INTERNATIONAL BUSINESS

The Great Divide: How the US and China Are Splitting the AI World

By Nikolaus Lang, Sylvain Duranton, Vladimir Lukic, Matt Langione, Rodrigo Ortiz Mena, Jona Lampert, and David Zuluaga Martínez

ARTICLE JUNE 30, 2026 15 MIN READ

The AI landscape today is dominated by the US and China, as both countries continue to put the weight of their private and public sectors behind the development of this technology.[1] But despite the superpowers' shared recognition of the importance of AI as a source of national power, their AI strategies have diverged. Each country is developing an AI technology stack designed to reduce its dependence on the other—which also means their stacks are increasingly incompatible. CEOs and policymakers across the globe may have to choose sides sooner than they think.

In 2024, we argued that the US was the clear leader in the burgeoning AI race, with China quickly gaining ground. Our updated, 2026 analysis across the six key enablers of AI supply—capital, talent, intellectual property (IP), data, energy, and compute—reveals the US has maintained its lead, fueled by its strength in talent and capital deployment. (See "Methodology.") But China has continued to close the gap on compute power and has systematically translated its IP strengths into a consistent "fast follower" approach to frontier model development, with clear focus on rapid adoption of AI across the real economy. (See Exhibit 1.)

– Methodology

Our research draws on an extensive quantitative comparison of the relative strength of the US, China, and select middle powers across the key enablers of AI supply (particularly of generative models and agentic systems built on them). We assess each country's strengths across 19 variables spanning the six key enablers of AI supply, drawing on a combination of public and proprietary data and analyses. These variables were selected as indicators or proxies for the national characteristics most critical for the development of AI.

Our analysis of comparative country strength in AI stretches back to 2024, when the geopolitics of AI were gaining global relevance. Our 2026 analysis reflects updated data and, in some cases, revisions that reflect the shifting center of gravity in the development trajectory of the technology.

Since our primary objective was to develop a relative sense of strength across enablers, we developed normalized scores by enabler and country or region in the following way:

- Each indicator leads to an absolute value by country or region. These values are not normalized by population, size of the economy, or other such factors, because competition in the supply of AI is largely a function of scale.
- Values for each indicator are then (linearly) normalized into country or region scores on a scale from 0 to 1, where 1 equals the highest actual value in our data set and 0 is set as absolute 0. Two indicators use a different anchor: for industrial electricity price, 0 is set at the lowest actual price; for mobile-broadband affordability, 0 is set at the UN Broadband Commission target of 2% of gross national income (GNI) per capita.
- For each country or region, we then take the average normalized score across the indicators associated with an enabler to generate the enabler score. A country would have a score of 1 on an enabler only if it had the

highest absolute value of all countries in our data set on every indicator associated with the enabler in question.

The following are the indicators we used to develop the stage-setting analysis of relative strength across the enablers of AI supply, with the sources for each indicated in parentheses. For further details on the analysis and data used in this study, please contact the authors.

Capital

- Venture capital funding from 2019 to 2026 (year-to-date), based on the observed AI-directed share of investments by venture capital funds, by country. In the case of China, this includes the sizeable pool of government VC funds devoted to AI. (PitchBook; Martin Beraja et al., "Government as Venture Capitalists in AI," NBER working paper, and personal communication with the authors)
- Corporate research and development (R&D) spending by the 20 largest technology companies, by country. (European Commission)
- Capital expenditure (CAPEX) by the same set of technology companies. (S&P Capital IQ, company reports)
- Sovereign wealth and public pension-fund investment power, adjusted for the share of assets under management allocated to equities and alternative investments (therefore excluding bonds, real estate and infrastructure investments, and risk-free assets). Whereas VC investments, corporate R&D, and CAPEX approximate actual spend on AI, this indicator widens the aperture to the country-level (i.e., sovereign) capital pools that could support strong national bets on AI. (Sovereign Wealth Fund Institute, sovereign wealth fund reports)

Talent

- Share of the top 2,000 AI researchers worldwide, based on the institution-affiliation of authors of leading AI publications. (AMiner)
- Share of the top 300 AI institutions. (AMiner)
- Size of the AI-specialized talent pool working within a country based on AI-related job titles. (LinkedIn)

IP

- Share of the top 100 most-cited AI publications, 2019–2025. This indicator illustrates countries' contributions to breakthrough research. (OpenAlex)
- Share of the top 10% most-cited AI publications, 2019–2025. This indicator captures countries' contributions to highly influential research more broadly, beyond the handful of landmark papers. (OpenAlex)
- Share of notable machine learning models developed since 2019 based on the country/countries associated with the developing organization(s). (Epoch AI)
- Average frontier-model capability, May 2025 through April 2026. This indicator uses the Epoch AI Capability Index, which combines scores across many benchmarks to allow models to be compared over time even as individual benchmarks reach saturation. (Epoch AI)
- AI patent quantity, measured as the number of AI patent families, 2020 to present. (LexisNexis PatentSight)
- Average AI patent quality, calculated by combining a patent's market coverage and its technological relevance based on the number of citations it receives from later patents, corrected for patent age. (LexisNexis PatentSight)

Data

- Total number of active handset-based and computer-based (i.e., connected by USB/dongle) mobile-broadband subscriptions, as proxy for relative magnitudes of digital data generation. (International Telecommunication Union)
- Mobile-broadband affordability, measured as the cost of a data-only 5 GB basket as share of gross national income (GNI) per capita in 2025. (International Telecommunication Union; UN Broadband Commission)

These indicators serve as a proxy for the quantity of local data available, through the number of mobile-broadband subscriptions, and the amount of data produced by each individual, through the mobile-broadband affordability measure, which are important enablers for the development of locally optimized AI models. The data availability score is the product of the normalized subscriptions and normalized affordability values. While these metrics are

directionally indicative of the degree of digitization and the volume of (digital) data produced in each country, they do not, however, account for other important factors, such as the regulatory flexibility of data uses or the level of contextualization of data (i.e., how easily data of different types and sources can be used in an integrated fashion).

