Trends in AI Supercomputers

Konstantin F. Pilz [1] **James Sanders** [2] **Robi Rahman** [2] **Lennart Heim** [3 2]

Abstract

Frontier AI development requires AI supercomputers with thousands of AI chips. Yet, analysis of developments in these systems is limited. We create a dataset of 500 AI supercomputers from 2019 to 2025 and quantify key trends. We find that computational performance of AI supercomputers has doubled every nine months, while hardware acquisition cost and power needs have doubled yearly. The leading system in March 2025, xAI's Colossus, had a hardware cost of $7B, and required 300 MW of power—as much as 250,000 households. While the public sector owned 60% of AI supercomputer performance in 2019, this share declined to only 20% by 2025, which may limit access to frontier capabilities for academic researchers. The United States dominates AI supercomputers, owning 75% of performance, suggesting a large degree of geographical concentration of compute. Our study provides visibility into AI infrastructure trends, allowing policymakers to make more informed AI governance decisions.

This is a shortened version accepted at ICML TAIG 2025. For the full version, see arxiv.org/abs/2504.16026.

1. Introduction

The computing power (compute) used to train notable AI models has increased at a rate of 4.2× per year since 2010, enabling new AI capabilities across many domains (Sevilla & Roldan, 2024). Exponentially increasing training compute relied on larger, higher-performance AI supercomputers (Hobbhahn et al., 2023; Frymire, 2024). Yet, data and analysis of trends in these systems is scarce. Existing public resources like the Top500 list or the ML-Perf benchmark rely on voluntary submissions and thus lack sufficient data to reliably analyze trends (Top500; Mattson et al., 2020).[1] We attempt to close this gap by collecting a dataset of more than 500 AI supercomputers between 2019 and 2025 and analyzing trends in performance, cost, power needs, and country and public/private distribution.

2. Methods

We define an AI supercomputer as a computer system capable of supporting large-scale AI model training, deployed on a contiguous data center campus. To qualify, a system must contain chips that can accelerate AI workloads such as NVIDIA's H100, Google's TPUv4, or other AI chips with features like FP16/INT8 precision support, dedicated matrix multiplication units, high-bandwidth memory, or documented use in training notable AI models. To limit the dataset to the most significant systems we additionally apply a performance threshold and only include systems that achieved at least 1% of the performance of the most powerful existing AI supercomputer at that time.

Using the Google Search API, existing compilations of AI supercomputers, and manual searches, we collected data on 501 leading AI supercomputers between 2019 and February 2025, and an additional 225 pre-2019 systems. For each AI supercomputer, we documented various features including chip specifications, the first operational date, reported performance, ownership, and location. Our dataset, along with regular updates, is available at epoch.ai/data/ai-supercomputers. We estimate that our dataset covers approximately 10% of aggregate performance across all AI chips produced through 2025, and about 15% of AI chip stocks held by major companies as of early 2025. Roughly half of the 25 largest AI training runs in Epoch AI (2025)'s notable models dataset as of March 2025 had a corresponding AI supercomputer in our dataset. Find an assessment of our data coverage in Appendix C.1.1.

Before analyzing trends, we filter our data to include only high-certainty operational systems. We then fit lin-log regressions and report 90% confidence intervals (CI) for all growth rates. For all regressions, we consider only systems that were among the top-10 leading AI supercomputers

[1]Georgetown University [2]Epoch AI [3]Centre for the Governance of AI. Correspondence to: Konstantin Pilz <kfp15@georgetown.edu>, Robi Rahman <robi@epoch.ai>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]For a review of other data sources of AI supercomputers, see Appendix A.

when they first became operational.

We provide details on our data collection and analysis methods in Appendix B.

3. Results

3.1. Computational performance has doubled every nine months

Between 2019 and 2025, the computational performance of leading AI supercomputers in 16-bit FLOP/s has doubled every 9 months (Figure 1). The most performant AI supercomputer in March 2025, xAI's Colossus achieved 1.98×10^{20} 16-bit FLOP/s, which was about 60× higher than Oak Ridge's Summit's 3.46×10^{18} FLOP/s, the leading AI supercomputer in 2019.

Two key factors drove the rapid performance growth: a yearly 1.6× (90% CI: 1.5–1.8×) increase in chip quantity and a 1.6× (90% CI: 1.5–1.7×) annual improvement in performance per chip. While systems with more than 10,000 chips were rare in 2019, companies deployed AI supercomputers more than ten times that size in 2024, such as xAI's Colossus with 200,000 NVIDIA H100 and H200 chips.

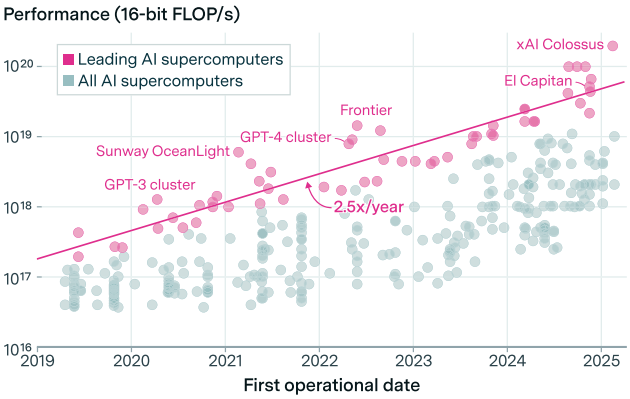

Figure 1. The performance of leading AI supercomputers (in FLOP/s, for 16-bit precision) has doubled every 9 months (a rate of 2.5× per year, 90% Confidence Interval (CI): 2.4–2.7×).

3.2. Power requirements have doubled every year

We assess the annual growth rate in power requirements of the leading AI supercomputers either based on reported power requirement or, if unavailable, by estimating the power requirement based on the number and type of AI chips.

We find that the power need of the leading AI supercomputers increased by 2.0× (90% CI: 1.8–2.2×) each year between 2019 and 2025. In January 2019, Summit at Oak Ridge National Lab had the highest power requirement with 13 MW. In 2024, the first systems began to cross the 100 MW threshold, and in March 2025, xAI's Colossus had the highest power requirement at an estimated 300 MW. For comparison, this is equivalent to the average power consumption of 250,000 U.S. households (EIA, 2024).[2]

3.3. Hardware cost has doubled every 13 months.

We analyze annual growth in the hardware cost for leading AI supercomputers based on either publicly reported cost figures or—if those are unavailable—by estimating the total hardware cost, based on the quantity of chips used and publicly available price data. We adjust all values for inflation and show our results in 2025 USD.

Hardware cost of the leading AI supercomputers increased by 1.9× every year between 2019 and 2025 (Figure 2). In 2019, Oak Ridge National Lab's Summit had the highest hardware cost with about $200 million. In 2024, the first systems crossed the $1B threshold and the most expensive AI supercomputer as of March 2025 was xAI's Colossus with an estimated hardware cost of $7 billion.

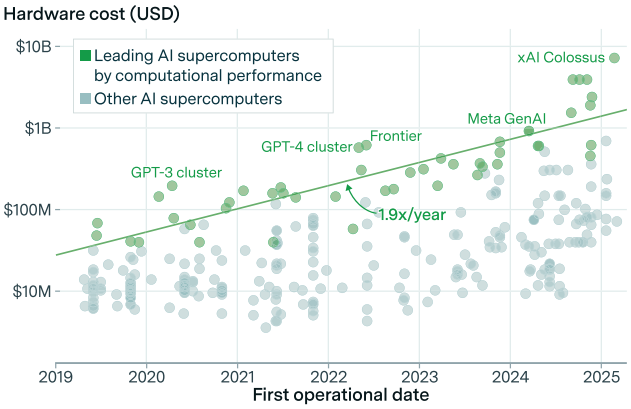

Figure 2. The hardware cost in 2025 USD of leading AI supercomputers has grown at a rate of 1.9× (90% CI: 1.8–2.1× per year) from 2019 to 2025.

3.4. Companies own the majority of AI supercomputers

For each AI supercomputer in our dataset, we classify the owner into one of three categories: public, private, and public/private (meaning the system has owners from both sectors or a private project received at least 25% of funding from a government.)

The private sector's share of total compute in our dataset rapidly increased from less than 40% in 2019 to about 80% in 2025 (Figure 3), while the public sector's share of AI supercomputers rapidly decreased from about 60% in 2019 to about 15% in 2025. Our data may even underestimate this shift, given that companies are less likely to publish data on their systems than public owners. However, note that public sector entities may still be able to access private

[2] 10,800 kWh /8760 h = 1.23 kW; 312 MW/ 1.23 kW = 250,000

sector AI supercomputers, given that many are available through cloud services.

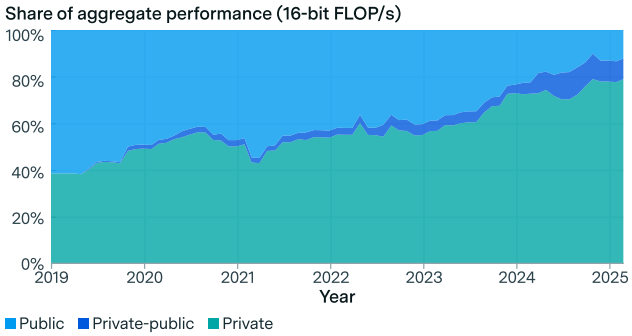

Figure 3. Relative performance shares of public and private sectors based on the owner of the AI supercomputer.

3.5. The United States hosts a dominant share of global AI supercomputer performance

When analyzing the distribution across countries, we find that in 2019, 70% of total comptuational performance in our dataset was in the United States, while 20% was in China (Figure 4).[3] Between 2019 and 2022, the Chinese share grew considerably, reaching about 40% at the start of 2022, although this may be an artifact of our incomplete data coverage. China's share has since diminished; in March 2025, the United States hosted around 75% of AI supercomputers by performance while China hosted around 15%.

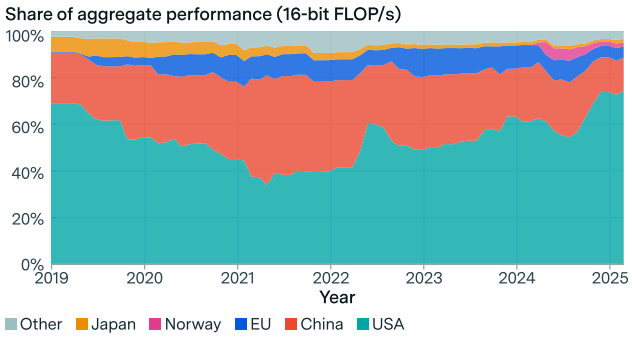

Figure 4. Share of aggregate 16-bit computing power by country over time from AI supercomputers in our dataset. We are visualizing all countries that held a more than 3% share at some point in time. See Appendix C.1.1 for a discussion of our data coverage.

[3]Physical location of an AI supercomputer does not directly determine access, given many of our systems are available through cloud services. Furthermore, location also does not necessarily determine ownership since AI supercomputers sometimes belong to companies headquartered abroad.

4. Discussion

4.1. AI supercomputer growth both relied on and enabled the increased economic importance of AI

The observed increase in AI supercomputer performance relied in part on long-standing improvements in chip design and manufacturing (Roser et al., 2023; Hobbhahn et al., 2023). However, the rapid 2.5× annual growth between 2019 and 2025 was only possible due to a rapid surge in investment as AI supercomputers developed from academic tools for scientific discovery to industrial machines running economically valuable workloads.

