



AI 现状：中国

人工分析
Q1 2025

人工分析 是领先且独立的AI基准测试和洞察提供商。我们支持工程师和公司了解AI能力，并就其AI战略做出关键决策。

我们的数据、见解和出版物基于我们对人工智能技术及其应用场景的全面基准测试。这包括从语言模型API的小时级性能测试到我们众包平台上的数百万次投票。

我们的公共网站，artificialanalysis.ai，广泛被人工智能创新领域的领先公司引用。如欲讨论本报告、我们的出版物或服务，请通过contact@artificialanalysis.ai 联系我们。

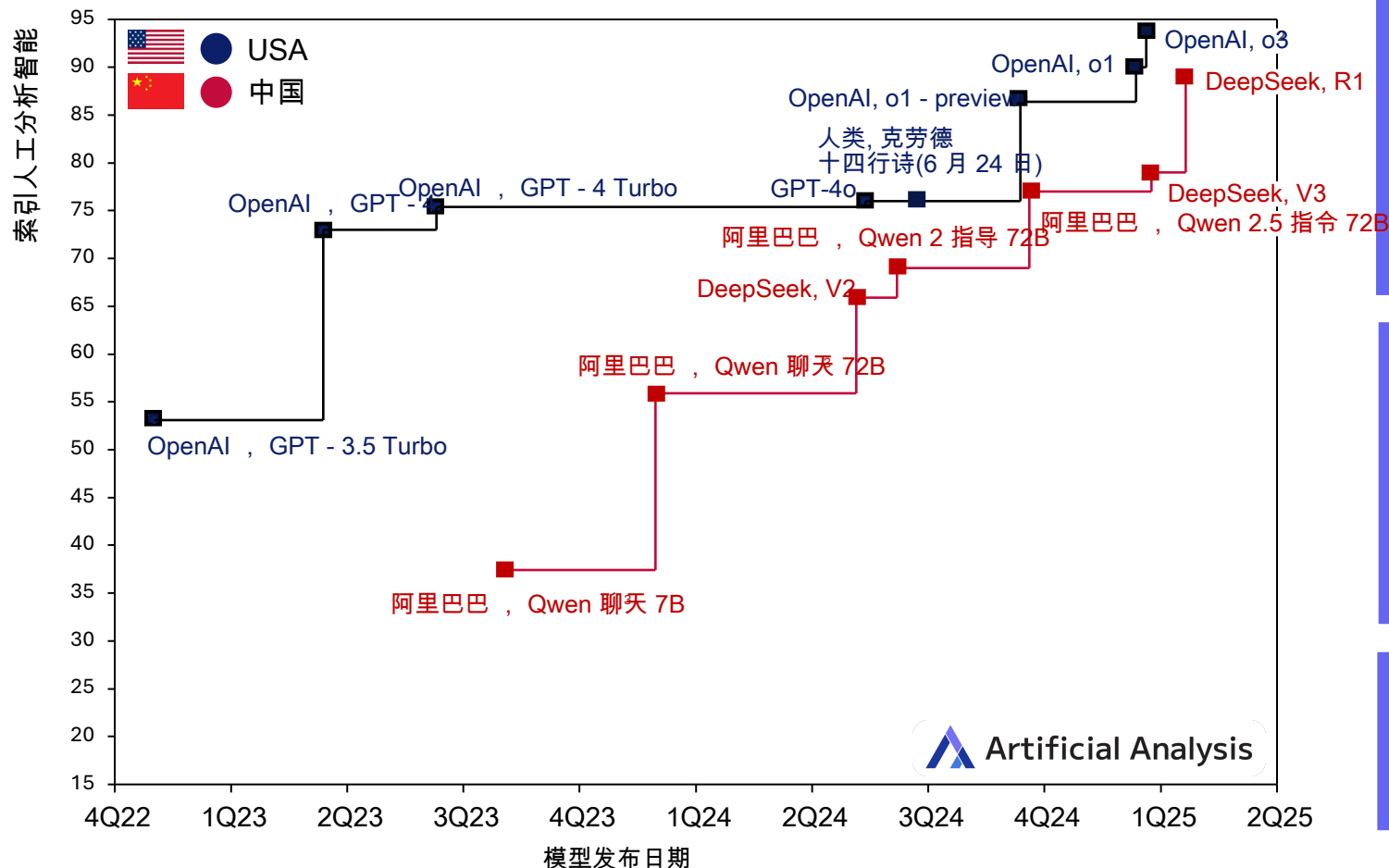




中国 AI 实验室逐渐赶上美国 AI 实验室；来自中国实验室的模型
随着 DeepSeek 的 R1 模型的发布，现在正在接近 o1 级智能

美国和中国：随着时间的推移，前沿语言模型情报¹

主要趋势



缩小差距：2024年最后几个月见证了多家顶尖中国AI实验室推出众多高性能模型。这导致了美国AI实验室在智能水平上的差距缩小。目前，一些中国模型已与美国顶级实验室的模型竞争。

