

Chapter 1 Technology

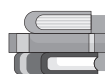




Lead-in

Before exploring the texts in this chapter, discuss the following questions, which will help you engage more deeply with the themes of technology, innovation, and society throughout the chapter.

1. To what extent do you agree with the statement that “AI is not merely a technology, but a mirror reflecting the values and biases of its creators”? How might this perspective influence its development and application in society?
2. If you could guide the future development of AI in one specific direction to address a pressing global or personal challenge, what would it be and why?
3. While AI promises significant advancements, it also raises widespread social concerns. What do you believe is the most critical ethical or practical issue we must tackle as AI becomes further integrated into everyday life?



Text Reading

Questions for Text A

1. How are China's AI firms innovating chip technology in creative and effective ways?
2. How do technology trade restrictions affect global innovation and consumers?
3. What are the possible effects of advancing chip technology from your perspective?



Text A

China's AI Firms Are Cleverly Innovating Around Chip Bans¹

Today's top artificial intelligence (AI) models rely on large numbers of

¹ Adapted from *The Economist*.

cutting-edge processors known as graphics processing units (GPUs). Most Western companies have no trouble acquiring them. Llama 3, the newest model from Meta, a social media giant, was trained on 16,000 H100 GPUs from Nvidia, an American chipmaker. Meta plans to stockpile 600,000 more before year's end. XAI, a startup backed by Elon Musk, has built a data center in Memphis powered by 100,000 H100s. And though OpenAI, the other big model maker, is tight-lipped about its GPU stash, it had its latest processors hand-delivered by Jensen Huang, Nvidia's boss, in April.

This kind of access is a distant dream for most Chinese tech firms. Since October 2022, America has blocked the sale of high-performance processors to China. And some Chinese firms have shifted their focus to making the most of limited resources. Their results are giving Western firms food for thought. Among the innovators is DeepSeek, a Chinese startup based in Hangzhou. Its model, DeepSeek-V2.5, holds its own against leading open-source models on coding challenges as well as tasks in both English and Chinese.

DeepSeek makes up for this shortfall in a number of ways. The first is that it consists of a number of different networks, each best suited to a different problem. This “mixture of experts” approach allows the model to delegate each task to the right network, improving speed and reducing processing time. Though DeepSeek has 236bn “parameters”—the virtual connections linking distinct bits of data—it uses less than a tenth at a time for each new chunk of information it processes. The model also compresses new data before it is processed. This helps it handle large inputs more efficiently.

DeepSeek is not alone in finding creative solutions to a GPU shortage. MiniCPM, an open-source model, comes in varieties with 2.4bn and 1.2bn parameters, respectively. Despite its small size, MiniCPM's performance on language-related tasks is comparable to large language models (LLMs) with between 7bn and 13bn parameters. Like DeepSeek's model, it combines a “mixture of experts” approach with input compression. Like other small models with fewer parameters, however, MiniCPM may not be terribly high-performing in areas outside its specific field of training. MiniCPM's tiny size makes it well-suited for personal devices. In August, its creators released a version of the model for mobile phones, which supports multiple languages and works with various types of data, from text and images to audio.

Similar approaches are being tried elsewhere. FlashAttention-3, an algorithm developed by researchers from Together.ai, Meta and Nvidia, speeds up the training and running of LLMs by tailoring its design for Nvidia's H100 GPUs.

JEST, another algorithm released in July by Google DeepMind, is fed smaller quantities of high-quality data for its initial training before being let loose on larger, lower-quality data sets. The company claims this approach is 13 times faster and ten times more efficient than other methods. Researchers at Microsoft, which backs OpenAI, have also released a small language model called Phi-3-mini with around 4bn parameters.



New Words

stockpile /'stɒkpail/ *v.* to keep adding to a supply of something for future use 大量储备

- An enormous volume of explosives was stockpiled inside one of the buildings.

tight-lipped /taɪt'lipt/ *adj.* unwilling to talk about something 口风紧的; 守口如瓶的; 紧闭双唇的

- He's been very tight-lipped about what happened at the meeting.

stash /stæʃ/ *n.* an amount of something that is kept in a secret place 储备; 存放物; 藏匿

- There were the stashes of booze, if nobody found them.

covet /'kʌvɪt/ *v.* to want to have something very much, especially something that belongs to someone else 觊觎; 垂涎 (尤指他人之物); 贪求; 渴望

- Gatlin covets my job, which he has been in line for twice before.

shortfall /'ʃɔ:tfɔ:l/ *n.* the difference between the amount you have and the amount you need or expect 不足之数; 缺口; 差额

- The drought caused serious shortfalls in the food supply.

delegate /'delɪɡeɪt/ *v.* to be chosen or elected by a group to speak, vote, etc. for them, especially at a meeting 授 (权); 把 (工作、权力等) 委托 (给下级)

- The authority to make financial decisions has been delegated to a special committee.

parameter /pə'ræmɪtə/ *n.* a set of fixed limits that control the way that something should be done 限制条件; 规范; 范围

- The central office sets the parameters that guide policies at the local level.

algorithm /'ælgərɪðəm/ *n.* a set of instructions that are followed in a fixed order and used for solving a mathematical problem, making a computer

program, etc. (尤指计算机使用的) 算法; 计算程序

- Music apps use algorithms to predict the probability that fans of one particular band will like another.



