State of AI: China

Artificial Analysis

Q1 2025

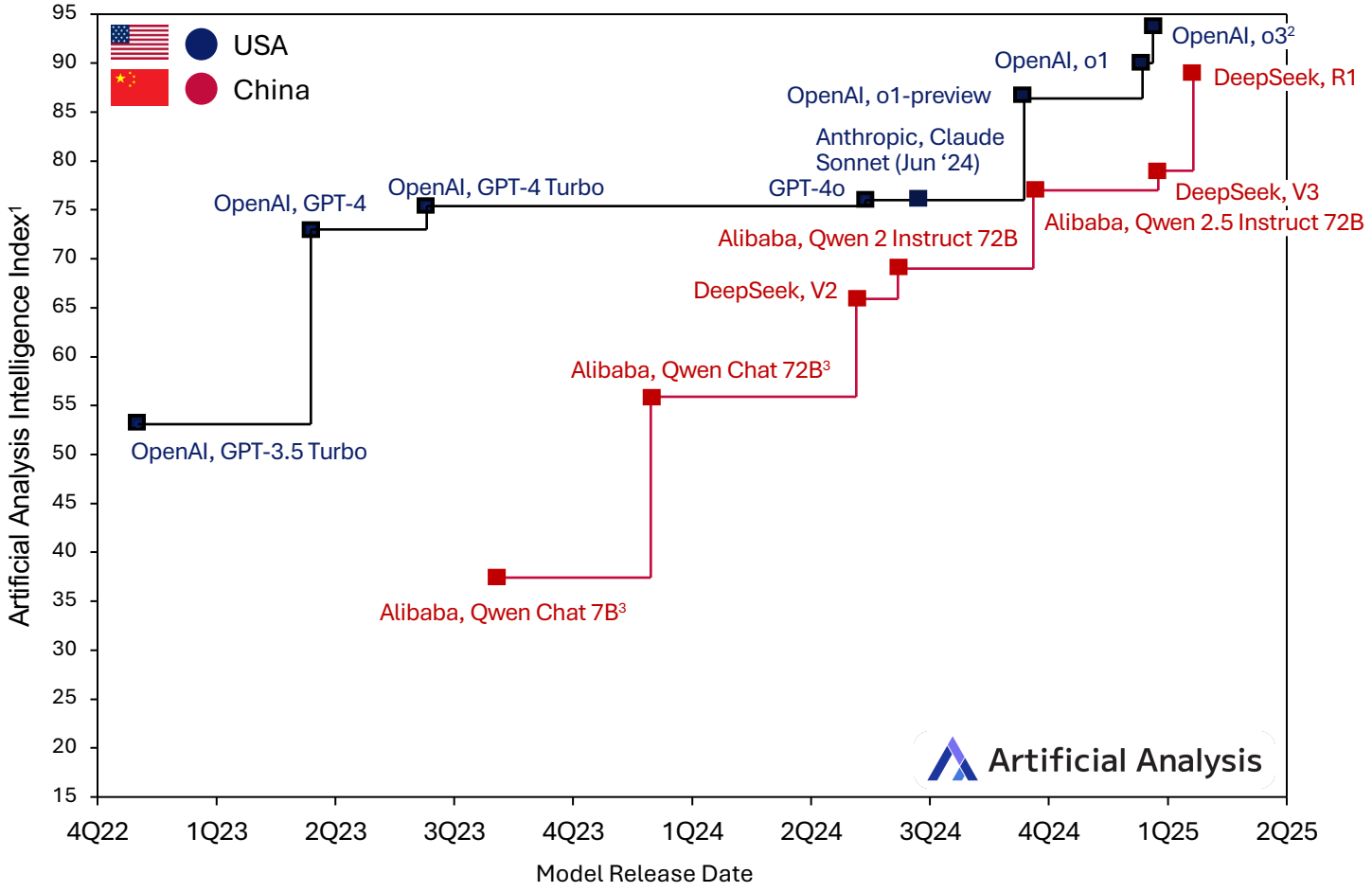
Artificial Analysis is a leading and independent AI benchmarking and insights provider. We support engineers and companies to understand AI capabilities and make critical decisions about their AI strategy.

Our data, insights and publications are grounded in our comprehensive benchmarking of AI technologies and use cases. This includes everything from hourly performance testing of language model APIs to millions of votes in our crowd-sourced arenas.

Our public website, artificialanalysis.ai, is widely referenced by companies leading innovation in AI. To discuss this report, our publications, or our services, please get in touch at contact@artificialanalysis.ai.

Chinese AI labs have progressively caught up to US AI labs; models from Chinese labs are now approaching o1-level intelligence with the release of DeepSeek's R1 model

US & China: Frontier Language Model Intelligence, Over Time[1]



Key Trends

Closing the gap: The final months of 2024 have seen the emergence of the numerous highly performant models from top Chinese AI labs. This has resulted in the delta between the level of intelligence offered by models from Chinese AI labs and US AI labs closing. Several Chinese models are now competitive with models from the top US labs.

