SYNCICAP ASSET MANAGEMENT

MARKET INSIGHTS

26 JULY 2024

A glimpse into the world of AI

Yixin GUO
Equity Analyst
SYNCICAP AM

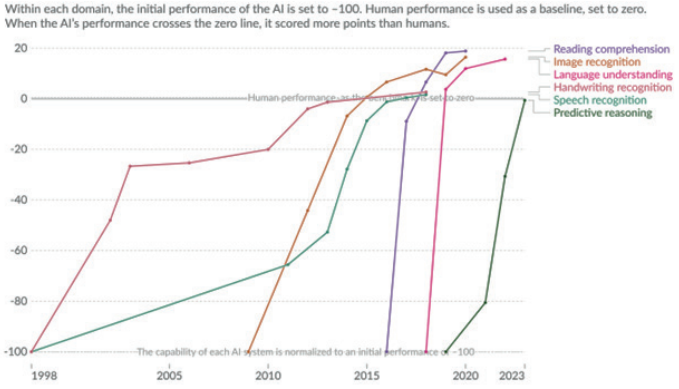
The trillion-dollar question to answer in the middle of the year 2024 – Is AI a bubble? Where are we in this AI-driven tech cycle? In this article we will discuss how AI will fundamentally reshape our work and life, why we are still early in this multi-decade technological movement, and why Asia is key in this overall supply chain.

First, we will briefly recap key AI development milestones since ChatGPT's launch in Nov 2022. Second, we will attempt to predict some future trends of the industry. Last, we will discuss our top AI picks for Asian markets.

SCALING LAW AND HUANG'S LAW – MODEL CAPABILITIES TO SCALE UP WHILE HARDWARE COST TO SCALE DOWN

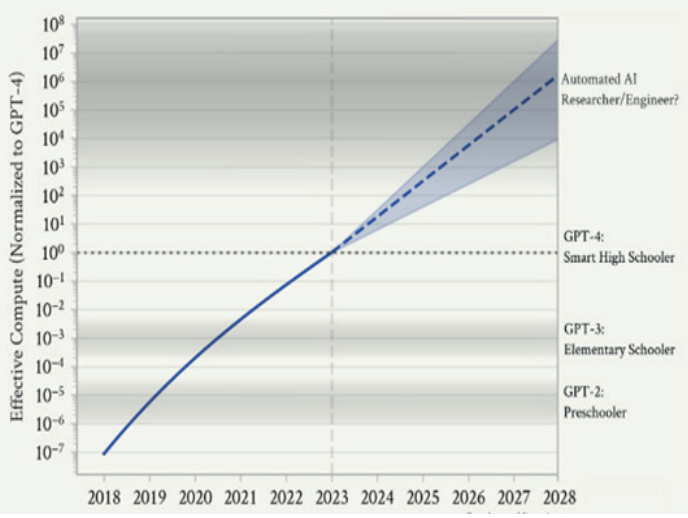
To begin with, we have advanced rapidly in our pursuit of AGI (Artificial General Intelligence). In 2019, the best model that we had is GPT-2, which can barely produce a few coherent sentences. Fast forward to 2024, our current best model, GPT-4, can already produce lengthy code snippets and reason through complex math problems. The magic behind the stellar leap in model performance is **scaling law,** which states that if we use several orders of magnitude more of data and computing resources to train our model, we can expect significant improvement in their capabilities. Ex-OpenAI Researcher Aschenbrenner predicts models will be able to do the work of AI researcher/engineer by 2027. The pursuit of AGI motivated cash-rich tech giants like Microsoft, Google and Meta to build unprecedentedly large training clusters, as they race against each other to approach the holy grail of building AGI. At the core of these datacenters is Nvidia's GPU, as they offer best-in-class performance and energy efficiency.

