WHAT GOES INTO THE MAKING OF A SENTENCE ON CHATGPT?

> **"THE SEMICONDUCTOR BUSINESS IS LIKE A TREADMILL THAT SPEEDS UP ALL THE TIME. IF YOU CAN'T KEEP UP, YOU FALL OFF."**
>
> MORRIS CHANG, FORMER CEO AND FOUNDER OF TSMC

Semiconductor Manufacturing: TSMC

The Taiwan Semiconductor Manufacturing Company (TSMC) makes cutting-edge microchips. This is a highly complex and error-prone process that only a few companies in the world have perfected. TSMC is the only company that successfully uses extreme ultraviolet (EUV) technology to print billions of transistors on coin-sized silicon wafers. There is only one company in the world that makes EUV machines: the Dutch company ASML. One machine comes with a price tag of $200 million.

Graphics Processing Unit Manufacturing: Nvidia

Nvidia designs the GPUs that provided the computational power for the generative AI boom. Nvidia is a "fabless" company – it doesn't own any manufacturing plants. All designs are manufactured by TSMC. A server comprising eight of Nvidia's cutting-edge H100 GPUs costs approximately $400,000. Following high demand, there is a massive shortage of these GPUs and Nvidia is the only supplier.

Cloud Service Provider: Azure

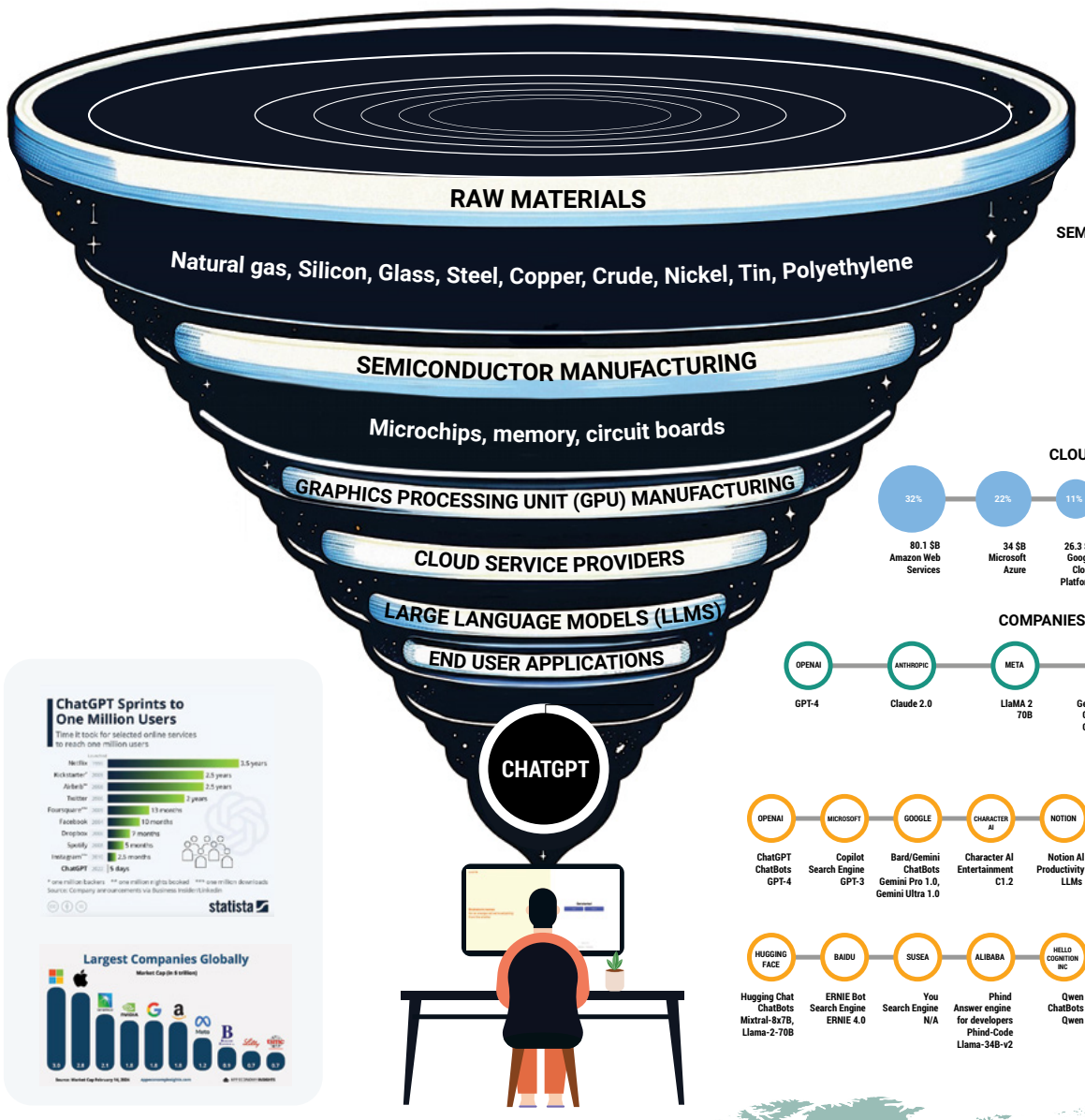
GPUs are very expensive and difficult to acquire. Instead of buying GPUs and assembling data centres, many generative AI companies opt to rent computational power from Cloud Service Providers. OpenAI has a partnership with Microsoft Azure: GPT-4 was trained on more than 10,000 Nvidia GPUs.