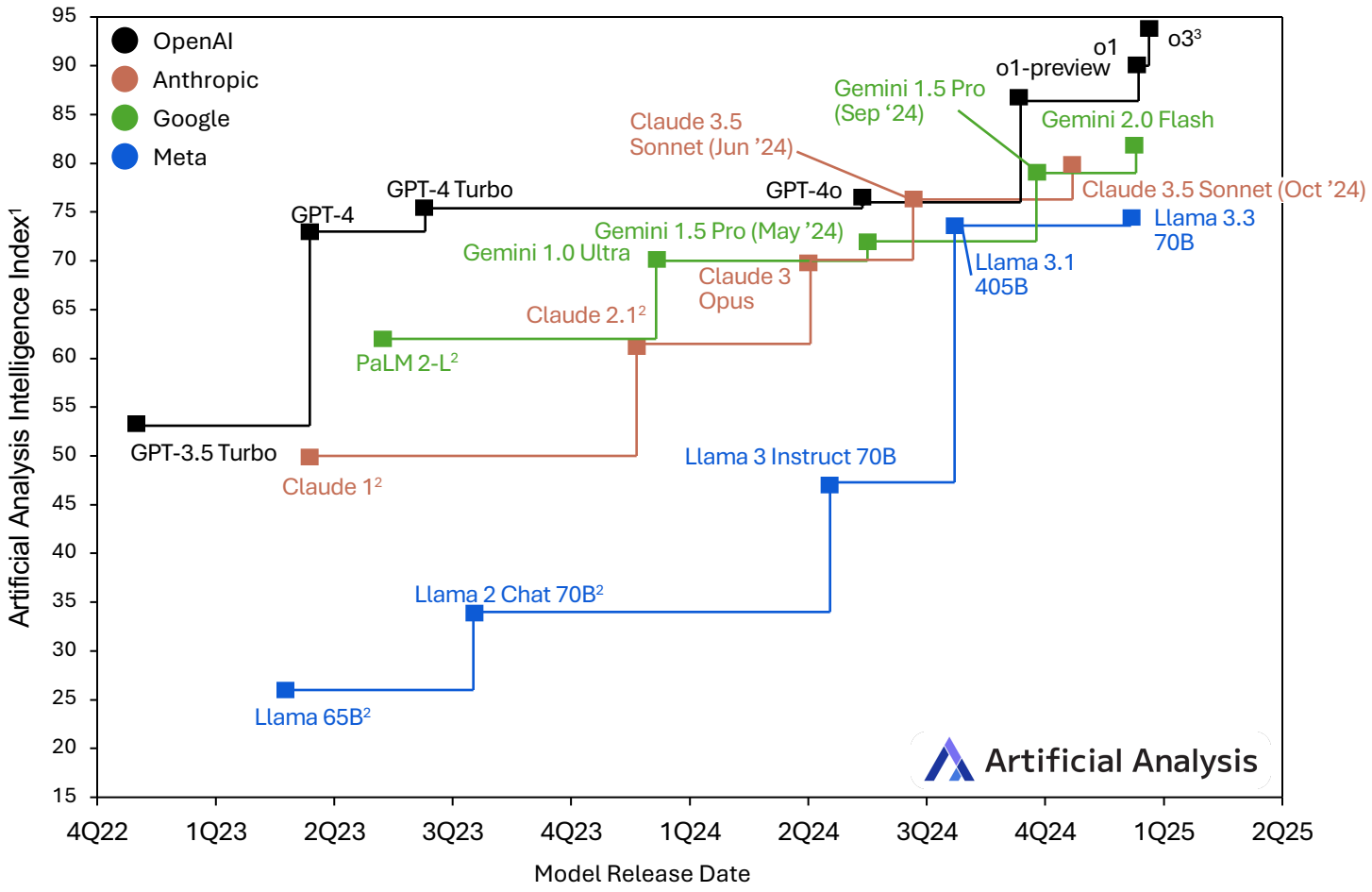
Reasoning models quickly becoming commonplace: Reasoning models (that "think" before answering) were first introduced by OpenAI in 3Q24. Within months, Chinese competitors, led by DeepSeek, have largely replicated the intelligence of o1. Several AI labs in China now have a frontier-level reasoning model.

Open models close in on the frontier labs: Open weights models, led by those from DeepSeek and Alibaba, have approached o1level intelligence.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. o3 Intelligence Index estimated by scaling measured Intelligence Index of o1. 3. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis

Since the launch of OpenAI's GPT-4 in early 2023, leading US AI labs have scrambled to catch up to OpenAI

Leading US AI Labs Frontier Language Model Intelligence, Over Time[1]



Key Trends

Competing labs catch up to OpenAI's GPT-4: OpenAI started the language model race in November 2022 with the launch of GPT-3.5 in ChatGPT; leading US labs have largely caught up with frontier models from OpenAI.

Big Tech closes in on the frontier labs: Models from Google and Meta are rapidly closing in on frontier models, with Gemini 2.0 Flash exceeding Claude 3.5 Sonnet and GPT 4o capabilities.

