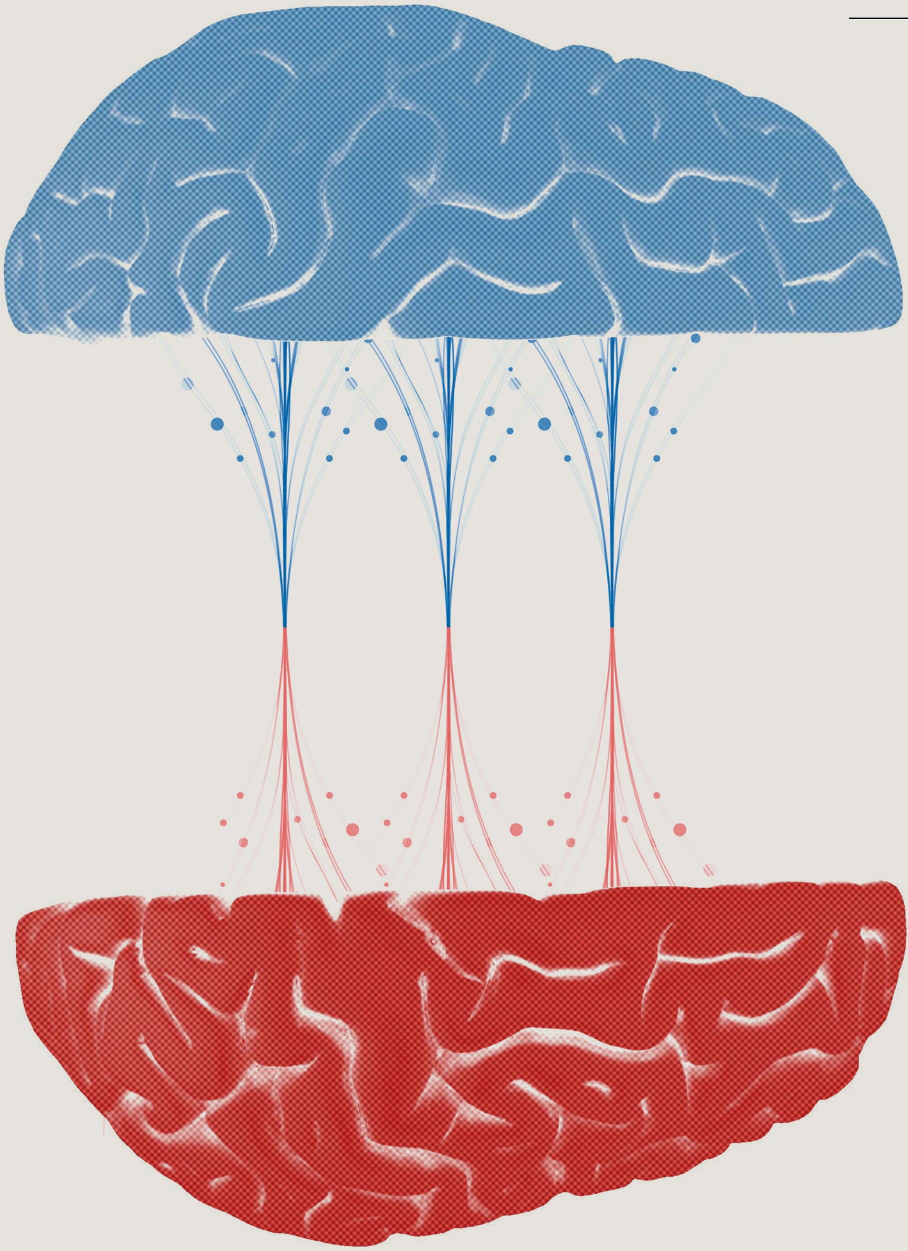
THREAT ANALYSIS

CHINA

Recorded Future®

By Insikt Group®

May 8, 2025



Measuring the US-China AI Gap

China's AI industry is advancing, and the competition is likely to be very close, but Beijing is unlikely to sustainably surpass the US's AI leadership by 2030.

China's AI models are likely currently three to six months behind US models, but there is potential for Chinese platforms to outperform the US at times and in certain sectors.

For China to succeed, it will likely need to overcome key challenges in computing, chip manufacturing, and diffusion, but the US must also work to maintain leadership.

Executive Summary

China has stated its ambition to become the world leader in artificial intelligence (AI) by 2030, a goal that encompasses not only the performance of individual AI models that often attract significant media attention but also AI innovation broadly and widespread adoption of AI for economic and geopolitical benefit. Based on an analysis of key industry pillars informing the US-China competition for AI supremacy — including government and venture capital (VC) funding, industry regulation, talent, technology diffusion, model performance, and compute capacity — Insikt Group assesses that China is unlikely to sustainably surpass the United States (US) on its desired timeline. Currently, China either trails behind or does not clearly lead the US in any of these pillars. The US-China AI competition is likely to become tighter, with China's AI industry likely being a close second to the US globally and its AI models possibly outperforming the US at times or in some sectors.[1] Chinese generative AI models likely lag behind US competitors by approximately three to six months as of this writing, based on Insikt Group analysis of publicly available Elo benchmarks, but potential new algorithmic breakthroughs along with agentic and collaborative AI systems could significantly sway the competitiveness of US or Chinese models before 2030.

China's government has sought to accelerate the development of a world-leading, globally influential AI industry since 2017, when authorities adopted a dedicated plan for achieving this goal. DeepSeek's unveiling of R1 in January 2025 was a key milestone in this endeavor. R1's capabilities reflect — and support continuing — noteworthy advances and lines of effort within China's AI ecosystem. China's AI research community is very likely benefiting from the government's supportive policy environment, government-led investment initiatives, access to an increasingly high-quality talent pool, and increasing links between academia and industry. Chinese AI companies like DeepSeek have realized performance gains by innovating and embracing open source. These companies have also become proficient at adopting techniques implemented by US peers and domestic rivals and prioritizing cost efficiency to remain competitive in domestic markets. Chinese open-source models are being adopted domestically and abroad, while inventors and organizations in China are filing more patents for generative AI applications in many key industries such as software, finance, and energy.

Despite this progress, China's AI industry faces important challenges. Total private sector investment in AI lags far behind the US, and central government funding likely trails US federal investment by a small margin. China's access to domestic AI-relevant talent is improving and the attractiveness of its domestic industry is increasing, but it is still insufficient, and the US likely retains its historical advantages. China's regulations, although pioneering, likely create a drag on innovation for teams with any interest in making their products public. Moreover, China's semiconductor industry — while achieving notable advances in spite of international export controls on specialized technology — is still unlikely to meet the rapidly growing demand for AI accelerator chips.

[1] This assessment of the competition's trajectory assumes that AI development in the US is likely able to sustain its current pace and advantages, but there are notable risks and uncertainties that the US will also have to navigate in coming years. These challenges are discussed further in the Outlook section of this report.

Maintaining leadership in the race to artificial general intelligence (AGI) is almost certainly perceived in China and the US as critical to their national security. To this end, China is almost certain to continue applying its kit of licit and illicit tools for advancing national development, to include economic espionage and foreign talent recruitment. The US and allied governments should closely monitor major developments by Chinese generative AI companies and public and private investment in research and development (R&D), in addition to monitoring indicators of AI diffusion, such as applications and patents. Western AI companies should consider proactively protecting themselves from model distillation and intellectual property (IP) theft. Western hardware manufacturers subject to export controls should continue improving end-customer due diligence processes to remain compliant and avoid selling to proxies of entities sanctioned by the US. Western government, academic, and corporate entities should adopt policies that facilitate the recruitment and retention of global AI talent.

Key Findings

- China's rapidly maturing AI ecosystem is very likely increasingly fostering collaboration between government, industry, and academia, and is supported by steady advances in semiconductor manufacturing.
- Government funding for AI-related technologies is likely on an upward trajectory in both the US and China as of early 2025, and China's overall government-led funding likely exceeds investment by US federal and state governments; however, total private-sector investment in AI companies in the US vastly outmatches private-sector investment in China.
- China's regulatory scheme likely hampers Chinese AI capabilities and extends development and deployment timelines — but only among developers aiming for public-facing products, meaning frontier advancements are unlikely to be impeded.
- The international AI talent pool likely continues to favor the US due to a continuing — though declining — immigration advantage and the quality of elite educational institutions, but the practical implications of this lead for AI competition are likely eroding.
- AI diffusion rather than innovation will very likely determine the "winner" in the competition to economically and geopolitically benefit from the technology, but whether the US or China has greater levels of diffusion is unclear, with one metric (patents) nevertheless showing China has a lead in many industries.
- According to Insikt Group's analysis of model benchmarks, Elo scores, and industry expert assessments, Chinese generative AI models likely now have a three to six-month performance gap behind US rivals, though this time lag is shortening.
- Closing the performance gap while being cost-competitive is very likely to pay off for China by driving the adoption of Chinese generative AI models domestically and abroad.
- Access to high-quality training data and IP is an increasingly contested domain where the US likely retains a competitive advantage; companies in both countries are likely leveraging user-generated content (UGC) to train generative AI models.
- Adopting open source is more prevalent among Chinese AI companies and likely enables China to diffuse its models more broadly than US proprietary models.

- US export controls have also almost certainly prompted the Chinese government to accelerate funding for its AI hardware and semiconductor industries and high-performance computing infrastructure for training and hosting AI models.
- China's semiconductor industry likely still faces a bottleneck in producing sub-7 nanometer (nm) chips, and it is almost certainly attempting to develop its own extreme ultraviolet (EUV) lithography tools using alternative techniques to advance domestic AI accelerator production.
- Although the current US presidential administration has signalled that maintaining the US's leading position in AI development is a priority, early actions to decrease public funding for science and target international students over alleged visa infractions likely risk undermining this goal.

Table of Contents

Background

The race to AGI, which is estimated by AI experts and industry leaders to be concluded within the next five to ten years, holds considerable stakes for national security and economic growth. A declining cost and general availability of human-level AI models — meaning that AI can perform most tasks previously done by humans, better than humans — will very likely disrupt the future of work while simultaneously leading to explosive economic growth, with AI estimated to contribute $15.7 trillion to the global economy by 2030. Researchers continue debating the existential risks and adversarial use of AI, ranging from facilitating chemical, biological, radiological, and nuclear (CBRN) threats to targeting democratic processes and supporting offensive cyber operations. The notion of "AI supremacy" captures the geopolitical stakes of the AGI race, with the US and China considered by many to be forerunners.[2]

On January 20, 2025, Chinese AI research company DeepSeek released R1, an open-source reasoning large language model (LLM) competing with OpenAI's then state-of-the-art o1 model. The following week, US AI hardware company Nvidia subsequently lost $593 billion in market value, marking the single largest market capitalization loss in US history. Media headlines compared DeepSeek's release of R1 to the Space Race's "Sputnik moment," referring to a rapid and unexpected shift in perceived adversary capabilities following the Soviet Union's surprise launch of the Sputnik satellite into orbit on October 4, 1957.

The US-Soviet "missile gap" theory is a likely more accurate Cold War analogy of current threat perceptions. Like missile production, the AGI race is not measured solely by innovation capacity (launching Sputnik) but also diffusion capacity, or the ability of national ecosystems to align capital, talent, and policies to rapidly convert innovations to economically productive processes (winning the Cold War). The missile gap analogy is also accurate insofar as US leaders long possessed inaccurate data on Soviet missile capabilities; it was only by 1961 that innovations in satellite imagery allowed decision-makers to obtain reliable data on this so-called "missile gap." Following the launch of the CORONA satellite project to monitor Soviet ICBM capabilities and Chinese nuclear testing, US President Lyndon Johnson admitted in 1967: "We've spent $35 billion or $40 billion on the space program. [...] We were doing things we didn't need to do. We were building things we didn't need to build. We were harboring fears we didn't need to harbor." Similarly, advances (whether real or speculative) in Chinese AI development are very likely fueling funding for AI research in the US, where AI capital expenditure is projected to reach over $320 billion in 2025, and influencing opinions on issues like AI regulation.

Maintaining an accurate threat perception of Chinese AI will almost certainly be key for major business and national security decision-makers in the US and its partners in the coming years. Despite the public perception of sudden and opaque progress via the lens of a "Sputnik moment," the development of China's AI ecosystem almost certainly bears the fruit of long-term investments, iterative innovation, and the necessity for efficiency gains in the face of hardened chip export controls.

[2] *https://archive.is/r1Exq*

China's AI Industry and Landscape

In July 2017, China's State Council issued the "New Generation Artificial Intelligence Development Plan" (AIDP; 新一代人工智能发展规划), which outlined an ambitious path to achieving "world-leading" AI theories, technologies, and applications and becoming the "world's primary AI innovation center" by 2030. To achieve these goals, the AIDP calls (in part) for the activities listed below. The AIDP also outlined key challenges China faced at the time — and to some extent continues to face — including a lack of breakthroughs in basic AI theory, algorithms, high-end chips, and other resources and outputs. Since that time, Chinese universities and businesses have become increasingly involved in AI conferences at the frontiers of AI research and have released hundreds of generative AI models. At the same time, China's semiconductor industry is steadily advancing its capabilities in spite of international restrictions.

According to the AIDP, China would seek — and in many ways has begun, as documented in this report — to:

- Support a development framework that enables industry, academia, and other sectors to pursue joint and cross-disciplinary innovation
- Seek breakthroughs in R&D, applications, and overall industry, including along the AI-relevant industrial chain
- Improve training and recruitment of talent, focusing on AI specifically as well as AI's application in other fields
- Promote AI applications and AI-related innovations across industries, enterprises, and services
- Coordinate support for AI across China's main science and technology (S&T) pillars, including key programs, megaprojects, bases, and talent plans
- Adopt a system of laws, regulations, and norms to aid the development of AI
- Disseminate news of China's AI progress to foster a society enthusiastic about, and supportive of, AI development

China's S&T innovation apparatus very likely emphasizes (though it is perceived to be deficient in) collaboration among government — including laboratories and state-owned enterprises [SOE]) — industry, and academia. Accordingly, entities from each of these sectors contribute to China's AI development enterprise. Entities from all three sectors are increasingly producing noted contributions to AI R&D and applications. Contributions to theoretical research can be observed in scholarly papers accepted to the prestigious AI-focused Conference on Neural Information Processing Systems (NeurIPS) since 2021. **Table 1** lists examples of government, industry, and academic entities from among the top fifteen institutions named as the affiliation of one or more co-authors on papers accepted by the conference (2021–2024). Note that there are no top-fifteen NeurIPS contributors from China that are SOEs; two are included in **Table 1** to indicate that some SOEs are nevertheless attempting and occasionally succeeding to contribute to frontier AI advancement.

