Infra

Welcome to LLMflation - LLM inference cost is going down fast ⬇️

Guido Appenzeller
Posted November 12, 2024

SUBSCRIBE

Share



To a large extent, it is a rapid decline in cost of the underlying commodity that drives technology cycles. Two prominent examples of this are Moore's Law and Dennard scaling, which help to explain the PC revolution by describing how chips become more

performant over time. A lesser known example is Edholm's Law, which describes how network bandwidth increases — a key factor in the dotcom boom.

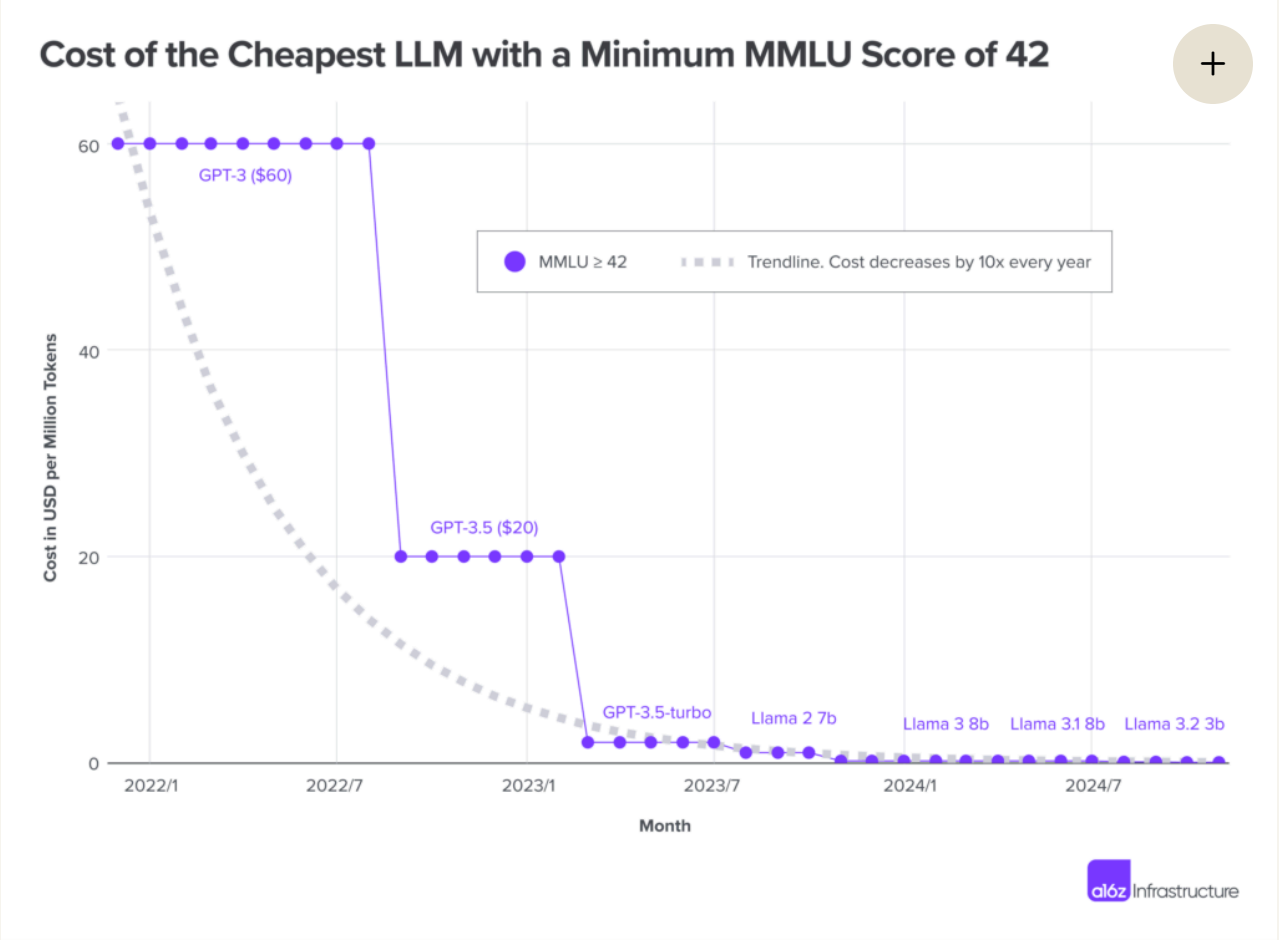
In analyzing historical price data since the public introduction of GPT-3, it appears that — at least so far — a similar law holds true for the cost of inference in large language models (LLMs). We're calling this trend LLMflation, for the rapid increase in tokens you can obtain at a constant price.