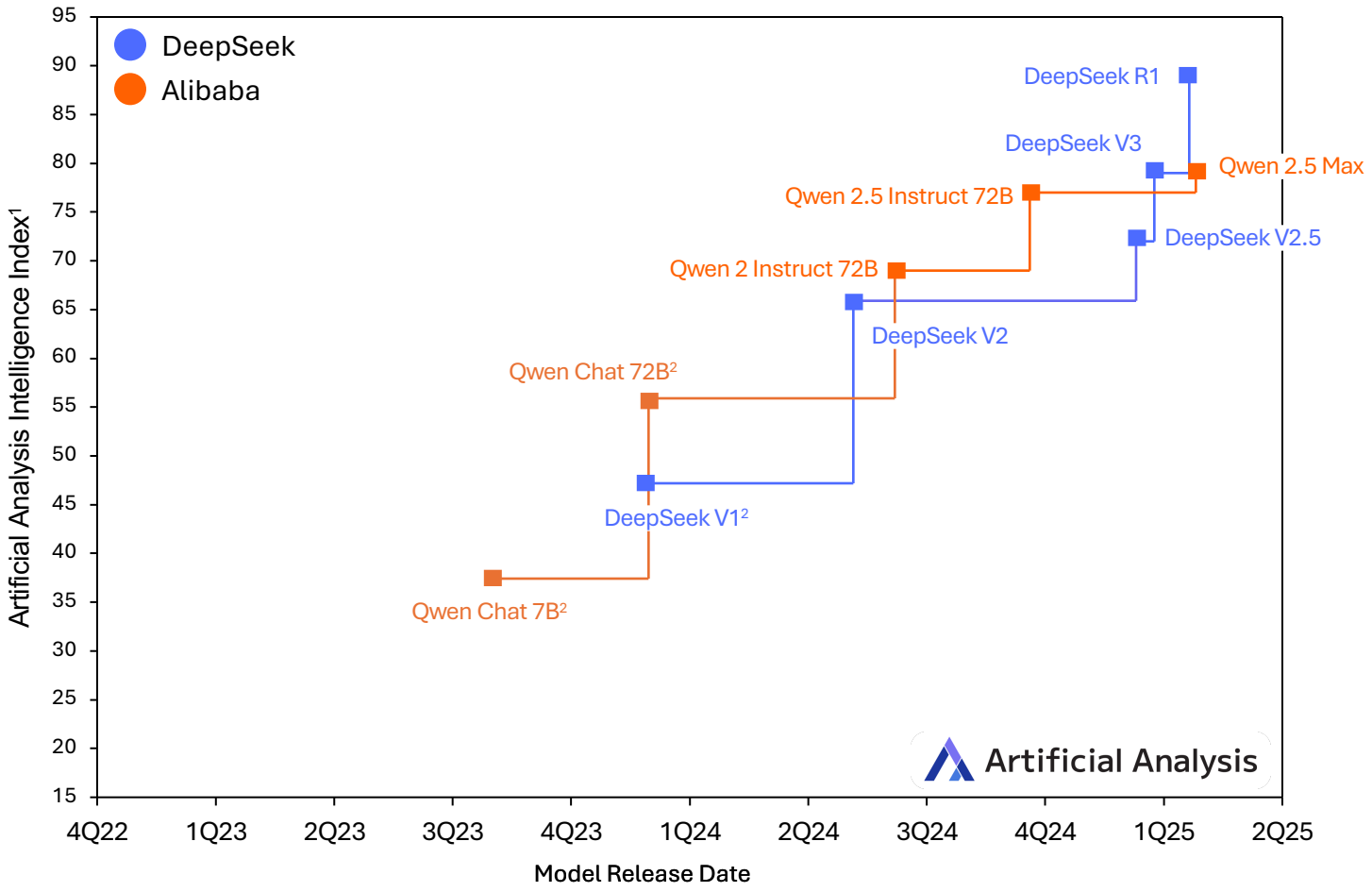
Sparks of intelligence beyond GPT-4: The final months of 2024 have seen the emergence of the first major intelligence leaps beyond GPT-4, led by OpenAI's o3. Topics including reasoning models, data quality and new reinforcement learning techniques have joined pre-training compute scaling as dominant levers for improving models.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis. 3. o3 Intelligence Index estimated by scaling measured Intelligence Score of o1.

Artificial Analysis

Leading Chinese AI labs DeepSeek and Alibaba have steadily released new models, with DeepSeek taking the lead from Alibaba in late 2024

Leading Chinese AI Labs Language Model Intelligence, Over Time[1]



Key Trends

Rapid improvements in intelligence: While Chinese AI labs joined the AI race later, they largely closed the intelligence gap with frontier US models in 2024. When OpenAI launched o1, Chinese labs produced a similarly performant model within months (DeepSeek's R1).

Leading with open weights models: Chinese AI labs, including Alibaba, DeepSeek and Tencent, have released open weights frontier models that are competitive with the leading models globally.