Large Language Model: OpenAI's GPT-4

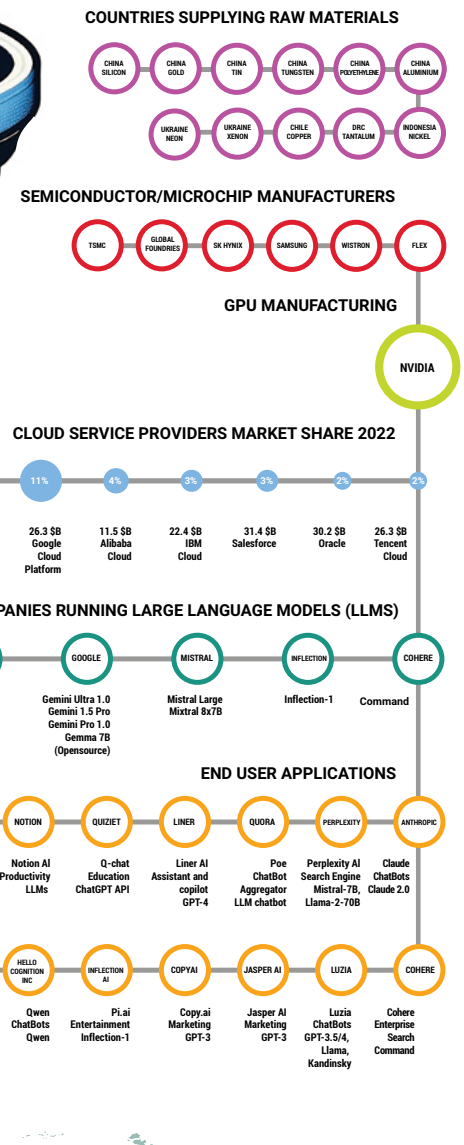
An LLM is an AI that uses neural networks trained on extensive text data to predict word sequences. It learns patterns during training and applies this knowledge during inference to perform language tasks like text generation and comprehension. GPT-4 was trained over 90 days which cost OpenAI more than $100 million.

