IDC 浪潮信息

2023 - 2024 年中国人工智能计算力发展评估报告

目录

IDC观点

1 **2023年是人工智能发展的重要转折年，企业正加速从业务数字化迈向业务智能化。**大模型的突破和生成式人工智能的兴起为企业实现产品/流程的革新提供先进生产工具，引领企业和产业迈入智能创新的新阶段。对于企业人而言，其将不再局限于思考“如何在产品/流程中增加智能化能力”，而需要更多关注“如何使用人工智能实现产品/流程的革新”。

2 **大模型和生成式人工智能的发展将引发计算范式之变、产业动量之变，以及算力服务格局之变。**未来几年，构建和调优生成式人工智能基础模型以满足应用需求，将为整个基础设施市场带来改变和发展机遇。

- 从计算范式角度而言，人工智能算力基础设施将持续向高性能、高互联等方向演进以更高的计算能力和链接速度加速实现大规模参数和数据集的训练和调优；不断提升算力泛在性，推进人工智能在云-边-端的覆盖，满足无处不在的智能化需求；通过优化计算架构、算法和软件栈，支持多元异构算力的协同，构建软硬件生态，加速计算技术的发展和创新。

- 从产业动量角度而言，基础模型技术的突破为人工智能产业的发展增加活力，催生新的玩家和投资机会，基础模型的持续迭代、调优、场景适配和部署、落地等对基础设施层、模型层、平台层和应用层提出新的需求，带来新的服务模式，降低人工智能技术的应用门槛，通过微调等方法实现与下游任务的适配，加速人工智能基础设施软件的开发、部署和应用，为用户和行业提供更多创新应用。

- 从算力服务角度而言，传统算力资源虚拟化共享复用的机制难以满足大模型时代企业对于集群式的高性能算力需求，生成式人工智能将加速企业更多地使用人工智能就绪的数据中心设施和人工智能服务器群，供应商需要具备提供定制化的、优化的基础设施服务能力，满足单个用户对训练和推理资源的独占式、大规模、长时间使用的诉求，并缩短部署时间、提高对数据和输出的控制，满足应用场景的需求，帮助企业实现成本优化。

3 **从感知智能到生成式智能，人工智能算力需求快速增长。**大模型和生成式人工智能的发展显著拉动了人工智能服务器市场的增长。IDC预计，全球人工智能硬件市场（服务器）规模将从2022年的195亿美元增长到2026年的347亿美元，五年年复合增长率达17.3%；在中国，预计2023年中国人工智能服务器市场规模将达到91亿美元，同比增长82.5%，2027年将达到134亿美元，五年年复合增长率达21.8%。从算力规模而言，预计到2027年通用算力规模将达到117.3 EFLOPS，智能算力规模达1117.4 EFLOPS；2022-2027年期间，预计中国智能算力规模年复合增长率达33.9%，同期通用算力规模年复合增长率为16.6%。

4 **中国市场对智能算力供给能力的衡量标准将加速演变，未来应用为导向、系统设计为核心将是算力升级的主要路径。**中国市场对于算力供给能力的评估指标将从硬件性能向应用效果转变，企业在获得算力服务的过程中，会增加对于诸如单位时间可处理Token数量、可靠性、时延、训练时间和资金成本、数据集质量等指标的关注。技术提供商需要以应用为导向，系统为核心，构建算力基础设施，提高算力利用率，提升诸如卡间互联、多节点间互联等水平，通过灵活可扩展的集群满足市场的需求。

5 **中国应持续提升基础大模型研发能力，通过逐步完善的人工智能工程化工具，加速应用落地。**目前，受政策支持、算力水平提升、数据资源庞大以及科研实力增强等利好因素的推动，中国在基础大模型方面取得一定成绩，但仍需加大在基础性技术方面的原创性突破，夯实底层模型和算法能力。在实践中，企业需要根据具体的任务和模型设计来决定参数量的大小，技术提供商需要从硬件、软件、算法、数据服务等多个维度入手，结合行业特点进行框架、模型、数据的垂直整合，提升大模型的准确性和可用性。

6 **基于液冷服务器构建绿色数据中心，推进人工智能算力可持续发展。**人工智能算力的不断提升加速对能耗问题的关注。从数据中心机柜功耗上来说，传统数据中心每机架功耗一般在3-10kW之间，而每台GPU服务器的功率可高达50kW，对于数据中心操作员和规划人员来说，需要依据计算需求对资源进行合理规划和分配，积极探索采用液冷等先进冷却方法，满足实现可持续发展提出的要求。IDC预计，2022-2027年，中国液冷服务器市场年复合增长率将达到54.7%，2027年市场规模将达到89亿美元。

7 **目前，中国的人工智能技术正加速迈入全面应用时代。**人工智能领域持续追求对技术的创新及增进，注重类人化和诸如机器学习、深度学习等技术的推进，以更好地处理现实生产、生活场景中的复杂问题。大模型和生成式人工智能的落地，也给各行业带来新的赋能。IDC认为，知识管理、对话式应用、销售和营销、代码生成等是企业应用生成式人工智能的主要领域；在特定业务部门或职能部门（营销、销售、采购等），生产力场景和垂直行业场景也有广泛的应用价值。未来，伴随特定领域行业大模型、多模态大模型、小样本学习、语音识别、计算机视觉等技术的突破，生成式人工智能的功能将实现持续进步，为企业带来新的能力。

8 **人工智能加速实现在行业和城市的渗透。**2023年，人工智能行业应用渗透度排名前五的行业依次为互联网、电信、政府、金融和制造。互联网行业依旧是人工智能技术应用和研发的主力军；在电信行业，云上人工智能能力的加速发展，进一步优化服务，支持电信网络的优化和智能化建设。在城市人工智能算力排行中，北京、杭州、深圳依然稳居城市排名前三位；其中，北京在大模型领域表现突出，聚集了大批大模型企业，推出诸多具有代表性的大模型及应用产品。此外，中国其他地区依然保持着对人工智能产业的热忱，正持续加大在相关领域的投资，不断推进人工智能产业的发展。

第一章
人工智能发展迈入新阶段

1.1 全球：生成式人工智能兴起，产业步入关键转折点

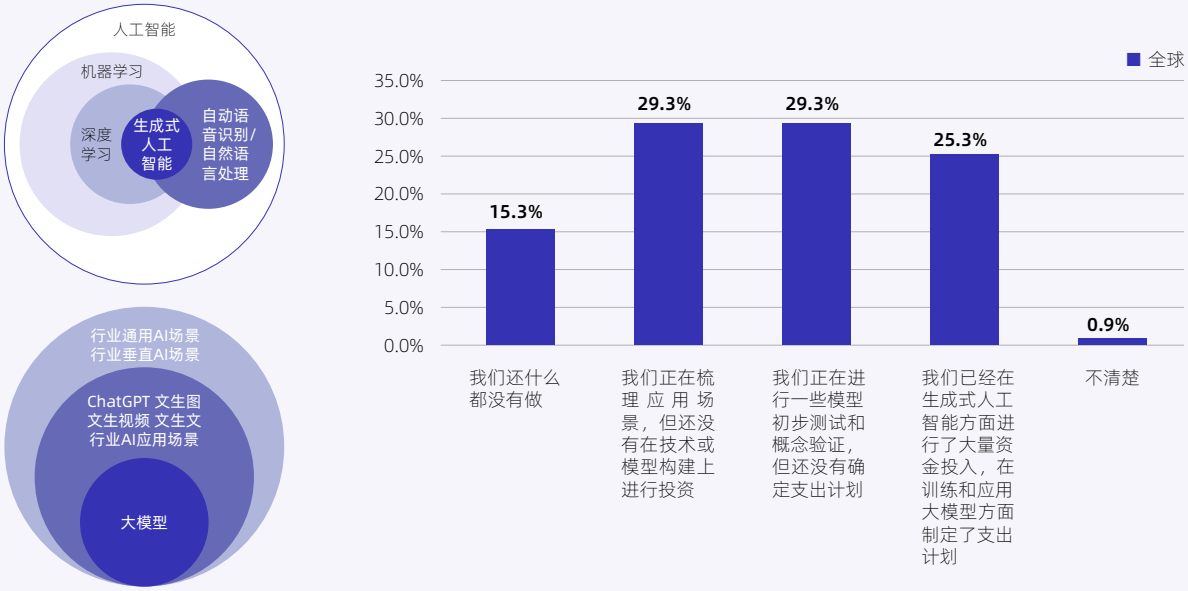
生成式人工智能和大模型将加速世界向智能化时代的迈进

2023年，人工智能实现了破圈式的发展。人工智能聊天机器人ChatGPT、AI编程工具GitHub CoPilot和图像生成系统Stable Diffusion 等生成式人工智能（Generative AI，Gen-AI）应用和工具产品的出现，为文本创建、图像生成、代码生成以及研发流程等工作带来全新的智能体验，极大提升生产力，提高生产水平。生成式人工智能应用的出现离不开大模型的支持。大模型是基于海量参数进行自监督学习的预训练模型，凭借更强的学习能力、更高的精准度以及更强的泛化能力，正在成为人工智能技术发展的焦点。

世界正在加速向智能化创新迈进。大模型及生成式人工智能的发展意味着人工智能正在从完成如图像识别、语音识别等特定任务，迈向拟人类智能水平，具备自主学习、判断和创造等能力。对于企业人而言，其将不再局限于思考“如何在产品/流程中增加智能化能力”，而需要更多关注“如何使用人工智能实现产品/流程的革新”。基于海量数据训练和模型调优，人工智能大模型具有更精准的执行能力和更强大的场景可迁移性，为人工智能在诸如元宇宙、城市治理、医疗健康、科学研究等综合复杂性场景中的广泛应用，提供更好的方案。

IDC调研发现，全球企业普遍关注并探索对生成式人工智能的布局，全球超八成被访企业已经开始展开相关实践行动，探索适用的落地场景；2023年，超过四分之一的企业已经对生成式人工智能技术进行了大量资金投入。

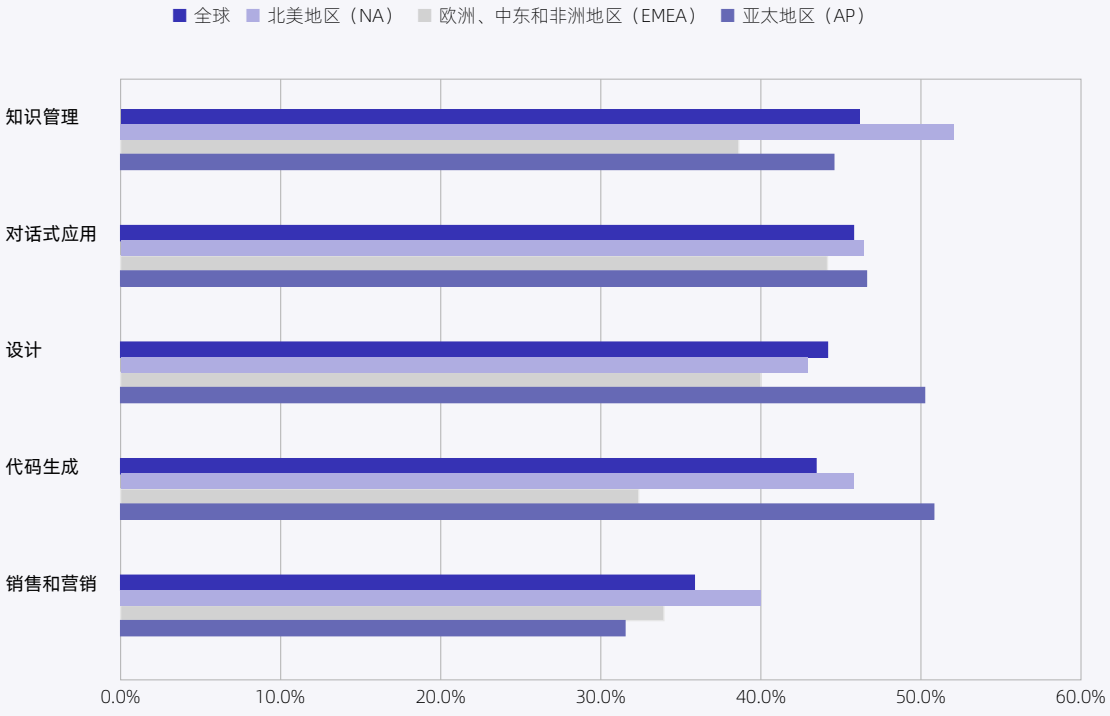
图1 全球企业对生成式人工智能的态度



来源：IDC，2023

从应用场景来看，IDC认为，知识管理、对话式应用、销售和营销、代码生成等是全球企业应用生成式人工智能的主要场景。其中，知识管理有望成为最有应用价值的生成式人工智能用例，通过人工智能手段，企业可实现对文本、图像和视频等知识内容的高效分析和管理，降低劳动密集型投入。