推理模型很快变得司空见惯：推理模型（在回答前进行“思考”的模型）最早由OpenAI在2024年第三季度引入。几个月内，以DeepSeek为代表的中国竞争对手已基本复制了o1的智能水平。目前，多家中国AI实验室已经拥有前沿级别的推理模型。

在前沿实验室附近打开模型：由DeepSeek和阿里巴巴领导的开放权重模型已经接近o1级智能。

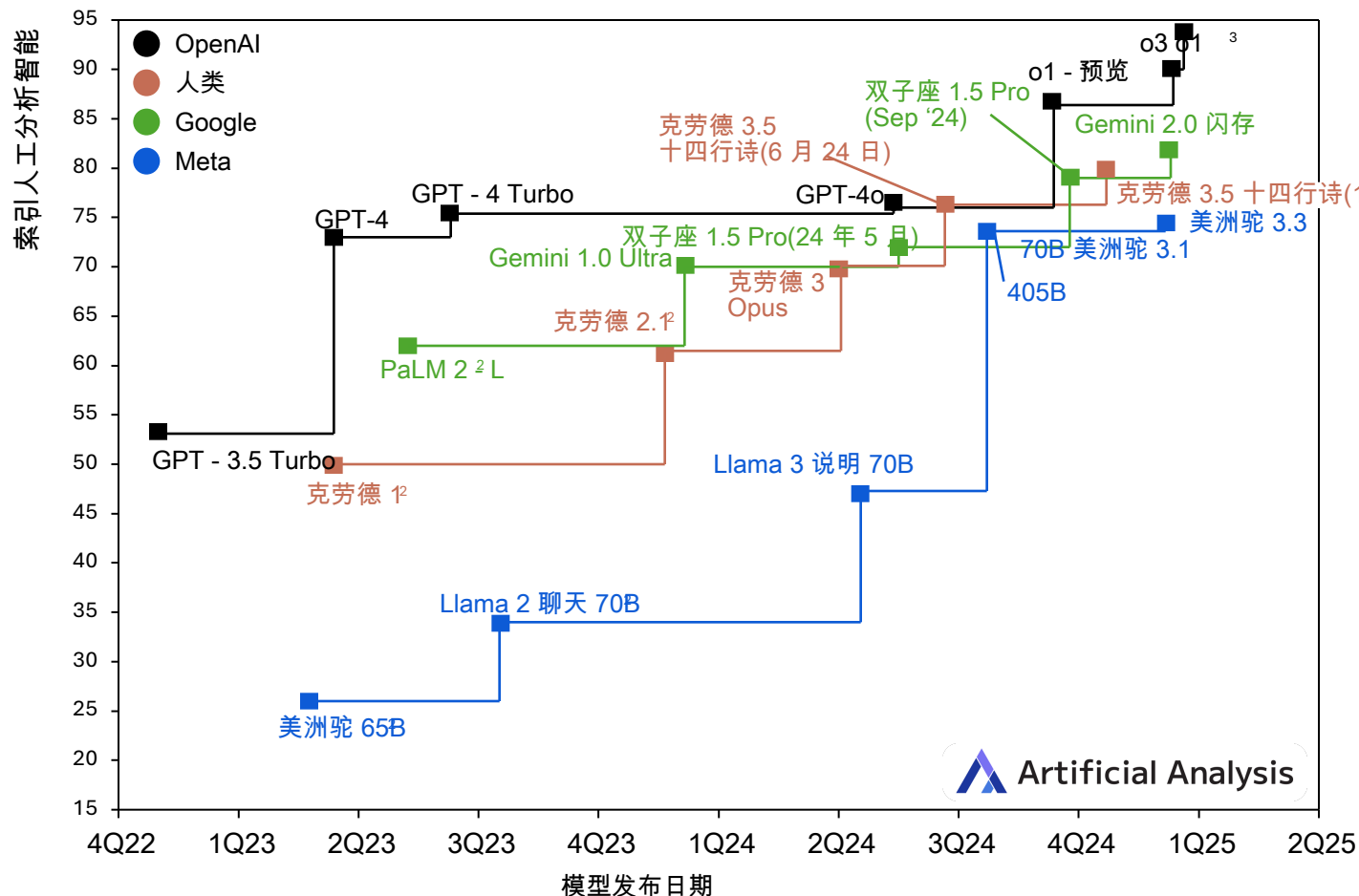
1. 人工分析智能指数：涵盖一系列语言模型智能和推理评估数据集的平均值。目前包括MMLU、GPQA Diamond、MATH-500及HumanEval。发布日期基于该模型的首次公开发布。2. o3 智能指数通过放大测量的 o1 的智能指数估算得出。3. 基于公司声明和可比结果估算。在可用的情况下，尚未通过人工分析进行独立基准测试



自 2023 年初推出 OpenAI 的 GPT - 4 以来，美国领先的 AI 实验室已经争先恐后赶上 OpenAI

随着时间的推移，美国领先的 AI 实验室前沿语言模型智能¹

主要趋势



竞争实验室赶上 OpenAI 的 GPT - 4 : OpenAI于2022年11月通过推出ChatGPT中的GPT-3.5开启了语言模型的竞争；美国领先的实验室已基本赶上OpenAI的前沿模型。