Leading AI supercomputers in 2019, like the U.S. Department of Energy's Summit and Sierra, were designed to handle a variety of workloads across different scientific domains and advance foundational research (Oak Ridge National Laboratory, undated). This changed in the early 2020s, when companies increasingly used AI supercomputers to train AI models with commercial applications, such as OpenAI's GPT-3 and GitHub's Copilot integration (Brown et al., 2020; Dohmke & GitHub, 2021). These demonstrations of AI capabilities led to a significant increase in investment, creating a record demand for AI chips (Our World in Data, 2024; Samborska, 2024; Richter, 2025).

As investments in AI increased, companies were able to build more performant AI supercomputers with more and better AI chips. This created a reinforcing cycle: increased investment enabled better AI infrastructure, which produced more capable AI systems, which attracted more users and further investment. The growth of AI supercomputers thus appears to have been both a result of increased funding and a driver of continued investment as AI supercomputers demonstrated their economic value.

4.2. U.S. dominance in global AI supercomputer distribution

We found that AI supercomputers are heavily concentrated in one country: Around three quaters of all AI supercomputer performance in our data was based in the United States as of March 2025 (Figure 4). This U.S. dominance likely resulted from AI supercomputers becoming increasingly commercialized and dominated by companies instead of governments or academia. Since U.S. companies dominated related industries, they were able to capture a large share of the AI supercomputer market. For instance, in 2019, three U.S. companies, AWS, Microsoft, and Google alone made up 68% of the global cloud computing market share (Gartner, 2020). American companies also played leading roles in key AI advances, including in recommender systems, scientific applications like AlphaFold, and LLM chatbots like ChatGPT (Dong et al., 2022; Jumper et al., 2021; OpenAI, 2022).

4.2.1. The United States will likely continue leading in AI supercomputers

The United States dominates not only AI supercomputers but also AI development, cloud services, and critically, the design and supply chain of AI chips (Sastry et al., 2024). This position has enabled the U.S. government to impose export controls on AI chips to China as well as other non-allied countries (Allen, 2022; Heim, 2025).

Yet, several challenges could reduce U.S. dominance: rapidly increasing power demands for AI infrastructure (Pilz et al., 2025; Fist & Datta, 2024; Mahmood et al., 2025), sovereign AI investments by countries like France, the UK, and Saudi Arabia (Reuters, 2025; UK DSIT, 2025; Benito, 2024), and China's significant investments in domestic AI chip production (Reuters, 2024). However, these challenges remain limited: Sovereign projects are small compared to U.S. systems, and Chinese efforts face significant obstacles due to restricted access to advanced lithography equipment (Grunewald, 2023).

Given U.S. control over critical semiconductor supply chain chokepoints and stated government policy to maintain AI leadership (The White House, 2025), U.S. dominance in AI supercomputers could continue for the foreseeable future, giving the United States an outsized role in shaping global AI development and governance.

4.3. Impacts of increased private sector concentration

Our finding that companies rapidly increased their share of AI supercomputers aligns with broader trends in AI research. Besiroglu et al. (2024) found that the share of large-scale AI models produced by academic institutions rapidly declined from 65% in 2012 to 10% in 2023. The increased dominance of industry likely resulted from AI models, and the AI supercomputers that powered them, becoming increasingly economically importan t(Section 4.1). This economic importance drove major private investments that enabled systems like xAI's Colossus with a hardware cost of $7B, while investment in government projects increased more slowly, with the hardware for the leading system, El Capitan, costing only $600M.

This private-sector concentration produces two significant consequences: First, it creates a significant barrier for academic researchers who have historically played vital roles in advancing AI methods and providing independent scrutiny. Although AI supercomputers available through cloud providers could enable researchers with access to large-scale compute resources, the costs of renting thousands of AI chips for sufficient durations remain prohibitively expensive for academic budgets (Heim & Egan, 2023). As a consequence, academic researchers are often forced to work on smaller, less capable models or narrower problems, potentially limiting the exploration of new algorithmic approaches and independent AI safety and interpretability research (Lohn, 2023; Besiroglu et al., 2024).

Second, as private companies control an increasing share of AI supercomputers, governments may struggle to monitor compute trends because companies are often less transparent about their compute ownership and use than academic or government labs. Given compute determines both AI development and deployment capabilities, limited data on AI infrastructure could make it more difficult for governments to track AI progress (Sastry et al., 2024). Additionally, limited data makes it harder for governments to assess national competitiveness on AI infrastructure and develop coherent national AI strategies.

To address these challenges, governments could require companies to report key AI infrastructure metrics like total compute capacity and gather intelligence on other countries' infrastructure, improving competitive assessment and potentially laying the groundwork for verifying potential future international AI agreements (Sastry et al., 2024; Baker, 2023).

5. Conclusion

We compiled a dataset of 500 AI supercomputers between 2019 and 2025 and found that performance, number of chips, power requirements, and hardware cost have all grown exponentially. The 2.5 $\times$ annual performance growth of AI supercomputers has enabled a rapid increase in training compute for frontier AI models, which has fueled significant advances in AI capabilities and driven further investment in infrastructure.

Our data also reveals that companies rapidly increased their share of total AI supercomputer performance from 40% in 2019 to more than 80% in 2025. This compute divide may hinder independent AI research and scrutiny, and complicate governments' oversight of AI development. The United States hosts approximately 75% of global AI supercomputer performance and will likely maintain this dominance through its control over the AI chip supply chain.

AI supercomputers have been a key driver of AI progress and represent a central component of the AI supply chain (Sastry et al., 2024). Our analysis provides valuable information about AI supercomputers' growth patterns, distribution, and resource requirements. Such information will be increasingly important for policymakers, and more generally for understanding the trajectory of AI.

Acknowledgements

We would like to thank the following people for their assistance, feedback, and contributions:

- David Owen for reliable guidance on scope and execution of this project as well as repeated feedback on the report.
- Qiong Fang and Veronika Blablová for substantial contributions to data collection.
- Lovis Heindrich, Terry Wei, David Atanasov for assistance with data entry and verification.
- Robert Sandler for figure design and Edu Roldán for figure editing.
- Luke Frymire for his work estimating power requirements for AI supercomputers and Ben Cottier for his work estimating hardware acquisition costs of AI supercomputers.
- Pablo Villalobos for reviewing our code.
- Caroline Falkman Olsson and Jessica P. Wang for typesetting
- Various people who reviewed our data and suggested additional systems to include.
- The Epoch AI team and everyone else who provided feedback and helpful discussions.

Impact Statement

This paper provides insights into the AI supercomputer landscape, thereby helping researchers and policymakers make more informed assessments of national competitiveness and AI governance. Increased visibility could also aid authoritarian governments in making better decisions, but we believe that the greater visibility provides a far larger benefit to open, democratic societies, and it is thus crucial to make this analysis publicly available.

References

Alexsandar K. Microsoft Acquired Nearly 500,000 NVIDIA "Hopper" GPUs This Year, December 2024. URL https://www.techpowerup.com/330027/microsoft-acquired-nearly-500-000-nvidia-hopper-gpus-this-year. Accessed 15-04-2025.

Allen, G. C. Choking off China's Access to the Future of AI, October 2022. URL https://www.csis.org/analysis/choking-chinas-access-future-ai. Accessed 20-04-2025.

Allen, G. C. Understanding the Biden Administration's Updated Export Controls, December 2024. URL https://www.csis.org/analysis/understanding-biden-administrations-updated-export-controls. Accessed 20-04-2025.

Ashkboos, S., Markov, I., Frantar, E., Zhong, T., Wang, X., Ren, J., Hoefler, T., and Alistarh, D. Quik: Towards end-to-end 4-bit inference on generative large language models. *arXiv:2310.09259*, 2023. URL https://arxiv.org/abs/2310.09259.

AWS. Amazon Web Services (AWS) - Cloud Computing Services, 2020. URL https://pages.awscloud.com/amazon-ec2-p4d.html. Accessed 15-04-2025.

AWS. Project Ceiba – Largest AI Super Computer Co-Built with NVIDIA, 2023. URL https://aws.amazon.com/nvidia/project-ceiba/. Accessed 15-04-2025.

AWS. AI Accelerator - AWS Trainium - AWS, undated. URL https://aws.amazon.com/ai/machine-learning/trainium/. Accessed 15-04-2025.

Baker, M. Nuclear arms control verification and lessons for AI treaties. *arXiv:2304.04123*, 2023. URL https://arxiv.org/abs/2304.04123.

BeautifulSoup. beautifulsoup4, 2025. URL https://pypi.org/project/beautifulsoup4/. Accessed 15-04-2025.

Benito, A. Saudi Arabia launches $100 Billion AI initiative to lead in global tech, October 2024. URL https://www.cio.com/article/3602900/saudi-arabia-launches-100-billion-ai-initiative-to-lead-in-global-tech.html. Accessed 20-04-2025.

Besiroglu, T., Bergerson, S. A., Michael, A., Heim, L., Luo, X., and Thompson, N. The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv:2401.02452*, 2024. URL https://arxiv.org/abs/2401.02452.

Borkar, R., Wall, A., Pulavarthi, P., and Yu, Y. Azure Maia for the era of AI: From silicon to software to systems — Microsoft Azure Blog, 2024. URL https://azure.microsoft.com/en-us/blog/azure-maia-for-the-era-of-ai-from-silicon-to-software-to-systems/. Accessed 15-04-2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020. URL https://arxiv.org/abs/2005.14165.

Champelli, P., Niiler, E., and Fitch, A. The Nvidia Chips Inside Powerful AI Supercomputers, 2024. URL https://www.wsj.com/tech/ai/nvidia-chip-technology-artificial-intelligence-006e29d4. Accessed 15-04-2025.

Chang, J., Lu, K., Guo, Y., Wang, Y., Zhao, Z., Huang, L., Zhou, H., Wang, Y., Lei, F., and Zhang, B. A survey of compute nodes with 100 TFLOPS and beyond for supercomputers. *CCF Transactions on High Performance Computing*, 6(3):243–262, 2024. URL https://link.springer.com/article/10.1007/s42514-024-00188-w.

Chen, M. and Chan, R. AMD to ship up to 400,000 new AI GPUs in 2024, say sources, 2023. URL https://www.digitimes.com/news/a20231205PD217/amd-ai-gpu-2024-us-china-chip-ban.html. Accessed 15-04-2025.

Chik, H. No sign of China's new supercomputers among world's Top500. South China Morning Post, November 2022. URL https://www.scmp.com/news/china/science/article/3180337/no-sign-chinas-new-supercomputers-among-worlds-top500. Accessed 20-04-2025.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., and Owen, D. The rising costs of training frontier AImodels. *arXiv:2405.21015*, 2024. URL https://arxiv.org/pdf/2405.21015v2.

Cushman & Wakefield. Data Center Development Cost Guide 2025, 2025. URL https://cushwake.cld.bz/Data-Center-Development-Cost-Guide-2025/8-9/. Accessed 15-04-2025.

Dickson, B. GPT-4.5 for enterprise: Are accuracy and knowledge worth the high cost?, January 2025. URL https://venturebeat.com/ai/gpt-4-5-for-enterprise-do-its-accuracy-and-knowledge-justify-the-cost/. Accessed 20-04-2025.

Dohmen, H. and Feldgoise, J. A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports, October 2023. URL https://cset.georgetown.edu/article/bis-2023-update-explainer/. Accessed 20-04-2025.

Dohmke, T. and GitHub. Introducing GitHub Copilot: Your AI Pair Programmer, June 2021. URL https://github.blog/news-insights/product-news/introducing-github-copilot-ai-pair-programmer/. Accessed 19-04-2025.