图2 生成式人工智能最具价值应用场景



来源：IDC，2023

目前，各国都在加强对大模型和生成式人工智能的布局和支持，以推动人工智能技术的快速发展和应用：

- **美国**持续推进各界在人工智能领域的快速发展，鼓励企业及科研机构积极创新，形成以科技巨头为引领的发展格局。通过推动基础研究和应用研究的发展，美国在基础大模型研发和生成式人工智能应用方面建立起优势，打造了现象级生成式人工智能产品，并将生成式人工智能技术广泛应用于行业领域和业务场景。2023年，白宫更新发布了《国家人工智能研发战略计划》，鼓励在控制安全风险的前提下，持续探索创新人工智能应用，促进研发投资，鼓励人才培养和产业合作。

- **欧洲地区**受到石油和天然气价格上涨、高通货膨胀、人员和技能短缺，以及供应链中断等因素影响，更重视技术对降本增效的积极推动作用，因而对智能化技术的关注较高。但欧洲地区整体对人工智能尤其是生成式人工智能在安全、隐私等方面的顾虑较多，故当前优先推动相关领域法律法规的建设和实施。2023年，欧盟批准《人工智能法案》，对涉及大量数据训练的人工智能系统提出了透明度和风险评估要求；欧盟还加强对人工智能伦理道德的监管，保护数据隐私和数据安全，加强对生成式人工智能的监管和审查，持续对自动驾驶系统、教育、移民和就业决策系统中的人工智能应用影响力进行评估。

- **在亚太地区，**中国、印度、新加坡、韩国、日本等国家积极制定国家人工智能战略，推进各项超大型生成式人工智能相关方案的落地。韩国政府重视人工智能基础设施和环境的发展，推动人工智能在各个领域的应用；日本政府通过加大投入和政策支持，推动生成式人工智能的研究和应用，并决定向企业提供资金补助，推进高算力基础设施的建设；

- **在中国，**政府加大了对生成式人工智能研究的支持力度，企业和科研机构也加速推动生成式人工智能的研究和应用，加速研究人工智能与实体经济、社会治理、民生服务等领域深度融合，促进技术的广泛应用和产业化发展；中国人工智能的发展在东南亚部分国家也逐步形成溢出效应，带动该地域相应产业的发展。

生成式人工智能和大模型将引发计算范式之变、产业动量之变、算力服务格局之变

计算范式之变：复杂的模型和大规模的训练需要大规模的高算力支持，这不仅需要消耗大量计算资源，而且对算力的速度、精度、性能也提出更高要求。基于持续演进的算力架构，不断满足基础大模型训练和推理应用过程中对计算、网络和存储的需求，以优质算力加速模型开发，通过提升算力泛在性满足无处不在的智能化需求，构建开放融合的软硬件生态，从而全面赋能智能创新。

- 高计算性能：市场对于更高性能的高端协处理器服务器，尤其是GPU服务器的需求将进一步提升，以满足大量高算力计算任务需求；

- 高互联：生成式人工智能工作负载也对连接人工智能服务器集群的网络结构提出更高的要求，推动网络架构的变革，以更好实现快速互联，加速模型的训练、调优和推理；

- 算力泛在性：人工智能算力还应提升其覆盖规模，以支持智能在终端、边缘、数据中心等位置的广覆盖，实现生成式人工智能推理能力在边缘、终端等位置的部署和应用；

- 多元化：应用场景多样性、硬件性能瓶颈、降低成本和提高能效需求等因素促使底层基础设施呈现多元化发展趋势，加强系统接口、互联协议、管理规范等方面的开放兼容显得很有必要。通过优化软件和硬件的协同，构建软硬件生态，方能加速计算技术的发展和创新。

产业动量之变：当前的模型市场中存在各式各样的模型类型，覆盖语言、图像、音频、音乐、代码、视频、3D、生物分子结构和行业特定模型等诸多领域，形成开源与专有模型并存的局面，这为企业的业务智能化发展提供基石：在创作相关的产业，大模型可重构工作方式，为创作方式提供诸多可能性，可自动生成文本、语音、视频等内容，还能激发创作灵感，节省时间和精力，提高创作效率；在自动驾驶领域，大模型可以辅助汽车和机器人更好地理解环境，做出智能决策；在零售领域，大模型可以从历史数据中学习用户的兴趣，提供定制推荐服务；在医疗领域，大模型可以通过对大量药物化合物进行分子模拟和预测，加速药物研发过程；在金融领域，大模型可以帮助金融机构更有效地进行风险评估和欺诈检测。

基础模型的发展正在成为促进整个人工智能产业保持活力的重要驱动因素。以模型为核心的变革性技术将带来丰富的市场机会，算力和数据支持、训练和调优、部署和应用等需求将促进人工智能产业的迅速发展。算法、应用、服务等诸多产业变量将成为创新的加速器，在算力生态链上的各个环节（包含基础设施层、模型层、人工智能平台层、人工智能应用层和服务层）催生出新的玩家、初创企业和投资机会：

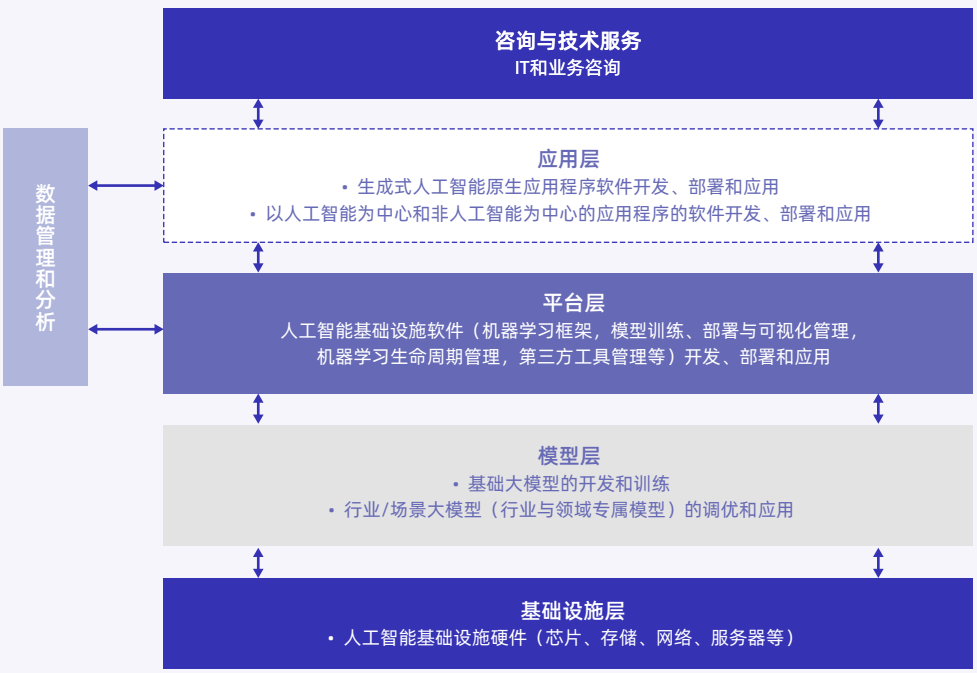
- 基础设施层：庞大且多样的市场应用将向算力供给提出更分化的需求，从芯片、互联技术、存储到基础设施软件，均需实现与市场需求的匹配，以满足不同应用场景、部署环境以及成本和性能的要求；

- 模型层和平台层：从基础大模型的研发、训练和调优，到行业化、场景化落地，各个环节都会创造出新的玩家和新的商业机会，通过将算力、模型、框架、基础设施软件平台等诸多能力调优，为行业需求提供匹配的解决方案，并在大模型赋能下，通过微调（Finetuning）等方法实现与下游任务的适配，在满足诸如“数据不出域”等行业特殊要求的前提下，落地大模型能力；

- 应用层：目前产品和服务以覆盖客户关系管理、知识管理、语音合成、生产力工具、图像设计、写作等领域为主，未来新玩家将持续入场，不断拓展生成式人工智能应用场景，面向金融、医疗、交通、科研、制造、自动驾驶等垂直领域提供更优的数据处理、训练和推理解决方案，以实现跨领域、智能化、个性化、可持续化的发展。

产业链间的不同参与者通过合作和创新，可共同推动大模型技术的发展和应用，促进整个产业生态的繁荣和发展。

图3 生成式人工智能生态价值链图谱



来源：IDC，2023

算力服务之变：生成式人工智能有望重构算力服务模式和市场格局。鉴于基础大模型的本地训练成本不菲，企业将更多地使用人工智能就绪的数据中心设施和生成式人工智能服务器集群，从而缩短部署时间，降低设施的长期投资成本。为算力服务市场带来新机会的同时，由于企业所需的算力服务质量和模式将会在大模型时代发生改变，也将为算力服务供应商带来新的挑战：

- 从基础设施层面来说，传统计算基础设施难以满足大模型时代对于算力、存储和网络的高性能需求，因此算力服务商需要从芯片、处理器、存储、网络、数据库、云原生架构等维度，对算力基础设施进行全面升级，满足用户在超大加速环境中对快速扩展的需求，提供可用、易用、高效的资源供给服务；

- 其次，人工智能算力需求会改变基础架构的算力调度和组合的方式。传统算力资源虚拟化共享复用的机制难以满足企业在大模型时代对于高性能集群式的算力的需求。在算力服务交付的过程中，供应商需要能够提供定制化的基础设施服务能力，满足单个用户对训练和推理资源的独占式、大规模、长时间使用的诉求，同时帮助用户实现成本控制；

- 算力服务商有必要基于软件栈提高模型训练效率，提高硬件利用率。除了硬件资源的提供，算力服务商还可基于用户需求对算法进行优化和改进，提高算法效率和准确性，通过“模型即服务”的方式，降低使用人工智能技术的门槛和成本，促进部署和应用。

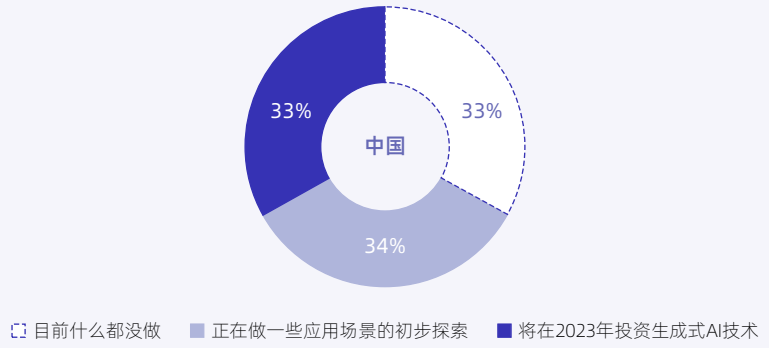
总之，市场对于算力服务需求的改变，将对算力服务商的商业模式和管理模式提出全新要求，生成式人工智能有望重构算力服务市场格局。在这个过程中，将会涌现更符合市场需求的算力供给方式，帮助企业应对算力需求，基于共享的基础设施，优化服务器利用率，获得更高的能效优势。

