IDC | 浪潮信息

2025年

中国人工智能计算力发展评估报告

目录

核心观点

算法是驱动是人工智能发展的核心引擎，决定了应用的智能上限，也牵引着算力的发展。2024年，o系列、Llama3、通义千问、R1等大模型不断升级，尤其是DeepSeek R1系列模型的发布，正是基于算法层面的极大创新，对中国乃至全球的人工智能产业带来深刻变革。一方面DeepSeek采用了大规模强化学习、多头注意力机制等算法创新，智能水平在美国高中数学竞赛邀请赛AIME、博士水平科学问答等测试中榜单上接近甚至超过了OpenAI的o1模型；另一方面，DeepSeek R1算法的创新也带来训练和推理阶段算力消耗的降低，训练算力只有Llama3的1/10，推理阶段缓存数据量降低了50倍，为在算力约束的条件下进行AI算法创新提供了一个全新思路，吸引了全球开发者，7天实现了活跃用户数破亿。

规模法则（Scaling law）在当前人工智能发展中仍然占主导地位，推高人工智能算力需求。目前规模法则正在从预训练扩展到了后训练和推理阶段，基于强化学习、思维链等算法创新在后训练和推理阶段更多的算力投入，可以进一步大幅提升大模型的深度思考能力。同时基于杰文斯悖论的现象表明，DeepSeek带来的算法效率的提升并未抑制算力需求，反而因更多的用户和场景的加入，推动大模型普及与应用落地，重构产业创新范式，带动数据中心、边缘及端侧算力建设。IDC数据显示，2024年全球人工智能服务器市场规模预计为1,251亿美元，2025年将增至1,587亿美元，2028年有望达到2,227亿美元，其中生成式人工智能服务器占比将从2025年的29.6%提升至2028年的37.7%。

中国智能算力发展水平增速高于预期。在中国，企业加速生成式人工智能布局和投入，IDC调研结果显示，目前42%的中国企业已经开始进行大模型的初步测试和重点概念验证，17%的企业已经将技术引入生产阶段，并应用于实际业务中，在未来18个月内，硬件升级将成为企业的首要投资目标。在旺盛的市场需求、丰富应用场景的驱动下，中国智能算力规模呈现增长态势。IDC最新预测结果显示，2025年中国智能算力规模将达到1,037.3 EFLOPS，并在2028年达到 2,781.9 EFLOPS，2023-2028年中国智能算力规模和通用算力规模的五年年复合增长率分别达46.2%和18.8%，较上一版本预期值33.9%和16.6%有显著提升。中国人工智能算力基础设施发展呈现出多元化、服务化、场景化、绿色化等特征。

大模型的开源趋势正在显著增强，成为加速AI普惠、降本增效的重要力量。开源模型，通过大幅降低训练部署成本并提供与闭源模型性能水平相当的能力，正成为推动人工智能技术普及和应用落地的重要力量。在过去的18个月里，全球领先的软件和云服务商发布了数十种开放和部分开放的基础模型，开源社区的协作和贡献正在成为加速技术创新的重要力量，开源框架作为人工智能开发的基础，其生态系统日益丰富。IDC预测，2025年，为了更快获得创新能力、运营主权、透明度和更低成本，将有55%的企业使用开源人工智能基础模型开发应用程序。

“扩容”与“提效”并行推动人工智能应用落地。为应对生成式人工智能和大模型应用扩展带来的数据、算力、模型、人才、成本等多方面挑战，尤其是算力基础设施的瓶颈，IDC建议企业采取“扩容”与“提效”并行策略，通过提升算力供给能力和质量，优化基础设施架构，增强数据支持和模型效率，系统性地提高算力利用率。**提高模型架构效率与增加原始计算能力同等重要，**更经济的先进模型将推动AI需求，增加效率——而不仅仅是原始计算能力——可能是真正的竞争优势。算法创新与模型迭代是提升模算效率的关键，通过算法创新，如模型剪枝、知识蒸馏、设计高效模型架构、分布式计算等方法，能在保障模型能力的前提下减少模型计算量和存储需求，降低同等精度水平下的算力成本，加速人工智能技术的应用和商业化进程。

人工智能算力服务市场蓬勃发展，算力供给模式不断创新。企业对智能算力基础设施和服务能力的需求正在发生深刻变化，传统算力技术架构和云服务模式难以满足新需求。生成式人工智能将推动企业更多使用人工智能就绪数据中心托管设施和生成式人工智能服务器集群，缩短部署时间，降低资本成本。这一变化挑战了传统算力服务的优势，带来新的市场机遇，促使算力服务商不断创新，提升技术水平和服务质量，并通过合作机制重新分配资源与市场，形成由数据中心服务商、云服务商、硬件制造商以及其他创新企业共同参与的产业生态，通过生成式人工智能 IaaS服务、算力租赁、算力共享、智算中心等算力供给模式满足多样化的智能算力需求。IDC数据显示，2024年中国智算服务市场整体规模达到50亿美元，2025年将增至79.5亿美元，2023-2028年五年年复合增长率达57.3%。

人工智能算力发展将坚持绿色可持续原则，液冷技术成为关注重点。IDC预测，2025年，人工智能数据中心IT能耗将达到77.7太瓦时（TWh），是2023年能耗量的两倍，2027年将增长至146.2太瓦时，2022-2027年五年年复合增长率为44.8%，五年间实现六倍增长。面对这一挑战，业界积极探索破局之道，液冷技术作为关键突破，可以显著提升计算密度，降低数据中心的总能耗，通过全栈液冷方案，推动算力设施在计算节点层面、机柜层面以及数据中心层面的绿色化和低碳化转型。IDC预测，2028年中国液冷服务器市场将达到105亿美元，2023-2028年五年年复合增长率将达到48.3%。

人工智能行业渗透度持续增加，城市走出各具特色的发展路径。人工智能行业渗透度排名前五的行业依次为：互联网、金融、运营商、制造和政府，其中，互联网企业在大模型的研发、应用及推广过程中持续发挥引领作用；金融行业进一步加深人工智能与风控、投资决策和个性化财富管理等场合的融合，排名从第四名攀升至第二名；制造业持续加速智能化转型，扩大人工智能技术在生产线、产品设计、运营和安监等场景的应用，排名从第五名提升至第四名。中国人工智能城市评估框架首次将大模型架构及生成式人工智能技术的投资、建设进度和规划布局纳入关键指标，评估结果显示，北京凭借其科研资源和人才优势成为人工智能创新中心，继续领跑发展，位居首位；杭州和上海分别位列第二和第三，其中上海凭借其国际化优势和政策支持，在推动人工智能世界级产业集群建设等方面表现出色，较前一年排名上升了一位。此外，广州、成都、天津、厦门等城市的排名均有所提升。

PART

第一章

全球及中国人工智能发展概述

1.1 全球：生成式人工智能成为重要新型工作负载，人工智能算力呈现五大发展趋势

全球人工智能市场持续呈现增长态势，成为各行业智能化升级的重要驱动力。生成式人工智能和大模型是推动人工智能技术迅猛发展的关键因素，深度学习、强化学习和迁移学习等核心技术的突破，使得模型在处理复杂任务时变得更加高效，进而在更多商业化场景中得以落地，并逐渐影响社会经济的方方面面。在技术创新、应用场景拓展的多重驱动下，全球企业对于人工智能技术的投资普遍提升，IDC预测，2025年全球2000强企业会将超过40%的IT预算投入到人工智能项目中，旨在推动产品和流程创新，并促成两位数的营收增长。

从区域角度来看：

- **美国**凭借政府支持、资本活力、技术基础和创新生态，在人工智能市场继续引领全球发展。美国政府持续加大对人工智能发展的支持力度，2024财年，美国“网络与信息技术研发计划”（NITRD）人工智能研发投资预算增长至31亿美元，占整体财年预算的近三分之一，相比于上一年提高了19.2%；2025年1月，美国政府公布“星际之门”国家级计划，该计划预计将投入5000亿美元用于美国国内人工智能基础设施建设，显示出其更倾向于支持人工智能发展而非严格监管的立场。对于企业而言，其对于人工智能的投资也在不断增加，如微软、Meta、谷歌等大型科技公司陆续宣布了数十亿美元的投资计划，用于建设和升级人工智能基础设施；医疗、金融和自动驾驶等行业也纷纷加大在人工智能技术研发上的投入。此外，基于良好的算力和数据支撑，美国在基础模型研发和应用方面也处于高水平，据IDC不完全统计，2024年1月至9月全球推出的四十多个重要开源语言模型中，三分之二来自于美国。

- **亚太地区**人工智能发展呈现出多样化和快速增长的特点。中国继续引领亚太地区人工智能市场发展，逐步成为全球人工智能强国，建立了多个国家级的人工智能实验室和研究中心，在研发更高算力服务器与芯片、开发生成式人工智能两项主线任务之外，全方位构建包含基础设施、算法工具、智能平台和解决方案的产业生态，持续以需求为牵引，加速智能技术的行业落地，并加快绿色技术研发，推进生成式人工智能的可持续发展；日本政府积极推进社会的智能化转型，2024年宣布设立规模超过650亿美元的投资基金，用于支持芯片和人工智能行业的发展，其中约420亿美金将用于下一代芯片的研发，并支持功率芯片的量产，2025年日本持续加速推进人工智能新法的相关立法准备工作，通过制定相应的法律法规来规范人工智能的应用和发展；新加坡在人工智能方面具备较强的学术能力与政府扶持力度，2024年预算案中，新加坡政府表示未来五年将投资10亿美元用于人工智能计算、人才和产业发展，2025年新加坡政府推出人工智能教育计划，所有中小学将开设人工智能课程，旨在培养学生的数字素养和人工智能知识。

- **欧洲**诸多国家政府在过去一年也积极推动人工智能技术的发展，同时出台了多项政策和法规，确保人工智能技术的安全使用。2024年8月，全球首部全面监管人工智能的法规《人工智能法案》开始生效，以更好规范人工智能应用的透明度和安全性，将对市场参与者的业务模式产生积极影响。2024年10月，英国成立了专门的监管创新办公室，加速人工智能等技术的创新和应用。欧盟委员会推出一揽子人工智能创新计划，其中包括计划在2025年初启动首批人工智能工厂的建设，这些工厂通过整合尖端计算能力、数据资源和人才，将为初创企业、中小型企业和科学家等开发可信、前沿的生成式人工智能模型提供计算、存储和数据等服务。目前，欧盟已经收到来自芬兰、卢森堡、瑞典、德国、意大利、西班牙和希腊等七个国家的人工智能工厂提案。在市场方面，欧洲企业在数字化和人工智能领域的投资持续加大，发展潜力逐步释放。

全球范围内人工智能技术的加速发展与生成式人工智能的持续创新密切相关，生成式人工智能正在成为企业重要新型工作负载。IDC全球调研数据显示，85%的企业认为生成式人工智能将与ERP、电子商务一样，成为企业重要的新型工作负载。基于大模型强大的计算能力和学习能力，生成式人工智能技术取得了突破性的进展，其能力可覆盖内容生成、数据增强、创意辅助等诸多应用场景，极大地提高了生产效率，为用户带来全新的体验，并进一步助推企业整体智能化发展，加速人工智能技术的广泛应用。IDC数据显示，目前全球超过70%的组织已经开始对生成式人工智能技术进行投资或处于初步测试阶段，已经有17%的组织将生成式人工智能应用和服务引入生产环节；2025年全球企业生成式人工智能支出预计将达到691亿美元，2028年超过2,022亿美元，2023-2028年五年年复合增长率为59.2%（人工智能市场整体同期年复合增长率为29.0%）。

伴随大模型技术的持续发展和生成式人工智能新兴应用场景不断涌现，全球人工智能算力发展正在呈现出新的发展趋势。

趋势一：规模法则（Scaling law）在当前人工智能发展中仍然占主导地位，推高人工智能算力需求

规模法则（Scaling law）目前正在从预训练扩展到后训练和推理阶段。基于强化学习、思维链等算法创新在后训练和推理阶段更多的算力投入，可以进一步大幅提升大模型的深度思考能力。

同时，基于杰文斯悖论的现象表明，DeepSeek带来的算法效率的提升并未抑制算力需求，反而因更多的用户和场景的加入，推动大模型普及与应用落地，重构产业创新范式，带动数据中心、边缘及端侧算力建设。DeepSeek系列模型的发布对中国乃至全球人工智能产业带来巨大变革，其通过技术普惠化、场景纵深化和算力泛在化三重路径，推动大模型的普及与应用落地，驱动算力需求增长。

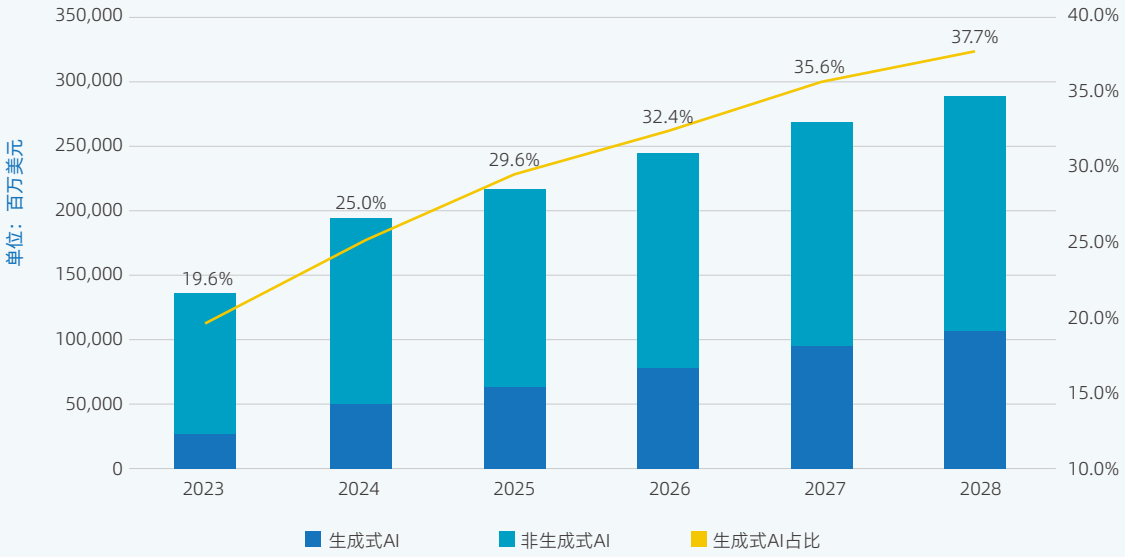
- **技术普惠化：**DeepSeek的核心技术不仅显著提升了模型性能，还大幅降低了算力消耗，为用户参与大模型应用生态创造了条件，引领一场从单纯算力扩张转向增效提质的产业变革。DeepSeek通过开源开放战略和轻量化部署，降低技术使用门槛，使更多企业和开发者能够便捷地获取先进技术，开源框架和工具的普及，让中小企业和个人开发者也能参与人工智能应用开发，推动技术民主化。

- **场景纵深化：**得益于其强大的语言处理能力、经济高效的训练过程以及对特定业务需求的高度适应性，DeepSeek在金融、医疗、汽车、电信等多个行业逐步落地，重构产业模式，提升企业的运营效率和服务质量，通过人工智能优化生产流程，降低成本的同时激发更大需求，验证了“杰文斯悖论”——效率提升带来成本下降，进而推动需求激增，形成良性循环。

- **算力泛在化：**DeepSeek通过其先进的算法优化和高效的模型性能，促进了人工智能技术在C端和B端用户中更广泛的应用，显著拉动了人工智能算力在数据中心、端侧及边缘侧的发展。在数据中心，DeepSeek不仅提高了训练和推理效率，降低了能耗，还因其高性能吸引了更多企业部署复杂人工智能解决方案；在端侧，DeepSeek提供了轻量化模型版本，使得智能手机、智能家居等设备能够执行如实时翻译、语音识别等复杂人工智能任务；在边缘计算领域，DeepSeek能够在边缘设备上执行关键分析任务，实现低延迟响应和分布式智能处理，减轻中央服务器的压力并提高系统的可靠性和安全性。

IDC数据显示，2024年全球人工智能服务器市场规模预计为1,251亿美元，2025年将增至1,587亿美元，2028年有望达到2,227亿美元，其中生成式人工智能服务器占比将从2025年的29.6%提升至2028年的37.7%。