Dong, Z., Wang, Z., Xu, J., Tang, R., and Wen, J. A brief history of recommender systems. *arXiv:2209.01860*, 2022. URL https://arxiv.org/abs/2209.01860.

Dongarra, J. J. The linpack benchmark: An explanation. In *International Conference on Supercomputing*, pp. 456–474. Springer, 1987. URL https://link.springer.com/chapter/10.1007/3-540-18991-2_27.

EIA. Frequently Asked Questions (FAQs), 2024. URL https://www.eia.gov/tools/faqs/faq.php?id=97&t=3. Accessed 20-04-2025.

Epoch AI. Data on Notable AI Models. Epoch AI Data Hub, 2025. URL https://epochai.org/data/notable-ai-models. Accessed 20-04-2025.

Financial Times. Nvidia to make $12bn from AI chips in China this year despite US controls, 2023. URL https://www.ft.com/content/b76ef55b-21cd-498b-ac16-5660908bb8d2. Accessed 15-04-2025.

Fist, T. and Datta, A. How to Build the Future of AI in the United States, October 2024. URL https://ifp.org/future-of-ai-compute/. Accessed 20-04-2025.

Fox, M. A single customer made up 19% of Nvidia's revenue last year. UBS thinks it's Microsoft, 2024. URL https://www.businessinsider.com/nvidia-stock-mystery-customer-microsoft-ubs-revenue-h100-gpu-chips-2024-5. Accessed 15-04-2025.

Frymire, L. The length of time spent training notable models is growing. Epoch AI, 2024. URL https://epoch.ai/data-insights/training-length-trend. Accessed 20-04-2025.

Galabov, V., Sukumaran, M., and Lewis, A. Data Center Server Market Insights and Forecast. Technical report, Omdia, 2025. URL https://omdia.tech.informa.com/collections/afcei005/data-center-server-market-insights-and-forecast. Accessed 20-04-2025.

Garreffa, A. Analyst says NVIDIA Blackwell GPU production volume will hit 750K to 800K units by Q1 2025, 2024. URL https://www.tweaktown.com/news/100980/analyst-says-nvidia-blackwell-gpu-production-volume-will-hit-750k-to-800k-units-by-q1-2025/index.html. Accessed 15-04-2025.

Gartner. Gartner Says Worldwide IaaS Public Cloud Services Market Grew 37.3% in 2019, August 2020. URL https://www.gartner.com/en/newsroom/press-releases/2020-08-10-gartner-says-worldwide-iaas-public-cloud-services-market-grew-37-point-3-percent-in-2019. Accessed 20-04-2025.

Grunewald, E. Introduction to AI Chip Making in China. Institute for AI Policy and Strategy, July 2023. URL https://www.iaps.ai/research/ai-chip-making-china. Accessed 20-04-2025.

Grunewald, E. AI Chip Smuggling is the Default, not the Exception. AI Policy Bulletin, January 2025. URL https://www.aipolicybulletin.org/articles/ai-chip-smuggling-is-the-default-not-the-exception. Accessed 20-04-2025.

Heim, L. Understanding the Artificial Intelligence Diffusion Framework, January 2025. URL https://www.rand.org/pubs/perspectives/PEA3776-1.html. Accessed 20-04-2025.

Heim, L. and Egan, M. Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS): Implications for Export Controls. Center for the Governance of AI, November 2023. URL https://cdn.governance.ai/Accessing_Controlled_AI_Chips_via_Infrastructure-as-a-Service.pdf.

Hobbhahn, M., Heim, L., and Aydos, B. Trends in Machine Learning Hardware, 2023. URL https://epoch.ai/blog/trends-in-machine-learning-hardware. Accessed 20-04-2025.

Hochman, T. Building Baseload: Reforming Permitting for AI Energy Infrastructure, 2020. URL https://www.thefai.org/posts/building-baseload-reforming-permitting-for-ai-energy-infrastructure. Accessed 20-04-2025.

Huang, Q., Bao, B., Liang, Y., Guo, X., Huang, M., Tekur, C., and Carilli, M. Introducing native PyTorch automatic mixed precision for faster training on NVIDIA GPUs, September 2020. URL https://pytorch.org/blog/accelerating-training-on-nvidia-gpus-with-pytorch-automatic-mixed-precision/. Accessed 20-04-2025.

IDC. Artificial Intelligence Infrastructure Spending to Surpass the $200bn USD Mark in the Next 5 years, According to IDC. Technical report, International Data Corporation, February 2025. URL https://www.idc.com/getdoc.jsp?containerId=prUS52758624. Accessed 20-04-2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with Alphafold. *Nature*, 596(7873): 583–589, 2021. URL https://www.nature.com/articles/s41586-021-03819-2.

Langston, J. Microsoft announces new supercomputer, lays out vision for future AI work - Source, 2020. URL https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/. Accessed 15-04-2025.

Lee, J. Microsoft Azure Eagle is a Paradigm Shifting Cloud Supercomputer, 2023. URL https://www.servethehome.com/microsoft-azure-eagle-is-a-paradigm-shifting-cloud-supercomputer-nvidia-intel/. Accessed 15-04-2025.

Lepton AI. The Missing Guide to the H100 GPU Market, January 2024. URL https://blog.lepton.ai/the-missing-guide-to-the-h100-gpu-market-91ebfed34516. Accessed 20-04-2025.

Lohn, A. Scaling AI — Cost and Performance of AI at the Leading Edge. Georgetown's Center for Security and Emerging Technology, December 2023. URL https://cset.georgetown.edu/publication/scaling-ai/.

Luszczek, P. Results, November 2024. URL https://hpl-mxp.org/results.md. Accessed 15-04-2025.

Mahmood, Y., Byrd, C., Somani, E., Pilz, K. F., and Heim, L. Possible Options for Unlocking and Securing U.S. Energy for AI Production, March 2025. URL https://www.rand.org/pubs/working_papers/WRA3883-1.html. Accessed 20-04-2025.

Martin, D. Google Was Third Biggest Data Center Processor Supplier Last Year: Research, 2024. URL https://www.crn.com/news/components-peripherals/2024/google-was-third-biggest-data-center-processor-supplier-last-year-research. Accessed 15-04-2025.

Mattson, P., Cheng, C., Diamos, G., Coleman, C., Micikevicius, P., Patterson, D., Tang, H., Wei, G.-Y., Bailis, P., Bittorf, V., et al. Mlperf training benchmark. *Proceedings of Machine Learning and Systems*, 2:336–349, 2020. URL https://arxiv.org/abs/1910.01500.

Meta. Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research, 2022. URL https://ai.meta.com/blog/ai-rsc/. Accessed 15-04-2025.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. Mixed precision training. *arXiv:1710.03740*, 2017. URL https://arxiv.org/abs/1710.03740.

Morgan, T. P. Stacking Up AMD MI200 Versus Nvidia A100 Compute Engines, December 2021. URL https://www.nextplatform.com/2021/12/06/stacking-up-amd-mi200-versus-nvidia-a100-compute-engines/. Accessed 15-04-2025.

Morgan, T. P. Energy Giant Eni Boosts Its HPC Oomph By An Order Of Magnitude, 2024. URL https://www.nextplatform.com/2024/01/24/energy-giant-eni-boosts-its-hpc-oomph-by-an-order-of-magnitude/. Accessed 15-04-2025.

Morgan, T. P. CoreWeave's 250,000-Strong GPU Fleet Undercuts The Big Clouds, 2025. URL https://www.nextplatform.com/2025/03/05/coreweaves-250000-strong-gpu-fleet-undercuts-the-big-clouds/. Accessed 15-04-2025.

Moss, S. Training Google's Gemini: TPUs, multiple data centers, and risks of cosmic rays. DCD, December 2023. URL https://www.datacenterdynamics.com/en/news/training-gemini-tpus-multiple-data-centers-and-risks-of-cosmic-rays/. Accessed 20-04-2025.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V. A., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., and Zaharia, M. Efficient large-scale language model training on gpu clusters using megatron-lm, 2021. URL https://arxiv.org/pdf/2104.04473.

Nolan, T. Nvidia h100: Are 550,000 gpus enough for this year?, August 2023. URL https://www.hpcwire.com/2023/08/17/nvidia-h100-are-550000-gpus-enough-for-this-year/. Accessed 19-04-2025.

NVIDIA. NVIDIA H100 Tensor Core GPU. NVIDIA Developer Blog, March 2023. URL https://www.nvidia.com/en-us/data-center/h100/. Accessed 20-04-2025.

NVIDIA. Introduction to NVIDIA DGX H100/H200 Systems, April 2025. URL https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html#power-specifications. DGX H100/H200 User Guide, Power Specifications section. Last updated April 10, 2025.

Oak Ridge National Laboratory. Summit, undated. URL https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/. Accessed 20-04-2025.

O'Brien, M. and Fingerhut, H. Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water, 2023. URL https://web.archive.org/web/20250220064324/https://apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-f551fde98083d17a7e8d904f8be822c4. Accessed 15-04-2025.

Oldham, M., Lee, K., and Gangidi, A. Building Meta's GenAI Infrastructure, 2024. URL http://web.archive.org/web/20240828225612/https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/. Accessed 15-04-2025.

OpenAI. Introducing ChatGPT. OpenAI Blog, November 2022. URL https://openai.com/index/chatgpt/. Accessed 20-04-2025.

Our World in Data. Annual global corporate investment in artificial intelligence, by type, 2024. URL https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type. Accessed 20-04-2025.

Patel, D., Nishball, D., and Knuhtsen, R. Mi300x vs H100 vs H200 Benchmark Part 1: Training – CUDA Moat Still Alive, January 2024. URL https://semianalysis.com/2024/12/22/mi300x-vs-h100-vs-h200-benchmark-part-1-training/. Accessed 20-04-2025.

Pilz, K. F., Mahmood, Y., and Heim, L. AI's Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness. *RAND Corporation*, (RR-A3572-1), 2025. URL https://www.rand.org/pubs/research_reports/RRA3572-1.html.

Pires, F. Chinese Companies Spend $5 Billion on Nvidia GPUs for AI Projects, 2023a. URL https://www.tomshardware.com/news/chinese-companies-spend-big-on-nvidia-gpus-for-ai-projects. Accessed 15-04-2025.

Pires, F. China's ByteDance Has Gobbled Up $1 Billion of Nvidia GPUs for AI This Year, 2023b. URL https://www.tomshardware.com/news/chinas-bytedance-has-gobbled-up-dollar1-billion-of-nvidia-gpus-for-ai-this-year. Accessed 15-04-2025.

Rahman, R. and Owen, D. Performance improves 12x when switching from FP32 to tensor-INT8. Epoch AI, January 2024. URL https://epoch.ai/data-insights/hardware-performance-trend. Accessed 20-04-2025.

Reuters. China sets up third fund with $47.5 bln to boost semiconductor sector. *Reuters*, 5 2024. URL https://www.reuters.com/technology/china-sets-up-475-bln-state-fund-boost-semiconductor-industry-2024-05-27/.

Reuters. Details of 110 Billion Euros in Investment Pledges at France's AI Summit. Reuters, January 2025. URL https://www.reuters.com/technology/artificial-intelligence/details-110-billion-euros-investment-pledges-frances-ai-summit-2025-02-10/. Accessed 20-04-2025.

Richter, F. Infographic: Nvidia's AI-Fueled Rally Hasn't Been Without Hiccups. Statista, January 2025. URL https://www.statista.com/chart/32358/nvidia-share-price/. Accessed 20-04-2025.

