EDITOR **MICHAEL ZYDA**
University of Southern California;
zyda@mikezyda.com

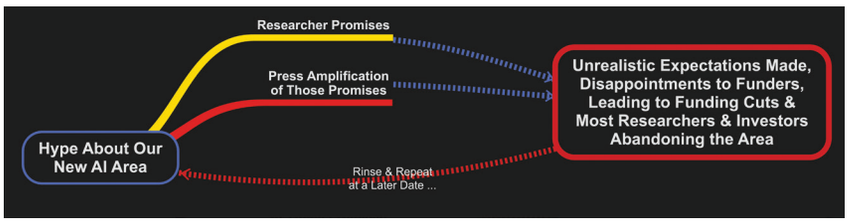

FIGURE 1. The AI hype cycle.

places are also far away from universities capable of graduating engineers that can actually build and run said data centers.

Is the economic crisis going to limit our ability to fly us to the moon to dream among the AI stars? (Figure 2)

Well, Nvidia stock was going to the moon, and now it's below the 2024 stock-split level, despite the promises of megacompanies to purchase millions of graphics processing units (GPUs) at a time!

We see Microsoft planning on investing US$80 billion into AI data centers in 2025. Additionally, we see Microsoft signing a 20-year deal with Constellation Energy to purchase power from the renovated Three-Mile Island nuclear power plant, with Constellation spending US$1.6 billion to restart the facility.

We see Google investing US$75 billion into AI data centers in 2025 after having spent US$52.5 billion in 2024 on AI and cloud infrastructure.

We see Meta Platforms investing US$65 billion in AI infrastructure in 2025, including a massive AI data center in Louisiana. Meta is also discussing a US$200 billion AI data center in states like Louisiana, Wyoming, and Texas.

Amazon (AWS) is investing over US$100 billion in 2025 to boost its AI and cloud infrastructures, including building new data centers and upgrading existing ones to support AI services.

We see xAI is part of a US$30 billion fund with Microsoft and Blackrock, an

COMMENTS?

If you have comments about this article, or topics or references I should have cited or you want to rant back to me on why what I say is nonsense, I want to hear. Every time we finish one of these columns, and it goes to print, what I'm going to do is get it up online and maybe point to it at my Facebook (mikezyda) and my LinkedIn (mikezyda) pages so that I can receive comments from you. Maybe we'll react to some of those comments in future columns or online to enlighten you in real time! This is the "Games" column. You have a wonderful day.

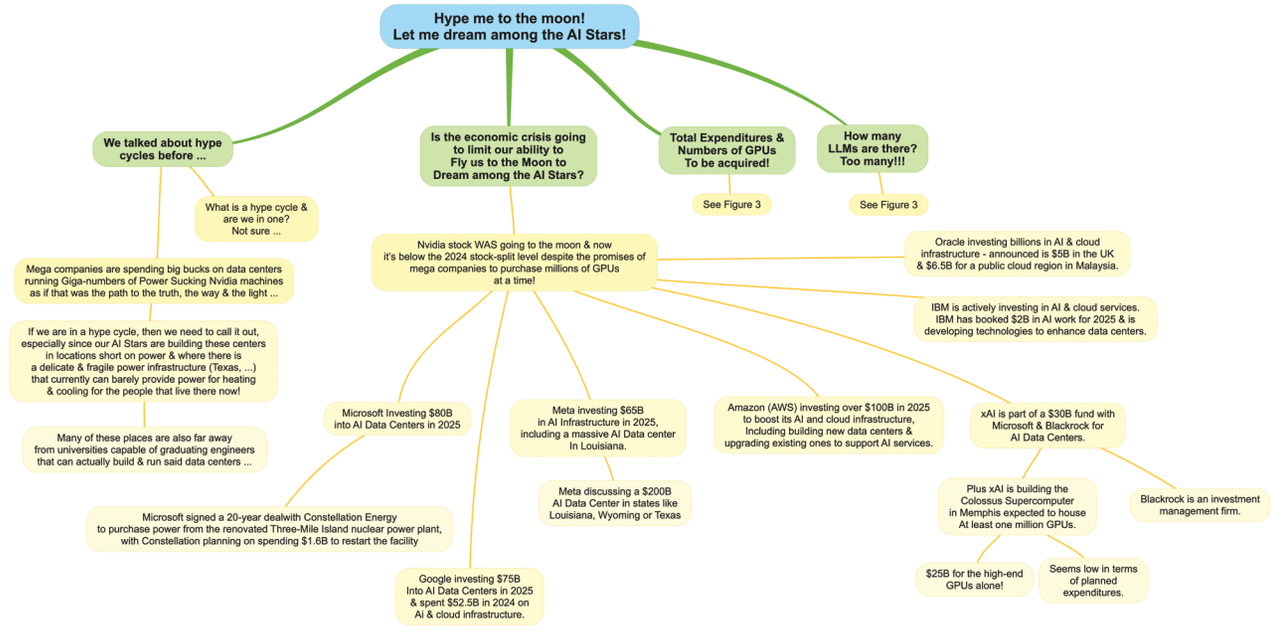

FIGURE 2. Hype me to the moon! Let me dream among the AI stars!