Energy

- Cost of electricity for a typical commercial/industrial user per kilowatt-hour. (Global Petrol Prices; Eurostat; GOV.UK)
- Lead time to grid connection for a new 50 MW data center. (International Energy Agency, AI Israel, IESO)

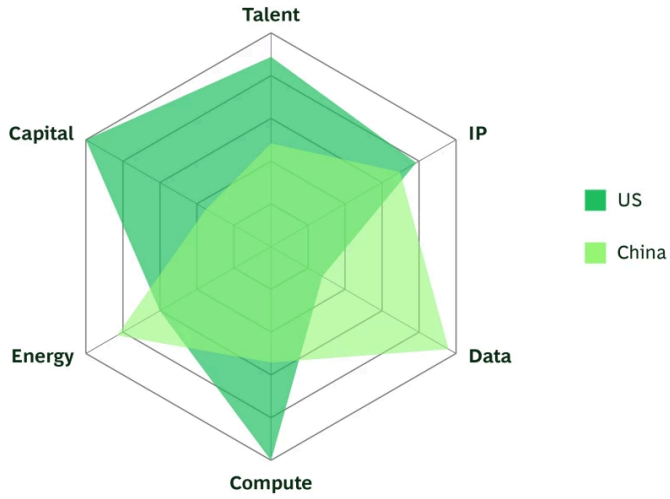
Access to affordable energy is among the biggest bottlenecks for data center buildouts, as shown by the delays countries like Japan and the US have faced in getting new capacity built and running. These two indicators capture both sides of the constraint: availability, via grid connection lead times, and affordability, via the cost of electricity.

Compute

- Existing data center capacity, including hyperscaler, colocation, and enterprise facilities, measured in gigawatts. While not all data center capacity is optimized for AI workloads, as noted in footnote 3 of the text, internationally comparable data on AI-specific compute capacity is not publicly available; we therefore rely on total data center capacity as a proxy for countries' AI compute capacity. (datacenterHawk, CBRE, S&P Global)
- Access to cutting-edge semiconductors optimized for AI workloads, such as NVIDIA's Blackwell and H200 chips. Access is scored 1.0 for countries and regions with no formal barriers, and 0.75 or 0.5 for progressively tighter levels of export-control restriction. NVIDIA's chips remain the industry gold standard, so regulatory barriers to access serve as a useful proxy for a country's ability to build out AI-optimized compute capacity. (US Department of Commerce)

EXHIBIT 1

The US and China Are the AI Superpowers



Source: BCG Institute analysis.
Note: See methodology sidebar for details on indicators by enabler.

As the US and China pursue AI sovereignty, the effects of their bifurcating technology stacks will be felt across geographies and sectors—defining the options leaders have available to engage with AI. In our 2024 study, we found that countries and companies could still mix and match across both superpower stacks. Since then, the two ecosystems have diverged to the point where neutrality may soon become harder to sustain, which changes the calculus. For most countries, the question is how best to develop AI resilience by ensuring robust access. AI is no longer something countries compete for, but something they compete *with*. For senior executives, the choice of which AI stack to use increasingly determines where a company can operate and how exposed it is to geopolitical volatility.

The US and China: The Great Bifurcation

In late 2024 we argued that China had effectively reached parity with the US on AI model performance. Then, in January 2025, the US announced "Stargate," a $500 billion AI infrastructure joint venture led by OpenAI, Oracle, SoftBank, and MGX, while China released its highly capable and low-cost DeepSeek R1 model.[2]

These events were early indications of the increasingly distinct AI ecosystems that the US and China are committed to developing and expanding. China is increasingly focused on foundational research (publications, patents), domestic chip development, and real-economy adoption (deployment across industry and public services). By contrast, the US dominates the model and infrastructure layers, pouring capital into frontier models and the compute power required to serve them.

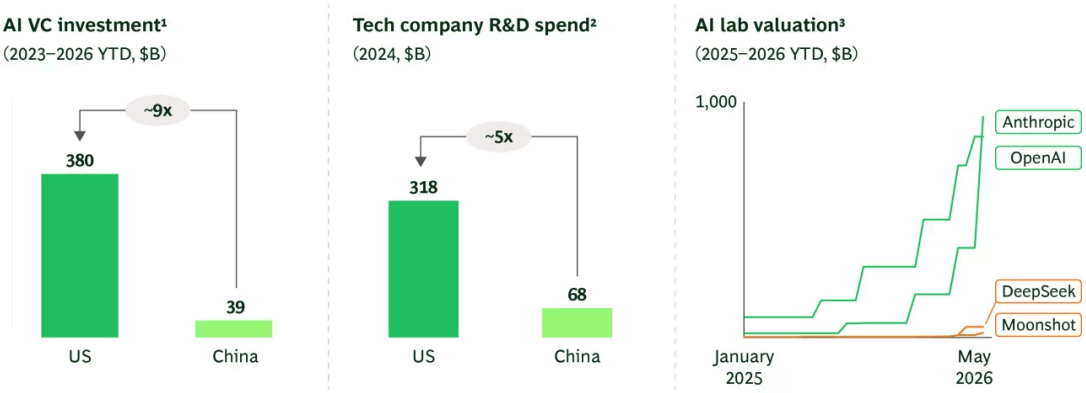
The US: Building Data Centers and Top-Performing Models with Big Capital

The US strategy can best be described as winning through scale. In practice, this has meant an acceleration in capital deployment over the past 18 months, with the center of gravity shifting from frontier model development toward the data center infrastructure needed to meet growing inference demand from enterprise uptake and the rise of AI agents.