图1 全球生成式人工智能和非生成式人工智能服务器市场规模预测



来源：IDC，2025

此外，大模型及应用正在驱动计算架构和数据中心变革：

- **在计算架构层面，**大模型的训练和应用通常需要处理大规模的数据集，这将增加对于高带宽的需求，以执行数据并行、流水线并行及张量并行等策略。为了满足大模型对计算资源的高需求，提升单节点的计算性能（Scale-up）变得至关重要，这包括增加单芯片或单个机架的计算能力。通常，配备8块高性能GPU的服务器可以支持具有2,000亿参数的大模型训练，而当插卡数量可扩展至72块高性能GPU时，则能够有效支持具有万亿参数的大模型训练，这将有效加速智能涌现的实现。其次，通过增加节点数量，实现计算能力的横向扩展（Scale-out），也正在被用于需要处理大规模数据集和复杂模型的应用场景。高速互联网络（以太网和硅光子技术）和分布式计算框架将有效支持千卡、万卡甚至十万卡的集群建设。通过构建具有更高性能的计算集群，支持更复杂的大模型计算和多样化的应用场景。此外，伴随大模型从训练阶段迈向应用阶段，推理工作负载将持续增加，面向应用和推理需求对芯片和系统架构进行设计愈加重要，大语言模型推理包含两个重要的阶段：预填充（Prefill）和解码（Decode），两个阶段处理token序列的长度不同，对计算和存储资源的访问频率和调度需求也不同，实操中往往采用P-D解耦部署策略，通过构建分离式算力资源池，缩短计算时间，降低计算成本，提高资源利用率。

- **在数据中心层面，**首先，提高集群系统的可用性和可靠性十分重要，从千卡集群到万卡、十万卡集群，节点故障几率会随集群规模增长而上升，数据中心需要更加高效的监控体系和先进的故障恢复机制，基于诸如智能显存分配、故障点恢复管理等技术，确保集群在发生节点故障时能够迅速响应，最小化停机时间。其次，应重视算力体系的兼容性和可扩展性建设，在执行模型训练、推理等工作任务时，CPU、GPU、ASIC等不同类型的计算资源各具优势，因此需要协同异构基础设施，将整个数据中心作为协同工作的有机体，整合多种计算资源，优化数据处理流程和模型训练效率，通过灵活的计算任务调度，高效执行人工智能任务。最后，随着单机柜性能大幅提升，能耗将持续攀升，通常GPU功耗在250W到700W之间，服务器单机柜功率可高达130KW，数据中心应持续优化能耗方案，通过优化空间规划、供电系统，并采用先进冷却技术，提高散热特性，应对能耗挑战。

趋势二：企业更加重视发挥平台价值，构建互联的生态体系

生成式人工智能代表了一种全新的技术范式，这种范式要求企业从硬件到软件、从开发工具到用户体验实现全面创新。若将生成式人工智能发展作为企业战略性工作负载，企业需要寻求新的供应商和合作伙伴支持生成式人工智能落地，IDC数据显示，全球85%的组织认为，需要制定全新的供应商/合作伙伴战略，在基础设施、软件、数据、云等维度获得不同的服务能力。

鉴于生成式人工智能技术栈复杂、供应链漫长，为企业提供低门槛的生成式人工智能应用开发平台越来越重要。通过平台整合服务能力，企业可获得模型构建和精排、应用开发与部署、数据管理等相关软件及工具，以及资源统筹和调度管理等服务和先进的行业智能化解决方案，从而有效加速先进技术落地和商业价值实现。生成式人工智能应用与开发平台不仅是技术工具的集合，更是互联生态的载体，其应具备开放性、互操作性、灵活性和适应性，通过围绕生态构建自身价值，帮助企业简化集成流程、实现资源高效扩展、推动跨供应商的一致性和互操作性，应对数据治理等方面面临的挑战。

趋势三：面向人工智能场景构建先进数据基础设施，并打造高质量数据集

生成式人工智能重塑了数据生命周期特征，数据的生成、采集、存储、处理和分析变得更加复杂。全球数据量持续增长，IDC数据显示，2024年全球产生的总数据量达到163ZB，2025年将增至201.6ZB，2028年将翻番至393.9ZB，2023-2028年五年年复合增长率为24.4%；生成式人工智能还将带来更多的混合内容生成和处理需求，目前，生成式人工智能生成的数据中，文本内容占比超35%，到2028年，图像和视频类数据占比将增加，超过75%的生成数据将均匀覆盖文本、图像和视频三种类型，此外，还有接近18%的生成数据为软件代码。企业需要面向训练和推理过程中的数据特征，构建先进数据基础设施，为数据收集、预处理、写入读出、稳定训练集、数据安全、推理结果使用等环节提供支撑，并根据数据量、访问模式及成本效益决定采用云存储、本地存储或混合存储方案，发挥先进存储介质和存储架构优势。

伴随数据逐渐成为企业的核心资产和重要生产要素，企业需要提升数据质量和数量，从而优化企业决策和业务流程。这一需求将促进企业使用数据增强工具、数据合成等方式提高数据质量。数据增强工具可以通过对现有数据进行扩展和优化，提升数据的多样性和代表性；合成数据通常由现有数据集或真实世界事件/流程的模拟数据创建而成，用来代替真实世界数据进行应用测试或人工智能训练，其比收集足够的真实世界数据更具成本效益和效率。通过这些服务，企业可以更好地管理和利用数据，支持人工智能应用的开发和部署，实现业务价值的最大化。

趋势四：优化策略制定、关注技术创新，加速实现投资高效回报

深度学习和生成式人工智能模型的规模和复杂性增加，使得支持这些模型的基础设施变得更加复杂、庞大且昂贵。为避免给企业造成不必要的财务负担，企业在投资人工智能基 础设施时，需要创建合适的投入产出比（ROI）模型，将投资回报率与生成式人工智能的应用案例和业务成果联系起来，通过加速应用部署，使高投资实现价值回报。生成式人工智能带来的成果既包括可通过KPI衡量的有形收益，如加速内容创作、提高客服效率和降低成本；也涵盖无形价值，比如提升员工体验、加强客户关系和忠诚度以及优化品牌营销。这些软性收益虽然不易量化，但对企业的长期成功至关重要。企业可以利用技术评估、项目组合管理和企业整体战略管理等方法，制定符合自身发展需求的生成式人工智能投资和应用策略。

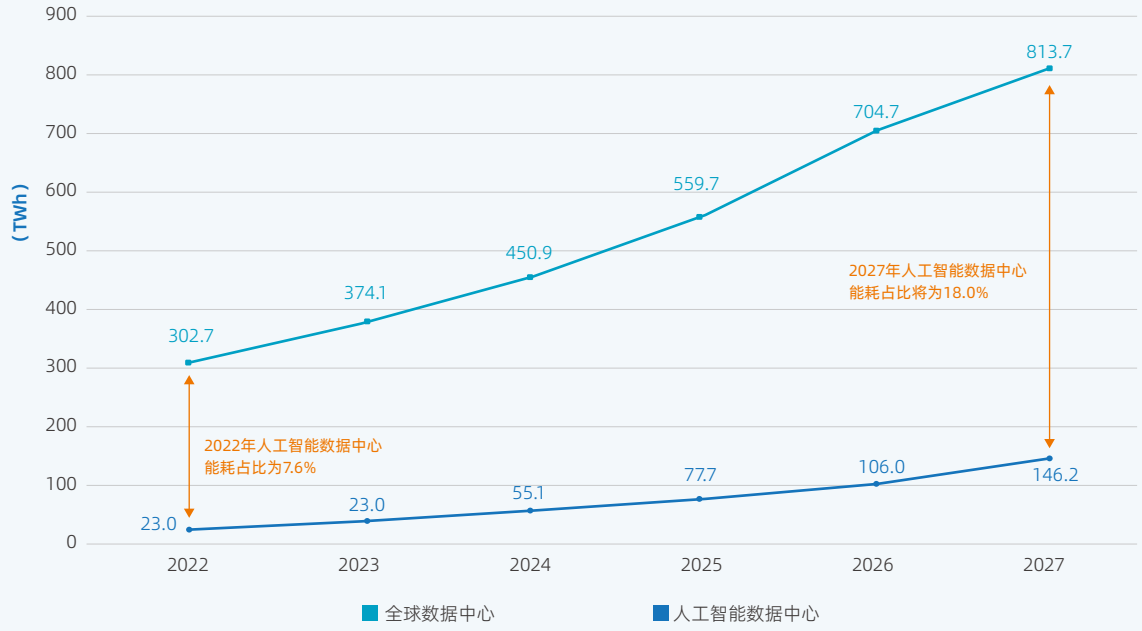
企业会更加重视成本效益更高的人工智能解决方案，采用有望降低人工智能基础设施要求和成本的新技术，如小语言模型（SLM）、窄人工智能模型和稀疏人工智能模型，以及通过检索增强生成（RAG）对较小模型进行定制的方法。

在硬件创新方面，厂商也正在推出先进的人工智能芯片方案，加速生成式人工智能工作负载的处理，提高性价比。同时，RISC-V作为一种开放源码架构，也可以为定制化人工智能硬件解决方案提供灵活性和成本效益。

趋势五：能耗挑战持续加剧，冷却技术不断创新

人工智能大模型技术的研发和应用带来了更高的能耗需求，IDC数据显示，2024年人工智能数据中心IT能耗（含服务器、存储系统和网络）达到55.1太瓦时（TWh），2025年将增至77.7太瓦时，是2023年能耗量的两倍，2027年将增长至146.2太瓦时，2022-2027年五年年复合增长率为44.8%，五年间实现六倍增长。

图2 全球数据中心及人工智能数据中心能耗预测，2022-2027



来源：IDC，2025

《欧洲绿色协议》、《美国新能源法案》、《数字中国建设整体布局规划》、《企业可持续发展报告指令》等可持续发展政策法规的施行，对现有数据中心基础设施提出了更严格的能耗要求。大模型的训练和优化作为能源密集型任务，需要高密度机架的支持，而这些机架的能耗已超出传统风冷的能力范围，促使越来越多的数据中心转向使用液冷技术。IDC预测，到2028年，60%的数据中心将采用微电网、定制硅芯片、液体冷却和加固结构等创新解决方案，以应对电力短缺和日益增长的可持续性要求。

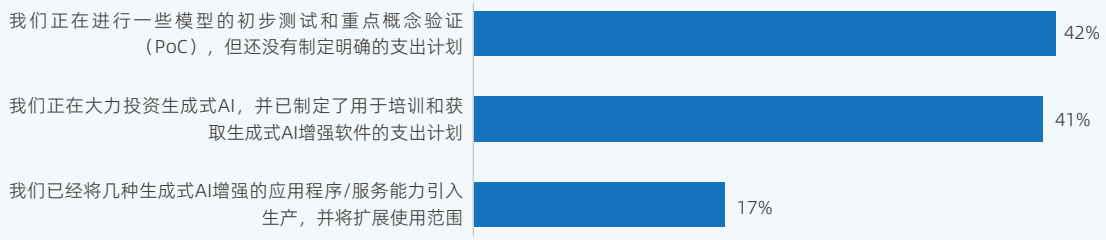
此外，数据中心正积极探索从储电能向储大模型算力的跨越性转变。这一转型的核心策略，在于通过大幅增加和优化IT基础设施，将原本静态储存的电能，转化为驱动大规模计算模型的动态算力。不再满足于电能的简单储备，而是致力于让算力靠近电力源头，利用先进的IT基础设施，如高性能服务器、高效能存储系统等，实现电能到算力的即时、高效转换。这些基础设施的升级与扩容，不仅提升了数据中心的算力水平，更为基于大模型的“智能蓄势”提供了稳定、强大的计算支撑。通过这一转变，智算数据中心不仅可以优化能源利用结构，减少能源浪费，更以就近、快速的方式，满足现代计算对算力的迫切需求，为数字经济的蓬勃发展注入强劲动力。

1.2 中国：系统性提高算力效能，加速智能涌现和智能应用

中国积极稳步扩大算力规模，探究生成式人工智能在行业中的应用

中国“两会”政府报告重点强调了数字基础设施建设、积极推进数字产业化和产业数字化、积极培育新兴产业和未来产业等七大加速发展领域。党的二十届三中全会也提出要推动实体经济和数字经济的融合发展，为高质量发展提供新动能。在政策的鼓励和引导下，中国企业将人工智能作为产业创新的抓手，加速探究生成式人工智能等先进技术在行业中的应用，越来越多的中国企业正在积极制定和实践人工智能转型战略。IDC调研显示，42%的中国企业已经开始进行大模型的初步测试和重点概念验证，17%的企业已经将技术引入生产阶段，并应用于实际业务中。

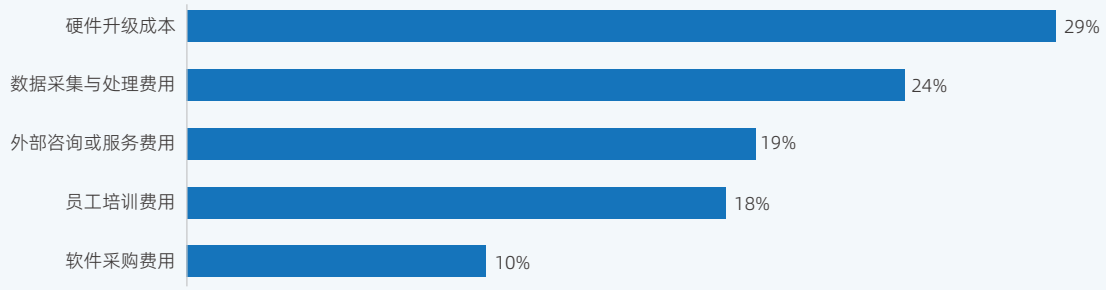
图3 中国企业生成式人工智能的应用现状



来源：IDC，2025

对大型模型及生成式人工智能需求的日益增长，正显著推动中国人工智能算力基础设施的快速发展，促使企业不断升级其硬件配置，通过采购高性能的计算设备、优化数据中心基础设施、提升存储和网络能力等，进一步支持复杂的人工智能运算任务。这一趋势不仅反映了市场对先进人工智能技术的迫切需求，也体现了中国企业在全球人工智能竞赛中的积极布局和投入。IDC调研结果显示，在未来18个月内，中国企业在生成式人工智能项目上的投资将首要集中在硬件升级方面。

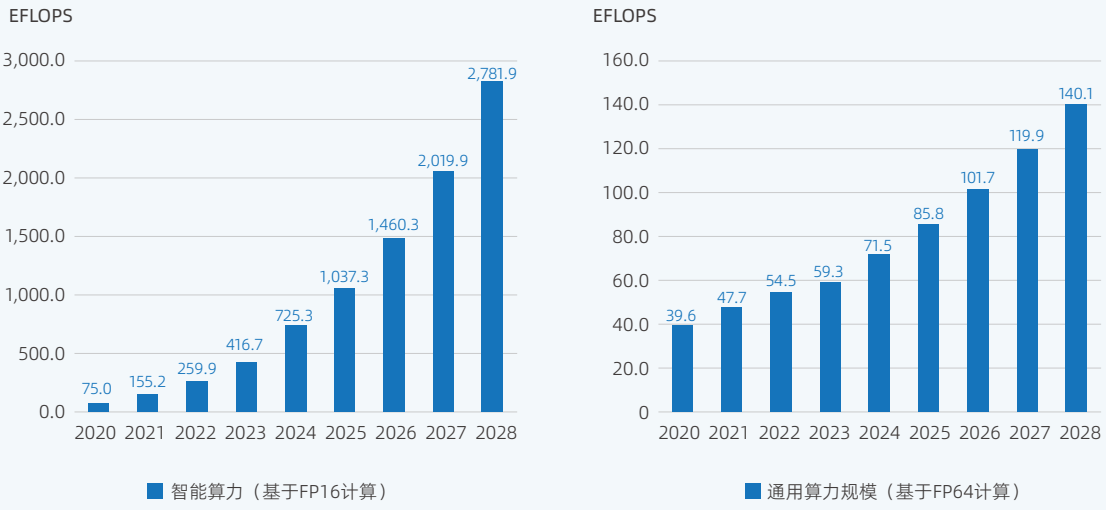
图4 未来18个月，中国企业生成式人工智能项目的主要支出方向



来源：IDC，2025

为评估中国算力规模发展现状和趋势，本报告基于《IDC中国加速计算服务器半年度市场跟踪报告》及智能加速卡半精度（FP16）相当运算能力数据，测算了中国智能算力规模。结果显示，2025年中国智能算力规模将达到1,037.3 EFLOPS，预计到2028年将达到2,781.9 EFLOPS。此外，本报告基于《IDC中国服务器市场季度跟踪报告》及CPU双精度（FP64）运算能力数据，测算了中国通用算力规模。2025年中国通用算力规模将达到85.8 EFLOPS，预计到2028年将达到140.1 EFLOPS。预测显示，2023-2028年期间，中国智能算力规模的五年年复合增长率预计达到46.2%，通用算力规模预计达到18.8%。较上一版本预期值33.9%和16.6%，均有显著提升。