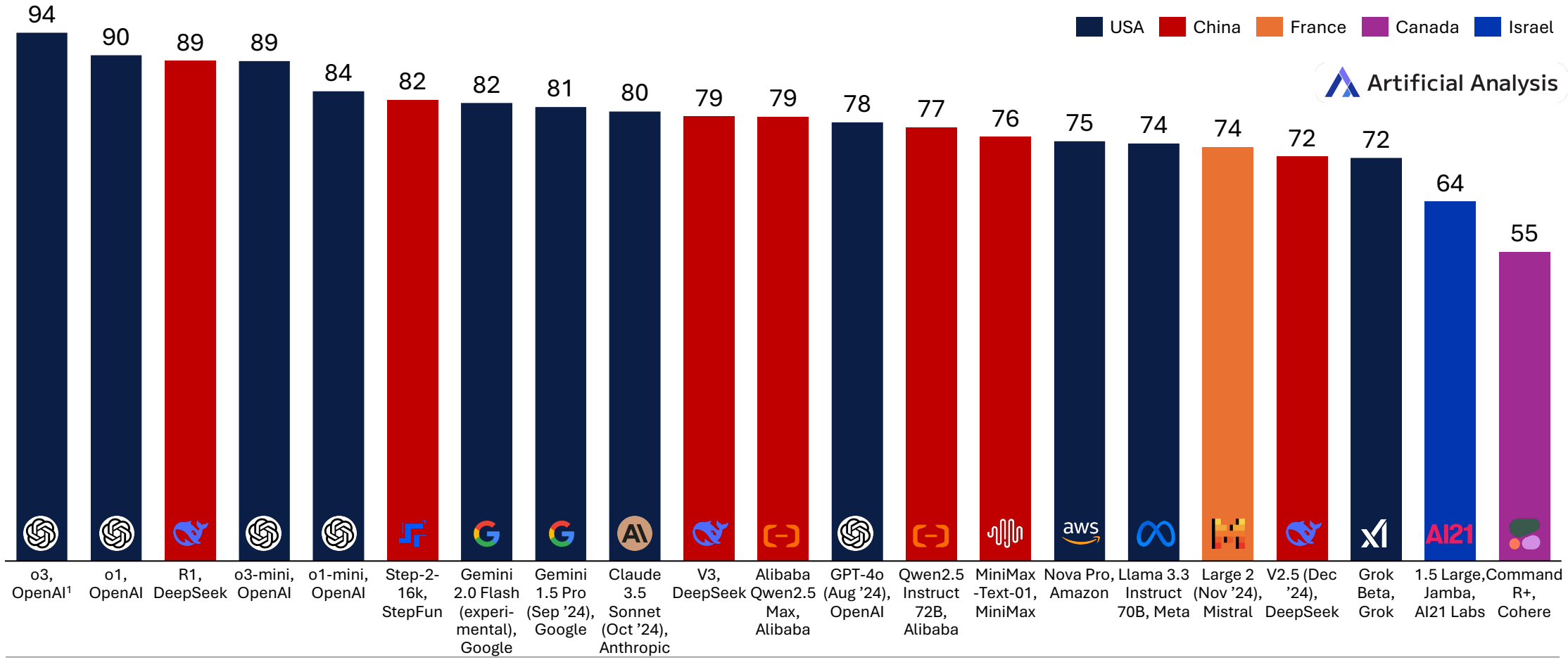
Potential leader in 2025: Early 2025 saw Chinese AI labs, including Alibaba, DeepSeek, MoonShot, Tencent, Zhipu, and Baichuan prolifically releasing frontier reasoning models. The release velocity and cadence suggest that Chinese AI labs are no longer laggards in 2025.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis

While the US maintains an overall lead in the intelligence frontier, China is no longer far behind. Few other countries have demonstrated frontier-class training

The Language Model Frontier: Country of Origin

Artificial Analysis Intelligence Index, Selected Leading Models (Early 2025), Non-exhaustive



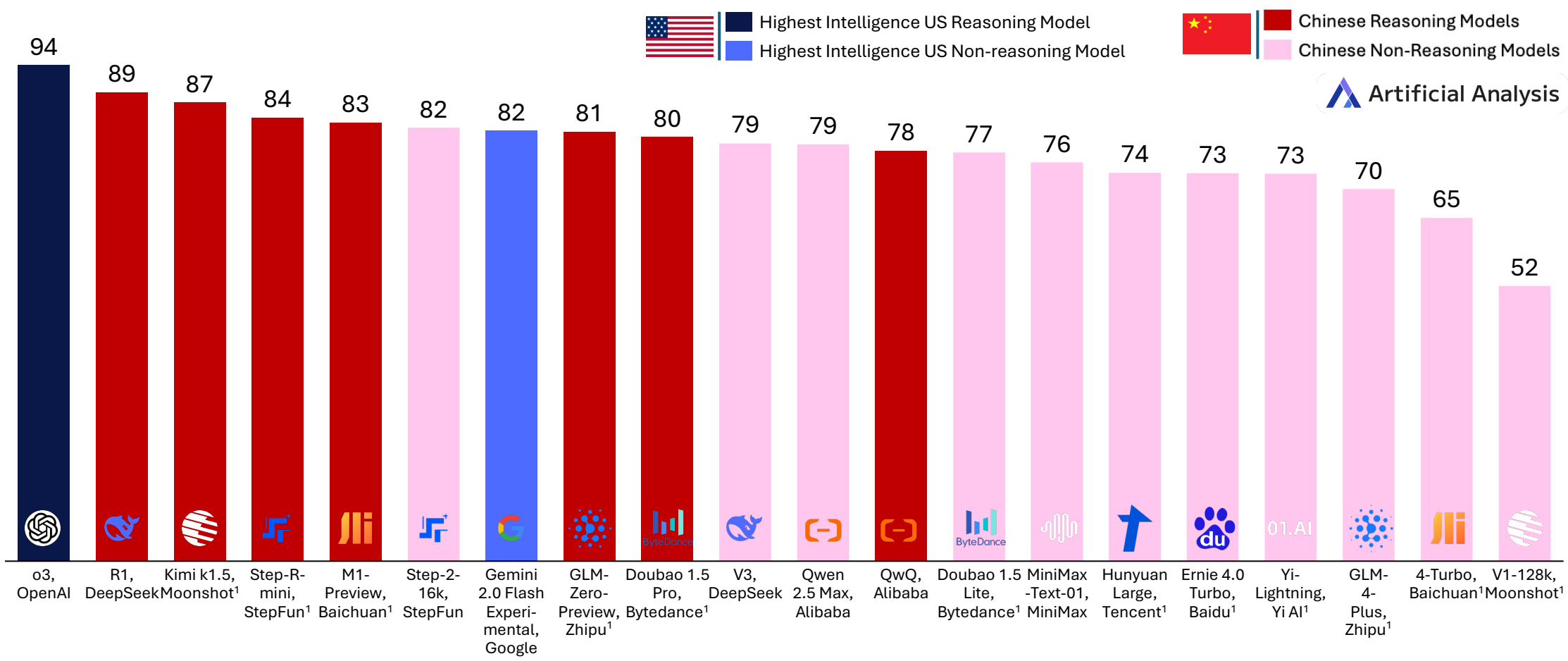
1. *Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis*
2. *A number of leading models from Chinese AI labs are excluded due to limited access or evaluation data*

