

# AI Is a 5-Layer Cake

Jensen Huang (✉)

NVIDIA Corporation, Santa Clara, CA 95051, USA

© The Author(s) 2026. This article is published by Higher Education Press.

AI is one of the most powerful forces shaping the world today. It is not a clever app or a single model; it is essential infrastructure, like electricity and the internet. AI runs on real hardware, real energy and real economics. It takes raw materials and converts them into intelligence at scale. Every company will use it. Every country will build it. To understand why AI is unfolding this way, it helps to reason from first principles and look at what has fundamentally changed in computing.

For most of computing history, software was prerecorded. Humans described an algorithm. Computers executed it. Data had to be carefully structured, stored into tables and retrieved through precise queries. SQL became indispensable because it made that world workable. AI breaks that model. For the first time, we

have a computer that can understand unstructured information. It can see images, read text, hear sound and understand meaning. It can reason about context and intent. Most importantly, it generates intelligence in real time.

Every response is newly created. Every answer depends on the context you provide. This is not software retrieving stored instructions. This is software reasoning and generating intelligence on demand.

## 1 AI as Infrastructure

When you look at AI industrially, it resolves into a five-layer stack (Fig. 1).

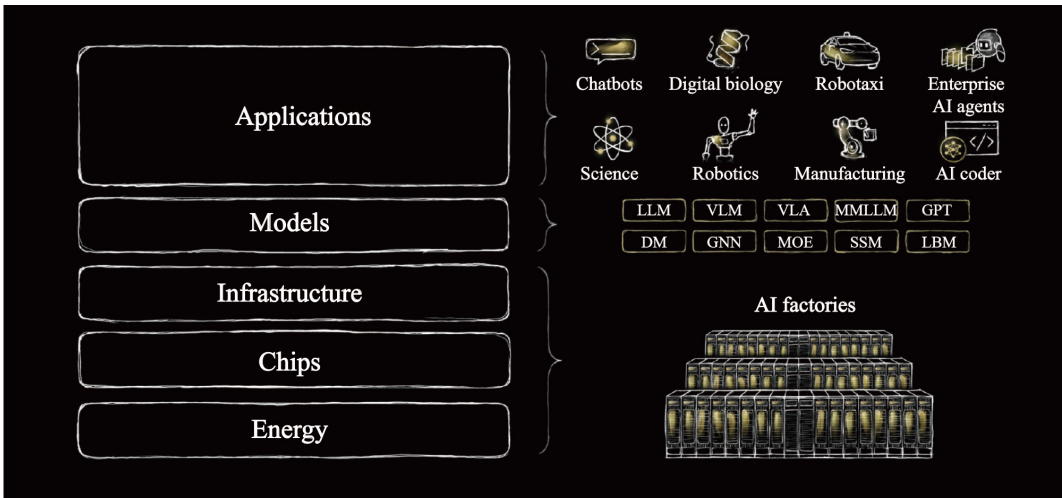


**Fig. 1** The five-layer AI stack: energy, chips, infrastructure, models, and applications.

### (1) Energy

At the foundation is energy. Intelligence generated in real time requires power generated in real time. Every token produced is the result of electrons moving, heat

being managed and energy being converted into computation. There is no abstraction layer beneath this.

### (2) Chips

Above energy are the chips. These are processors

designed to transform energy into computation efficiently at massive scale. AI workloads require enormous parallelism, high-bandwidth memory and fast interconnects. Progress at the chip layer determines how fast AI can scale and how affordable intelligence becomes.

### (3) Infrastructure

Above chips is infrastructure. This includes land, power delivery, cooling, construction, networking and the systems that orchestrate tens of thousands of processors into one (Fig. 2).



**Fig. 2** An illustration of the physical infrastructure supporting AI.

### (4) Models

Above infrastructure are the models. AI models understand many kinds of information: language, biology, chemistry, physics, finance, medicine and the physical world itself. Language models are only one category. Some of the most transformative work is happening in protein AI, chemical AI, physical simulation, robotics and autonomous systems (Fig. 3).



**Fig. 3** Computing hardware in an AI factory.

### (5) Applications

At the top are applications, where economic value is created. Drug discovery platforms. Industrial robotics. Legal copilots. Self-driving cars. A self-driving car is an AI application (Fig. 4).

### (6) Conclusions

That is the five-layer cake:

Energy → chips → infrastructure → models → applications.

Every successful application pulls on every layer beneath it, all the way down to the power plant that keeps it alive. We have only just begun this buildout. We are a few hundred billion dollars into it. Trillions of dollars of infrastructure still need to be built.

