

# 为何生成式AI下一阶段将要求更高算力？

*世界正从单纯训练生成式AI（Gen AI）模型转向大规模应用。许多人认为这意味着消费者边缘计算增加，而数据中心计算将减少。但在2026年，这两者都不太可能发生。*

人们普遍预计，到2026年，生成式AI计算将从目前主要依赖海量数据训练模型，转向利用这些模型来思考并回答企业和消费者的问题、提示及任务，这一过程被称为“推理”（inference）。许多人推测，这种计算工作负载（即“计算”）的转变意味着AI生态系统将需要专门针对推理任务优化的芯片，而这些（或许成本更低的）芯片可能会被部署在大型数据中心之外的边缘设备上。这甚至意味着我们未来可能需要更少、更小，或至少是不同类型的计算中心，并减少相关支出。

德勤的看法略有不同：推理工作负载确实将成为2026年的热点，占有计算量的约三分之二（2023年为三分之一，2025年为一半）。<sup>1</sup> 并且推理优化芯片的市场规模将在2026年增长至超过500亿美元。但德勤同时预测，大部分计算仍将依赖于价值2,000亿美元或更多的尖端、昂贵且高功耗的AI芯片，这些芯片主要部署在价值4,000亿美元或更多的大型数据中心，或价值500亿美元或更多的企业本地化解决方案中，而这些解决方案使用的芯片和机架与数据中心相同，而非边缘设备上的芯片。这意味着我们可能仍需要当前正在规划的所有数据中心和企业本地AI工厂，以及这些数据中心所需的全部电力资源。

## 不断增长的AI计算需求

尽管对新模型训练算力的需求增长可能已经放缓（很可能仍在增长，但增速低于2023年和2024年），<sup>2</sup> AI模型正通过训练后改进的先进技术持续演进。这些方法，加上海量的推理查询，很可能意味着计算需求将持续增长而非减少。换句话说，尽管得益于摩尔定律，用于计算的芯片每年都在变得更加高效，但计算需求的增长速度更快，预计到2030年每年将增长4至5倍。<sup>3</sup>

## 用于初始训练的计算需求增长正在放缓

2020年的一篇论文表明，模型规模越大、训练数据越多、采用越先进的AI加速芯片，产出的结果就越好：这构成了生成式AI的首个扩展定律。<sup>4</sup> 到2022年和2023年，训练模型的参数量已从10亿级增长至千亿级，并进一步达到万亿规模。<sup>5</sup>

2024年，两个问题开始浮现：可用的训练数据并非无限，而不断增大的模型正呈现出收益递减的趋势：十倍的训练数据可能仅能产生一个比前一版本略好一点的“最先进的”AI模型，甚至可能毫无提升。<sup>6</sup>与此同时，更小且更高效的AI模型似乎有可能利用更少的数据、更短的时间、更少的资金和更少的芯片，开发出真正达到顶尖水平的AI模型。<sup>7</sup>

如果训练规模的增长放缓，AI计算将越来越侧重于推理。让一个大语言模型对文档进行摘要（推理的一个例子）所消耗的计算资源，仅相当于训练该模型所需资源的一小部分。然而，逻辑在于，随着数十亿消费者和企业员工更频繁地发起此类请求，所有这些推理任务累积起来，将使整体计算工作负载从训练转向推理。其中部分请求可以在消费者和企业设备（如智能手机和个人电脑）上处理，而正如德勤在2024年预测的那样，2025年已有数亿台搭载了支持本地AI加速芯片的个人电脑和智能手机售出。<sup>8</sup>此外，由于推理的计算密集程度低于训练，数据中心内部或许可以采用专门针对推理优化的芯片。这类芯片成本更低，每次推理所消耗的能量也少于用于大规模训练的高性能AI芯片，并且可能不需要配备大量昂贵的共封装内存。<sup>9</sup>

上述所有情况均发生在2025年，并可能在2026年持续。德勤2025年的调查显示，无论是在美国还是全球范围内，使用生成式AI的消费者数量都在增加，且每日使用的人数也在上升。<sup>10</sup>诸如个人电脑和智能手机等边缘设备越来越多地搭载了AI加速器。许多针对推理任务优化的芯片（专用集成电路，ASIC）已被设计、制造，并正在数据中心和部分边缘设备中部署。这些芯片包括但不限于来自Meta、Google、Amazon、Intel、AMD、Qualcomm、Groq、SambaNova、Cerebras和Graphcore的产品，其中一些基于Broadcom的封装解决方案，由设计方提供处理核心。<sup>11</sup>尽管这些不同芯片的销售数据并未公开，但德勤认为，这些芯片在2025年的总收入合计将超过200亿美元，并在2026年达到500亿美元或以上。<sup>12</sup>