Riken Center for Computational Science. About Fugaku — RIKEN Center for Computational Science, undated. URL https://www.r-ccs.riken.jp/en/fugaku/about/. Accessed 15-04-2025.

Roser, M., Appel, C., and Ritchie, H. What is Moore's Law? Our World in Data, 2023. URL https://ourworldindata.org/moores-law. Accessed 20-04-2025.

Samborska, A. Investment in generative AI has surged recently. Bloomberg, June 2024. URL https://ourworldindata.org/data-insights/investment-in-generative-ai-has-surged-recently. Accessed 20-04-2025.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing power and the governance of artificial intelligence. *arXiv:2402.08797*, 2024. URL https://arxiv.org/abs/2402.08797.

SemiAnalysis. Datacenter industry model. Commercial database, 2024. URL https://semianalysis.com/datacenter-industry-model/. Accessed 20-04-2025.

Sevilla, J. and Roldan, E. Training Compute of Frontier AI Models Grows by 4-5x per Year, 2024. URL https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year. Accessed 20-04-2025.

Shah, A. Top500: China Opts Out of Global Supercomputer Race, 2024. URL https://thenewstack.io/top500-chinas-supercomputing-silence-aggravates-tech-cold-war-with-u-s/. Accessed 15-04-2025.

Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A., Holecek, B., Koomey, J. G., Masanet, E. R., and Sartor, D. A. 2024 united States Data Center Energy Usage Report. Technical report, Lawrence Berkeley National Laboratory, 2024. URL https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf. Accessed 20-04-2025.

Shilov, A. Nvidia sold half a million H100 AI GPUs in Q3 thanks to Meta, Facebook — lead times stretch up to 52 weeks: Report, 2023a. URL https://www.tomshardware.com/tech-industry/nvidia-ai-and-hpc-gpu-sales-reportedly-approached-half-a-million-units-in-q3-thanks-to-meta-facebook. Accessed 15-04-2025.

Shilov, A. Nvidia to reportedly triple output of compute gpus in 2024: Up to 2 million h100s, October 2023b. URL https://www.tomshardware.com/news/nvidia-to-reportedly-triple-output-of-compute-gpus-in-2024-up-to-2-million-h100s. Accessed 19-04-2025.

Shilov, A. Nvidia's H100 AI GPUs cost up to four times more than AMD's competing MI300X, February 2024. URL https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-h100-ai-gpus-cost-up-to-four-times-more-than-amds-competing-mi300x-amds-chips-cost-dollar10-to-

dollar15k-apiece-nvidias-h100-has-peaked-beyond-dollar40000. Accessed 15-04-2025.

Tal, E., Viljoen, N., Coburn, J., and Levenstein, R. Our next-generation Meta Training and Inference Accelerator, 2024. URL https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/. Accessed 22-04-2025.

Tekin, A., Tuncer Durak, A., Piechurski, C., Kaliszan, D., Aylin Sungur, F., Robertsén, F., and Gschwandtner, P. State-of-the-art and trends for computing and interconnect network solutions for HPC and AI. Technical report, 2021. URL https://prace-ri.eu/wp-content/uploads/State-of-the-Art-and-Trends-for-Computing-and-Interconnect-Network-Solutions-for-HPC-and-AI.pdf. Accessed 20-04-2025.

The Information. AI Data Center Database, 2025. URL https://www.theinformation.com/projects/ai-data-center-database. Accessed 20-04-2025.

The White House. Executive Order on AI Infrastructure Development. The White House Briefing Room, February 2025. URL https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2025/01/14/executive-order-on-advancing-united-states-leadership-in-artificial-intelligence-infrastructure/. Accessed 20-04-2025.

Top500. Top500 List. Online database. URL https://www.top500.org. Undated. Accessed 20-04-2025.

Trueman, C. xAI's Memphis Supercluster has gone live, with up to 100,000 Nvidia H100 GPUs, 2024. URL https://web.archive.org/web/20241009045341/https://www.datacenterdynamics.com/en/news/xais-memphis-supercluster-has-gone-live-with-up-to-100000-nvidia-h100-gpus/. Accessed 15-04-2025.

UK DSIT. AI Opportunities Action Plan. Technical report, UK Government, March 2025. URL https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan. Accessed 20-04-2025.

U.S. Bureau of Industry and Security. Addition of Entities to the Entity List and Revision of an Entry on the Entity List. Technical report, U.S. Department of Commerce, June 2019. URL https://www.federalregister.gov/documents/2019/06/24/2019-13245/addition-of-entities-to-the-entity-list-and-revision-of-an-entry-on-the-entity-list. Accessed 20-04-2025.

U.S. Bureau of Labor Statistics. Producer Price Index by Industry: Data Processing, Hosting and Related Services, 2025. URL https://fred.stlouisfed.org/series/PCU518210518210. Accessed 15-04-2025.

U.S. Department of Commerce. Commerce Adds Seven Chinese Supercomputing Entities to Entity List for their Support to China's Military Modernization, and Other Destabilizing Efforts, April 2021. URL https://www.commerce.gov/news/press-releases/2021/04/commerce-adds-seven-chinese-supercomputing-entities-entity-list-their. Accessed 20-04-2025.

Vultr. Pioneering the Future of AI with AMD Instinct™ MI300X GPUs, Broadcom, and Juniper Networks — Vultr Blogs, 2024. URL https://blogs.vultr.com/Lisle-data-center. Accessed 15-04-2025.

Wei, L. U.S. Think Tank Reports Prompted Beijing to Put a Lid on Chinese Data - WSJ, 5 2023. URL https://www.wsj.com/world/china/u-s-think-tank-reports-prompted-beijing-to-put-a-lid-on-chinese-data-5f249d5e. Accessed 18-04-2025.

A. Review of existing data sources

A.1. The Top500 list and its limitations for AI supercomputers

The Top500 list has been the primary leaderboard for tracking supercomputer performance since its inception in 1993. It ranks systems based on their performance in solving linear equations using the LINPACK benchmark (Dongarra, 1987). While this benchmark has provided a consistent, long-term method for comparing traditional high-performance computing (HPC) systems, it has several significant limitations when applied to AI supercomputers:

- Participation in the Top500 list is voluntary, leading to significant gaps in reporting. Companies, particularly cloud providers, which own many of the largest AI supercomputers, face limited incentives to report their AI supercomputers. Running the LINPACK benchmark diverts valuable supercomputer and engineer time from more economically valuable uses like AI training or deployment. Instead of reporting to the Top500, companies sometimes independently publish promotional blog posts about their systems (Langston, 2020; Meta, 2022; AWS, 2023), while often maintaining ambiguity about the number and size of their largest systems to avoid giving competitors unnecessary information about their strategies. Additionally, Chinese owners stopped reporting any systems to the Top500 list in 2022, presumably to reduce scrutiny and avoid U.S. sanctions (Shah, 2024).
- LINPACK is not an AI benchmark. It measures performance on linear equations requiring high-precision 64-bit number formats (Dongarra, 1987), while modern AI workloads run on much lower precision formats (16-bit, 8-bit, or even 4-bit for some inference workloads[4]). While performance on different precision formats was formerly highly correlated, the introduction of tensor cores for lower precision formats on AI accelerators led to drastically faster performance increases in these formats (Hobbhahn et al., 2023; Rahman & Owen, 2024). This divergence means LINPACK performance does not accurately capture a supercomputer's performance for AI workloads.[5] New benchmarks like HPL-MxP and ML-Perf better capture AI-relevant performance but have not been widely adopted (Luszczek, 2024; Mattson et al., 2020).

Besides the Top500, no major datasets of supercomputers exist, meaning that previous analyses of supercomputers, such as Hochman (2020), Tekin et al. (2021) and Chang et al. (2024) have exclusively relied on the Top500 list. While these analyses offer useful insights into changes in components, performance, and energy efficiency of traditional supercomputers, the limitations of the Top500 lists discussed above mean the observed trends do not comprehensively capture AI supercomputers.

A.2. Commercial databases of AI supercomputers

Some analysts, like SemiAnalysis and The Information, have private databases of AI supercomputers that are available for paid subscribers. Furthermore, some companies such as Omdia offer trackers of AI chip shipments (SemiAnalysis, 2024; The Information, 2025; Galabov et al., 2025). These databases are typically focused on providing business intelligence. Thus, they do not assess historical trends and may not capture data from non-industry sources. Furthermore, these databases usually do not disclose their methods and sources and do not make the analysis of their data publicly available.

B. Detailed Methods

B.1. Data collection process

We relied on systematic Google searches and publicly available datasets to find potential AI supercomputers. For each potential AI supercomputer, we conducted an additional search to find and verify all relevant publicly available data about it.

Search methodology:

a) We used the Google Search API to search for terms such as "AI supercomputer" and "GPU cluster" in consecutive 12-day windows (1-1-2019–1-3-2025). We additionally conducted year-by-year country searches (e.g., "Albania AI supercomputer").
   - Although our study period begins in 2019, we also conducted a similar, pared-down Google search for January

[4] Or even 4-bit precision for some inference workloads (Ashkboos et al., 2023).

[5] For instance, Microsoft's Eagle and Japan's Fugaku have comparable performances on LINPACK (5.6×10^{17} FLOP/s vs 4.4×10^{17} FLOP/s), but given that Fugaku does not contain any GPUs or other chips optimized for low-precision performance, they diverge by almost an order of magnitude on FP8 performance (2.9×10^{19} FLOP/s vs 4.3×10^{18} FLOP/s) (Lee, 2023; Riken Center for Computational Science, undated).

2016–January 2019 in order to be able to determine which AI supercomputers were in the top 10 by computational performance at the start of 2019. For this, we reduced our search terms by roughly 80% to lower the number of records to look through.

b) We parsed the top results with the Beautiful Soup Python package and used GPT-4o via the OpenAI API to extract system names and chip counts of any AI supercomputers mentioned.
c) We grouped entries by name in a spreadsheet, deduplicated, verified all potential AI supercomputers manually, and added those that fit our inclusion criteria to our dataset.
d) Find additional details about the Google Search methods in Section B.2.

Additional sources:

a) Top500 list, inferring AI chip counts from reported accelerator cores.
   - Many systems in the Top500 did not contain AI chips; however, those that did usually listed the 'Accelerator/Co-Processor' type and the total number of 'Accelerator/Co-Processor Cores.' Since we knew the number of cores for each AI chip model, we calculated the implied AI chip count for the system by dividing the number of cores by the cores per AI chip. We verified this method by checking it for AI supercomputers in the Top500 with previously known AI chip counts.
   - We considered all Top500 entries from June 2014 to November 2024 (but included only those that qualified for our inclusion criteria between 2017 and 2025).
b) Epoch AI's notable AI models dataset.
c) Published compilations of Chinese AI supercomputers (redacted, please reach out).
d) A small number of entries from a project on sovereign compute resources led by Aris Richardson (publication forthcoming).
e) MLCommons Results.
f) gpulist.ai (last accessed January 2025).
g) Articles and newsletters shared by colleagues, such as from SemiAnalysis, Transformer, and Import AI.

Remaining components:

- We built our initial dataset via Google Alerts for the keyword "AI supercomputer" (June 2023–Aug 2024)
- Two Chinese-language analysts conducted targeted searches of systems in China and Hong Kong (see Appendix B.3).
- Our main data collection focused on AI supercomputers that first became operational between 2019 and 2025. However, we also included AI supercomputers that became operational between 2017 and 2019 if they met the standard inclusion criteria, or if they were operational before 2017 and were at least 1% as large as the largest known supercomputer in January 2017.
- We collected various additional sources for details on specific supercomputers using the Perplexity API.
- For over 500 key supercomputers, a team member did an additional verification of the entry (marked as true in the 'Verified Additional Time' field). This focused on systems that were especially large for their time, most Chinese systems, and any outliers.

