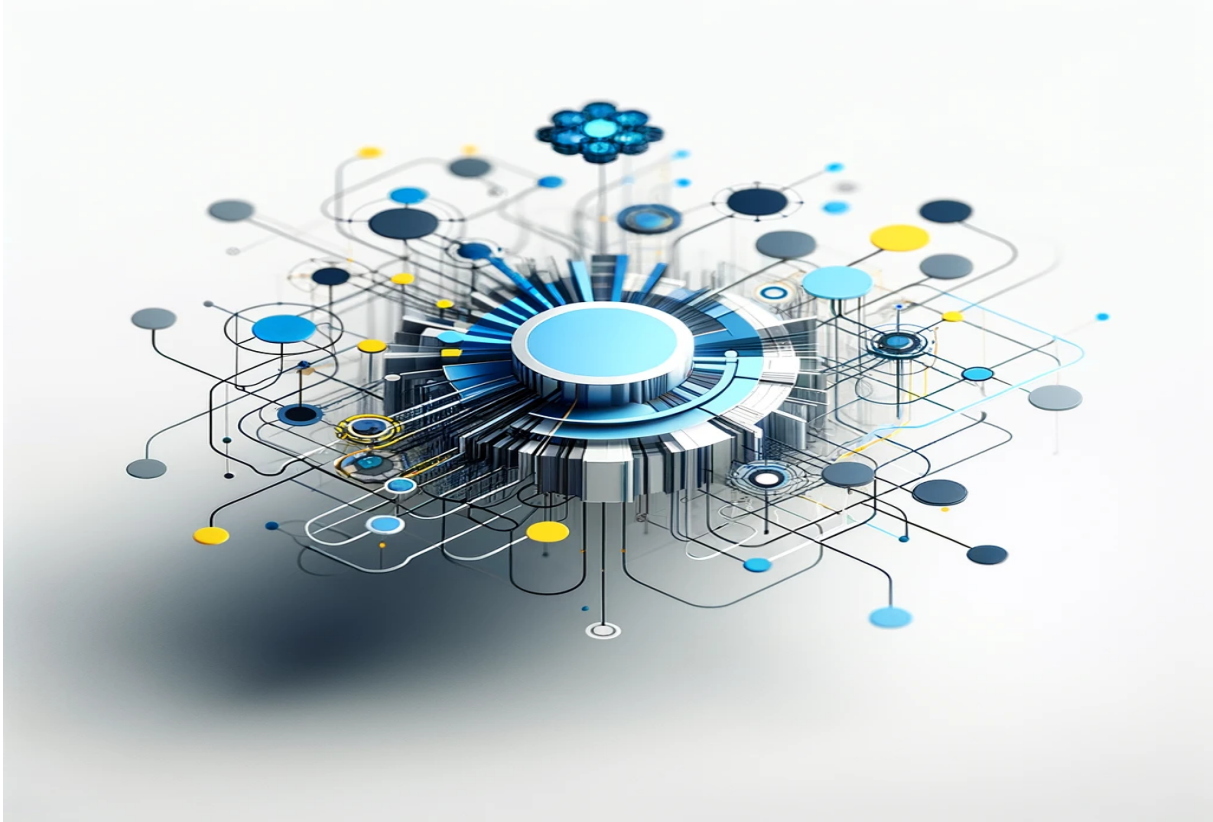


From AI Bubble to AI Superstructures

The Wall Street Journal published an article a few days ago entitled "The AI Revolution is Already Losing Steam" by Christopher Mims. As you might guess from the title, Mims projects future disappointment in what he calls the "AI hype train."

Mims builds his case on three assumptions: the slowing rate of AI improvement, the high costs and energy consumption associated with training and running AI models, and the potential for AI to become commoditized.

He draws the conclusion that the AI industry is heading towards significant disappointment in terms of its capabilities and financial returns. Additionally, Mims suggests that today's massive investment in AI infrastructure could parallel the investment in fiber optics in the late 1990s, which was overbuilt initially and did not realize returns for the original investors.

Mims makes some good points, such as that AI will be commoditized and that large language models (LLMs) are running out of new data to train on. With knowledge of the

early stages of previous technology revolutions, raising a flag of caution about the investment returns of their early stages is certainly warranted.