Big Tech 关闭了前沿实验室：Google和Meta的模型正迅速逼近前沿模型，其中Gemini 2.0 Flash超过了Claude 3.5 Sonnet和GPT 4o的能力。

GPT - 4 以外的智能火花：2024年最后几个月见证了除了GPT-4之外的重大智能飞跃，由OpenAI的o3引领。推理模型、数据质量以及新的强化学习技术等主题已与预训练计算扩展一起成为提升模型性能的主要杠杆。

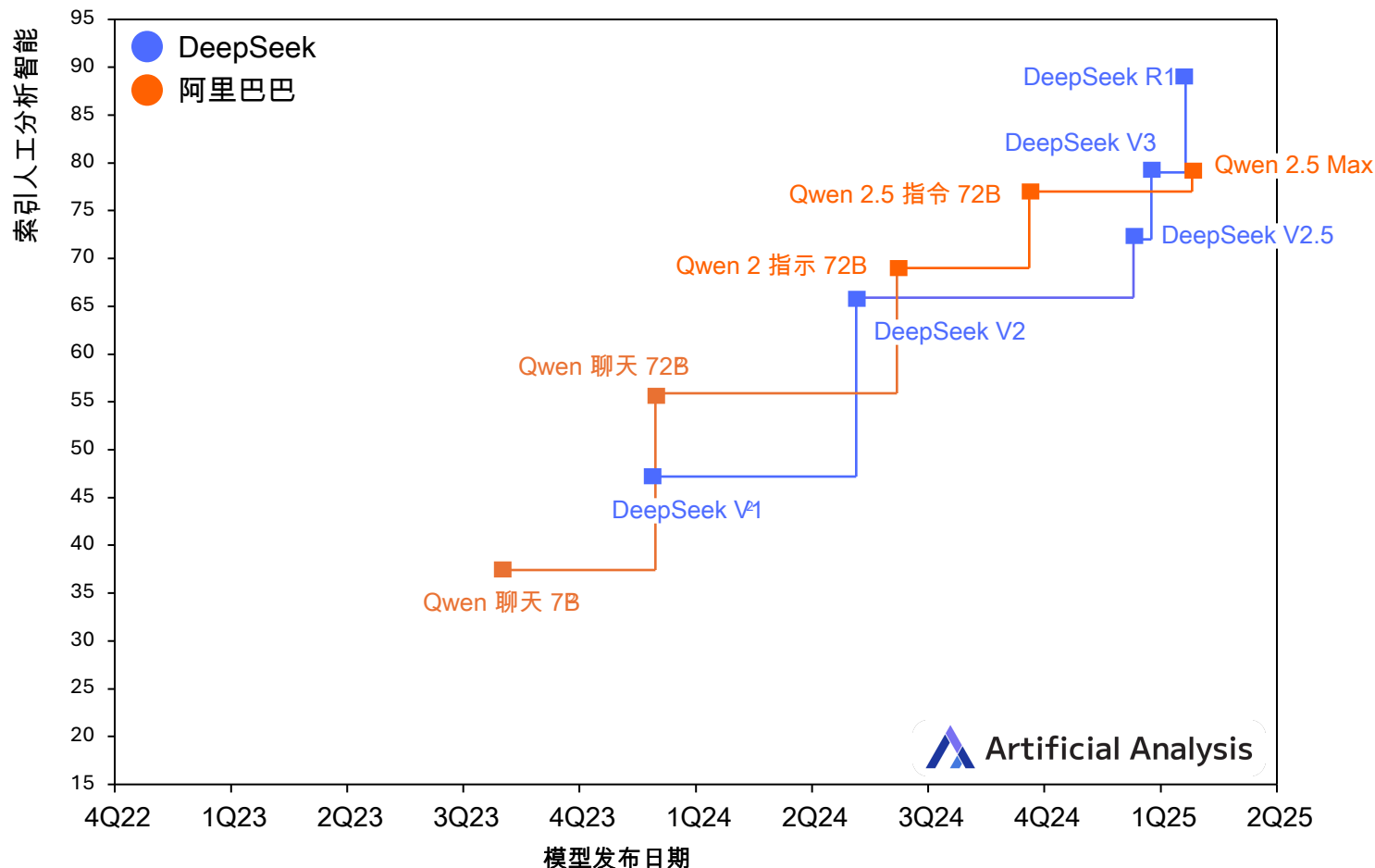
1. 人工智能分析指数：涵盖了多种语言模型智能和推理评估数据集的平均值。目前包括MMLU、GPQA Diamond、MATH-500及HumanEval。发布日期基于该模型的第一次公开发布。2. 估计依据公司声明和可比结果（如有），尚未由Artificial Analysis独立基准测试。3. o3 智力指数通过缩放测量的智力得分 o1 来估计。