既然如此，为何我们仍需在数据中心部署耗电量巨大、单价高达3万美元甚至更高的芯片（到2028年，此类芯片总成本或将达到约4,000亿美元），<sup>13</sup>而仅2026年数据中心的建设成本就预计高达4,000亿美元，且到同年，其年度潜在成本可能攀升至1万亿美元？<sup>14</sup>

## AI模型的训练比过去更加复杂

第一个扩展定律的目的是生成“更好”的AI模型，而且在最初几年效果非常显著。这种最初的扩展形式过去被称为训练，现在则称为“预训练”，用于生成基础模型。

事实证明，还有两种方法可以构建更优秀的模型：一种是后训练扩展，包括各种技术，如微调、剪枝、量化、蒸馏、基于人类反馈以及越来越多地基于AI反馈的强化学习，以及合成增强数据。<sup>15</sup>另一种是测试时扩展，也称为长时思考，即模型在被提问后，通过多种技术在推理过程中逐步进行思考，例如思维链提示、基于多数投票的采样、搜索，甚至一些后训练技术。<sup>16</sup>变得更加准确，拥有更多选择、更优质的来源以及更少的错误信息。<sup>17</sup>

## 新的高能耗AI技术可能会超过能效提升的速度

首先，训练后扩展和测试时扩展似乎已成为新的常态：目前大多数AI公司都采用这些方法以多种方式提升AI模型的性能。<sup>18</sup>

其次，这两者都非常消耗AI算力。据估计，训练后阶段总体使用的计算量是训练原始基础模型的30倍，而长链推理所消耗的算力更是一次简单推理（例如让AI总结一封邮件）的100倍以上。<sup>19</sup>

第三，由于这两种扩展技术被广泛使用且计算资源消耗大，这对AI数据中心及其地理位置和电力需求、用于AI数据中心的芯片（以及需要执行AI的其他场所）、往年的AI芯片、边缘设备等均产生了影响。

## 关于德勤预测2025年及以后AI数据中心所需芯片的简要回顾

数据中心已经存在了几十年。事实上，全球有数千万平方英尺的数据中心，多年来每年都有价值数百亿美元的半导体组件被售出，用于数据中心。<sup>20</sup> 但新的AI数据中心以及实现这些数据中心的新半导体，往往与昨日的数据中心和半导体截然不同，简直是天壤之别。

下一代AI数据中心的建设成本可能每年高达数千亿美元，并消耗数百吉瓦的电力。在大多数此类数据中心中，冷却方式可能与以往几代数据中心不同，电源和电压可能不同，内部通信技术也可能不同，甚至地板可能需要更厚，以支撑密度更高、重量更大的服务器机架。或许最重要的是，新型AI服务器机架不再采用以CPU为中心、内存就近配置的服务器架构，而是主要由一种名为图形处理单元（GPU）的专用芯片构成。<sup>21</sup> 这些组件通常配备专用的高带宽内存（HBM），并与GPU紧密集成，同时还配有专门的CPU来协调庞大的AI计算工作负载。许多组件都是针对新一代AI数据中心的独特需求和规模而设计的。<sup>22</sup>

直到2006年，高端GPU仍被认为仅适用于游戏电脑和游戏主机，而非数据中心。<sup>23</sup> 大多数数据中心的任务由CPU即可很好地完成，因为CPU主要是串行处理器，任务按顺序执行。一些高性能计算机，即“超级计算机”，则配备了称为“大规模并行处理器”的特殊芯片，能够同时执行数百个任务。然而，这些芯片的成本通常是游戏用GPU或数据中心CPU的数十倍甚至数百倍。

2009年，科学家们注意到游戏用的GPU同样是并行处理器，于是尝试在高端GPU上运行机器学习模型，这些GPU与游戏电脑中使用的完全相同。<sup>24</sup> 它们运行良好，几年之内，经过特别优化的GPU（与游戏版本略有不同）开始被用于一些数据中心和部分本地设备上，以执行机器学习AI任务。<sup>25</sup> 但直到2018年，该市场的规模每年仍仅为数十亿美元级别。<sup>26</sup>