Mims is one of the better technology columnists, but I think he has it wrong on AI. Not the "hype train" part, which might be making markets a bit too frothy. But rather the "losing steam" part. Where he sees problems with AI, I see strengths. For example, he misses a key development that is already occurring – the transformative potential of leveraging the commoditization of AI models by integrating and orchestrating multiple models into a cohesive system. What Mims thinks is a weakness is actually a strength. Meanwhile, I am not aware of any technology revolution that did feature dramatically declining costs and consistently increasing efficiency over time.

LLMs: The Netscape of AI

Many of the pessimistic forecasts about AI today are focused on the current state of LLMs. They naturally get all the attention today because we can interact with them in a way that is innate to us—through speech. Built into the negativity is an assumption that LLMs need to grow ever larger based on training with increasingly large amounts of text and maybe there just isn't much more left.

I'd like a dollar for every time I have read about "peak" anything and it was wrong (it almost always is). But, really, in terms of language skills, at least in English, how much better do LLMs have to be? While LLMs embed an enormous amount of information, their primary function is conversational engagement. They are already astonishingly good at understanding language and quickly generating well formulated and thorough responses. Meanwhile, AI is getting really good at near-perfect information retrieval, even from a massive, multi-modal corpus.

The (Much) Broader AI Landscape

But LLMs are only a small subset of the total AI picture. Most AI models are not LLMs. A large variety of specialized AI models, often referred to as machine learning (ML) models, play crucial roles in various applications. The number of different types of AI models and their applications is ever-expanding, covering areas such as computer vision, speech recognition, time series forecasting, and more. Some easily recognizable examples include:

1. Drug Discovery Models

   Application: These models analyze biological data to discover new drug candidates, predict their effects, and optimize their chemical structures.
   Example: DeepMind's AlphaFold, which predicts protein folding, significantly aiding in understanding diseases and developing new treatments.

2. Speech Synthesis Models

   Application: Used to create highly realistic synthetic voices for virtual assistants, audiobooks, and customer service automation.
   Example: Google's WaveNet, which generates natural-sounding speech and has been integrated into Google Assistant and Google Translate.

3. Autonomous Navigation Models

   Application: Enable self-driving cars, drones, and robots to navigate through environments autonomously.
   Example: Tesla's Autopilot system, which uses a combination of computer vision, sensor fusion, and reinforcement learning to drive vehicles without human intervention.

4. Augmented Reality (AR) Models:

   Application: Enhance real-world environments with overlaid digital information for applications in gaming, education, and industrial training.
   Example: Microsoft HoloLens, which uses AI to integrate holograms with the physical world, providing an immersive AR experience.

5. Predictive Models

   Application: Predict future events and trends, such as stock market movements, weather patterns, and disease outbreaks.
   Examples:
   - Stock Market Prediction - Models like those developed by Kensho Technologies use AI to analyze financial data and predict market trends.
   - Weather Forecasting - IBM's Watson, which integrates vast amounts of meteorological data to improve weather prediction accuracy (I still don't trust weather forecasts!).
   - Disease Outbreak Prediction - BlueDot uses AI to track, predict, and analyze the spread of infectious diseases, as demonstrated during the COVID-19 pandemic.

These examples showcase the diverse and exciting applications of AI beyond LLMs. But the models I am thinking about are even more specialized and focused - one might call them "AI Primitives." Fish around on Hugging Face to see a wide variety of open-source models easily available for anyone to use today. As of today, **Hugging Face lists 691,093 models**.

AI Superstructures

While individual AI models have impressive capabilities, they are most powerful as components of integrated *AI systems*. AI systems are composed of several different AI

models, with different functions, all being orchestrated by another AI model (or collection of them). In fact, ChatGPT and the other well-known LLMs are already composed of several integrated AI models.

Developers have made a lot of progress, but they are in the early stages of learning how to mix and match AI models to make wholes that are far greater than the sum of their parts. The power of future AI will be in how different models are combined, integrated, and orchestrated into what I call *AI Superstructures*.

AI-based autopilot systems for vehicles are an example of this multi-model approach. The different types of AI models in these systems include:

1. Vision Models - These models process inputs from the car's cameras to create a comprehensive 3D view of the vehicle's surroundings. They detect obstacles, lane markings, traffic lights, and other road users.