Even with this shift, significant capital continues to flow into the development of frontier models. Since 2023, US-based startups have raised around $380 billion in AI-related venture capital, and incumbent tech giants invested more than $300 billion in R&D in 2024 alone. A sizeable portion of those funds is invested in AI and adjacent innovations. The depth of these capital pools exceeds every other country's and is reflected in the valuations of frontier model developers. (See Exhibit 2.)

EXHIBIT 2

The US Leads in Capital Deployment for AI

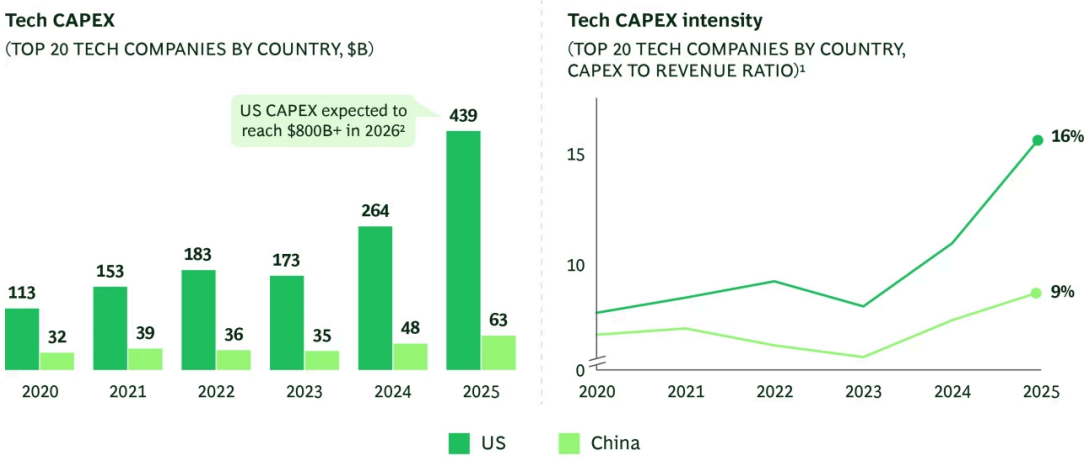


Source: BCG Institute analysis.
¹Values for China based on PitchBook data as baseline plus consideration for the government's role as VC investor based on the NBER working paper "Government as Venture Capitalists in AI" by Martin Beraja et al.; $185 billion invested by the Chinese government 2000–2023 has been distributed across the years using the same distribution that results from the PitchBook dataset; for years beyond 2023 we assumed the same growth/decline for government VC as results from the PitchBook data; the resulting decline in VC between 2023–2026 YTD compared to the prior period is consistent with an overall decline in the Chinese VC landscape as also reported by the OECD, PitchBook, and Preqin.
²By top 20 tech companies in each country.
³Based on most recent valuation in private markets through June 1, 2026.

What's new is the sheer scale of spending by US tech giants on the infrastructure to serve those models: CAPEX by the top technology companies surpassed $400 billion in 2025—compared with $63 billion in China—and is estimated to exceed $800 billion in 2026. (See Exhibit 3.) That spending is expanding what is already the world's largest data center base with 50+ GW of capacity at the end of 2025, compared to 31 GW in China and 12 GW in the EU.[3] This extraordinary level of investment is responding to skyrocketing demand. For example, Google's monthly inference volume grew about 50-fold year-on-year, and the three largest US cloud providers now hold more than $1.4 trillion in contracted future revenue, more than double a year ago and in significant measure driven by AI.