Useful Expressions

get/lay one's hands on something: to catch someone or find something 得到 / 找到某物

- I would read any book I could lay my hands on.

give somebody food for thought: to make someone think seriously about something 引起 (某人) 深思

- The professor's comment on how technology might reshape human relationships in the next decade gave all the students food for thought.

hold one's own (against somebody/something) (in something): to perform well enough to compete effectively, resist pressure, or avoid being outmatched 坚守立场; (使) 立于不败之地

- Despite being the youngest competitor in the debate tournament, 16-year-old Lila held her own against seasoned adult debaters and even won two rounds.

be down to somebody/something: to be someone's responsibility or decision 是……的责任; 由……决定

- It's down to me to find a suitable person for the job.

Exercises



I. Choose the best option to answer each of the following questions according to Text A.

1. What resource did Meta primarily use to train its Llama 3 model?
 - A) Central processing units.
 - B) Chips designed for mobile devices.
 - C) Servers used for cloud computing.
 - D) Graphics processing units.

2. Why are Chinese tech companies struggling to get high-performance processors?
 - A) The cost of these processors is too high.
 - B) The U.S. government restricts the sales to China.
 - C) They have a preference for older technology.
 - D) Local manufacturers don't produce such processors.
3. What is the "mixture-of-experts" approach used by DeepSeek?
 - A) It hires experts from global tech firms to design its AI models.
 - B) It combines data from different industries to train its models.
 - C) It shares GPU resources with other Chinese AI startups.
 - D) It uses multiple networks, each specialized in a different task.
4. What advantage does MiniCPM have over large language models (LLMs)?
 - A) It has far more parameters than most large LLMs.
 - B) It performs better in all AI-related tasks than LLMs.
 - C) It is suitable for personal devices like mobile phones.
 - D) It requires no GPU support to run efficiently.
5. What benefit does the "scarcity mindset" (last paragraph) provide, according to Nathan Benaich?
 - A) It leads to higher profits for companies.
 - B) It builds stronger international partnerships.
 - C) It drives the development of more efficient tech.
 - D) It speeds up the production of GPUs.

II. Discuss the following questions with your classmates.

1. Some argue that "software optimization plus alternative hardware" is a temporary fix rather than a long-term solution. Others believe it could spark a new paradigm of efficient computing that outlasts any single hardware generation. Which view do you find more convincing for the next ten years?
2. Will growing technological diversity across regions lead to fragmented digital experiences for ordinary users, such as on smartphones or social media? Why or why not?

3. Is it ever acceptable for companies to bypass regulations in the name of innovation? Why or why not?



Text B

Artificial Intelligence: DeepSeek, and You Shall Find¹

The market reaction, when it came, was brutal. On January 27th, as investors realised just how good DeepSeek's V3 and R1 models were, they wiped around a trillion dollars off the market capitalization of America's listed tech firms. Nvidia, a chipmaker and the chief shovel-seller of the artificial-intelligence (AI) gold rush, saw its value fall by \$600bn. Yet even if the Chinese model-maker's new releases rattled investors in a handful of firms, they should be a cause for optimism for the world at large. DeepSeek shows how competition and innovation will make AI cheaper and therefore more useful.

DeepSeek's models are practically as good as those made by Google and OpenAI—and have been produced at a fraction of the cost. Barred by American export controls from using cutting-edge chips, the Chinese firm undertook an efficiency drive, even reprogramming the chips it used to train the model to eke out every drop of power. The cost of building an AI model that can stand toe-to-toe with the best has plummeted. Within days, DeepSeek's chatbot was the most downloaded app on the iPhone.

The contrast with America's approach could not be starker. Sam Altman, the boss of OpenAI, has spent years telling investors—and America's new president—that vast sums of money and computing power are needed to stay at the forefront of AI. Investors have accordingly been betting that a handful of firms stand to reap vast monopoly-like rents. Yet if fast followers such as DeepSeek can eat away at that lead for a fraction of the cost, then those profits are at risk.

Nvidia became the most valuable listed company in the world thanks to a widespread belief that building the best AI required paying through the nose for its best chips (on which its profit margins are reported to exceed 90 percent). No

¹ Adapted from *The Economist*.

wonder DeepSeek's success led to a stockmarket drubbing for the chipmaker on January 27th. Others in the data-center business are also licking their wounds, from Siemens Energy (which would have built the turbines to power the build-out) to Cameco (which would have provided the uranium to fuel the reactors to turn the turbines). Had OpenAI been listed, its stock would surely have taken a tumble as well.

Yet there are far more winners than losers from the DeepSeek drama. Some of them are even within tech. Apple will be cheering that its decision not to throw billions at building AI capabilities has paid off. It can sit back and pick the best models from a newly commoditized selection. Smaller labs, including France's Mistral and the Emirati TII, will be racing to see if they can adopt the same improvements, and try to catch up with their bigger rivals.

