



**Statement before the
Committee on Science, Space, and Technology of
the U.S. House of Representatives**

“DeepSeek: A Deep Dive”

A Testimony by:

Gregory C. Allen

Director, Wadhvani AI Center, CSIS

Tuesday, April 8, 2025

2318 Rayburn House Office Building

Chair Obernolte, Ranking Member Stevens, and distinguished Members of the Committee, thank you for inviting me to testify today. The Center for Strategic and International Studies (CSIS) does not take policy positions, so the views represented in this testimony are my own and should not be taken as representing those of my current or former employers.

I currently serve as the director of the Wadhvani AI Center at CSIS, where I lead a team conducting policy research at the intersection of technology, economics, and national security. Prior to CSIS, I spent three years working at the U.S. Department of Defense Joint Artificial Intelligence Center, where I most recently served as the director for strategy and policy. My primary professional background is corporate strategy roles in technology-driven industries. On March 7, 2025, I published a report through CSIS titled “DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race,” and many of my remarks today reflect my conclusions from that research effort.

For my testimony today, I hope to offer a useful perspective on the origins of DeepSeek’s AI models and what DeepSeek’s progress in AI model development, as well as Huawei’s progress in AI chip design and manufacturing, mean for U.S. competition with China.

DeepSeek did not come out of nowhere. Its parent company, High-Flyer Capital Management, has roots in AI-enabled high frequency trading that provided a strong technical foundation in terms of both computing infrastructure and workforce talent. Algorithmic trading firms frequently own and operate their own data center infrastructure, which they use to develop proprietary investment strategies powered by machine learning artificial intelligence. These efforts routinely include intense research on developing proprietary and unorthodox data center engineering methods for improved speed and efficiency. DeepSeek’s heritage from High-Flyer’s machine learning research and intensive computing optimization techniques are evident in some of the strategies that it has pursued to develop its AI models.

Many media and policy commentators in the United States and China have argued that DeepSeek’s efficiency gains are evidence that U.S. export controls have failed to restrain China’s AI sector and that DeepSeek proves that huge numbers of cutting-edge U.S. AI chips are not required for developing and serving high-performance AI capabilities. There is some merit to aspects of the first claim, but the second is wrong.

To begin, DeepSeek’s V3 research paper states that their models were trained on 2,788,000 GPU-hours using Nvidia H800 chips, which, at an estimated cost of 2 dollars per GPU-hour, equates to \$5.576 million. This was only the final, successful pretraining run. It did not include the compute cost of the hundreds of experiments that were run beforehand to generate the insights necessary for that final run, nor did it include the compute costs associated with post-training fine-tuning or inference compute workloads. Thus, it is not an apples-to-apples fair comparison to say that DeepSeek’s AI model development cost only \$5 million while American firms require hundreds of millions of dollars.

The H800 is a degraded version of the H100 chip that Nvidia specifically developed for the Chinese market to comply with the U.S. export controls imposed on October 7, 2022. In particular, those controls restricted the sale of chips that exceeded performance thresholds across two metrics: total processing power and interconnect speed.

Government sources told me that the U.S. government knew that these restrictions would prevent Nvidia from shipping its (at the time) market-leading A100 chip and upcoming H100 chips. They inferred that Chinese customers would be restricted to using the V100, which was first introduced in 2017 and has significantly lower performance than its successors. Additionally, the U.S. government assumed that, if Nvidia were to develop a new chip specifically for the Chinese market that exceeded export control performance thresholds in one metric but not the other, this would require the typical multiyear AI chip development timeline.

However, Nvidia had a mechanism for post-manufacturing modification of its existing chip products. Specifically, each Nvidia chip is designed for redundancy, recoverability, and defect tolerance to minimize the impact of manufacturing defects. Nvidia blew fuses on A100 chips to reduce their interconnect speed (but not their processing power) below the export control performance thresholds, thus creating the A800 product lines that were legal to export to China.

The Biden administration ultimately realized that continued sales of the A800 and H800 chips meant that their policy would not have the intended impact on China's AI ecosystem. Former government officials told me that this was clear internally by December 2022. However, it took the administration 12 months after their initial package of export controls to act. The U.S. government modified the export control performance thresholds in October 2023 to block exports of A800s, H800s, and any non-Nvidia chips with comparable performance to China. During the year-long period in which A800 and H800 chip exports to China were legal, an enormous number of them were sold to China. Precise sales figures are not available, but Nvidia's disclosed sales to China between October 31, 2022, and October 31, 2023, exceeded 9 billion dollars.

If DeepSeek were to admit that it used H100s in its data center to train its models, it would be confessing to activity illegal under U.S. law, since the H100 has never been legally available in China and cannot even legally be purchased by Chinese-firm subsidiaries outside of China. Thus, if DeepSeek did engage in illegal activity, it would have an incentive to conceal this fact. This raises the question of whether DeepSeek lied in its publications when it claimed to have used H800 chips exclusively.

