FEATURE

DeepSeek burst on the scene – and may be bursting some bubbles.

Why building big AIs costs billions – and how Chinese startup DeepSeek dramatically changed the calculus

Author Ambuj Tewari
Professor of Statistics, University of Michigan
Published: January 29, 2025, 9:08am EST
Disclosure statement - Ambuj Tewari receives funding from NSF and NIH.
Partners - University of Michigan provides funding as a founding partner of The Conversation US.

State-of-the-art artificial intelligence systems like OpenAI's ChatGPT, Google's Gemini and Anthropic's Claude have captured the public imagination by producing fluent text in multiple languages in response to user prompts. Those companies have also captured headlines with the huge sums they've invested to build ever more powerful models.

An AI startup from China, DeepSeek, has upset expectations about how much money is needed to build the latest and greatest AIs. In the process, they've cast doubt on the billions of dollars of investment by the big AI players.

I study machine learning. DeepSeek's disruptive debut comes down not to any stunning technological breakthrough but to a time-honored practice: finding efficiencies. In a field that consumes vast computing resources, that has proved to be significant.

Where the costs are

Developing such powerful AI systems begins with building a large language model. A large language model predicts the next word given previous words. For example, if the beginning of a sentence is "The theory of relativity was discovered by Albert," a large language model might predict that the next word is "Einstein." Large language models are trained to become good at such predictions in a process called pretraining.

Pretraining requires a lot of data and computing power. The companies collect data by crawling the web and scanning books. Computing is usually powered by graphics processing units, or GPUs. Why graphics? It turns out that both computer graphics and the artificial neural networks that underlie large language models rely on the same area of mathematics known as linear algebra. Large language models internally store hundreds of billions of numbers called parameters or weights. It is these weights that are modified during pretraining.

Large language models consume huge amounts of computing resources, which in turn means lots of energy.

Pretraining is, however, not enough to yield a consumer product like ChatGPT. A pretrained large language model is usually not good at following human instructions. It might also not be aligned with human preferences. For example, it might output harmful or abusive language, both of which are present in text on the web.

The Conversation brings you analysis from scientists and medical doctors.

The pretrained model therefore usually goes through additional stages of training. One such stage is instruction tuning where the model is shown examples of human instructions and expected responses. After instruction tuning comes a stage called reinforcement learning from human feedback. In this stage, human annotators are shown multiple large language model responses to the same prompt. The annotators are then asked to point out which response they prefer.

It is easy to see how costs add up when building an AI model: hiring top-quality AI talent, building a data center with thousands of GPUs, collecting data for pretraining, and running pretraining on GPUs. Additionally, there are costs involved in data collection and computation in the instruction tuning and reinforcement learning from human feedback stages.

All included, costs for building a cutting edge AI model can soar up to US$100 million. GPU training is a significant component of the total cost.

The expenditure does not stop when the model is ready. When the model is deployed and responds to user prompts, it uses more computation known as test time or inference time compute. Test time compute also needs GPUs. In December 2024, OpenAI announced a new phenomenon they saw with their latest model o1: as test time compute increased, the model got better at logical reasoning tasks such as math olympiad and competitive coding problems.

Slimming down resource consumption

Thus, it seemed that the path to building the best AI models in the world was to invest in more computation during both training and inference. But then DeepSeek entered the fray and bucked this trend.

DeepSeek sent shockwaves through the tech financial ecosystem.

Their V-series models, culminating in the V3 model, used a series of optimizations to make training cutting edge AI models significantly more economical. Their technical report states that it took them less than $6 million dollars to train V3. They admit that this cost does not include costs of hiring the team, doing the research, trying out various ideas and data collection. But $6 million is still an impressively small figure for training a model that rivals leading AI models developed with much higher costs.

The reduction in costs was not due to a single magic bullet. It was a combination of many smart engineering choices including using fewer bits to represent model weights, innovation in the neural network architecture, and reducing communication overhead as data is passed around between GPUs.

It is interesting to note that due to U.S. export restrictions on China, the DeepSeek team did not have access to high performance GPUs like the Nvidia H100. Instead they used Nvidia H800 GPUs, which Nvidia designed to be lower performance so that they comply with U.S. export restrictions. Working with this limitation seems to have unleashed even more ingenuity from the DeepSeek team.

DeepSeek also innovated to make inference cheaper, reducing the cost of running the model. Moreover, they released a model called R1 that is comparable to OpenAI's o1 model on reasoning tasks.

They released all the model weights for V3 and R1 publicly. Anyone can download and further improve or customize their models. Furthermore, DeepSeek released their models under the permissive MIT license, which allows others to use the models for personal, academic or commercial purposes with minimal restrictions.

Resetting expectations

DeepSeek has fundamentally altered the landscape of large AI models. An open weights

model trained economically is now on par with more expensive and closed models that require paid subscription plans.

The research community and the stock market will need some time to adjust to this new reality.