1.2 中国：人工智能产业加速创新，机遇与挑战并存

生成式人工智能背景下，中国人工智能市场展现出极高活力

从企业角度而言，中国企业对生成式人工智能的接受度普遍较高。据IDC调研，67%的中国企业已经开始探索生成式人工智能在企业内的应用机会或已经开始进行相关资金投入。

图4 中国企业对生成式人工智能的态度

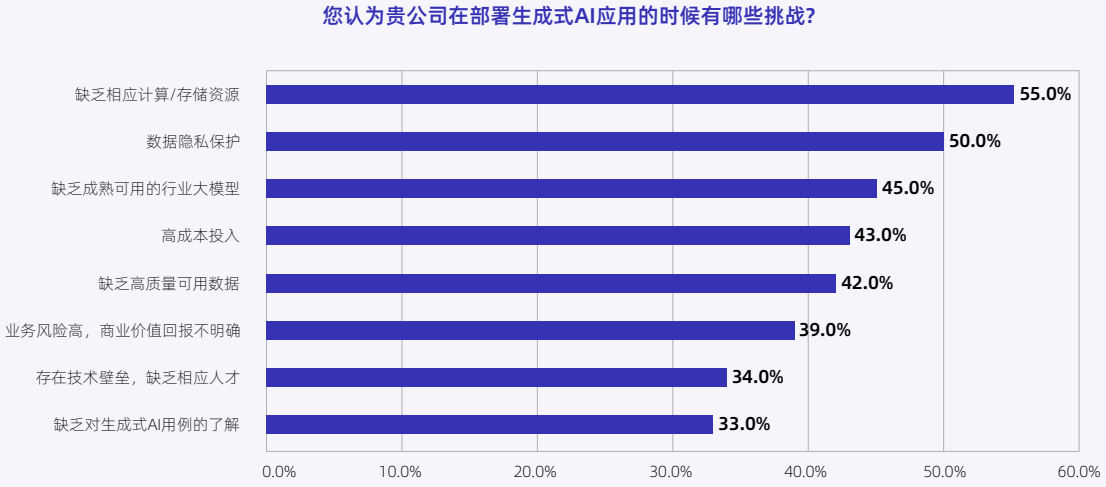


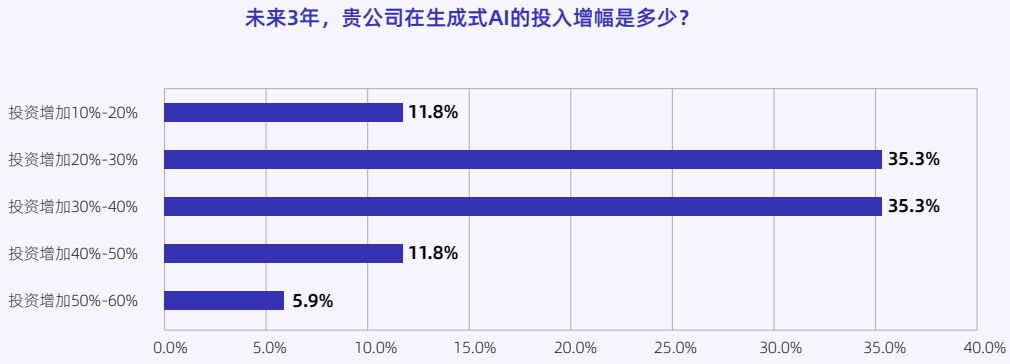
来源：IDC，2023

据调研，中国企业尤其认可生成式人工智能在加速决策、提高效率、优化用户和员工体验等维度带来的价值，并将在未来三年持续提高投入力度，超过七成企业增幅达到20%-40%；但与此同时，企业需要直面计算、存储等资源短缺、行业大模型可用性待提升以及投入成本高等问题带来的压力。

图5 中国企业对生成式人工智能的态度

来源：IDC，2023





来源：IDC，2023

从技术厂商角度而言，目前，国内诸多互联网巨头、科技企业及研究机构纷纷宣布在生成式人工智能的领域进行产业布局，国产大模型进入集中发布期，已拉开“百模大战”的序幕，通用类大模型（含语言类、视觉类和多模态大模型等）、任务大模型（含代码生成和生命科学等）以及行业大模型持续拓展应用领域，深化场景落地，不断探索商业价值，解决科研难题，助力产业升级。据公开信息，截至2023年10月，中国累计发布两百余个大模型，发布地主要集中在北京，其中以科研院所和互联网企业为开发主力军。

随着新算法、新应用的提出，人工智能产业生态呈现出高度活力，丰富的应用场景和潜在行业用例，将对大模型迭代和调优、行业和场景适配以及应用软件功能设计提出新的需求。当下中国大模型技术已经在自然语言处理、机器视觉和多模态等领域具备高度活力；面向未来，中国应持续关注基础大模型等基础性技术的原创性突破，以获得国际竞争力。可以预测，大模型应用将带来诸多产业化变革，因此，夯实底层模型和算法能力，对未来人工智能原生应用的质量和生态竞争力将起到决定性作用。

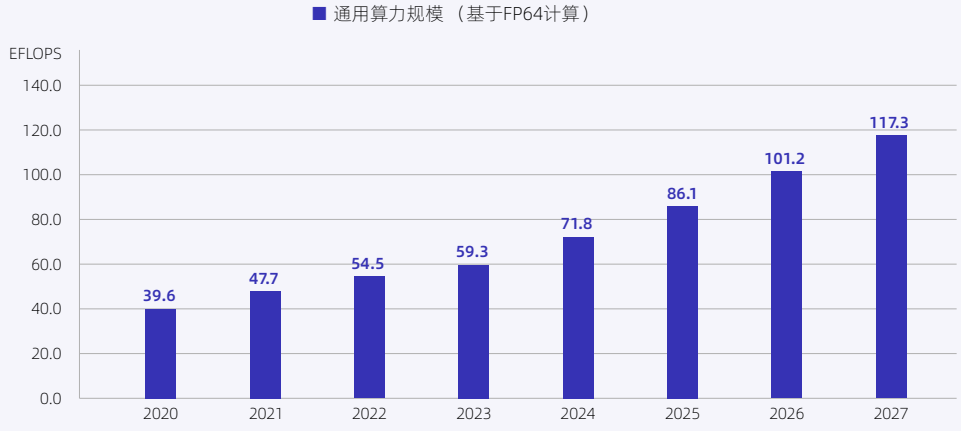
从政策角度而言，中国人工智能相关法规体系的完善与技术发展保持同频，政府将持续加强监督，疏导市场健康成长。2023年7月，中华人民共和国国家互联网信息办公室通过了《生成式人工智能服务管理暂行办法》 ，进一步明确生成式人工智能服务的指导方针，新法规已于 2023 年 8 月 15 日生效，将为实现发展与安全之间的平衡提供重要参考。

中国算力市场在摸索中蓬勃发展

生成式人工智能对中国人工智能服务器市场的发展带来了明显的拉动作用。丰富的应用场景和对技术创新迭代的热忱，让中国市场对于人工智能服务器的关注度和需求量均明显增长。IDC数据显示，2023年上半年，中国人工智能服务器市场规模达到30亿美元，同比增长55.4%。

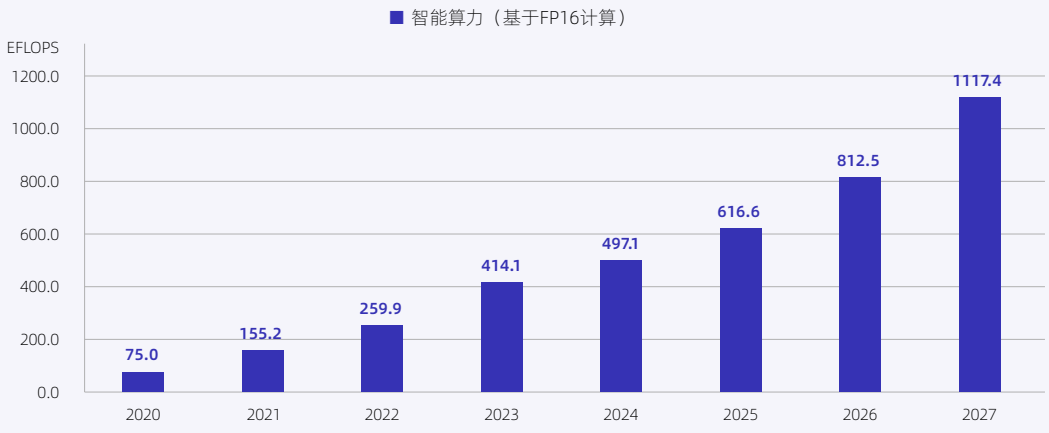
为评估中国算力规模发展现状和趋势，本报告基于《IDC中国加速计算服务器半年度市场跟踪报告》及智能加速卡半精度（FP16）相当运算能力数据，测算了中国智能算力规模。结果显示，2022年中国智能算力规模达259.9每秒百亿亿次浮点运算（EFLOPS），2023年将达到414.1 EFLOPS，预计到2027年将达到 1117.4 EFLOPS。此外，本报告基于《IDC中国服务器市场季度跟踪报告》及CPU双精度（FP64）运算能力数据，测算了中国通用算力规模。2022年中国通用算力规模达 54.5 EFLOPS，预计到2027年通用算力规模将达到117.3 EFLOPS。2022 - 2027年期间，中国智能算力规模年复合增长率达33.9%，同期通用算力规模年复合增长率为16.6%。

图6 中国通用算力规模及预测，2020-2027



来源：IDC，2023

图7 中国智能算力规模及预测，2020-2027



来源：IDC，2023

大模型的发展提升了智能算力的需求，中国的人工智能算力平台将呈现多元化发展趋势，整体市场也将充满机遇。同时，针对国内市场面临单芯片算力的瓶颈问题，以系统化思维构建算力基础设施平台，保障算力调度，优化大模型研发效率，成为破局之法和发展趋势。这也将加速中国市场对智能算力供给能力衡量标准的演变：用户对算力供给能力的评估指标将从对基础设施硬件性能的关注，迁移以及扩展至与应用需求和结果相关的维度上，如单位时间可处理Token的数量、可靠性、时延、训练时间和资金成本、数据集质量等。对于技术提供商而言，他们需要构建以应用为导向、系统为核心的算力供给能力，提高算力利用率，提升诸如卡间互联、多节点间互联等水平，支持灵活稳定扩展和弹性容错，积极打造通用的人工智能软件和硬件平台，以先进的系统性能力满足市场的应用需求。

适度超前部署算力资源，重点关注普惠和绿色

在中国，政府积极加大投资和支持，推动人工智能技术和应用的发展。在此背景下，算力基础设施建设成为一个重要环节，被纳入国家新基建范畴。在适度超前的指导思想下，国家正加大对人工智能算力基础设施的投资。目前，互联网企业、电信运营商，以及各级政府均积极投入到智算中心的建设之中。据不完全统计，截至2023年8月，全国已有超过30个城市建设智算中心，总建设规模超过200亿。

与此同时，人工智能、尤其是生成式人工智能，对能源消耗提出了更高要求，这让绿色节能成为先进技术落地的重要关注点。模型训练或人工智能应用程序开发以及应用阶段，即文本/聊天、图像或视频大量生成阶段，都会消耗大量能量，产生大量热量。从数据中心机柜功耗上来说，GPU服务器每台机架的功率可高达50kW，这将与传统数据中心每机架7kW功耗的行业平均水平形成鲜明对比。数据中心原有供电网络需要升级改造，以匹配生成式人工智能对基础设施的需求。