While there is no public evidence that large-scale Nvidia chip smuggling occurred prior to October 2023, industry sources told me that, by early 2024, large-scale H100 chip smuggling operations were underway. In mid-2024, journalists at The Information interviewed participants in eight distinct H100 smuggling networks, each of which provided evidence that they had completed

H100 smuggling transactions worth more than \$100 million.¹ These networks continue to be active, with increasingly sophisticated techniques for evading detection. For example, a December 2024 investigation by The Information found that

When notified of an upcoming inspection, smugglers have duplicated the serial numbers of the servers with Nvidia chips they've purchased from Supermicro [and already smuggled to China] and attached them to other servers they had access to.²

More recent reporting by the Wall Street Journal suggests that this smuggling now includes the latest generation of Nvidia Blackwell AI chips.³ China is betting that its network of smugglers and shell companies can find the leaks in the Commerce Department's Bureau of Industry and Security (BIS) export control enforcement barrier. As long as Congress continues to neglect BIS by providing grossly inadequate resources compared to the size and importance of its mission, China has a reasonable expectation of success.⁴ BIS needs not only more money, but also more skilled staff, more enforcement agents, and better enabling technology, especially in data analysis.

Industry analyst Ben Thompson has pointed to strong evidence that DeepSeek did in fact use H800s as it claimed: Many of DeepSeek's algorithmic and architectural improvements are ideal for maximizing the effective use of computing resources under conditions of limited interconnect bandwidth. At a minimum, this strongly suggests that DeepSeek uses many H800s in its computing infrastructure and that a model with the performance of V3 can indeed be trained exclusively on H800s.

This does not prove, however, that DeepSeek exclusively uses H800s in its overall computing infrastructure. Indeed, some reporting in Chinese news media claims that DeepSeek did train its slightly more recent R1 model on Nvidia H100 chips that are banned from China under U.S. export controls. The semiconductor consulting firm SemiAnalysis, citing anonymous industry sources, recently wrote that DeepSeek has a total of 50,000 Hopper generation GPUs, a category that includes H100s, H800s, and H20s. SemiAnalysis specifically claimed that it has evidence that High-Flyer's/DeepSeek's computing infrastructure includes at least 10,000 H100s, 10,000 H800s, 30,000 H20s, and 10,000 A100s as part of its total computing stack. SemiAnalysis further estimated that High-Flyer/DeepSeek spent \$1.63 billion in GPU server capital expenditures alone (i.e., excluding other data center construction and operating costs).⁵

In terms of absolute performance DeepSeek's January 2025 AI models are best understood as being comparable to the best American models available in the summer of 2024. DeepSeek did introduce, however, many improvements related to computational efficiency for both AI training

¹ <https://www.theinformation.com/articles/nvidia-ai-chip-smuggling-to-china-becomes-an-industry>

² <https://www.theinformation.com/articles/u-s-prompts-nvidia-supermicro-probe-into-how-chips-ended-up-in-china>

³ <https://www.wsj.com/tech/china-nvidia-blackwell-chips-ai-531fed0c>

⁴ <https://www.csis.org/analysis/mismatch-strategy-and-budgets-ai-chip-export-controls>

⁵ <https://semianalysis.com/2025/01/31/deepseek-debates/>

and inference. Some of these were already well understood by American firms but not disclosed publicly, while others were genuinely new innovations. To the general public and media, both types were perceived as new.

While I do believe that DeepSeek should serve as a wakeup call for America, the extent of the media coverage on DeepSeek was out of proportion to its technical achievements. This reflects in part the fact that users of free AI models had not experienced an exposed chain-of-thought reasoning feature, which customers clearly enjoyed, as well as interest in the fact that the model came from China. This helped drive DeepSeek as a major topic of consumer and media interest, which in turn drove the tech stock price drop.

Beyond chip smuggling, the greatest strategic challenge for the United States is the potential for China to produce AI chips domestically at sufficient quantity and quality to build AI data center infrastructure that is competitive with the United States.

DeepSeek is not in and of itself the most significant threat to U.S. leadership in AI. Instead, the greater challenge arises from the possibility of China having a domestic ecosystem for producing its own AI chips at large scale and integrating them into Chinese data center training, as well as running inference for DeepSeek and other AI models.

As U.S. technology firms are planning hundreds of billions of dollars in AI data center infrastructure investments, it is worth remembering that—for those investments to be possible—companies like TSMC must manufacture enough AI chips to fill those data centers. In the case of companies such as Nvidia, their revenue growth in recent years is less than it would have otherwise been due to shortages of TSMC production capacity.

The Biden administration took many steps designed to definitively cut China's AI chip designers off from TSMC production capacity. Most recently, on January 15, the Commerce Department announced the final tranche of Biden administration export controls, often referred to as "the Foundry Rule." The Foundry Rule moved advanced chip production to a white-list system that will likely make it impossible for Chinese AI firms to access TSMC capacity to produce chips above export control performance thresholds even when operating through complex shell company arrangements. However, TSMC manufactured a strategically significant quantity of chips on behalf of Huawei via shell companies prior to the rule going into effect.

That effectively means that China's long-term future in AI is closely tied to its ability to produce AI chips domestically. The Biden administration sought to hamstring China's domestic production of advanced chips by restricting the sale of advanced semiconductor manufacturing equipment, including from other countries.