B.2. Google search methodology

We conducted automated Google searches spanning from January 2019 to March 2025 for consecutive 12-day windows, using various keywords related to AI supercomputers. For each search term, we collected different amounts of results based on their utility in finding relevant information:

- "AI Supercomputer": 30 Google results
- "AI Supercomputer cluster": 30 Google results
- "AI Supercomputer news": 20 Google results
- "AI Supercomputer cluster news": 20 Google results
- "GPU Cluster": 20 Google results
- "Compute Cluster": 10 Google results
- "V100 Cluster": 10 Google results
- "A100 Cluster": 10 Google results
- "H100 Cluster": 10 Google results

We parsed all websites using the BeautifulSoup (2025) Python library and used GPT-4o from the OpenAI API to search for

information on all mentioned AI supercomputers (see prompt below).

Our searches yielded over 20,000 unique websites, resulting in approximately 2,500 potential AI supercomputer mentions after deduplication. For each unique AI supercomputer, we used the Perplexity API to collect additional data sources (see prompt below).

GPT-4o Prompt for Initial Extraction

Here is the text from a webpage that potentially contains some information about AI supercomputers. Please list the names of any AI supercomputer clusters that are listed in this article, separated by semicolons if there are multiple. If you know the company/organization name that owns/runs it, you should write the supercomputer name as the company/organization name, followed by the name of the cluster. If the cluster does not have a name, simply refer to it with 'UNNAMED' and include any identifiable information given. Please include any information about the number and type of AI chips (e.g. GPUs or TPUs) in square brackets after the cluster name. Say '[NOINFO]' if there is no information in the article about chip type or quantity. For example, a response might look like 'OpenAI Stargate [NOINFO]; Frontier [37,632 AMD MI250X]; Microsoft UNNAMED Arizona H100s [50,000 NVIDIA H100s]'. You should only list AI supercomputer clusters and associated chip information, nothing else. If there are no supercomputer clusters mentioned in the article, just reply with 'None'. If you can't access or read the article, just reply with 'Could not access article'. However, this should be rare, and mainly only happen if the article is paywalled. Do not mention any other details. Article text: {TEXT HERE}

Perplexity Prompt for Detailed Information

Tell me all the details you can about the {SUPERCOMPUTER NAME} supercomputer, including but not limited to: What type of AI accelerator chips (eg GPUs, TPUs, etc) do they use (be as specific about the exact type of chip as possible)? How many do they have, if any? When was it completed, or when is it expected to be completed? When was it first announced? What is the timeline for any updates/iterations to this supercomputer? Where is it located? (be as specific as possible) How many AI FLOP/s could it do? Who operates it? Who uses it? Who owns the supercomputer? Please list several organizations if it is a joint partnership, and list if these organizations are or part of government, academia, industry, or something else? Are there multiple supercomputers that could go by roughly this name? Have there been different versions/iterations of this supercomputer?

B.3. Approach for finding Chinese AI supercomputers

We decided to redact our approach to finding Chinese AI supercomputers and avoid providing identifying information about them throughout the paper to preserve data sources. We take this step as a precautionary measure because Chinese websites cited in public reports have been redacted or replaced with malware in the past (Wei, 2023).

B.4. Power requirements

We calculated the peak power demand for each AI supercomputer with the following formula:

$$\textit{Chip TDP} \times \textit{number of chips} \times \textit{system overhead} \times \textit{PUE}$$

We collected Thermal Design Power (TDP) for most chips when publicly available, though we did not find the TDP for some Chinese chips and custom silicon such as Google's TPU v5p. We considered both primary and secondary chips when counting the number and types of chips. We used a 2.03× multiplier for non-GPU hardware to account for system overhead (additional power needed for other server components like CPUs, network switches, and storage), based on NVIDIA DGX H100 server specifications (NVIDIA, 2025). We also factored in Power Usage Effectiveness (PUE), which is the ratio of total data center power use to IT power use (with a minimum value of 1). According to the 2024 United States Data Center Energy Usage Report (Shehabi et al., 2024), specialized AI datacenter facilities had an average PUE of 1.14 in 2023, which is 0.29 lower than the overall national average of 1.43. We adjusted the trend for all datacenter facilities to estimate the average PUE of AI datacenters by subtracting 0.29 from the overall values reported by Shehabi et al. (2024) (Table 1).

Table 1. AI data center power usage effectiveness (PUE) over time, adapted from (Shehabi et al., 2024).

YEAR	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
PUE	1.31	1.29	1.26	1.22	1.20	1.18	1.17	1.14	1.12	1.10

The full formula we use is:

$$\begin{aligned} Power = [&(Primary\ AI\ chip\ TDP \times Primary\ AI\ chip\ quantity) \\ &+(Secondary\ AI\ chip\ TDP \times Secondary\ AI\ chip\ quantity)] \\ &\times Server\ overhead\ factor \times Datacenter\ PUE \end{aligned}$$

We base some of the reported power values in our dataset on the top 500 list. However, the list reports average power utilization during the benchmark, rather than peak power requirement. To determine peak power, we compare peak and average power for supercomputers where we have both, find that they differ on average by a factor of 1.5, and scale all the Top500 reported power figures by this factor. We then multiply by the PUE in the given year to find peak power demand for the entire system.

B.4.1. LIMITATIONS WITH OUR POWER DATA

We rely on owner-reported power estimates for 15% of the AI supercomputers in our dataset. These reported figures lack standardization—some may represent only critical IT load at theoretical maximum utilization, while others include complete data center infrastructure overhead (accounting for power conversion losses and cooling requirements).

For the remaining 85% of systems, we estimate the power requirements as detailed in the previous section. A key limitation of our current approach is the application of a uniform 2.03× multiplier for all chip types to account for additional system hardware. Future analyses would benefit from developing chip-specific overhead multipliers that better reflect the varying cluster-level power requirements across different AI chip and cluster architectures.

To check for consistency between reported and estimated power values, we plotted the correlation below (Figure 5). The correlation coefficient of 0.98 indicates our values are highly correlated.

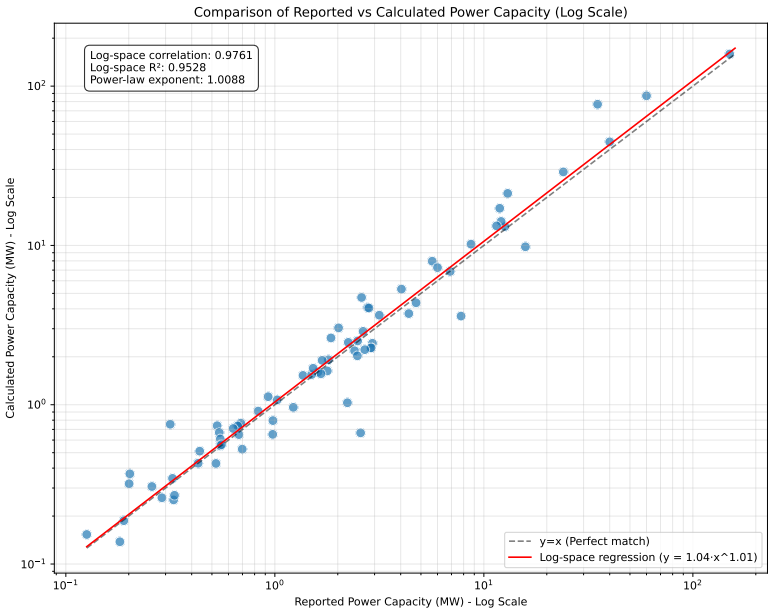

Figure 5. Comparison of power requirements for AI supercomputers that reported it, versus our calculations of power requirements based on chip type and count.

Note that our methods assess theoretical peak power usage when all the processors are fully utilized and not power consumption. The average power consumption of an AI supercomputer is usually only a fraction of its peak.

B.5. Hardware cost

We use the publicly reported total hardware cost of the AI supercomputer in our analysis whenever it is available. When it is unavailable, we estimate this cost based on the chip type, quantity, and public chip prices. The procedure used to estimate costs is adapted from Cottier et al. (2024). Using Epoch AI's dataset of hardware prices, we select the latest known price of the chips used in the AI supercomputer, from before the system's first operational date. For each type of chip, we multiply the cost per chip by the number of chips, multiply by factors for intra-server and inter-server overhead, and then sum these costs if there are multiple types of chips. Intra-server cost overhead was estimated in Cottier et al. (2024) for the NVIDIA P100 (1.54×), V100 (1.69×), and A100 (1.66×), based on known DGX and single-GPU prices near release. We use the mean of these factors (1.64×) for all chips, to estimate server prices, including interconnect switches and transceivers. Then, we adjust for the cost of server-to-server networking equipment, which was estimated to be 19% of final hardware acquisition costs.

Additionally, we apply a discount factor of 15% to the final hardware cost of the AI supercomputer to account for large purchasers of AI chips often negotiating a discount on their order. We discuss limitations with this estimate and our cost data in the next section.

Our final formula for estimating hardware cost is as follows:

$$\begin{aligned} \textit{Hardware acquisition cost} = [&(\textit{Primary AI chip cost} \times \textit{Primary AI chip quantity}) \\ &+ (\textit{Secondary AI chip cost} \times \textit{Secondary AI chip quantity})] \\ &\times \textit{Intra-server overhead} \times \textit{Inter-server overhead} \times \textit{Discount factor} \end{aligned}$$

In this formula, our intra-server overhead, or "chip-to-server" factor, is 1.64×, our inter-server overhead, or "server-to-cluster" factor, is 1.23×, and our discount factor is 0.85×.

Notably, our cost figures refer only to the hardware acquisition cost of the AI supercomputer, and not costs required for maintenance, electricity, or the cost of the datacenter hosting it.[6]

All cost values are adjusted for inflation into 2025 USD, using the producer price index for the Data Processing, Hosting, and Related Services industry, reported by the Federal Reserve Bank of St. Louis (U.S. Bureau of Labor Statistics, 2025). We divided pre-2025 cost figures by the price index value at its closest reported date and multiplied by the price index value in January 2025. Our trends refer to values in 2025 USD.

B.5.1. Limitations with our hardware cost data

Our cost data for AI supercomputers has several important limitations:

1. We found reported cost figures for only a limited subset of AI supercomputers, with data predominantly from public sector systems rather than industry deployments.

2. The reported figures may diverge from true costs in multiple ways.

- They sometimes represent planned contract costs rather than final realized expenditures.
- Contract figures may bundle additional expenses, such as multi-year operational costs, that should be excluded from our analysis.
- When uncertainty about the precise meaning of reported costs is too high, we excluded the data, though some ambiguity likely remains.

3. We also encountered challenges with estimating hardware costs based on chip quantities and prices.

- Our price dataset lacks information for some GPU types, particularly custom silicon, though it does cover most common GPUs.
- Google does not sell TPUs, so our price data for them is based on comparison of their performance and manufacturing

[6] A 2025 estimate of the cost of datacenters puts them at $11.7 million per MW. This could be combined with our power requirement estimates to get an estimate of hardware plus datacenter acquisition (Cushman & Wakefield, 2025).

costs with those of NVIDIA chips that have similar technical specifications.
- Most GPU suppliers do not publish wholesale prices, forcing us to rely on third-party retailer prices and reports from experts that can vary significantly by vendor and time.
- We use the most recent listed price for each GPU, but prices fluctuate substantially with market conditions, so our limited time-series data means some AI supercomputer costs may be mismatched with the prices actually paid for the chips.