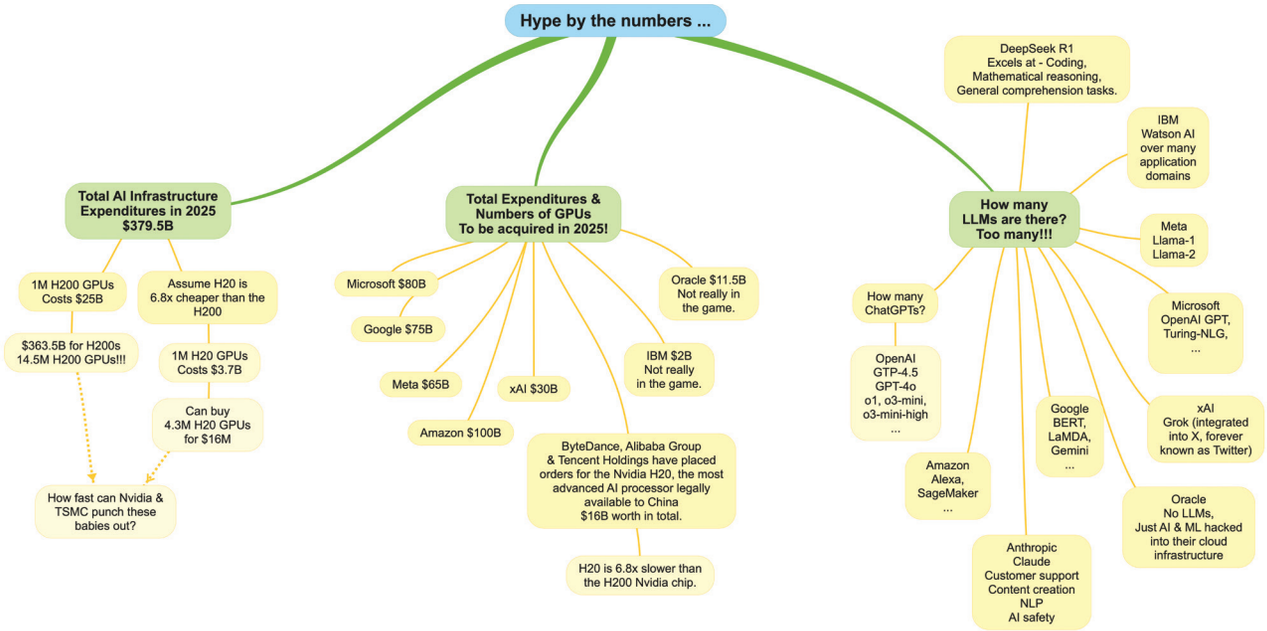

FIGURE 3. Hype by the numbers.

investment management firm, for AI data centers. Plus, xAI is building the Colossus Supercomputer in Memphis, which is expected to house at least one million GPUs, with US$25 billion being spent on the high-end GPUs alone! xAI's planned expenditures seem low in comparison to some of the other megacompanies in this space.

Other companies dabbling in this space are Oracle, which is investing in AI and cloud infrastructure. US$5 billion of that investment is in the United Kingdom and US$6.5 billion is in Malaysia for a public cloud region.

IBM is actively investing in AI and cloud services. IBM has booked US$2 billion in AI work for 2025 and is developing technologies to enhance data centers.

HYPE BY THE NUMBERS

Well, if we have hype rather than big business development, we would expect to hear a lot of noise and wild promises but not see large expenditures. So, let's look at total expenditures and the numbers of GPUs to be acquired in 2025. In Figure 3, we see Microsoft (US$80 billion), Alphabet/Google (US$75 billion), Meta Platforms (US$65 billion), Amazon (US$100 billion), xAI (US$30 billion), ByteDance/Alibaba group/Tencent Holdings (US$16 billion), IBM (US$2 billion), and Oracle (US$11.5 billion). So, total AI infrastructure expenditures in 2025 is about US$379.5 billion, not a small number; in fact, it is a very nice number in terms of infrastructure spending.

> If we are in a hype cycle, then we need to call it out.

We want to figure out how many GPUs this is going to buy. We already know that 1 million Nvidia H200 GPUs is about US$25 billion. If we subtract out the Chinese GPU investment, this is US$363.5 billion for H200s or 14.5 million H200 GPUs!

China is restricted to the purchase of H20 GPUs, which are 6.8 times slower than the H200 Nvidia GPUs. So, ByteDance/Alibaba Group/Tencent Holdings can purchase some 4.3 million H20 GPUs for their US$16 million investment.

So, the real question then becomes how fast can Nvidia and Taiwan Semiconductor Manufacturing Corporation punch these GPUs out? Can they even make them all in 2025? We asked ChatGPT[a] what was Nvidia's production capacity in 2025, and it said no number was released for 2024 or 2025. ChatGPT did state that in 2023, Nvidia shipped around 3.76 million data center GPUs, which is way short of the 14.5 million H200s being ordered in 2025. The 3.76 million number is about the number that the three Chinese AI companies require, 4.3 million H20 GPUs.

