Export Controls and the Future of China’s AI Compute Capacity

Fernando Ruiloba Portilla

Fall 2025

Frontier model performance has advanced significantly, resulting in impressive capabilities that confer substantial advantages to organizations and nations developing these models. In the quest to achieve a first-mover advantage, companies around the world have raced to pour in money into the industry, while governments attempt to develop policies that balance the protection of their national interests while ensuring sustainable and safe development of this new technology.

Consequentially, policymakers have sought to understand China’s frontier model training capabilities, which are directly linked to the country's compute resources. To estimate China's compute capacity for training frontier AI models by 2027 under current export controls, it is necessary to identify the main factors that influence this capability.

Frontier models are AI models with a large number of parameters, trained on enormous datasets using mind-boggling amounts of computing power to obtain state-of-the-art performance on various benchmarks. The usual lifecycle of a frontier model consists of an experimentation phase, which involves testing multiple architectures and hyperparameters to find the optimal configuration, a training phase which involves extracting meaningful relationships over the entire dataset and adjusting the internal model weights accordingly, and a deployment phase in which users leverage the trained model to perform inference and obtain desired outputs. While the training process is highly compute-intensive, it only needs one run to determine the optimal set of model weights. On the other hand, the deployment phase focuses on distributing those weights and enabling scalable inference across various applications.

Given these developments, an important question arises: what drives model performance? Kaplan et al. [1] from OpenAI point to three factors driving model performance: compute, number of parameters, and dataset size. Another study by Gunasekar et al. [2] points towards data quality as an efficient tool for achieving high performance with smaller models. Innovations in model architecture, such as the “mixture of experts” technique used by China’s Deep Seek, can also increase compute efficiency [3]. In general, researchers have found that model performance predictably increases with the amount of compute used to train the models. [4]

So how much compute exactly does it take to train a frontier model? Training compute is usually measured in floating-point operations (FLOPs) over the final training run. In recent years, some of the largest models were trained with up

to 6e+23 FLOPs. And according to some estimates this number doubles around every 9-10 months. [5] If this trend continues, by 2027 the biggest models could be using as much as 1.2e+24 FLOPs during training. Hardware and energy requirements also double every 12 and 13 months, respectively. [5]

Given the large amount of computing power needed to train frontier models, only some countries hold enough AI compute capacity. As of 2025, the United States accounts for about 75% (9.1e+20 FLOPs/s) of total AI supercomputer performance, with China in second place at 15% (1.9e+20 FLOPs/s). [6] In addition, American companies were involved in developing 338 of the 476 notable AI models and trained 18 of the 25 largest AI models by training compute. [6] The very large models were almost exclusively published by industry corporations, and 17 were published by DeepMind, Google AI, OpenAI, or Microsoft. [5] Some sources suggest that U.S. companies also lead in total number of users. [6]

While China is working to narrow this gap, it encounters significant challenges. In 2022, the US government set export controls on AI and semiconductor technology. This restriction on AI chips helped the US keep its lead in overall compute capacity and pushed China toward self-reliance. [7] One AI policy goal pursued by Beijing is the development of domestic alternatives to NVIDIA chips. Although export controls on semiconductor manufacturing equipment (SME) have hindered China's capacity to mass-produce domestic AI chips [7], Chinese SME firms have achieved notable market share growth across various SME segments since 2019. However, lithography remains a key SME segment in which China has yet to establish any market presence. [8] Extreme ultraviolet lithography, the equipment necessary to produce the most advanced chips, has

been particularly elusive, with only one US-aligned company, ASML, able to master it at scale. [9] In addition to chips, China is also pursuing domestic alternatives to US software such as NVIDIA's CUDA, Meta's Pytorch and Google's Tensorflow. However, these frameworks still lag far behind in terms of adoption and capabilities, thus their capacity to reduce reliance on US technology is limited. Chinese AI companies are also finding ways around export controls to obtain banned Nvidia GPUs, such as stockpiling chips, smuggling, and setting up data centers globally. [7] In 2024, Huawei was able to secure 3 million chip dies from TSMC through a proxy company, enabling them to produce their new Ascend chips and effectively gain the computing power equivalent of 1 million export-controlled NVIDIA H100s. [10]

With all these factors in mind and if everything remains the same, I do not expect China to significantly expand its AI compute capacity by 2027. Even with a Chinese EUV breakthrough, fabs can take years to build, thus native capabilities for producing the most advanced chips are off the table. One argument is that China can scale its AI computing capabilities with weaker chips, but this view fails to consider that the performance gap might be of a few orders of magnitude. These constraints compound over time. With fewer chips, fewer experiments can be run, leading to fewer insights and worse performance in an industry with network effects and winner-takes-all dynamics. [11] In short, Chinese AI firms are at a constant disadvantage.