2022年，生成式AI的大语言模型（LLMs）的发展需要更加专业化的GPU，而且通常要求将这些GPU与一种相对较新且专用的内存（HBM）集成在同一个先进的封装中。<sup>27</sup> 这些GPU+HBM组件还需要一个设备来协调和管理数据流。经过优化的专用CPU（虽然核心架构相似，但不同于计算机、智能手机或数据中心中的CPU）也是生成式AI数据中心的重要组成部分，此外还有多个其他可能同样关键的组件。到2025年，全球几乎所有的Top 500超级计算机都采用了类似的GPU、专用内存和CPU组合。<sup>28</sup> 从某种意义上说，正在建设的超大规模AI数据中心可以被描述为一种专用超级计算机的版本。



# 小结：更多的算力需求可能对AI生态系统产生影响

企业和高管应当为未来计算需求持续增长做好准备——特别是在大型数据中心和企业AI工厂领域，这种增长部分源于训练后阶段及测试阶段的扩展需求。尽管面向推理优化的芯片和边缘处理将实现增长，但对超大规模数据中心和企业级AI专用设备的投资仍不可或缺。值得注意的是，“针对推理优化”并不等同于低功耗：近期某款专为推理预填充计算优化的产品通过采用GDDR7显存替代HBM显存，但其单机柜功率需求高达370千瓦，达到同厂商训练版本产品功率密度的近三倍！<sup>29</sup>

**AI数据中心：**预计到2026年，全球AI数据中心的资本支出将达到4,000亿至4,500亿美元。<sup>30</sup> 其中超过一半的支出用于设备内部的芯片（2,500亿至3,000亿美元），<sup>31</sup> 其余部分则包括土地、建设、电力和许可等所有其他费用。此外，预计到2028年，AI数据中心的资本支出将上升至1万亿美元。<sup>32</sup> 当年AI芯片市场规模超过4,000亿美元。<sup>33</sup> 尽管预训练阶段的增长正在放缓，且计算资源正转向推理阶段，但训练后扩展、测试时扩展以及使用量增加所带来的计算需求表明，全球可能仍需要大量数据中心，从2025年约3,000亿至4,000亿美元增长到2028年约1万亿美元的规模，这一趋势是基本合理的。

**AI数据中心的位置：**预训练一个拥有100万亿参数的大语言模型可能需要数周时间，且对微小的中断极为敏感：关键组件的故障或处理器之间过高的延迟交接，都可能导致所有工作成果的丢失，而不得不重新开始。大多数基础模型的预训练通常集中在同一栋建筑或园区内进行，所有服务器和机架都位于同一地点。然而，越来越多的AI计算任务现在可以在美国不同地区的数据中心，甚至在全球范围内的多个数据中心之间协同完成。<sup>34</sup> 此外，数据中心的规模可能从吉瓦级到较小规模的推理数据中心不等，这些数据中心可用于部署完全训练好的模型，且通常更靠近城市区域，以帮助降低延迟。这有助于形成不断增长的需求。**主权AI解决方案**（每个国家或地区都拥有位于国内甚至本地运营的AI计算能力），以及作为混合云一部分的企业本地边缘解决方案。<sup>35</sup>

**AI数据中心的电力需求：**总体而言，进行三种扩展方式的更多AI数据中心仍将需要大量电力。但与需要一次性完成训练过程的预训练相比，后训练和推理时的扩展在一定程度上相对“可中断”。这使得AI公司能够参与需求响应计划，通过将任务转移到其他数据中心位置或降低处理器时钟频率，在用电高峰期减少电力需求。<sup>36</sup> 据估计，提高此类负荷的灵活性，可使大型新建数据中心有助于维持电网的可靠性和经济性。<sup>37</sup>

AI的训练和推理可以分布式进行，这意味着数据中心无需集中在一个州或一个国家，而是可以更均匀地分布在全球各地，从而分散电力需求。

**AI数据中心中的芯片：**有些人可能将AI芯片市场视为零和博弈。这种观点通常表现为“虽然我在预训练基础模型时不得不花费数万美元购置搭载HBM的先进GPU，但随着计算重心向推理转移，或许可以采用更廉价、专为推理优化且减少HBM配置的芯片？”

