THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

RESEARCH

FOR THE WORLD

How secure is ChatGPT's supply chain?

Published September 2024

Dr Nils Peters,
LSE Fellow in Economic Sociology, Department of Sociology, LSE.

Generative AI systems like ChatGPT are becoming increasingly embedded in our lives, but how secure is the infrastructure driving the AI age? **Nils Peters** has been mapping the Large Language Model supply chain, exposing the power dynamics at play in this growing market.

The extraordinary rise of **Large Language Models** (LLMs) like ChatGPT or Microsoft's Copilot has raised ethical questions around the impacts their use might have on society. Less focus, however, has been placed on the material infrastructures that enable these systems to function. What do we really know about the LLM supply chain and, given the huge financial and power costs of running these systems, how sustainable is machine-learning in the long term?

Questions like these are behind new research by Dr Nils Peters, LSE Fellow in Economic Sociology. His work, recently showcased at **LSE Festival 2024**, explores **the material infrastructure that enables these models to work**. Through mapping what goes into the creation of these AI applications, Dr Peters identifies the power dynamics at play, as companies around the world compete to create ever-more advanced AI systems.

"When we look at the industry, while we can see the user side of things with models like ChatGPT, a lot less is known about the steps that come before – the making of the 'behind the scenes' – and so this really intrigued me," he says.

"We started by looking at the computing power and then tracing that further and further back. If you dig a bit deeper into each company, it becomes this very complex web of connections."

While the majority of us will only interact with LLMs through user-friendly applications like ChatGPT, there are many stages, managed by different companies around the world, that must happen before this end point is reached. Dr Peters's research highlights how easily issues like shortages of a particular material, or production issues with a particular company, could impact our ability to access

LLMs – important information for anyone tasked with mitigating any potential disruptions to AI systems.

"I think that's something we realised during the pandemic, when there were shortages of things like semiconductors which impacted car manufacturing, or fertiliser, which had an impact on products used in fridges and cooling operations," he says.

"Supply-chain issues have become much more present in the way we think about the economy and so, when looking at the AI industry, the logical next step would be to have a more comprehensive understanding of what the supply chains are so we can anticipate issues like bottlenecks and shortages that are so hugely disruptive."

> If you dig a bit deeper into each company, it becomes this very complex web of connections.

What is generative AI and how much does it cost?

LLMs represent a remarkable advance in AI – capable of creating original content like text, imagery and video, and adapting as they learn. Trained on the "**Common Crawl**" – a collection of 4.5 billion webpages – they are capable of offering increasingly sophisticated content on an enormous number of topics. But while models like ChatGPT currently offer free versions alongside subscriptions, they are a costly enterprise to run, requiring cutting-edge technology and immense amounts of power.

At their core, LLMs are reliant on large numbers of powerful graphics processing units (GPUs), with companies like Meta and OpenAI estimated to train their models on anything between 25,000 and 50,000 GPUs. The power required to run these systems is vast – data centres based in Ireland now consume **more energy than all its residential homes combined** – and then there is the cost of training, with a **single training run for GPT-3 estimated to cost as much as $4.6 million**. With a **market value predicted to rise to $1.3 trillion by 2032**, however, one can understand why companies like Meta continue to push the boundaries of their technology.

"When there are financial and material constraints like these, the capacity to exercise power becomes paramount. Given the increasing use of these models, it is vital that we understand how the industry operates and where this power currently lies," says Dr Peters.

Using a process of mapping the flow of goods and money, analysis of company databases and trade press, and data visualisation, Dr Peters identifies three "power dimensions" he believes policymakers should focus on: supply-chain constraints, geopolitical tensions and financing.

> Just a few companies in the world can make technology at that level. That's deeply intriguing, but also concerning.

Supply-chain constraints

As we become more reliant on generative AI, all parts of the supply chain underpinning these systems become increasingly important globally. There are six key stages in the LLM supply chain, says Dr Peters: the supply of raw materials; the manufacture of semiconductors/microchips; the manufacture of GPUs; the provision of cloud services; building LLMs; and the creation of end-user apps.

Disruption at any of these stages would cause issues; however, Dr Peters finds that not all parts of the chain are equal, with many companies providing end-user apps and far fewer operating at the technical and manufacturing stages. Through data visualisation (see below), Dr Peters shows clearly which parts of the chain are particularly vulnerable, and which actors in the chain hold the most power.