In fact, the price decline in LLMs is even faster than that of compute cost during the PC revolution or bandwidth during the dotcom boom: **For an LLM of equivalent performance, the cost is decreasing by 10x every year.** Given the early stage of the industry, the time scale may still change. But the new use cases that open up from these lower price points indicate that the AI revolution will continue to yield major advances for quite a while.

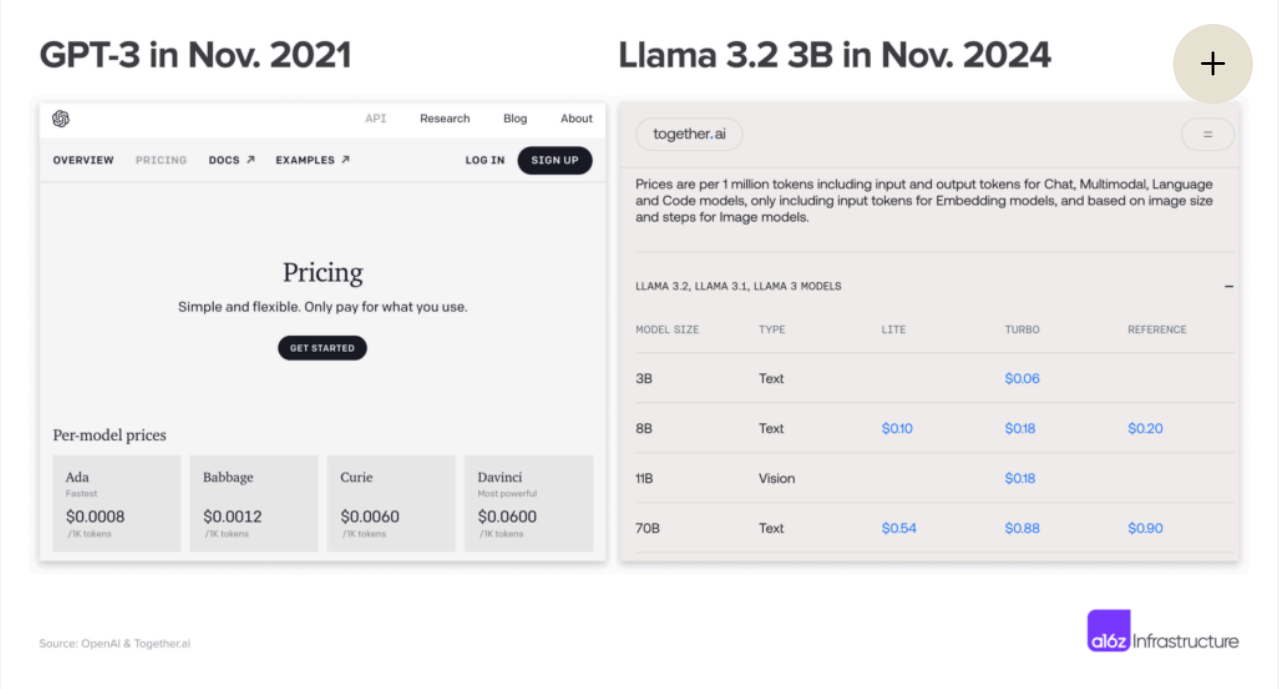
The methodology

In determining this trend, we looked at the performance of LLMs using MMLU scores, as reported by the model creators or external evaluations. LLMs usually price per million tokens (on average, one word is the equivalent of 1-2 tokens), and we obtained historical pricing data for the models from the Internet Archive. If the price differed for input and output tokens, we took the average of the two. To simplify the search, we limited our search to models from OpenAI, Anthropic, and Meta's Llama from third party inference providers.