Around the world, we are seeing chip factories, computer assembly plants and AI factories being

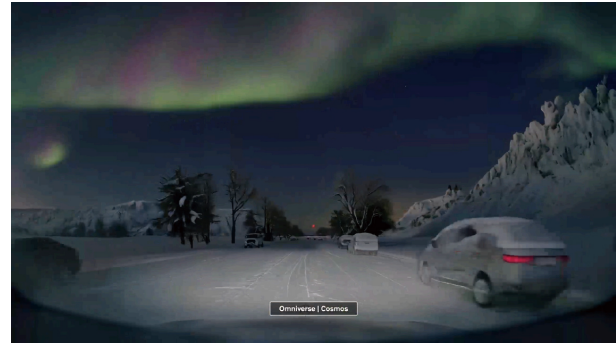


**Fig. 4** A simulated driving scene illustrating an application of physical AI.

constructed at unprecedented scale. This is becoming the largest infrastructure buildout in human history. The labor required to support this buildout is enormous. AI factories need electricians, plumbers, pipefitters, steelworkers, network technicians, installers and operators. These are skilled, well-paid jobs, and they are in short supply. You do not need a PhD in computer science to participate in this transformation. At the same time, AI is driving productivity across the knowledge economy. Consider radiology. AI now assists with reading scans, but demand for radiologists continues to grow. That is not a paradox. A radiologist's purpose is to care for patients. Reading scans is one task along the way. When AI takes on more of the routine work, radiologists can focus on judgment, communication and care. Hospitals become more productive. They serve more patients. They hire more people.

Productivity creates capacity. Capacity creates growth.

## 2 What Changed in the Last Year

Dramatically. For the first time, applications built on AI began generating real economic value.

Applications in drug discovery, logistics, customer service, software development and manufacturing are already showing strong product-market fit. These applications pull hard on every layer beneath them.

Open source models play a critical role here. Most of the world's models are free. Researchers, startups, enterprises and entire nations rely on open models to participate in advanced AI. When open models reach the frontier, they do not just change software. They activate demand across the entire stack (Fig. 5).

DeepSeek-R1 was a powerful example of this. By making a strong reasoning model widely available, it accelerated adoption at the application layer and increased demand for training, infrastructure, chips and energy beneath it.

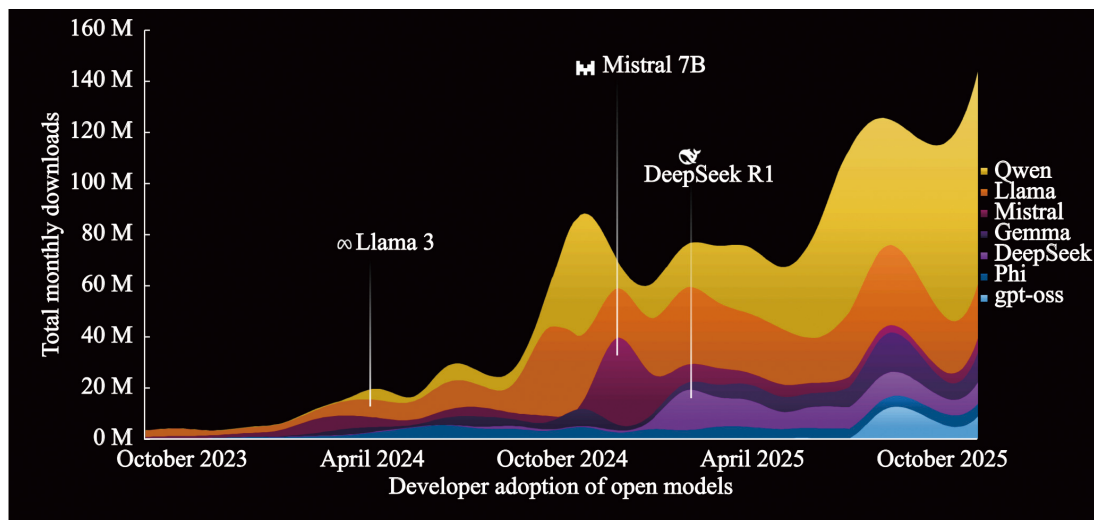


Fig. 5 Developer adoption of open models, shown as total monthly downloads.

### 3 What This Means

When you see AI as essential infrastructure, the implications become clear.

AI starts with a transformer LLM. But it's much more. It is an industrial transformation that reshapes how energy is produced and consumed, how factories are built, how work is organized and how economies grow.

AI factories are being built because intelligence is now generated in real time. Chips are being redesigned because efficiency determines how fast intelligence can scale. Energy becomes central because it sets the ceiling on how much intelligence can be produced at all. Applications accelerate because the models beneath them have crossed a threshold where they are finally useful at scale.

Every layer reinforces the others. This is why the buildout is so large. This is why it touches so many industries at once. And this is why it will not be confined to a single country or a single sector. Every company will use AI. Every nation will build it.

We are still early. Much of the infrastructure does not yet exist. Much of the workforce has not yet been trained. Much of the opportunity has not yet been realized.

But the direction is clear. AI is becoming the foundational infrastructure of the modern world. And the choices we make now, how fast we build, how broadly we participate and how responsibly we deploy it, will shape what this era becomes.

**Data availability** No datasets were generated or analyzed for this work.

**Acknowledgments** This research received no external funding. The authors declare that no generative artificial intelligence (AI) or AI-assisted technologies were used in the preparation, analysis, or writing of this manuscript. This article is sourced from the NVIDIA Blog, and reprinted in this publication with authorization.

**Competing interests** The authors declare no conflicts of interest.



Leadership Award.

**Jensen Huang** is a Member of the National Academy of Engineering of the United States. As the Founder and Chief Executive Officer of NVIDIA Corporation, he has received numerous honors and awards, including the Robert N. Noyce Award (the highest honor of the Semiconductor Industry Association, SIA), the IEEE Founders Medal, and the Dr. Morris Chang Exemplary