"When we trace the process back, we start to see how difficult and constrained the supply chain becomes. At one end, many companies provide end-user applications, but that narrows quickly. Globally, there are just a handful of companies providing the raw computing power, and when we look at GPU manufacturing, this narrows to just one company – **NVIDIA**," he explains.

Who holds the power in the AI era?

"From the NVIDIA point down, we see an extremely constrained supply chain, with just a few companies in the world that can make technology at that level. That's deeply intriguing, but also concerning.

"There are manufacturing capacities that make operations difficult to scale up, so that is a real constraint, and of course shortages in any of the materials used to make the chips would also create delay. But our findings also raise questions of power, and the potential of geopolitical disruption."

Being the one company responsible for the world supply of high-end GPUs gives US-based NVIDIA a huge amount of power, while also making GPU production the most vulnerable part of the LLM supply chain. Matters are made more complex, however, because while NVIDIA is only responsible for GPU design, the manufacture of these chips takes place in Taiwan, outsourced to another company, **TSMC**.

"This global element also adds a huge geopolitical dimension to LLM production. There is the conflict between China and Taiwan. Then there's issues of foreign policy – for example, US policy has recently changed to try to limit exports of these types of chips to China. Any kind of conflict of course would also not just impact the production and supply of GPUs, but have a massive impact on world GDP. And then Taiwan is also prone to earthquakes – so that adds a different element of risk again," Dr Peters says.

> [Focusing] on the material and financial processes is a very good way of thinking of alternative futures, or identifying how we can possibly intervene to change current power dynamics.

Follow the money

Production issues and global politics aside, the vast sums required to finance the development, construction, training and operation of LLMs also has implications, not just around who is able to work in this field, but also in setting the terms around how they must operate. Dr Peters points to finance as the third "power dimension" policymakers should be concerned with.

"Just like with previous technological revolutions, it becomes very obvious very quickly that a lot of the upsides of this new technology also comes with big downsides. The power that a handful of companies now have is one of these. The flow of finance behind these LLMs is another," Dr Peters says.

While companies like Meta or Google are big enough to finance this from their existing business, newcomers like **OpenAI** or **Anthropic** are reliant on sourcing funding from external sources. The enormous sums needed to not just develop, but also maintain, LLMs places these companies in the position where their need to raise the capital required to operate means they are reliant on then meeting the requirements their investors place on them.

Because of the huge costs of developing, manufacturing and running these models, the sector is vulnerable to the pressures required when guided by commercial interests and the need to turn a profit. Understanding the flow of finances behind the work is vital if these vulnerabilities are to be countered, argues Dr Peters.

He cites the case of OpenAI as an example of how financial considerations can overpower altruistic ambitions. "OpenAI was founded as a non-profit, and the idea was really to build AI in a way that is safe and ethical. In order to develop ChatGPT, however, they needed to attract investment, and so in 2019 they established a for-profit arm. Today, OpenAI is closed source, meaning that the company closely guards access to its innovations," he says.

Given the extraordinarily high cost of developing and running generative AI, this closed-source, restrictive model has become common. "Not only does this restrict the use of its AI to those who can pay – which is the opposite of their founding ambitions – but you have to wonder whether there's enough sort of care and deliberation in releasing these things when you have to meet demands from investors," Dr Peters says.

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

How will AI change the world?

With AI advancing so quickly it might appear that the path it will take is now set, but Dr Peters argues against this idea. "This project has highlighted to me that there's a possibility to challenge the idea that all of this is inevitable and moving towards this one technological future, because as soon as you start teasing apart all the components behind the apps that we use, we start to see there are lots of ways that this could have gone differently.

"This focus on the material and financial processes is a very good way of thinking of alternative futures, or identifying how we can possibly intervene to change current power dynamics," he concludes.

With generative AI not just here to stay but evolving all the time, it is imperative that policymakers focus not just on how to ensure it benefits rather than disadvantages society, but also that these increasingly important services remain operational and are not disrupted at times of crisis. Understanding the complexities and the major players in today's AI landscape is a vital first step towards answering these pressing questions. ■

Research Festival 2024's exhibit on Dr Nils Peters work, asked "What goes into the making of a sentence on ChatGPT?"

Dr Nils Peters was speaking to Jess Winterstein, Deputy Head of Media Relations at LSE.