Sector	Entity	No. of Papers (Rank)
Academia	Tsinghua University	643 (1st)
Academia	Peking University	458 (2nd)
Industry	Huawei Technologies	228 (6th)
Industry	Tencent AI Lab	197 (7th)
Government (Research Institute)[3]	Shanghai AI Laboratory	141 (11th)
Government (Research Institute)	Institute of Automation of the Chinese Academy of Sciences	118 (12th)
Government (SOE)	Intelligent Science and Technology Academy of China Aerospace Science and Industry Corporation	9 (N/A)
Government (SOE)	China Telecommunications Corporation	2 (N/A)

Table 1: *Chinese government, industry, and academic entities with the most accepted number of NeurIPS submissions (2021–2024) (Source: PaperCopilot)*

Shanghai AI Laboratory (SHLAB; 上海人工智能实验室) exemplifies how China's S&T system supports inter- and intra-sector collaboration. SHLAB was created in 2020 as an almost certainly state-affiliated "new-type scientific research institution" (新型科研机构). Described as a "large-scale comprehensive research base" (大型综合性研究基地), SHLAB offers open-source information repositories and platforms that facilitate AI development in specific areas. For example, its OpenMMLab is an "open-source computer vision algorithm system" that seeks to facilitate academic and industry applications; it claims to have users from more than 100 countries and territories (**Figure 1** shows reported user affiliations). SHLAB's website further asserts strategic partnerships with at least thirteen universities across China. The research institution is also a significant contributor to China's "technical research on AI safety, and it has also begun producing accompanying policy papers" to advance AI safety. In July 2023, China's National Artificial Intelligence Standardization General Working Group (国家人工智能标准化总体组) announced the creation of a large model-focused working group (大模型专题组) led by SHLAB and comprising the companies Baidu, Huawei, Qihoo 360, China Mobile, iFlytek, and Alibaba.

[3] Shanghai AI Lab is a "new-type scientific research institution" (NSRI; 新型科研机构). It is almost certainly state-affiliated, though the exact structure of the organization is not clear and lies beyond the scope of this research. NSRIs (also known as NRDI; 新型研发机构) can have very diverse institutional structures and do not necessarily receive government funding. For readers who may judge it incorrect to include Shanghai AI Lab in Table 1 as a government entity, the next government research institute to appear in the NeurIPS data is the CAS Institute of Computing Technology (90 accepted papers; ranked 19th among other entrants from China).

***Figure 1:** Institutions reportedly using or affiliated with users of OpenMMLab as of September 2024 (Source: OpenMMLab)*

According to information released by the Cyberspace Administration of China, 302 generative AI service models are fully registered in China as of January 2025. Noteworthy frontier model companies contributing to China's LLM capabilities — most often from the private sector, with some seen in the NeurIPS data — include DeepSeek, Alibaba, Baidu, 01.AI, Tencent, Stepfun, ByteDance, Infinigence AI, and ModelBest. Signifying the important role of academia-industry linkages, China's "new AI tigers" — Baichuan AI, MiniMax, Moonshot AI, and Zhipu AI — were founded by Tsinghua University faculty and graduates.

China's semiconductor industry plays a key role in supporting the country's ecosystem for AI development, which is almost certainly critical to achieving the AIDP's goals. Faced with explicit[4] efforts by the US to curtail China's advances in both AI and semiconductors, China's support — including financial support — for the domestic semiconductor industry has almost certainly accelerated.[5] One vehicle of state-led investment, the multi-phase China Integrated Circuit Industry Investment Fund ("Big Fund"; 国家集成电路产业投资基金) had minority ownership in at least 74 domestic semiconductor companies as of February 2023. Huawei is a particularly important and rising contributor to China's semiconductor industry, but multiple Chinese companies are making progress or experimenting with methods (**Table 2**) that — while far from cutting-edge capabilities and likely currently insufficient for domestic demand — likely have the potential to alleviate China's dependence on the US and others in the long term. In March 2025, Bloomberg reported (citing unnamed sources) that China's Ant Group had found a way to train AI models using domestically produced chips at a cheaper cost but with similar results compared to those obtained by using US chips.

[4] See 2023 remarks by then-US Secretary of Commerce Gina M. Raimondo (relevant remarks begin at approximately 9 minutes and 20 seconds).
[5] U.S.-China Economic and Security Review Commission, "Chapter 3: U.S.-China Competition in Emerging Technologies", in 2024 Report to Congress of the U.S.-China Economic and Security Review Commission (November 2024), p. 181.

China's Semiconductor Development
Huawei is now mass-producing 7-nanometer (nm) Ascend 910B and 910C AI chips, marking significant progress despite yield rates reportedly at only 40% — though Huawei's yield was reportedly at 20% only a year before.
China is now forecasted to become the world's single largest source of IC wafer manufacturing by 2026, likely led by growth in 12-inch wafer fabs by Big Fund investees like Semiconductor Manufacturing International Corporation (SMIC), Hua Hong Semiconductor, Runpeng Semiconductor (润鹏半导体, a subsidiary of China Resources Microelectronics), Tiancheng Advanced (天成先进), Yandong Microelectronics (燕东微), and Zensemi (广州增芯). [6] [7] [8] [9]
Chinese companies are currently exploring alternative techniques to manufacture extreme ultraviolet (EUV) lithography equipment, with patents by the Harbin Institute of Technology (哈尔滨工业大学) and Shanghai Micro Electronics Equipment (SMEE, "上海微电子") exploring laser-induced discharge plasma (LDP, "激光诱导放电等离子体") and laser-produced plasma (LPP, "激光产生的等离子体") EUV sources, respectively.[10] [11]
China is progressing on domestic production of deep ultraviolet (DUV) lithography equipment, including recent breakthroughs in domestic manufacturing of DUV photoresists by Hubei Dinglong (鼎龙) and Shenzhen Rongda (容大感光).

Table 2: *Evidence of development and milestones in China's domestic semiconductor industry (Source: Insikt Group)*

Economic espionage, whether state-led or otherwise, also very likely supports China's progress in advancing its AI and semiconductor industries. It would be inaccurate to attribute all progress to this activity, but government-directed and tacitly allowed espionage almost certainly remains an important instrument in China's toolkit for achieving modernization. In October 2023, intelligence leaders of the Five Eyes alliance noted AI as one sector targeted by China's illicit actions. In February 2025, the US Department of Justice indicted a former software engineer at Google over alleged attempts between mid-2022 and mid-2023 to steal proprietary information on hardware and software enabling supercomputing in support of large AI models. The engineer allegedly intended to use the information to support his own and another early-stage AI-related company. In November 2024, a former employee of South Korea's SK Hynix Inc. was convicted of printing and seeking to retain proprietary documents on semiconductor manufacturing solutions in mid-2022, likely to facilitate her role in a new position at Huawei. Although not always a form of espionage, China's AI sector very likely further benefits from concerted efforts to attract and recruit foreign talent. Additionally, there is evidence that Chinese companies have circumvented foreign platforms' terms of service to enhance their own models through model distillation techniques.

[6] *https://www.jiemian[.]com/article/10611806.html*
[7] *https://www.sohu[.]com/a/844386326_121266081*
[8] *https://finance.sina[.]com[.]cn/roll/2024-11-22/doc-incwxxhe3843052.shtml*
[9] *https://www.eet-china[.]com/mp/a343946.html*
[10] *https://finance.sina[.]com.cn/stock/relnews/cn/2025-01-21/doc-inefsqmi8939277.shtml*
[11] *https://www.scmp[.]com/tech/big-tech/article/3278235/chinese-chip-making-shows-progress-new-euv-patent-domestic-lithography-champion*

Government and Venture Capital (VC) Funding

Funding for AI-related technologies is likely on an upward trajectory in both the US and China as of early 2025. Direct comparisons are difficult due to definitional issues and — especially with regard to China — data intransparency, but US federal government funding for civilian (non-military) AI likely exceeds the Chinese central government's funding by up to several billion USD per year, especially in light of evidence of declining expenditure by Beijing until approximately 2023 (likely in part due to COVID-19). The balance of government-led investment when including activity at the state and provincial levels likely leans in China's favor, where guidance funds combine public and private capital to invest approximately $16 billion annually in likely AI-related companies (broadly defined). Total private-sector investment in AI companies in the US far outstrips private-sector investment in China.

According to Federal Budget IQ, the US federal government has increased its AI and IT R&D spending from $8.2 billion in 2021 to $10.4 billion in 2024, a nearly 27% increase. The 2025 fiscal year budget earmarked $11.2 billion for AI and IT R&D spending. Removing spending by the US Department of Defense (DoD; $2.035 billion) and the Defense Advanced Research Projects Agency (DARPA; $1.41 billion), US federal spending on AI and IT R&D for civilian applications represents approximately $7.33 billion. Within this, the budget for "core" AI ($1.955 billion) is the largest component, though funding for the likely AI-relevant area of high-capacity computing is also rising quickly. The US federal budget for "core" and "cross-cut" AI in fiscal year 2025 combined is $2.8 billion (excluding US DoD and DARPA).

In China, total central government funding for AI is much more opaque. **Figure 2** and **Figure 3** show likely AI-related investment in basic, applied, and experimental R&D over recent years for two of China's largest government-run institutions: the National Natural Science Foundation of China (NSFC) and the Chinese Academy of Sciences (CAS). Specifically, **Figure 3** presents data from the budgets of three CAS institutes particularly focused on AI. **Figure 4** shows notionally earmarked funds for research to be conducted under China's "New Generation Artificial Intelligence" Megaproject (科技创新 2030—"新一代人工智能"重大项目). In addition to these data points and trends, the two organizations supervising nearly half of China's state key labs (SKLs) — the Ministry of Education and CAS — cumulatively budgeted 2.4 billion RMB ($327 million) under associated R&D line items.[12] Of this, approximately 29% would go to SKLs in the information and communications technology field, assuming funds were distributed equally among all SKLs supervised by these organizations and not accounting for other known unknowns.

Insikt Group's review of the data summarized above and evidence from other central government funding channels leads us to a low-confidence estimate that China's central government spends on the order of several billion USD annually on likely AI-related R&D, roughly similar to the US government's spending on "core" and "cross-cut" AI (excluding US DoD and DARPA).[13] Because "core" and "cross-cut" AI have a clear definition and estimating China's funding relies on imperfect visibility into the research program it targets, it is likely that US federal government funding for AI is higher than that

[12] Based on Insikt Group's review of relevant budget documents. Sources held by Recorded Future.
[13] Insikt Group's full assessment of Chinese central government investment in AI is available via the Recorded Future Intelligence Cloud, here: China's Central Government AI Investment Likely Totals Several Billion USD Per Year.

of China's central government. Total Chinese government-led investment in AI — via guidance funds — likely exceeds US cumulative federal and state investment, however (discussed further below).

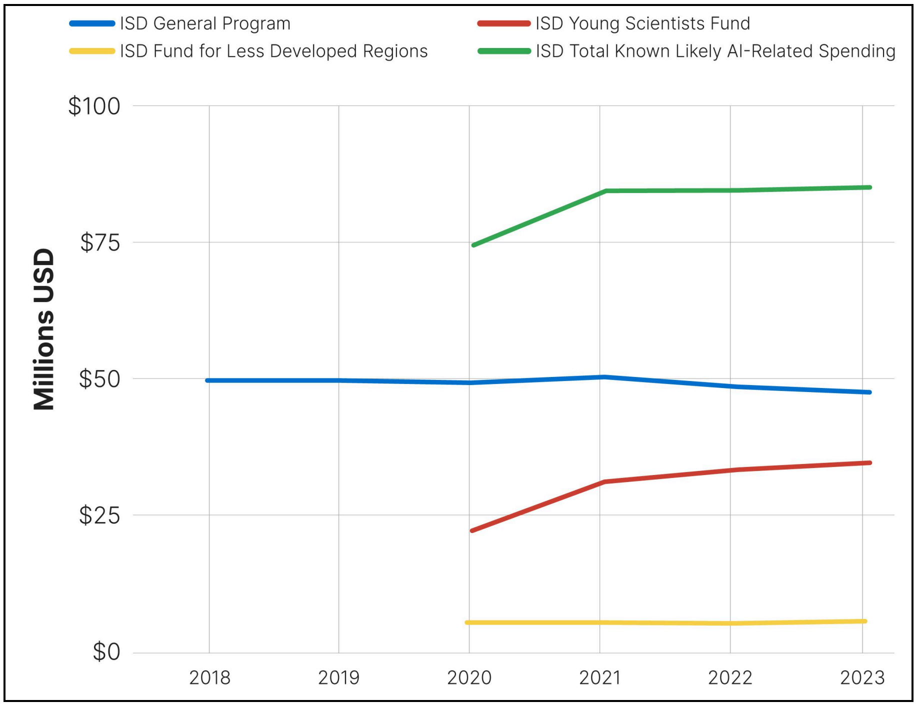


***Figure 2:** NSFC National Natural Sciences Fund Information Science Department (ISD) R&D expenditures (actual) for AI and automation, 2018–2024 (RMB all years converted to 2025 USD) (Source: Insikt Group[14])*

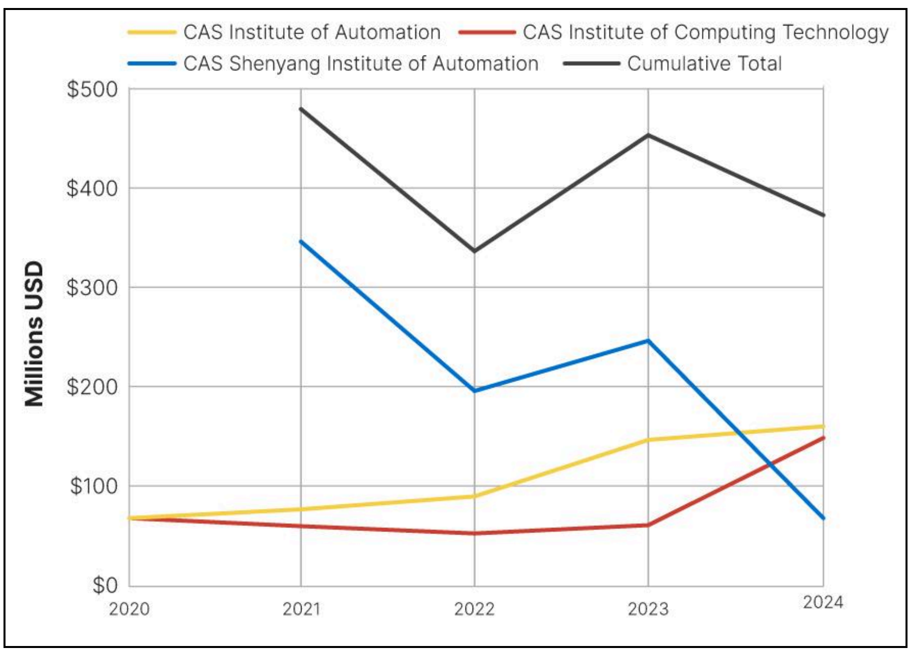


***Figure 3:** R&D expenditures (2018–2023 actual; 2024 budgeted) for AI-focused CAS research institutes, 2018–2024[15] (RMB all years converted to 2025 USD) (Source: Insikt Group[16])*

[14] Based on Insikt Group's review of relevant budget documents. Sources held by Recorded Future.
[15] ISD refers to NSFC's Information Science Department.
[16] Based on Insikt Group's review of relevant budget documents. Sources held by Recorded Future.