Artificial Analysis

As of early 2025, several Chinese AI labs have demonstrated or claimed frontier-level intelligence, with seven releasing models featuring reasoning capabilities

The Language Model Frontier: Models by Chinese AI Labs

Artificial Analysis Intelligence Index, Leading Models (Early 2025), Non-exhaustive



1. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis

Artificial Analysis

The leading Chinese Big Tech firms are actively competing in the AI race and have released AI language models as well as models across other modalities

Non-Exhaustive

</> Open Weights LLM

Frontier Models by Chinese Big Tech Firms

		Alibaba.com	Baidu 百度	ByteDance	HUAWEI	Tencent 腾讯
	Description	Large ecommerce player and Hyperscaler (Alibaba Cloud), largest shareholder of Ant Group	China's largest search engine, and operator of Wenxin Yiyan, an AI chatbot with a reported ~300m users	Parent company of Douyin (TikTok) and Toutiao, one of China's leading news applications	Global telco leader and one of the world's largest smartphone manufacturers	Parent company of Riot Games and WeChat, the 'all-in-one 'super app' of China; Hyperscaler with their Tencent Cloud offering
	AI Strategy *(high-level)*	• Release open weights models • More recently launched proprietary models • Offer inference on Alibaba Cloud	• Actively integrating proprietary models into search platform • Long time leader in self-driving AI	• Develop proprietary models and integrate across their consumer platforms	• Develop proprietary, domain-specific models and offer on Huawei Cloud	• Release open weights models and offer proprietary models on Tencent Cloud
Best LLM[4]	Non-Reasoning	Qwen 2.5 Max **Intelligence: 79**	Ernie 4.0 Turbo **Intelligence: 76**[5]	Doubao 1.5 Lite **Intelligence: 77**[5]	Pangu 5.0 Large	Hunyuan Large **Intelligence: 74** </>
	Reasoning	QwQ **Intelligence: 78**[5] </>	–	Doubao 1.5 Pro **Intelligence: 80**[5]	–	–
Other Models	Text to Speech	✓	✓	✓	✓	✓
	Speech to Speech	–	–	✓	–	–
	Image Generation	✓	✓	✓	✓	✓
	Video Generation	✓	✓	✓	✓	✓
	3D Generation	–	–	✓	–	✓
	Primary Consumer Apps	Tongyi Qianwen	Wenxin Yiyan, Wenxin Yige	Doubao	Celia	Yuanbao, Yuanqi
	Valuation (US$)	235B[1]	32B[1]	300B[2]	128B[3]	469B[1]

Other Firms with AI Ambitions

Kunlun Tech
SHE: 300418 (Mkt Cap: $6B)[1]
Beijing-based internet group with >300m MAUs; owner of the Opera browser. Launched the SkyWork series of models and AI accelerators

360 Security (Qihoo 360)
SHA: 601360 (Mkt Cap: $11B)[1]
China's largest provider of Internet and mobile security products. Launched the Zhinao series of models under the 360 AI brand

iFlytek
SHE: 002230 (Mkt Cap: $16B)[1]
Leading voice AI company in China with >14,000 employees. Launched the Spark series of models

Meituan
HKG: 3690 (Mkt Cap: $115B)[1]
China's leading shopping platform with >600m DAUs. Cofounder Wang Huiwen returned to lead AI efforts. Investor in multiple frontier AI labs

Xiaomi
HKG: 1810 (Mkt Cap: $123B)[1]
China's leading consumer electronics brand. Launched the MiLM series of small models. Recently poached Luo Fuli, DeepSeek researcher, to run AI lab. Investor in multiple frontier AI labs