领先的人工智能实验室DeepSeek和阿里云（Alibaba）持续发布新的模型，在2024年底，DeepSeek超越了阿里云。

随着时间的推移，中国领先的AI实验室语言模型智能¹

主要趋势



迅速提高智力： 尽管中国的AI实验室较晚加入AI竞赛，但在2024年，它们在智能方面与前沿的美国模型之间的差距显著缩小。当OpenAI推出o1时，中国实验室在几个月内开发出了一个性能相当的模型（DeepSeek的R1）。

领先的开放式重量模型： 中国的人工智能实验室，包括阿里云、深视和腾讯，发布了开放权重前沿模型，这些模型在全球领先的模型中具有竞争力。

2025 年的潜在领导者： 2025年初，包括阿里云、深视、明略、腾讯、智谱和通义等中国人工智能实验室纷纷发布了前沿推理模型。发布的速度和频率表明，中国人工智能实验室在2025年已不再是落后者。

1. 人工分析智能指数：涵盖一系列语言模型智能和推理评估数据集的平均值。目前包括MMLU、GPQA Diamond、MATH-500及HumanEval。发布日期基于模型的首次公开发布。

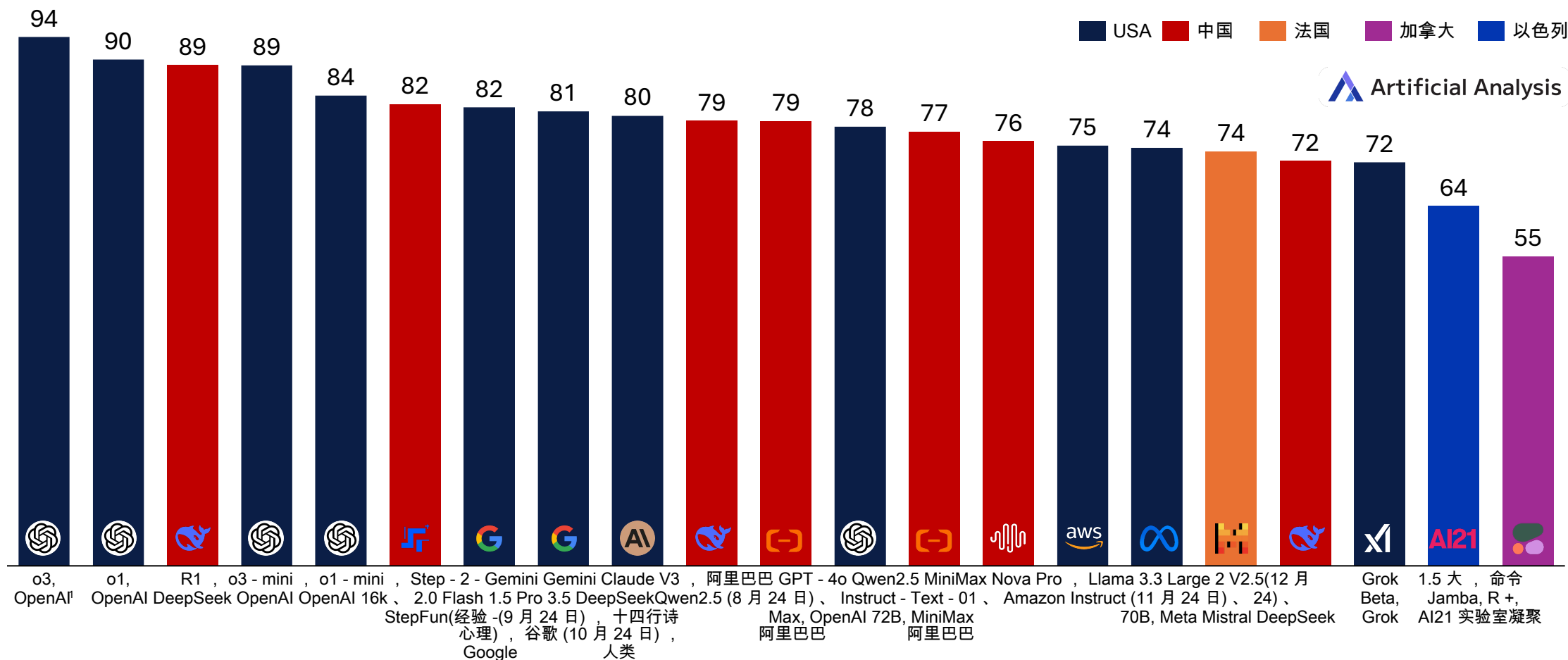