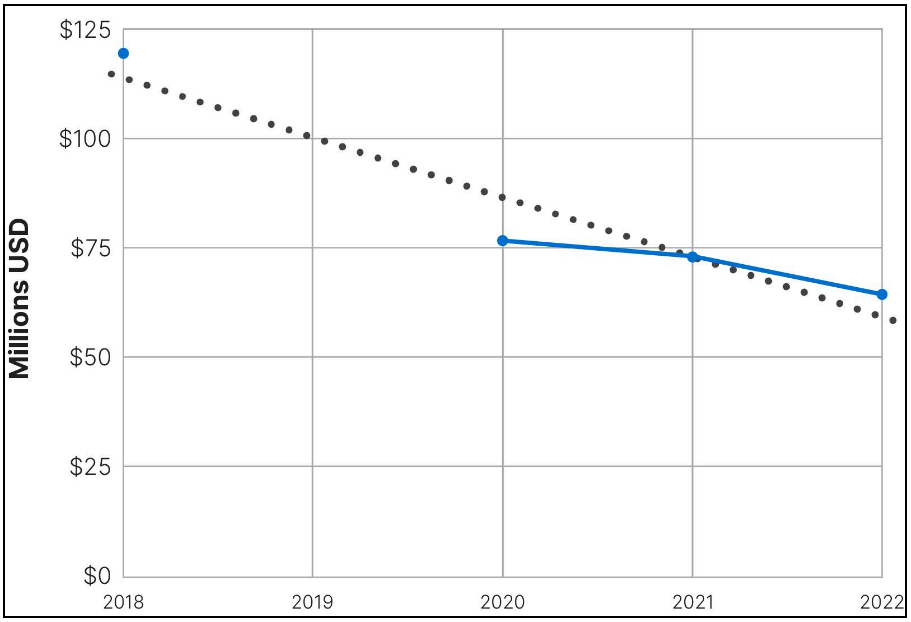


Figure 4: *Annual proposed funding for projects announced under the "New Generation Artificial Intelligence" Megaproject, 2018 and 2020–2022 (data not available for 2019; the dotted line shows the five-year trend); all RMB converted to 2025 USD (Source: Insikt Group[17])*

The declining trend through the early 2020s is likely related to the impact of COVID-19 on China's economy, but should also be viewed cautiously given missing data from other central government funding channels. As of early 2025, central government actors have signaled increasing interest in AI. In January 2025, Bank of China announced a five-year plan for financially supporting the AI industry value chain with a minimum of one trillion RMB ($138 billion).[18] In March 2025, China's National Development and Reform Commission announced the establishment of a National Innovation Investment Guidance Fund (国家创业投资引导基金). The new fund is being dubbed an "aircraft carrier-level" (航母级) fund and will seek to raise one trillion RMB ($138.2 billion) while focusing investment on start-up, early-stage, small-, and medium-sized enterprises in "frontier domains" like AI, quantum science, and future energy. Indicating its focus on the concept of "patient capital," the fund's duration is being set at twenty years — reportedly about five years longer than most other guidance funds.

Government guidance funds, such as the one announced in March 2025, have been established at multiple levels of government in China and mix funds from governmental and non-governmental sources (including corporate sources). This creates a challenge when attempting to isolate just central government funding. However, research published by the National Bureau of Economic Research (NBER) has found that more broadly across central, provincial, and local governments, these guidance funds have "invested in 9,623 unique AI firms through more than 20,000 transactions, in total 184 billion USD" between 2000 and 2023, with the majority of transactions occurring since 2013.[19] Annualizing this

[17] Based on Insikt Group's review of relevant documents. Sources held by Recorded Future.
[18] Importantly, activity under this initiative may largely be a recategorization of existing investments rather than new capital allocation.
[19] AI investment may be higher. The NBER study also states that the "government VC funds invested in 912 billion USD in the decade leading up to the second quarter of 2023," where "23% [$209.8 billion] of the total investment amount of the government VCs fund" went to AI firms. Insikt Group analysts could not determine the reason for this apparent discrepancy.

estimate over the 2013–2023 period points to an average spend of $16.7 billion per year, though the NBER study's method of identifying AI companies was very broad and potentially leads to an overestimate.[20] There are state-level initiatives in the US, such as New York's Empire AI consortium that has received $565 million in funding from public and private sources, but they likely do not match sub-national funding in China, which is steadily increasing. In January 2025, two state-backed Chinese investment companies announced the creation of a National Artificial Intelligence Industry Investment Fund (国家人工智能产业投资基金) with an initial registered capital of 60 billion RMB ($8.2 billion).

Although China's government may be ahead in terms of total AI investment — when the likely small lag in central government spending is compared to the likely larger lead in sub-national investment through guidance funds — total private-sector investment in AI companies in the US vastly outmatches private-sector investment in China. This is seen in data (**Figure 5**) collected by both Stanford University's Center for Human-Centered Artificial Intelligence (Stanford HAI) and the Emerging Technology Observatory (ETO), a project under Georgetown University's Center for Security and Emerging Technology (CSET). Note that there is likely overlap between the aforementioned investments of China's guidance funds and the private-sector investment data; that is, this data likely does not exclude investments in AI companies that come from government sources. This private-sector investment data is narrower, however; ETO's data only reflects investments into 2,100 Chinese AI companies compared to the NBER study's over 9,000 companies.

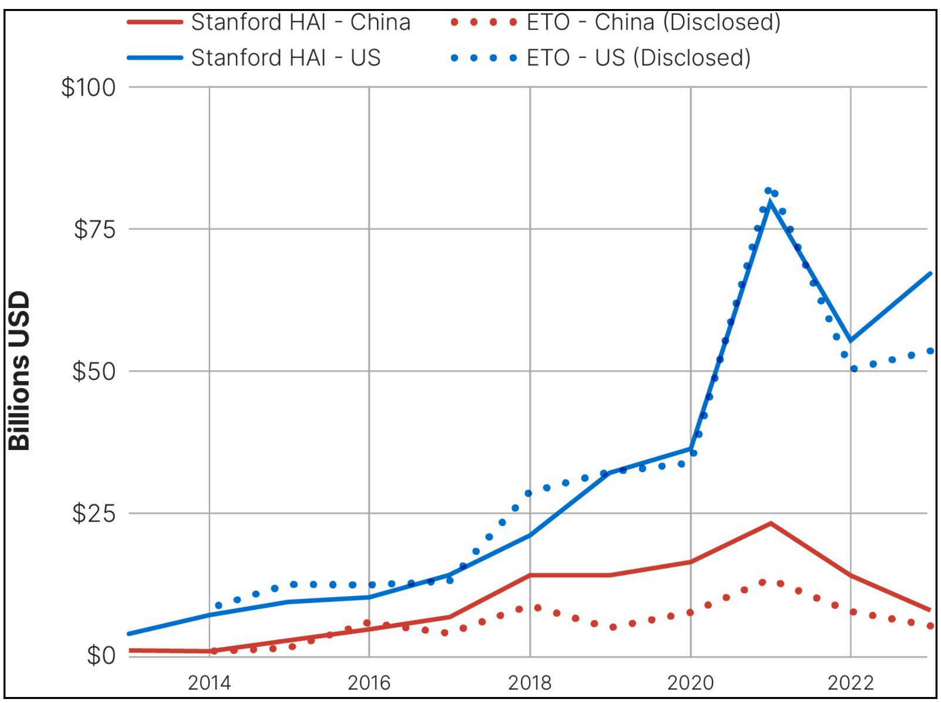


Figure 5: *Disclosed private-sector investment in China and the US, 2013–2023 (Sources: Stanford HAI; ETO)*

[20] The study identified AI companies by searching registered business scopes for a broad set of keywords. Keywords included targeted phrases like "Artificial Intelligence" (AI), "Neural Networks," and "Machine Learning," but also expansive terms such as "Data Science," "Image Processing," and "Robotics."

AI Regulation

Although necessary, industry regulations are a potential obstacle to accelerating AI development. The US government has been slow to advance specific limits on AI or duties for developers. Conversely, the Chinese government became a first mover in AI regulation globally in 2021 and 2022. China has established a regulatory framework that places limits and duties on AI and developers for the purposes of protecting consumer rights, the public good, and the government's security interests. China's regulatory scheme likely creates a drag on Chinese AI capabilities and extends development and deployment timelines, but only among developers aiming for public-facing products. Existing regulations in China likely do not impair the development of cutting-edge advances for non-public use cases, except to the extent that teams may proactively choose methods that align with the standards for public products.

In the US, there have been federal and state-level executive orders urging "responsible" and "ethical" practices, but no comprehensive law at the federal level. One of US President Donald Trump's first acts in his second term was to repeal an executive order issued by former president Joe Biden that had sought to coordinate a federal approach to "safely" developing and using AI. Proposals in the US Congress reflect an emphasis on "voluntary guidelines and best practices for AI systems," likely to avoid limitations that might inhibit innovation. AI restrictions are emerging at the state level, but enacted laws often focus on consumer privacy and the risk of automated user "profiling." Tennessee's "Ensuring Likeness Voice and Image Security (ELVIS) Act" specifically targets generative AI by prohibiting imitation of people's voice or likeness without their permission, or creating the tools to do so. New York has also issued guidelines governing the "acceptable use of artificial intelligence technologies." A California state bill that would have mandated safety testing for 32 of the world's top 50 AI companies was ultimately vetoed. This patchwork of state-based regulation is likely to continue to be the foundation of US policy toward AI, wherein AI innovators will likely gravitate toward less-regulated jurisdictions.

The core of China's framework is a two-stage technology registration process. The first step is the "algorithmic filing" (or "algorithmic registration"; 算法备案) referenced in Provisions on the Management of Algorithmic Recommendations in Internet Information Services (2021; Art. 24),[21] Provisions on the Administration of Deep Synthesis Internet Information Services (2022; Art. 19) [1], and Interim Measures for the Management of Generative Artificial Intelligence Services (2023; Art. 17). This step is likely a formality, but applies to service providers, technical support, back-end technology, and consumer-facing products related to most types of algorithms as well as "deep synthesis" (content generation algorithms; in essence, generative AI).

The second step of the registration process — "generative artificial intelligence service filing" (生成式人工智能服务备案; also called the "large language model coming online filing" [大语言模型上线备案] and other variations) — is more intensive and entails two levels of direct government testing that can take

[21] An updated final version effective March 2022 contains the same provision.

months or longer and require developers to make adjustments to their technology before approval. There may also be "ongoing communication between [government authorities] and AI service providers" after this second stage of filing, similar to how traditional information platforms are regulated. However, these steps are likely only required for public-facing products.

Algorithm and generative AI developers who intend for their products to be public-facing must, per regulations, take steps to ensure their systems perform in the ways listed below. The aforementioned Interim Measures hold service providers accountable for the content their models produce (Art. 14) and instruct developers to seek training data that is truthful, accurate, and objective (Art. 7; within China's context, this means aligned with government viewpoints).

- Uphold core socialist values and not produce content that, for example, incites subversion, endangers national security, harms the national image, or advocates terrorism and ethnic discrimination
- Prevent discrimination by race, ethnicity, nationality, faith, and other characteristics during development, including in the selection of training data
- Respect and protect intellectual property rights, "commercial ethics," and commercial secrets
- Prevent endangerment of the physical and psychological well-being of people, such as preventing infringement of specific people's image, reputation, and privacy
- Increase the accuracy and reliability of generated content, where relevant to the specific technology

Authorities in China are further in the process of developing a recommended national standard for the "basic security requirements for generative artificial intelligence service." It is currently in draft form, but aligns with and — by its nature of being an industry standard — promotes steps to ensure compliance with existing regulations. For example, the draft's recommended standards assert that developers should establish compliant training data by rejecting sources that fail one of two checks for more than 5% "illegal and unhealthy information."

The Chinese government — with industry input — has, however, likely intentionally left room for innovation within its regulatory framework. For example, whereas the promulgated Interim Measures only applies to public-facing generative AI technologies, the draft version applied the above-listed requirements to all "research, development, and use of products with generative AI functions" (Art. 2). An early draft version of the aforementioned standard would have required the use of models included in China's "algorithm registry" — very likely precluding access to international models — while the current draft (issued in May 2024) does not.

China's framework likely allows at least some generative AI development teams space to make cutting-edge advances free from some security-minded restraints, even if certain changes in training regimes and other aspects must be made before final products can be sold to customers or released to the public. Anecdotally, this appears to be the case for some companies. In July 2024, an employee of a "top AI start-up in Beijing" told the Financial Times that the company's "foundational model is very, very uninhibited ..., so security filtering is extremely important" — seemingly referring to a process whereby sensitive material is culled from training data before further development of a releasable product. Per the Financial Times' reporting, other companies build a layer of security that sits on top of the model, striving to replace sensitive responses with refusals in real-time.

At the same time, research teams that develop public-facing platforms, or who believe they may do so in the future, are likely negatively impacted by China's current regulatory environment. The capabilities they achieve are likely impaired compared to competitors, at least to a limited extent or in specific contexts. Industry analysts in China reportedly discuss a "safe in, safe out" approach aligned with the aforementioned recommended national standard. DeepSeek implemented this approach with V2. The company acknowledged that this decision negatively impacts its model's performance:

> During pre-training data preparation, we identify and filter out contentious content, such as values influenced by regional cultures, to avoid our model exhibiting unnecessary subjective biases on these controversial topics. Consequently, we observe that DeepSeek-V2 performs slightly worse on the test sets that are closely associated with specific regional cultures. For example, when evaluated on MMLU, although DeepSeek-V2 achieves comparable or superior performance on the majority of testsets compared with its competitors like Mixtral 8×22B, it still lags behind on the Humanity-Moral subset, which is mainly associated with American values.

Talent Pool

The US and China are almost certainly the world leaders in terms of access to, and cultivation of, science, technology, engineering, and mathematics (STEM) talent. Historically, the US has held primacy in this aspect of the bilateral technology competition due to excellent post-secondary educational institutions and inbound immigration, including from China, in related fields of expertise. The US likely retains an edge over China in terms of AI talent and quality of education at higher levels, but that advantage and its practical implications for judging technological disparity have likely begun to erode as China's domestic industry expands and the quality of its research institutions improves.

Access to STEM Talent

In terms of the overall S&T workforce, China has more full-time employee researchers than the US as of 2023 — 2.4 million versus 1.7 million. For AI specifically, however, the workforce is likely much closer in size, and who leads in recent years is not clear. Studies from 2017 and 2018 found that China had a talent pool of about 39,000 AI researchers compared to the US's 78,000 researchers. Both the US and China face shortfalls in AI talent relative to industry demand.

Looking at post-secondary degrees conferred in STEM fields[22] and where scholars choose to study or work offers an imperfect proxy for understanding more recent and long-term trends in likely talent pool evolution. STEM degrees issued in China outnumber those in the US at all post-secondary levels

[22] Note that how STEM and related clusters of study like "Science and Engineering" are defined for analytical purposes varies between organizations. Numbers between different sources are therefore not directly comparable.

(bachelor's, master's, and doctoral) since at least 2015 (**Table 3** provides approximate numbers for 2019). While the closing gap appears to reflect that — at the level of highly educated talent — the disparity may not be numerically great, most US graduates at this level, especially those in fields likely to be AI-relevant, are not US citizens or permanent residents; the majority are likely foreign-born talent.

STEM Degree Level (2019)	US	China
Bachelor's	>400,000	>1.5 million
Master's	~150,000	>200,000
Doctoral	>30,000	>35,000

Table 3: *STEM degrees issued in 2019 by universities in the US and China (Source[23])*

This narrowing gap not only reflects personal choices in whether to obtain advanced degrees but also very likely reflects mobility patterns among STEM graduates. In 2024, the US National Science Foundation's (NSF) National Science Board (NSB) found the US to be "the most popular destination for internationally mobile students, hosting 15% of all international students worldwide in 2020." Moreover, among post-secondary degrees, persons on temporary US visas were more likely to earn "science and engineering" (S&E) degrees — which is broadly defined, including psychology and social sciences — than US citizens and permanent residents, especially at the doctoral level and specifically in the fields most related to emerging technologies. According to the NSF NSB, in 2021, temporary visa holders "earned more than half of doctoral degrees in computer and information sciences (59%), engineering (60%), and mathematics and statistics (54%)."