China's alliance of Huawei (AI chip designer), SMIC (AI chip manufacturer), and CXMT/XMC (high-bandwidth memory manufacturers) have recently made strategically significant progress in advancing domestic production of AI chips.

Domestically producing large quantities of AI chips will require China to domestically replicate multiple segments of the AI chip value chain. The most important links are AI chip design, advanced node logic chip manufacturing, and advanced node high-bandwidth memory (HBM) manufacturing.

Like the United States, China has many different companies working on AI chip design, including Huawei, Cambricon, Biren, and more. However, Huawei is unambiguously in the strongest position with its Ascend AI chip product line.

The most advanced logic chip manufacturer in China is SMIC. SMIC's SN2 facility in Shanghai is the sole facility in China with an active 7 nm logic chip production line and has been producing 7 nm chips since July 2021, more than a year before the first tranche of the Biden administration's semiconductor equipment export controls went into effect. SMIC and Huawei are now working to bring a 5 nm node into scaled production but must do so without access to Extreme Ultraviolet (EUV) lithography equipment, since China has no local producer of EUV lithography machines and since export controls have prevented such machines from ever being exported to China. Industry sources told me that, in early 2020, ASML was poised to ship EUV tools to China and that SMIC was planning to work with key research labs in Europe, such as the Interuniversity Microelectronics Centre (IMEC), to help develop their EUV-based manufacturing process.

In December 2024, industry sources told me that SMIC currently has enough immersion deep ultraviolet (DUV) lithography equipment supplied by the Dutch company ASML to produce 85,000 FinFET wafers per month (WPM) across both SN1 (which focuses on 14 nm node production) and SN2 (which focuses on 7 nm and 5 nm production). This acquisition of lithography tools reportedly took effect before Dutch DUV lithography export controls went into effect in mid-2023.

However, the bottleneck in expanding 7 nm (which in SMIC's node naming system is called "N+2") production capacity has not been lithography but rather U.S. tools for etching, deposition, inspection, and metrology. Some of this equipment is restricted on a country-wide basis, meaning that it cannot be legally sold anywhere in China. However, other types of this equipment were restricted only on an end-use and end-user basis. This means that the equipment can be sold to some customers in China but not others. In some cases, this even means that it can be sold to some facilities of a particular customer, but not others. In such cases, relocating the equipment from one facility to another would require a new export license in order to be legal. But SMIC's production of 7 nm chips using U.S. equipment is already illegal, and both SMIC and other Chinese firms always have the option to choose illegal activity, particularly since such activity frequently enjoys the active support of the Chinese government.

Industry sources told me that SiEn, Pensun, and Huawei's fab in Dongguan all were able to legally acquire the needed etching, deposition, and inspection/metrology equipment that SMIC needs for two reasons: (1) the equipment was not restricted on a country-wide basis to all of China and (2) the equipment was restricted on an end-use and end-user basis, but SiEn and Pensun told U.S. firms that it would exclusively be used for producing chips less advanced than 14 nm. These firms also denied any affiliation with Huawei. Government officials told me that in such circumstances, the equipment can often be sold under a no-license required status.

According to the sources, SiEn and Pensun, however, did not have sufficient customers providing demand for using all of the equipment they had purchased, and so some of it was never used operationally in their fabs. They purchased the equipment as a stockpiling move in anticipation of future export controls. The source described this as a "buy everything you can, while you can" strategy.

The SMIC SN2 facility needed the equipment. Since SiEn and Pensun were not making economically productive use of the equipment, they were amenable to a sale. This sale was negotiated in Q4 of 2024 and completed in Q1 of 2025. The sources are under the impression that all of the desired equipment is currently either installed at SMIC SN2 or on-site awaiting installation.

As a result of the successful in-country equipment transfer, SMIC expects to achieve 50,000 7 nm wafers per month (WPM) by the end of 2025. If all of this capacity was devoted to manufacturing Ascend AI chips that would imply the production of millions of Ascend 910C chips annually. However, SMIC is unlikely to devote all of its 7 nm capacity to Ascend chips. Huawei needs that 7 nm capacity for its chips for smartphones, laptops, data centers, and telecommunications equipment. Moreover, SMIC has other customers besides Huawei. Still, the point remains that Huawei is likely poised to dramatically expand Ascend production in the near future.

Huawei's Ascend chips continue to face challenges in terms of a lack of compatible AI software that is driving low utilization of purchased chips. However, this could change if DeepSeek's open-source community enthusiasm improves Huawei's CANN software ecosystem competitiveness with Nvidia's Compute Unified Device Architecture (CUDA). DeepSeek may have both the technical knowledge and the open-source community enthusiasm to finally start generating momentum around Huawei's competitor to CUDA, which Huawei refers to as its Compute Architecture for Neural Networks (CANN). For a company like DeepSeek, migrating all AI workloads from CUDA to CANN would likely be a multiyear project. The greater maturity of the CUDA software ecosystem currently makes Nvidia chips more attractive, but this could change over the next few years. If it does, it would have major implications for U.S. AI competitiveness both inside and outside of China.

Thank you for the opportunity to testify today, and I look forward to your questions.