图5 中国智能算力和通用算力规模及预测，2020-2028



来源：IDC，2025

中国人工智能算力基础设施发展主要特征

在旺盛的市场需求、丰富应用场景的驱动下，中国人工智能算力基础设施呈现出快速发展的趋势，并表现出如下发展特征：

- **多元化：**由于人工智能在国内的应用场景较为复杂，同时受到地缘、供应链等因素影响，人工智能芯片类型与厂商呈现出多元化的趋势，GPU、CPU、DSA、ASIC等多种芯片被广泛应用在人工智能的训练与推理中，不少场景实现了多类型芯片的混合使用。在芯片厂商方面，诸多国内厂商开始崭露头角，提供了大规模的算力集群方案。

- **服务化：**为满足不同的算力需求，人工智能算力服务不断革新。生成式人工智能IaaS服务可为企业按需提供容量，支持灵活的模型训练和内容生成；算力租赁使用户按需租赁计算资源，降低成本并提高灵活性；算力共享通过资源池化和动态调度，实现资源共享和成本分摊；智算中心集成高性能的计算、存储和网络资源，提供高可用的一站式服务，支持大规模人工智能应用。这些服务化模式不仅提高了资源利用率和灵活性，还降低了用户的使用成本，推动了人工智能技术的广泛应用。

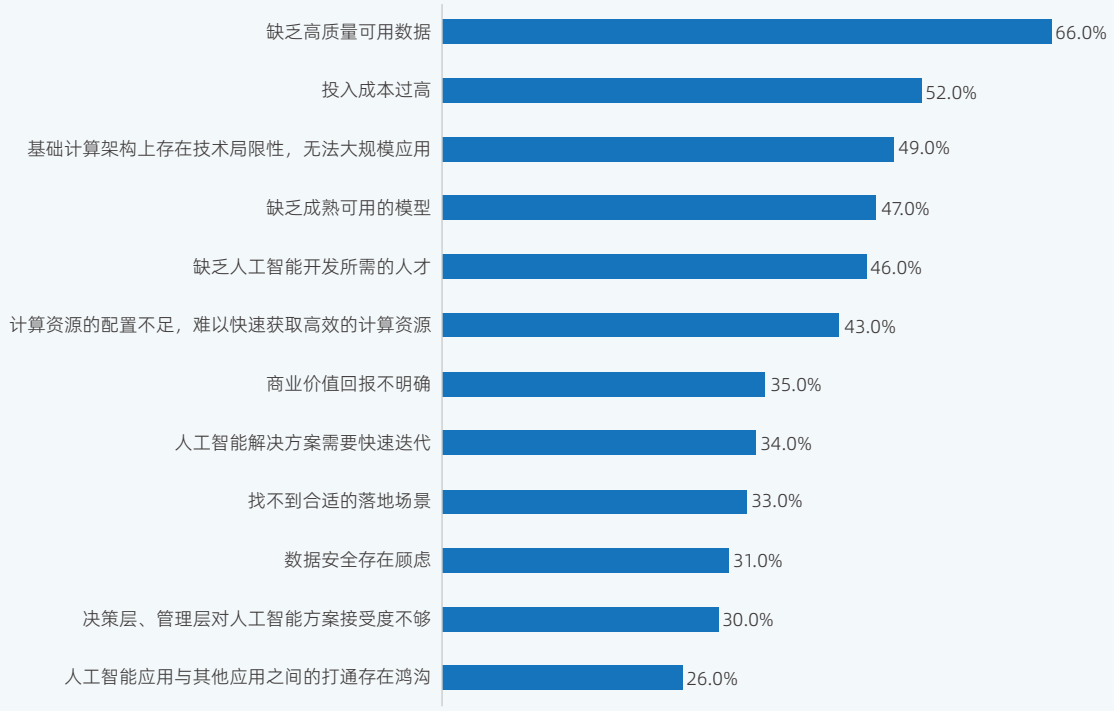
- **场景化：**人工智能算力基础设施的多样化旨在应对不同行业和应用场景的多样化需求。各行业对于数据处理的需求各具特色，例如金融行业需要高安全性和低延迟的计算环境，医疗行业需要处理大量医学影像数据，制造业需要实现实时生产优化，互联网行业需要处理大规模用户数据和内容推荐。这些需求对底层架构提出了高性能、低延迟、高安全性、可扩展性和成本效益等新要求。通过资源池化、智能调度、多租户支持、异构计算和边缘计算等技术，人工智能算力基础设施能够灵活应对这些需求，确保资源的高效利用和业务的稳定运行，从而推动人工智能技术在各行业的广泛应用。

- **绿色化：**在双碳目标的指引下，全国范围内对绿色发展的重视程度和投资力度不断加大。政策上来说，一方面国家对于数据中心的新建审批及能耗要求上持续趋严；另一方面也会通过诸如电费分段计价等引导数据中心绿色化发展。技术上来说，越来越多的数据中心采用先进的计算、存储和网络架构，提高计算效率；使用液冷服务器等节能技术，提高散热效率，降低PUE值至1.3甚至1.15以下，满足可持续发展政策的要求。

中国人工智能算力发展面临的挑战与解决方案

IDC研究发现，随着生成式人工智能和大模型逐步扩大应用，企业将面临来自数据、算力、模型、人才、成本等多方面的挑战。其中，算力基础设施是关键议题，企业当下面临的相关挑战包括但不限于计算架构难以支持大规模应用、与基础设施建设和维护相关的高昂成本投入，以及高性能的计算资源的不足。

图6 企业在应用人工智能过程中面临的挑战



来源：IDC，2025

企业在人工智能大模型训练、推理阶段，会面临不同的算力挑战。对于持续开展大模型训练和研发的企业和研究机构而言，他们需要完成大量计算任务，推高算力需求，将长期处于高性能算力供不应求的状态；随着大模型和生成式人工智能技术在实际应用场景中落地，企业普遍面临以推理负载为主的算力需求，在推理阶段，算力分配和调度是主要问题，推理任务的算力需求具有波动性，难以预测和管理，导致资源分配不均衡，缺乏有效的算力分配和调度机制，导致算力资源的局部浪费和整体利用率低下，这不仅影响了人工智能基础设施的算力效率，也增加了整体成本。

此外，对于调整技术发展路径的科技企业或者行业巨头而言，如放弃自研大模型转用第三方模型，结束大模型训练转向模型推理，或通过模型剪枝、量化等方法降低模型算力需求，可能会出现算力盈余的情况。同时，在智算中心的积极建设的过程中，部分中心也出现了在实际运营中算力利用率未达预期的情况。

为应对以上挑战，IDC认为在人工智能算力发展过程中应采取“扩容”和“提效”并行的策略：

提升算力供给能力，提高算力供给质量

- **增强算力资源的可获得性：**增加智算中心的数量，实现充足的多元算力供给，支持人工智能技术的研发和应用，并在不同地区合理规划智算中心分布，实现不同区域协调发展。同时智算中心呈现出规模化发展趋势，政府智算中心单期算力规划可达千P级及以上，而运营商和互联网公司的智算中心则致力于实现万卡及以上的算力规模部署。应面向未来适度超前规划并建设智算中心，支持科技创新和业务扩展，重视规模扩大的同时，还要注重技术先进性。

- **优化算力基础设施架构：**采用先进的计算架构，提升单计算节点性能，提高计算效率。优化内存层次结构，减少数据传输延迟，增强数据处理速度。利用智能调度算法合理分配计算任务，优化集群管理方面，确保资源高效利用。

- **推动产业聚集效应：**充分考量区域产业基础和发展目标，通过集中建设智算中心，吸引相关产业链上下游企业集聚，形成规模效应和技术溢出效应，降低低单位算力成本的同时，带动地区经济发展和技术进步。

系统性地提高算力利用率

- **提高模型算力效率：**通过算法创新，如模型剪枝、知识蒸馏、设计高效模型架构、分布式计算等方法，在保障模型能力的前提下减少模型计算量和存储需求，降低同等精度水平下的算力成本，促进模型的应用落地。

- **增强数据支持：**提高数据质量可以减少无效计算，提高模型训练和推理的效率，提升整体计算性能和结果准确性。可通过建立高质量的数据集，并构建统一的数据存储和访问接口，简化数据流动与共享，为人工智能模型训练提供强有力的支持。此外，基于先进的加密技术和访问控制机制等技术，可以保护数据和模型免受未授权访问或攻击，提高数据安全。

以浪潮信息源2.0-M32为例，其创新性地提出和采用了“基于注意力机制的门控网络”技术，构建包含32个专家（Expert）的混合专家模型（MoE），大幅提升模型算力效率，显著降低在模型训练、微调和推理所需的算力开销，单Token下训练和推理所需的算力资源仅为Llama-70B的1/19，显著提升效率。

2024年12月，深度求索（DeepSeek）发布了拥有6710亿参数的DeepSeek-V3模型，不到一个月后，发布了通过强化学习优化的DeepSeek-R1推理模型，在英语、代码、数学等多个基准测试中表现优异，迅速登上Hugging Face排行榜榜首。DeepSeek模型采用专家混合（MoE）架构，允许模型在保持高表达能力的前提下，大幅减少计算量和内存占用，提高训练效率和推理速度，使人工智能技术更加经济高效，便于广泛应用，公开信息显示DeepSeek-V3模型的开发时间仅为两个月，开发成本不到600万美元。DeepSeek-R1 API服务定价为每百万输入tokens 1元（缓存命中）/4元（缓存未命中），每百万输出tokens16元，大约是OpenAI o1运行成本的3%。

PART

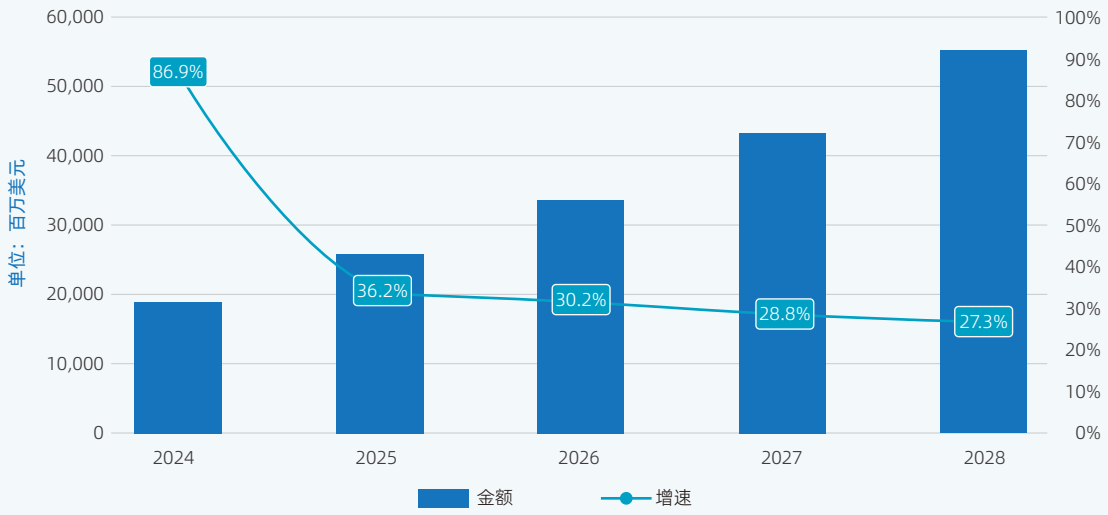
第二章

人工智能算力及应用

2.1 芯片和服务器：向高性能与高效能方向演进，重视开放多元体系建设

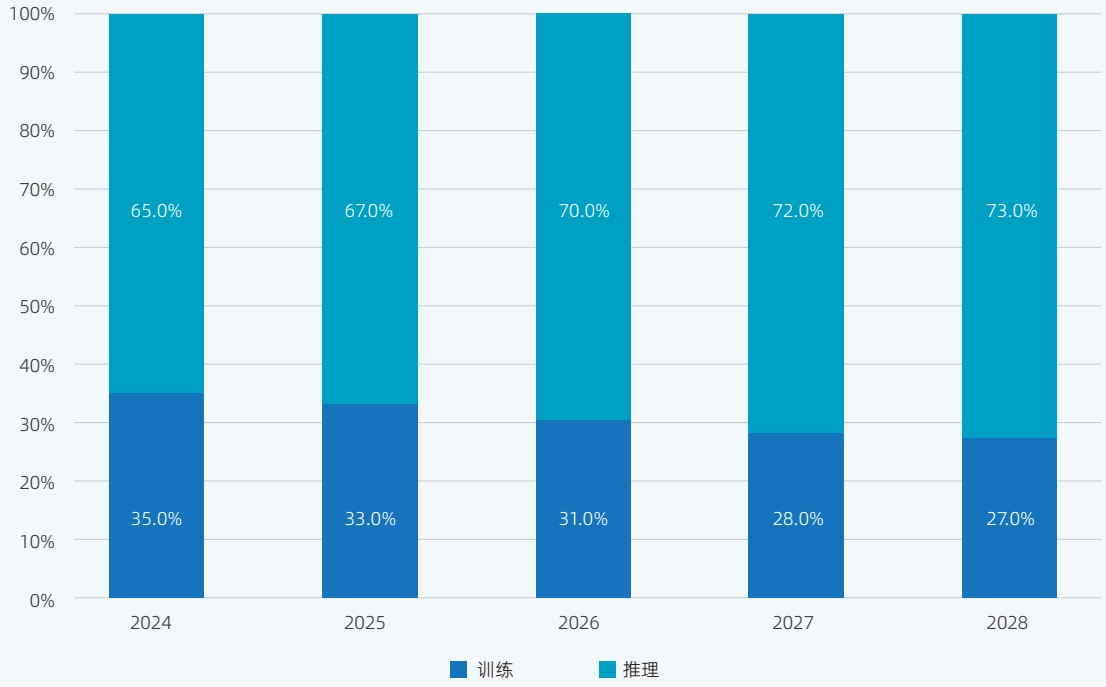
大模型兴起和生成式人工智能应用显著提升了对高性能计算资源的需求，人工智能服务器作为支撑这些复杂人工智能应用的核心基础设施，市场规模也持续扩大。根据IDC报告，2024年中国人工智能算力市场规模达到190亿美元，2025年将达到259亿美元，同比增长36.2%，2028年将达到552亿美元。随着模型的成熟以及生成式人工智能应用的不断拓展，推理场景的需求日益增加，推理服务器的占比将显著提高。IDC数据显示，预计到2028年，推理工作负载占比将达到73%。

图7 中国人工智能服务器市场预测，2024-2028



来源：IDC，2025

图8 中国人工智能服务器工作负载预测，2024-2028



来源：IDC，2025

随着大模型训练和推理任务的复杂性和规模不断增加，人工智能服务器朝着更高性能和更高效能方向演进。人工智能服务器在算力提升、功耗优化和硬件加速等领域实现突破，为大规模数据处理和执行深度学习等复杂计算任务提供了强有力的支撑。人工智能服务器不断集成更多的智能管理和优化功能。例如，通过引入智能调度和资源管理系统，可以动态分配计算资源，优化任务执行效率，减少资源浪费。这些功能的提升，使得人工智能服务器能够更好地支持大规模并行计算和实时数据处理，满足大模型训练和推理的高性能需求。

人工智能服务器的生态系统建设也将成为市场发展的关键因素，硬件与软件的协同作用将进一步提升算力效率和应用场景的适应性。企业将在技术研发、定制化解决方案和生态建设上持续投入，以满足日益复杂的行业需求。