4. Given limited data, we assume that all AI supercomputers have the same overhead costs, but this is unlikely, particularly for systems built five years ago.

5. The discount factor is another significant source of uncertainty. Price negotiations generally occur privately, making reliable estimates difficult, and discounts vary substantially by supplier, purchaser, chip type, and time. For simplicity and due to data limitations, we apply a constant 15% discount rate across all AI supercomputers, but we expect the true rate to vary significantly by AI supercomputer. We selected this rate because it best aligns with the difference between our cost estimates and reported costs, and stated estimates of discount rates.[7] However, as stated above, our reported cost data is itself biased. Our universal discount rate likely overestimates costs for major purchasers like U.S. national labs[8] and the largest GPU buyers while underestimating costs in other scenarios.

As a consequence of these limitations, we estimate that a 90% confidence interval for the true hardware cost value is +/- 0.5 orders of magnitude (within a factor of $\sim 3\times$) of our estimate.

B.6. Figures and regressions

For all figures and regressions, we filtered the dataset as follows:

1. We excluded 99 AI supercomputers where the "Exclude" field is marked. 85 of these systems are outside of our definition because they do not meet our performance threshold. We also excluded 14 systems for other reasons, such as because we decided the chips they used did not qualify as AI chips.
2. We further excluded 92 AI supercomputers marked as "Possible duplicates". (We try to only mark systems as potential duplicates if we think there is a $>25\%$ chance they are a duplicate.)
3. We further excluded 36 AI supercomputers where "Single cluster" is marked as "No" or "Unclear".
4. We excluded 15 AI supercomputers where "Certainty" is lower than "Likely".
5. We excluded 113 AI supercomputers where "Status" is "Planned", i.e., systems that were not yet operational as of March 2025.

In total, we include 470 out of the 825 systems in our dataset in the analysis. Of these, 389 became operational in 2019 and after.

For all regressions, we consider the 57 AI supercomputers that were in the top-10 by 16-bit FLOP/s and became operational between January 2019 and March 2025.[9]

For our distribution figures we consider all 470 systems remaining after filtering, including those that became operational before 2019. We exclude AI supercomputers that were superseded by newer entries after the newer entry's first operational date.

B.7. Adequately representing performance gains from using lower precision units

Values in calculations for AI training (such as model weights, gradients, and updates) can be represented in different precisions. This is analogous to how you may represent the same number as "\$15,228,349,053.84" or "\$15 billion",

[7] Citi Analysts imply that Microsoft received a 33% discount compared to other purchasers, who paid what we would count as the full price (Shilov, 2024). If these groups buy equal amounts of chips, this implies an average discount of 16% (Morgan, 2021).

[8] NextPlatform implies that the Oak Ridge National Lab Summit supercomputer got close to a 50% discount on the cost of their GPUs, and that industry partners have historically paid (Morgan, 2024) $1.5\times$ to $2\times$ more for chips than National Labs.

[9] In some figures we specify that we are showing trends for the 59 AI supercomputers that were in the top-10 considering highest performance across 32, 16, and 8-bit precisions.

depending on the context. In this example, the first representation has a much higher precision than the second, but it also takes more memory to store.

Until the 2010s, AI training primarily used relatively high-precision 32-bit number formats but moved to 16-bit representation in the late 2010s[10] and began to move to 8-bit in 2024, thanks to new hardware supporting these precisions and algorithmic innovations to use the new number formats efficiently (Huang et al., 2020; NVIDIA, 2023). Given that working with values in lower precisions requires less memory and computations, AI chips offer much faster performance for calculations in lower precisions.

The shift in precision used for training in our study period makes it challenging to adequately display performance trends in our data.

- If we showed the highest available performance across these three precisions (Max OP/s)[11] it may seem like AI supercomputers that supported 8-bit precision in the early 2020s were more powerful than they actually were in practice, since 8-bit precision was not widely used to train AI models then.[12]
- Instead, we limit our analysis to performance in 16-bit precision (16-bit OP/s), which 92% of the AI supercomputers included in our analysis support.[13] However, we acknowledge that only considering 16-bit performance does not adequately show the performance gains AI companies achieved by moving to lower precision.

In practice, we find that trends in a) Max OP/s and b) 16-bit OP/s are mostly consistent (Appendix D. We thus use 16-bit OP/s as the default for our trend analysis.[14]

Meanwhile, we decided to use Max OP/s for our inclusion criteria, i.e., to select whether or not a given system has at least 1% of the performance of the leading operational AI supercomputer.

We include an overview table showing all metrics in each of 16-bit FLOP/s, 8-bit OP/s, and Max OP/s in Appendix D.1.

C. Limitations

This section summarizes some overall limitations of our data. We discuss limitations with specific parts of our data in the methods section (Appendix B).

C.1. Summary of limitations

C.1.1. We likely only cover about 10-20% of all AI supercomputers within our definition

We use four references to assess our coverage:

- **Coverage by chip production:** Our dataset likely covers 20–37% of all NVIDIA H100s produced until 2025, about 12% of all NVIDIA A100s produced, and about 18% of all AMD MI300X produced. Meanwhile, we estimate we cover less than 4% of Google's TPUs and very few custom AI chips designed by AWS, Microsoft, or Meta. We also only cover about 2% of NVIDIA chips designed to be sold in China (including the A800, H800, and H20). Our average coverage of the six chip types we assessed is 11%.
- **Coverage by company:** The coverage of different companies varies considerably, from 43% for Meta and 20% for Microsoft to 10% for AWS and 0% for Apple. The coverage of Chinese companies is particularly poor. Our average coverage of 8 major companies is 15%.
- **Coverage of total 16-bit FLOP/s in China:** Between end of 2020 and end of 2024 we cover between 10-20% of total Chinese 16-bit FLOP/s based on an estimate by IDC (2025).
- **Coverage of largest training runs:** Our dataset contains a matching AI supercomputer for about half of the largest training runs as of March 2025 reported by Epoch AI (2025). However, we only find official confirmation that the

[10] Micikevicius et al. (2017) is an early example of mixed-precision training which moved the most computationally expensive operations to 16-bit.

[11] OP/s stands for operations per second.

[12] Specifically, we are unsure when 8-bit training first became widespread. Developers usually do not report what precisions they use to train their models, making it difficult to assess when newly available formats were widely adopted.

[13] For comparison, 96% of AI supercomputers have a performance for Max OP/s (performance across 32, 16, and 8-bit precisions) The remaining AI supercomputers either lack performance data or we only found a performance for 64-bit precision.

[14] Notationally, we generally refer to 16-bit performance as FLOP/s (instead of "OP/s"), since this is more common terminology.

system was used for the specific training run for one-third of all models. Coverage of Chinese training runs is slightly better compared to all training runs.

Overall, we estimate we cover between 10 and 20% of all AI supercomputers as of early 2025. For more details on our coverage, see Appendix C.3.

C.1.2. We lack data for key properties

- We cannot reliably determine when an AI supercomputer was first operational. In most cases, we use the date an AI supercomputer was first reported as existing as the "first operational" date. However, owners may sometimes wait several months before publicly announcing their AI supercomputer, or they may announce a system even if it is not yet available. We expect that most of our "first operational" dates will be a few weeks to a few months later than the real date the AI supercomputer came online.
- We sometimes need to make assumptions about basic system facts. For instance, owners sometimes report vague chip quantities such as "EC2 UltraClusters are comprised of more than 4,000 latest NVIDIA A100 Tensor Core GPUs" (AWS, 2020), or "With thousands of MI300X GPUs available, clusters of any size can be deployed for reliable, high-performance computing." (Vultr, 2024). To include such AI supercomputers, we try to make reasonable estimates of the system's chips and performance and explain our reasoning in the notes field.
- Our data is incomplete. Some fields in our dataset are only filled for a fraction of systems, such as reported power requirement, reported hardware cost, and location. However, our data captures key statistics like performance and first operational date for more than 95% of all AI supercomputers that are included in our dataset.

C.1.3. Key reasons for low coverage

Why do we only cover 10–20% of all AI supercomputers? The following factors contribute to our low data coverage:

a) Companies often choose not to report their AI supercomputers publicly. While companies may benefit from increased public and investor attention when they publish information on large AI supercomputers, they may also prefer to keep this information private to maintain ambiguity about their competitive position.
b) Companies may only report their largest AI supercomputers. A large fraction of all chips are sold to hyperscalers that have more limited incentives to publish information about their AI supercomputers. While they may benefit from publishing information about their largest systems, they have no incentives to publish about the number and size of smaller AI supercomputers.
c) Even if an owner publishes information about an AI supercomputer, our search methods may not find it, especially if the information is published in a language other than English or Chinese.
d) Chinese companies may try to avoid scrutiny from U.S. regulators, both for chips that they legally imported, such as NVIDIA's A800 and H800, as well as illegally imported chips like NVIDIA's A100 and H100. Chinese companies may have smuggled more than 100,000 AI chips last year (Grunewald, 2025). See Appendix C.2.4 for a longer discussion.

C.2. Detailed limitations

This section discusses some of the limitations of our data and analysis in more detail.

C.2.1. Defining AI supercomputers is challenging

Ideally, our dataset would only capture systems that can efficiently run large-scale AI training workloads. However, it is difficult to develop a practical definition that captures only such systems based on limited publicly available data. Additionally, some companies, including Google DeepMind and OpenAI, have used AI chips distributed across multiple data center campuses to train large models (Moss, 2023; Dickson, 2025). To adequately include relevant AI supercomputers, we considered the following four definitions:

a) AI chips within a single building
b) AI chips on a single data center campus
c) AI chips within a fixed proximity (e.g., 2 or 5 miles)
d) No distance limit; an AI supercomputer is any system capable of training large models.

We decided to use definition (b), given the following considerations: The single building (a) may miss cases where well-connected accelerators span multiple buildings on the same campus. A fixed proximity definition (c) is not feasible in

practice since we do not know the precise physical location of most of the AI supercomputers in the dataset. Finally, a functional definition (d) is difficult to scope because assessing if a given AI supercomputer meets certain thresholds for performance, connectivity, and integrated operation requires data on network architecture and connections between AI supercomputers that public reports almost never provide. At the same time, we think it is useful to include AI supercomputers that meet the theoretical performance threshold but lack adequate network infrastructure, given it is comparatively easy to retrofit the networking equipment (see Appendix C.2.2).

We thus adopt the contiguous campus definition (b), where accelerators on a contiguous campus linked by high-bandwidth networks operate as a single AI supercomputer. However, there are two remaining limitations to this definition:

- **Limited data:** Public reports seldom include details on facility boundaries or network topology, making it hard to verify the contiguous nature of a campus.[15] When we are unsure if a reported system may span several campuses, we mark the field "Single Cluster" as "Unclear" (20 entries). We mark the "Single Cluster" field as "No" if we think the report most likely refers to a decentralized system (8 entries).
- **Decentralized training:** Our dataset currently does not capture the fact that AI developers may use multiple AI supercomputers for a training run. To assess which AI supercomputers may be most suitable for decentralized training, we would need additional information on the network bandwidth between them.

C.2.2. Theoretical performance does not necessarily correspond to usefulness for large-scale training

Systems may lack sufficient networking for efficiently running AI training. Public performance figures do not guarantee efficient large-scale training. Some AI supercomputers may suffer from inadequate networking, which can reduce utilization and prolong training runs (Narayanan et al., 2021). However, systems with inadequate networking infrastructure can easily be upgraded by changing the network fabric, usually at a fraction ($\sim$10–20%) of the total AI supercomputer cost (Lepton AI, 2024).