Test scores of AI systems on various capabilities relative to human performance



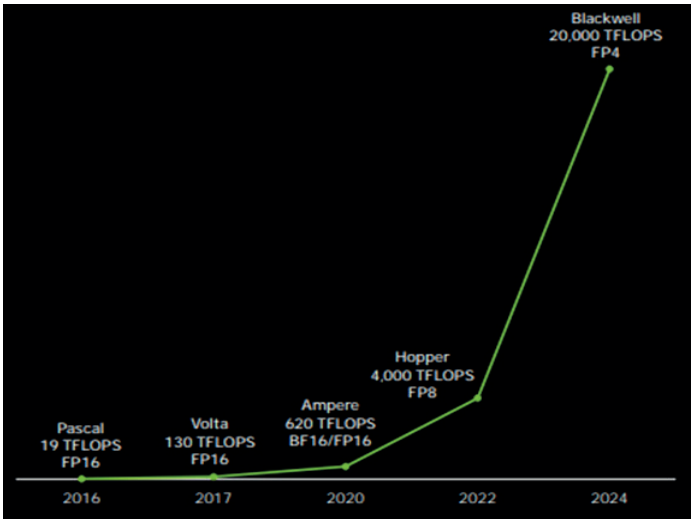
Sources: Our World in Data, end-2023

Base Scaleup of Effective Compute



Source: Leopold Aschenbrenner's blog, 2024

1000X AI Compute in 8 years

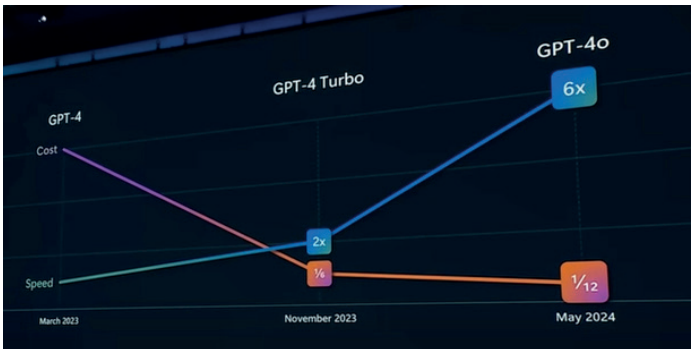


Source: Nvidia GTC, 2024

There are mainly 2 phases in deploying AI systems. First, **training** is needed to obtain a model that can be used for different purposes. Then, **inference** is conducted every time the trained model is accessed. For example, each time when we ask ChatGPT, the model is carrying out one inference task.

On the **training** side, total training cost for the latest generation model continues to surge, as scaling law commands that models will grow rapidly in parameter count and data used for training, which require much more compute resources (GPU, networking etc) to train the next-generation model. For example, to train GPT-4, the current SOTA (state of the art) model, we need 8000 Nvidia's Hopper GPU. However, to train GPT-5, roughly 100,000 Hopper GPUs are needed, representing a 10x increase in hardware cost. While current infrastructure is mostly limited to adopting Nvidia's solution, in the future as more GPU or ASIC (Application Specific Integrated Circuit) vendors compete against Nvidia, there is more room for hardware cost improvement.

On the **inference** side, the cost is declining quickly. The cost of inference for GPT-4 today is just one-twelve of the cost one year ago, which is driven by both enhanced hardware performance and better software optimizations. **Huang's law,** named after Jensen Huang, exerts that the computing power of each GPU chip has doubled every year, which means it has increased by 1000-fold within a decade, with the critical driver being using lower-precision numbers to do calculations. Thus, software techniques like quantization (representing model parameters with lower-precision figures) can leverage on hardware gains to reduce inference cost. Thus, while the cost of training the latest model is steadily on rise, the inference cost will gradually decrease, which paves way for wider LLM adoption.



Source: Microsoft Build, 2024

WE ARE LESS WORRIED ABOUT DATA SUFFICIENCY

As models rapidly grow in size, more people begin to worry if we have sufficient data to train future models. Currently, we have effectively utilized all public text data for training, which has been the single biggest data source. However, there are still many methods that we can **extract more data** to support the needs for scaling law. Enterprises keep their **proprietary data** within their data lakes, so model companies like OpenAI are actively negotiating with private firms to make use of their data. On top of that, we can use legacy models (eg GPT-3.5) to produce data to train future models, known as **synthetic data.** This is analogous to how AlphaZero, the AI model that reached superhuman level at Go, has iteratively becomes better at the game by playing against itself. To mimic humans that can learn complex concepts with just a handful of examples, neuromorphic techniques like **experience replay** are also employed, which allows the model to develop more insights using the same set of data. Thus, while data sufficiency continues to be a controversial topic, we are less worried about the problem, at least within the next 2 to 3 years.