2. 基于公司声明及可比结果进行估算，尚未由人工分析独立基准测试。



在美国继续在情报前沿领域保持总体领先地位的同时，中国与之的距离已不再遥远。很少有其他国家能够展示出同等水平的训练能力。

语言模式前沿：起源国

人工分析智能指数，精选领先模型(2025 年初)，非详尽



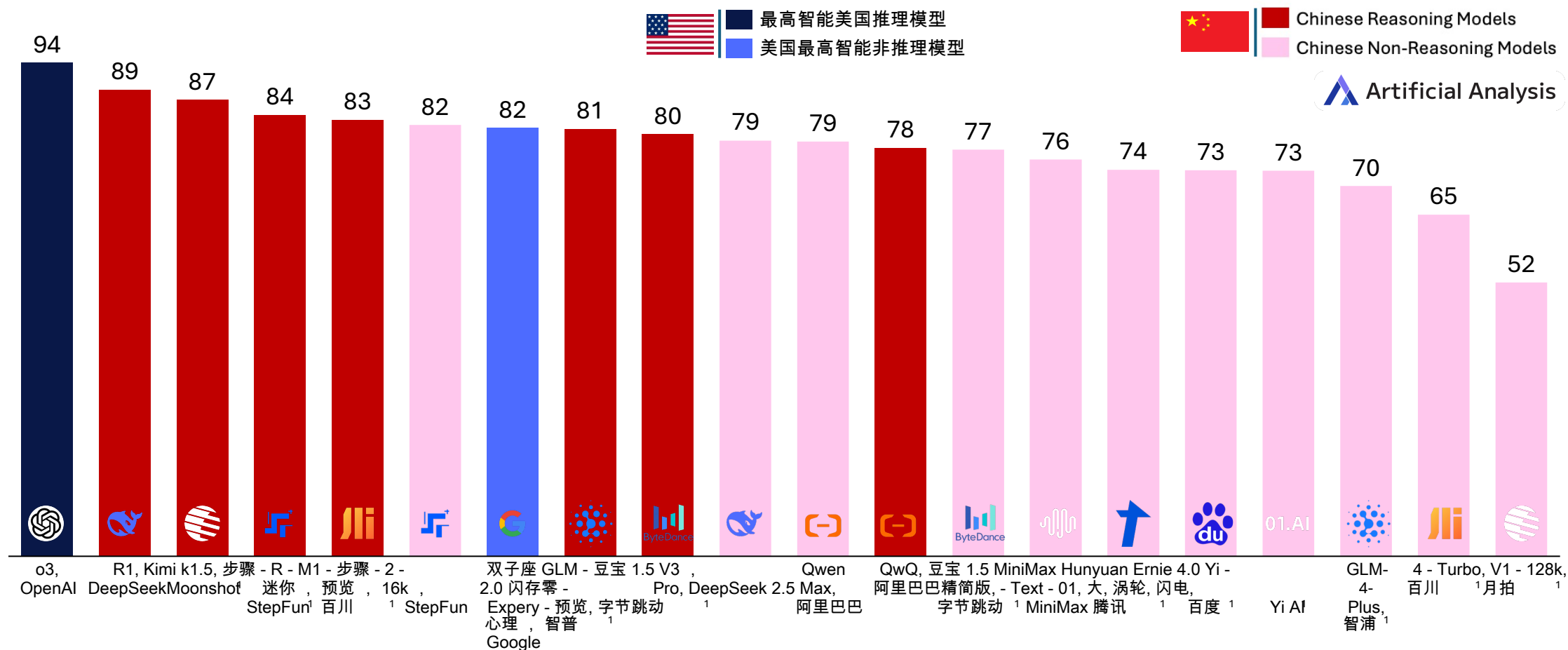
1. 根据公司声明和可用的类似结果估算得出，尚未经过人工分析独立基准测试。
2. 由于访问限制或评估数据有限，排除了部分中国AI实验室的领先模型。



截至2025年初，几家中国人工智能实验室已展示或声称达到了前沿水平的智能，其中七家发布了具备推理能力的模型。

语言模型前沿：中国 AI 实验室的模型

人工分析智能指数，领先模型(2025 年初)，非详尽



1. 本估计基于公司的声明及可比结果 (尚未由Artificial Analysis进行独立基准比较)。

非穷尽

</> 打开权重 LLM

其他有 AI 野心的公司

昆仑科技 SHE : 300418(
市值 : 60 美元) ¹

360 安全 (奇虎 360)