2. Sensor Fusion Models - These models combine data from various sensors (cameras, radar, ultrasonic sensors) to improve the accuracy and reliability of the information about the car's environment.

3. Path Planning Models - These models determine the best route and actions the car should take based on the processed sensor data. They make decisions about acceleration, braking, and steering.

4. Control Models - These models execute the decisions made by the path planning models. They control the car's actuators to perform specific actions like turning the steering wheel, applying brakes, or accelerating.

5. Environmental Modeling - This involves creating a dynamic map of the car's surroundings, updating in real-time as the car moves and interacts with the environment.

6. Central Planner (Planning and Control Model) - This model acts as the final decision-maker for the autopilot system. It integrates information from the vision, sensor fusion, path planning, and environmental models to make holistic driving decisions. It then instructs the control models on what specific actions to perform.

I know, we have been hearing promises of truly automated vehicles (AVs) for a while. But the full self-driving (FSD) features already available in cars today are impressive. And the AV robo-taxis that can be seen in some cities were unimaginable only a few years ago. AI systems will eventually master this challenge. There might be a day when seeing a human drivimg a car will seem strange.

Cheer on AI Commoditization

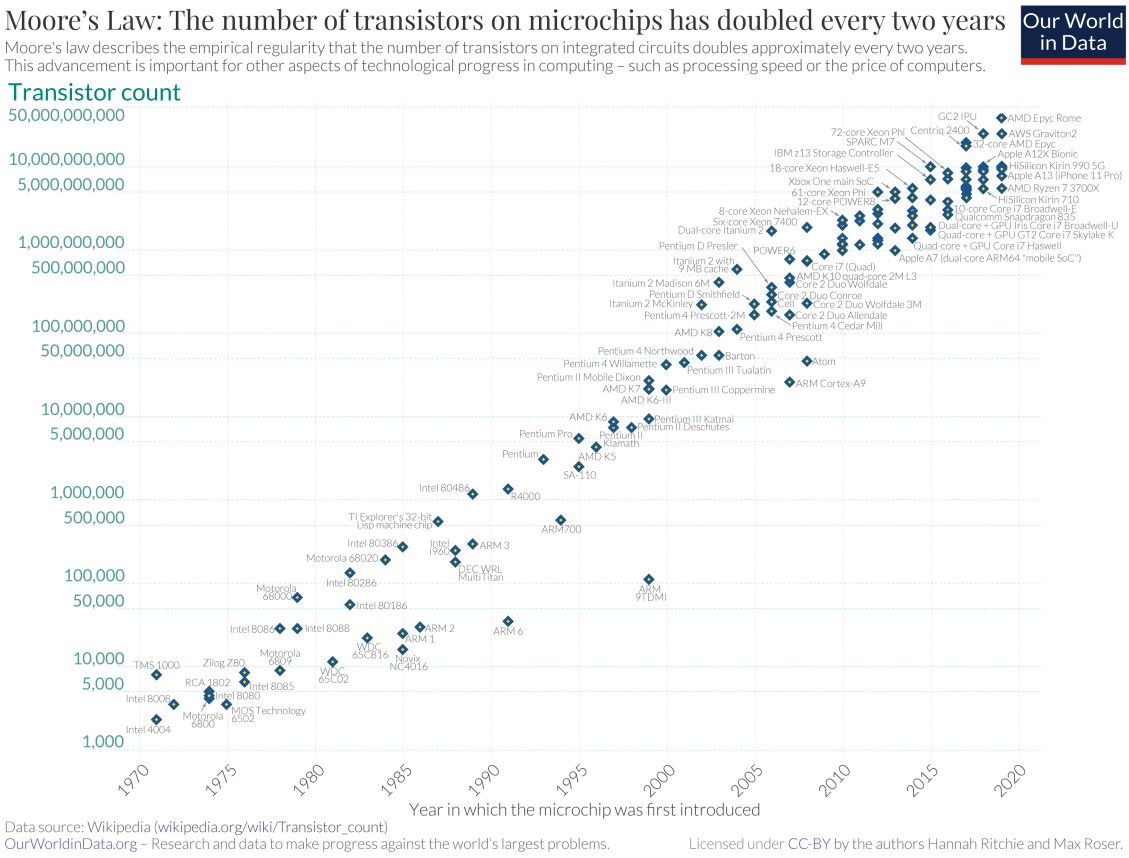
Is the commoditization of AI a bad thing for investors? Actually, it's a requirement for the AI ecosystem to blossom. While a company like Tesla has the resources to develop their own proprietary ML models for their AV system, the commoditization of AI/ML models democratizes the ecosystem for millions of developers. The ever-increasing supply of easily accessible and low-cost AI models unleashes the creativity of developers to build AI superstructures that we have yet to even imagine. (In fact, I'd add that another key aspect that I have not yet seen mentioned will be standardization.)

