MBI Deep Dives

BY **MBI DEEP DIVES** IN **OPENAI** — DEC 21, 2025

Sam Altman's Explanation

Jensen Huang in his recent Joe Rogan podcast had an interesting story from early years of Nvidia:

> “**we convinced ourselves** that chip is going to be great. And so I had to call some other gentleman. So I called TSMC...and I explained to them what we were doing. And I explained to him (Morris Chang) I had a lot of customers. **I had one**, you know, Diamond Multimedia...the demand's really great, and we're going to tape out a chip to you, and I like to go directly to production because I know it works.
>
> And they said, "Nobody has ever done that before. Nobody has ever taped out a chip that worked the first time. And nobody starts out production without looking at it."
>
> But I knew that if I didn't start the production, I'd be out of business anyways. And if I could start the production, I might have a chance.
>
> ...as we were starting the production, Morris flew to United States. He didn't so many words asked me so, but **he asked me a whole lot of questions that was trying to tease out do I have any money** but he didn't directly ask me...**so the truth is that we didn't have all the money but we had a strong PO from the customer and if it didn't work some wafers would have been lost. I'm not exactly sure what would have happened but we would have come short, it would have been rough**.”

History doesn’t repeat, but it does rhyme. Two and half decades later, there is another American CEO who is trying to convince everyone that they have a lot of demand! Perhaps the key distinction is while both Nvidia and TSMC back the[illegible]e hardly a footnote in the tech industry, OpenAI today is at the front and cente[illegible] perhaps the most consequential technological revolution in history. If their demand forecast is substantially off, the value destruction in “AI trade” can be whole lot larger than if Nvidia couldn’t pay TSMC in mid 1990s.

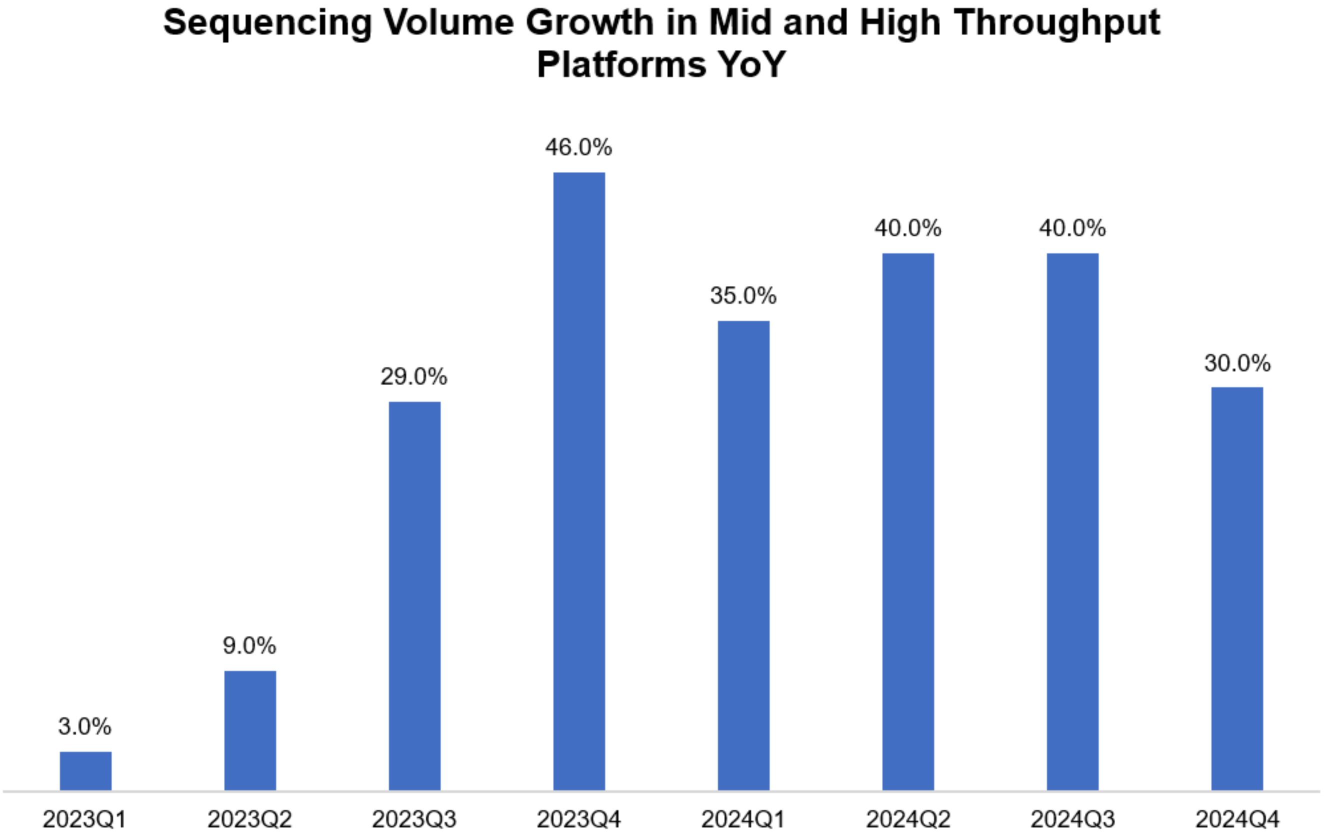
In a recent appearance on Big Technology podcast, Sam Altman was asked about OpenAI’s $1.4 trillion “commitment” to various players in the AI value chain. This was a good podcast with thoughtful, reflective answers from Sam Altman. It is worth listening to the entire episode, but I will focus primarily on his comments regarding infrastructure commitment. Here’s the excerpt on this point:

> “…my learning in the history of this field is once the squiggles start and it lifts off the x-axis a little bit, we know how to make that better and better. But that takes huge amounts of compute to do. So that’s one area—throwing lots of AI at discovering new science, curing disease, lots of other things.
>
> A kind of recent, cool example: we built the Sora Android app using Codex. **They did it in less than a month.** They used a huge amount—one of the nice things about working at OpenAI is you don’t get any limits on Codex. They used a huge amount of tokens, but they were able to do what would normally have taken a lot of people much longer. And Codex kind of mostly did it for us. And you can imagine that going much further, where entire companies can build their products using lots of compute.
>
> People have talked a lot about video models pointing towards these generated, real-time generated user interfaces that will take a lot of compute. Enterprises that want to transform their business will use a lot of compute. Doctors that want to offer good, personalized health care that are constantly measuring every sign they can get from each individual patient—you can imagine that using a lot of compute.
>
> It’s hard to frame how much compute we’re already using to generate AI output in the world, but these are horribly rough numbers, and I think it’s undisciplined to talk this way, but I always find these mental thought experiments a little bit useful. So forgive me for the sloppiness.
>
> Let’s say that an AI company today might be generating something on the order of 10 trillion tokens a day out of frontier models. More, but it’s not like a quadrillion tokens for anybody, I don’t think. Let’s say there’s 8 billion people in the world, and let’s say on average, the average number of tokens outputted by a

> person per day is like 20,000—these are, I think, totally wrong. But you can then start—and to be fair, we'd have to compare the output tokens of a model provider today, not all the tokens consumed—but you can start to look at this, and you can say, we're gonna have these models at a company be outputting more tokens per day than all of humanity put together, and then 10 times that, and then 100 times that.
>
> In some sense, it's like a really silly comparison, but in some sense, it gives a magnitude for how much of the intellectual crunching on the planet is human brains versus AI brains, and those relative growth rates there are interesting.

This answer is a good encapsulation of why analyzing companies in the AI value chain has become more challenging. Altman's explanation for demand is not non-sensical at all; it is certainly possible the average user in 5-10 years will utilize order of magnitude more tokens per day than we are today. But the more challenging aspect is to project how pricing of such token will evolve over time. While thinking about compute demand, I keep thinking about what I pointed in my Illumina Deep Dive:

> "I do find it quite interesting that Illumina's revenue will be basically flat from 2021 to 2026. As alluded earlier, studying Illumina is bit of a cautionary tale how Jevons paradox doesn't really absolve us from difficult questions. Just to give you perspective, while Illumina's core revenue slightly declined in both 2023 and 2024, their sequencing volume data kept growing at a pretty healthy rate. Given cost of sequencing fell faster thanks to transition to higher throughput instruments as well as due to potentially competitive factors, Illumina couldn't grow their revenue. **When cost of sequencing kept falling, Illumina's customers did increase volume of sequencing materially but that wasn't enough to outweigh the pricing pressure**.



Illumina's sequencing volume growth data

Later in the podcast, Altman shared how they plan to reach profitability:

> “As revenue grows and as inference becomes a larger and larger part of the fleet, it eventually subsumes the training expense. So that’s the plan. Spend a lot of money training but make more and more. If we weren’t continuing to grow our training costs by so much, we would be profitable way, way earlier. But the bet we’re making is to invest very aggressively in training these big models.”

The only problem is...that is everyone’s plan! **The optimal strategy for Google is to keep inference prices so low that it remains very hard for inference revenue to subsume training runs for all the other model developers**. If OpenAI had a monopoly in building frontier models, you can bet their inference revenues would be easily able to supersede training costs and perhaps make respectable operating margins. But if everyone remains in lock-step in the red queen race of training the next model while the pricing for inferences keeps falling precipitously, the **economics** can remain far from compelling. It’s a risky bet when such questions are still pretty much up in the air, especially at $830 Billion valuation.

But hey, it worked out just fine for Nvidia even though Jensen too didn’t really have demand lined up for his chips. The age old American audacity of "just go for it" without having all the answers may still have its final say!