芯片市场似乎将不再是“非此即彼”，而是走向“两者兼有”。仅用于推理或针对推理优化的芯片很可能会迎来显著增长；但与此同时，通常最适合基础模型预训练、后训练以及测试时扩展（计算需求包含训练与推理的混合）的芯片，仍将是那些配备高带宽内存（HBM）、每块售价数万美元、性能强大且耗电量高的GPU。对于采购这些芯片的用户而言，到2026年它们的价格可能更加昂贵，预计先进制程晶圆的成本将上涨50%。<sup>38</sup>

**面向智能手机和PC等消费类或企业级设备的边缘AI：**如前所述，数以亿计的智能手机和个人电脑正在出货并被购买，这些设备均配备了NPU：<sup>39</sup>专用于处理AI推理任务的专用芯片或CPU芯片的部分模块（NPU部分价值几美元至几十美元），在功耗合理的情况下进行了优化。

然而，目前NPU的性能仅足以支持前文讨论的单次推理任务（如“总结这封邮件”等）。因此，德勤预测，2026年几乎所有的AI计算仍将主要在大型AI数据中心，或由企业拥有的相对昂贵的高端AI服务器上完成，而非在个人电脑和智能手机上进行。至少在当前所处的高速增长、抢占市场的阶段，成本优化的混合架构似乎并非供应商或企业的优先考虑事项。此外，诸如测试时扩展（test-time scaling）等技术对于绝大多数消费者应用场景，甚至大多数企业级设备端应用场景而言，都可能显得过于冗余。未来某一天，计算机和智能手机或许会发挥更重要的作用，但在2026年之前这不太可能发生。

最近，一家AI公司推出了一款具备推理能力的生成式AI模型，可在个人电脑上本地运行。目前尚不清楚该模型的实际效果如何，对电池续航会产生何种影响，或有多少个人电脑用户更倾向于在本地而非通过云端使用AI功能。<sup>40</sup>

**边缘AI与企业本地化解决方案的应用：**那些通常被部署于全球大型AI数据中心的高性能、高功耗GPU，搭配HBM和协同工作的CPU模组，如今也可供希望采用本地化、混合式、更具韧性的生成式AI计算架构的企业使用，尤其适用于训练后的应用场景。出于对成本、知识产权归属、数据主权、系统韧性和定制化需求的考虑，企业可以花费约30至50万美元购置一台配备约八块GPU（以及HBM和CPU）的设备，以实现特定规模的AI训练与推理任务。<sup>41</sup>或者，他们可以花费300万至500万美元购买一个配备多达72个最先进GPU（以及HBM和CPU）的机架，其性能更强大。<sup>42</sup>或者，他们甚至可以花费数千万美元购买多个机架，以实现更强大的功能。<sup>43</sup>德勤预测，到2026年，这一本地混合企业市场规模将超过500亿美元。

**用于机器人、无人机和自动驾驶汽车的边缘AI：**到2026年规模仍然相对较小，但存在多种需要实时、本地设备推理的应用场景。这些场景涵盖从无人机、机器人到自动驾驶汽车等多个领域。目前所使用的芯片种类繁多：大多数无人机配备的是相对基础且低功耗的AI推理芯片。<sup>44</sup>而大多数自动驾驶汽车使用的GPU芯片解决方案，其性能仅略低于数据中心所采用的芯片。<sup>45</sup>这个非AI工厂市场的规模可能仍然相对较小（2026年不足50亿美元）<sup>46</sup>但可能成为要大得多，尤其是如果机器人市场腾飞，这种情况可能会发生，但很可能是在2030年之后。<sup>47</sup>

值得注意的是，我们仍处于AI发展的初期阶段。截至2025年夏季，对AI算力的需求（从而对更多数据中心、企业本地解决方案以及更高性能AI芯片的需求，无论是用于预训练、后训练、测试时扩展还是推理）增长十分迅猛，即便人们持续努力提升算法效率，这一趋势仍未减缓。<sup>48</sup>在某个阶段，新的技术可能会取得突破，更先进的AI模型将能够在成本更低的芯片上良好运行，从而减少对数据中心的需求并降低功耗。但这不会发生在2026年。

# 中国AI算力市场前景

## 中国AI算力不断扩容

中国十五五规划，统筹布局全国一体化算力网络，建设国家级AI算力中心，推动算力资源普惠化，降低社会算力总成本。推动中国AI算力的不断扩容。同时中国工业和信息化部于2025年5月印发《算力互联互通行动计划》，旨在加快构建算力互联互通体系，实现不同主体、不同架构的公共算力资源标准化互联，提高公共算力资源使用效率和服务水平，促进算力高质量发展。