Commoditized AI models will cost a tiny fraction of Google's Gemini API pricing. They might sometimes be free to download. They will be like software primitives - building blocks with narrow and discreet functions - and it will be possible to combine them in an infinite number of possible configurations and flows. In fact, it will be possible to orchestrate them, changing these configurations on the fly.

How will all of this become so cheap that we can call AI models commodities? Well, the way it always does.

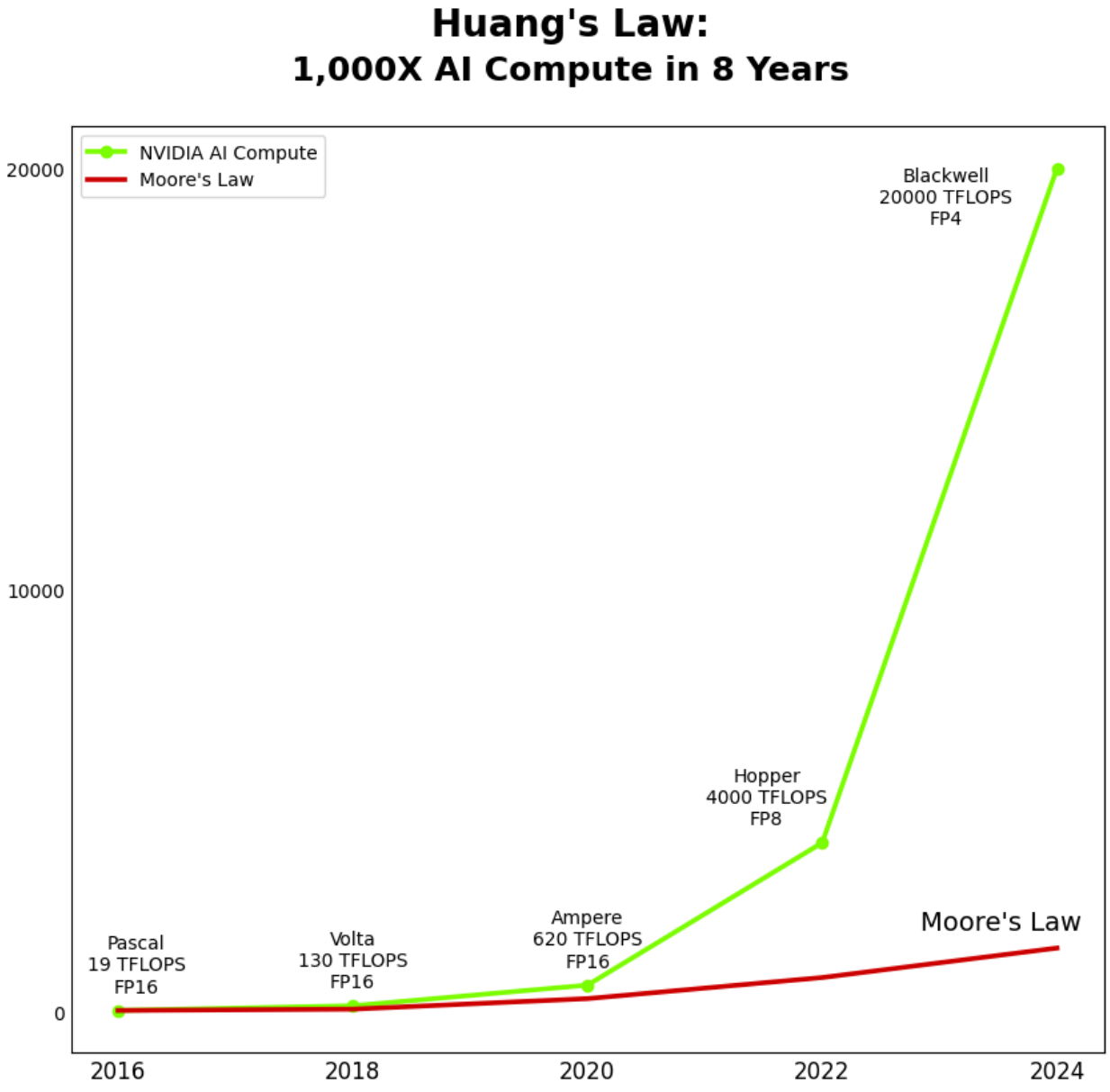
Moore's Law Never Dies ... but it might need an upgrade