SHA : 601360(市值 : \$11

中国最大的互联网和移动安全产品提供商推出了360 AI品牌的Zhinao系列模型。

科大讯飞

SHE : 002230(市值 : \$16

中国领先的语音 AI 公司，拥有超过 14,000 名员工。推出 Spark 系列模型

美团

香港：3690(市值：\$115

中国的领先购物平台，日活跃用户超过600万。联合创始人王慧文回归领导AI项目。投资多家前沿AI实验室。

小米

香港：1810(市值上限：1

中国的领先消费电子品牌推出了MiLM系列小型型号。最近挖来了DeepSeek研究员罗浮利，负责AI实验室。投资了多个前沿AI实验室。



在中国大型科技公司和中国的支持下，中国的 AI 初创公司政府，已经开发了一些世界领先的开放重量模型

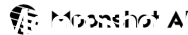
非穷尽

中国 AI 老虎和初创企业的前沿模型

</> 打开权重 LLM



国家支持实体

 MINIMAX  Moonshot AI  零一万物 YI AI  deepseek  智谱 AI  百川智能 SAICHUAN AI  阶跃星辰 Stepfun								
Description		中国 AI 虎和 Talkie 的出版商 AI 应用程序 (在美国下载 1H24)	中国 AI 老虎 2M 中文字符上下文窗口模型; 中国最好的 - 资助 AI 启动基于可用信息	中国 AI 初创公司 专注于较小的语言模型 由 李开复(作者，前负责人 谷歌中国)	中国 AI 实验室 起源于 以 AI 为中心 定量 贸易公司	中国 AI 老虎 近 700k 企业和开发者用户	中国 AI 虎与 专注于医疗 AI 建立的模型 王小川 (前 CEO， 搜狗)	第一个中国 AI 启动开发一个 万亿参数模型; 由 姜大新 (前 - 首席科学家, 微软研究院 Asia)
最佳 LLM	非推理	MiniMax - Text - 01 情报：76	V1-128k 情报：52	Yi - Lightning 情报：73	V3 情报：79	GLM - 4 - Plus 情报：70	百川 4 - Turbo 情报：65	Step - 2 - 16k 情报：82
	推理	-	Kimi k1.5 情报：87	-	R1 情报：89	GLM - Zero - 预览 情报：81	百川 M1 - 预览 情报：83	Step - R - mini 情报：84
其他型号	文本到语音	✓	-	-	-	-	-	✓
	演讲到演讲	-	-	-	-	✓	✓	-
	图像生成	-	-	-	✓	✓	-	✓
	视频生成	✓	-	-	-	✓	-	-
	3D 生成	-	-	-	-	-	-	-
主要消费者 Apps		HailuoAI 聊天, HailuoAI 视频	Kimi	YiChat	DeepSeekChat	ChatGLM	白小英	Yuewen, PopDuck
募集资金(\$)		0.85B ¹	1.67B ²	0.2B ³	未知	1.12B ⁴	1.04B ⁵	未知
著名投资者 (非详尽)		 Tencent 腾讯  Alibaba.com  HONGSHAN 红杉中国  高榕 GAO RONG VENTURES	 Tencent 腾讯  Alibaba.com  小红书  ZhenFund  Microsoft  HONGSHAN 红杉中国  启宏基金	 Alibaba.com  mi  SINO VATION VENTURES  顺为 SHUNWEI	 幻方量化 HIGH-FLYER QUANT	 Alibaba.com  Tencent 腾讯  QIMING 启明创投  LightSpeed 光速创投  美团 MEITUAN  清华控股 TSINGHUA HOLDINGS  美团 MEITUAN	 Alibaba.com  Tencent 腾讯  清华控股 TSINGHUA HOLDINGS  SUSV  顺为 SHUNWEI  mi	 Tencent 腾讯  QIMING 启明创投  上海国有资本投资有限公司 Shanghai State-owned Capital Investment Co., Ltd.

1. 人工分析智能指数：涵盖一系列语言模型智能和推理评估数据集的平均值。目前包括MMLU、GPQA钻石版、MATH-500及HumanEval。2. 根据公司声明及可比结果估算，尚未经过Artificial Analysis独立基准测试。(Pitchbook，2024年3月) 3. (Pitchbook，2024年8月) 4. (Pitchbook，2024年12月) 5. (Pitchbook，2024年7月) 6. (Pitchbook，2024年8月)

不断升级的监管限制已禁止出口高 向中国提供 AI 加速器(1 / 2)

监管限制		未释放	无许可证 需要	NAC 许可证 需要	推定 否认
NVIDIA GPU 建筑	模型	预控制	2022 年 10 月控制 ²	2023 年 10 月 控件 ^{3,4}	AI 扩散规则 ⁵
	宣布 有效 ¹		7-Oct-22 21-Oct-22	17-Oct-23 17-Nov-23	13-Jan-25 15-May-25
Blackwell	B200				
	B100				
料斗	H100				
	H200				
	H800				
	H20				
Lovelace	L40S				
	L4				
	L40				
	L20				
安培	L2				
	A100				
	A800				
	A40				
消费者 GPU	A30				
	RTX 6000 Ada				
	RTX 4090				
	RTX 4090D				
	RTX 3090				

Commentary

- NVIDIA 对2022年10月和2023年10月发布的两项控制措施迅速反应，发布了符合规定要求的Hopper GPU变体。特别是在H100和A100被禁止出口到中国后，NVIDIA推出了具有有限互联功能的H800和A800（完整Hopper系列规格详见附件）。
- 2023 年 10 月的管制继续禁止向中国出口 H800 和 A800，导致 NVIDIA 开发 H20 继续向中国客户销售Hopper一代GPU。H20在计算能力上相对有限（148 TFLOPs），而H100则达到了989 TFLOPs。

1. 生效日期是指最近的合规日期 2. BIS 3. 乔治敦 CSET
4. 联邦登记册 5. BIS

不断升级的监管限制已禁止向中国出口高端 AI 加速器(2 / 2)