1. Market cap as per Reuters (aa 31 Jan 25) 2. ByteDance is a private company. Valuation by Reuters 3. Huawei is a private company. Valuation by Reuters (2023) 4. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. 5. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis

Chinese AI startups, with the support of Chinese Big Tech firms and the Chinese Government, have developed some of the world's leading open weights models

Non-Exhaustive

Frontier Models by Chinese AI Tigers and Startups

</> Open Weights LLM State Backed Entity

		MINIMAX	Moonshot AI	零一万物 01.AI	deepseek	ZHIPU·AI	百川智能 BAICHUAN AI	阶跃星辰 Stepfun
	Description	China AI Tiger and publisher of Talkie AI app (4th most downloaded in US in 1H24)	China AI Tiger with 2M Chinese character context window model; China's most well-funded AI startup based on available information	Chinese AI startup focused on smaller language models founded by Lee Kai-Fu (author, former head of Google China)	Chinese AI lab originating out of an AI-focused quantitative trading firm	China AI Tiger with nearly ~700k enterprise and developer users	China AI Tiger with a focus on medical AI models founded by Wang Xiaochuan (ex-CEO, Sogou)	First Chinese AI startup to develop a trillion-parameter model; founded by Jiang Daxin (ex-Chief Scientist, Microsoft Research Asia)
Best LLM[1]	Non-Reasoning	MiniMax-Text-01 </> **Intelligence: 76**	V1-128k **Intelligence: 52**	Yi-Lightning **Intelligence: 73**	V3 </> **Intelligence: 79**	GLM-4-Plus **Intelligence: 70**	Baichuan 4-Turbo **Intelligence: 65**	Step-2-16k **Intelligence: 82**
	Reasoning	–	Kimi k1.5 </> **Intelligence: 87**	–	R1 </> **Intelligence: 89**	GLM-Zero-Preview **Intelligence: 81**	Baichuan M1-Preview **Intelligence: 83**	Step-R-mini **Intelligence: 84**
Other Models	Text to Speech	✓	–	–	–	–	–	✓
	Speech to Speech	–	–	–	–	✓	✓	–
	Image Generation	–	–	–	✓	✓	–	✓
	Video Generation	✓	–	–	–	✓	–	–
	3D Generation	–	–	–	–	–	–	–
	Primary Consumer Apps	Hailuo AI Chat, Hailuo AI Video	Kimi	YiChat	DeepSeek Chat	ChatGLM	Bai Xiaoying	Yuewen, PopDuck
	Funding Raised ($)	0.85B[2]	1.67B[3]	0.2B[4]	Unknown	1.12B[5]	1.04B[6]	Unknown
	Notable Investors *(non-exhaustive)*	Tencent 腾讯, Alibaba.com, HongShan 红杉中国, 高榕 GAORONG VENTURES	Tencent 腾讯, Alibaba.com, 小红书, Microsoft, HongShan 红杉中国, ZhenFund 真格基金	Alibaba.com, SINOVATION VENTURES, 顺为 SHUNWEI	幻方量化 HIGH-FLYER QUANT	Alibaba.com, Tencent 腾讯, HongShan 红杉中国, QIMING VENTURE PARTNERS, Lightspeed China Partners, Meituan 美团, 清华控股 TSINGHUA HOLDINGS	Alibaba.com, Tencent 腾讯, 清华控股 TSINGHUA HOLDINGS, SOSV, 顺为 SHUNWEI	Tencent 腾讯, QIMING VENTURE PARTNERS, 上海国有资本投资有限公司

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. 2. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis 2. Pitchbook (Mar 2024) 3. Pitchbook (Aug 2024) 4. Pitchbook (Dec 24) 5. Pitchbook (Jul 24) 6. Pitchbook (Aug 24)

Artificial Analysis

EXPORT RESTRICTIONS TIMELINE

Escalating regulatory restrictions have banned the export of high-end AI accelerators to China (1/2)

Regulatory Restrictions

Unreleased | No Licence Required | NAC License Required | Presumption of Denial

NVIDIA GPU Architecture	Model	Pre-Controls	October 2022 Controls[2]	October 2023 Controls[3,4]	AI Diffusion Rules[5]
	Announced		7-Oct-22	17-Oct-23	13-Jan-25
	Effective[1]		21-Oct-22	17-Nov-23	15-May-25
Blackwell	B200				
	B100				
Hopper	H100				
	H200				
	H800				
	H20				
Lovelace	L40S				
	L4				
	L40				
	L20				
	L2				
Ampere	A100				
	A800				
	A40				
	A30				
Consumer GPUs	RTX 6000 Ada				
	RTX 4090				
	RTX 4090D				
	RTX 3090				

1. Effective date refers to latest compliance date 2. BIS 3. Georgetown CSET
4. Federal Register 5. BIS

Commentary

- **NVIDIA reacted quickly to both the October 2022 and October 2023 controls by releasing Hopper GPU variants that complied/comply with the regulations.** Specifically, after the H100 and A100 were banned for export to China, NVIDIA released the H800 and A800 with limited interconnect (see appendix for full Hopper generation specifications).
- **The October 2023 controls went on to ban export of the H800 and A800 to China,** leading to NVIDIA **developing the H20** to continue selling a Hopper-generation GPU to Chinese customers. The H20 has limited compute (148 TFLOPs) compared to the H100 (989 TFLOPs)