Advantageously for the US, the top-tier AI researchers almost certainly represented in these statistics tended to stay in the US and work for US-based institutions. Data on top-tier AI researchers who participated in the December 2022 NeurIPS conference, compiled for MacroPolo's Global AI Talent Tracker in 2023, strongly suggests that a sizable portion of undergraduates from China elect to study in the US at the graduate level, after which only a small portion returned to China for their careers. Similarly, NSF NSB finds that as of 2021, "more than half of doctorate-level computer and mathematical scientists and engineers ... working in the United States were born outside the country," and that many PhD-level graduates remain in the US for "significant amounts of time" after graduation.

This trend, which for many years likely constituted a talent advantage for the US, is likely declining. MacroPolo's research into NeurIPS conference participants showed that, from 2019 and 2022, China's ability to retain and employ AI talent domestically strengthened. In part, this likely reflects the maturity of China's education system during that period — the AIDP ushered in a wave of new support for AI education and AI-related post-secondary programs that continued to evolve through 2024 — and growing domestic demand for AI-related skills. As discussed below, DeepSeek is an example of this. Another example is Manycore Tech (群核科技) — a Chinese start-up focused on spatial models. After

[23] Xiaomei Yan, Tianzuo Yu, and Yizhe Chen (eds.), *Education in China and the World: Achievements and Contemporary Issues* (Shanghai Jiao Tong University Press and Springer, 2024), 397, 405, 407 (Figures 7, 14, 17), *https://link.springer[.]com/chapter/10.1007/978-981-99-5861-0_9*.

DeepSeek's debut and recognition alongside DeepSeek as a "dragon" of Hangzhou, Manycore is reportedly receiving an increasing number of applications from "top talents" graduating from elite universities in China, as well as from Stanford University and Harvard University. MacroPolo has observed a similar trend among AI researchers from India, who were increasingly staying in their home country.

This decline is also likely the result of increasing geopolitical tensions; an October 2024 working paper published by NBER found a decreasing probability between 2015 and 2019 that Chinese STEM PhD students would enroll at US universities and a decreasing probability that Chinese doctoral graduates from US programs would stay in the US, arguing these and other findings "strongly suggest" a "chilling effect" caused by geopolitical factors. The researchers further found that this talent was choosing to move to anglophone countries other than the US. In 2019, about 370,000 Chinese students were enrolled in US universities, compared to about 277,000 in the 2023–2024 academic year, with increased scrutiny — rooted in US national security concerns — of Chinese students in STEM fields likely being a factor in this declining trend.

Finally, authorities in China and Chinese technology companies are almost certainly seeking to recruit foreign talent, including from the US, in the fields of AI and other relevant areas like semiconductors. Multiple US-based AI-focused scientists have moved to China since 2016, some renouncing US citizenship in the process. However, as documented in a May 2024 report published by CSET, there is some evidence that those who accept recruitment offers associated with China's government-supported talent plans are often not the best of the best. Moreover, CSET, citing other research, asserts that scientists who "return to China may, on average, become less productive" (based on one's number of academic publications) than they are in the US, although neither the report nor its principal source examines why this is the case.[24]

Faced with a likely shrinking pool of immigrant talent, the US will likely become more reliant on domestic resources (the potential of remotely outsourcing AI-related jobs notwithstanding). This is likely to affect foundational research more than AI-related commercially oriented development activities, but in the long term is likely detrimental to the US's development of AI capabilities. Stanford University's Emerging Technology Review 2025 does not distinguish between immigrants and US-born talent but notes that "universities report that the number of students studying in AI who are joining the industry, particularly start-ups, is increasing at the expense of those pursuing academic careers and contributing to foundational AI research." The NSF NSB nevertheless describes foreign graduates as a "vital" source of STEM capabilities within the US economy.

Quality of Accessible Talent

A numerical advantage in access to STEM graduates and the S&T workforce is not the only factor to consider; the quality of educational programs available and the excellence of institutions (academic or corporate) conducting AI-related research are also critical. Below the post-secondary level, China's

[24] Ji Young Hwang, "Return or Stay? Impact of Reverse Brain Drain: The Case of the Youth Thousand Talents Plan in China," Master's thesis submitted to Korean Development Institute (School of Public Policy and Management, 2022). Available for download: *https://archives[.]kdischool[.]ac[.]kr/bitstream/11125/46596/1/Return%20or%20stay%3F%20impact%20of%20reverse%20brain%20drain.pdf*.

education system appears to succeed over the US based on international competitions, likely contributing to a relatively stronger foundation for, and interest in, STEM fields as potential talent ages. From 2017 to 2021, Chinese middle school students consistently achieved more gold medals than their American counterparts in STEM olympiads.[25] At the high school level, China ranked number one globally in STEM olympiads from 2015 to 2023, according to aggregate performance results calculated by the Center for Excellence in Education. In 2024, the US ranked first for the first time since 2001.

At the post-graduate level, however, US universities enjoy long-standing top positions in global rankings based on quality of output metrics. EduRank is an "independent metric-based ranking of 14,131 universities from 183 countries" that uses data on scholarly publications and citations to compare universities in specific fields. As of March 2025, seven out of the top ten universities publishing on AI-related topics are located in the US. Among these in the top three positions are Stanford University, University of California - Berkeley, and the Massachusetts Institute of Technology. Only one university from China — Tsinghua University — currently makes the top ten per EduRank's methodology. The second China-based institution on EduRank's list, Harbin Institute of Technology, holds the 25th spot globally. Despite rapidly expanding numbers of AI programs offered at Chinese universities, as of 2022, many reportedly lacked high-quality strong curricula, datasets, computing power, and other resources.

Rising quality among China's AI-focused research institutions (academic or corporate) can be observed via participation in the aforementioned NeurIPS conferences since 2021. There has been a sharp rise in the number of accepted papers and quality of submissions by Chinese researchers between 2021 and 2024, which likely signals a stronger engineering and research pool in the field of AI and generative AI specifically. According to Paper Copilot, four Chinese universities (Tsinghua University, Zhejiang University, Shanghai Jiao Tong University, and Peking University) were among the top ten organizations with the most accepted papers to NeurIPS 2024. While the US remains the leader in submissions, Chinese entities have nearly quadrupled (3.77x) their quantity of total accepted papers at NeurIPS between 2021 and 2024, reaching over 3,700 papers in the last four years — still trailing the estimated 7,479 papers from US entities accepted during the same period.

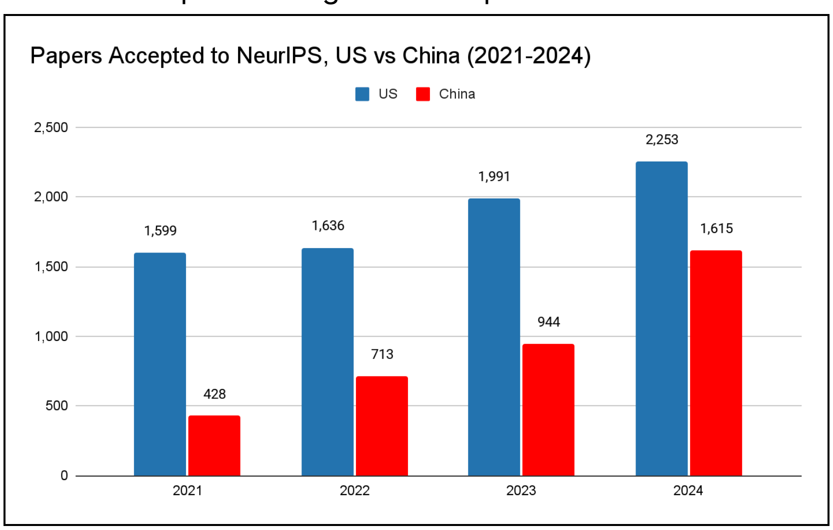


***Figure 6**: Papers accepted at NeurIPS, US vs China (2021–2024) (Source: PaperCopilot)*

[25] Ibid., 401 (Figure 9).

Increasingly frequent (4.03x) cross-sector collaboration among Chinese institutions participating in NeurIPS since 2021 is also likely resulting in better submissions: a paper co-authored by researchers from ByteDance and Peking University on visual autoregressive modeling won "Best Paper" at NeurIPS 2024. Researchers from Xiamen University, Tsinghua University, Shanghai AI Laboratory, and Microsoft Research Asia also won "Best Paper Runner–Up" for proposing a novel language model, RHO-1.

Some observers in China are critical of the country's ability to cultivate high-quality talent. For example, an op-ed by Fang Shou'en, the head of the Chinese Communist Party (CCP) committee at Tongji University — researchers from which have authored or co-authored more than 48 papers accepted by NeurIPS since 2021 — highlights that the number of PhD advisors in the AI field is growing, but many lack industry experience. Fang notes that many companies in China lament a lack of "practical skills" among graduates who studied AI. He also asserts that industry-academia collaboration in China continues to fall short of the level of such activities seen in the US.

Despite these challenges, China likely increasingly has the capacity to cultivate and utilize top talent within its borders, signaling decreasing reliance on the old model of development in which innovation relied on Chinese students going abroad to obtain advanced skills and then returning to contribute those skills domestically. DeepSeek's research team likely exemplifies this trend. Ryan M. Allen of Soka University of America examined the backgrounds of 31 "core contributors" to published DeepSeek papers, finding that all but two obtained their undergraduate and graduate degrees from Chinese universities. Only one previously studied in the US (a third researcher who was previously a visiting student at Stanford University notwithstanding). Moreover, Allen found that, among graduate school advisors to the DeepSeek team for which information could be identified, only four had studied in the US (plus a fifth advisor who studied in Singapore). While speaking at a Stanford University-Hoover Institute event in February 2025, Amy Zegart (who, among other positions, is a senior fellow at Stanford HAI) cited the work of an unnamed researcher to challenge the "myth" of returning overseas talent.[26] The report behind these remarks was released in late April 2025, stating that 111 out of 201 members of DeepSeek's research team had been "trained and affiliated exclusively at Chinese institutions throughout their careers — evidence of China's growing capacity to develop world-class AI talent domestically without relying on Western expertise."

AI Diffusion

Innovation — breakthroughs at the frontiers of sciences — is important, but without diffusion — the adoption of new technologies — very likely has little significance in terms of which country "wins" the AI competition by leveraging the technology for long-term economic and geopolitical benefit. Broadly, the US and China have very different approaches to supporting the diffusion of AI capabilities. Diffusion in the US gives precedence to a market-led, "bottom-up" approach that prioritizes private enterprises and academia, while China likely aims to encourage diffusion through "top-down" government planning and funding at the central and local levels.

[26] Relevant remarks begin at approximately 36 minutes and 30 seconds).

Although the US likely retains a first-mover advantage, China's AI diffusion capacity is likely to increase quickly due to the aforementioned government-led approach, coupled with private-sector support that is similar to developments in the US. Reportedly high trust in AI within China, as well as state-led efforts to encourage positive views of AI (as seen in the AIDP) and current lean toward open source (discussed further in **The Model Gap** section, below), is also likely to facilitate faster adoption. Whether diffusion in China exceeds or will exceed diffusion in the US is ultimately unclear, but by one metric — patent applications — China is leading in overall volume.

Government Support

China's approach toward state guidance is likely (but not assuredly) to translate into an advantage for the adoption of AI. Following the release of the aforementioned AIDP in 2017, sub-national governments have released their own downstream policies to promote local AI industries and strategies. Localized policies focus on accelerating the adoption of AI capabilities in ways that best use regional specializations, existing industries, and other geographic factors:

- Beijing is promoting AI for applications such as "government services" ("政务服务"), "judicial services" ("司法服务"), and "content security" ("内容安全"), including multimodal analysis for "large-scale content review work" ("大规模内容审核").
- Shaanxi province is focused on developing AI applications for defense by fostering collaboration between local universities such as Xi'an Jiaotong University with defense companies like Xi'an Tianhe Defense Technology (天和防务) for "intelligent exoskeletons" ("智能骨骼机器人"), accelerating AI-driven development of unmanned aerial vehicles ("培育智能无人机产业") for defense and agriculture, and encouraging the "proactive association of generic large model development organizations" ("主动联合通用大模型开发机构").[27]
- Yunnan, which borders Vietnam, Laos, and Myanmar, seeks to use AI to develop products and services for Southeast Asian partners and has deployed AI to inspect power substations in mountainous regions, reducing local government spending.

To what degree these and other government efforts will achieve their goals is unclear. Application and experimentation in some areas are already observable. At least 72 local governments throughout China have already deployed DeepSeek in some fashion. On the issues of "content security" and "content review," a leaked database of articles in need of categorization shows how an LLM is being used to automate the mass classification of content as "military developments," "social developments," and other subjects, likely as one step in a multi-stage censorship process to manage public sentiment. China's military hospitals are reportedly using DeepSeek to support doctors in developing treatment plans.

Over the long term, at least some aspects of these government policies are likely to succeed. For instance, an innovation-to-diffusion success is China's "large-scale aircraft" (大型飞机) megaproject that was introduced in the early 2000s as part of the National Medium- and Long-Term S&T Development Plan (2006–2020) (国家中长期科学和技术发展规划纲要(2006—2020年)). Development of

[27] Source downloaded from *https://files.parkworld[.]net/files/8dc5929fce04efb/download*

what would become China's first domestically produced "large passenger plane" — the C919 — began shortly afterward. As of December 2024, C919s had carried one million passengers. Jeffrey Ding, assistant professor of political science at George Washington University, notes other successes, including mobile payment applications and high-speed rail, while also pointing to evidence of challenges in other areas also promoted by the state. Comparing broad-range indicators, Ding finds that (up to at least 2020) there is a consistent gap between China's capacity for innovation and its capacity to adopt new technologies. As of 2020, he also identified evidence that US adoption of some technologies, such as AI-relevant cloud computing and storage, exceeded adoption in China.

Corporate Diffusion Via AI Platforms

Questions over the potential impact of government support and market adoption notwithstanding, tech companies in China are enthusiastically building platforms to translate generative AI capabilities into commercial software-as-a-service (SaaS). Cloud providers and others are partnering with AI labs to integrate LLMs into their tools and tailoring AI services to sector-specific applications. For example, Baidu offers the Qianfan "large model platform" ("千帆大模型平台") designed to simplify enterprise adoption of generative AI via model-as-a-service (MaaS), providing developers with access to Baidu's own models (Ernie) and third-party open-source models (like DeepSeek and Meta's open-source models), retrieval augmented generation (RAG), model finetuning to specific application domains, and agentic workflows. Qianfan is also now integrated with one of the most popular LLM developer frameworks, LangChain. Tencent Cloud's TI platform, Alibaba Cloud's Model Studio, and Huawei's ModelArts also provide similar services, such as granting access to Huawei's AI chips via Ascend Cloud Service, enterprise access to Alibaba's Qwen models, and ModelArts offering iFlytek's Spark LLM models. Individual AI companies like iFlytek are also entering the AI services market by tailoring their models to specific use cases such as translation, transportation, video conferencing, and virtual anchors.