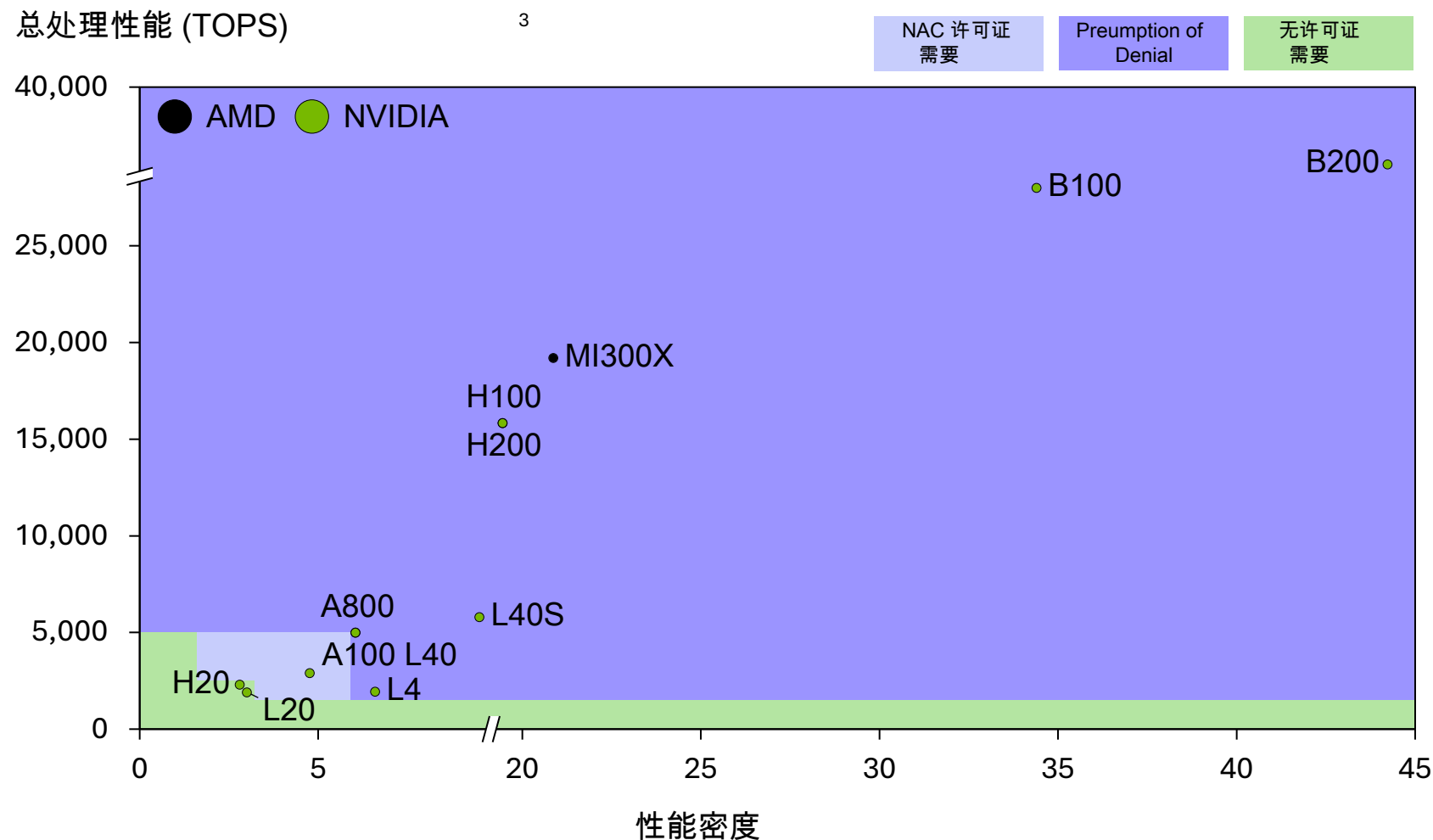
监管限制

Rule	Summary	日期 ¹	影响 ²
2022 年 10 月 控件 ³	初始限制 在前沿 GPU 上。两种性能和互连阈值必须为 GPU 被破坏受限制。	宣布：22 年 10 月 7 日	限制分类
		有效：21 - Oct - 22	总加工性能 (TPP)
			互连带宽
2023 年 10 月 控件 ^{4,5}	修订框架以防止变通办法。限制出口基于 TPP 或中国的 GPU 性能密度 (PD)	宣布：23 年 10 月 17 日	Groupings
		有效：17 - Nov - 23	第 1 组：拒绝的推定
			第 2 组：限制性 NAC 许可审查
BIS 最终 Rule ⁶	通过以下方式打击间接进口中国附属芯片制造实体	宣布：24 年 12 月 2 日	第 3 组：无限制
		生效日期：31 - Dec - 24	判据 (数据中心 GPU)
		更新时间：24 年 1 月 16 日	TPP ≥ 4,800 或 TPP ≥ 1600W/DPD ≥ 5.92. 2,400 ≤ TPP ≤ 4,800W/DPD ≥ 1.6 或 TPP ≥ 1,600W/DPD ≥ 3.2. TPP < 1,600 或 PD < 3.2
AI 扩散 Rule ⁷	广泛的三层许可 框架隔离访问 GPU 由国家	宣布：25 年 1 月 13 日	没有影响限制芯片
		有效：15 - 5 月 25 日	来自先进芯片行业的 140 家实体 (大多数是中国人) 现在面临着拒绝并在 12 月 24 日添加到实体列表
AI 尽职调查 Rule ⁸	AI 扩散的配套 KYC 规则 Rule	宣布：25 年 1 月 16 日	3 级国家(包括中国) 实际上面临对先进 AI 芯片的禁令
		生效日期：24 年 1 月 31 日	所有向这些三级国家出口的受控芯片现在都需要出口许可证，在审查期间被推定为否认 2 级国家现在面临大量 AI 芯片订单的限制 要求公司对其客户进行类似 KYC 的合规性检查，并遵守 AI 扩散规则