人工智能芯片作为算力产业的关键基础设施要素，呈现多元化发展趋势。多元化的人工智能芯片可针对不同的应用场景进行优化设计，例如，针对深度学习训练的GPU和TPU，能够提供大规模矩阵运算的高效支持；而针对推理任务的ASIC，则在功耗和延迟方面表现出色，适合部署在边缘计算设备中。此外，人工智能芯片在技术创新、生态系统建设和产业发展等多个方面展现出蓬勃发展的态势。在技术方面，中国人工智能芯片针对深度学习优化架构设计，推进架构创新，并重视软硬件协同优化，通过编译器、运行时环境、开发工具链等一系列配套软件的支持加速硬件潜能的发挥。在生态系统建设方面，诸多厂商推出开放平台，提供丰富的API接口和预训练模型库，降低使用门槛，并构建云-边-端一体化解决方案，形成完整的计算链条；活跃的技术社区加速知识共享和技术交流，帮助厂商改进产品和服务；此外，人工智能芯片企业还与其他行业领导者建立战略合作伙伴关系，共同探索新的应用场景和服务模式。未来全球人工智能芯片市场将持续扩大，中国作为重要市场之一，也将迎来高速增长期。政府出台一系列促进产业发展的扶持政策，包括资金投入、税收优惠和知识产权保护等。在大力政策支持下，芯片厂商正在加速新技术研发，推进产业链上下游联动，推动人工智能芯片技术的突破和广泛应用。

总的来说，中国推动算力产业的转型与市场发展，通过人工智能服务器和人工智能芯片的双轨并行发展，正在构建一个高性能、多元化和高效的算力基础设施体系。这一体系不仅能够满足当前大模型训练和推理所需的爆发式增长的算力需求，还为未来人工智能技术的创新和应用提供了坚实的基础。随着算力产业的不断发展，中国在全球人工智能领域的竞争力也将进一步提升，推动智能化社会的全面到来。

2.2 存储和网络：分布式存储与全闪存提升性能，先进网络架构优化数据访问速度

随着人工智能技术的飞速发展，企业对数据存储的需求达到了前所未有的高度。IDC数据显示，2024年全球企业在人工智能存储领域的支出达到67亿美元，2025年将增至76亿美元，2028年有望达到102亿美元，2023-2028年五年年复合增长率为12.2%。

大模型训练和生成式人工智能技术的应用对存储市场产生了显著影响。首先，算力的增强推动了对存储性能需求的升级。存储系统需要具备更大的容量、更快的读写速度、更低的延迟、更高的可靠性和更佳的灵活性，以支持高效的数据处理与模型训练，同时适应不断增长的数据量和扩大模型规模的需求。在这个过程中，分布式存储架构凭借其性能线性扩展的优势，成为训练场景的主要选择。随着算力集群规模扩大，存储带宽需相应提升，传统集中式存储和串行运算模式已难以满足需求。分布式存储与并行运算的结合，为人工智能领域提供了新的解决方案。分布式存储系统通过将数据分散在多个物理节点上，提供冗余备份、无限扩展性和并行访问，提升数据可靠性和容错性，系统可以迅速从其他节点恢复数据，确保应用的连续运行；支持PB级甚至EB级数据扩展，满足人工智能模型对海量数据的需求；支持并行访问，多个节点可以同时读取和写入数据，降低数据访问延迟。

其次，存储系统需要能够灵活应对复杂的数据治理和应用需求。当训练出的大模型应用于具体行业时，需要将行业内的各种数据与大模型结合，而这些海量数据来源广泛、类型多样且分布在不同地点，企业需要高效完成这些数据的汇集和预处理。除分布式存储架构外，多协议支持将成为关键。存储系统需要支持多种数据访问协议以及具备无损互通非结构化数据的能力，兼容不同数据格式和传输方式，从而提高数据管理的效率和灵活性，满足复杂应用场景的需求。

未来，随着推理工作负载的增加，存储系统将更加注重快速读写数据、实时响应推理任务、支持大量并发访问以及确保数据的高可用性和完整性。通过采用NVMe SSD等高性能存储介质，结合数据缓存和预取机制，设计灵活的扩展方案，并实施智能数据管理和优化策略，可以帮助企业提高存储利用率和性能，优化数据治理，加速数据价值挖掘。

全闪存存储方案凭借其卓越的数据传输速度、更低的能耗以及更高的单位物理空间容量，在人工智能市场中展现出强劲的增长势头。特别是QLC SSD（四层单元固态硬盘），预计将在以读取为中心的工作负载及应用程序中，以及缓存和分层架构的企业SSD市场中占据越来越重要的地位。PCIe接口协议正逐步占据更大的市场份额，代表下一代高性能网络技术的NVMe over Fabric（NVMe-oF）也开始崭露头角。NVMe-oF能够提供服务器与网络存储间极低的访问延迟，从而充分发挥NVMe SSD的优势并加速其普及。

生成式人工智能和大模型的发展带来了计算集群规模的提升，从万卡扩展到十万卡，大模型在训练和推理过程中会生成大量的临时数据，这些数据需要在不同计算节点之间快速传输和处理：基于远程直接内存访问（RDMA）技术可实现数据在设备间的直接传输，而不用经过传统网络协议，从而提高传输速度；在数据中心层面，400G以太网技术已经开始在新型数据中心得以采用。面对不断增长的数据传输需求，需要全面提升网络运力效率，构建面向人工智能的运力体系，该体系应具备如下特征：

- **高带宽：**高带宽能够显著提升数据传输速度，减少训练和推理时间，提高整体效率。目前网络速率已经可达到400G/800G，1.6T是超大规模数据传输和高效能需求的下一步计划，未来行业目光将投向3.2T乃至更高速率，以应对数据中心内部服务器之间和跨数据中心场景下的数据互联。

- **低延迟：**低延迟对于实时数据处理和快速响应至关重要，能够显著提升系统的性能和用户体验。RoCEv2的出现，使AIGC集群可以基于RDMA技术扩展到超大规模，较传统以太网的通信方式大大降低了延迟。

- **高可靠：**高可靠性能够保证数据传输的稳定性和连续性，避免因网络故障或升级导致数据丢失和服务中断。从传输层的冗余路由和智能流量控制，到链路层的链路聚合与自动故障切换，再到物理层的硬件冗余设计和环境监控防护，通过多层次的设计和技术手段，确保网络系统和数据传输在任何情况下都能稳定、连续地运行。

- **端网协同：**在大规模人工智能计算中，端网协同至关重要，能够显著提升数据传输的效率和可靠性，减少数据传输过程中的延迟和丢包率。新一代以太网技术的优化将不再局限于单一的网络设备或组件，而是覆盖数据端到端流程，从交换设备到网卡，再到软件端的全面优化。通过诸如智能网卡等技术的应用，减轻主处理器的负担，提高整体系统性能。

- **负载均衡：**随着网络规模的扩大，传统的路由技术已难以满足需求。新的路由技术需要引入先进的负载均衡算法，确保数据流在网络中的均匀分布，避免某些节点过载。同时，这些算法还需具备高效的路径选择能力，以最优路径传输数据，减少延迟。

- **拥塞控制：**在大规模网络中，拥塞是一个常见现象。拥塞控制技术通过监测网络流量，动态调整数据传输速率。先进的拥塞控制算法能够快速响应网络状态变化，避免因网络拥塞导致的数据传输延迟和丢包，确保数据传输的稳定性和高效性。

- **全网可视：**网络可视化技术提供了对网络状态的实时监控和管理能力。通过可视化工具，网络管理员可以直观地了解网络拓扑、流量分布和性能指标，及时发现和解决潜在问题，优化网络性能。

这些升级不仅提升了网络性能，还增强了系统的灵活性和可靠性，为人工智能技术的发展提供了坚实的基础。

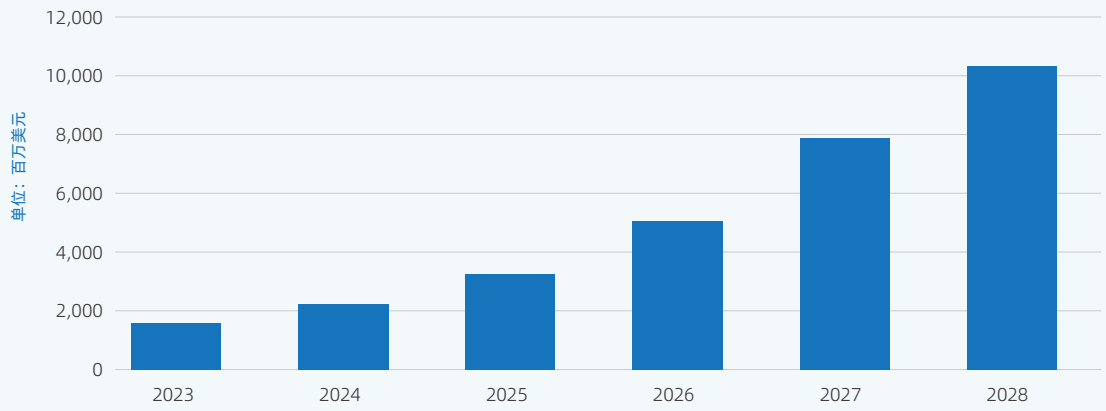
2.3 可持续数据中心：液冷技术成为关注重点，聚焦智能算力散热革命

数据中心作为现代信息技术的基础设施，其能源消耗问题日益受到关注。随着人工智能、大数据和云计算等技术的快速发展，数据中心的能耗不断增加，对环境和经济都带来了巨大压力，可持续性发展成为行业关注的焦点。

为了解决人工智能工作负载带来的功耗和热挑战，业界积极探索多种形式的技术手段，共同研发更高效的能源管理和数据冷却技术，推动算力设施在计算节点应用如冷板、相变浸没液冷服务器、处理器垂直供电、超高转化率大功率氮化镓（GaN）电源等解决方案，提升算、电综合利用率；在机柜层面采用超高密度液冷机架实现集中部署，并融入创新的基于人工智能的智能控制技术，提高运维效率，通过落实“三总线盲插技术”、毫秒级多重漏液监测、负压式CDU等手段保障机柜运维安全；此外，数据中心积极引入可再生能源、智能能效管理系统实现绿色化、低碳化转型，加强电力与算力在绿色技术创新方面的合作。

液冷技术的发展是实现这一转型的重要技术突破，它可以显著降低数据中心的总能耗，提高计算密度，加速训练和推理的速度，提高算力利用率。该技术可将数据中心PUE值降至1.1以下，并减少了散热设备所需的空间，这不仅节省了空间资源，还在相同面积内实现了更多的服务器配置，极大提高了数据中心的效率和性能。此外，随着技术的不断进步，液冷技术还将与数据中心的其他技术相结合，如人工智能、物联网等，实现更加智能化、自动化的散热管理，推动数据中心向更加绿色、可持续的方向发展。IDC预计，2028年中国液冷服务器市场将达到105亿美元，2023-2028年五年年复合增长率将达到48.3%。

图9 中国液冷服务器市场规模及预测，2023-2028

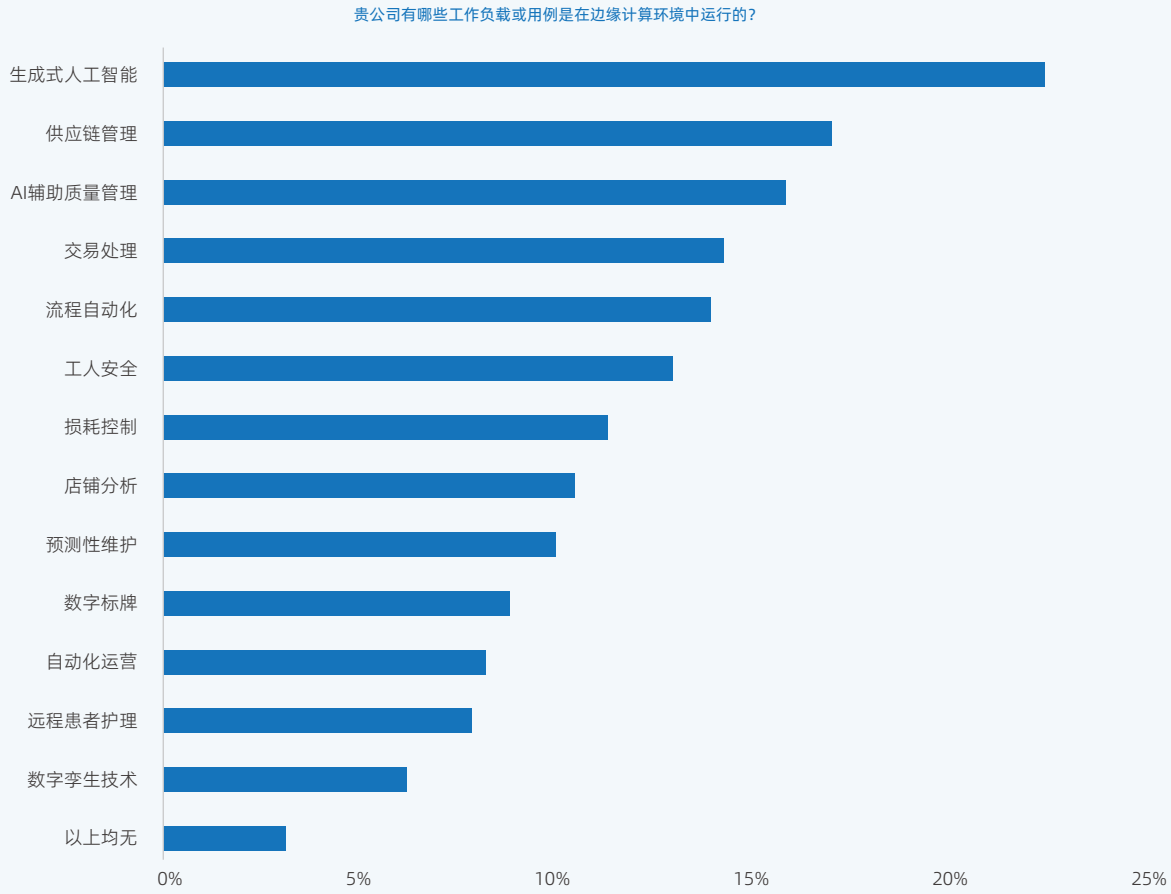


来源：IDC，2025

2.4 边缘计算：大模型的部署向边缘迁移，智慧边缘加速模型推理

边缘计算将在更广泛的IT战略中发挥关键作用。人工智能将逐步向边缘侧或端侧倾斜，未来企业级大模型有可能将越做越小，成为可搭载于边缘侧的计算设备，从而促进大模型在各种边缘场景下发挥更大的价值。IDC研究显示，生成式人工智能正迅速成为企业在边缘计算环境中最广泛应用的工作负载。

图10 全球边缘基础设施支持广泛的数字业务和运营计划



来源：IDC，2025

随着人工智能技术的日益成熟，其应用场景也在不断拓展。然而，传统的云端人工智能处理模式在面临海量数据、实时性要求高以及网络带宽限制等挑战时，显得力不从心。因此，人工智能向边缘侧或端侧的迁移成为了一种必然趋势。边缘人工智能通过直接在数据源头进行处理和分析，不仅减少了数据传输的延迟和成本，还有效保障了数据隐私和安全，为智能应用的广泛部署提供了可能。

在这一背景下，企业级大模型的设计和应用也迎来了新的变革。为了适应边缘侧的计算环境和资源限制，大模型正朝着更加精简、高效的方向发展。通过模型压缩、剪枝、量化等技术手段，大模型能够在保持较高精度和性能的同时，显著减小模型体积和计算复杂度，从而轻松搭载于边缘侧的计算设备。这一变化不仅降低了部署成本，还提高了模型的灵活性和可扩展性，为边缘智能应用的快速发展奠定了坚实基础。

边缘计算通过本地化数据处理、减少网络延迟、保障数据隐私、优化资源分配和增强系统弹性等多方面的优势，显著提高了大模型的算力效率和实用性。首先，本地化数据处理使得大模型能够直接在数据源头进行实时分析，减少了数据传输的延迟和成本；其次，通过减少网络延迟，边缘计算确保了智能应用的即时响应和高效运行；同时，数据隐私的保障也为企业级应用提供了更加安全、可靠的环境；此外，资源分配的智能化优化和系统弹性的增强，使得边缘计算能够根据实际情况动态调整资源分配，最大化利用资源，确保大模型在各种场景下都能发挥出最佳性能。