The graph below shows the cost of the cheapest model for any month that would give us a minimum MMLU score of 42.

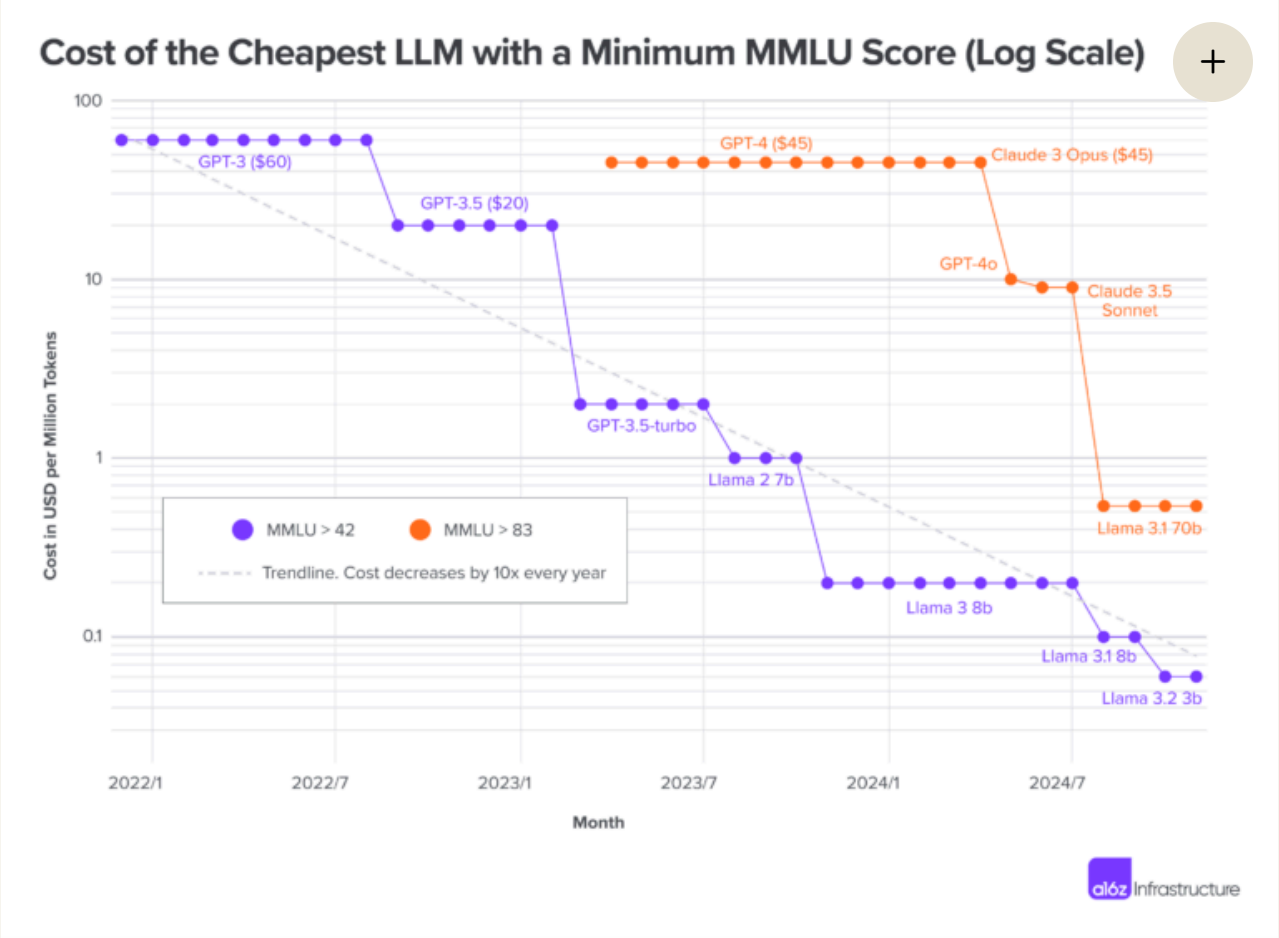


When GPT-3 became publicly accessible in November 2021, it was the only model that was able to achieve an MMLU of 42 — at a cost of $60 per million tokens. As of the time of writing, the cheapest model to achieve the same score was Llama 3.2 3B, from model-as-a-service provider Together.ai, at $0.06 per million tokens. **The cost of LLM inference has dropped by a factor of 1,000 in 3 years.**



If we pick a higher MMLU score of 83, we have less data because models of this quality level have only existed since GPT-4 came out in March of 2023. Since then, however, the price for models at this level have come down by about a factor of 62.

In the logarithmic plot below, we can see that the trend of a 10x decrease every year (the dashed line) is a fairly good approximation of the cost decline across both MMLU performance levels.



While we think the overall result is valid, the methodology is far from perfect. Models can be easily contaminated or intentionally trained on the MMLU benchmark. We also, in some cases, could only find multi-shot data for MMLU (although we are not including any chain-of-thought results in our data). And other models and fine-tunes may have been slightly more cost-effective at any given time. All that said, **there is no question we are seeing an order of magnitude decline in cost every year.**

Will LLM prices continue to decline at this rate?

This is very hard to predict. In the PC revolution, cost decreased to a large degree as a function of Moore's Law and Dennard's Law. It was easy to predict that as long as these laws held and transistor counts and frequencies increased, price drops would continue. In our case, however, the decrease in the cost of LLM inference is caused by a number of independent factors:

- **Better cost/performance of the GPUs** for the same operations. This is a result of Moore's Law (i.e., the increasing number of transistors per chip), as well as structural improvements.
- **Model quantization.** Initially, inference was done at 16-bits, but for Blackwell GPUs we expect 4-bit to become common. That is a net increase of at least 4x in performance, but likely more as less data movement is required and arithmetic units are less complex.