ABUNDANT GEN AI APPLICATIONS RESHAPE OUR WORK AND LIFE

Currently, generative AI is already reshaping our working and living habits. Internet giants are using DLRM (Deep Learning Recommendation Models) to suggest more relevant contents to users, such as ads or social posts. Companies have been using LLMs to assist on tedious tasks like translation or text summarization. Programmers are using coding copilots to accelerate application development, as AI helps catch bugs and suggest improvements. In the future, better AI performance coupled with lower inference cost will undoubtedly unlock new markets that we have not conceived of, like how the proliferation of Internet leads to e-commerce platforms. Generative AI will also be one of the key drivers for economic growth, as Goldman Sachs estimates that the technology will account for a 0.4pp increase in GDP growth in the U.S., 0.2-0.4pp in other developed markets, and 0.1-0.2pp in advanced emerging markets over the next decade.

AI DEVELOPS MULTIMODAL UNDERSTANDING AND GENERATION

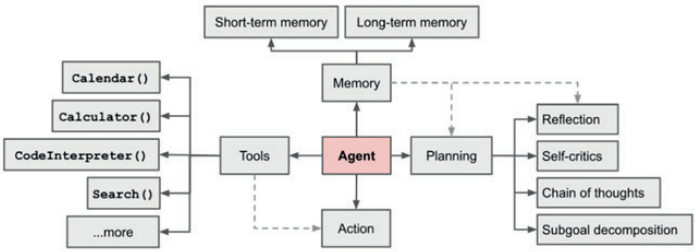
How will AI continue to evolve? The first step will be to **bring more sensory organs to AI.** We are not only training the models on texts, but also on various forms of data, such as images, audio, and video, to make the model smarter. Like how Andrew Parker, a zoologist has proposed, during the Cambrian explosion, the emergence of vision was crucial for early animals to not only find food and avoid predators, but also to evolve and improve. Similarly, allowing AI to see data beyond plain text should drive further breakthroughs. The Swiss psychologist Jean Piaget published theory in 1952, believing children develop cognitive capabilities through sensory experiences such as sight, hearing, touch, taste, and smell, and through interactions with the physical world. Although today AI systems are still limited in sensory perception capabilities, the development of humanoid robots powered by LLMs will likely provide increased and enriched interactions between the models and the physical world, making us excited about the long-term intelligence potential of AI.

AI DEVELOPS SYSTEM-2 STEP-BY-STEP REASONING CAPABILITIES

We have been also exploring how models can think more like humans. For example, researchers incorporate methods such as "chain of thought prompting", to enable the models to think step-by-step when solving a complex problem. This method is analogous to **System-2** thinking, which refers to slow, deliberate, cognition process, in contrast to **System-1** thinking, which is fast, automatic, and intuitive thinking (more analogous to how LLMs function today). Researchers have found that when using chain of thought prompting, model performance improves on arithmetic reasoning, symbolic reasoning, and commonsense reasoning.

AI AGENT AS PROMISING VEHICLE TO PURSUE AGI

The lower cost of computing, advancements in multi-modal understanding and generation, more efficient model architectures, and enhanced reasoning and planning capabilities will all contribute to the development of advanced **AI agents.** AI agent is considered as a promising vehicle to pursue human-level artificial intelligence. Open AI's Lilian Weng describes a LLM powered AI agent as a system having LLM as brain, having short-term memory and long-term memory, able to do planning with multiple steps, and able to use tools such as API calls. Some early proof-of-concept demo, such as Devin, shows the potential of AI agent, which is designed to individually solves complex coding tasks. Researchers are also developing multi-agent systems, as they discovered that an agent specializing on a narrow domain of task outperforms generalist agents, analogous to the "division of labor" view first brought up by Adam Smith. In the future, we believe we will envision the harmonious co-existence between multi-agent systems and humans.