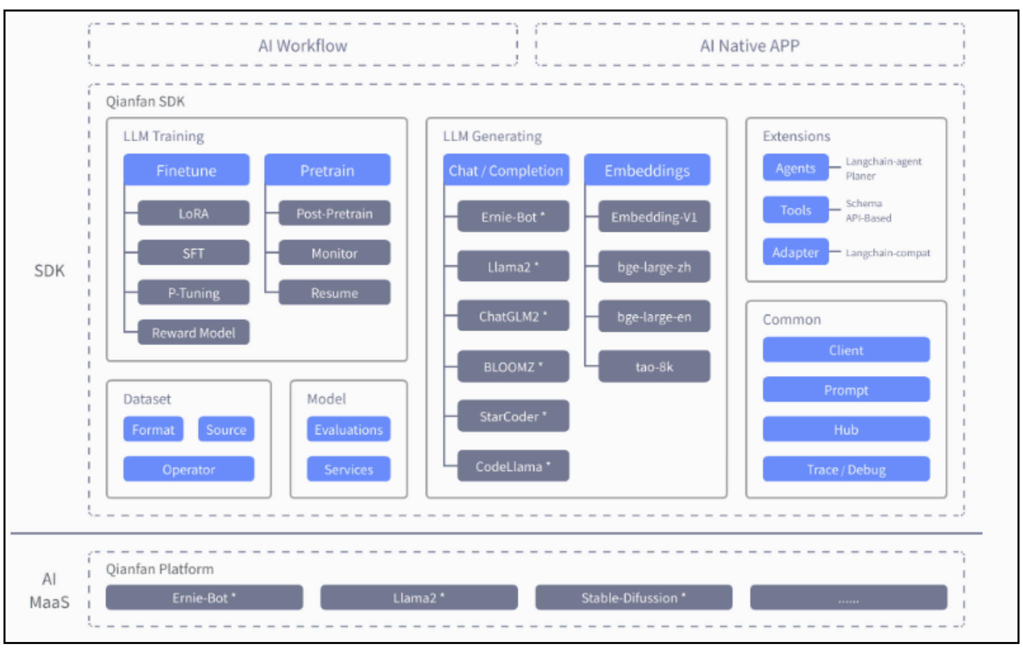


***Figure 7**: Summary of services and models provided by Qianfan (Source: Qianfan SDK Documentation)*

Further, these platforms are at the epicenter of domestic price wars, likely lowering barriers of entry to application development and accelerating the diffusion of generative AI capability to a broader number of users. As of November 2024, ByteDance's Doubao models had over 60 million monthly active users (MAU), and ByteDance slashed its prices for Doubao Pro models to prices 99.8% lower than OpenAI's GPT-4 in December 2024. In May 2024, Alibaba announced that its Qwen models had "attracted over 90,000 enterprise deployments" through Alibaba Cloud's Model Studio platform. In December 2024, Alibaba cut its prices for Qwen-VL-Max by 85% to $0.41 per million input tokens.

These activities in China are similar to those in the US, particularly as many tech giants (including Microsoft, Alphabet, Apple, and Amazon) are heavily invested in frontier model AI companies. For example, Microsoft's Azure AI Foundry now serves OpenAI models while Amazon Web Services (AWS) now serves Anthropic's Claude models in AWS Bedrock. In turn, companies seeking to integrate generative AI into their product suite can purchase access to these SaaS platforms. Companies like KPMG, Airbus, and UBS likely use the former (OpenAI via Azure AI Foundry) while Pfizer, Perplexity, and the European Parliament likely use the latter (Anthropic via AWS).

AI Patent Output

Metrics such as patent submissions across both the US and China can provide a proxy for measuring how each country applies AI in practice. Many AI breakthroughs are now published to public code repositories like GitHub, uploaded to preprint platforms like arXiv, or kept as trade secrets. Despite this, patenting is likely an indicator of diffusion to other sectors. The US Patent and Trademark Office (USPTO) uses patents to measure AI diffusion in the US economy. China has taken the lead over the US in patenting generative AI applications in many industries, including in sectors like software, finance, energy, and gaming, likely indicating a lead in diffusion.

AI patents submitted by entities in China have seen a sharp rise. Since the release of AIDP, Chinese AI patent applications have seen an average annual growth rate of 43%, with 22% of applications concerning natural language processing (NLP), according to an April 2024 report by China's Ministry of Industry and Information Technology (MIIT). As of 2024, China is a clear leader in patenting generative AI models and their applications. According to the World Intellectual Property Organization (WIPO), thirteen of the top twenty generative AI patent holders between 2014 and 2023 are Chinese, and only four are from the US. WIPO claims that Chinese companies lead patent claims in sectors like software (Tencent, Ping An, Baidu, Alibaba, ByteDance), telecommunications (Huawei), gaming (NetEase), finance (Bank of China), and energy management (State Grid Corporation of China). China's patent quality has traditionally been lower than other countries, but is improving; the country's patent grant ratio stood at 55% in 2023, a 25-point increase from 2019.

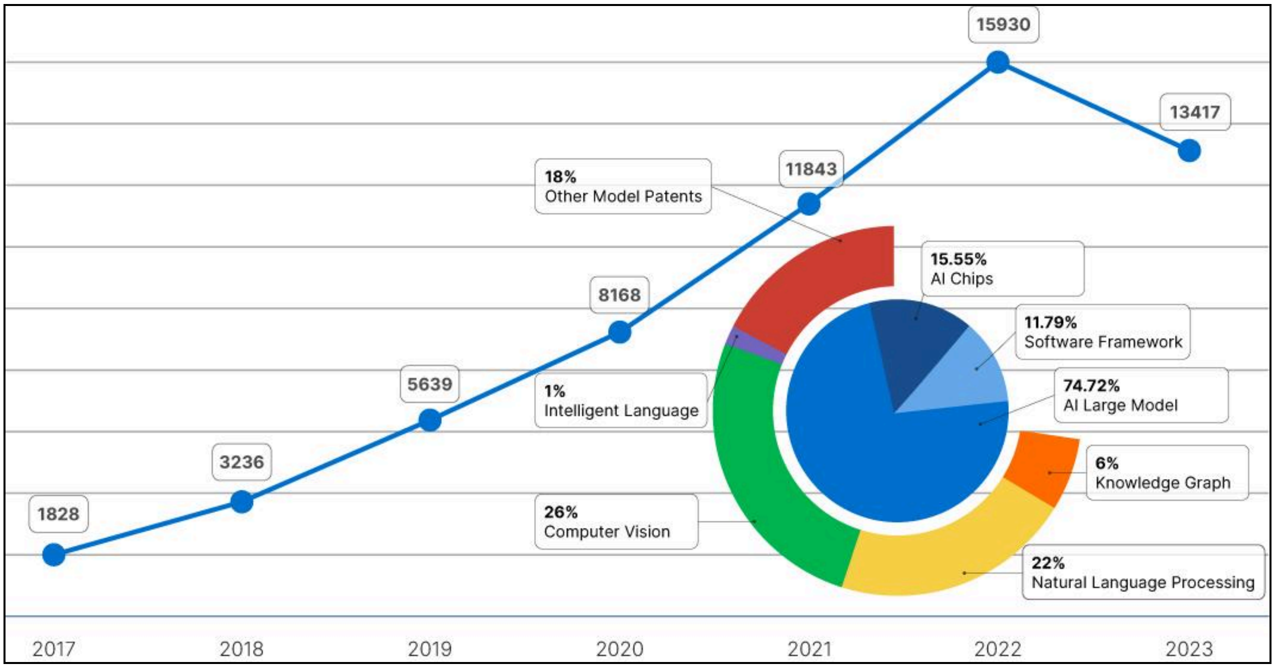


***Figure 8**: Annual patent applications by Chinese companies for AI models, chips, and frameworks and distribution of applications, 2017–2023 (Source: MIIT; translated and adapted by Insikt Group)*

China is also likely to adopt policies seeking to protect and accelerate the diffusion of generative AI intellectual property domestically. China's National Intellectual Property Administration (国家知识产权局) announced on March 8, 2025, that it is taking steps to "further strengthen the protection of [AI] intellectual property rights" and build a domestic patent pool for large AI models, in conformity with initiatives set out in the 2017 AIDP.

Narrowing the Technology Gap

In 2017, China's AIDP set out three milestones in AI competition: catching up with "advanced" AI capabilities by 2020, achieving "major breakthroughs" by 2025, and reaching "world-leading levels" by 2030. Following DeepSeek's release of R1, the Chinese government almost certainly views its 2025 goal as achieved and will seek to deepen its investments to achieve its 2030 goal of producing "world-leading" models. Chinese companies are rapidly improving model performance and cost, and, as a result, positioning themselves closer to their US counterparts at a fraction of the cost, leading to increasing adoption of Chinese models domestically and abroad.

Access to high-quality training data and compute infrastructure are likely to remain contested domains where US advantages persist. China's competitive advantage likely stems from Chinese AI companies' embrace of open sources and focus on cost efficiency, which facilitates diffusion and helps overcome export controls. Both countries will also likely continue benefiting from open science standards, as AI research conferences like NeurIPS demonstrate strong collaboration between US and Chinese entities.

Long-term AI hardware bottlenecks will likely limit Chinese companies' ability to continue benefiting from neural scaling laws, however, as US export controls are likely curtailing Chinese AI companies' access to cutting-edge AI hardware. China's investments in domestic semiconductor manufacturing and high-performance computing (HPC) infrastructure signal a clear intent to overcome export controls and become self-sufficient, though current achievements likely still fall short of these goals in significant ways, as analyzed below.

China is unlikely to overcome the sub-7nm bottleneck by developing domestic EUV lithography tools by 2030. Even if a domestic EUV capability emerges in the next five years, China will unlikely be able to scale to meet the demand for AI accelerators by that date. Thus, while Chinese capabilities have very likely made significant progress toward goals outlined by the AIDP, we assess it as unlikely that China will realize its goal to surpass US AI capabilities by 2030 due to constraints like compute capacity.

The Model Gap

In the last year, China has likely closed the gap in model performance according to industry standard benchmarks and evaluations due to models developed by DeepSeek, Alibaba, 01.AI, and Moonshot AI. Based on available data, measured according to crowdsourced user ratings, benchmark performance, and industry experts' assessments, Chinese frontier models likely now have a three to six-month performance gap behind their state-of-the-art US counterparts. Additionally, Chinese LLMs are significantly more cost-efficient than their US counterparts, which is almost certainly granting the Chinese AI ecosystem disproportionate adoption.

DeepSeek's recipe for success can be very likely generalized to other Chinese AI companies that achieved significant performance gains in the last year, including Moonshot AI's Kimi k1.5 and Alibaba Cloud's Qwen models. These companies have improved their models by implementing innovations pioneered by rivals in addition to prioritizing cost efficiency to compete domestically and overcome challenges posed by US export controls on AI accelerator chips. Chinese AI capabilities likely benefit from open science standards and collaborations with US entities, as demonstrated by co-authorship networks for papers accepted at major AI conferences like NeurIPS.

Acquiring new training data to continue scaling LLMs is almost certainly a persistent challenge for AI companies across both countries and is likely to remain a contested process that benefits companies with access to large amounts of user-generated content (UGC) such as Meta (via Facebook and Instagram) or ByteDance (via TikTok/Douyin). Open source is very likely an accelerant in the race to AGI, and will likely also benefit both sides. AI companies in both countries are increasingly choosing to release state-of-the-art, open-weights models (such as DeepSeek R1 and Llama 3 at the time of their releases), citing factors such as geopolitical influence and attracting talent. Additionally, model distillation is likely a growing risk for US AI companies and likely represents an opportunity for Chinese companies to obtain performance gains at a reduced development cost. Another option for overcoming training data constraints would be to use synthetic (artificially generated) data, an approach Alibaba is reportedly exploring.

Model Performance

Cumulative innovations in model architecture and algorithmic efficiency have very likely supported Chinese models' closing of the performance gap with the US. DeepSeek R1's precursors, V2 and V3, already implemented novel techniques such as multi-head latent attention (MLA), which reduces memory usage by 5% to 13% for LLMs, and a new mixture-of-experts (MOE) architecture dubbed "DeepSeekMOE" (later improved by the company's Deep Expert Parallelism library), which reduces the number of parameters activated by large models at runtime and facilitates training without access to large HPC clusters. In V3's technical report, DeepSeek engineers disclosed that they optimized the network performance of their GPU clusters by bypassing the industry standard high-level language developed by Nvidia, CUDA, and optimizing their compute infrastructure directly using Parallel Thread Execution (PTX), a lower-level instruction set. Moonshot AI's Kimi k1.5 and Alibaba Cloud's Qwen models show similar innovations, such as using reinforcement learning (RL) on chain-of-thought (CoT) tokens for Kimi k1.5.

Such innovations have almost certainly enabled DeepSeek and other Chinese companies to steadily raise their models to standards previously set by US rivals. An analysis of top Elo ratings for US and Chinese models on LMsys ChatArena demonstrates that Chinese models like Alibaba Cloud's Qwen and 01.AI Yi models showed early competitive gains even prior to DeepSeek's release of R1 (**Figure 9**).

Since January 2024, the gap between the top-performing US and Chinese AI models has fluctuated between three and six months (**Figure 10**), with DeepSeek R1 narrowing the gap to under three months in January 2025 by achieving a similar Elo score to OpenAI's o1, which was initially released in September 2024. Other Chinese models had previously temporarily reduced the gap to under three months, with 01.AI yi-large-preview model achieving a similar Elo score in May 2024 as Anthropic's Claude 3 Opus had in February 2024. Notably, this finding is similar to assessments made by other industry experts and academics. All model Elo scores used in our analysis are included in **Appendix B**. Insikt Group assesses that, while this gap may continue fluctuating between three and six months in the short term, unforeseen algorithmic breakthroughs can significantly impact the competitiveness of US or Chinese models in the long term.