第二章
人工智能算力及应用

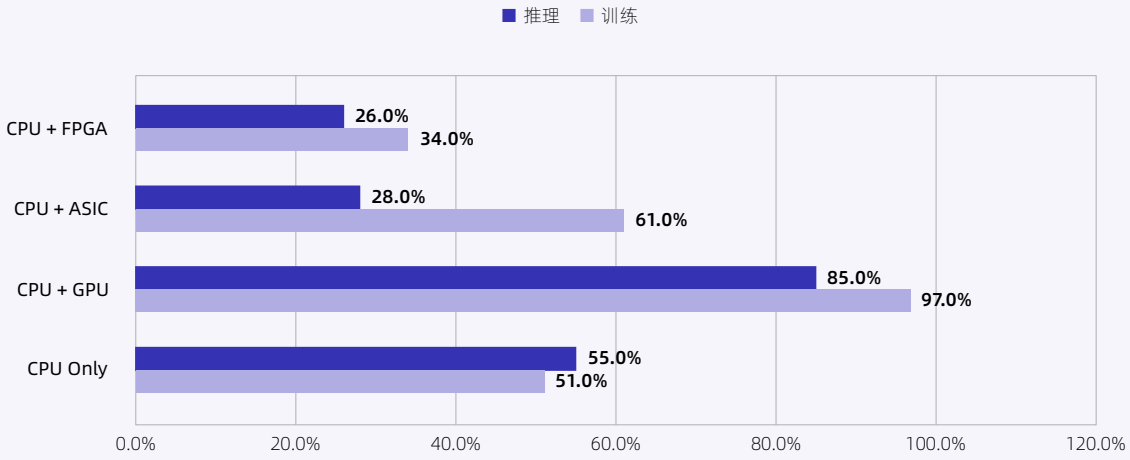
2.1 芯片：满足多场景高质量应用需求

人工智能芯片广泛应用于人工智能领域的各个方面，其所具备的高性能等特性可更好地支持复杂的算法，满足实时处理需求。通过硬件能力和软件能力的紧密结合，人工智能芯片具备自我学习和自适应计算能力，可以根据场景进行优化和调整，以更好地适应应用需求。

除主流大型芯片制造商外，全球还有大批创业公司专注于人工智能芯片技术的研发和应用。与此同时，全球一些大型互联网科技企业，特别是规模庞大的云服务提供商，也在人工智能芯片领域投入大量资源进行自研，并与相关产业链上的企业进行合作，加速自身在人工智能芯片领域的能力提升，满足其对硬件的定制化需求，并降低成本。

从技术发展视角来看，异构计算仍然是芯片发展趋势之一。异构计算通过在单一系统中利用不同类型的处理器（如CPU、GPU、ASIC、FPGA、NPU等）协同工作，执行特定任务，以优化性能和效率，更高效地利用不同类型的计算资源，满足不同的计算需求，比如，通过发挥GPU并行处理能力，可以提高模型，尤其是大模型的训练速度和效率；在数据预处理、模型调优等阶段，可以使用CPU进行计算和决策，或在控制和协调计算资源（如GPU、FPGA等）的工作过程中使用CPU，以确保计算过程的顺利进行；此外，可通过使用FPGA进行推理加速，从而将模型实现在边缘设备的部署，以开展更快速的实时推理工作。IDC调查研究显示，截至2023年10月，中国市场普遍认为“CPU+GPU”的异构方式是人工智能异构计算的主要组合形式。

图8 人工智能训练和推理工作负载选用的计算架构



来源：IDC，2023

在中国，芯片市场机遇与挑战并存。算力需求的提升给本土芯片厂商的发展提供了较大的空间，带来新的机遇。IDC预计，2023年中国人工智能芯片出货量将达到133.5万片，同比增长22.5%。

- **持续升级：**大模型发展等利好因素为中国人工智能芯片企业提供了良好的发展机遇，推动其在芯片设计、算法优化、生产制造等方面不断升级，加速创新研发，持续提升产品性能进步。

- **多元细分：**人工智能大模型应用呈多样化趋势发展，这就需要不同类型的芯片满足场景需求，除了CPU、GPU等传统的计算芯片外，国内芯片市场也在向更细分、更专业化的方向发展。

- **政策支持：**中国政府出台了一系列措施，鼓励和支持人工智能芯片研发和产业发展，包括资金支持、税收优惠、科研机构合作、人才激励等。各地区可根据自身需求和特点采取措施。这些政策措施为芯片产业的发展提供了良好的环境和条件，将有效促进中国人工智能芯片产业的快速发展，提升竞争力。

但与此同时，中国芯片产业发展也面临着一些挑战，其中以技术突破、人才培养、知识产权保护等方面的问题尤为突出。以封装技术为例，3D封装等技术的出现意味着高端芯片赛道上的竞争无须再仅围绕摩尔定律下的晶体管工艺能力展开，而是可以从新的角度切入，达成电路密度提升的目标，进而实现性能的升级，封装工艺突破正在成为中国芯片制造的新课题；此外，芯片产业发展不仅依赖硬件能力，还需要构建与硬件匹配的软件生态，包括操作系统、中间件和工具链等，当下诸多本土芯片技术储备和生态能力仍围绕小模型时代的识别式人工智能展开，难以匹配大模型和生成式人工智能发展所需的软件生态、模型框架、性能需求，因此本土人工智能芯片仍需在发展、继承和竞争中成长。未来，中国人工智能芯片产业需要进一步加强技术研发和创新，加强生态体系建设，培养更多高素质高技术的人才，加强国际合作与交流，提高自主研发和创新能力，以推动人工智能芯片行业可持续发展的目标。

2.2 服务器：高算力和高能效受到持续关注

从感知智能到生成式智能，人工智能算力需求快速增长。IDC认为，生成式人工智能和大模型发展正在成为人工智能算力市场发展的加速器。从感知智能到生成式智能，人工智能越来越需要依赖“强算法、高算力、大数据”的支持。模型的大小、训练所需的参数量等因素将直接影响智能涌现的质量，人工智能模型需要的准确性越高，训练该模型所需的计算力就越高。以ChatGPT模型为例，公开数据显示，其所使用的GPT-3大模型所需训练参数量为1750亿，算力消耗为3640PF-days（即每秒运算一千万亿次，运行3640天），需要至少1万片GPU提供支撑。据统计，当模型参数扩大十倍，算力投入将超过十倍，模型架构、优化效率、并行处理能力以及算力硬件能力等因素均会影响具体增加的倍数。

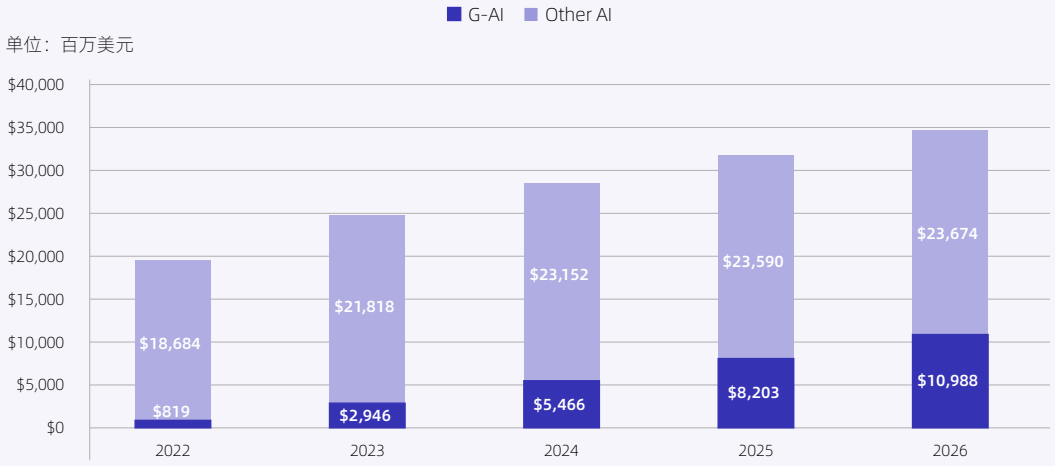
表1 大模型训练算力当量

模型名称	BERT-Large	GPT-2	GPT-3	T-5	MT-NLG	PaLM	PaLM-2	Switch-Transformer	Chinchilla	LLaMA	源1.0
参数量	3亿	15亿	1750亿	110亿	5300亿	5400亿	3400亿	1.6万亿	700亿	650亿	2450亿
算力当量	2.4PD	8.7PD	3640PD	26PD	9900PD	29000PD	85000PD	46PD	6795PD	6330PD	4095PD

来源：公开资料，浪潮信息，2023

由于大模型对计算能力和数据的高需求，其所需要的服务器设施将在人工智能基础设施市场中占据越来越大的份额。IDC预计，全球人工智能硬件市场（服务器），将从2022年的195亿美元增长到2026年的347亿美元，五年年复合增长率达17.3%；其中，用于运行生成式人工智能的服务器市场规模在整体人工智能服务器市场的占比将从2023年的11.9%增长至2026年的31.7%。

图9 全球人工智能服务器市场规模预测（含生成式人工智能和非生成式人工智能服务器），2022-2026

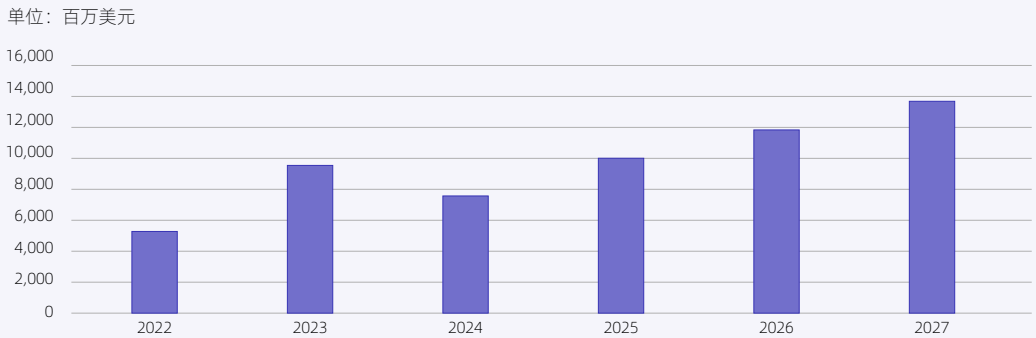


来源：IDC，2023

目前，中国的人工智能服务器发展取得了快速且显著的进展。由于人工智能是国家发展战略的重要部分，在当前数字经济的时代背景之下，服务器已经延伸到多个应用领域；人工智能服务器作为快速发展的新兴领域，市场规模也在不断增长。与此同时，国家相关部门陆续出台支持行业发展的相关文件。例如，2023年2月，中共中央、国务院印发《数字中国建设整体布局规划》，提到系统优化算力基础设施布局，以促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局，在政策层面为人工智能服务器需求量的增长提供保障。

从需求侧来说，在国内数字基础建设不断加快的带动下，人工智能服务器行业也保持快速增长。各大相关企业相继进行布局，加之中国人工智能应用场景的逐步落地，对算力的需求量快速增长，人工智能服务器在服务器整体市场中的比重越来越高。同时，中国的企业和研究机构积极进行人工智能服务器的技术研发和创新。这包括高性能的处理器、大容量的内存、高速的存储器和高效的冷却系统等领域的创新，以满足对计算能力和数据处理速度的需求。IDC预测，未来市场需求量也将会实现大幅度上升，预计2023年，中国人工智能服务器市场规模将达91亿美元，同比增长82.5%，2027年将达到134亿美元，五年年复合增长率为21.8%。随着生成式人工智能任务的不断增加，市场对于高性能和高能效的人工智能服务器需求将持续增长。未来的人工智能服务器将注重提高计算能力和处理效率，以适应更复杂、更大规模的人工智能应用。