EXHIBIT 3

The US Speeds Ahead on Tech CAPEX

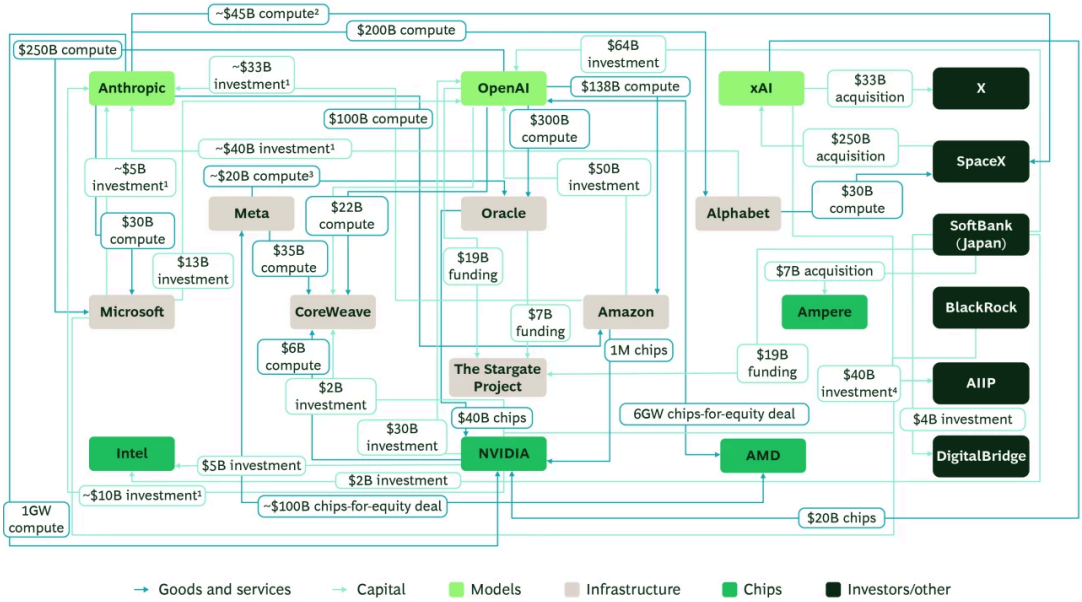


Source: Capital IQ; company quarterly earnings and annual reports (Huawei, Oracle, Alibaba); BCG Institute analysis.
Note: US companies are Apple, Microsoft, NVIDIA, Amazon, Meta, Alphabet, Broadcom, Salesforce, Oracle, Adobe, ServiceNow, Accenture, IBM, AMD, Cisco, Qualcomm, Texas Instruments, Danaher, Intuit, Palantir; Chinese companies are Tencent, CATL, NetEase, Luxshare, Hikvision, Cambricon, NAURA, Will Semiconductor, SMIC, Techtronic, ZTE, DiDi, Kuaishou, Inovance, Hygon, Huawei (Huawei CAPEX from "Additions" line in Note 14 (PP&E) of annual reports; revenue from annual reports. CNY converted at FRED annual average rate), Alibaba, Xiaomi, Baidu, NARI.
¹Weighted average across companies.
²Based on latest company guidance as of June 1, 2026, and estimated growth rate where company guidance not available.

The seemingly unstoppable rise in both CAPEX and VC investment in AI reflects the tightly integrated network of cross-investments and commercial commitments that has come to characterize the US tech ecosystem. Since 2024, deals worth more than $1.5 trillion have been announced between US AI labs, chip designers, hyperscalers, and capital providers: equity stakes, multi-year compute contracts, and chip purchase agreements that link the major players to each other on multiple sides at once. (See Exhibit 4.) These investments are part of a concerted effort by big tech companies to ensure that suppliers across the AI value chain have the necessary capital to keep up with skyrocketing demand. While these integrated relationships and capital commitments are based on real, rapidly growing revenues, they create systemic risk in the US tech ecosystem: a default, write-down, or supply disruption at any major player could ripple quickly across the rest.

EXHIBIT 4

The Self-Reinforcing US AI Ecosystem



Source: BCG Institute analysis.
Note: As of June 1, 2026; non-exhaustive illustration of US ecosystem; includes deals that span multiple years into the future; companies may fall into more than one category; includes deals above $1B; not all companies shown are US companies; equity stakes sized at initial value of investment, not current value.
[1]Announcements of "up to" these figures.
[2]$1.25B per month until May 2029.
[3]Media reported figure, deal size not officially confirmed.
[4]$40B investment includes other investors, e.g., MGX.

The inflow of capital to AI has been aided by the US government's emphasis on promoting bundled, exportable "AI stacks" of chips, models, cloud, and software. While maintaining targeted restrictions, US policy is more focused on outpacing through speed, scale, and ecosystem reach.

The US's light-touch regulatory strategy, however, has been tested by the arrival of increasingly powerful models with national security implications. In April 2026, Anthropic announced it was delaying the public release of its Claude Mythos model to allow the US government and select companies to prepare for the model's cyber capabilities. A version of Anthropic's model with built-in safeguards was released to the public in June 2026; days later, the US government issued an export control directive for Anthropic to suspend all access to the model by any foreign national, whether inside or outside the US (with the net effect that the model had to be suspended for all customers), indicating that national security concerns may in some instances take precedence over the global diffusion of the most capable AI models.

China: Prioritizing Adoption of Cost-Optimized AI

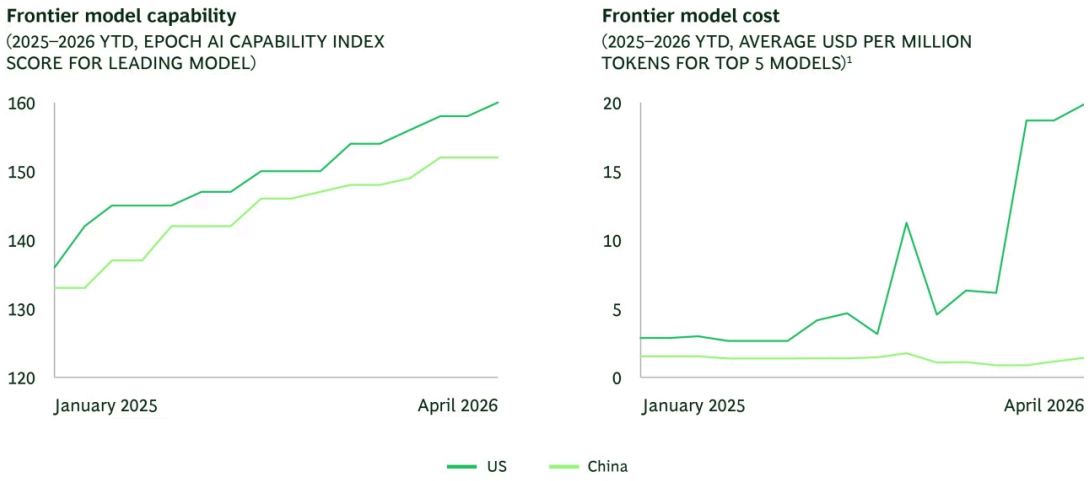
China is pursuing a visibly different trajectory. While its investments remain sizeable, it is not mobilizing a comparable share of its economy to match the US at the frontier of model

development or data center expansion. Instead, its strategy has focused on accelerating domestic adoption of AI while decreasing its reliance on foreign chips.