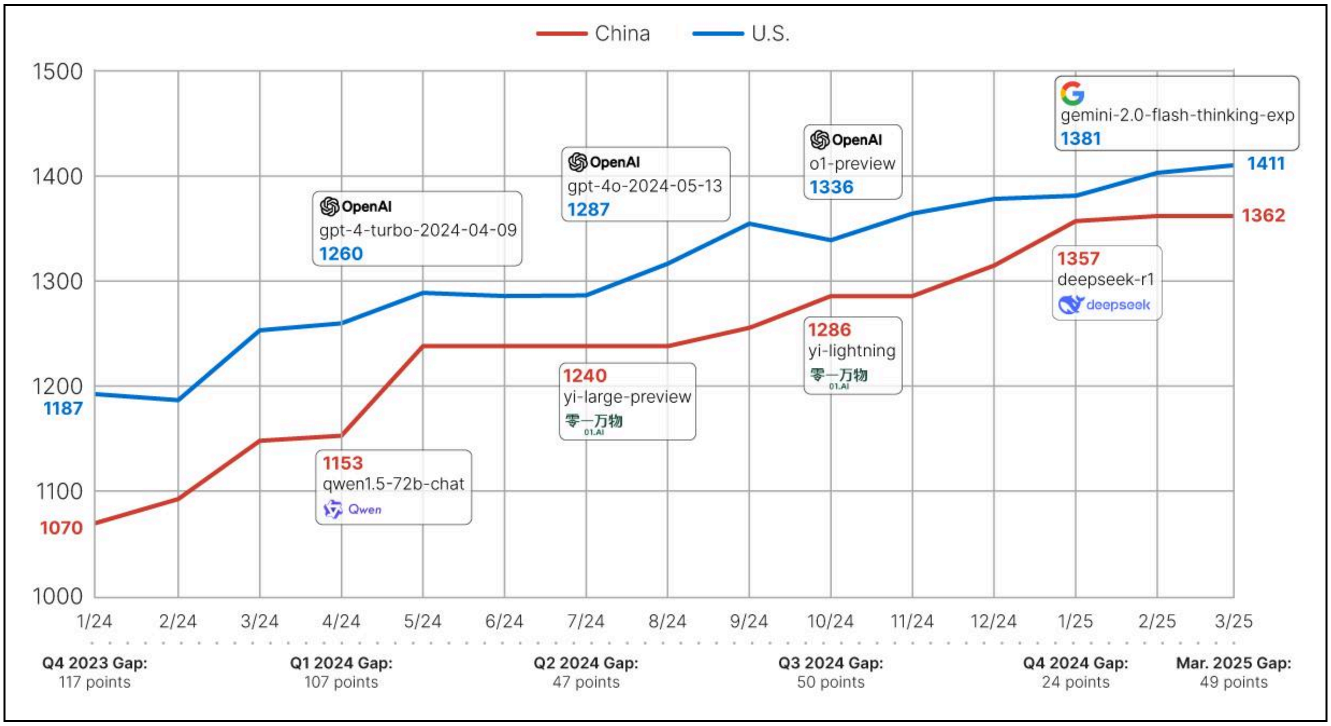


***Figure 9**: Top monthly Elo ratings for US models (blue) and Chinese models (red) (Source: [LMsys ChatArena](#))*

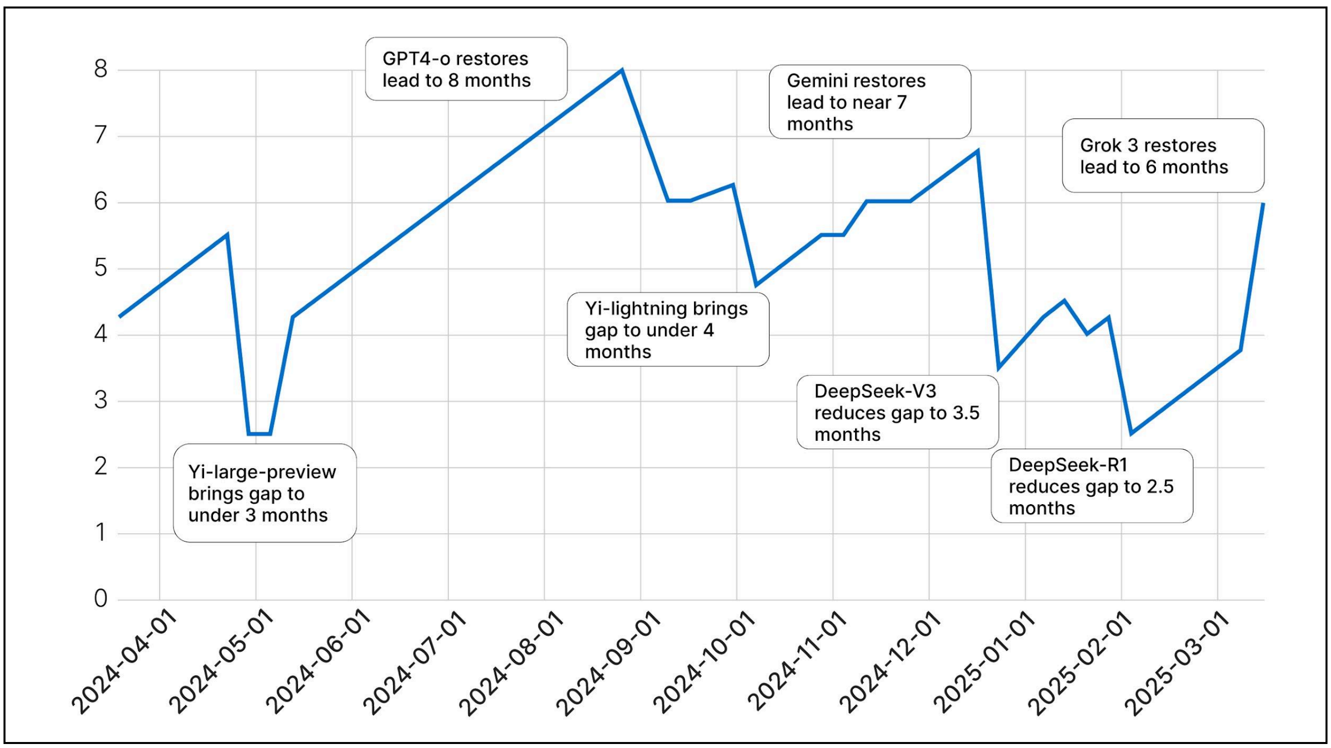


***FIgure 10**: Chinese models' lag on US models (months) (Source: [LMsys ChatArena](#))*

Chinese AI companies are now also proficient at adopting US and Chinese competitor-developed cutting-edge techniques for efficiency and performance gains, resulting in similar progressions in model performance across Chinese AI companies (**Figure 11**). Alibaba Cloud's latest Qwen 2.5 model series implemented techniques like grouped-query attention (GQA), first developed by Google Research in 2023, and Group Relative Policy Optimization (GRPO), a reinforcement learning technique introduced by DeepSeek in April 2024, to achieve significant efficiency gains. Alibaba Cloud's latest Qwen 2.5-series model, Max, now competes with models like Anthropic's Claude 3.5 Sonnet and OpenAI's GPT-4o. Such algorithmic efficiency gains allow models to increase performance and decrease cost. Cheap models with larger context windows can be distilled into higher-performance reasoning models using techniques like reinforcement learning on chain-of-thought (CoT) prompting, which require large amounts of output tokens but significantly increase performance in reasoning tasks.

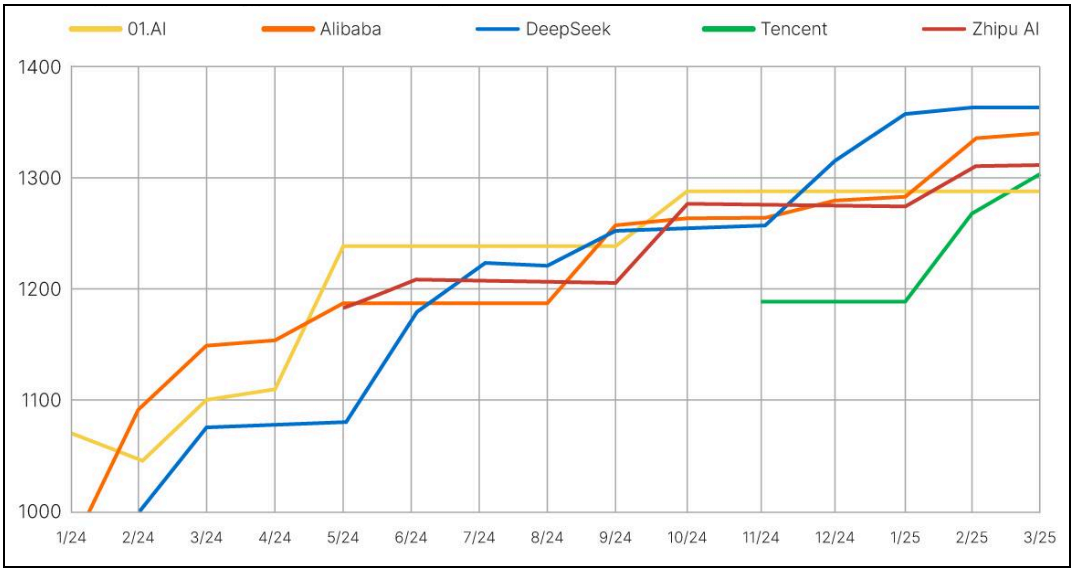


***Figure 11**: Top monthly Elo ratings, Chinese AI companies (Source: LMsys ChatArena)*

Cost efficiency is very likely an important metric to measure the capability of a national ecosystem to train and operate models that can be readily used and scaled for real-world applications, including those of strategic interest such as national security and intelligence applications. Lower inference costs are very likely a competitive advantage for China and its ability to diffuse generative AI capabilities domestically, even if performance gaps persist between the top US and Chinese models on benchmarks like Massive Multitask Language Understanding (MMLU) and Graduate-Level Google-Proof Q&A (GPQA). Chinese AI companies are also likely incentivized to cut costs even on state-of-the-art models due to domestic price wars, resulting in price cuts between "50 and 88%."

Closing the performance gap while remaining competitive on cost is very likely already paying off for Chinese AI companies. Companies that develop AI applications outside of China are more likely to replace US models with "satisficing" Chinese models that operate above required performance thresholds but are significantly cheaper to operate — dubbed the "access effect." Consumers seeking a close alternative to OpenAI's o1 reasoning model, for example, may select DeepSeek R1, which is on average 27x cheaper per million input and output tokens (**Table 4**).

Model	Release Date	MMLU	GPQA	USD / Million Input Tokens	USD / Million Output Tokens
o1	December 17, 2024	91.8%	75.7%	15	60
deepseek-r1	January 20, 2025 (34 days)	90.8%	71.5%	0.55 (27.4x cheaper)	2.19 (27.3x cheaper)
claude-3-5-sonnet-20241022	October 22, 2024	90.4%	67.2%	3	15
qwq-32b-preview	November 28, 2024 (37 days)	70.97%	65.2%	0.15 (20x cheaper)	0.20 (75x cheaper)

***Table 4**: Cost comparison between leading US (white) and Chinese models (red) (Source: LLMStats)*

Chinese LLMs are almost certainly diffusing abroad. Following DeepSeek's release of R1, some European startups reportedly switched to R1 and US AI providers Microsoft, Nvidia, AWS, and Perplexity began offering the model to their customers. OpenRouter, a platform that measures token throughput from users to commercial and open-source models hosted by third-party providers, recorded a significant jump in total tokens processed for all DeepSeek models (including all variants of R1 and V3) versus all OpenAI models (including all variants of o1, GPT4, and GPT3.5 Turbo, **Figure 12**).[28] Chinese provider SiliconFlow, which serves DeepSeek models using Huawei's domestically produced Ascend 910B AI chips, reportedly saw a thirty-fold increase in traffic following R1's release. Focusing on cutting costs for end users is likely an effective strategy for increasing market share and extending influence, even if Chinese models have a noticeable (but diminishing) lag behind frontier US models in terms of performance.

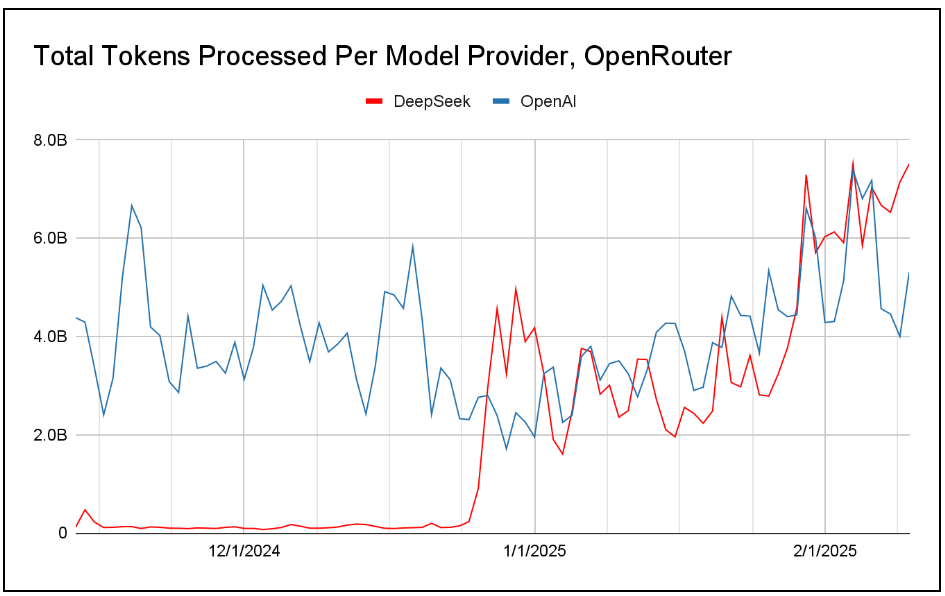


***Figure 12**: Token throughput measured by OpenRouter for DeepSeek and OpenAI models (Source: OpenRouter: 1, 2)*

[28] Token throughput includes total input and output tokens to different model providers. Please note that this data likely reflects usage via third-party providers of commercial and open-source providers, and does not include usage statistics from AI companies' own platforms. Based on DeepSeek's disclosure that it serves approximately 800 billion tokens per day, OpenRouter likely represents <1% of total traffic to DeepSeek models.

Data Acquisition

Acquiring large amounts of high-quality data to train generative AI models is imperative for both US and Chinese AI companies to remain competitive. Researchers have estimated that LLMs trained on increasingly larger corpora of human text will likely exhaust the availability of high-quality data between 2026 and 2032 based on known scaling laws. The strategic nature of this problem is evidenced by increased investments made by AI companies to license UGC, such as Google's reported $60 million per year deal with Reddit (with the latter reaching a similar agreement with OpenAI) and Google's deal with Stack Overflow, demonstrating the value of high-quality UGC for training models on human conversation and programming tasks.

Chinese companies are likely also using UGC to train generative AI models. For example, TikTok owner ByteDance has published several AI models trained to generate realistic videos of humans, such as OmniHuman-1 and CyberHost. The former's research paper does not disclose what type of video and audio data researchers used to train the model, only that the researchers obtained "18.7K hours of human-related data for training" and used public datasets like CelebV-HQ for test data. UGC is also valuable for providing context to AI models: Chinese social media network RedNote — which saw an increase in US users following a short-lived TikTok ban — reportedly uses UGC to provide context to its AI search product, Diandian ("点点"), similarly to xAI's Grok model.

Chinese AI companies train their models on English-language content and use shared internet crawl datasets with US AI companies such as CommonCrawl (and derived datasets, such as RefinedWeb and CCNet). Chinese researchers are also open-sourcing pretraining datasets with data primarily acquired from Chinese data sources, although there likely remains a significant gap in the size and quality of available Chinese-language datasets. In April 2024, academic researchers from the Fudan, Peking, and Shanghai Jiao Tong universities, in collaboration with international researchers, open-sourced the Massive Appropriate Pretraining Chinese Corpus (MAP-CC). The dataset, which contains 1.2 trillion tokens, of which 800 billion are Chinese tokens, was subsequently used to train Chinese-Tiny-LLM, "marking a shift towards focusing on the Chinese language for LLM development," according to the authors. For comparison, also in April 2024, Meta announced that it had trained Llama 3, the most capable open-source model at the time, using a training dataset of 15 trillion tokens, approximately twelve times the size of MAP-CC.

Open-Source Models

The development of open-source (or "open-weights") generative AI models is very likely an accelerant in the US-China AGI race. Open-source AI research and development is very likely viewed by both countries as a vector for economic competition and geopolitical influence. China's 2017 AIDP states that the country should "advocate the concept of open source sharing" (倡导开源共享理念) as a basic principal, including through the development of open-source software platforms.

US companies are sustaining increased pressure from Chinese competition to adopt open-source strategies. Meta has publicly declared that its decision to open-source its Llama models is due to the company's belief that "widespread adoption of American open-source AI models serves both economic and security interests." Similarly, OpenAI CEO Sam Altman declared shortly after DeepSeek's release that he believes OpenAI is "on the wrong side of history" and needs "to figure out a different open source strategy." US companies must, however, balance the decision to open-source models with the risks of proliferating dual-use technologies, as evidenced by the PLA's reported adoption of Meta's open-source Llama models.