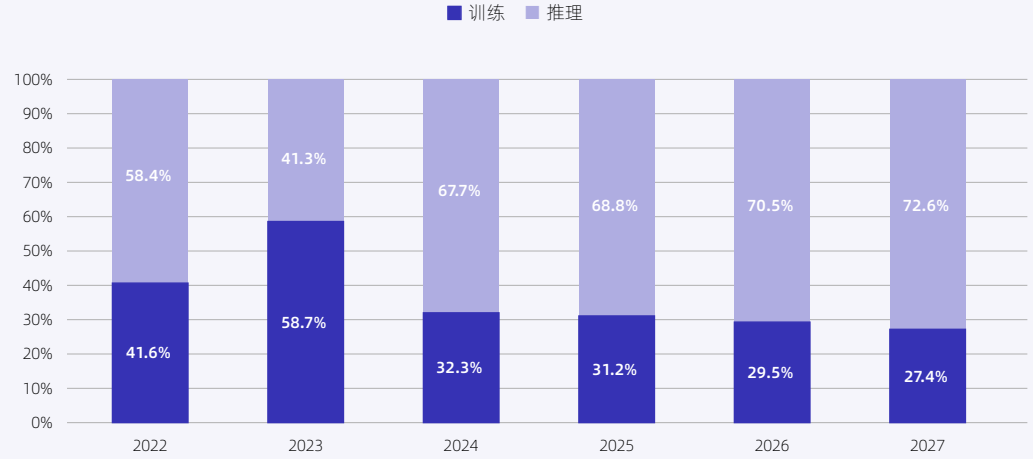
图10 中国加速计算服务器市场预测，2022-2027



来源：IDC，2023

从工作负载来看，2023年，大模型的兴起推动了训练服务器的增长速度。IDC数据显示，在中国，2023上半年训练工作负载的服务器占比达到49.4%，预计全年的占比将达到58.7%。随着训练模型的完善与成熟，模型和应用产品逐步进入投产模式，处理推理工作负载的人工智能服务器占比将随之攀升。IDC预计，到2027年，用于推理的工作负载将达到72.6%。

图11 中国人工智能服务器工作负载预测，2022-2027



2.3 算法和模型：加速模型迭代以探索行业实践

伴随模型参数不断扩大，智能涌现将为人工智能发展产生深远影响。当前，人工智能大模型处于高速发展阶段，Open-AI、谷歌、Meta、微软等在技术实力、资金、人才基础等方面具有优势的大型科技企业正推动大模型的发展，千亿乃至万亿级参数量加速智能涌现。

在这个过程中，我们发现：

- **大语言模型成为人工智能突破的核心之一。**大语言模型的出现使得人工智能在自然语言处理领域取得了显著的进展，目前已经在翻译文本、生成文章、回答问题、生成对话等任务场景下具备优质的性能和表现。国内人工智能大语言模型目前主要聚焦基础大模型训练，众多企业和机构也在不断尝试各种不同技术路线的大语言模型。但单纯依赖通用大语言大模型无法为创新企业提供持续的竞争力，垂直领域的数据、面向场景的模型优化、工程化的解决方案，才是将人工智能落地、建立竞争优势的关键。

- **大模型技术发展推动多模态模型不断升级迭代。**伴随深度学习、强化学习、迁移学习等多种技术的发展，多模态大模型正在成为人工智能领域的发展趋势之一。多模态大模型能够实现图像、文本、语音等模态之间的统一表示和相互生成，具有广泛的应用范围，覆盖自然语言处理、图像识别、语音识别、多媒体处理等诸多领域，诸如GPT-4等多模态大模型，可以在很多专业领域表现出类人类的水准，实现了突破性发展。未来，基于技术的不断突破，多模态将持续拓展各行业场景下的融合应用。我们看到，头部厂商持续布局多模态大模型领域，在注重模型整体通用性的同时，也在不断提升子领域的优化体验和技术升级。

大模型可通过自主学习和改进，不断刷新任务完成的质量。参数越多，模型可以学习到的特征和模式也会更多，但是，智能涌现不仅只与参数量有关，还受到模型设计、数据集选择、训练方法、模型架构、任务类型和计算资源等多重因素的影响。因此，在实践中，企业需要根据具体的任务和模型设计来决定参数量的大小，算力服务商需要从硬件、软件和算法等多个维度提供全面服务，共同提升大模型的准确性和可用性。以Megatron-DeepSpeed框架为例，Megatron是基于Transformer开发的、采用混合精度训练的模型，支持并行、多节点训练，通过与DeepSpeed深度学习加速优化库结合，创建了一个支持数据并行（DP）、张量并行（TP）和流水线并行（PP）的3D 并行系统，这使得千亿级参数量以上的大规模语言模型的分布式训练变得更简单、高效和有效。

持续提升人工智能框架技术的易用性和灵活性，加速人工智能在多元化应用场景的落地。人工智能所面对的下游行业纷繁复杂，应用场景多种多样，需求千差万别。定制化的开发模式造成项目开发成本高，开发周期长，难以适应变化。大模型通常具有更强的泛化能力，可在一定程度上解决模型定制的问题。预训练大模型正成为人工智能产业发展的重要选择，即：基于海量行业数据和知识，通过强大算力集群，预先训练基础模型，并结合应用场景的数据和各类需求，通过“预训练大模型+任务微调”的方式，进行更高效率的“工业化”开发。在实现软硬件协同优化、分布式计算以及云边端全场景部署等目标过程中，市场也对人工智能框架技术提出更高的要求。

此外，开源框架技术作为人工智能领域的操作系统，具有核心地位。人工智能开发者对于开源框架的依赖度非常高，尤其关注框架技术的易用性、稳定性、灵活性和可扩展性。开源框架虽然可以简化模型的构建，但要实施一个大型人工智能应用，不仅需要开源框架能力，还需要对数据采集、整理、模型训练，以及训练后的模型调优、迭代，乃至云边协同等一系列环节进行投入。当大模型在小算力设备上运行时，还需要将模型进行压缩，诸如此类的应用需求需要一整套工具来支持，这将促进具有一定壁垒性的开发工具链的打造。这意味着，开发者或用户在使用某一开发工具落实大模型项目时，就会更倾向于使用配套的框架技术。从长期上来看，这将促成框架技术的生态化发展，市场玩家将会越来越聚焦，其他技术产品的市场占有率会越来越少。目前，全球来说，TensorFlow具有广泛应用基础，PyTorch市场接受度持续上升，在产业界生产环境中继续呈现出后来者居上的竞争态势；在中国，主流的深度学习框架以百度飞桨PaddlePaddle等为代表，在学术界和工业界有较高的认知度和应用性。

2.4 AI软件基础设施：加速大模型的应用落地

人工智能持续发展离不开底层服务支撑和软件平台优化。越来越多的行业用户关注到框架或者平台产品中的大模型能力，但在通往应用和大规模落地过程中，还需不断面对算力、数据、效果、成本等多维度带来的挑战。

- **需要充足的算力资源：**大模型技术创新和应用需要基于海量数据集，在拥有成百上千加速卡的人工智能服务器集群上对千亿级参数人工智能大模型进行分布式训练，这对算力资源的规模提出了极高的要求。算力不足意味着无法处理庞大的模型和数据量，因此无法有效支撑高质量的大模型技术创新。

- **需要高效的算力供给：**相比普通的人工智能训练，大模型的训练技术考虑的问题更加复杂，对于基础设施的要求也更高。大模型训练在带来海量的算力需求的同时，还需要在算力平台设计上考虑到一个问题——庞大的算力节点规模将带来算力使用效率的衰减。大规模人工智能计算集群上的训练算力效率会直接影响到模型训练时长以及算力消耗成本。因此，如何发挥大模型算力平台效能、抑制性能损耗，对于提升生成式人工智能研发创新效率至关重要。

- **需要优质的数据服务：**优质数据集是训练高质量模型的关键，面对海量的数据规模，在大模型预训练阶段如能做到精准、高效的数据清洗、集成、变换和规约，将大大提高数据质量，减少噪音和错误数据的影响，从而有效提升算法的准确性和泛化能力。而数据标注是耗时耗力的工作，中小型人工智能企业或需要垂直模型的企业可能缺乏相应的人力、物力和时间。即使项目开始落地，由于生产过程中现场数据质量参差不齐，开发人员能力高低不一，数据处理质量会受到影响；此外，一旦项目运维过程中应用环境稍有改变，数据标注和模型训练工作就需要重做。

随着大模型从基础研发走向应用落地，人工智能软件基础设施的重要性和价值进一步凸显。大模型预训练完成了“从0到1”的技术统一，而大模型在通往“从1到100”的应用和大规模落地过程中，还需不断面对算力、数据、效果、成本等多维度带来的挑战，标准化的全栈基础软件和工作流是支撑大模型基础研发和应用落地的核心环节。

- **提供全栈全流程支持：**人工智能软件基础设施是算力和应用之间的中间层软件基础设施，包含系统环境部署、算力调度保障和模型开发管理等多个层次，覆盖了数据准备、模型训练、模型部署、产品整合等诸多环节。通过构建系统性的架构和全栈的解决方案，可更好全流程支持企业实现模型的应用落地。

- **充分释放算力资源：**大模型技术创新和应用需要基于海量数据集，在拥有成百上千加速卡的人工智能服务器集群上对千亿级参数人工智能大模型进行分布式训练，这对算力资源的规模提出了极高的要求。算力不足意味着无法处理庞大的模型和数据量，也就无法有效支撑高质量的大模型技术创新。人工智能软件基础设施需要从计算系统的各个层面充分释放算力资源。

- **实现广泛的兼容适配：**相比普通的人工智能训练，面向大模型的人工智能软件基础设施涉及的层次和范围更广，包括从集群环境的部署、管理和监控到算力和计算任务额的管理与调度保障，从数据清洗和数据管理到模型的预训练和微调等。在此过程中，必然会涉及到多种软件工具的集成整合。因此，只有打造开放、兼容、解耦的人工智能软件基础设施，才能更好地适应用户需求。

总体来说，未来人工智能软件基础设施将呈现如下发展趋势：

- 为多元化算力，提供更多无缝的兼容性适配，以应用为导向升级人工智能算力体系；

- 结合不同行业的特点做框架、模型、数据的垂直整合，支持人工智能能力在行业的落地；

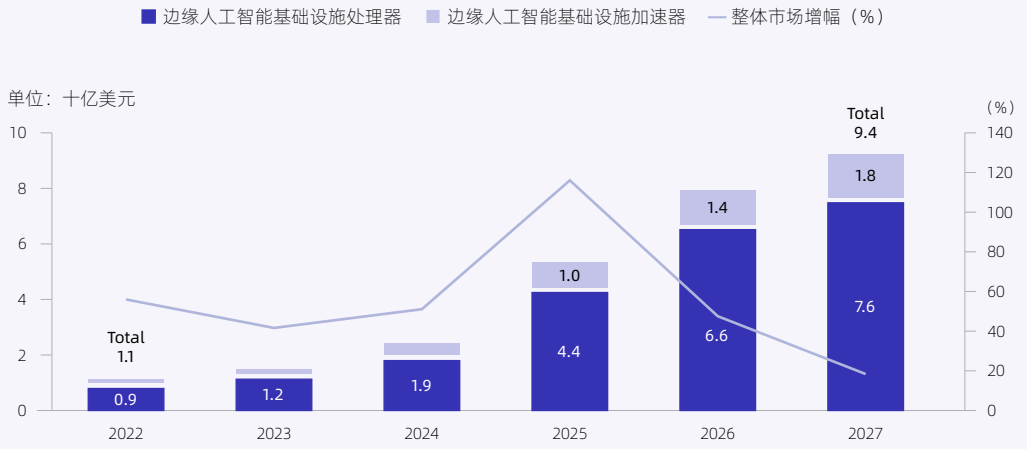
- 降低人工智能开发门槛，推进低代码或智能辅助编程工具的发展，优化API接口性能；