With each new technological advancement, the cost per performance metric of that technology tends to decline over time. This phenomenon can be observed across various fields of technology. For semiconductors, we have the famous Moore's Law, which states that the number of transistors on a microchip doubles approximately every two years, resulting in a proportional increase in performance and a decrease in cost per transistor. Consequently, the cost per transistor in silicon chips, such as CPUs and GPUs, has declined exponentially.

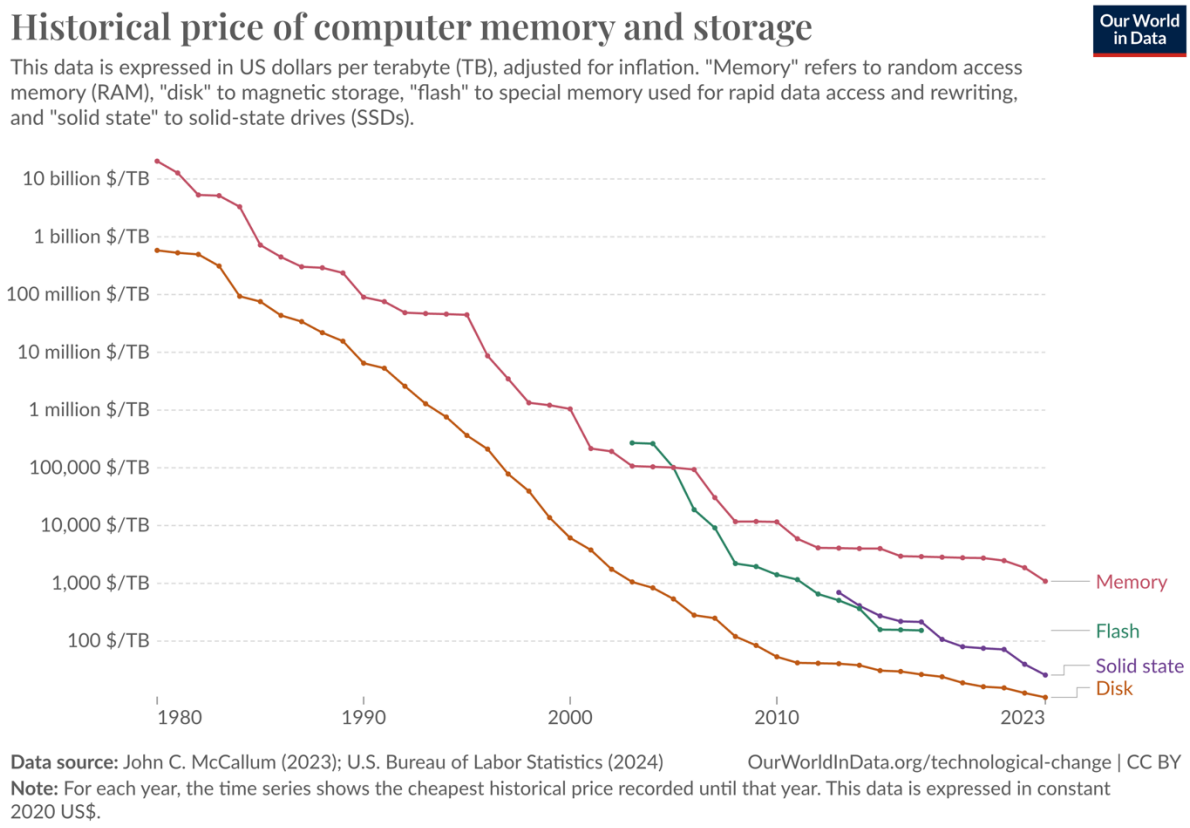


Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.