1. TPP 以 Tera Operations per Second 计量，PD 以 TPP / 模头尺寸计量。2. 生效日期是指最新合规日期
3. BIS4. 乔治敦 CSET 5. 联邦登记册 6. BIS7. 联邦登记册 8. BIS9. 联邦登记册

美国出口管制限制了英伟达领先加速器的出口 基于性能和密度阈值；H20 和 L20 低于这些阈值，可以自由出口

美国禁止向中国出口加速器^{1,2}



Commentary

- H20和L20是当前唯一不超出总处理性能或性能密度阈值的NVIDIA数据中心级AI加速器。

- 虽然H20加速器目前在中国市场上可以销售，但特朗普政府已开始就将该芯片列入受限名单的可能性进行初步讨论，这表明了潜在的变化。限制芯片的范围可能会进一步扩大。

1. [半分析](#) 2. [乔治敦 CSET](#)

3. 总处理性能 (TPP，单位为tera操作每秒)，性能密度测量为TPP/晶粒尺寸



人工分析

你好 @ artificialanalysis. ai

[https: // 人工分析. ai /](https://人工分析.ai/)

法律声明:

版权所有 © 2025 人工分析公司。保留所有权利。

本文件包括但不限于此处包含的数据、分析和见解，仅由Artificial Analysis提供以供信息参考。这些信息基于通过各种来源收集的数据，包括但不限于我们在网站上进行的第一方基准测试和调查。虽然Artificial Analysis力求确保信息的准确性和可靠性，但信息按“原样”提供，可能不完整或不及时更新。内容不应被视为专业建议，接收者被鼓励在根据本信息做出任何决定之前进行自己的研究和分析。您访问或使用本文件即表示同意受Artificial Analysis的条款和服务条款约束，可在我们的网站上查看。

附录：加速器硬件规格(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	AMD MI300X	AMD MI325X
Initial 发布日期	1Q23	1Q23	1Q23	2Q23	4Q23	2Q24	2Q24	1Q25	1Q25	4Q23	4Q24
记忆	80GB HBM3	94GB HBM3	80GB HBM2e	80GB HBM2e	96GB HBM3	141GB HBM3e	141GB HBM3e	192GB HBM3e	384GB HBM3e	192GB HBM3 256MB 片上 SRAM	256GB HBM3e 256MB 片上 SRAM
内存带宽	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
电源 / TDP	700W	350-400W	350W	350W	400W	600W	700W	1,000W	2,700W	750W	1000W
BF / FP16 TFLOP (密集)	989 TFLOP	835 TFLOP	756 TFLOP	756 TFLOP	148 TFLOP	835 TFLOP	989 TFLOP	2, 250 TFLOP	5, 000 TFLOP	1, 307 TFLOP	1, 307 TFLOP
芯片到芯片互连	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™
模块类型	SXM	PCIe	PCIe	PCIe	SXM	PCIe	SXM	SXM	SXM		
Process Node	台积电 4N	台积电 4N	台积电 4N	台积电 4N	台积电 4N	台积电 4N	台积电 4N	台积电 4NP	台积电 4NP	台积电 5N	台积电 5N
源 URL	https://resources.nvidia.com/en-us-张量-core/nvidia-张量核-gpu-数据表	https://resources.nvidia.com/en-us-张量-core/nvidia-张量核-gpu-数据表	https://www.nvidia.com/containers/大坝/en-zz/解决方案/gtcs22/data-center/h100/PB-11133-001_v01.pdf	https://lenovopress.lenovo.com/lp1814.pdf	https://viperatech.com/shop/nvidia-hgx-h20/	https://www.nvidia.com/en-us-数据中心/h200/	https://www.nvidia.com/en-us-数据中心/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/nt-dam/amd/en/documents/本能-技术-文档/数据-床单/AMD-本能-mi300x-data-sheet.pdf	https://www.amd.com/content/dam/amd/en/documents/nt-dam/amd/en/documents/本能-技术-文档/产品-三角旗/本能-mi325x-数据表.pdf

1. Grace Blackwell 超级芯片包括两个 Blackwell GPU 和一个 NVIDIA Grace ARM CPU