- 推进数据维度的开源，实现数据的开源、开放，为全社会创造价值。

2.5 边缘智能：以广泛的部署推进智能的延伸

边缘对人工智能和机器学习的依赖程度越来越高。随着边缘计算逐步进入稳健发展期，使用单一边缘技术构建的应用难以充分发挥其价值，边缘计算与云计算、5G、区块链等其他技术的协同与融合需求将进一步增加；边缘人工智能、5G边缘计算、边缘即服务等成为边缘计算技术的未来发展趋势。IDC报告数据显示，全球边缘人工智能基础设施处理器和加速器的增长将从2022年的11亿美元增加至2027年的94亿美元，五年年复合年增长率（CAGR）为52.3%。

图12 全球边缘人工智能基础设施处理器和加速器市场规模，2022-2027



来源：IDC，2023

传感器、摄像机、物联网设备和机器正在源源不断地产生更多的数据，一方面，边缘智能可以让数据在靠近创建的位置实现分析和处理；另一方面，通过云-边协作，可以将人工智能算法和模型的学习和优化成果部署到边缘侧，不断提高边缘侧的执行效果。

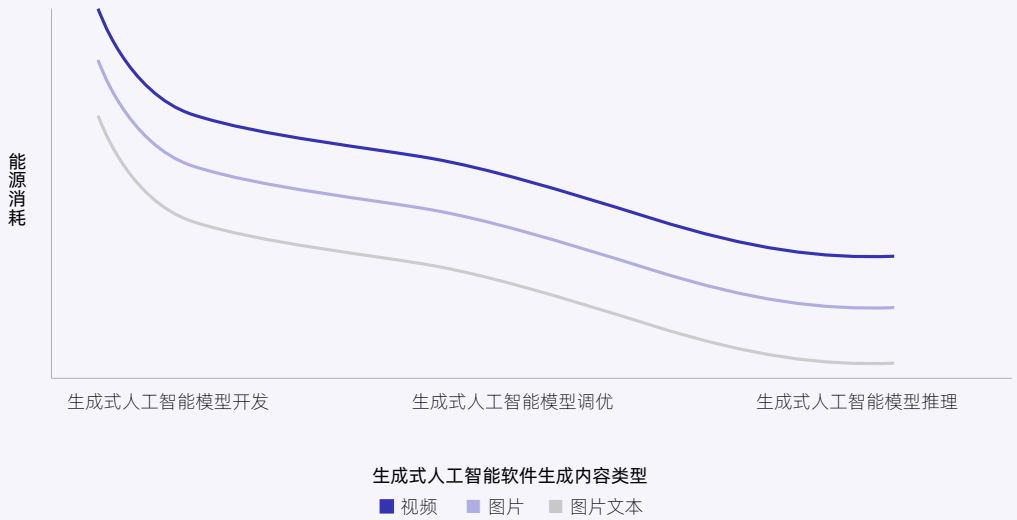
在中国，对用户而言，边缘智能具有降低时延、提高数据隐私、降低运营成本等优势，已经在工业、交通、医疗、零售、农业等行业实现广泛应用。在制造业领域，它可以为企业提高工厂设备的可维护性和可预测性。在交通领域，通过在道路沿线部署边缘智能设备，可以实现对道路交通情况的实时监测，通过与智能交通管理系统的联动，及时缓解城市交通拥堵。在科研领域，可以帮助科研人员实时监测实验过程和结果，加速实验方案的实施，提高实验的可靠性和准确性。此外，边缘智能可以通过本地加密等技术，在一定程度上满足智能应用过程中“数据不出域”的要求，更好地保护数据隐私，避免数据泄露。同时，由于边缘设备可以独立运行，即使网络连接不可靠，也可以保证业务的正常运行。

在构建边缘智能的过程中，企业需要制定全面的人工智能战略，进而更好地实现数据收集和处理，保障隐私和安全，优化模型部署和维护，提高软、硬件的集成和协同，在覆盖边缘和数据中心的异构复杂混合环境中，实现基于人工智能的创新。同时，供应商侧的相关知识和经验共享，对于企业加速拓宽边缘智能的视线也至关重要。

2.6 绿色算力：基于液冷服务器构建可持续发展数据中心

伴随人工智能算力和存力的不断提升，芯片的功耗正越来越高，发热量也越来越大。总体来说，训练阶段比推理阶段更耗电；同时，生成的内容不同，所需的功率也存在差异（通常，视频和图像等富媒体内容比单纯文本需要消耗更多的电力）。对于数据中心操作员和规划人员来说，需要依据计算需求对资源进行合理规划和分配，节约能源消耗，同时应针对高密度机架数据中心来优化电路设计，积极探索模块化设计和部署，并采用液冷等先进冷却方法、借助节能硬件和软件技术，满足可持续发展提出的要求，以通过“优先考虑数字化”和“环境、社会和治理（ESG）”战略来推动业务价值发展。

图13 生成式人工智能模型训练和推理能源消耗概念趋势图

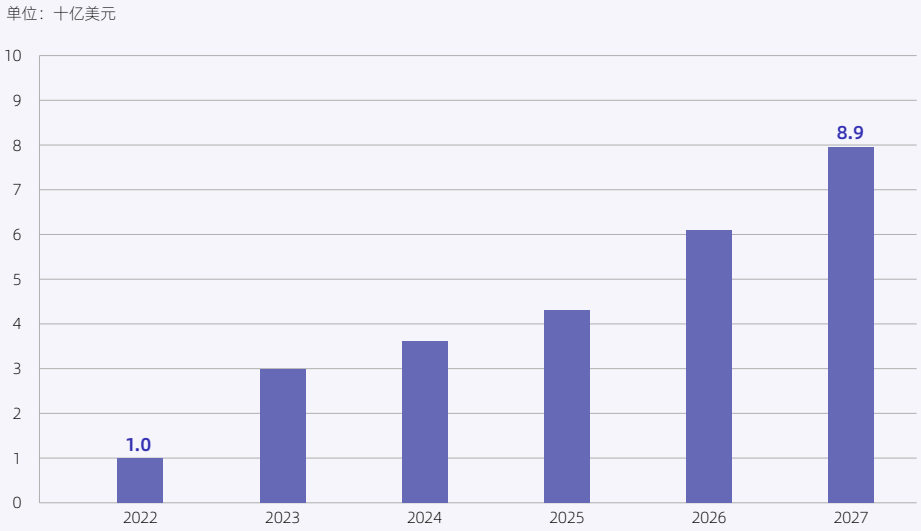


来源：IDC，2023

能耗越高的数据中心，所要付出的电力成本将越高。在中国，面对大量涌现的人工智能大模型应用以及国家“双碳战略”和“东数西算”的逐步实施，为实现算力建设和能源消耗成本间的有效平衡，通过提升散热效率来降低能耗的液冷数据中心受到市场的关注。

液冷服务器正快速迭代。IDC预计，2022-2027年，中国液冷服务器市场年复合增长率将达到54.7%，2027年市场规模将达到89亿美元。2022年，中国液冷服务器数量尚不到服务器总量的3%，渗透率在10%左右。进入2023年至今，主流IT设备厂商均已公开表明将加大研发力度并加快产品迭代速度，尤其重视液冷数据中心系统架构、液冷部件及接口、液冷基础设施、液冷监控系统的标准制定等维度的提升，加速液冷算力规模化部署，优化液冷产业链生态伙伴协同发展，加速液冷在数据中心的规模化落地。

图14 中国液冷服务器市场预测，2022-2027



来源：IDC，2023

在通用算力市场中，液冷数据中心制冷解决方案主要有三条技术路线，即冷板式、浸没式和喷淋式。IDC数据显示，2022年，中国液冷服务器市场中，冷板服务器占到了90%。这种技术路线通过冷板制冷方案，对CPU、内存和磁盘等高功耗的部件进行接触式降温。冷板方案在对原有基础设施进行改造的投入和难度方面具有优势，同时具有较高成熟度和较好商用基础。浸没式在散热效率和单机柜功率、空间利用率等方面比冷板式更具优势，但是受限于基础设施改造、建设成本、电子氟化液或其他冷却液的成本及可维护性等因素，目前发展仍相对缓慢。喷淋式与浸没式类似，同样适用于结构承重经过特殊加固的新建项目，不同之处在于：喷淋式方案中目前单机柜最大负载为48KW，应用范围相对狭窄。

越来越多的行业标准相继推出，也在带动生态链的有序发展。Intel等全球技术供应商提出在OCP（开源计算项目）下为数据中心液冷用快速接头制定UQD全球标准，希望能在液冷数据中心防喷快换接头的快速更换方面提供开放标准，以提高液冷数据中心的快速部署能力以及减轻运维难度。在中国，尽管市场对数据中心制冷有旺盛的需求，但液冷技术从宏观上看仍处于发展早期阶段，产业生态建设仍有待完善。各方应全力打造高水平液冷生态链，构筑开放生态，引领形成统一标准，打造液冷生态的主导者、设计者、构筑者，推进产业生态成熟。

2.7 人工智能算力服务和云：根据算力需求优化服务模式

近年来，人工智能的广泛应用提出了更高的算力需求，也让算力提供的使用方式发生了重大改变。过去十年，企业IT基础架构的部署从传统采购模式加速向公有云上迁移，越来越多的用户开始利用人工智能aaS服务带来的便利、快速灵活地部署相关应用。然而近年来，人工智能和应用提出了更高的算力需求，让算力提供方式发生了重大改变。人工智能应用带来的算力使用方式具有算力资源占用集中化、技术门槛更高等显著特征。一个大模型的训练往往需要大量集群的算力提供支持，传统人工智能就绪度低的数据中心环境和运维管理状态难以满足算力、网络和安全等方面的需求；应用侧需求的改变也将催生算力服务模式的更新。

2023年，大模型和生成式人工智能快速发展，将给人工智能算力服务市场带来新的机遇：

- 其一，不管是大模型的训练或推理，都需要相对更大的算力和相应的投入。当前，生成式人工智能正处在起步阶段，主要的投入来自超大规模互联网企业，随着应用在各行业的不断深入，更多的用户将涉足这一领域，而对于短期内不考虑或不具备自建人工智能算力数据中心能力的用户而言，使用人工智能算力服务的方式是理想的选择。

- 其二，超大规模云服务器提供商和人工智能解决方案提供商具有更强的技术能力，具备生成式人工智能和大模型开发的技术基础，并有能力进行快速迭代。因此，人工智能算力服务能够帮助各行业的中小企业实现生成式人工智能技术的可用，从而快速为自身业务发展赋能。