2.5 算法和模型：算法创新与模型迭代解锁更高算力利用率，实现卓越性能与效率

算法是驱动是人工智能发展的核心引擎，决定了应用的智能上限，也牵引着算力的发展。2024年，o系列、Llama3、通义千问、R1等大模型不断升级，尤其是DeepSeek R1系列模型的发布，正是基于算法层面的极大创新，对中国乃至全球的人工智能产业带来深刻变革。一方面DeepSeek采用了大规模强化学习、多头注意力机制等算法创新，智能水平在美国高中数学竞赛邀请赛AIME、博士水平科学问答等测试中榜单上接近甚至超过了OpenAI的o1模型；另一方面，DeepSeek R1算法的创新也带来训练和推理阶段算力消耗的降低，训练算力只有Llama3的1/10，推理阶段缓存数据量降低了50倍，在算力约束的条件下进行AI算法创新提供了一个全新思路，吸引了全球开发者，7天实现了活跃用户数破亿。

多模共存，行业和企业落地成为重点。中国大模型发展呈现多模型共存的繁荣景象，市场上既有通用型大模型，也有针对特定行业或应用的专业模型。多样化生态不仅满足了不同应用场景的需求，也促进了技术进步和市场细分。随之而来的则是大模型的场景化落地需求快速增长，企业普遍希望能够根据行业数据或是内部数据对于模型进行微调，以提高模型的针对性和有效性，增强数据安全与隐私保护，提升模型在特定应用场景中的性能，从而更好释放大模型的价值。

提高模型架构效率与增加原始计算能力同样重要。杰文斯悖论指出当技术进步提高了资源利用效率时，该资源的总消耗量反而可能会增加，这一规模同样适用于人工智能领域。**更经济的先进模型将推动AI需求，增加效率——而不仅仅是原始计算能力——可能是真正的竞争优势，更多公司致力于开发在不要求过多计算资源的情况下平衡性能和效率的模型。**大模型的压缩、量化、蒸馏技术可以在不牺牲模型精度的前提下，显著提升推理吞吐量，减少内存、算力等关键资源的消耗，提高模算效率，降低计算成本。合适的压缩剪枝技术可减少模型参数数量50%以上，而模型性能损失很小。对于特定应用场景，采用专用的硬件与软件进行加速，能够实现高效加速并提高大模型在训练和推理的效率。同时，这些优化措施也提升了企业大模型的投资回报率，加速人工智能技术的应用和商业化进程。

端侧大模型和人工智能推理蓬勃发展。端侧模型以其较低的参数量和高效的计算能力，使得在资源受限的人工智能PC、智能手机等设备上运行复杂的人工智能模型成为可能。这些模型不仅能够处理文本，还在图像和语音识别等多模态领域展现了巨大的潜力，同时依托人工智能推理的发展，可将应用拓展到即时消息生成、实时翻译、会议摘要、医疗咨询、科研支持以及自动驾驶等场景，成为未来智能设备的核心，为用户带来更加智能、便捷和个性化的体验。

多模态大模型依旧是当前大模型训练和开发的重要方向。随着技术的快速演进，大模型正在从文本类、图片类等单模态向多模态、跨模态演进。多模态大模型是提高模型性能和泛化能力的一种有效方式，将成为未来发展重点之一。其能够同时综合处理和分析文本、图像、音频等各种模态的信息，有助于克服单一模态数据的局限和偏见，并将其融合以完成复杂的实际任务。在自动驾驶、内容推荐、虚拟现实和增强现实、远程医疗、虚拟助手、虚拟实验室、媒体和智能家居等场景中，多模态大模型能够更好地理解实际场景，输出更加准确的结果，展现出巨大的应用潜力和社会价值。

大模型的开源趋势正在显著增强。开源模型成为推动人工智能技术普及和应用落地的重要力量。在过去的18个月里，全球领先的软件和云服务商发布了数十种开放和部分开放的基础模型，开源社区的协作和贡献正在成为加速技术创新的重要力量，开源框架作为人工智能开发的基础，其生态系统日益丰富。开源不仅促进了生态繁荣，还降低了研发成本，使得中小企业和研究机构能够专注于应用开发和优化。以DeepSeek为例，企业可以访问一个性能水平相当的开源模型，并大幅降低了训练和部署这些模型的成本。这对于资源有限的中小企业和个人开发者来说尤为重要，使得更多组织能够负担得起先进人工智能技术的应用，加速了人工智能技术的普及。这种开放共享的方式加速了技术迭代，激发了更多的合作与创新，为整个行业的快速发展注入了新的活力。IDC预测，2025年，为了更快获得创新能力、运营主权、透明度和更低成本，将有55%的企业使用开源人工智能基础模型开发应用程序。

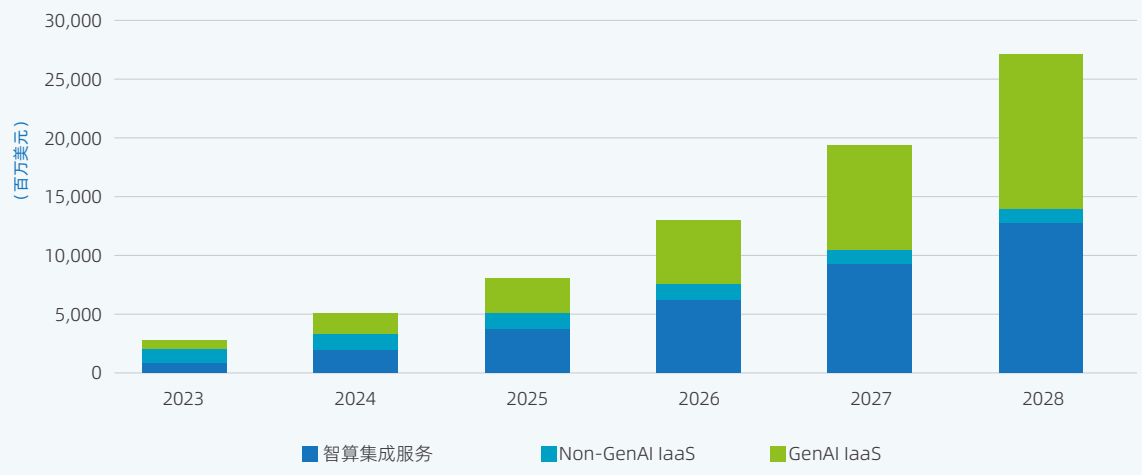
2.6 人工智能算力服务：构建全栈服务体系，加速大模型应用落地

随着科技的飞速发展，企业对于智能算力的基础设施和服务能力的要求正在发生深刻变化，传统算力技术架构和云服务模式难以满足新的需求，生成式人工智能将使企业更多使用人工智能就绪数据中心托管设施和生成式人工智能服务器集群，从而缩短部署时间，降低长期设施的资本成本。

这些变化尽管挑战了传统算力服务的既有优势，但也为算力服务市场带来了新的机遇与变革。这意味着云服务商除了需要不断创新，提升技术水平和服务质量，还须开辟更加开放、灵活的赛道，适应市场的快速发展。在当前技术体系下，人工智能服务器的高昂投资成本已成为行业面临的显著挑战，亟需新的合作机制来重新分配资源与市场。在此过程中，众多技术和资本纷纷加入人工智能算力服务的竞争。在人工智能时代，市场中的参与者需要重新考虑如何分配资源、降低成本并共享收益。这促使企业间寻求合作，共同分担投资压力，共享市场成果，形成数据中心服务商、云服务商、硬件厂商以及其他创新企业共同参与的生态体系，为用户提供人工智能算力资源，并通过技术创新和服务优化，满足用户多样化的算力服务需求，通过资源池化、动态分配和智能调度等技术手段，突破传统算力供给模式的局限性，提高资源利用率和灵活性。

智算服务市场正在快速发展。IDC定义下的智算服务是指以GPU、ASIC等人工智能专用算力为主的基础设施服务，主要包括智算集成服务和智算基础设施即服务（AI Infrastructure as a service, 简称AI IaaS）。智算集成服务主要是指厂商在帮助客户建设私有智算基础设施过程中提供的咨询、集成、开发、运维等专业和管理服务；AI IaaS是指供应商以租赁形式为客户提供一站式智能算力服务，并由供应商提供后续的运营及运维保障。其中AI IaaS市场又分为面向生成式人工智能（简称GenAI IaaS）和非生成式人工智能（简称Non-GenAI IaaS，如传统渲染、仿真、视联网推理等业务场景）两个细分市场。IDC数据显示，2024年中国智算服务市场整体规模达到50亿美元，2025年将增至79.5亿美元，2028年将达到266.9亿美元，2023-2028年五年年复合增长率为57.3%。其中，智算集成服务市场及GenAI IaaS市场是未来重要的两个增量市场，五年年复合增长率分别达到73%和79.8%，预计至2028年智算集成服务市场规模占比可达47%，GenAI IaaS市场规模占比达48%。

图11 中国智算服务细分市场规模预测，2023-2028



来源：IDC，2025

随着算力基础设施的搭建完成，市场和用户对资源利用率的关注度将会显著提升，运营和应用成为核心焦点。高效的建设、管理和运维能力对于智算服务的成功至关重要，能够确保资源的最优配置和高效利用。此外，智算服务商在这一过程中积累的成功运作经验也变得尤为重要，这些经验不仅能够提升服务质量，还能为用户提供更具竞争力的解决方案。因此，具备强大建管运能力和丰富运作经验的厂商将在新型算力服务市场中占据优势地位，推动行业的持续发展和创新。

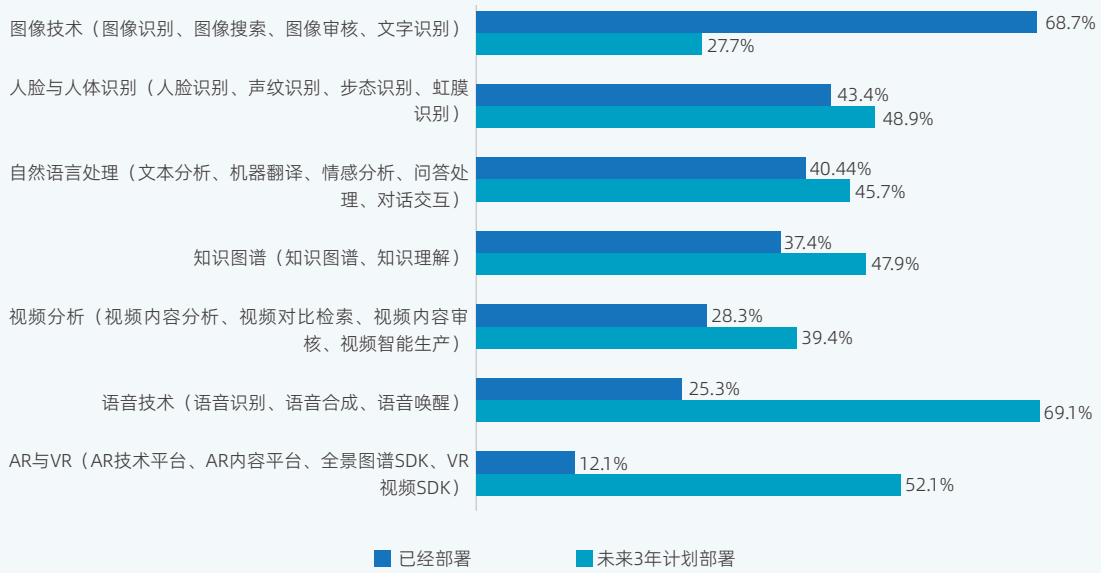
2.7 应用：积极探索人工智能应用场景，加速智能对于业务发展的价值转化

全球人工智能市场规模持续攀升，在这个千亿级别的市场下，各个细分领域都取得了显著的进展，其中生成式人工智能市场是主要的增长细分领域，企业支出的五年（2023-2028年）年复合增长率高达59.2%。从行业支出角度来看，过去一年，全球生成式人工智能IT基础设施的投入主要集中在互联网、运营商和IT服务这三大行业里，但从未来预测数据来看，这三大行业的占比将逐渐减少，进而转向政府、金融、制造、教育和医疗等行业。这意味着生成式人工智能将逐渐从大模型训练向推理迈进，尽管这将是一个长期的过程，但前景依然被看好。IDC认为，向“无处不在的人工智能”的过渡将在个人层面、业务或职能部门层面以及特定行业背景下出现一系列新的由人工智能驱动的应用场景。人工智能发展已进入一个全新的阶段。

人工智能技术应用及算力需求

在人工智能单点技术应用方面，IDC调研显示，图像技术成为当下最主要的应用技术类型，人脸与人体识别紧随其后，自然语言处理位列第三。

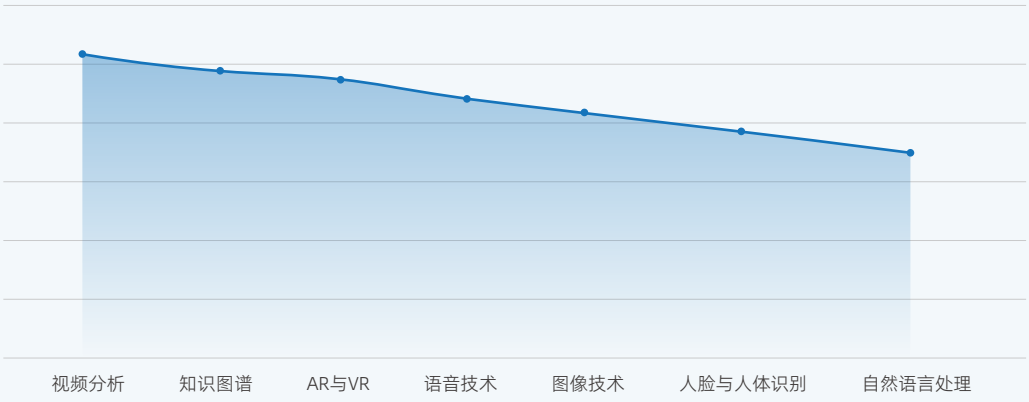
图12 企业已部署及未来三年计划部署的人工智能单点技术



来源：IDC，2025

根据单点技术应用及企业IT资源消耗调研结果显示，视频分析成为占用企业IT资源最多的工作负载。这归因于其需处理海量图像数据，执行复杂计算与实时处理，对计算和存储资源要求极高。除此之外，知识图谱和AR/VR以及图像技术是参与调研样本企业中另外三个高算力消耗的单点技术。知识图谱涉及大量关系数据处理和复杂查询；AR/VR需要高性能图形处理与实时渲染；图像技术则涵盖广泛的图像处理和分析任务。这些应用均对企业IT资源提出了严峻挑战，推动了对高效能计算和存储解决方案的需求。

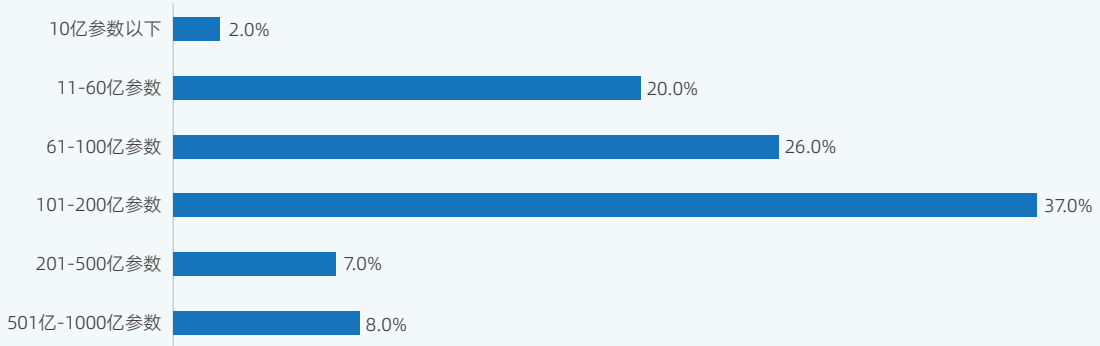
图13 单点技术场景对IT资源的占用情况



来源：IDC，2025

随着生成式人工智能和大模型的广泛应用，模型参数量的增加对算力提出了更高的要求。目前企业使用的生成式人工智能模型的平均参数量分布如下：37%的模型拥有100-200亿参数，26%的模型参数量在60-100亿之间，而8%的模型参数量超过500亿。生成式人工智能模型参数量的增加对算力的需求显著提升，企业需要在硬件配置、软件框架和优化技术等方面进行综合考虑，以确保模型的高效训练和推理。