- **Software optimizations** that reduce the amount of compute required and, equally importantly, reduce the required memory bandwidth. Memory bandwidth previously was a bottleneck.
- **Smaller models.** Today, we have a 1-billion-parameter model that exceeds the performance of a 175-billion-parameter model just 3 years ago. A major reason for this is training the models on a larger number of tokens, far beyond what was considered optimal based on Chinchilla scaling laws.
- **Better instruction tuning.** We have learned a lot about how to improve models after the pre-training phase, with techniques such as reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO).
- **Open source.** Meta, Mistral, and others have introduced open models that can be hosted by competing, low-cost model-as-a-service providers. This reduced profit margin across the value chain, which lowered prices.

There is no doubt we will see rapid advancements in some of the areas, but for others, like quantization, it is less clear. So while the cost of LLM inference will likely continue to decrease, its rate may slow down.

Another important question is whether this rapid decrease in cost is a problem for LLM providers. For now it seems like they are willing to concede the low end of the market and instead focus their efforts on the highest-quality tier. Interestingly enough, OpenAI's leading model today, o1, has the same cost per output token as GPT-3 had at launch ($60 per million).

That said, the rapid decrease of LLM inference cost is still a massive boon for AI in general. Every time we decrease the cost of something by an order of magnitude, it opens up new use cases that previously were not commercially viable. For example, humans can speak around 10,000 words per hour. If someone were to speak for 10 hours a day, every day of the year, they could now use a GPT-3-class LLM to process all the words they said for about $2 per year. Processing the entire Linux kernel (about 40 million lines of code) would cost under $1.

Text-to-speech models are equally cheap, so building a simple voice assistant is now essentially free from an inference perspective.