What would happen if export controls were implemented differently, such as simultaneously pursuing tight export controls over AI chips and more lax controls over cloud computing services? I believe this approach could increase China's reliance on US technologies and feed into the positive reinforcement

loop that helps enhance US-controlled technologies. Cloud computing also enables stronger control mechanisms by facilitating user and workload identification and allowing for swift access revocation. This might seem like a desirable policy goal for US policy-makers, but it could also be a big deterrent for adoption by Chinese companies. Nevertheless, I remain confident that the advantages of developing frontier AI models using state-of-the-art hardware would surpass the potential drawbacks for Chinese AI companies, particularly in light of China's considerable data resources and talent pool. This would also prevent Chinese companies from missing out on critical network effects. In this scenario, I would expect China’s total use of compute capacity for training frontier AI models to grow in line with current trends, at around 65% per year. [12] That would put its overall AI compute capacity at around 3.135e+20 FLOPs.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu and D. Amodei, "Scaling Laws for Neural Language Models," 23 January 2020. [Online]. Available: https://arxiv.org/abs/2001.08361. [Accessed 1 November 2025].

[2] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. d. Rosa, O. Saarikivi, A. Salim, S. Shah and H. S. Behl, "Textbooks Are All You Need," 20 June 2023. [Online]. Available: https://arxiv.org/abs/2306.11644. [Accessed 1 November 2025].

[3] C. Metz, "How Did DeepSeek Build Its A.I. With Less Money?," New York Times, 12 February 2025. [Online]. Available: https://www.nytimes.com/2025/02/12/technology/deepseek-ai-chip-costs.html. [Accessed 1 November 2025].

[4] L. Heim, M. Anderljung, E. Bluemke and R. Trager, "Computing Power and the Governance of AI," 14 February 2024. [Online]. Available: https://www.governance.ai/analysis/computing-power-and-the-governance-of-ai. [Accessed 1 November 2025].

[5] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos, "Compute Trends Across Three Eras of Machine Learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022.

[6] K. F. Pilz, J. Sanders, R. Rahman and L. Heim, "Trends in AI Supercomputers," 23 April 2025. [Online]. Available: https://arxiv.org/abs/2504.16026. [Accessed 11 November 2025].

[7] K. Chan, G. Smith, J. Goodrich, G. DiPippo and K. F. Pilz, "Full Stack: China's Evolving Industrial Policy for AI," 26 June 2025. [Online]. Available:

https://www.rand.org/pubs/perspectives/PEA4012-1.html. [Accessed 3 November 2025].

[8] J. Feldgoise and H. Dohmen, "Inside Beijing's Chipmaking Offensive. Where Is China Gaining Ground?," CSET, 14 July 2025. [Online]. Available: https://cset.georgetown.edu/article/inside-beijings-chipmaking-offensive/. [Accessed 27 10 2025].

[9] J. Wentz and A. Lin, "Breakthroughs or Boasts? Assessing Recent Chinese Lithography Advancements," 24 September 2025. [Online]. Available: https://www.csis.org/blogs/strategic-technologies-blog/breakthroughs-or-boasts-assessing-recent-chinese-lithography. [Accessed 3 November 2025].

[10] L. Heim, "China's AI Models Are Closing the Gap—but America's Real Advantage Lies Elsewhere," 2 May 2025. [Online]. Available: https://www.rand.org/pubs/commentary/2025/05/chinas-ai-models-are-closing-the-gap-but-americas-real.html. [Accessed 3 November 2025].

[11] J. Schneider and L. Ottinger, "AI Geopolitics in the Age of Test-Time Compute w Chris Miller + Lennart Heim," 6 January 2025. [Online]. Available: https://www.chinatalk.media/p/ai-geopolitics-in-the-age-of-test. [Accessed 3 November 2025].

[12] W. Hunt, S. M. Khan and D. Peterson, "China's Progress in Semiconductor Manufacturing Equipment: Accelerants and Policy Implications," CSET, 2021.

[13] L. Heim, "Understanding the Artificial Intelligence Diffusion Framework: Can Export Controls Create a U.S.-Led Global Artificial Intelligence Ecosystem?," RAND, 14 January 2025. [Online]. Available: https://www.rand.org/pubs/perspectives/PEA3776-1.html. [Accessed 1 November 2025].

[14] L. Heim, "Crucial Considerations for Compute Governance," 24 February 2024. [Online]. Available: https://blog.heim.xyz/crucial-considerations-for-compute-governance/. [Accessed 1 November 2025].