图14 目前使用的生成式人工智能模型平均参数量

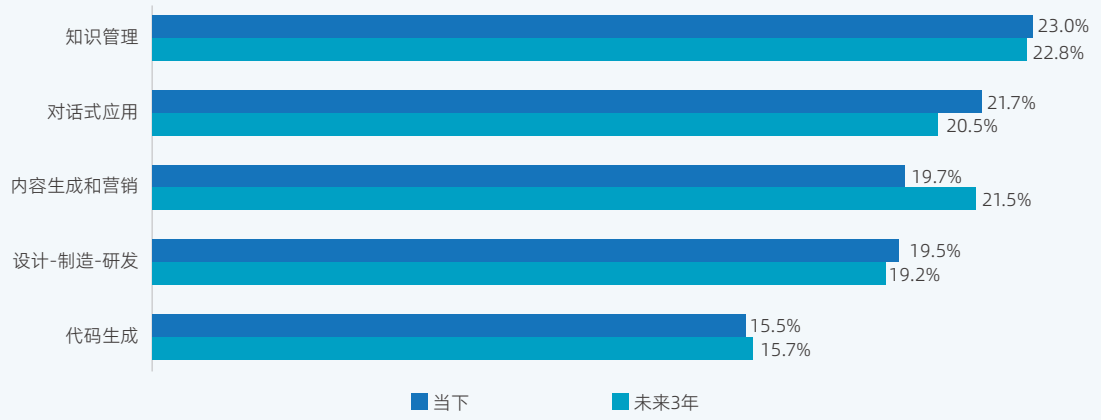


来源：IDC，2025

生成式人工智能用例分析

IDC调研结果显示，知识管理、对话式应用、内容生成和营销是企业当下应用生成式人工智能最具发展潜力的领域。这些应用通过分析和生成数据、图像、文字、情感及代码等内容，为企业带来新的增长机会。

图15 当下和未来三年企业应用生成式人工智能最具潜力的场景



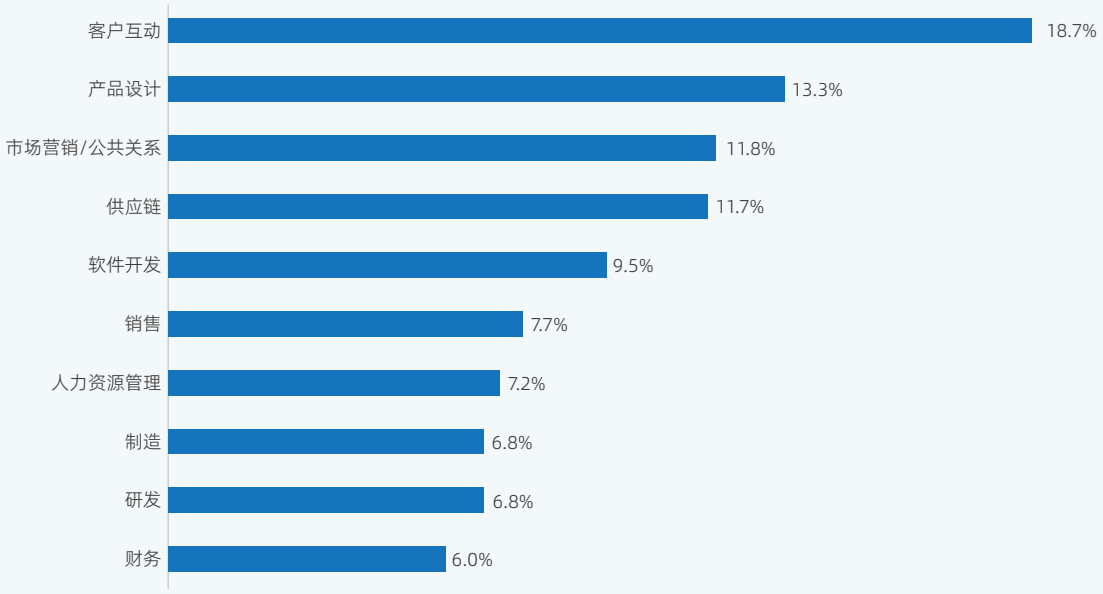
来源：IDC，2025

生成式人工智能用例有三种主要类型，包括生产力类应用、业务部门或职能部门类应用，以及行业场景应用：

- **生产力类应用**是针对具体工作任务设计的应用，用于提升劳动生产率和优化业务流程。通过自动化复杂任务（如设计生成、内容总结和生成、代码生成等），生成式人工智能使个人能够专注于高价值活动。此外，生成式人工智能可以高效处理大量数据，生成适当的响应和建议，从而优化生产流程和工作流，提高运营效率。越来越多的生成式人工智能功能正在被集成到现有的很多应用程序中，例如微软360 Copilot或谷歌的Duet，从而为企业和个人带来更高的生产率和更好的用户体验。

- **业务部门或职能部门类应用**倾向于将一个模型或多个模型与企业数据集成，供特定部门如市场营销、销售、采购等使用。这些业务功能用例需要与成熟企业应用和平台集成，其能力受到客户数据、产品数据、知识库等业务数据的约束。IDC调研结果显示，未来18个月，企业优先应用生成式人工智能的业务部门或职能部门场景依次为：客户互动，产品设计和市场营销。目前客户互动集中应用在智能客服场景，已经在金融、电商、通信、医疗和教育等多个行业的智能工作流程中广泛适用。通过优化服务流程和提高响应效率，生成式人工智能不仅增强了客户满意度，还在降低人力资源成本方面取得了显著成效。生成式人工智能在产品开发与设计方面的应用主要集中于文生图、图生视频、文生视频领域，并形成了较为稳定的业务流。其中，文生图与图生视频由于更加稳定、可预测的表现，成为较为主流的应用。

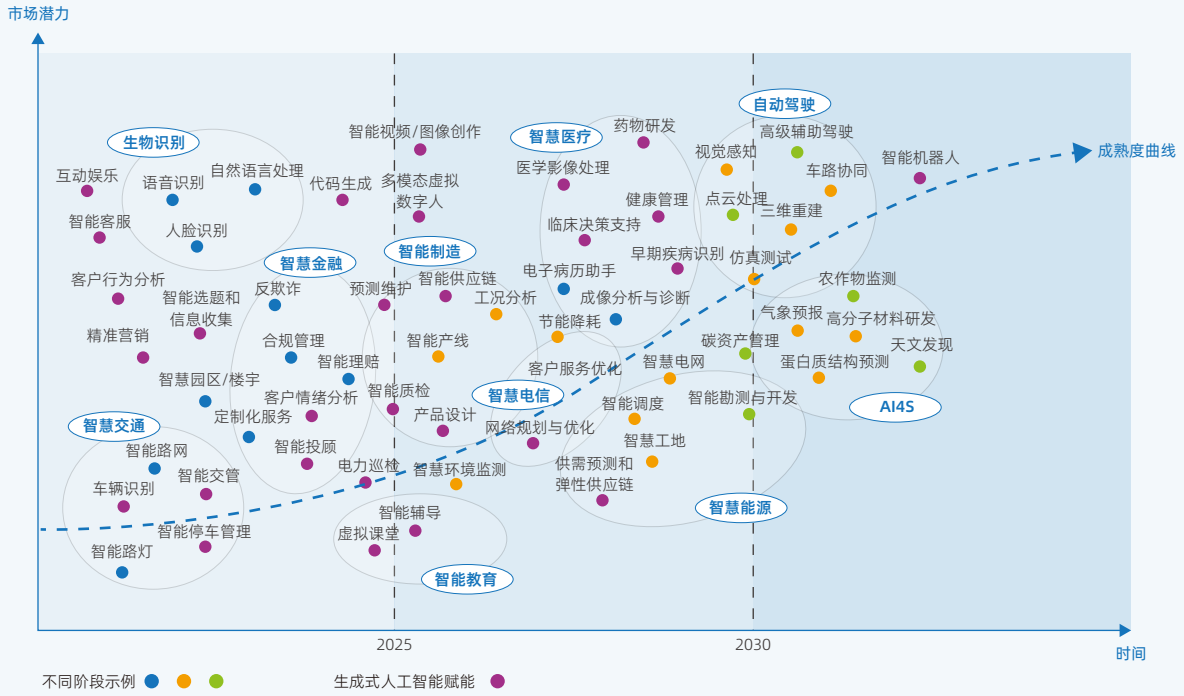
图16 未来18个月企业应用生成式人工智能最具潜力业务领域/部门



来源：IDC，2025

- **行业场景应用**是生成式人工智能在各个行业中的特殊场景应用，这些特定的应用场景通常需要高度的定制化才能实现。在垂直领域内利用特定数据、针对具体场景优化模型以及提供工程化解决方案，是实现生成式人工智能实际应用的基础，同时也是构建企业竞争优势的核心。

图17 中国人工智能应用场景发展



来源：IDC，2025

与往年相比，技术创新成果加速了行业应用的进展，人工智能在自动驾驶、制造、金融、城市建设、科研教育、能源等多个领域展现了突出的应用效果和潜力：

- 技术进步大幅提升车辆感知和决策能力，使得自动驾驶的安全性和易操作性达到了新的高度，推动了该领域超越预期的快速发展；
- 在制造业中，人工智能技术持续优化生产流程，降低企业成本，提高制造产品质量；
- 在金融领域，人工智能技术在风险管理、欺诈检测和个性化金融服务等场景中发挥重要作用；
- 在智慧城市建设过程中，人工智能改进了交通管理和环境监控方式，提升治理效率；
- 科研方面，人工智能助力数据分析和生成，支持个性化教学和智能辅导系统的发展；
- 在能源领域，人工智能优化能源生产和分配过程，提高能源利用效率，推动智慧能源体系建设。

此外，生成式人工智能在行业中的渗透范围进一步扩大，在视频/图片创作、多模态数字人、代码生成、互动娱乐、营销等场景中，实现泛行业式的广泛应用。

就行业特殊场景而言：

- 在自动驾驶领域，生成式人工智能为用户提供更丰富的驾驶体验，实现在智能座舱、车辆状态检测、环境检测、故障诊断、路径规划和演练数据合成等场景的应用，并支持企业内部汽车研发工作，覆盖汽车设计、代码开发和内部知识库建立等场景，提高工作效率；
- 在金融行业，生成式人工智能被广泛应用于客服、数字人助手、风险审核、贷款审核、电子营销和财富管理等场景，提升了服务效率和客户体验；
- 制造行业尝试使用大模型与生成式人工智能能力辅助产品设计、供应链管理、智能客服，此外通过构建专家大脑加速智慧工厂建设；
- 在医疗领域，生成式人工智能支持远程手术，覆盖门诊、医务科室、设备管理和人员管理等多个场景，提升医疗服务质量和管理效率。

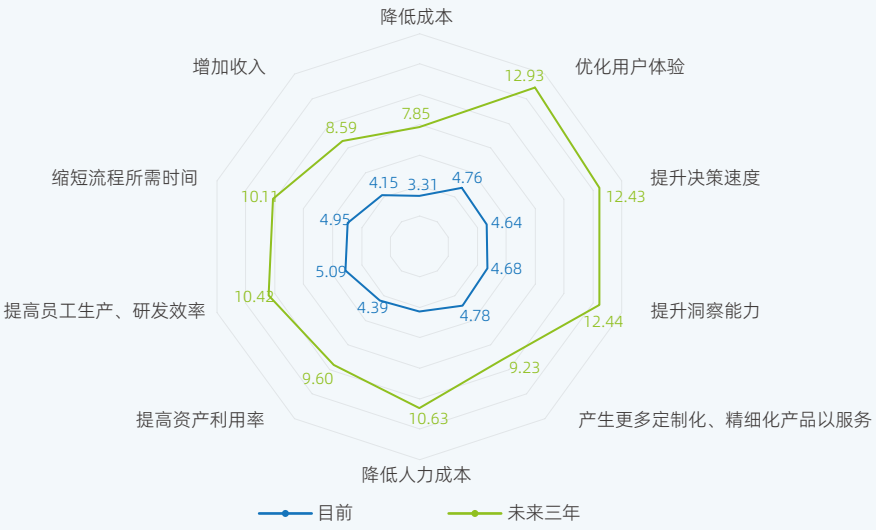
同时生成式人工智能与传统人工智能形成了互补共存的状态，在有着明确规则、需要广泛提炼特征并进行分类与预测的场景中，传统人工智能仍继续扮演着不可或缺的角色。

与此同时，具身智能机器人的发展引发市场关注，由于其可以在动态环境中进行实时决策和行动，提高任务执行的效率和效果，正在成为机器人未来发展的重要方向，预计将在工业、医疗、服务等领域具有广泛的应用前景。

人工智能应用价值分析

IDC调研显示，目前除了提升生产和研发效率外，人工智能带来的主要业务变化和重要业务成果包括：缩短流程时间、创造新的产品与服务，以及优化用户体验。未来，企业对人工智能的应用将在提升企业洞察、助力决策等维度提供更显著价值。

图18 目前及未来三年应用人工智能对企业带来的价值



来源：IDC，2025

最佳实践

招商银行：深入推进企业数智化发展，打造数字员工助力提质增效

招商银行由招商局于1987年在深圳蛇口创建，是中国境内第一家完全由企业法人持股的股份制商业银行。成立37年来，招商银行已成为拥有商业银行、金融租赁、基金管理、人寿保险、境外投行、消费金融、理财子公司等金融牌照的银行集团。

挑战及解决方案

随着招行数字金融科技的发展和业务的不断扩张，如何提升工作效率、优化资源配置，成为了一大挑战。为实现业务流程超级自动化，提高工作准确性、生产效率和降低运营成本，支持银行线上化、数字化、智能化发展，招商银行引入RPA（机器人流程自动化）技术，打造海螺RPA+平台，构建灵活、好用、覆盖面强的平台和工具。

招商银行海螺RPA+是集成RPA、AI、Open API和大模型的自动化解决方案，拥有业务流程录制、封装编码等功能，可提供可视化设计，具备AI扩展性、服务化、智能资源匹配、监控预警和操作回溯等能力，可替代重复人工操作，实现降本增效控风险。

招商银海螺RPA+平台的构建为员工带来了全新的工作模式，实现文化的传承、数字化的扩容和空间的优化：

- **“数字员工”海螺RPA+：**将人工智能和机器人流程自动化技术相结合，实现RPA技术向“数字员工”的转变，具备处理常规性、流程化、重复度高的工作内容的能力。基于此，员工可以把更多精力放在创造更有价值的新内容上，从而提升整体工作效能。

- **生态建设与人才培养：**通过提供全面的培训支持，帮助员工快速掌握RPA+的使用方法。建设了完善的官网社区与在线沟通交流群，构建了完整的培训体系，根据员工的不同需求因材施教，以确保每个人都能够充分发挥自己的潜力，提高工作效率。

- **云+数字化多点开花：**海螺RPA+的独特之处在于它的静默画中画功能，实现了员工和“数字员工”共享一台设备同时工作。可以大大提高员工整体交付价值与设备资源利用率。此外，借助云机器人，员工可以在移动端发起工作请求，“数字员工”在云端进行远程处理，进一步提升工作效率。

项目收益

最新统计结果显示，海螺RPA+平台注册用户数已超过18,000人，活跃用户超过17,000人，平台开发者超过6,000位，场景应用超过8,900个，拥有27,000多个机器人，处理上千万条业务。通过这一数字化转型方案，招商银行有效提高了业务处理效率，降低了运营成本，并支持了其数字化和智能化的发展战略：拥有数字化能力的员工可以利用RPA技术和工具对业务流程进行梳理和优化，增强创造力，消除流程中的冗余和重复步骤，使业务流程更加高效、顺畅；引入人工智能和机器学习、大模型技术，对业务流程进行智能分析和预测，优化流程或降低错误风险，为决策人员提供更有力的依据，提高招行的运营能力，进而增强竞争优势和市场地位。