This shift is premised on China's ability to remain competitive at the level of foundation models despite its limited access to the highest-end computing power. Indeed, since the release of DeepSeek R1 in January 2025, top-performing Chinese models have kept up with US leaders. (See Exhibit 5.) Importantly, this isn't just about DeepSeek; China has consolidated a proper ecosystem of AI startups and big tech incumbents developing models—including Alibaba, ByteDance, Moonshot, Z.AI, MiniMax, and Xiaomi. The open-weight nature of most Chinese models has helped consolidate this ecosystem, as labs build directly on each other's weights and architectures.

EXHIBIT 5

China Builds Powerful, Cost-Optimized Models

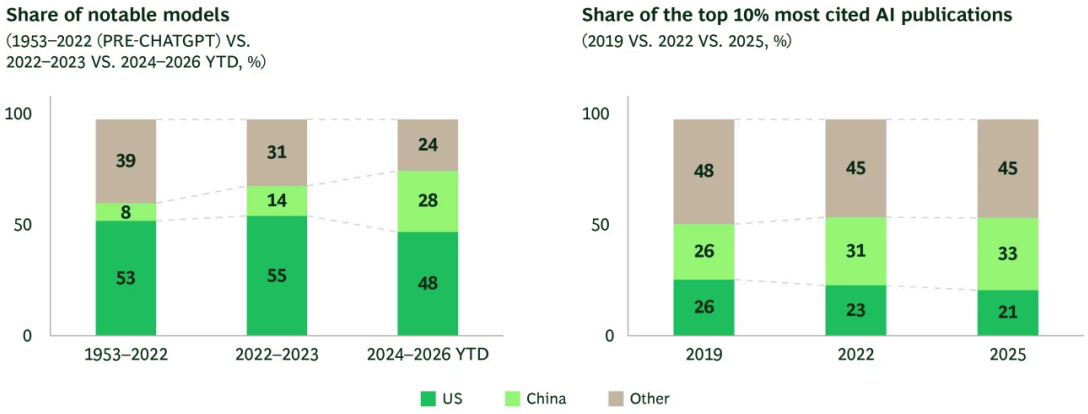


Sources: Epoch AI; Artificial Analysis; BCG Institute analysis.
[1]Assuming 3 to 1 ratio of input to output tokens, which are differently priced.

China has stayed close to the US on model quality benchmarks while making its models multiple times cheaper than US counterparts, driven by open weights for most leading models and architectural efficiencies from novel model engineering. Moonshot's Kimi K2.6, China's top model in May 2026, runs at $1.71 per million tokens, versus $11.25 for GPT-5.5, the US's top model. These frontier advancements are underpinned by a deep and improving research base: China now accounts for roughly one third of the world's top 10% most-cited AI publications, ahead of the US, and is the clear leader in the number of AI patents, a reflection of its focus on economic applications of AI. (See Exhibit 6.)

EXHIBIT 6

China's Share of Models and AI IP Is Growing



Sources: Epoch AI; OpenAlex; BCG Institute analysis.
Note: Epoch AI defines notable models as those that meet any of the following criteria: state-of-the-art improvement on a recognized benchmark, highly cited, historically relevant, and/or enjoying significant use. Figures may not add up to 100% due to rounding.

In parallel, China is now focused on developing a homegrown supply chain to lessen reliance on overseas chips. Even as the US has allowed more exports of high-end chips to China, major Chinese tech firms, including NVIDIA's largest Chinese customer, ByteDance, have not been allowed to deploy these chips. Additionally, state-funded data centers are now required to use only domestic chips. The bet is that induced reliance on Huawei's Ascend series will accelerate progress on domestic design and manufacturing of frontier semiconductors. The first proof point arrived in April 2026, when DeepSeek released V4, optimized for inference on Huawei's Ascend chips—though still trained at least partially on chips designed by NVIDIA.

This trend may accelerate the bifurcation of AI technologies all the way down to the hardware layer. More cost-efficient Chinese models running on cheaper Chinese chips could become an attractive package for any country looking to deploy AI affordably.

China's ability to export its AI stacks is bolstered by its close trade and investment partnerships with countries in the Global South: China is now the primary trading partner of 78 countries in the region, a ~50% increase since 2015, and has invested heavily in infrastructure through its Belt and Road Initiative, building, financing, or operating over a third of Africa's commercial ports. For countries that already rely on China to finance their debt and to build and operate their ports, highways, and airports, adopting a Chinese AI stack may be the path of least resistance. Even for countries with close ties to the US, the cost-efficiency of the Chinese stack may become too attractive to ignore, especially as AI costs have emerged as a key concern for many corporations, one that will only grow with the prevalence of token-intensive agentic workflows. The results are beginning to show at the model level: In late 2025, Chinese open-source models surpassed US ones in total downloads.