EXPORT RESTRICTIONS TIMELINE

Escalating regulatory restrictions have banned the export of high-end AI accelerators to China (2/2)

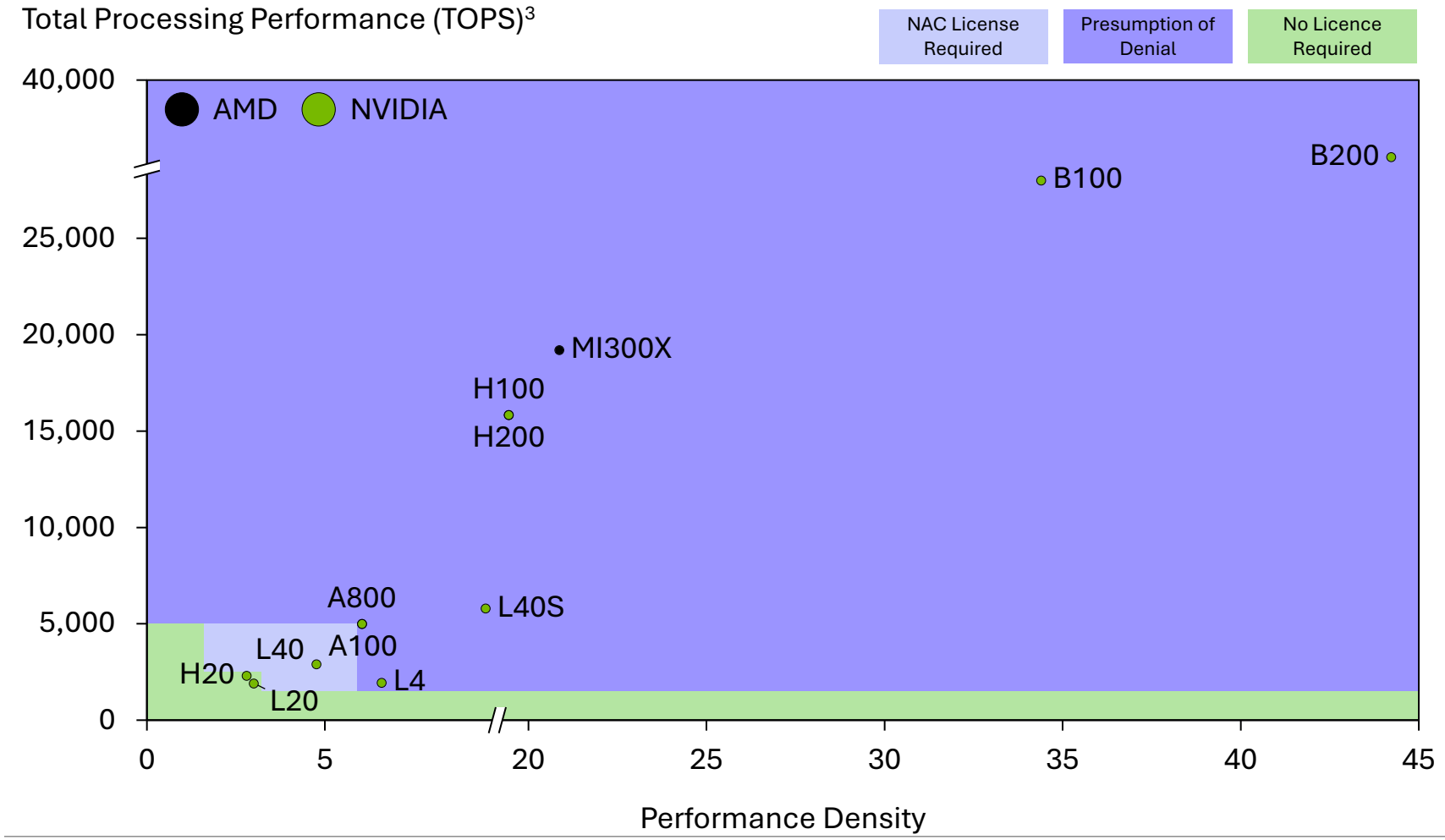
Regulatory Restrictions

Rule	Summary	Dates[1]	Impact[2]	
October 2022 Controls[3]	Initial restrictions on frontier GPUs. Both performance and interconnect thresholds had to be breached for the GPU to be restricted.	Announced: 7-Oct-22 Effective: 21-Oct-22	**Restriction Classification** Total Processing Performance (TPP) Interconnect Bandwidth	**Criterion** TPP ≥ 4,800 TPP ≥ 600 GB/s
October 2023 Controls[4,5]	Revised framework to prevent workarounds. Restricted exports of GPUs to China based on TPP or Performance Density (PD)	Announced: 17-Oct-23 Effective: 17-Nov-23	**Groupings** Group 1: Presumption of denial Group 2: Restrictive NAC licensing review Group 3: No Restrictions	**Criterion (Datacenter GPUs)** TPP ≥ 4,800 or TPP ≥ 1600 *AND* PD ≥ 5.92. 2,400 ≤ TPP ≤ 4,800 *AND* PD ≥1.6 or TPP ≥ 1,600 *AND* PD ≥ 3.2. TPP <1,600 or PD < 3.2
BIS Final Rule[6]	Crackdown on indirect imports by Chinese-affiliated chip manufacturing entities	Announced: 2-Dec-24 Effective: 31-Dec-24 Updated: 16-Jan-24	• Did not impact restricted chips • 140 entities (majority Chinese) from advanced chip sector now face a presumption of denial and added to Entity List in Dec 24[7]	
AI Diffusion Rule[8]	Extensive three-tiered licensing framework segregating access to GPUs by countries	Announced: 13-Jan-25 Effective:15-May-25	• Tier 3 countries (including China) face a de facto ban on advanced AI chips • All exports of controlled chips to these Tier 3 countries now require an export license, subject to a presumption of denial during review • Tier 2 countries now face limitations on large orders of AI chips	
AI Due Diligence Rule[9]	Companion KYC rule for AI Diffusion Rule	Announced: 16-Jan-25 Effective: 31-Jan-24	• Requires companies to conduct KYC-like compliance checks on their customers and comply with the AI Diffusion Rule	

1. TPP measured in Tera Operations per Second, PD measured as TPP / Die Size. 2. Effective date refers to latest compliance date
3. BIS 4. Georgetown CSET 5. Federal Register 6. BIS 7. Federal Register 8. BIS 9. Federal Register

US export controls restrict export of leading Nvidia accelerators based on performance and density thresholds; the H20 and L20 fall below these thresholds and can be freely exported

US Accelerators Prohibited for Export to China[1,2]



1. SemiAnalysis 2. Georgetown CSET
3. Total Processing Performance (TPP) measured in Tera Operations per Second, Performance Density measured as TPP / Die Size