Performance on AI training depends on the software stack. Our analysis compares theoretical performance across hardware types. In practice, actual performance depends on the software stack and how well the hardware supports it. For instance, despite having a higher theoretical performance, SemiAnalysis assessed that AMD's MI300X is less useful for large-scale AI training than NVIDIA's H100 (Patel et al., 2024). This software ecosystem gap becomes especially significant when evaluating AI supercomputers across different hardware platforms, as systems based on Chinese AI chips may not achieve their theoretical potential without the mature software infrastructure that NVIDIA's CUDA provides.

Theoretical performance does not fully capture AI inference performance. Our database focuses on systems suitable for AI training. A system's computation performance is not a good proxy for how well it can run AI inference workloads. NVIDIA's H20, for instance, delivers comparable inference performance to the H100 on certain workloads despite having only 1/7th the raw computational power, due to its high memory bandwidth. We recommend differentiating between FLOP/s (or OP/s for 8-bit and lower) when assessing training capabilities and memory bandwidth in Byte/s when assessing inference and long-context capabilities.

C.2.3. Limitations with our Chinese data

Despite involving Chinese speakers in our data collection, we encountered several significant challenges in gathering comprehensive data on Chinese AI supercomputers.

1. **Official announcements often lack key data**, such as information on chip type and quantity. Furthermore, reported performance values often do not include precision.
2. **Sources sometimes report aggregate data for several AI supercomputers.** Computing zones that consist of several separate data center campuses sometimes report total computing capacity at an aggregate level rather than breaking down by individual AI supercomputers.
3. **Different conventions.** Chinese sources sometimes use different metrics and reporting standards than Western conventions, sometimes reporting the number of server racks that we cannot easily convert to chip numbers.

While we encounter similar issues for AI supercomputers in other countries, they are particularly common in China. However, we estimate that our database covers 10–20% of Chinese AI supercomputer performance, which is similar to our

[15] We found it particularly challenging to verify this for reports from companies and for AI supercomputers in China.

coverage estimate for U.S. data (see Appendix C.3).

C.2.4. Chinese owners may have become more secretive about their AI supercomputers, but this has not impacted our data coverage

In the late 2010s and early 2020s, Chinese supercomputer announcements frequently led to U.S. sanctions, with companies like Sugon, Phytium, and several national supercomputing centers being added to the Entity List due to concerns about military use of these systems (U.S. Bureau of Industry and Security, 2019; U.S. Department of Commerce, 2021). This is likely what caused China to release less information about its AI supercomputers. In 2022, China stopped submitting any systems to the Top500 list (Chik, 2022).

In October 2022, the United States first introduced export controls on AI chips and semiconductor manufacturing equipment with the goal of slowing down Chinese advances in AI (Allen, 2022). These export controls were strengthened in October 2023 and December 2024 by fixing loopholes and further restricting Chinese import of chip manufacturing tools (Dohmen & Feldgoise, 2023; Allen, 2024). These actions may have incentivized Chinese owners to further increase secrecy about their AI supercomputers to reduce scrutiny from the United States, particularly if they deployed smuggled AI chips.

However, the effects of increased Chinese secrecy on our data coverage are limited. While we see a decrease in the number of Chinese systems added to our database in 2021 and 2022, the number of Chinese systems increased again in 2024 (Figure 6). Comparing the aggregate performance in our database with IDC (2025)'s estimate of total 16-bit FLOP/s in China indicates that our coverage was consistently between 10 and 20% of Chinese performance (see Table 4).

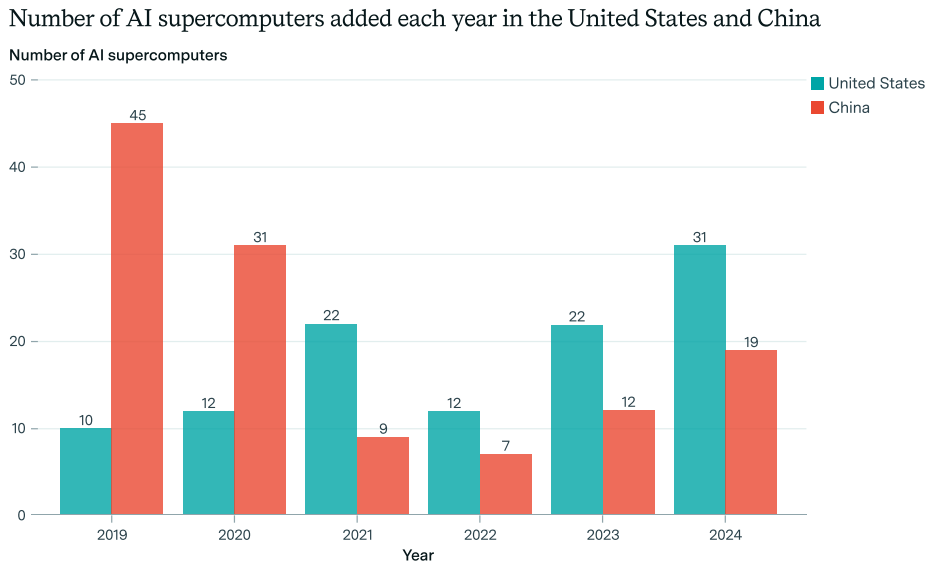

Figure 6. Number of Chinese and U.S. systems added each year.

C.3. Comparing our data with public reports

To assess what fraction of AI supercomputer capacity we capture in the dataset and how our coverage differs between chip types and companies, we compare our data to four sources of public information:

- Estimates of the total production of AI chips.
- Estimates of the total AI chip stock of companies.
- An estimate of the total 16-bit FLOP/s in China by IDC (2025).
- The fraction of the largest publicly known AI models that were likely trained on an AI supercomputer in our dataset.

C.3.1. Estimating the coverage of all AI supercomputers based on total chip production

One relevant reference point for our coverage is what fraction of total production we cover for different chip types (Table 2). While some AI chips may be sold to individuals and small research groups, we expect that the vast majority of all AI chips

will be used in AI supercomputers that would fall within our definition.

Table 2. Variation of coverage by chip type based on public reports of AI chip production until 2025. Note that the public estimates may include chips in AI supercomputers that are not yet operational or that are otherwise outside of our inclusion criteria. Our full dataset includes potentially existing and planned systems and has a higher coverage. Note that we explicitly search for the H100, A100, and V100 in our automated methodology. This may marginally increase our coverage of these three chip types compared to others.

CHIP TYPE	PUBLIC ESTIMATE	DATASET	IMPLIED COVERAGE
H100/H200	2.5M – 4.5M[16]	830K	36.5% – 20.3%
A100	1.5M – 3M[17]	234K	16.1% – 8.1%
H20	1M[18]	– [19]	0%
H800/A800	>200K[20]	2K	<1.5%
AMD MI300	400K[21]	72K	18%
GOOGLE TPUS	>4M[22]	95K	<4%
OTHER CUSTOM SILICON	?[23]	4K	?%
TOTAL	9.6 – 13.1M	1.2M	9.2% – 12.5%

Based on the public sources used in the table, our dataset covers between 20% and 37% of all NVIDIA H100s produced until late 2024.[24] However, coverage is much worse for NVIDIA's H20, A800 and H800, Google's TPUs, and other custom silicon chips. The average coverage is about 10%. (Note that the table above only includes confirmed operational AI supercomputers. Our dataset also contains planned AI supercomputers that make up another 920k H100s and 33k MI300X. Some of those may include chips already included in the production volume estimates.)

Table 2 reveals that our dataset likely covers H100, A100, and MI300 equally well, whereas coverage of Google's TPUs and other custom silicon chips is significantly worse. This is expected, given that NVIDIA and AMD sell their chips to a wide range of customers, incentivizing them to report about successful projects to attract more customers. Meanwhile, Google and other hyperscalers only deploy their chips within the company, offering limited incentives to publish more than a few large AI supercomputers.

C.3.2. COVERAGE BY COMPANY

Another reference point for our coverage is comparing our chip numbers to the publicly reported numbers of chips acquired by different companies (Table 3). We expect that hyperscalers deploy most of their AI chips in AI supercomputers covered by our definition, since even when primarily running inference workloads, they usually deploy thousands of AI chips in the same data center. (Note that the March 2025 inclusion threshold was at 2,000 H100-equivalents but was below 1,000 H100-equivalents until August 2024.)

[16]Public sources estimate that NVIDIA shipped about 500k H100s in 2023 and 2 million in 2024, for a total of 2.5 million H100s (Nolan, 2023; Shilov, 2023b). However, Garreffa (2024) estimates NVIDIA produced up to 1.5 million H100s in Q4 of 2024. Assuming NVIDIA produced about 1M H100s on average per quarter in 2024 yields a total of 4.5 million H100s.

[17]Reports on how many A100s NVIDIA produced are limited, but the company reportedly shipped 500k in Q3 2023 (Shilov, 2023a). The A100 was first produced in 2020 and likely reached peak production in 2023 before demand reduced in 2024. It thus seems plausible that NVIDIA produced between 1.5 – 3 million A100s until 2025.

[18]Financial Times (2023)

[19]We capture 30k H20s that DeepSeek likely owns, but exclude these from the analysis because we are uncertain if they are in the same location.

[20]Public reports indicate Chinese companies spent $5 billion on NVIDIA H800 and A800 in 2023 (Pires, 2023a), indicating at least 200k of these chips imported (conservative estimate assuming $25k average price per chip (Champelli et al., 2024) .)

[21]AMD to ship up to 400,000 new AI GPUs in 2024 (Chen & Chan, 2023).

[22]Google's internal TPU production likely reached 2 million TPUs in 2023 (Martin, 2024), although public data is severely limited, given Google does not sell TPUs to outside companies. Assuming a similar production in 2024, there would be at least 4 million TPUs.

[23]Microsoft, AWS, and Meta all developed their own custom silicon AI chips deployed in-house (Borkar et al., 2024; AWS, undated; Tal et al., 2024), but we were unable to find trustworthy public estimates of the total numbers.

[24]We do not account for H100s produced in 2025, since these would unlikely be installed in any systems before our March 1st cutoff.

Table 3. Public reports of number of chips owned for various companies at the end of 2024 and comparison with our dataset[26]

COMPANY	PUBLIC CLAIM	OUR DATASET	IMPLIED COVERAGE
META	350K H100S	149K	42.8%
MICROSOFT	475K – 855K H100[27]	118K	14% – 25%
AWS	200K H100S (IN 2024)	–	0%
GOOGLE	170K H100S (IN 2024)	8K	4.7%
APPLE	180K H100S[28]	–	0%
COREWEAVE	175K GPUS[29]	57K	22.8%
BYTEDANCE	310K HOPPERS[30]	8K	3%
TENCENT	230K HOPPERS (IN 2024)	–[31]	0%
TOTAL	2.09M – 2.47M	0.34 M	13.8 – 16.3%

Note: Public estimates cannot be verified and only serve as an approximate assessment of coverage. Some sources are inconsistent with others.

Table 3 shows that our coverage differs considerably between companies. While we cover almost half of Meta's H100s, we cover only 5% of Google's and none of Apple's H100s. Our data is particularly limited for Chinese hyperscalers. However, Table 3 does not consider AI supercomputers we cover based on reported performance, but for which we lack the specific chip type. This is especially common for Chinese systems.