In fact, Moore's Law might need an upgrade. In terms of AI compute power, the curve for GPUs from Nvidia is currently far exceeding Moore's Law. This acceleration is being called "Huang's Law." Jensen Huang showed the following graphic in at the COMPUTEX 2024 event in Taipei on June 2 (I added the title):



To take some related examples: In 1980, the cost of 1 megabyte (MB) of DRAM was about $411, whereas by 2020, it had dropped to about $0.004 per MB. Similarly, the cost per gigabyte (GB) of storage has seen a steep decline. In 1980, the cost of 1 GB of storage was approximately $437,500, whereas by 2020, it had decreased to less than $0.02 per GB.



The cost per unit of Internet bandwidth has also declined significantly. In 2000, the cost of 1 Mbps of Internet bandwidth was around $1,200 per month, while by 2020, it had dropped to about $0.50 per month. This trend is not limited to these examples but is generally true for other aspects of computing, such as data transfer rates and power efficiency.

This pattern holds true for all types of technology, where the cost per unit of productivity or performance declines rapidly over time. For software, the cost of deployment and scaling has also decreased significantly due to advances in cloud computing and open-source platforms.

Now, consider the likely path of the GPU-based platforms for training AI models and inference:

The cost of each generation of semiconductors drops by approximately 30% per year, so that five years later, today's most advanced semiconductors cost significantly less. In a recent interview, Jensen Huang suggested that Nvidia's Blackwell GPUs will cost between $30,000 and $40,000 each when they launch later this year. All of what we see in AI today was trained on the current and older generations of Nvidia GPUs, and these models are already impressive. Blackwell GPUs are better than the current generation by a factor of approximately 5x and will enable even more powerful training.

These Blackwell GPUs will likely only cost around $5,000 five years from now (maybe substantially less due to competing GPU's; sorry Nvidia investors), with improvements in energy consumption likely included. While the most advanced work will be done on the newest GPUs available five years from now, which will likely be many times more powerful than the Blackwell, infrastructure based on Blackwell and its competitors will still be extremely powerful and will be available at far lower costs than today.

Imagine what you will find on Hugging Face in 5 years.

Cheer on the Bubble

It is very conceivable that we are in an AI infrastructure bubble. Today's massive investment in Nvidia's $35k GPUs might never realize the expected return for investors. This phase of investment could look something like the initial investment in long haul fiber optic networks in the late 1990's.

However, remember that all that fiber-based Internet backbone investment eventually led to the emergence of massive businesses that were not even conceived of in those early days. Investing in Cisco in early 2000 would have been painful. Investing in Worldcom even worse. But, if you had invested in Amazon, or in Google or Facebook a few years later, you would be more than happy with the results.

While a lot of the AI driven data center investment occurring today might not yield big returns for the companies that made them, I hope they continue. Those investments are a necessary phase to get to the truly revolutionary advancements.

Today's AI infrastructure investment is driving the R&D to keep advancing ever more powerful GPUs. As it always does, the cost per computational unit will rapidly decline along with increasing energy efficiency. Combined with the commoditization of AI models, these trends portend massive advancements in the capabilities of multi-modal AI-based systems. Imagine what can be done with 1,000 (maybe 1 million) different specialized AI models, all integrated into AI superstructures designed to do something extremely useful. Or, indeed, to do several useful things at once.

I might be off the mark to some degree. Perhaps by a lot. Yet, it seems to me that, at least in a general sense, this is where we are heading. While I have my doubts that there will be much return on spending billions on Nvidia GPU's today, I feel confident in the future returns from businesses – that we cannot yet envision – created by building AI superstructures from a readily available cornucopia of commoditized AI models.

I could be in the AI hype bubble. But I am willing to put my faith in, and money on, organic human ingenuity and creativity to make synthetic intelligence into something pervasive and

powerful. Eventually even just taken for granted, like we do now when we flip a switch to turn on a light.

What do you think? Is AI running out of steam? Is it truly an industrial revolution, or is it just a “hype train”?

More importantly, do you believe the future of AI lies in big, monolithic LLM models trained on ever larger corpora of data, or in plug-and-play commoditized models that can be configured and integrated into AI superstructures?