Application: ChatGPT

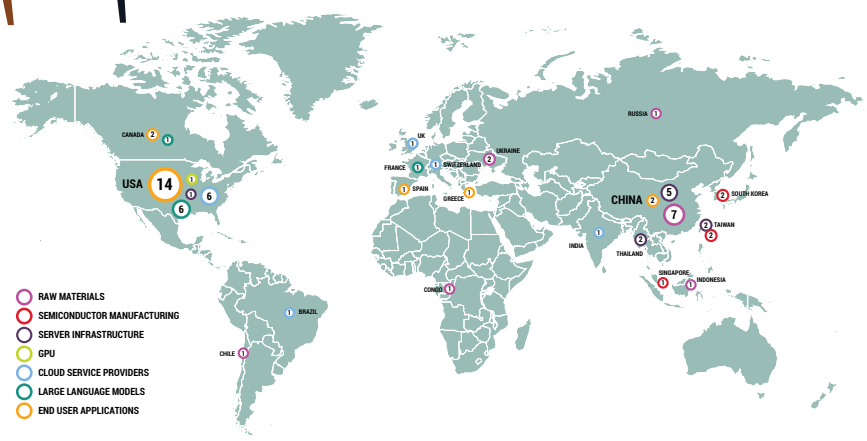
Generating a sentence on ChatGPT is only possible because of a complex and international supply chain of know-how and technology. The ease at which text appears in the chat box could easily disguise the fact that this is the biggest computational undertaking in human history of humanity and highly capital-intensive.



THE SUPPLY CHAIN BOTTLENECK

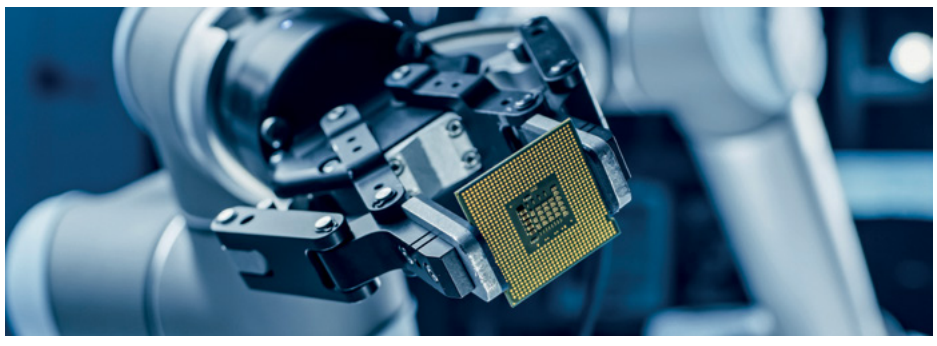


Generative AI is poised to make a fundamental difference to the functioning of economy and society. While ethical implications and consequences for the nature of work are at the forefront of debates about its impact, the technological and financial dimensions of the rise of large language models like ChatGPT have mostly remained unexamined. Dr Nils Peters' research, with Yuanling Liao, explores the infrastructure that lies behind the biggest computational undertaking in human history, revealing the power dynamics at play.



GEOGRAPHY OF SUPPLY CHAIN POWER

The generative AI supply chain is global, with companies in the US sitting at the top of the value chain. This includes LLM companies like OpenAI and "fabless" chip makers like Nvidia. Semiconductor manufacturing is concentrated in East Asia. Microchips from Taiwan provide almost 40 per cent of the world's new computational power every year, and two Korean companies produce almost half of the world's memory chips (Chris Miller, Chip Wars).



SCAN TO HEAR MORE FROM NILS ABOUT HIS RESEARCH

> **"WHEN TECHNOLOGY MOVES THIS FAST, IF YOU'RE NOT REINVENTING YOURSELF, YOU'RE JUST SLOWLY DYING. YOU'RE SLOWLY DYING, UNFORTUNATELY, AT THE RATE OF MOORE'S LAW, WHICH IS THE FASTEST OF ANY RATE THAT WE KNOW."**
>
> JENSEN HUANG, CEO AND CO-FOUNDER OF NVIDIA

FUN FACTS

- Many Large Language Models (LLMs) are trained on a dataset called the "Common Crawl", containing 3.35 billion web pages or 454 tebibyte of uncompressed content.
- Training runs for LLMs are highly time-and capital-intensive. Meta's LLaMA model released in February 2023, for instance, used over 2,000 Nvidia GPUs on 1.4 trillion tokens (750 words is about 1,000 tokens). The training run took about 21 days. The estimated cost is over $2.4 million.
- Chip designs have become so small that companies like ASML have had to invent printing techniques at the edge of known physics. They use extreme ultraviolet (EUV) with a wavelength of 13.5 nanometres, the size of five DNA strands laid side by side.