## AI算力互联互通不断增强提效

到2026年，中国将建立较为完备的算力互联互通标准、标识和规则体系。在设施互联层面，将规模部署新型高性能传输协议，以提升算力节点间的网络性能；在资源互用层面，建成国家、区域及行业级算力互联互通平台，统一公共算力资源标识，实现全国主要算力服务商资源的接入与汇聚；在业务互通层面，推动计算、存储、网络等业务的协同调度，实现跨主体、跨架构、跨地域的算力智能供给；在应用创新层面，开展算力互联网试验网试点，探索普惠化的算力服务模式。这一系列举措旨在打通算力资源流动的关键环节，最终构建形成全国一体化调度、高效集约利用的算力互联网，为产业数字化升级提供坚实、普惠的算力底座。

## 国产芯片追赶，与国际巨头差距不断缩小

在外部环境制约和国家战略推动下，国产AI芯片正在加速崛起。以华为昇腾、寒武纪、阿里巴巴含光等为代表的本土力量正不断缩小与国际巨头的差距。未来随着国产芯片在推理赛道的进一步渗透、超节点/集群技术的规模化部署以及产业链闭环生态的完善，有望在更多细分市场实现追赶甚至领先。

## 国产化替代不断加快

全国多个智算中心项目中，国产化率大幅提升，中国移动呼和浩特智算中心AI芯片国产化率超85%，成渝（成都）天府数据中心国产芯片占比不低于60%，浙江东阳人工智能算力中心更是明确了70%的国产化目标。在AI算力中，国产化替代正不断加速。

**Duncan Stewart**  
Canada

**Arpan Tiwari**  
United States

**Karthik**  
India

**Jeroen Kusters**  
United States

**Girija Krishnamurthy**  
Global

**Roger Chung**  
China

**Deb Bhattacharjee**  
United States



## ENDNOTES

1. Rodrigo Liang, "Scaling AI without breaking the grid: The path to sustainable innovation," World Economic Forum, Jan. 3, 2025.
2. Michelle Weaver, "Big debates: The AI evolution," Morgan Stanley, Jan. 10, 2025.
3. Josh You and David Owen, "How much power will frontier AI training demand in 2030?" Epoch.AI, Aug. 11, 2025.
4. Jared Kaplan et al., "Scaling laws for neural language models," OpenAI, Jan. 23, 2020.
5. Amazon Web Services, "What are foundation models?" accessed Sept. 19, 2025.
6. Ashu Garg, "Has AI scaling hit a limit?" Foundation Capital, Nov. 27, 2024.
7. Aixin Liu et al., "Deepseek-v3 technical report," arXiv preprint arXiv:2412.19437 (2024).
8. Chris Arkenberg, Duncan Stewart, Gillian Crossan & Kevin Westcott, "On-device generative AI could make smartphones more exciting—if they can deliver on the promise," *Deloitte Insights*, Nov. 19, 2024; IDC Media Center, "Worldwide smartphone market forecast to grow 1% in 2025, driven by accelerated 3.9% iOS growth, according to IDC," Aug. 27, 2025; Gartner, Inc., "Gartner says artificial intelligence ("AI") PCs will represent 31 percent of worldwide PC market by the end of 2025," press release, Aug. 28, 2025.
9. Amazon Web Services, "AWS Inferentia," accessed Sept. 19, 2025.
10. Paul Lee and Clare Mortimer, "How citizens use devices and AI: what government needs to know," Deloitte UK, Aug. 29, 2025; Steve Feinberg, et al., "In the gen AI economy, consumers want innovation they can trust: Deloitte's 2025 Connected Consumer Survey," Deloitte, Sept. 25, 2025.
11. Wylie Wong, "Data center chips in 2024: Top trends and releases," Data Center Knowledge, April 11, 2024; Reen Singh, "AI inference chips latest rankings: Who leads the race?" Uvation, July 11, 2025; Broadcom Inc., "3.5D XDSiP AI Accelerator Platform," accessed Oct. 23, 2025.
12. Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components, and market sizes. This analysis is due to be published in December 2025.
13. Skye Jacobs, "NVIDIA Blackwell server cabinets could cost somewhere around \$2 to \$3 million each," TechSpot, July 28, 2024.
14. Beth McKenna, "2 key things from AMD's earnings call that investors should know," The Motley Fool, Feb. 1, 2024; Dell'Oro Group, "AI infrastructure spending sustains strong growth momentum," press release, Feb. 5, 2025.
15. Kari Briski, "How scaling laws drive smarter, more powerful AI," NVIDIA, Feb. 12, 2025.
16. Ibid.
17. Jonathan Farrington, "What is chain of thought prompting – AI prompt engineering," Silicon Dales, July 24, 2025.
18. Briski, "How scaling laws drive smarter, more powerful AI."
19. Ibid.
20. "Data centers: Computing risks and opportunities for U.S. real estate," S&P Global, Oct. 22, 2024; Equinix, Inc., "Form 10-K: Annual report for fiscal year ended Dec. 31, 2023," Feb. 16, 2024; Digital Realty Trust, Inc. and Digital Realty Trust, L.P., "Form 10-K: Annual report for fiscal year ended Dec. 31, 2023," Feb. 23, 2024.