Chinese companies are also adopting open source, marking a cultural shift across both sides of the Pacific. In recent interviews, the CEOs of MiniMax and DeepSeek have both declared their preference for publishing models in open source, with MiniMax CEO Yan Junjie (闫俊杰) stating that "open-sourcing can accelerate technological evolution" and DeepSeek CEO Liang Wenfeng (梁文锋) declaring that "closed-source moats are fleeting" and that open source "is cultural, not just commercial. Giving back is an honor, and it attracts talent." Established forums such as the Chinese Software Developer Network (CSDN) are becoming epicenters of open-source AI development and knowledge sharing among Chinese students and software engineers, including implementation guides and courses. CSDN also periodically runs polls to measure developer sentiment on new AI models and technologies, such as a poll on using DeepSeek R1 for software development. Developers found that, for example, DeepSeek R1 helps cut down development time on software projects, but the model currently lacks support from popular software integrated development environments (IDE)s.

Releasing open-source models can also be as profitable as retaining proprietary ownership. Unlike many other Western AI companies likely operating at a loss, DeepSeek disclosed that it had reached a theoretical cost-profit ratio of 545% per day across its AI services and applications while serving approximately 608 billion input tokens and 168 billion tokens over 24 hours. Spurred by DeepSeek's success and reported figures, new entrants to the Chinese AI market will therefore be more likely to adopt open source as a viable commercial strategy in order to compete with established closed-source players.

Additionally, open-source models accelerate the diffusion of advanced AI capabilities via model distillation. DeepSeek itself has admitted to using model distillation by using R1's outputs to boost the performance of other open-source models like Meta's Llama and Alibaba's Qwen models. Model distillation is also a growing risk for frontier AI companies as it allows challengers to use the outputs from more advanced "teacher" language models to train weaker "student" models, which can be subsequently released in open source. Microsoft and OpenAI alleged that DeepSeek used this technique to train R1, an open-source model, using OpenAI model outputs.

AI Conferences

As mentioned above, China is increasing its influence at prestigious AI conferences like NeurIPS. Submissions to such conferences also demonstrate that AI research is an increasingly collaborative process between China and the US. Co-authorship networks for papers accepted by NeurIPS between 2021 and 2024 demonstrate strong ties across US and Chinese entities (although collaboration between US and European entities at NeurIPS is likely higher). Reciprocal knowledge sharing and collaborative research are common and likely a result of increasingly shared talent pipelines. Research also demonstrates that cross-country collaboration on AI research has a greater effect than strictly collaborating domestically. Thus, cross-country talent movement and collaboration likely lead to a quantitative and qualitative increase in both countries' research and engineering pools, a dynamic that will likely continue narrowing the gap between both countries' AI capabilities.

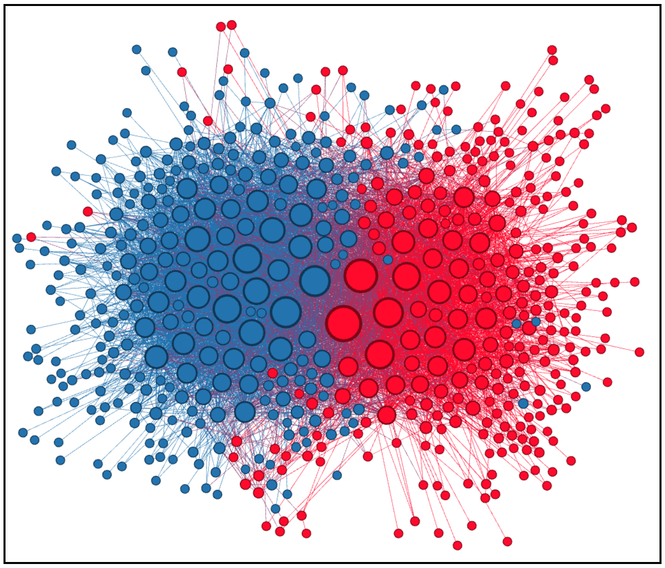


***Figure 13**: Co-authorship network, US (blue) and China (red), papers accepted to NeurIPS between 2021 and 2024 (Source: PaperCopilot — size denotes PageRank scores)*

The Compute Gap

Compute availability is almost certainly one of the single most important resources in the race towards AGI. Recent advances in LLMs stem from the discovery of neural scaling laws (such as the Chinchilla and Kaplan scaling laws), which infer that scaling the size of models and training datasets directly results in performance improvements. Scaling parameters and training data by a factor of 10x relative to previous models, among other innovations, resulted in the training of GPT-3 in 2020, the precursor to OpenAI's GPT-3.5 (the latter was the first model used for the company's release of ChatGPT in November 2022). Although companies are likely increasingly exploring training techniques that do not require ever-larger compute bandwidth — something that DeepSeek's approaches have shown the value of — the availability of computing resources like GPUs to train and host large and complex generative AI models at scale is still very likely a major bottleneck for AI companies, resulting in massive investments in GPU suppliers like Nvidia.

Since at least 2008, China's lag behind the US in leading-edge semiconductor production has averaged five years — a gap likely to persist due to significant barriers in manufacturing sub-7nm chips. Chinese foundries remain constrained by a lack of access to EUV lithography tools following US export bans in 2018. Export controls targeting China's AI development intensified in December 2024 by targeting "AI accelerators" like GPUs. Successive US export controls have almost certainly prompted the Chinese government to accelerate its funding of multibillion-dollar projects to develop the country's domestic semiconductor industry, HPC infrastructure, and energy grids. At the same time, the rise of reasoning models like o1 and R1 will likely create a greater interest in building low-latency inference infrastructure rather than large training clusters, which may threaten many of China's data center investments in rural regions.

Despite early signs of progress — including advances in chip yields and silicon wafer manufacturing, experimental EUV techniques, and improved DUV lithography — Insikt Group assesses that China's semiconductor industry is likely at least five years behind market leader TSMC, particularly in manufacturing sub-7nm chips. China is almost certainly working to bypass ASML's monopoly on EUV lithography technology by developing alternative techniques, although scalability to mass production remains unproven. As a result, China is unlikely to develop and scale domestic EUV lithography tools prior to 2030, matching assessments made by ASML CEO Christophe Fouquet.

Nevertheless, domestically produced AI accelerators have already entered the Chinese market. Huawei's domestically produced Ascend 910B and 910C chips, which reportedly rival Nvidia's H100 chips, entered mass manufacturing in 2024 and 2025 following improvements in chip yield rates. In June 2024, iFlytek announced its Xinghuo 4 model as the first Chinese LLM to be entirely trained on Huawei infrastructure.

Semiconductor Manufacturing

US export controls have almost certainly accelerated China's subsidizing of its domestic semiconductor and integrated circuit (IC) industry. The Chinese government's "Made In China 2025" plan, enacted in 2015, sought to increase domestic semiconductor production by 70% by 2025. To achieve this goal, in 2014, the Chinese government created the highly centralized 139 billion RMB ($21 billion) "Big Fund." A second phase committed an additional 204 billion RMB ($28 billion) in 2019. While government initiatives like the Big Fund have almost certainly accelerated the development of China's semiconductor industry, the sector will likely continue facing a bottleneck in the production of leading-edge chips without access to EUV lithography tools.

By February 2023, the Big Fund had minority stakes in at least 74 domestic semiconductor companies and invested in over 2,793 entities. According to the Semiconductor Industry Association (SIA), 43.4% (approximately $21.2 billion) of the first two phases of the Big Fund were invested in integrated device manufacturers (IDMs) such as Yangtze Memory Technologies Corp (YMTC), and 26.3% (approximately $12.8 billion) in chip foundries like SMIC and Hua Hong Semiconductor. In May 2024, the Chinese government announced the third phase of the fund, which will extend until May 2039 and is expected to spend 344 billion RMB ($47 billion) on areas like silicon wafer production. The Chinese government is also estimated to have provided below-market loans (at least $4.85 billion between 2014 and 2018) and helped establish "more than 15" local government funds worth $25 billion.

Chinese domestic chip production has very likely advanced closer to the leading edge of semiconductor manufacturing as a result of these investments. SMIC now produces 7nm chips for Huawei's Ascend 910B AI accelerator, which is likely now an increasingly viable alternative to Nvidia accelerators. Additionally, manufacturing capacity and efficiency are increasing: Huawei's AI chip yield, or the number of chips produced by wafer — a key semiconductor manufacturing efficiency metric — has reportedly increased from 20% to 40%, with plans to reach the industry standard of 60%. Huawei is also aiming to increase its production capacity of Ascend 910B chips by 50% (from 200,000 chips in 2024 to 300,000 in 2025) and begin manufacturing of newer 910C chips (100,000 projected in 2025), which reportedly deliver 60% of Nvidia's H100 performance according to DeepSeek researchers. Huawei reportedly claims its 910B chip "achieves an efficiency [of] 80% compared to Nvidia's A100" when training LLMs.

Despite these advances, Chinese companies likely still face a bottleneck in producing sub-7nm chips, while the West presses on with fabricating 2-nm chips. To manufacture sub-7nm chips, Chinese semiconductor companies need access to EUV lithography tools, 99% of which are manufactured by Dutch company ASML, and are subject to US export controls. According to ASML CEO Christophe Fouquet, US export controls banning the sale of EUV tools to China mean that it "will lag ten to fifteen years behind the West." However, this gap will almost certainly close over time: in 2024, Intel CEO Pat Gelsinger claimed that SMIC's production of 7nm chips shows that the Chinese foundry currently has a five-and-a-half-year lag on TSMC, and corroborated Fouquet's assessment that China likely has a ten-year lag on the leading-edge, with TSMC introducing 2nm (N2) process nodes in 2025.

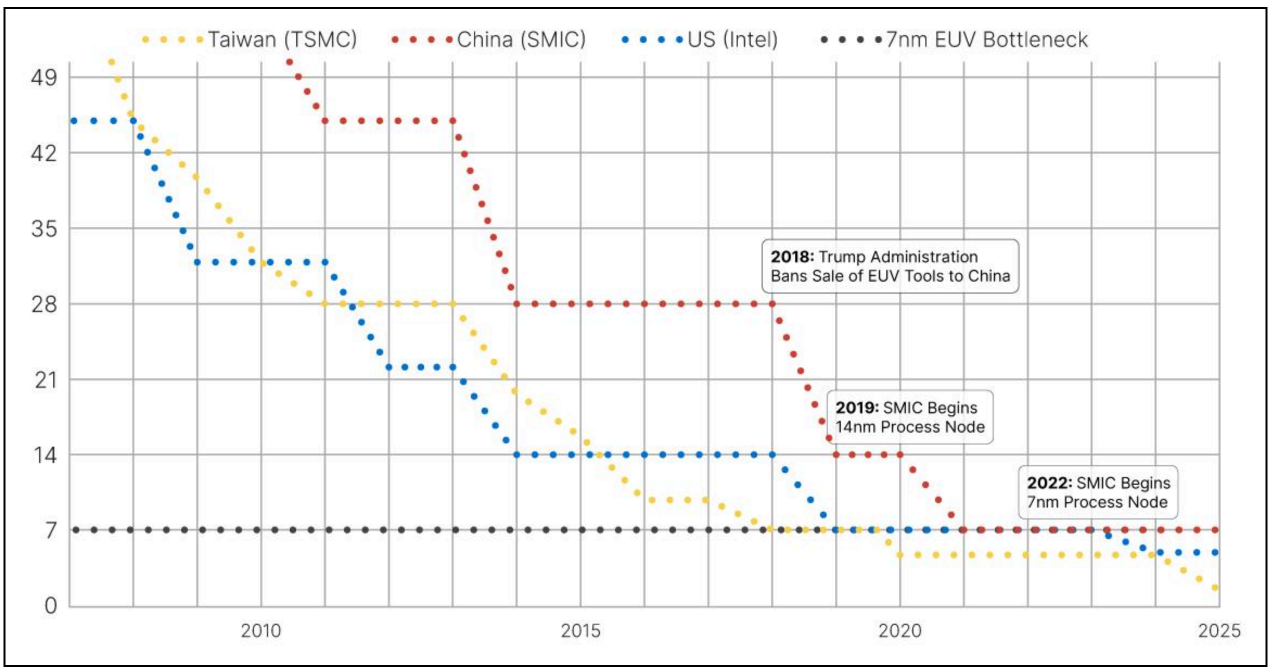


***Figure 14**: Year reached for integrated circuit manufacturing process by nanometer, TSMC, Intel, and SMIC (Source: Recorded Future, TSMC, WikiChip)*[29]

[29] Please note that process nodes are not definitive measures for progress in semiconductor manufacturing, and "sizes" have been decoupled from physical transistor gate lengths. Other factors such as transistor density and chip yields influence manufacturing performance. Process nodes in Figure 14 were collected as "equivalents" to TSMC nodes; for example, Intel's "10nm" process nodes have roughly equivalent density to TSMC's 7nm nodes.

High-Performance Computing

China is almost certainly deploying funding for HPC clusters to sponsor its domestic AI sector. In order to train LLMs, computing infrastructure must be closely co-located in HPC clusters in order to limit network latency and save on energy and cooling costs, which remain critical for such infrastructure to remain economically viable. US AI companies have begun building HPC "megaclusters" dedicated exclusively to training generative AI models: the largest current AI cluster, owned by xAI and dubbed "Colossus," uses 200,000 Nvidia H100s. The company is reportedly planning to expand to 1,000,000 GPUs.

Other US AI companies are almost certainly following suit: on January 21, 2025, OpenAI announced "Project Stargate," a $500 billion AI infrastructure project over the next four years; Meta CEO Mark Zuckerberg announced on January 24, 2025, that Meta is planning to spend over $65 billion overall on AI and reach over 1,300,000 GPUs by the end of the year; and Google is expecting to spend $75 billion on AI infrastructure in eleven data center regions. Both the US and Chinese governments are also competing on building increasingly large HPC clusters for use in national security applications, such as Lawrence Livermore National Laboratory's "El Capitan" supercomputer, which helps the US government oversee nuclear stockpiles, and China's Tianhe-3, developed by the National University of Defense Technology (NUDT).

According to SemiAnalysis and CSIS, by the end of 2025, the US will possess over 14.31 million AI "accelerators," while China will only have an estimated 4.8 million. CSIS argues, however, that China's ability to concentrate AI capabilities and deploy national-level industrial policy could enable it to rapidly catch up and surpass US capacity. China's "Eastern Data Western Computing" (EDWC, "东数西算") project, for example, attempts to concentrate HPC capabilities by connecting China's data-rich coastal regions in Eastern China to resource and energy-rich regions in Western China like Guizhou, Ningxia, and Gansu. The 43.5 billion yuan ($6.1 billion) project was first announced in 2021 and aims to deliver 300 exaflops of computing power by 2025 across eight regional data center hubs. As of August 2024, 63% of the projected 1.95 million racks (approximately 1.23 million racks) were occupied. Ten clusters across eight data center hubs were established in the following locations, with an estimated planned server capacity of over 3.6 million servers:

Clusters	Relevant Hubs
Horinger Cluster (和林格尔集群)	Inner Mongolia Hub (内蒙古枢纽)
Zhongwei Cluster (中卫集群)	Ningxia Hub (宁夏枢纽)
Qingyang Cluster (庆阳集群)	Gansu Hub (甘肃枢纽)
Tianfu Cluster (天府集群)	Chengdu-Chongqing Hub (成渝枢纽)
Chongqing Cluster (重庆集群)	Chengdu-Chongqing Hub (成渝枢纽)
Gui'an Cluster (贵安集群)	Guizhou Hub (贵州枢纽)
Zhangjiakou Cluster (张家口集群)	Beijing-Tianjin-Hebei Hub (京津冀枢纽)
Wuhu Cluster (芜湖集群)	Yangtze River Delta Hub (长三角枢纽)
Yangtze River Delta Ecological Green Integrated Development Demonstration Zone Cluster (长三角生态绿色一体化发展示范区集群)	Yangtze River Delta Hub (长三角枢纽)
Liuguan Cluster (留关集群)	Guangdong-Hong Kong-Macao Hub (粤港澳枢纽)

Table 5: *EDWC clusters and hubs (Source: Insikt Group)*

Over thirty companies have signed on to the project, including China Mobile, China Telecom, Huawei Cloud, Tencent Cloud, Alibaba Cloud, Kuaishou, Volcano Engine (owned by ByteDance), and Amazon Web Services China. Data centers built by these companies in EDWC hubs and their server capacity estimates are included in **Appendix A**.[30]

Part of these investments are almost certainly aimed at supporting generative AI development. Tencent's investments in the Yangtze River Delta Hub (长三角枢纽) are reportedly aiming to create a cluster of more than 800,000 servers intended to help train the company's HunyuanVideo generative AI models. The cluster will also likely host offices for other Tencent AI-related entities such as Keen Security Lab and YouTu AI research lab. Chinese companies are also making progress on improving data center network infrastructure for training generative AI models, although they are likely still relying on US hardware. In 2024, Alibaba Cloud engineers published a paper on Alibaba's High Performance Network (HPN), a "2-tier, dual-plane architecture capable of interconnecting 15K GPUs within one Pod," an alternative to Equal-Cost Multi-Path (ECMP) routing and used specifically to train LLMs like its Qwen models. The paper also specifies that Alibaba Cloud continues using Nvidia's NVLink technology, and possibly uses Broadcom's Tomahawk ASICs.