平台通过降低技术门槛、提供员工培训、促进生态建设，推动了员工的数字化技能提升，并优化了业务流程管理。

- **构建智能工作空间：**为员工提供个性化、低门槛、可视化的流程编辑器，帮助员工快速高效创建业务场景。通过智能化的任务管理和调度系统，提升数字员工资源利用率。
- **打造数字化人才：**提供数字化技能培训，满足员工数字化转型需求。鼓励员工利用数字员工辅助工作，提高效率和准确性。通过智慧虚拟助手提升客户服务和内部支持的工作效率。
- **以数据驱动企业发展：**提供实时准确的数据，为管理层提供有力支持，助力流程优化，提高流程效率，减少人为错误和重复工作。通过实时监控预防潜在风险和问题，确保流程持续合规。
- **无缝、高效：**应用数字员工消减系统鸿沟，减少业务需求与技术排期冲突。数字社区，让每位员工皆可成为数字员工的开发者和使用者。目前，使用海螺RPA进行自主开发的人员中近六成为业务人员。

西湖大学：以计算之力，为科学家助力

西湖大学是由施一公院士领衔创办的、聚焦前沿科学研究的研究型大学，成立于2018年。西湖大学的目标是成为世界一流的研究型大学，鼓励科学家们打破学科壁垒，探索人工智能与各学科交叉融合，为科研创新提速，重点发展科学、工程、医学等领域的高水平基础与应用研究。

为此，西湖大学与浪潮信息合作，打造了西湖大学人工智能集群，为校内大部分算力需求提供服务，为“科学家+AI”展现了无限可能：

- **非编码RNA研究：**非编码RNA约占人类转录组的98%，不仅参与生物体的各种基本生命过程，而且与很多重大疾病的发生密切相关。非编码RNA的数目非常庞大，而且在生命体里是高度动态的，可以跟很多其他生命分子相互交付、相互调节。面对数量庞大又动态的非编码RNA分子，如果单纯采用传统实验方式研究，需要耗费大量时间和精力也很难分析其中复杂的调控关系，找出调控规律。而人工智能技术能够高效解析测序数据，分析出其中的调控关系，找到调控规律，解决了以前单靠实验解决不了的问题，大大加速了研究进程，对于大量的疾病的治疗会带来福音。
- **演绎算法：**演绎算法是把自然的演化规则引入到人工智能领域，以解决复杂系统的优化和决策问题。近两年遵循scaling law的大模型发展火热，模型越做越大，带来的能耗问题也引发了业界的担忧。因此西湖大学的科学家希望通过演化和发育的方式，让人工智能像生物智能一样自然演化，以更低的能耗产生更高的智能。最终研发出有自主学习能力、更类人的人工智能系统，为实现通用人工智能探索出一条新路径。
- **人工智能心理咨询师“小天”：**“小天”是西湖大学研发的心理咨询大模型，基于西湖大学自研的多模态通用大模型“西湖大模型”研发。西湖大学科学家认为，“EQ让模型更有温度，更深地理解并满足人的需求。”经过大量的语料积累和真实心理咨询案例学习，加上自研的情感计算和共情模块，小天能带有感情地倾听和沟通。目前“小天”已能达到中级心理咨询师的水平。

在智算力的驱动下，科学研究正在迎来第五范式即“科学智能”（AI for Science）时代。“科学智能”不仅大幅提升了科研效率和准确性，还革新了科研范式，让人类能够挑战更复杂的难题，让很多科学创新的发现，从不可能成为可能。

长城汽车：着力发展智能化，持续提高用户体验

当前，中国汽车行业正通过智能化走向全球化。近年来，长城汽车凭借其深厚的技术积累和创新能力，成为中国汽车智能化发展的代表性企业。这一切得益于长城汽车在人工智能技术的应用、平台和硬件等方面的战略性布局和前瞻性投资。

- **应用：**长城汽车的智能座舱系统基于智能感知和流畅操作，为用户提供更好的驾驶体验。通过摄像头、传感器等设备收集数据，结合人工智能算法，识别驾驶者的情绪、疲劳状态、视线焦点等，从而提供更加个性化的服务；基于生成式人工智能语音助手，不再局限于简单的问答形式，而是能够理解上下文，进行更自然的对话；座舱软件实现办公协同、旅行规划、生成式人工智能应用、内容聚合搜索等功能，使汽车从一个单纯的交通工具转变为一个智能的移动生活空间。此外，基于自主研发的SEE端到端智驾大模型，长城汽车推出全场景NOA智驾系统，覆盖从高速公路到城市道路，从行车到泊车全场景，为用户提供安全、高效的辅助驾驶体验。

- **平台：**长城汽车智能座舱系统搭载了新一代CUX软件平台，以其卓越的硬件兼容性，适配多种车型，简化了传统一车一款一研发的复杂流程。通过CUX平台，长城汽车能够对不同车型的算力需求进行整体规划和分配，实现资源的优化配置，并根据车型发布计划制定详细的算力需求节点清单。此外，借助车联网技术，CUX平台还能实现车辆的远程监控和诊断，进一步提升用户体验和车辆管理效率。

- **硬件：**在硬件方面，长城汽车采用“本地+云端”算力相结合的方式，提高数据处理效率和驾驶安全性。在智能驾驶和智能座舱系统中采用高性能硬件设备，如车规级全固态激光雷达、高性能CPU、GPU和NPU等，为系统和座舱软件的稳定运行提供了有力保障；而智能座舱、智能语音交互、智能驾驶等先进功能，通过部署车载以太网接入云端算力，提高车辆内部通信的速度和可靠性，为智能驾驶和车联网功能提供坚实的硬件基础。

智能化的发展需要跨领域的紧密合作。长城汽车通过与科技企业、通信运营商、芯片制造商等多方合作，实现资源共享和优势互补，共同推进智能网联汽车技术的进步，以加速智能化转型和创新。目前，长城汽车与浪潮信息在算力中心构建与车载计算系统研发等领域展开深度合作，充分发挥浪潮信息在服务器系统架构、散热、电源、高速信号等关键领域的技术优势，持续升级智能座舱的硬件性能与数据处理能力。

未来，长城汽车将持续提高智驾大模型的性能、迭代速度和泛化能力，充分发挥人工智能算力、云计算和大数据方面的优势，推动智能化技术在智能驾驶领域的落地应用，为用户提供更智能、更安全和个性化的驾驶体验。

PART

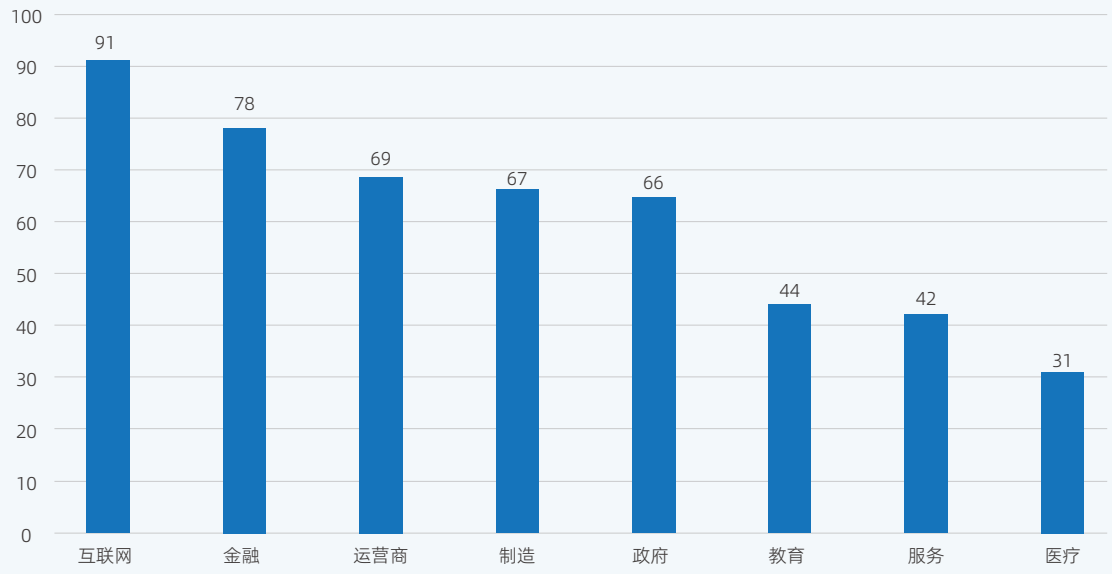
第三章

人工智能算力发展评估

3.1 行业排名

各个行业对于人工智能技术的应用愈加重视，持续加大相关投资与研发力度，加深人工智能渗透度。通过评估人工智能技术场景和应用场景成熟度、投资规模等维度，IDC对人工智能的行业渗透度进行评估，排名前五的行业依次为：互联网、金融、运营商、制造和政府。此外，人工智能在教育、医疗、能源等行业的应用也可圈可点。其中，互联网企业在大模型的研发、应用及推广过程中持续发挥引领作用；金融行业积极探索大模型与业务的融合，进一步发挥人工智能技术在风险控制、决策优化和金融产品推荐等场景中的价值，排名从第四名攀升至第二名；人工智能持续重塑制造业生产流程，加速产品设计、管理维护、安全监督等场景的智能化升级，赋能高端装备、工业机器人、汽车、航空航天、船舶等制造业重点领域的发展，排名从第五名提升至第四名；在教科研行业，人工智能尤其是生成式人工智能在提升教学效率、促进个性化学习、优化资源配置等方面贡献重要力量，引领科研新范式，渗透度排名从第八名提升至第六名。

图19 中国人工智能行业应用渗透度



来源：IDC，2025

互联网

互联网企业在大模型的研发、应用及推广过程中依旧起到了引领的作用，并积极探索应用场景创新，基于自身良好的数据基础和技术能力加速AI agent（智能体）研发，推进人工智能原生应用开发，覆盖问答、写作、客服、路线规划、生活指导、学习助手、角色扮演、视频生产、图片企业智能客服、智能销售分析等多个场景， 并发布相关开发平台，满足其他行业用户的使用需求。互联网行业将继续加大对人工智能技术的投资，以提高用户体验，优化企业收入模式，并通过针对性设计的智能计算力部署，实现资源的最大化利用，降低运营成本。未来，建议互联网企业面向大模型落地和推理场景构建先进算力基础设施，提升数据处理和实时分析能力，满足大规模用户和复杂应用的需求。

金融

金融行业通过对人工智能计算力的深化应用，显著提升了运营效率和风险控制能力，同时推动了产品创新，为金融服务的普惠化和精细化发展奠定了基础。智能风控系统利用人工智能对海量数据进行实时分析，在欺诈检测和信用评估中展现出卓越的精准性；量化交易策略通过高性能的计算能力和机器学习优化，缩短了市场决策的时间差并提升投资回报；智能投顾借助自然语言处理技术，实现了个性化财富管理方案的快速定制，提升了客户粘性与满意度。此外，金融机构正在利用大模型和生成式人工技术辅助信息处理和业务决策，支持内部运营、金融知识理解与生成、政策研报解读、信贷审批、理财顾问等场景。在金融科技的支持下，算力需求主要集中在实时处理和大规模建模领域。特别是在高频交易和金融预测模型中，算力需求有大幅度提升。人工智能驱动的动态定价、市场预测和风险管理等场景，正逐步成为金融机构对算力依赖的核心领域。随着监管技术（RegTech）的普及和绿色金融的崛起，人工智能计算力将成为金融机构提升合规效率和支持可持续发展的关键驱动。未来，建议持续加强金融数据中心的安全性和合规性，采用人工智能技术，提升风险管理能力。

运营商

电信行业正通过人工智能的全面赋能，不断提升网络管理效率和用户服务水平。中国运营商正在部署人工智能技术进行智能客服、网络优化及预测性维护，这些应用显著改善了用户体验并提升了网络的稳定性。例如，在网络管理方面，人工智能帮助运营商实时监测、分析网络状态，通过异常检测和自适应优化保障网络高效运行，特别是在覆盖更大范围的5G网络中表现尤为突出。此外，预测性维护利用机器学习算法识别设备潜在故障，从而提前采取措施，降低维护成本并减少服务中断。人工智能将在动态资源分配、服务个性化和超大规模网络管理中发挥更重要作用，预计将助力运营商实现收入增长。此外，基于人工智能的精准营销技术正在帮助运营商深入洞察用户需求，提供定制化服务和套餐，进一步优化服务升级。未来，建议运营商可提高高性能计算资源利用率，加速应用场景挖掘，优化网络资源调度，支持更多低延迟、高带宽的智能应用和服务。

制造

制造业加速智能化转型，通过人工智能技术优化制造生产流程，减少材料浪费并显著提高设备维护的精准性；利用人工智能视觉技术实现质量检测自动化，替代人工检测，大幅减少缺陷率；通过机器学习模型对生产过程中的关键参数进行优化，缩短生产周期。目前，大模型技术已经在智能搜索、知识管理、视觉安监等场景实现初步应用，未来，随着RAG和Agent、MoE等技术逐步成熟，制造业企业可通过大模型、小模型、机理模型和业务应用的组合，深化大模型与业务的融合，提升数据准备度，推进工业互联网平台建设，整合边缘计算和云计算能力，完善智能问答、智能工厂、智能营销预测、智能供应链管理等能力建设。

政府

在政务服务领域，人工智能正快速融入社会治理和民生服务等领域，推动服务效率和质量的提升。在社会治理方面，人工智能通过数据分析、实时监测、决策支持和自动化流程，优化公共服务，增强安全管理，促进透明度和公众参与，实现社会治理的全面优化。例如，利用人工智能技术进行环境监测，实时分析空气质量和污染源，推动精准治理。在民生服务方面，随着老龄化社会的加速到来，智能居家养老成为新兴需求。人工智能技术正在支持远程健康监测、智能家居辅助设备的开发，为独居老人的生活安全与健康管理提供便利。在交通出行方面，人工智能和大数据结合的车联网（V2X）技术可实现交通流量的实时监测和调度，优化信号灯控制，缓解拥堵问题。交通预测模型和自动驾驶技术也开始得到实践应用，逐步推进智能交通的发展。基于强算力的智能交通系统能够利用生成式人工智能技术预测交通状况，为驾驶者提供实时路况信息，提升出行效率。电动汽车和自动驾驶车辆的普及进一步提升了交通管理系统对于实时数据处理和计算能力的需求。未来，建设政府可加速统一平台建设，整合数据资源，提升数据共享和智能决策能力，推动智能化建设。

教育

人工智能尤其是生成式人工智能在提升教学效率、促进个性化学习、优化资源配置等方面贡献重要力量。在教学场景中，自适应学习系统通过高性能的计算资源实现了对学生行为数据的实时分析，提供个性化学习路径，提升了学生的学习效率；虚拟现实与增强现实技术依托强大的算力支持，使沉浸式学习成为可能，尤其在工程、医学等领域的实验教学中展现了显著优势。人工智能辅导系统已在全国范围内普及，特别是在教育资源分配不均的地区，为学生提供了随时随地的智能问答和学习指导。在教科研领域，人工智能正在引领科研新范式：生成式人工智能可以通过自动化数据分析和报告生成，提升科研效率；生成研究假设和实验设计，辅助科研创新；提供文献综述和参考文献管理，简化研究过程；此外，人工智能在生命科学和物质科学领域的应用愈加广泛且深入，在计算物理中，人工智能用于模拟和预测物理现象，加速新材料和药物的发现；在材料设计中，通过机器学习算法，人工智能可以预测材料性能，优化设计，缩短研发周期；在组学分析中，人工智能提升了基因组学和蛋白质组学数据处理的速度和准确性，发现新的生物标志物和治疗靶点。这些应用显著提升了教科研的效率和创新能力。目前，教育行业的人工智能计算力需求正在快速攀升，智能学习终端、教育数据中心和虚拟教学平台成为教育算力消耗的主要来源，以支持传统教育模式的转型和智能化科研的发展。未来，建议持续优化高性能的计算资源供给，支持更大规模的在线教育、虚拟实验室和智能教学的应用。