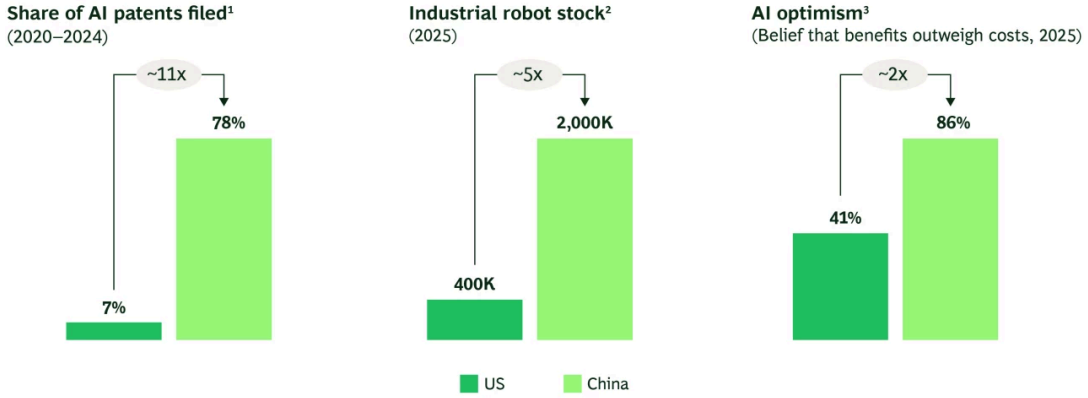
This emphasis on cost-optimized AI is at the core of China's strategy to aggressively push for the diffusion of this technology across its economy, anchored in the government's "AI+" initiative. Introduced in 2024, the "AI+" initiative explicitly targets the large-scale integration of AI into

manufacturing, services, public administration, and everyday life, with policy goals of more than 70% penetration of AI-enabled "intelligent terminals" and agents by 2027 and more than 90% by 2030.

Robotics is one of the clearest physical expressions of this push, with China installing 54% of the world's industrial robots in 2024 and naming embodied intelligence as a key focus industry in its latest five-year plan. While China's robot density is still half of the US level, it is expected to overtake the US by 2030 at current growth rates. China has also experienced a rapid uptick in domestic AI usage, projected to process over half of the world's tokens in Q2 2026 despite making up only ~17% of the population. This impressive adoption speed is driven in part by an optimistic view of AI: more than 85% of China's population agrees that products and services using AI have more benefits than drawbacks, compared with less than 45% in the US. The result is a fundamentally different trajectory than that of the US: less about pioneering technical developments at the frontier, more about rapid adoption to capture productivity gains throughout the real economy. Ultimately, China's combination of a low-cost model stack and rapid diffusion may prove self-reinforcing, accelerating growth for Chinese AI. (See Exhibit 7.)

EXHIBIT 7

China Is Betting on Accelerated AI Adoption



Sources: Ipsos (via Stanford HAI); International Federation of Robotics; LexisNexis PatentSight; BCG Institute analysis.
[1]Based on location of inventor of patent family.
[2]In part driven by the industry mix of the respective economies: manufacturing share of the economy is ~25% in China and ~10% in the US.
[3]Percentage of respondents agreeing that "products and services using AI have more benefits than drawbacks."

The Anatomy of Stack Choice

The diverging US and Chinese strategies are pulling their AI ecosystems into two increasingly incompatible stacks. Inside the two superpowers, the room for companies and governments to mix across the two has almost closed; in the rest of the world, it is starting to shrink—faster than many companies might be ready for.

The AI technology stack has four layers, dominated by major players in the US and China, that work together to power any AI solution. For example, when you ask a chatbot a question, you are interacting with the application at the top of the stack, which passes your request to the underlying model that processes your question and generates an answer. That model is hosted by a cloud, which sends your question to the model and returns the answer to your screen. The cloud, in turn, runs on specialized chips housed in data centers.

Across all four layers of the tech stack sits a governance and security overlay—data rules, privacy constraints, export controls, procurement conditions, and national-security priorities—alongside the technical characteristics of each. Together, these political and technical constraints increasingly determine which layers can be combined, where they can be deployed, and how easily systems can move across borders.

Inside the US and China, at the model layer, frontier systems from one superpower are either unavailable to customers in the other or coming under increasing political scrutiny. At the cloud layer, the market is dominated by domestic companies, with negligible share for outside providers. At the chip layer, a company that builds on NVIDIA's CUDA software cannot seamlessly run on Huawei's Ascend chips, since most production AI code would need rewrites to work on Huawei's CANN software.

For multinational corporations operating in China, adopting separate stacks has become an existential imperative, illustrated by Apple partnering with Alibaba for its "Apple Intelligence" in China while partnering with OpenAI in all other markets.

While open-weight models (particularly Chinese ones) may still be fungible across stacks, it is reasonable to expect that each ecosystem will tend to optimize for its own hardware. Over time, this is likely to increase the engineering cost of crossing over, even when nothing formally prevents it. As a result of such self-reinforcing dynamics, companies may have to start making choices at the level of end-to-end stacks rather than layer-by-layer.

Companies outside the superpowers retain greater ability to mix tech stacks, as they typically face fewer regulatory hurdles and the technical costs of doing so remain manageable (for now). However, as the geopolitical competition intensifies for AI-driven industrial advantage and offensive cyber capabilities, the superpowers are likely to exert increasing political pressure on third countries to pick a side.

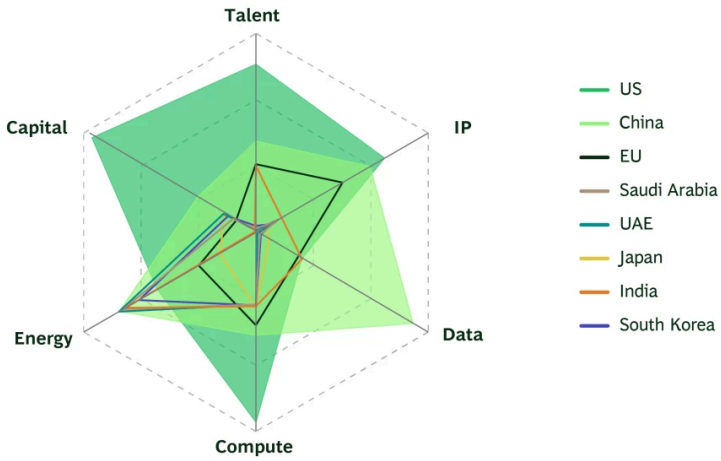
Middle Powers: Limited Choices and Diverging Strategies

The rest of the world is paying close attention to the development of the US and Chinese stacks. For most countries, decisions are focused on which layers of the stack to procure or even allow from which companies (and, by implication, the superpowers with which they are aligned). But for a select number of countries, the deepening bifurcation fuels the more ambitious goal of creating domestic alternatives for some layers of the tech stack to avoid a forced choice of superpower on which to become dependent.