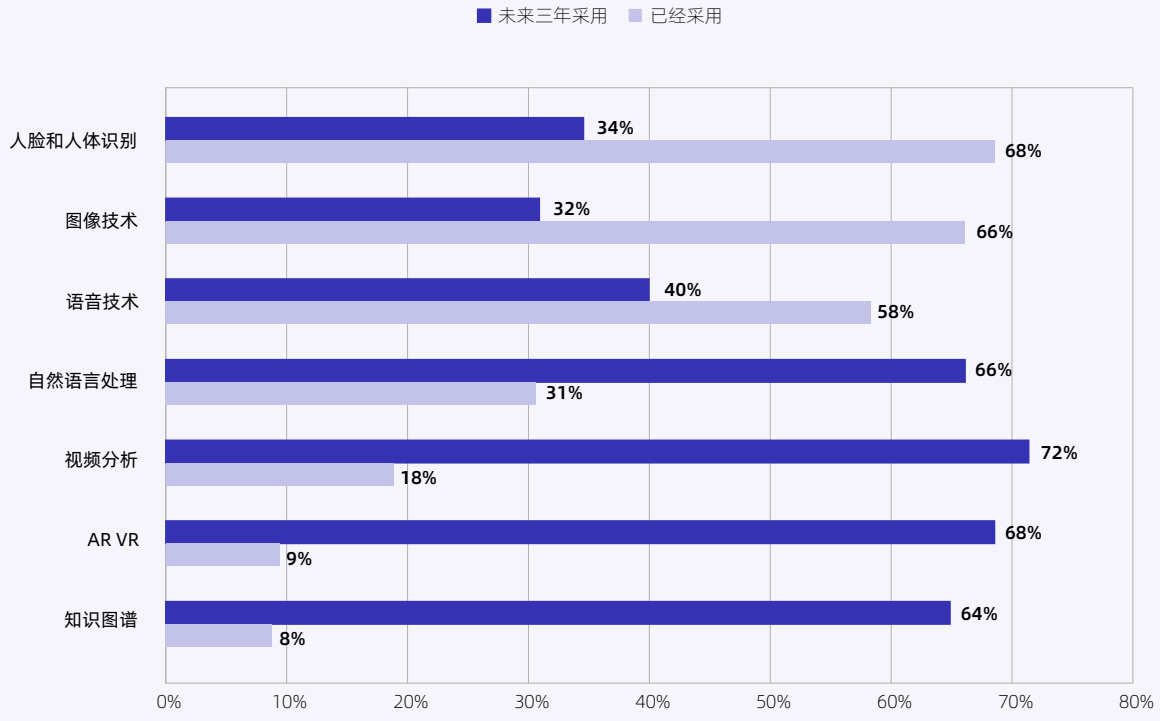
未来，人工智能算力服务的发展将与人工智能生态的发展息息相关。过去几年来，人工智能在数据、算法和算力方面日益成熟，行业应用更加丰富，加之人工智能产业链各个厂商做出的贡献，使得人工智能算力服务的应用场景更加丰富和成熟。当前用户最大的痛点在于：面对偏通用的模型，如何将人工智能技术更好地应用到自身企业业务场景，“最后一米”的距离看似很近，实则需要各个人工智能生态中合作伙伴大量的技术和时间投入，同时降低开发门槛，突破这些瓶颈，让人工智能应用更好地赋能各行各业。

2.8 应用：企业积极投入以满足大模型时代的应用需求

近年来，人工智能从单点应用到多元化、从通用场景到行业特定场景，不断深入，飞速发展。由生成式人工智能引发的热潮，也在迅速扩散。人工智能应用场景在近年的发展中，逐步开始变得更加多样化，对人工智能的需求也逐渐从单一功能转向为多维度，多思维，多模态以及多场景。

在人工智能单点技术应用方面，根据 2023年 IDC 针对企业对于人工智能技术的应用现状调研的结果，计算机视觉仍为最主要的应用技术类型，以生物识别和图像技术为主，语音技术的应用程度紧随其后，而自然语言处理仍处在相对早期的发展阶段，从调研的样本来看，已经采用的企业不超过三成，但从未来三年计划采用的情况来看，自然语言处理类应用将快速落地，66%的企业表示将在未来三年采用该应用场景。

图15 企业已部署及未来三年计划部署的人工智能单点技术，2023



来源：IDC，2023

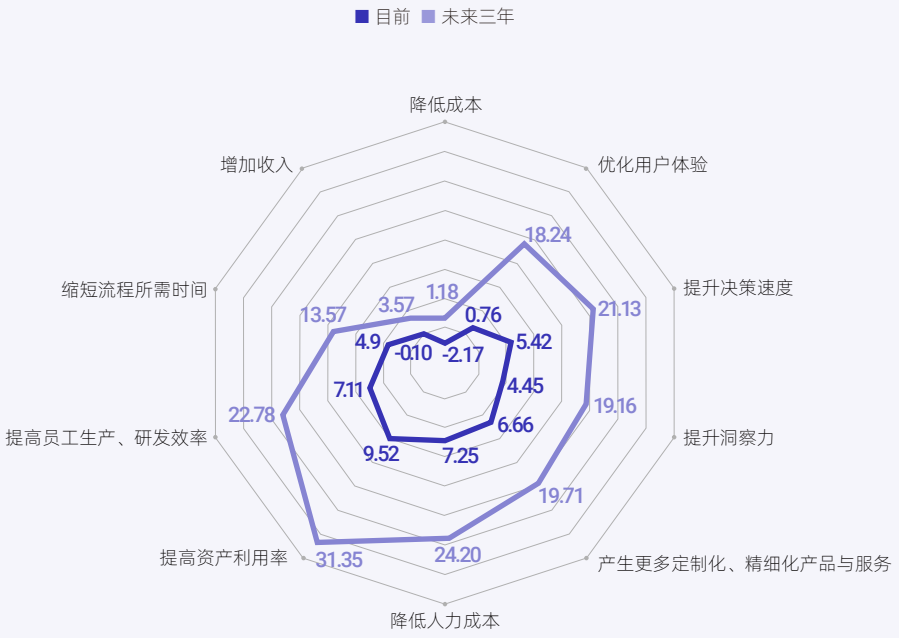
同时，本次调研结合单点技术应用及企业IT资源消耗发现，自然语言处理对企业IT资源占用最大，未来随着这一应用普及度的提高，该特征将更为显著。除此之外，图像技术、知识图谱和AR/VR是参与调研样本企业中另外三个高算力消耗的单点技术。

图16 单点技术场景对IT资源的占用情况

来源：IDC，2023

对于企业而言，人工智能对企业带来的价值正愈加显著，尤其体现在提高资产利用率、提高员工生产及研发效率、提升产品与服务三方面。IDC本次调研显示，企业在未来三年由人工智能所产生的价值将大幅提升，尤其在提高资产利用率、降低人力成本、提升洞察力、提升决策速度和优化用户体验等几个方面。

图 17 人工智能目前及未来三年对企业产生的价值



来源：IDC，2023

从行业的角度看，与往年相比，人工智能技术已全面迈入应用加速时代，持续追求对于技术的创新，更注重类人化和机器学习、深度学习等人工智能技术，分析处理真实数据以及复杂问题。生成式人工智能的落地，也给各行业带来新的动能。

- **智慧金融行业：**金融行业对人工智能投入高，增速快。人工智能技术的引入也打破了传统金融行业存在的流程复杂、周期较长、工作单一且重复、审批方式以及流程耗费时间长等问题。因此，金融业对于人工智能算力、算法的需求日益增高。以银行业为例，伴随当前国内经济大环境以及宏观调控等因素，信贷及风险管控压力骤增，实现风险管控，反欺诈以及基于RPA的流程自动化是银行目前的布局重点。各家银行依托大数据建立专属的信贷评级系统以及审批系统，可极大程度上降低借贷风险，并优化借贷流程。近年来，银行业已在风控、营销、客服等多个场景熟练使用人工智能技术，有效提高了工作效率，降低了各项成本，提升了客户服务满意度，同时实现了“便捷化”、“智能化”及“绿色化”；随着人工智能技术的发展，业务流程自动化管理是银行业发展的重心和趋势，其能够有效减少信息录入、核验、审批等简单重复的工作，加速个性化金融服务的拓展，从而减少运营成本以及提高客户满意度。

- **智能制造：**人工智能的革新以及算法、算力的更新迭代加速推进了制造业的发展进程。基于海量大数据，基础设施全面实现智能化、绿色化，在避免消耗大量人力成本的同时也助力技术的不断创新和升级。智能产线、预测性维护、自动化生产、运营优化和实时监控是目前制造企业使用率最高的应用场景。未来，智能决策、远程操作，以及围绕生成式人工智能的应用场景将加速落地。

    - 智能产线：在人工智能技术的赋能下，制造业企业可以实现产品性能和质量的可视化预测以及自动化复检；制造业推进供应链管理自动化，同时了解内外部发展趋势，结合市场发展现状、自我产品定位以及相关竞品的优劣势对产品进行定位和研发，协助制定全新设计方案和销售方案。

    - 预测性维护：通过收集和分析设备数据，人工智能可以预测设备故障的可能性和维护需求。这种预测性维护可以减少设备停机时间，提高生产效率。

    - 自动化生产：人工智能可以控制和优化生产线的自动化过程。例如，使用机器学习算法来优化生产计划，调整机器参数和改进供应链管理，以提高生产效率和产品质量。

    - 运营、工况分析和优化：通过对大量数据的收集和分析，人工智能可以揭示生产和运营中的潜在问题并提供优化建议，帮助制造厂商优化生产过程，降低成本并改进产品设计。

    - 实时监控：通过传感器和物联网技术，人工智能可以实时监测设备和生产线的状态。通过收集和分析大量数据（如温度、压力、震动等）以及设备的工作状态和性能指标，帮助制造企业及时发现问题并采取措施，以避免生产中断或出现质量问题。

    - 智能决策：人工智能可以通过学习和优化算法，实现设备和生产线的自主决策；根据实时数据和预设的规则，自动调整参数，优化生产计划，及时响应异常情况，有效提高生产线的自适应性和效率。

    - 远程操作：人工智能技术可以通过远程控制技术，实现对设备和生产线的远程操作。制造商可以通过云平台或移动应用监控和控制设备，进行远程调整和优化，极大程度提高生产线的灵活性和响应能力，减少人工干预，降低生产成本。

- **智慧医疗：**人工智能在医疗行业的应用正日益增长，对医疗服务的质量、效率和准确性产生了深远的影响。例如，人工智能在医学诊断和影像学方面发挥了重要作用。通过深度学习技术和大数据分析，人工智能能够对医学图像，如CT扫描、MRI和X射线等进行自动分析和解读，有助于辅助医生提供更准确的诊断，快速识别潜在的疾病特征，减少漏诊和误诊的风险。此外，人工智能可改善对患者的监测和个性化治疗。通过传感器、健康监测设备和实时数据分析，人工智能能够监测患者的生理参数、活动水平和病情变化，并基于这些数据提供个性化的治疗建议、预测疾病发展趋势，并提前预警患者的恶化风险，从而优化患者管理、提升治疗效果。另外，人工智能在新药研发和药物治疗方面也具有潜力。通过分析大量的医学文献、基因组学数据和临床试验结果，人工智能可以发现新的治疗目标和药物候选物，加速新药研发的过程，并帮助制定更有效的治疗策略。人工智能能为患者提供更智能化和个性化的医疗服务。通过虚拟助手、聊天机器人和语音识别技术，人工智能可以为患者提供24/7的医疗咨询和居家远程医疗服务；还可以根据患者的个人健康数据和历史记录，提供个性化的预防和康复建议，促进患者的健康管理和自我护理。

- **AI4S（人工智能应用于科学）**：人工智能在科学领域的应用也取得了诸多阶段性成果。基于先进的算法和模型，科学家、研究机构和相关企业可加速数据分析和处理，基于模型和算法对原子运动规律、物质性质等进行预测和模拟，也可对医学图像、天文图像等进行更好的识别和理解，加速实验的自动化和智能化，实现自动化合成、自动化表征等。目前，人工智能已经在材料科学、生物医学、环境和气象、海洋、航空航天、化工等领域实现落地和应用，为科学研究的发展带来更多的可能性。

生成式人工智能应用在2023年快速发展，未来将进一步赋能各行各业。由于生成式人工智能可为实际应用注入增强功能，引来商业领域的广泛关注。通过把大模型能力和应用需求结合，结合场景或业务数据，可加速生成式人工智能向行业领域的渗透。IDC认为知识管理、对话式应用、销售和营销、代码生成等是企业应用生成式人工智能的主要领域。