Commentary

- The H20 and L20 are the only current NVIDIA data center-class AI accelerators that do not exceed either the Total Processing Performance or Performance Density threshold.
- While the H20 accelerator is currently available for sale in China, the Trump administration has started preliminary conversations around the potential inclusion of the chip on the restricted list, suggesting that **there may be a further broadening of the scope of restricted chips**

Artificial Analysis

hello@artificialanalysis.ai

https://artificialanalysis.ai/

Appendix: Accelerator hardware specifications (NVIDIA Hopper, NVIDIA Blackwell, AMD)

	NVIDIA H100 (SXM)	NVIDIA H100 (NVL)	NVIDIA H100 (PCIe)	NVIDIA H800 (PCIe)	NVIDIA HGX H20	NVIDIA H200 (NVL)	NVIDIA H200 (SXM)	NVIDIA B200	NVIDIA GB200[1]	AMD MI300X	AMD MI325X
Initial Release Date	1Q23	1Q23	1Q23	2Q23	4Q23	2Q24	2Q24	1Q25	1Q25	4Q23	4Q24
Memory	80GB HBM3	94GB HBM3	80GB HBM2e	80GB HBM2e	96GB HBM3	141GB HBM3e	141GB HBM3e	192GB HBM3e	384GB HBM3e	192GB HBM3 256MB on-chip SRAM	256GB HBM3e 256MB on-chip SRAM
Memory Bandwidth	3.35 TB/s	3.9 TB/s	2 TB/s	2 TB/s	4 TB/s	4.8 TB/s	4.8 TB/s	8 TB/s	16 TB/s	5.3 TB/s	6 TB/s
Power/TDP	700W	350-400W	350W	350W	400W	600W	700W	1,000W	2,700W	750W	1000W
BF/FP16 TFLOPs (Dense)	989 TFLOPs	835 TFLOPs	756 TFLOPs	756 TFLOPs	148 TFLOPs	835 TFLOPs	989 TFLOPs	2,250 TFLOPs	5,000 TFLOPs	1,307 TFLOPs	1,307 TFLOPs
Chip-to-chip Interconnect	900GB/s NVLink™	600GB/s NVLink	600GB/s NVLink	400GB/s NVLink	900GB/s NVLink	900GB/s NVLink™	900GB/s NVLink™	1,800 TB/s NVLink™	3,600GB/s NVLink™	7X128GB/s Infinity Fabric™	7X128GB/s Infinity Fabric™
Module Type	SXM	PCIe	PCIe	PCIe	SXM	PCIe	SXM	SXM	SXM		
Process Node	TSMC 4N	TSMC 4N	TSMC 4N	TSMC 4N	TSMC 4N	TSMC 4N	TSMC 4N	TSMC 4NP	TSMC 4NP	TSMC 5N	TSMC 5N
Source URL	https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet	https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet	https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf	https://lenovopress.lenovo.com/lp1814.pdf	https://viperatech.com/shop/nvidia-hgx-h20/ https://wccftech.com/nvidia-h20-ai-gpu-china-mass-produced-q2-2024-full-compliance-us-policies/	https://www.nvidia.com/en-us/data-center/h200/	https://www.nvidia.com/en-us/data-center/h200/	https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid	https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid	https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300x-data-sheet.pdf	https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/instinct-mi325x-datasheet.pdf

1. Grace Blackwell superchip includes two Blackwell GPUs and a NVIDIA Grace ARM CPU