21. Shubham Sharma, "Going beyond GPUs: The evolving landscape of AI chips and accelerators," VentureBeat, Sept. 26, 2024.
22. Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components, and market sizes. This analysis is due to be published in December 2025.
23. Eric Reed, "History of NVIDIA: Company and stock," SmartAsset, May 22, 2024.
24. Rajat Raina, Anand Madhavan, and Andrew Y. Ng, "Large-scale deep unsupervised learning using graphics processors," In Proceedings of the 26th Annual International Conference on Machine Learning, 2009.
25. NVIDIA, "NVIDIA delivers massive performance leap for deep learning, HPC applications with NVIDIA Tesla P100 accelerators," press release, April 5, 2016.
26. Hannah Wilson, "NVIDIA facts and statistics (2025)," Investing.com, Aug. 28, 2025.
27. Hannah Wilson, "NVIDIA facts and statistics (2025)," Investing.com, Aug. 28, 2025.
28. Top 500, "June 2025," June 2025.
29. Ray Wang, "NVIDIA's new Rubin CPX targets future of large-scale inference," Futurum, Sept. 18, 2025.
30. In 2025, Deloitte Consulting LLP performed an analysis of the data center market, including a rough bill of materials for the various components, and market sizes. This analysis is due to be published in December 2025.
31. Omdia, "New Omdia forecast: AI data center chip market to hit \$286bn, growth likely peaking as custom ASICs gain ground," Aug. 28, 2025.
32. Anthony Di Pizio, "Jensen Huang predicts annual data center spending will hit \$1 trillion by 2028. Here's the ultimate semiconductor ETF to buy right now." The Motley Fool, May 1, 2025.
33. Dave Lawler, "Exclusive: 'Massive ten-year' AI boom is just starting, AMD CEO says," Axios, Sept. 17, 2025.
34. Paul Mah, "AI training is going to multiple data centers," CDO Trends, Sept. 11, 2024.
35. Chris Thomas, Akash Tayal, Duncan Stewart, Diana Kearns-Manolatos, and Iram Parveen, "Is your organization's infrastructure ready for the new hybrid cloud?" Deloitte Insights, June 30, 2025.
36. Mike Robuck, "Google strikes deals for flexible AI data centre power use," Mobile World Live, Aug. 5, 2025.
37. Tyler H. Norris, Tim Profeta, Dalia Patino-Echeverri, and Adam Cowie-Haskell, "Rethinking load growth: Assessing the potential for integration of large flexible loads in US power systems," Nicholas Institute for Energy, Environment & Sustainability, Duke University, February 2025.
38. Anton Shilov, "TSMC could charge up to \$45,000 for 1.6nm wafers — rumors allege a 50% increase in pricing over prior-gen wafers," Tom's Hardware, June 4, 2025.
39. Francisco Jeronimo, "The rise of gen AI smartphones," IDC, July 5, 2024.
40. Dan Shipper, "Vibe check: OpenAI drops two new open-weight models," Every Media, Aug. 5, 2025.
41. Cyfuture Cloud, "NVIDIA DGX H100 price 2025: Cost, specs, and market insights," Cyfuture Cloud Knowledgebase, accessed October 2025.
42. Tae Kim, "NVIDIA's multi-million dollar AI servers are getting more expensive," Barron's, Aug. 28, 2025.
43. Skye Jacobs, "NVIDIA Blackwell server cabinets could cost somewhere around \$2 to \$3 million each," TechSpot, July 28, 2024.