医疗

人工智能辅助诊断已成为医疗领域的核心应用。基于强算力的大模型能够实时处理和分析海量医学影像，在早期癌症筛查、心血管疾病诊断和神经系统疾病识别中，大幅度提升诊断准确率，显著降低误诊率。此外，通过人工智能系统还能实时生成个性化诊疗建议，将缩短医生决策时间，提高诊疗效率。药物研发领域因算力技术的突破取得长足进步。人工智能驱动的分子筛选与药物设计平台显著缩短了新药研发周期，将潜在药物的筛选时间从传统的几年缩短至数月。中国药企在抗癌药物和罕见病治疗领域也取得了多项创新成果。智慧医院正在快速崛起。通过算力技术整合患者数据、设备运行状态和临床路径，智慧医院实现了患者管理自动化、资源调配智能化。远程医疗与虚拟健康助手基于强算力模型，覆盖了更多偏远地区，将医疗资源辐射范围扩大，极大改善了医疗公平性。未来，建议可加速区域性医疗人工智能算力中心的建设，整合医院和研究机构的数据和计算资源，支持大规模医疗数据分析、模型训练和智能诊断系统，提升医疗服务质量和效率。

3.2 地域排名

本报告针对不同城市在人工智能投资规模（包括人工智能算力投资规模，人工智能其他投资规模，未来投资计划）、人工智能相关政策支持力度（包括人工智能相关政策扶持力度、政策落地情况和实施进展）、人工智能技术成熟度（包括人工智能技术应用成熟度、第三平台技术应用成熟度、数据平台成熟度），以及劳动供给（包括人工智能相关技术人员数量和水平、AI企业人数/企业数量、未来人才储备）等维度的情况，并基于持续研究和最新用户调研，进行综合评估。鉴于生成式人工智能与大模型技术的战略重要性日益凸显，本年度的城市评估框架将城市在大模型及生成式人工智能技术领域的投资力度、建设进度、政策支持、应用水平、规划布局等因素纳入关键评价指标体系。最新评估结果显示，北京凭借大量人才、成熟的企业和有力的政策扶持，继续领跑人工智能算力发展，位居首位；杭州和上海分别位列第二位和第三位。此外，深圳、广州、南京、成都、济南、天津、厦门等城市在人工智能领域也具有较为突出的表现。

相较前一年，上海、广州、成都、天津、厦门五座城市排名有所提升。上海凭借其国际化优势和政策支持，加速人才引进，吸引顶尖专家，夯实技术创新基础；广州持续吸引大量人工智能企业和创新实体，特别是在智能制造、智慧城市、医疗健康等领域，展现了强大的产业聚集效应，推动了技术创新与应用的快速发展；成都与厦门则通过加速算力基础设施建设，凭借云计算中心和智算中心为人工智能、大数据和科研领域提供了关键支持；天津侧重高校协作，为当地人工智能提供了大量优质人才。

图20 中国人工智能计算力发展评估--城市排行

来源：IDC，2025

北京

北京持续位列第一。北京聚集了大批大模型企业，凭借其有力的政策支持及实力强盛的公司，吸引了大批人工智能领域人才。由北京大学国家发展研究院与智联招聘联合发布的《人工智能大模型对我国劳动力市场潜在影响研究：2024》指出，北京市在2024年上半年招聘的人工智能相关岗位数量占全国的19.1%，而对应的投递人数占比更达14.3%。北京各企业推出诸多具有代表性的大模型及应用产品，为中国大模型研发和应用提供强劲动力。此外，强大的基础设施建设为企业的研发和部署提供了坚实的支持。在政策方面，北京市政府加大对人工智能行业的政策支持和资金投入，2024年7月印发实施《北京市推动“人工智能+”行动计划（2024-2025年）》，从“标杆应用、示范应用、商业应用”等三个维度制定规划，加速人工智能应用，构

建大模型赋能的经济社会发展蓝图，8月，海淀区发布了《中关村科学城人工智能全景赋能行动计划（2024-2026年）》，通过先行先试、全域覆盖、全面辐射等方式，推动人工智能在全行业、全领域的落地应用。此外，北京将推进国家数据管理中心、国家数据资源中心和国家数据流通交易中心的建设，夯实数字基础设施建设，其中，海淀区作为北京数据基础制度先行区，积极建设北京人工智能公共算力平台，建成人工智能数据运营平台，汇聚高质量数据近2PB，支撑区人工智能企业的发展需求。目前北京市备案大模型105款、已建公共智能算力2.2万P，这些因素共同造就了北京在国内人工智能算力领域的领先地位。

杭州

排名第二的杭州在人工智能领域的持续创新举措巩固了其作为全国人工智能技术和应用中心的地位。杭州通过政策、应用和基础设施的协同发展，加速经济数字化转型，强化在科技革命中的领先地位，成为中国人工智能创新的标杆城市。政策方面，“521”人才引进计划和《浙江省“人工智能+”行动计划（2024–2027年）》成为发展核心，推动基础研究、技术应用、人才培养和政策优化。在项目应用上，杭州在智能医疗、金融科技、智慧交通等领域成果显著，此外，智能零售和智能制造领域的技术突破，助力杭州在多个行业树立人工智能应用典范。基础设施建设方面，杭州智算中心和云计算平台取得新进展，云计算与大数据中心的升级增强了数据处理能力和安全性，而绿色数据中心项目的启动体现了可持续发展的承诺。

上海

上海在人工智能领域展现了强劲的发展势头。在政策层面，上海市印发《关于人工智能“模塑申城”的实施方案》，旨在深入贯彻国家关于发展“人工智能+”的战略部署，加快建设《上海市促进人工智能产业发展条例》，该方案设定了到2025年底的多个发展目标，包括智能算力规模突破100 EFLOPS、形成50个行业开放语料库示范应用成果以及设立3-5个大模型创新加速孵化器等。在人才引进方面，上海市政府还推出了《上海市重点产业领域人才专项奖励实施办法（征求意见稿）》，旨在吸引和培养集成电路、生物医药、人工智能等八大重点产业领域的优秀人才，促进相关产业发展与创新。在行业应用方面，上海聚焦智智能医疗、金融服务、智慧城市和智能制造等领域的发展，通过大模型的应用加速行业智能化转型，提升各领域的服务质量和效率 。此外，上海正在加速数据中心和云计算平台的建设与扩容，专注于提升人工智能和大数据的处理能力。同时，市内多个数据中心正在进行升级改造，提升数据处理能力和安全性，以应对不断增长的数据需求。凭借这些综合措施和项目的推进，上海在人工智能技术应用和创新能力方面不断提升，位列排行榜第三位，较前一年上升一名。

深圳

深圳位列第四。深圳在人工智能领域继续深化创新布局，巩固其作为全国技术前沿城市的地位。在政策支持方面，深圳市政府推出了首支区级人工智能专项基金，规模达到10亿元，专注于“人工智能产业化”和“产业人工智能化”，将发展重心放在语音识别、人脸识别以及人工智能芯片等关键领域，旨在通过强有力的资金支持，加速人工智能技术的创新与应用。在智能制造、金融科技、智能医疗和智慧城市等领域，深圳的人工智能技术均取得了显著进展。在加速数据中心与云计算平台布局的同时，多个现有数据中心也在进行升级改造，以提高安全性和数据处理能力，满足日益增长的市场需求。在教育与人才培养方面，深圳的高校与企业合作推出了多项人工智能专业和培训项目。

广州

广州在排行榜中位列第五。广州在人工智能领域持续加快发展步伐，巩固其作为南方科技中心的地位。政策方面，广州市政府推出了《数字广州建设总体规划》，提出全领域推动“五位一体”数字化转型，以支持人工智能技术的研发和应用，特别是在智能医疗和智慧城市建设方面，力争2030年建成数字中国标杆城市。广州加速了数据中心的建设与升级，新建的数据中心项目将重点服务于人工智能和大数据应用。同时，现有数据中心正在进行改造，提升安全性和处理能力，以应对日益增长的市场需求。

南京

南京在排行榜中位列第六。南京凭借政策支持、基础设施建设及产业创新，稳固其算力采购市场第六大城市的地位。南京市发布了《人工智能创新发展行动计划（2024–2026年）》与配套政策，计划在三年内投入超过60亿元人民币，目标是打造规模达600亿元人民币的人工智能核心产业。在这些举措的推动下，南京成为国家人工智能创新应用先导区，具备强大的科技与人才优势。在行业应用上，南京通过“人工智能+”行动计划构建了多个创新场景，涵盖智能医疗、智慧交通和城市治理等领域。多项签约项目落地南京智谷，包括大模型与边缘计算等前沿领域，为南京人工智能产业注入新的增长动能。

成都

成都位列第七，在人工智能计算领域表现出不断加速发展的势头。相较于前一年，成都排名提升了两位，显示出其在西南地区科技创新中的核心地位。成都市政府推出的《四川省人工智能产业链总体工作方案（2024-2027年）》为其算力基础设施建设提供了强力支持。四川省计划实施多个算力项目，预计投资高达559.89亿元，其中包括31个关键基础设施建设项目。此外，成都还出台了创新的“算力券”机制，以降低算力使用成本并推动科研和技术发展。成都的算力基础设施建设也在持续升级，成都云计算中心和智算中心为人工智能、大数据和科研领域提供了关键支撑。2024年，成都的目标是实现算力资源的高效分配与应用，以支撑未来五年的科技与产业创新。同时，成都还通过加强与高科技企业的合作，在加速智能化转型方面发挥了重要作用。全市的智算中心不仅提高了数据处理能力，还为智能制造、医疗健康和智能交通等领域提供了强大的技术支持，推动了产业的跨越式发展。

济南

排名第八。济南承载了中国近一半人工智能服务器的产能，在人工智能领域持续加快发展，巩固其作为山东省科技创新中心的地位。政策方面，济南市政府推出了《济南市新一代人工智能高质量发展行动计划》，未来三年将推动人工智能产业发展，设定四大行动和17项任务，目标是打造多个产业集聚区，建设人工智能中心，促进新兴技术协同研发，实施智慧交通示范工程，并加强人才培养，提升整体产业能力。济南企业与本地高校开展合作，推进智能项目的研发和实际应用。

天津

天津凭借着政策支持、高校人才输送和算力底蕴闯入榜单，位列第九。经过“新一代人工智能产业发展三年计划”，天津夯实了自身算力基础。在京津冀协同发展的背景下，抓住风口，厚积薄发。通过加速产学研合作，助力智能制造发展。在城市管理方面，天津实施的“智慧交通解决方案”通过人工智能技术优化交通流量，预计交通拥堵时间减少30%。另外，天津正在推进滨海新区人工智能产业园的建设，预计将引入超过200家人工智能相关企业，创造5000个就业机会。

厦门

厦门位列第十名，得益于其在数字经济和人工智能领域的持续发力。政策方面，厦门市出台《厦门市加快数字经济发展行动计划（2024-2025年）》，提出多项具体措施，包括构建完善的集成电路产业体系，推进智能制造和服务业数字化转型，以及厦门数据港的建设，统筹部署厦门人工智能计算中心、赋能中心、大模型创新中心等新算力基础设施，到2025年全市智能计算中心算力可达到1.1EFLOPS以上。此外，加速建设福建省人工智能产业园厦门园区，拓展无人驾驶、智慧安防、智慧交通、智慧医疗等优势和特色应用场景。在人才培养上，厦门大力发展人工智能教育，目前全市有13所高校开设人工智能相关专业，为数字经济发展储备人才。

另外，银川和苏州的表现也可圈可点。银川市作为“交换中心+枢纽节点”双中心，积极推进“算力之都”建设，布局多个智能技术与算力资源深度融合的项目。目前，银川市“算力之都”建设已签约近50个重点算力产业项目，总金额达380.5亿元，已形成智能算力3000P以上。苏州市通过发布《“人工智能+”创新发展试验区行动方案》等政策，设定了到2027年产业规模突破3000亿元及生成式人工智能服务深度合成服务算法国家级备案数达50个等目标；苏州还积极举办供需对接会，推动大模型备案和标准体系建设，营造了良好的产业发展生态环境，这些努力为人工智能在苏州的发展奠定了坚实基础。

PART

第四章

IDC建议

考虑到中国人工智能市场的现状，IDC针对行业用户和人工智能解决方案提供商分别提出了如下行动建议，希望对中国人工智能的发展和生态的成熟有所裨益。

对行业用户的建议

- **制定长远战略规划，全面准备人工智能应用落地**

  人工智能是企业创新力的重要评估标准。在发展人工智能时，企业应首先制定明确且与长期发展目标一致的战略规划。明确人工智能在业务流程中的角色、预期的投资回报率，以自身业务场景为出发点，关注业内人工智能应用的成功案例，确定哪些用例具有最高级别的战略优势。企业需要全面评估技术，确保现有的IT基础设施能够支持人工智能项目的部署，并考虑未来的扩展需求，以确保技术设施既满足当前需求又具备未来发展潜力。

- **选择合适的合作伙伴，按需获取人工智能算力服务**

  人工智能应用，尤其是生成式人工智能应用的落地将耗费大量的时间和资金成本，对于缺乏智算基础设施建设资金或技术能力的企业来说，面对智算中心、云服务商和本地数据中心提供的多样化算力资源，企业应坚持长期主义，遵循融合适用的原则寻找最佳算力供给方案，并合理规划算力采购的来源分配。如前文所述，人工智能算力服务不同于通用算力，其新的需求将改变算力服务供应市场格局，企业在选择算力服务提供商时应保持开放的态度，从服务商规模、项目经验、数据积累、定制化能力、智能算法水平等多个维度来评估各家供应商的能力，从而获得最适合自己的智算服务能力。

- **以用定建，提高算力利用率**

  在进行人工智能基础设施建设之前，企业应根据具体的业务需求和应用场景来确定建设规模，避免资源浪费。运营过程中，应根据实际需求逐步扩展计算资源，保证成本效益的最大化。同时，通过优化算力基础设施架构、提高模型算力效率、增强数据支持等方式，提高现有计算资源的利用效率。此外，还应面向推理场景构建算力基础设施，以支持生成式人工智能落地，并通过精细化的算力监测和智能调度，持续提高资源利用率，有效控制成本。

- **发挥系统创新优势，加速人工智能应用**

  人工智能是紧耦合的系统。人工智能应用落地不仅需要算力基础设施，还需要算法，数据，以及运维等全方位就绪：先进的硬件设备可以提供强大的计算能力，超级人工智能以太网可以通过智能网卡提升网络效率，创新的算法可以提高模型的准确性和效率，高质量的数据可以增强人工智能系统的学习能力，高效的软件平台可以简化开发流程。系统化的创新不仅加速了人工智能应用的落地和商业价值的实现，还提升了企业的技术能力、竞争力和市场响应能力，帮助企业加速实现智能涌现与人工智能应用。

对解决方案提供商的建议

■ **聚焦人工智能开发应用平台，降低人工智能使用门槛**

人工智能解决方案提供商需要为用户提供丰富的大模型选择，匹配行业用户最适用的模型，打造灵活的生成式人工智能应用与开发平台。不同行业用户通常处在不同的人工智能应用成熟度阶段，因此，解决方案提供商应该尽可能提供灵活的定制功能和数据支持，强调可扩展、灵活的IT基础设施解决方案，基于算法创新、硬件重构和软件定义解决算力资源不足的问题，这些解决方案需具有可观的成本效益、强大的安全功能并易于集成，降低用户的部署门槛，尤其是那些处在人工智能应用早期阶段的用户。

■ **优化用户的人工智能应用成本**

解决方案提供商应采用透明的定价模型，并清楚地传达总体拥有成本（TCO），以帮助行业用户更好地匹配自身预算与人工智能计划。解决方案提供商可以通过灵活定价模式的全栈解决方案，帮助组织管理成本，同时实现其投资回报率目标。此外，结合高能效的液冷数据中心解决方案，可以在高算力密度场景下提供更好的长期投资回报率。

■ **继续建立广泛的生态合作**

人工智能应用需要完善的算力、算法和数据的支撑。无论是传统AI还是最新的生成式AI能力的产业化落地，都将依赖产业中多方的共同建设，形成一个完整的解决方案产业生态链。在这一生态中，算力基础设施提供商、平台提供商和行业解决方案提供商都是重要参与者，建立广泛的合作联接将极大地助力人工智能产业化发展。

关于 IDC

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC帮助IT专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC在全球拥有超过1100名分析师，他们针对110多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在IDC超过50年的发展历史中，众多企业客户借助IDC的战略分析实现了其关键业务目标。IDC是IDG旗下子公司，IDG是全球领先的媒体出版、会展服务及研究咨询公司。

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

blogs.idc.com

www.idc.com