C.3.3. COVERAGE OF CHINESE DATA

To assess data coverage of AI supercomputers in China, we compare the aggregate 16-bit performance of all Chinese systems in our database to the total Chinese 16-bit performance published in a 2025 report by market intelligence firm International Data Corporation (IDC, 2025). We find that we cover between 10 and 20% of Chinese 16-bit performance between the end of 2020 and the end of 2024 (Table 4). Not all 16-bit performance would likely fall under the definition of our database, so actual coverage of AI supercomputers is likely somewhat higher.

Table 4. FP16 Performance

	OUR DATA	IDC	IMPLIED COVERAGE
2020	1.05×10^{19}	7.50×10^{19}	14%
2021	1.88×10^{19}	1.55×10^{20}	12%
2022	3.46×10^{19}	2.60×10^{20}	13%
2023	4.18×10^{19}	4.17×10^{20}	10%
2024	1.46×10^{20}	7.25×10^{20}	20%

We were unable to find reliable total performance estimates for other countries, so we had to limit our coverage analysis by FLOP/s to Chinese data.

[25]Note we only include systems in our analysis if we are confident they exist in a single site rather than a distributed system. E.g., AWS announced 20k H100 clusters in 2023, but did not explicitly say whether or not those were on the same data center campus.

[26]Note we only include systems in our analysis if we are confident they exist in a single site rather than a distributed system. E.g., AWS announced 20k H100 clusters in 2023, but did not explicitly say whether or not those were on the same data center campus.

[27]Microsoft likely made up 19% of total 2023 NVIDIA revenue (Fox, 2024). We assume they maintained a 19% share of revenue throughout 2024, and bought a mix of NVIDIA data center products that is approximately equal to NVIDIA's sales mix. Based on estimates for H100 shipments in our previous section, this indicates Microsoft owns between 475k and 855k H100s.

[28]∼2,500 servers in 2023 and 20,000 servers in 2024 * 8 GPUs per server = 180k.

[29]Estimate, given the claim that most of the 250,000 total GPUs said to be H100s and some H200s (Morgan, 2025).

[30]About 50k H100 in 2023 and 240k in 2024 (Pires, 2023b; Alexsandar K, 2024).

[31]We identified two Tencent AI supercomputers but were unable to identify the performance or hardware used.

C.3.4. Coverage of AI supercomputers used in the largest training runs

To check how well our dataset covers the AI supercomputers used for known large training runs, we check which of the 25 largest training runs in Epoch AI's notable AI models dataset (as of 1 March 2025) correspond to AI supercomputers in our dataset (Epoch AI, 2025). (Note that our dataset uses the models dataset as a data source. To avoid circularity, we distinguish between systems reported independently from the training run and systems included in our dataset based exclusively on the reports of the training run.)

We find that for about half of the largest AI training runs, we capture an AI supercomputer that could have plausibly been used or was confirmed to be used in the training run (Figure 7; Table 5).

Our data coverage is slightly better for Chinese AI supercomputers, where we find plausible AI supercomputers for about two-thirds of all reported models (Figure 7; Table 6).

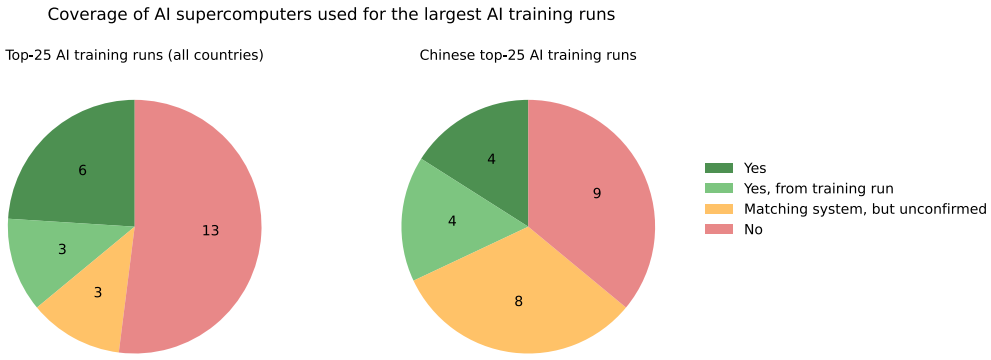

Figure 7. Coverage of AI supercomputers used for the largest AI training runs according to Epoch AI's notable models dataset. "Yes, from training run" indicates we cover the AI supercomputer, but only based on reports about the training runs itself. "Matching system, but unconfirmed" means an AI supercomputer in our dataset was likely used by the model developer but we find no public reports on whether or not the system was actually used for the training run.

Table 5. Coverage of largest AI training runs (all countries) according to Epoch AI's notable model dataset

TRAINING RUN	COVERED	NOTE
GROK-3	Yes	Trained on Colossus in Memphis, Tennessee
GEMINI 1.0 ULTRA	Yes, from training run	
GPT-4O	No	
LLAMA 3.1-405B	Yes	Presumably trained on Meta GenAI 2024a or 2024b (Oldham et al., 2024)
CLAUDE 3.5 SONNET	No	
GLM-4-PLUS	No	
CLAUDE 3.7 SONNET	No	
GROK-2	Matching AI supercomputer, but unconfirmed	Trained on the Oracle Cloud. "Oracle OCI Supercluster H100s" matches the description of the training details (Trueman, 2024)
DOUBAO-PRO	No	
GPT-4 TURBO	No	Possibly trained on same AI supercomputer as GPT-4, but no confirmation
MISTRAL LARGE 2	No	
GPT-4	Yes	Likely trained on Iowa AI supercomputer (O'Brien & Fingerhut, 2023). Entered in the dataset as "Microsoft GPT-4 cluster"
NEMOTRON-4 340B	Matching AI supercomputer, but unconfirmed	"NVIDIA CoreWeave Eos-DFW" appears to match the training description
CLAUDE 3 OPUS	No	
GEMINI 1.5 PRO	No	We capture several systems from Google, but none were likely used for this model
GLM-4 (0116)	No	
MISTRAL LARGE	Yes	Likely used Leonardo
ARAMCO METABRAIN AL	No	
INFLECTION-2	Yes, from training run	
INFLECTION-2.5	No	We capture several of Inflection's systems, but none were confirmed
REKA CORE	Yes, from training run	
LLAMA 3.1-70B	Yes	Presumably trained on Meta GenAI 2024a or 2024b (Oldham et al., 2024)
LLAMA 3-70B	Yes	Trained on Meta GenAI 2024a or 2024b (Oldham et al., 2024)
QWEN2.5-72B	Matching AI supercomputer, but unconfirmed	
GPT-4O MINI	No	

Table 6. Coverage of AI supercomputers used for the largest AI training runs in China according to Epoch AI (2025) as of March 2025.

MODEL	COVERED
GLM-4-PLUS	No
DOUBAO-PRO	No
GLM-4 (0116)	No
QWEN2.5-72B	Matching AI supercomputer, but unconfirmed
TELECHAT2-115B	Matching AI supercomputer, but unconfirmed
DEEPSEEK-V3	Yes
DEEPSEEK-R1	Yes
MEGASCALE (PRODUCTION)	Yes, from training run
SENSECHAT	Yes
QWEN2.5-32B	Matching AI supercomputer, but unconfirmed
HUNYUAN-LARGE	No
QWEN2-72B	Matching AI supercomputer, but unconfirmed
YI-LARGE	No
DEEPSEEK-V2.5	Matching AI supercomputer, but unconfirmed
YI-LIGHTNING	Yes, from training run
QWEN1.5-72B	Matching AI supercomputer, but unconfirmed
QWEN-72B	Matching AI supercomputer, but unconfirmed
XVERSE-65B-2	No
HUNYUAN	No
LUCA 2.0	No
QWEN2.5-CODER (32B)	Matching AI supercomputer, but unconfirmed
BLUELM 175B	No
ERNIE 3.0 TITAN	Yes
MEGASCALE (530B)	Yes, from training run
XTRIMOPGLM -100B	Yes, from training run

D. Additional data

D.1. Overview of trends in different precisions and by sector.

Table 7. Overview of key trends between 2019 and March 2025. Square brackets indicate the 90% confidence interval. Note 8-bit trend is only starting in July 2021.[33]

LEADING AI SUPERCOMPUTERS (INCLUDING BOTH PUBLIC AND PRIVATE)			
	16-BIT OP/S	**8-BIT OP/S**	**MAX OP/S**
PERFORMANCE GROWTH	**2.54** [2.35–2.74]	**2.60** [2.31–2.93]	**2.55** [2.34–2.78]
NUMBER OF CHIPS	**1.60** [1.45–1.78]	**1.69** [1.47–1.94]	**1.46** [1.29–1.64]
PERFORMANCE PER CHIP	**1.60** [1.49–1.71]	**1.54** [1.42–1.67]	**1.77** [1.62–1.94]
HARDWARE COST	**1.92** [1.76–2.11]	**1.99** [1.72–2.30]	**1.76** [1.58–1.97]
COST-PERFORMANCE	**1.36** [1.29–1.42]	**1.37** [1.29–1.45]	**1.51** [1.43–1.60]
POWER	**1.95** [1.77–2.15]	**2.12** [1.85–2.42]	**1.78** [1.60–1.99]
ENERGY EFFICIENCY	**1.34** [1.25–1.43]	**1.26** [1.20–1.32]	**1.51** [1.39–1.63]
LEADING PRIVATE AI SUPERCOMPUTERS			
	16-BIT OP/S	**8-BIT OP/S**	**MAX OP/S**
PERFORMANCE GROWTH	**2.69** [2.47–2.92]	**3.17** [2.78–3.61]	**3.00** [2.76–3.27]
NUMBER OF CHIPS	**1.82** [1.66–2.00]	**2.14** [1.85–2.47]	**1.83** [1.65–2.03]
PERFORMANCE PER CHIP	**1.50** [1.44–1.57]	**1.48** [1.36–1.61]	**1.65** [1.55–1.76]
HARDWARE COST	**2.06** [1.88–2.26]	**2.39** [2.09–2.73]	**2.05** [1.86–2.26]
COST-PERFORMANCE	**1.33** [1.28–1.39]	**1.32** [1.26–1.39]	**1.47** [1.41–1.54]
POWER	**2.16** [1.98–2.35]	**2.57** [2.26–2.93]	**2.16** [1.96–2.37]
ENERGY EFFICIENCY	**1.27** [1.23–1.31]	**1.23** [1.19–1.28]	**1.40** [1.34–1.46]
LEADING PUBLIC AI SUPERCOMPUTERS			
	16-BIT OP/S	**8-BIT OP/S**	**MAX OP/S**
PERFORMANCE GROWTH	**1.86** [1.60–2.15]	**1.79** [1.46–2.19]	**1.90** [1.63–2.22]
NUMBER OF CHIPS	**1.21** [0.98–1.50]	**1.20** [0.96–1.49]	**1.11** [0.89–1.38]
PERFORMANCE PER CHIP	**1.56** [1.34–1.82]	**1.48** [1.31–1.67]	**1.75** [1.45–2.11]
HARDWARE COST	**1.40** [1.25–1.57]	**1.34** [1.09–1.65]	**1.38** [1.20–1.58]
COST-PERFORMANCE	**1.41** [1.28–1.56]	**1.48** [1.32–1.66]	**1.51** [1.32–1.73]
POWER	**1.41** [1.17–1.70]	**1.38** [1.10–1.74]	**1.31** [1.07–1.61]
ENERGY EFFICIENCY	**1.38** [1.19–1.61]	**1.33** [1.19–1.47]	**1.56** [1.28–1.90]

[32]We assess this trend only after 50 AI supercomputers in our dataset support 8-bit precision.
[33]We assess this trend only after 50 AI supercomputers in our dataset support 8-bit precision.