We've gauged the potential of a number of middle powers by looking at how they fare across the same six enablers of AI supply. (See Exhibit 8 and Exhibit 9.) Their relative strengths—and limitations—inform their different strategies for building leverage within the AI tech stack.

EXHIBIT 8

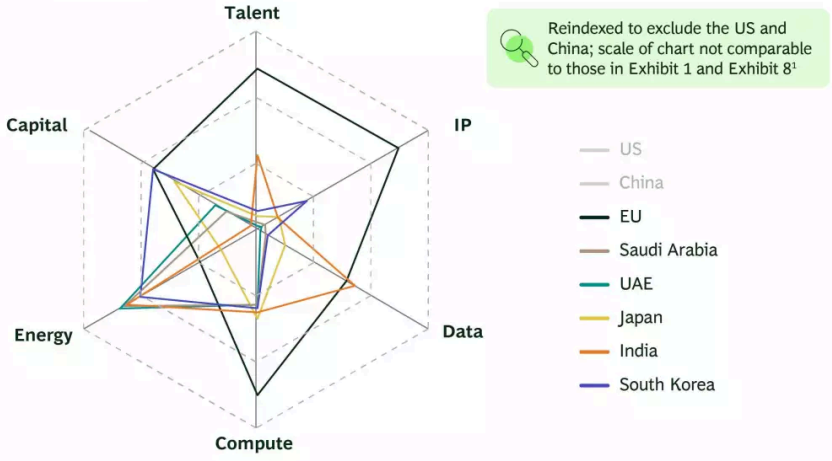
AI Middle Powers Lag Behind the US and China



Source: BCG Institute analysis.
Note: See methodology sidebar for details on indicators by enabler.

EXHIBIT 9

The EU Is Strongest Among Middle Powers



Source: BCG Institute analysis.
Note: See methodology sidebar for details on indicators by enabler.
[1]The chart is not evenly scaled to the versions in Exhibit 1 and Exhibit 8, as each dimension is independently reindexed to the highest value in our dataset excluding the US and China.

The EU is primarily focused on two goals: backing its champion model developer, Mistral, and securing sovereign compute capacity that doesn't depend on US hyperscalers. Mistral is continuing to improve its models, and coordinated efforts are starting to scale European compute. But the gap compared to the US is wide on both counts. As of early 2026, Mistral's revenue, valuation, and total capital raised each sat at roughly 2% of OpenAI's, and while its model capabilities have remained ~3-10 months behind the frontier, Mistral currently ranks 16th in AI lab rankings.[4]

Compute capacity faces the same problem, with US hyperscaler CAPEX for 2026 alone projected to be more than three times larger than the EU's multi-year InvestAI program (deploying €200 billion alongside the buildout of 19 AI Factories—shared AI-optimized supercomputers with additional support services—by 2027). However, Europe could become a pole for other like-minded countries seeking an option outside the US and Chinese stacks but lacking the scale to build one alone, as suggested by the newly announced $20 billion combination of Cohere and Aleph Alpha, Canada's and Germany's leading enterprise AI labs.

Japan is buying a seat at America's table. Constrained by limited data center capacity and without a domestic market the size of Europe's to feasibly sustain a domestic model ecosystem, Japan is relying on capital to influence the US ecosystem. Domestic efforts like Rakuten's Japanese-optimized AI models remain targeted in scope, geared toward local language and cultural context rather than generalist, frontier capability. SoftBank has participated in roughly $90 billion of AI dealmaking since 2024, including more than $60 billion invested in OpenAI, securing privileged inclusion in the US-led ecosystem. They are expanding their strategy beyond the US as well, with SoftBank planning to invest "up to" $87 billion in AI infrastructure in France. Potential IPOs by leading US AI players could reshape this strategy, increasing the financial value of investments while potentially reducing their strategic leverage within the ecosystem.

The UAE and Saudi Arabia are pursuing ambitious AI strategies to attract top technical talent as well as develop homegrown models. Most critically from a geopolitical standpoint, they are positioning themselves as critical infrastructure hubs, replicating the Gulf's playbook in trade, logistics, and finance. The UAE is scaling regional compute through US-linked bets, most notably a 5 GW AI Campus in Abu Dhabi. Saudi Arabia is moving along the same lines, announcing roughly $33 billion in investments under Vision 2030 through its PIF-owned HUMAIN vehicle, anchored by an NVIDIA partnership for hundreds of megawatts of AI factories. The current conflict is testing and highlighting the criticality of infrastructure resilience.

India is large enough to refuse the binary, at least for now. It is the second largest market for both OpenAI and Anthropic—with more than 100 million weekly ChatGPT users—giving it real commercial weight with AI companies. It is using that position to assemble capabilities across multiple ecosystems at once—working with the US on infrastructure, the EU on governance, and the UAE on compute, while engaging China, albeit to a lesser degree, through multilateral platforms such as BRICS. The result is active engagement across competing ecosystems rather than alignment with any one ecosystem. But the costs of maintaining ties across ecosystems, politically and economically, may mount as AI stacks become increasingly self-reinforcing.