However, the rise of cost-efficiency models like DeepSeek R1 has likely reduced demand for major high-performance clusters. In March 2025, MIT Technology Review reported that data center utilization in China is only about 20% of capacity and that GPU rental prices have declined, citing Chinese media and sources. The emergence of the test-time compute paradigm (used by reasoning models like R1 and o1) means that AI companies will likely prioritize lower network latency to improve LLM performance during inference, rather than building massive HPC clusters for pre-training models. Top-down projects like the EDWC have likely led to an oversupply of computing in rural regions, given that AI companies

[30] *https://archive.is/3HVOS*

now likely prefer to be closely co-located to their inference infrastructure to reduce latency. Moreover, many of China's AI data centers are reportedly low quality, being constructed quickly and below industry standards.

Evasion of US Export Controls

There is almost certainly a growing black market for banned US AI hardware in China, and AI companies are almost certainly hoarding stocks of hardware acquired prior to export controls. According to SemiAnalysis, DeepSeek has access to an estimated 10,000 H800s and 10,000 H100s, both of which are now banned from exporting to China. According to CSIS and The Information, at least "eight distinct H100 smuggling networks" have continued to supply Chinese companies (which US officials allege includes Deepseek) with banned Nvidia chips since early 2024, using Singapore as a likely entry point.

Chinese companies have very likely exploited loopholes in export controls to acquire foreign equipment for their domestic semiconductor supply chain, such as wafer fabrication equipment (WFE). Huawei's spending on WFE reportedly grew 27% year-on-year and was estimated to reach $7.3 billion in 2024, making Huawei the fourth-largest spender on WFE. Huawei likely bypassed previous US export rules by investing in companies across the semiconductor supply chain that had not yet been listed on BIS's Entity List. Many of these entities were subsequently added in December 2024. The updated list stated that the entities contributed to Huawei's efforts to "support China's government's goal of indigenous production of 'advanced-node ICs [integrated circuits]' to support its military modernization." According to SemiAnalysis, unsanctioned entities and shell companies are frequently established in close geographical proximity with sanctioned entities, allowing the new, unsanctioned companies to freely import and transfer WFE to companies with known or alleged control from Huawei. Huawei has also been accused of using shell companies or proxies to directly import TSMC chips, some of which were reportedly found in some of its Ascend 910B chips, raising questions about whether Huawei has managed to fully solve domestic manufacturing.

Alternative EUV Techniques

Chinese companies are almost certainly attempting to develop their own EUV lithography tools to surpass the sub-7nm bottleneck using alternative techniques. In January 2025, researchers at the Harbin Institute of Technology (哈尔滨工业大学) announced that they had developed a "compact and efficient EUV light source" using laser-induced discharge plasma (LDP, "激光诱导放电等离子体"). SMEE reportedly filed a patent in March 2023 that plans to use a laser-produced plasma (LPP, "激光产生的等离子体") EUV source, a technique first pioneered by ASML and used in the company's commercial NXE and EXE-series machines. Chinese companies are also making progress in other sections of the lithography supply chain, such as DUV photoresists: Hubei Dinglong (鼎龙), Shenzhen Rongda (容大感光), and Xuzhou B&C (徐州博康) are edging closer to domestic production of high-end argon fluoride (ArF) and krypton fluoride (KrF) photoresists, further lowering China's reliance on imports. Other major Chinese photoresist manufacturers include Beijing Kehua (北京科华), Shanghai Sinyang (上海新阳), Suzhou Gaojing (苏州高晶), Ningbo Jiangfeng (宁波江丰), and Jiangsu Nata (江苏南大).[31]

[31] *https://archive.is/valtT*

DUV Production and Imports

China is also very likely attempting to bolster its manufacturing capacity for "trailing-edge" or "legacy" semiconductors, including chips that are 28nm or more according to the US CHIPS Act. While unlikely to be relevant to cutting-edge infrastructure for LLM training and inference, legacy chips remain ubiquitous in defense and automotive systems. Chinese dominance in this area will therefore likely remain a flashpoint in broader US-China economic competition. In December 2024, SMEE announced the start of manufacturing of its SSA/800-10W 28nm lithography machines, marking a milestone in domestic production of DUV lithography tools.

Between 2022 and 2024, China also accelerated its imports of DUV lithography equipment in preparation for US restrictions, accounting for up to 49% of Dutch lithography manufacturer ASML's revenue in Q1 2024. However, the company stated it expects Chinese revenue to decline to 20% following US export rules.

Mitigations

- Western AI companies should consider closely monitoring major developments by Chinese generative AI companies in addition to ecosystem-level indicators of public and private investment, as well as patents related to AI development.
- Companies should mitigate adversarial model distillation by monitoring for high-volume API queries from suspicious sources, such as proxies or IP addresses located in China.
- Companies should mitigate IP theft by implementing model watermarking, strengthening access control measures, and bolstering insider risk programs.
- Government agencies, academic institutions, and companies should adopt measures that support their country's ability to attract, recruit, and retain global AI talent.

Outlook

The Soviet Union's launch of Sputnik, coupled with the "missile gap" myth, almost certainly intensified the US–Soviet rivalry by prompting major federal investments in America's space program and ICBM capabilities. Similarly to Sputnik, DeepSeek R1's release has likely rapidly reshaped public perceptions of Chinese AI. In reality, DeepSeek and other Chinese AI companies' success likely reflect a near decade-old Chinese ambition to become the world's leading AI power by 2030. Beyond the hype, maintaining an accurate understanding of the US and China's respective AI innovation and diffusion capacities, and thus the state of US-China AI competition, will almost certainly stay crucial for steering future investments in US AI and R&D capabilities.

US-China AI competition will almost certainly intensify, particularly in relation to generative AI and LLMs. Unless US companies embrace open source and drastically lower their prices, Chinese models will almost certainly see increased adoption worldwide while making continued improvements. The US government and AI companies will likely use increasing competitive pressure from China to justify the deregulation of the US AI industry. Efforts to recruit global AI talent through current or reformed immigration pathways will likely be a point of contention between Silicon Valley and the Trump administration. While the US AI industry thus far has been mostly funded via private and VC investments, mounting pressure is likely to revitalize public investment in computing infrastructure and energy grids to support model training and inference.

China's ability to sustain AI innovation and diffusion will almost certainly depend heavily on its ability to scale its domestic semiconductor industry and overcome long-term bottlenecks in leading-edge semiconductor manufacturing, which requires access to technologies like EUV lithography. Even if China were to develop domestic EUV lithography tools by 2030, it remains unlikely to scale a hypothetical EUV capability rapidly enough to catch up with the manufacturing capacity of Taiwanese and US foundries and meet the growing demand for AI accelerators. If planned US investments in semiconductors, energy, and AI computing infrastructure come to fruition, the computing gap with China will likely continue widening by 2030. Investment in China's AI industry is unlikely to match or exceed such investment in the US, but innovative approaches, espionage, and an academic environment increasingly capable of cultivating high-quality talent will likely support China's continued AI advancement over the long term.

Maintaining a competitive advantage over China in AI will likely grow as a priority for the current and future US administrations, particularly as generative AI capabilities increasingly lead to concrete economic, military, and intelligence applications. However, it remains to be seen whether and to what extent the administration of US President Donald Trump will support AI with public funding. The administration has signaled that sustaining the US's AI leadership is a priority, but has also taken contradictory actions on that front. Further, regardless of how debates over high-skill immigration evolve, the Trump administration's methods for addressing US immigration challenges will likely dampen the interest of potential students and professionals who might consider participating in the US economy. Thus, the US's immigration advantage over China and other countries is likely to continue declining at an accelerated rate in the coming years. However, risks to the semiconductor supply chain — including the administration's tariff policies and the possibility of China using force to capture Taiwan — could undermine the US's AI industry.

Still, the US will very likely be spurred by perceptions that China's access to such capabilities will be used as leverage in diplomacy, coercion, or social control within its growing spheres of influence like the Digital Silk Road (DSR). Relatedly, the impact of AI on geopolitical power will only likely continue to increase. In the short term, the likely rise of autonomous AI agents in the next one-to-two years will almost certainly confer significant competitive advantages for countries that adopt and operationalize agentic capabilities as part of their national security apparatus and economies. In the medium-to-long term, a hypothetical winner of the race to achieve and adopt AGI will almost certainly wield significant geopolitical and economic power at a global scale.

Very likely with such economic and geopolitical benefits in mind, China embarked on an ambitious plan to become the world leader in AI by 2030. As of mid-2025, however, the country does not clearly lead the US on any industry support pillar examined in this report. Given the headwinds it faces, and assuming the US is able to maintain its own progress and advantages, China is unlikely to sustainably surpass the US on its desired timeline. The gap between China and the US will likely become tighter and China is likely to be at least a close second to the US. However, this race will almost certainly continue to be marked by unforeseen breakthroughs similar to the “surprising” release of DeepSeek. Areas to watch that likely have the potential to subvert this long-term assessment of China’s position by 2030 include Chinese or US companies developing new AI architectures and techniques that lead to a sea change in model training and performance, pioneering ways to overcome constraints on training data by using synthetic content, and achieving agentic and collaborative AI systems that enable emergent properties and capabilities, especially in the field of robotics.

Appendix A: Known East-West Data Centers

Company	Location	Hub	Reported Planned Server Capacity
Tencent Cloud	Gui'an Qixing (2021)	Guizhou Hub (贵州枢纽)	1,000,000[32]
Tencent Cloud	Huailai Ruibei	Beijing-Tianjin-Hebei Hub (京津冀枢纽)	300,000
Tencent Cloud	Dongyuan	Beijing-Tianjin-Hebei Hub (京津冀枢纽)	300,000
Tencent Cloud	Qingpu (2023)	Yangtze River Delta Hub (长三角枢纽)	800,000[33]
Tencent Cloud	Chongqing (2018)	Chengdu-Chongqing Hub (成渝枢纽)	200,000[34] [35]
Alibaba Cloud	Zhangbei (2016)	Beijing-Tianjin-Hebei Hub (京津冀枢纽)	100,000[36]
Alibaba Cloud	Ulanqab (2020)	Inner Mongolia Hub (内蒙古枢纽)	300,000[37]
Huawei Cloud	Ulanqab (2022)	Inner Mongolia Hub (内蒙古枢纽)	300,000[38]
Kuaishou	Ulanqab (2023)[39]	Inner Mongolia Hub (内蒙古枢纽)	300,000[40]

[32] *https://www.telecomreview[.]com/articles/wholesale-and-capacity/5386-largest-huawei-cloud-data-center-operates-over-1-million-servers*
[33] *https://www.yicaiglobal[.]com/news/20230510-05-tencent-takes-steps-toward-chinas-biggest-data-center*
[34] *https://www.datacenterdynamics[.]com/en/news/work-on-second-phase-of-tencents-chongqing-data-center-to-start-by-years-end/*
[35] *https://www.chinadaily.com.cn/a/201808/22/WS5b7d13eda310add14f387308.html*
[36] *https://clever-electronic[.]com/en/service/case/fcf7d54d-0141-4c77-956e-7c421102dc9c.html*
[37] *https://www.idcnova[.]com/html/1/59/69/687.html*
[38] *https://www.huaweicentra[l].com/huawei-cloud-debuted-five-data-centers-in-china/*
[39] *https://ir.kuaishou.com/news-releases/news-release-details/kuaishou-smart-cloud-ulanqab-data-center-has-been-officially*
[40] *https://baxtel.com/data-center/kuaishou-unlanqab-campus*

Appendix B: Monthly US-China Top Elo Scores

Elo scores are sourced from LMsys ChatArena.

Year/Month	US	China	Gap
2024-01	1195 (gpt-4-turbo, OpenAI)	1070 (yi-34b-chat, 01-ai)	125
2024-02	1187 (gpt-4-1106-preview, OpenAI)	1093 (qwen1.5-72b-chat, Alibaba)	94
2024-03	1253 (claude-3-opus-20240229, Anthropic)	1148 (qwen1.5-72b-chat, Alibaba)	105
2024-04	1260 (gpt-4-turbo-2024-04-09, OpenAI)	1153 (qwen1.5-72b-chat, Alibaba)	107
2024-05	1289 (gpt-4o-2024-05-13, OpenAI)	1238 (yi-large-preview, 01-ai)	51
2024-06	1287 (gpt-4o-2024-05-13, OpenAI)	1240 (yi-large-preview, 01-ai)	47
2024-07	1287 (gpt-4o-2024-05-13, OpenAI)	1240 (yi-large-preview, 01-ai)	47
2024-08	1316 (chatgpt-4o-latest, OpenAI)	1239 (yi-large-preview, 01-ai)	77
2024-09	1355 (o1-preview, OpenAI)	1256 (qwen2.5-72b-instruct, Alibaba)	99
2024-10	1336 (o1-preview, OpenAI)	1286 (yi-lightning, 01-ai)	50
2024-11	1365 (gemini-exp-1121, Google)	1287 (yi-lightning, 01-ai)	78
2024-12	1379 (gemini-exp-1206, Google)	1315 (deepseek-v3, DeepSeek)	64
2025-01	1381 (gemini-2.0-flash-thinking-exp-01-21, Google)	1357 (deepseek-r1, DeepSeek)	24
2025-02	1403 (early-grok-3, xAI)	1362 (deepseek-r1, DeepSeek)	41
2025-03	1411 (grok-3-preview-02-24, xAI)	1362 (deepseek-r1, DeepSeek)	49

Recorded Future reporting contains expressions of likelihood or probability consistent with US Intelligence Community Directive (ICD) 203: Analytic Standards (published January 2, 2015). Recorded Future reporting also uses confidence level standards employed by the US Intelligence Community to assess the quality and quantity of the source information supporting our analytic judgments.

About Insikt Group®

Recorded Future's Insikt Group, the company's threat research division, comprises analysts and security researchers with deep government, law enforcement, military, and intelligence agency experience. Their mission is to produce intelligence that reduces risk for customers, enables tangible outcomes, and prevents business disruption.

About Recorded Future®