Moreover, efficiency gains are unlikely to result in less spending on AI overall. The Jevons paradox—the observation that greater efficiency can lead to more, not less, use of an industrial input—will surely come into play. The possible applications for a language model with computing costs as cheap as DeepSeek's (\$1 per million tokens) are vastly more numerous than those for Anthropic's (\$15 per million tokens). Many uses for cheaper AI are as yet unimaginable.

Even Nvidia may not suffer too much in the long run. Although its market clout may be diminished, it will continue to sell chips in vast quantities. Reasoning models, including DeepSeek's R1 and OpenAI's o3, require much more computing power than conventional large language models to answer questions. Nvidia will be hoping to supply some of that.

However, the real winners will be consumers. For AI to transform society, it needs to be cheap, ubiquitous, and out of the control of any one country or company. DeepSeek's success suggests that such a world is imaginable. Take Britain, where Sir Keir Starmer, the prime minister, has unveiled a plan to use AI to boost productivity. If he does not need to pay most of the efficiency gains back to Microsoft in usage fees, his proposal has a better chance of success. When producers' rents vanish, they remain in the pockets of users.

Some have begun to suggest that DeepSeek's improvements don't count because they are a consequence of "distilling" American models' intelligence into its own software. Even if that were so, it remains a groundbreaking innovation. The ease with which DeepSeek found greater efficiency will spur competition. It suggests many more such gains are still to be discovered.

For two years the biggest American AI labs have vied to make ever more marginal improvements in the quality of their models, rather than models that are cheap, fast and good. DeepSeek shows there is a better way.



New Words

capitalization /kæpɪtəlaɪ'zeɪʃən/ *n.* the total value of a company's shares on a stock exchange (公司在股票市场上的) 资本总额; 股本

- After the tech company announced its breakthrough in AI technology, its capitalization surged by over \$50 billion in just one week, making it one of the top 10 most valuable firms on the stock exchange.

rattle /'rætl/ *v.* to worry someone or make someone nervous 使慌乱; 使紧张

- The creaking upstairs was starting to rattle me.

fraction /'frækʃən/ *n.* a small part of something, or a small amount 极小的部分

- They can produce it at a fraction of the cost of traditional methods.

bar /bɑː/ *v.* to prevent something or someone from doing something or going somewhere, or not to allow something 阻止; 禁止

- A lack of relevant work experience bars many recent graduates from applying for senior positions in this company.

plummet /'plʌmɪt/ *v.* to fall very quickly and suddenly 骤然下降; 暴跌

- Several large rocks plummeted down the mountain.

chatbot /'tʃætɒt/ *n.* a computer program designed to have a conversation with a human being, especially over the Internet 聊天机器人

- When customers visit the online store outside of business hours, a friendly chatbot pops up automatically to answer common questions about shipping times, return policies, and product availability.

monopoly /mə'nɒpəli/ *n.* (an organization or a group that has) complete control of something, especially an area of business, so that others have no share 垄断; 专营

- The pharmaceutical giant's monopoly on the life-saving drug has sparked public outrage, as it has raised the price beyond the reach of most patients.

margin /'mɑːdʒɪn/ *n.* the profit made on a product or service 利润; 赚头; 差额

- Our increased profits are due to improved margins and successful cost control.

drubbing /'drʌbɪŋ/ *n.* a beating or serious defeat, especially in sports (尤指体育比赛中的) 痛击, 惨败

- Norwich took a severe drubbing at the hands of Manchester United.

uranium /ju'reɪniəm/ *n.* a heavy white metal that is radioactive and is used to produce nuclear power and nuclear weapons 铀 (一种放射性化学元素)

- It has therefore been concluded that it is necessary to develop an alternate target design for such high intensities of the uranium beam.

clout /klaʊt/ *n.* the power and influence over other people or events (影响他人决定的) 力量; 权势; 影响力

- The Queen may have privilege but she has no real political clout.

ubiquitous /ju:'bɪkwɪtəs/ *adj.* found or existing everywhere 无处不在的

- The mobile phone, one of the most ubiquitous consumer-electronic appliances, is about to enter a new age.

unveil /ʌn'veɪl/ *v.* to show or introduce something new or make it known publicly for the first time (首次) 透露 (新计划); 展示 (新产品)

- The company recently unveiled a test version of its new search engine.

vanish /'vænɪʃ/ *v.* to disappear or stop existing suddenly 消失; 不存在; 灭绝

- We ran after the thief, but he had vanished into thin air.

distill /dɪ'stɪl/ *v.* to heat a liquid until it changes to a gas and then make it liquid again by cooling 蒸馏

- Some strong alcoholic drinks such as whisky are made by distilling.

groundbreaking /'graʊnd,breɪkɪŋ/ *adj.* new and likely to have an effect on how things are done in the future 创新的; 开创性的

- His latest movie is interesting, but not groundbreaking.

spur /spɜ:/ *v.* to encourage an activity or development or make it develop faster 促进; 刺激

- Rising consumer sales have the effect of spurring the economy.



Useful Expressions

eke out something: to use something slowly or carefully because you only have a small amount of it 精打细算地维持; 尽量节约使用