The community will continue to build amazing applications around this technology, and we are super excited to partner with the founders who create the breakthrough companies that bring them to market. It is a great time to be an entrepreneur! ■

About the Contributor



Guido Appenzeller

is an investor at Andreessen Horowitz, where he focuses on AI, infrastructure, open source technology, and silicon.

More From this Contributor +

Expert News by a16z

We have built a network of experts who are deeply rooted in technology and how it's shaping our future. Subscribe to our newsletters to receive their perspectives.

Type your email... Subscribe

By subscribing you agree to Substack's Terms of Use, our Privacy Policy and our Information collection notice

substack

See All Newsletters

Views expressed in "posts" (including podcasts, videos, and social media) are those of the individual a16z personnel quoted therein and are not the views of a16z Capital Management, L.L.C. ("a16z") or its respective affiliates. a16z Capital Management is an investment adviser registered with the Securities and Exchange Commission. Registration as an investment adviser does not imply any special skill or training. The posts are not directed to any investors or potential investors, and do not constitute an offer to sell — or a solicitation of an offer to buy — any securities, and may not be used or relied upon in evaluating the merits of any investment.

The contents in here — and available on any associated distribution platforms and any public a16z online social media accounts, platforms, and sites (collectively, "content distribution outlets") — should not be construed as or relied upon in any manner as investment, legal, tax, or other advice. You should consult your own advisers as to legal, business, tax, and other related matters concerning any investment. Any projections, estimates, forecasts, targets, prospects and/or opinions expressed in these materials are subject to change without notice and may differ or be contrary to opinions expressed by others. Any charts provided here or on a16z content distribution outlets are for informational purposes only, and should not be relied upon when making any investment decision. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by a16z. While taken from sources believed to be reliable, a16z has not independently verified such information and makes no representations about the enduring accuracy of the information or its appropriateness for a given situation. In addition, posts may include third-party advertisements; a16z has not reviewed such advertisements and does not endorse any advertising content contained therein. All content speaks only as of the date indicated.

Under no circumstances should any posts or other information provided on this website — or on associated content distribution outlets — be construed as an offer soliciting the purchase or sale of any security or interest in any pooled investment vehicle sponsored, discussed, or mentioned by a16z personnel. Nor should it be construed as an offer to provide investment advisory services; an offer to invest in an a16z-managed pooled investment vehicle will be made separately and only by means of the confidential offering documents of the specific pooled investment vehicles — which should be read in their entirety, and only to those who, among other requirements, meet certain qualifications under federal securities laws. Such investors, defined as accredited investors and qualified purchasers, are generally deemed capable of evaluating the merits and risks of prospective investments and financial matters.

There can be no assurances that a16z's investment objectives will be achieved or investment strategies will be successful. Any investment in a vehicle managed by a16z involves a high degree of risk including the risk that the entire amount invested is lost. Any investments or portfolio companies mentioned, referred to, or described are not representative of all investments in vehicles managed by a16z and there can be no assurance that the investments will be profitable or that other investments made in the future will have similar characteristics or results. A list of investments made by funds managed by a16z is available here: https://a16z.com/investments/. Past results of a16z's investments, pooled investment vehicles, or investment strategies are not

necessarily indicative of future results. Excluded from this list are investments (and certain publicly traded cryptocurrencies/ digital assets) for which the issuer has not provided permission for a16z to disclose publicly. As for its investments in any cryptocurrency or token project, a16z is acting in its own financial interest, not necessarily in the interests of other token holders. a16z has no special role in any of these projects or power over their management. a16z does not undertake to continue to have any involvement in these projects other than as an investor and token holder, and other token holders should not expect that it will or rely on it to have any particular involvement.

With respect to funds managed by a16z that are registered in Japan, a16z will provide to any member of the Japanese public a copy of such documents as are required to be made publicly available pursuant to Article 63 of the Financial Instruments and Exchange Act of Japan. Please contact compliance@a16z.com to request such documents.

For other site terms of use, please go here. Additional important information about a16z, including our Form ADV Part 2A Brochure, is available at the SEC's website: http://www.adviserinfo.sec.gov.