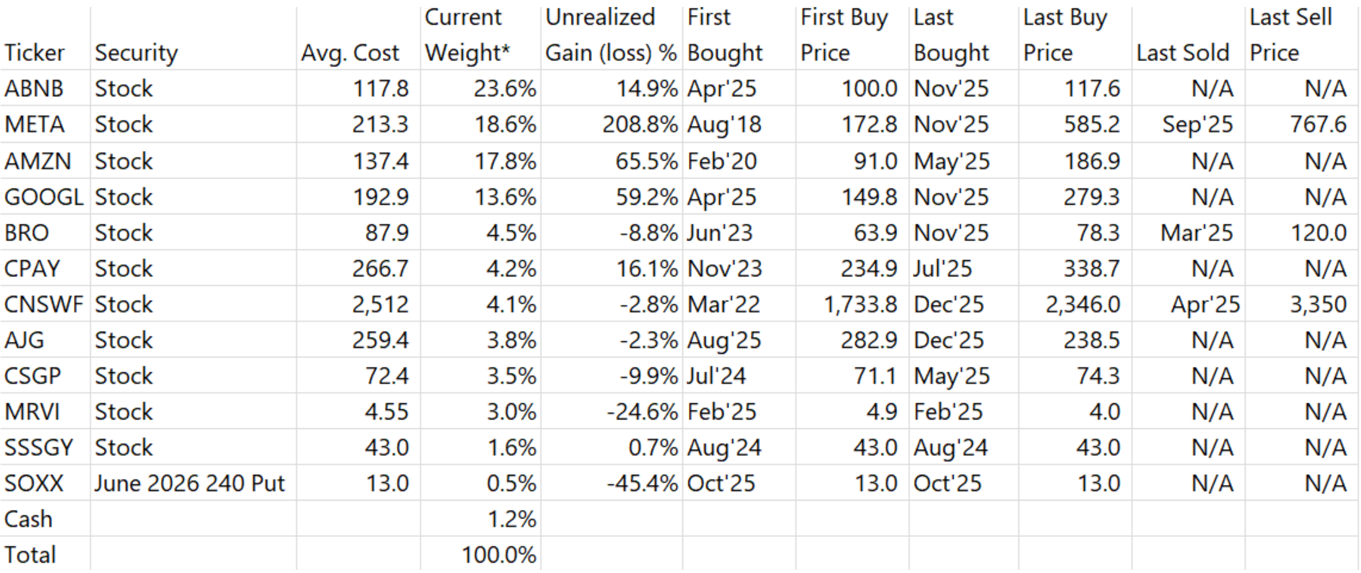
*In addition to “Daily Dose” (yes, **DAILY**) like this, MBI Deep Dives publishes one Deep Dive on a publicly listed company every month. You can find all the 65 Deep Dives [here]()*

Subscribe

Current Portfolio:

Please note that these are **NOT** my recommendation to buy/sell these securities, but just disclosure from my end so that you can assess potential biases that I may have because of my own personal portfolio holdings. Always consider my write-up my personal investing journal and never forget my objectives, risk tolerance, and constraints may have no resemblance to yours.

My current portfolio is disclosed below:



Ticker	Security	Avg. Cost	Current Weight*	Unrealized Gain (loss) %	First Bought	First Buy Price	Last Bought	Last Buy Price	Last Sold	Last Sell Price
ABNB	Stock	117.8	23.6%	14.9%	Apr'25	100.0	Nov'25	117.6	N/A	N/A
META	Stock	213.3	18.6%	208.8%	Aug'18	172.8	Nov'25	585.2	Sep'25	767.6
AMZN	Stock	137.4	17.8%	65.5%	Feb'20	91.0	May'25	186.9	N/A	N/A
GOOGL	Stock	192.9	13.6%	59.2%	Apr'25	149.8	Nov'25	279.3	N/A	N/A
BRO	Stock	87.9	4.5%	-8.8%	Jun'23	63.9	Nov'25	78.3	Mar'25	120.0
CPAY	Stock	266.7	4.2%	16.1%	Nov'23	234.9	Jul'25	338.7	N/A	N/A
CNSWF	Stock	2,512	4.1%	-2.8%	Mar'22	1,733.8	Dec'25	2,346.0	Apr'25	3,350
AJG	Stock	259.4	3.8%	-2.3%	Aug'25	282.9	Dec'25	238.5	N/A	N/A
CSGP	Stock	72.4	3.5%	-9.9%	Jul'24	71.1	May'25	74.3	N/A	N/A
MRVI	Stock	4.55	3.0%	-24.6%	Feb'25	4.9	Feb'25	4.0	N/A	N/A
SSSGY	Stock	43.0	1.6%	0.7%	Aug'24	43.0	Aug'24	43.0	N/A	N/A
SOXX	June 2026 240 Put	13.0	0.5%	-45.4%	Oct'25	13.0	Oct'25	13.0	N/A	N/A
Cash			1.2%							
Total			100.0%							

**Based on closing prices as of December 19, 2025 (time-weighted YTD: +6.5%); Since inception (August 24, 2018) time-weighted annualized return CAGR +16.5%*

***Disclaimer:** All posts on "MBI Deep Dives" are for informational purposes only. This is NOT a recommendation to buy or sell securities discussed. Please do your own work before investing your money.*

← **PREVIOUS ISSUE**

Veeva Update

NEXT ISSUE →

Veeva's Opportunity in Clinical Software