44. Qualcomm, “[Flight RB5 5G platform](#),” accessed Sept. 19, 2025.
  45. Ali Kani, “[NVIDIA DRIVE Thor strikes AI performance balance, uniting AV and cockpit on a single computer](#),” NVIDIA, Sept. 20, 2022.
  46. There are a variety of suppliers for chips for driving assistance, but as one example, NVIDIA’s auto segment is at a US\$2 billion run rate as of August 2025: Pras Subramanian, “[NVIDIA's auto business surges 69% from self-driving tech](#),” Yahoo Finance, Aug. 25, 2025.
  47. Karthik Ramachandran, et al, “[AI for industrial robotics, humanoid robots, and drones](#),” Deloitte Insights.
  48. Jameel Rogers, “[AI chips for data center and cloud to exceed US\\$400 billion by 2030](#),” IDTechEx, May 8, 2025.
- 

## ACKNOWLEDGMENTS

The authors would like to thank **Brandon Kulik, Amy Scimeca, Karan Aggarwal, Kate Hardin, Diana Kearns-Manolatos, Mike Luk, Baris Sarer, Jason Chmiel, Dan Hamling, Jan Nicholas, Jordan Bish, Nitin Mittal, Rohit Tandon, Nicholas Merizzi, and Dan Littmann** for their contributions to this article.

Cover image by: **Jaime Austin**; Adobe Stock

## 关于德勤

德勤中国是一家立足本土、连接全球的综合专业性服务机构，由德勤中国的合伙人共同拥有，始终服务于中国改革开放和经济建设的前沿。我们的办公室遍布中国31个城市，现有超过2万名专业人才，向客户提供审计、税务、咨询等全球领先的一站式专业服务。

我们诚信为本，坚守质量，勇于创新，以卓越的专业能力、丰富的行业洞察和智慧的技术解决方案，助力各行各业的客户与合作伙伴把握机遇，应对挑战，实现世界一流的高质量发展目标。

德勤品牌始于1845年，其中文名称“德勤”于1978年启用，寓意“敬德修业，业精于勤”。德勤全球专业网络的成员机构遍布150多个国家或地区，以“因我不同，成就不凡”为宗旨，为资本市场增强公众信任，为客户转型升级赋能，为人才激活迎接未来的能力，为更繁荣的经济、更公平的社会和可持续的世界开拓前行。

Deloitte（“德勤”）泛指一家或多家德勤有限公司，以及其全球成员所网络和它们的关联机构（统称为“德勤组织”）。德勤有限公司（又称“德勤全球”）及其每一家成员所和它们的关联机构均为具有独立法律地位的法律实体，相互之间不因第三方而承担任何责任或约束对方。德勤有限公司及其每一家成员所和它们的关联机构仅对自身行为承担责任，而对相互的行为不承担任何法律责任。德勤有限公司并不向客户提供服务。请参阅[www.deloitte.com/cn/about](http://www.deloitte.com/cn/about)了解更多信息。

德勤亚太有限公司（一家担保责任有限公司，是境外设立有限责任公司的其中一种形式，成员以其所担保的金额为限对公司承担责任）是德勤有限公司的成员所。德勤亚太有限公司的每一家成员及其关联机构均为具有独立法律地位的法律实体，在亚太地区超过100个城市提供专业服务，包括奥克兰、曼谷、北京、班加罗尔、河内、香港、雅加达、吉隆坡、马尼拉、墨尔本、孟买、新德里、大阪、首尔、上海、新加坡、悉尼、台北和东京。

本通讯中所含内容乃一般性信息，任何德勤有限公司、其全球成员所网络或它们的关联机构并不因此构成提供任何专业建议或服务。在作出任何可能影响您的财务或业务的决策或采取任何相关行动前，您应咨询合格的专业顾问。

我们并未对本通讯所含信息的准确性或完整性作出任何（明示或暗示）陈述、保证或承诺。任何德勤有限公司、其成员所、关联机构、员工或代理方均不对任何方因使用本通讯而直接或间接导致的任何损失或损害承担责任。

© 2025。欲了解更多信息，请联系德勤中国。

CQ-017-CN-25



这是环保纸印刷品