Source: Lilian Weng's blog 2023

AI DRIVES TRANSFORMATIONAL CHANGE IN ROBOTICS

AI also drives transformational change in robotics, which will unlock unlimited resources and productivity gains, creating a seismic shift in the global economy. For example, researchers have been giving robots the capabilities of vision, planning, and the capabilities of interaction with humans, enabled by LLMs. For instance, a team of Google developed a project in which the robot can act by reasoning through current states of the environment and robot's ability. When given the prompt "I spilled my coke, can you help?" the robot evaluates what are the available tools in the environment that can be useful to help and what actions it can possibly take, and can pick up a sponge to help, instead of something irrelevant.

A HETEROGENEOUS WORLD WHERE LARGE MODELS AND SMALL MODELS CO-EXIST

Models need to be commercialized and put into products. People are making specialized **small language models** tailored for specific tasks or workflows. For example, Apple uses a ~3B parameter on-device model trained/specialized for specific tasks like summarization, proofreading, or generating mail replies. Techniques to make the large models into smaller models such as distillation or pruning have greatly evolved. We see the future as a heterogeneous world where researchers continue to develop more competent cloud-based models with more emergent behaviors, while smaller edge-based models with better security and lower latency are deployed on our PC and mobile devices.

WE ARE OPTIMISTIC ON GENERATIVE AI

Many researchers believe digital neural networks are conceptually analogous to biological brains. Parameters of models are analogous to brains' synapses, or the connection between neurons. A human brain has 100 trillion synapses. The more the synapses, the higher the intelligence. Today, GPT-4 is believed to have 1.8 trn parameters, scaling to 100 trn is the direction where we want to test the limit of scaling law.

To conclude, while skepticism continues to revolve around the sustainability of scaling law and the commercialization timeline of generative AI, we are still more optimistic on the overall trend. Hardware cost will continue to decline under Huang's law, enabling wider adoption of generative AI. On top of that, software breakthroughs like training with more modalities of data or allowing the model to self-train itself unlock more capabilities. The advancement of physical AI might be the last missing piece in realizing AGI, as robots help seamlessly integrate into our society to solve more complicated tasks.

TAIWAN, KOREA AND JAPAN ARE AT THE HEART OF THE SUPPLY CHAIN

How to participate in this AI journey? Many investors might think they can only invest in US companies to benefit from this, however, there are many Asian companies enabling the great journey of AI development. **Taiwan, Korea** and **Japan** at the heart of the supply chain. Let's dive into several examples.

TSMC (2330-TW) holds a de facto monopoly position in manufacturing the Nvidia GPU chip, the custom silicon that the US cloud providers are making, and the networking chips. In addition, it enables a broader AI theme by manufacturing the chip in the mobile phones and PCs, where AI applications will see their manifestation.

SK Hynix (000660-KR) is the biggest high bandwidth memory (HBM) chips provider in the world, a specialized memory chips to deal with the high bandwidth requirements of training and inference when we are running the AI models today.

Disco (6146-JP) is the world leader in grinders and dicers used in manufacturing semiconductor chips. Disco's machines play critical roles in manufacturing these highly sophisticated HBM, which requires grinding the chips to ultra-thin level. Only machines that exert high precision and high flatness are able to do that. Disco with its 50 years of history in precision semiconductor equipment is a dominant player enabling the manufacturing of HBM.

SYNCICAP ASSET MANAGEMENT

Syncicap AM is a joint-venture between Ofi Invest group (66%) and Degroof Petercam Asset Management (34%). It was licensed on 4 October 2021 by the Hong Kong Securities and Futures Commission.
Syncicap AM specializes in emerging markets and offers a beachhead into Asia from Hong Kong. It also manages a range of emerging markets funds offered to European investors by Ofi Invest Asset Management.

The figures cited deal with past years. Past performances are not a reliable indicator of future performances.
Investing in financial markets involves risks, including risk of capital loss. Source of indexes cited: www.bloomberg.com