A separate group of economies is working from narrower strengths. For example, South Korea controls roughly 80% of the global supply of high-bandwidth memory (HBM), a specialized memory component that is packaged together with virtually every AI logic chip made by Taiwan's TSMC. The UK, meanwhile, pairs world-class talent—ranking third on number of notable machine-learning models and impactful AI papers since the 1950s—with a dominance in a key niche of the value chain through Arm, whose chip design IP is used in roughly half of the CPU compute at the top hyperscalers. The UK has also been home to several significant AI startups, namely DeepMind, but a number of these have been acquired by larger US companies. These are narrower bets than Europe's or the Gulf's, but they are bets the superpowers cannot easily route around.

From Geopolitical Exposure to Resilience

The walls between the US and Chinese AI technology stacks are going up faster than most companies are reorganizing around them. A stack choice that looks like cost optimization today can become a strategic liability tomorrow. CEOs navigating this landscape should consider three key questions.

- **How does your company engage with the US and China stacks?** A firm whose customers, data, and regulators sit primarily in one superpower will likely follow the local stack—for US companies operating in the US and Chinese companies operating in China, the path, at least for now, is clear. A multinational operating meaningfully on both sides may find that running two distinct AI stacks is the only workable path. A company based in neither superpower faces a more open choice, but one that increasingly comes with strings, as procurement rules, export controls, and national strategies shape which combinations are viable. While the window remains open, there is an advantage to exploring both stacks, as different stacks may be optimal for different markets. This is particularly relevant for European companies that may be wary of overdependence on US technology. The stakes are even higher for companies whose products or services have been optimized for specific models: switching may be technically costly or regulatorily infeasible.
- **How does your company engage with non-US/China technology?** The US and China will continue to dominate frontier models and large-scale compute power, but real alternatives exist elsewhere in the stack: cost-optimized or custom models for specific use cases (particularly when off-the-shelf models underperform), locally tailored applications, domestic cloud services, and regional compute capacity. For companies based outside the two superpowers, consider whether non-superpower alternatives for each tech stack layer

provide value from localization and diversification that justifies the additional costs and different performance profiles. This matters most for companies based in countries or regions with robust national AI strategies their firms can plug into.

- **How do you prepare for a landscape that keeps shifting?** Build resilience in your AI tech stack by applying three principles: redundancy, modularity, and heterogeneity. Redundancy means having fallback options for the layers most exposed to disruption, such as contracting with two frontier model providers so a price hike or policy change impacting any one doesn't strand critical workflows. Modularity means designing your stack so a shock at one layer doesn't cascade through the others, for example by building applications in a way that lets you switch from one model provider to another without rewriting the underlying code. Heterogeneity means ensuring those fallbacks aren't exposed to the same risks, such as pairing a US closed-weight model with an open-weights alternative that can be self-hosted, or running primary workloads on a US hyperscaler's infrastructure with a sovereign-cloud option for regulated workloads.

These principles are harder to apply as the bifurcation deepens, and some break down entirely in a fully bifurcated world. Prioritizing them now won't make you immune, but it buys time to pivot. Alongside the principles, leaders must maintain a watch list of shifts—tightening export controls or technical breakthroughs in either ecosystem—that would force them to activate these principles.

AI and the technologies that power it are irreversibly intertwined with the geopolitical environment; the more dependent a company is on technology, the more exposed to volatility. What separates strong companies from those that are more exposed to risk is whether they can translate this awareness into geopolitical muscle, building resilience that holds up as the superpower bifurcation deepens.

The authors would like to thank Azeem Azhar, founder of Exponential View*; Wenwei Peng, post-doctoral fellow at Harvard University's Department of Economics; and Leonid Zhukov, director of the BCG X AI Science Institute, for their generous contributions to our research.*

---

business. For more ideas and inspiration from the Institute, please visit our website and follow us on LinkedIn and X (formerly Twitter).

Authors

Nikolaus Lang

Managing Director & Senior Partner; Global Leader, BCG Institute; Global Vice Chair, Global Advantage Practice
Munich

Sylvain Duranton

Managing Director & Senior Partner; Global Leader, BCG X
Paris

Vladimir Lukic

Managing Director & Senior Partner; Global Leader, Tech and Digital Advantage
Boston

Matt Langione

Managing Director & Partner
Boston

Rodrigo Ortiz Mena

BCG Institute Ambassador, Geopolitics & Society Lab
New York

Jona Lampert

Consultant
Doha

David Zuluaga Martínez

Senior Director, BCG Institute
New York

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group bridges the gap between ambition and outcomes for the world's leading companies and organizations. We are built for this era of unprecedented change — bringing strategic clarity rooted in over 60 years of deep domain knowledge, combined with applied AI shaped by our practitioners. BCG works shoulder-to-shoulder with CEOs across industries and geographies to deliver transformative impact at scale: stronger returns, transferred capabilities, and change that sticks. For more information, visit bcg.com.

1 Generative AI has been the focal point of the commercial and geopolitical competition since 2024, but agentic and physical AI are rising in relevance—hence this piece's broader use of "AI" rather than generative AI alone.

2 OpenAI has since pivoted toward leasing capacity from third parties rather than owning data centers directly.

3 Not all data center capacity is optimized and utilized for AI purposes. However, data on AI-specific data center capacity is not yet available. Analysis by Exponential View on the US's and China's fleet stock of H100-equivalent GPUs—a measure of each country's capacity to meet computing demand for AI—indicates that the US's AI optimized computing power is 8x the size of China's.

4 Model capabilities are based on Epoch AI's Capability Index, which does not yet include the latest Mistral model "Mistral Large 3," while LMArena's lab rankings includes latest models.