So, in order for Nvidia to meet the planned 14.5 million GPU orders for 2025, it has to have increased its production capacity by 3.9 times over its 2023 production capability! Wow! Lots of demand but four times too small a production capacity in 2023, but maybe reasonable in 2025! So, we are not hype, but with the market crash and the current economic mess, we

[a]All of the numbers in AI infrastructure investments were searched for using ChatGPT 4o on 4 April 2025.

don't see this easily happening as the placement of money by companies onto the GPU production bus may no longer be possible.

HOW MANY LARGE LANGUAGE MODELS ARE THERE?

There are way too many large language models (LLMs)! I am not sure that it is possible for me to count them all. I listed the obvious ones, the ones from the megacompanies in the news (Figure 3).

OpenAI is the starting point, and the question to ask is, how many ChatGPTs are there? It seems there is a new one with a special purpose almost every week, and most people I know are using ChatGPT of some version. The biggest issue with ChatGPT though is its legal problem over the licensing of the training data it used in creating its GPT line of models.

Amazon has Alexa and SageMaker for their "AI models," and they are very primitive—they are nowhere near ChatGPT's capabilities. Amazon does offer Anthropic's Claude models on AWS. OpenAI's GPT models are not available on AWS. AWS does offer its Amazon Nova and Amazon Titan AI models on AWS as well as other foundation models through Amazon Bedrock.

Anthropic is the one most people ask me about with respect to investing. Anthropic's Claude LLM is listed as good for customer support, content creation, natural language processing, and AI safety. So, it seems it has several great reasons for its existence and usage.

Google has its own model, Gemini, and there is also Gemini Pro. Gemini is good for natural language understanding, conversation, and other capabilities. In my experience, Gemini seems less capable than ChatGPT, and I have found that often Gemini gives me made-up or just plain wrong answers for the question I asked. Besides Gemini, Google has other LLMs, BERT, LaMDA, and others built internally as part of their research efforts.

xAI has Grok, which is integrated into X (forever known as Twitter). xAI's Grok most likely will die as the CEO has stopped paying attention to his quiver of companies—hopefully he will be successful in flying to Mars. The Grok model focuses on real-time information retrieval and conversation. Grok is "designed to be helpful, honest, and harmless, integrating data from X (forever known as Twitter)." We all laughed at that statement, so I had to put it into this article.

Microsoft licenses OpenAI's GPT series of models and additionally has its own Turing-NLG. Turing-NLG is a powerful natural language generation model known for its ability to generate human-like text, understand context and assist in various applications, like chatbots, summarization, and content creation. Turing-NLG is a very capable model and will succeed as Microsoft is a megacompany with a long history of success.

Meta Platforms has its Llama series of models, which are known to be outstanding for natural language processing. The Llama series of models are available on Amazon AWS through Amazon Bedrock.

DeepSeek's R1 model excels at coding, mathematical reasoning, and general comprehension tasks. It is a Chinese model that has had some recent success. DeepSeek's R1 models are available on Amazon Bedrock. U.S. companies may avoid using it because it is from China, which is just the way it goes. That is sad.

IBM Watson AI works well and has been deployed by IBM over many application domains. IBM is mostly not in the news with respect to LLMs, but they have a long history of deploying technology quietly.

Oracle says it uses its LLMs embedded into their cloud infrastructure for enterprise solutions, but there is rare press about their LLM efforts.

So, lots of models to choose from which makes it difficult, but choosing OpenAI's GPT series, Google's Gemini once it has been cleaned up, Anthropic's Claude, Microsoft's Turing-NLG, or Meta's Llama seem like reasonable choices. It comes down to what is your favorite company. That favorite company might just be Amazon AWS as it provides standardized access to almost all of the foundational models! Eventually, I believe these models will merge in capability and will all seem like the same thing, especially since we will most likely end up getting them through Amazon AWS, just like our groceries ...

Well, after all that, we see a great business in building GPU hardware to be placed in all of the promised data centers, but we don't know when all of those centers will receive their GPUs to be installed there. We see there are plenty of great LLMs, many with overlapping purposes and differing qualities of outputs that will be repaired over time. So, I want to say NOT on the hype label, except for statements by that CEO that needs to get to Mars soonest!!!

ACKNOWLEDGMENT

The author wishes to thank those readers who have gotten to the end of this bimonthly column without finding all of the deliberate and accidental errors. And I apologize for my personal commentaries, but I think they are very important for our futures. And I really do hope that a select section of mankind ends up on Mars!!!

REFERENCE

1. M. Zyda "Large language models and generative AI, oh my!," *Computer*, vol. 57, no. 3, pp 127–132, Mar. 2024, doi: 10.1109/MC.2024.3350290.

MICHAEL ZYDA is an emeritus professor of practice at the University of Southern California, Los Angeles, CA 90007 USA. Contact him at zyda@mikezyda.com.