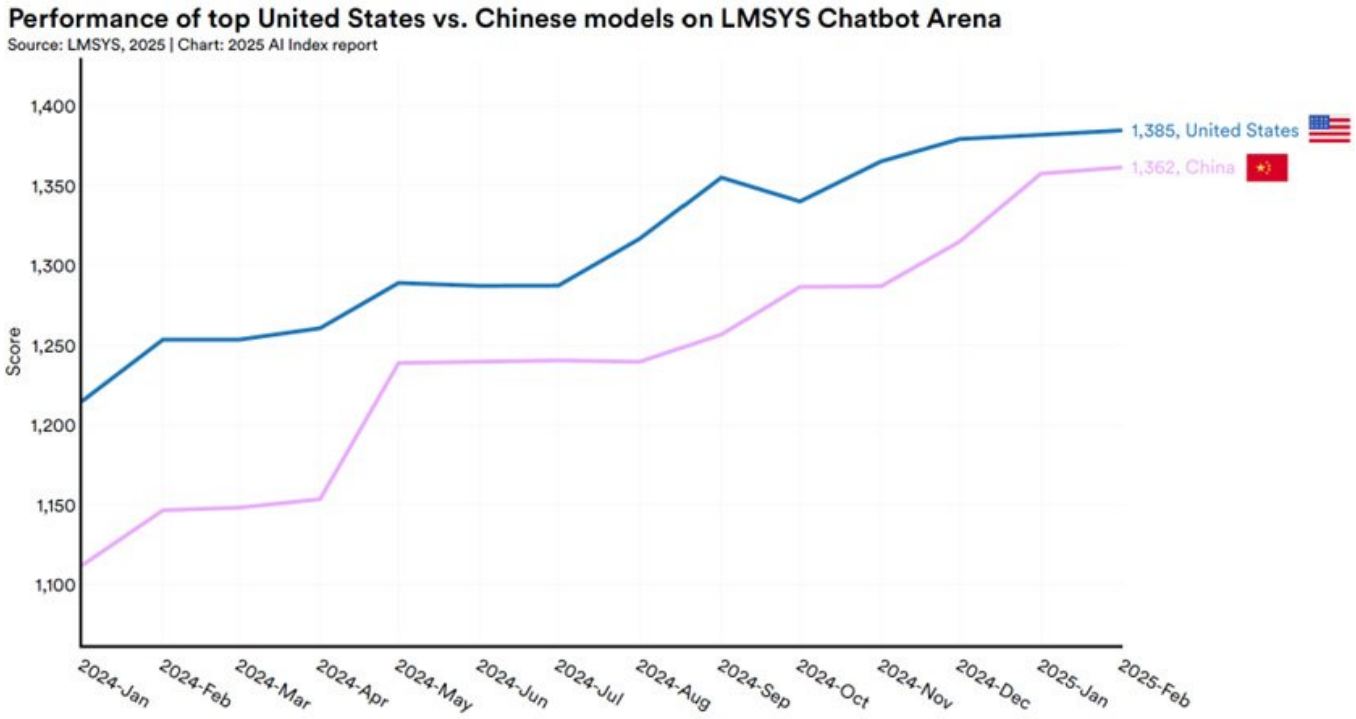
In the race for AI supremacy, the leaderboard is no longer a foregone conclusion. The latest LMSYS Chatbot Arena results show the U.S. maintaining a performance edge, but the story is one of rapid convergence. As we can see in figure 4, in early 2024, American models outscored Chinese counterparts by more than 100 points. In the LMSYS framework, that's not trivial; it reflects *clearly preferred user responses* across a wide range of tasks. Think of it like the Elo rating system in chess: a 100-point gap suggests a decisive, repeatable advantage. But fast forward to early 2025, and that margin has shrunk to just 23 points—a level that, in practice, means users often can't reliably distinguish which model is "better" in a head-to-head encounter. That kind of tightening doesn't just happen; it's the result of material leaps in model quality, likely driven by better pretraining data, architectural tuning and scaled compute on both sides of the Pacific.

This narrowing gap is more than just a statistical footnote; it's geopolitical narrative encoded in model performance. U.S. companies like OpenAI, Anthropic and Google DeepMind continue to define the frontier, bolstered by a blend of research culture, infrastructure and private capital. But China, with its vertically integrated AI champions and state-aligned strategy, is gaining ground faster. The LMSYS data reflects this clearly: while U.S. models gained about 160 points over 13 months, Chinese models gained more than 250.

And this isn't just about chatbots getting "smarter"—it's about real-world deployment. A 20-point difference might not even register in user experience, especially once models are fine-tuned for local language, regulations or sectoral needs. That's where the race gets more complicated and more consequential.

If you zoom out, this isn't about which model aces trivia better; it's about who shapes the decision-making infrastructure of the modern world. Whether it's medical diagnostics, financial modeling or autonomous defense systems, these models are becoming the engines behind institutions. A 100-point lead meant American models were the obvious first choice. A 20-point lead? That's margin-of-error territory, especially when geopolitical alignment, data sovereignty and regional control come into play. The U.S. still leads, but China is no longer trailing in the rearview mirror. It's close enough to draft behind, and in a long enough race, that changes everything.

Figure 4: U.S. vs. China in the Geopolitical AI Race



The Investment Case for AI Ubiquity

For long-term investors, the collapsing cost curves in AI aren't just technical trivia; they're macro signals. We've seen this movie before: when inputs to value creation fall by orders of magnitude, output explodes. The falling cost of inference is to AI what Moore's Law was to computing and mobile bandwidth was to the internet. And just like those shifts, the biggest returns may accrue not to the firms building the substrate, but to those building the everyday tools people will use without even realizing there's AI behind them. As token prices race toward zero, expect monetization models, enterprise adoption curves and even regulatory regimes to shift with them. This is no longer about building AI products; it's about building in an AI-shaped world.

Important Information Related to this Article

Source for all data in this blog post, including figures: Maslej et al., "The AI Index 2025 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 4/25.