- 从业务职能场景来说，通过将一个模型（或多个模型）与企业数据集成，可供特定业务部门或职能部门（营销、销售、采购等）使用，满足职能场景的业务需求；

- 从生产力场景来说，通过具体工作任务设计，如生成文案、图片、视频等内容，加速软件开发，可将生成式人工智能功能注入现有应用，提高生产力；

- 从行业场景来说，基于足够大的训练数据集或与行业生态伙伴合作，共享数据，围绕具体模型，以定制化方式构建特定集成架构，满足行业需求。

结合用例来看：

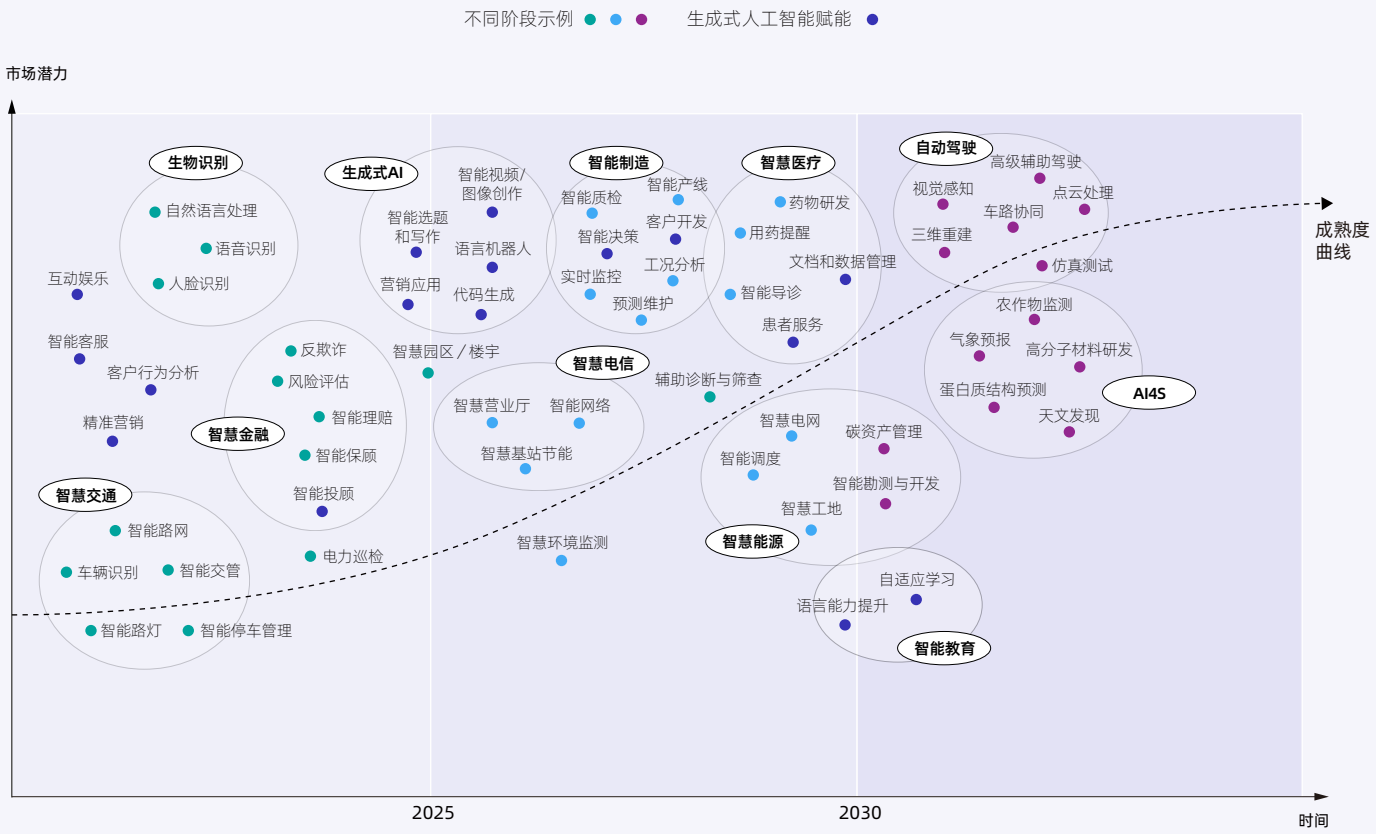
- 金融行业是对生成式人工智能使用较早的行业之一，IDC全球一项调研结果显示，样本中超过半数的金融机构计划在2023年在生成式人工智能技术上进行投入，只有10%的金融机构表示，他们目前没有试验计划，或者目前没有使用生成式人工智能的计划。在金融行业脱颖而出并最容易实现的用例多数围绕用户体验、知识管理和应用开发这几方面，包括智能投顾、自动化客服（如聊天机器人和语音机器人）、风险评估、报告自动化生成、代码生成应用等。生成式人工智能在金融行业有着巨大的前景，但金融作为监管最为严格的行业之一，对数据安全和隐私有极高的要求，本次参与调研的金融机构表示，对数据安全和隐私的顾虑是他们采用生成式人工智能时的最大阻碍，如何解决这一问题，对未来技术的应用发展至关重要；

- 在制造业领域，生成式人工智能应用处于早期阶段，但由于其能够为客户生成个性化内容和营销体验，在改善客户体验方面具有巨大潜力。主要用例涉及潜在客户开发、生产流程优化、产品设计辅助、设备预测性维护、供应链管理等。以潜在客户开发为例，制造业的客户开发通常涉及多个接触点，包括网站、邮件、推荐和社交网络。生成式人工智能可以分析线索，一旦产生潜在客户，就通过评估大量数据来考量客户质量和潜力；如果潜在客户与制造企业理想的客户画像相匹配，就能进行资格预审，从而帮助企业获得优先权。在制造行业应用过程中，生成式人工智能需要解决的关键挑战包括：数据的可用性和质量，客户对销售人员参与度的期望，以及数据和隐私法的要求。总的来说，人工智能应用可以大幅度提高制造业的效率、质量和创新能力，带来更多的机遇和竞争优势；

- 在医疗领域，医疗机构也正积极地了解与生成式人工智能相关的应用，主要涵盖五大领域：临床文档和数据管理、患者服务、工作流程和资源优化、员工支持、医疗保健服务和商业应用。相较于金融和制造业而言，医疗行业在生成式人工智能的投入占比相对较低，但其未来潜力不容忽视。

此外，生成式人工智能与大模型也在能源（能源消费趋势预测、能源存储和分发、环境监测）、零售（库存管理和预测、个性化推荐、客户行为）、教育（语言学习、教学辅助、学习评估）等领域实现推进。随着技术的不断发展和进步，未来还会出现更多新的应用场景。

图18 中国人工智能应用场景发展，2023



来源：IDC，2023

最佳实践

案例一

美图：深耕生成式人工智能，提前布局，顺势发展，广泛赋能

美图是一家成立于2008年的中国科技公司，秉承着“让科技与艺术美好交汇”的使命，美图公司致力于打造优秀的影像与设计产品，让图片、视频、设计、数字人的制作变得更简单，并通过美业解决方案助力产业数字化升级。凭借其多年积淀的人工智能技术能力以及对用户需求及市场趋势的敏锐洞察，美图公司在中国以及其他国家和地区的用户积累量已达亿万级。

2010年，为了进一步开展计算机视觉、深度学习、机器学习等人工智能相关领域的研发，美图成立了美图影像研究院（MT Lab），积累人工智能技术能力，提升研发能力和学术影响力。同时，通过与产品、设计等团队形成高效协作机制，以技术和业务紧密结合的方式共同推进先进人工智能技术在产品中的落地，以核心技术创新赋能企业业务发展。

通过发挥人工智能的技术优势，美图将AIGC相关能力成熟应用在图片、视频、设计、数字人等应用场景中，覆盖视觉创作、商业摄影、专业视频编辑、商业设计等领域，不断推出新功能，旨在全面提升影像行业的生产力。

- 图片场景：利用AIGC能力，实现图生图、文生图，以及AI绘画、AI写真、AI人像、AI发型、AI精修、AI妆容等一系列功能。
- 视频场景：利用AIGC能力，实现视频的智能化生成和渲染，其中AI动漫功能通过提升帧与帧之间的连续性，保证生成效果的稳定性，精准实现视频人物的表情和姿态控制。
- 设计场景：在应用中具备AI海报、AI商品图以及AI模特试衣等功能，满足电商行业海报生成、商品图展示和模特换装等场景的智能化需求，提升工作效率的同时，控制成本支出。
- 数字人场景：基于美图AI演员、AI主播等数字人功能，呈现更加逼真的效果，为影视制作等行业提供数字生产工具。

2023年6月，美图正式对外发布自研AI视觉大模型MiracleVision（奇想智能），为美图全系产品提供人工智能模型能力，也助力美图形成由底层、中间层和应用层构建的人工智能产品生态，降低使用门槛、满足专业设计需求的同时，逐步落地电商、广告、游戏、动漫、影视等行业，助力行业用户工作流提效。

大模型的研发和应用对基础架构提出更高的要求，美图通过与浪潮信息等设备服务商合作，升级基础架构能力，提高算力、存储、网络性能，满足视觉大模型的训练和应用需求。同时，发挥诸如云原生服务的基础构建等能力，提高模型训练平台的利用率，助力整个高算力集群性能的稳定、高效输出。

2023上半年，美图完成了由生活场景到生产力场景的进化，并持续推进商业模式创新，逐步为生产力场景中的设计师用户和企业级用户提供更优的产品体验。例如，美图旗下产品美图设计室利用AI技术为众多商家提供AI商拍服务，助力近百万中小电商卖家降本增效。

在AIGC助力下，美图月活跃用户数和会员数持续增长，以订阅为主要变现模式的影像与设计产品业务收入继续攀升，为利润提升做出贡献。2023年上半年企业营业收入和利润均实现明显增长，同比增幅分别达29.8%和320.4%。月活跃用户数达2.47亿，同比增长2.5%。同时，企业内部新产品开发速度与功能研发速度也得以提升。

接下来，美图将持续迭代自研大模型和AIGC产品能力，从算法、算力、数据三维度夯实人工智能发展基础，做好技术人才储备，与合作伙伴密切协作，推动人工智能对企业产品能力、服务能力的提升。

案例二

网易：沉淀关键技术能力，推进人工智能能力的场景化应用

网易是一家成立于1997年的中国互联网科技公司，业务范围覆盖游戏、音乐、电子邮件服务、电商平台以及教育等领域。随着智能创新时代的来临，网易看到大模型和生成式人工智能发展给企业、行业和社会带来的机遇。在此背景下，网易一方面加速关键技术的研发和突破，以生成式人工智能技术加速业务发展，另一方面加速人工智能能力的场景化应用，推进技术对行业和产业的赋能。网易伏羲是网易旗下专业从事游戏与人工智能研究和应用的机构，是网易发展人工智能技术的一个重要缩影。

伏羲采取以应用为核心的策略，加速人工智能在实际生产环境中的应用和落地：

- 设计创作场景：网易伏羲“丹青约”绘画平台面向美术资产生产管线，旨在辅助美术工作者解决设计灵感创作问题，加速图片作品生成，提升创作效率，实现至少50%以上的工作效率提升。

- 元宇宙场景：基于网易瑶台元宇宙场景搭建编辑器，实现多场景、强互动、沉浸式的虚拟空间搭建，在虚拟商场、虚拟娱乐场景以及虚拟旅游等方面实现成熟落地。

- 工程制造场景：基于网易有灵智能体平台，完成数据训练、建模、以及智能体构建等任务，将智能能力迁移和应用到工程制造场景当中，打造出具备“自我进化”能力的设备。

在实现人工智能认知和决策能力的行业应用过程中，企业往往面临门槛高、周期长、应用落地难等问题。伏羲有灵平台作为人机协作任务平台，为用户提供AOP（Agent-Oriented-Programming，面向智能体编程）框架，通过提供统一对接人工和机器的规范接口和服务，支持用户通过平台自动构建业务领域的数据闭环，可针对行业细分场景定制化构建并训练领域特殊的智能体。此外，平台为开发者统一解决了复杂分布式系统的计算、网络和存储等资源问题，并支持将云端资源延展和下沉到边缘端，为开发者提供极简编程体验。

与传统方案相比，在数据构建和能力发展方面，有灵平台展现了明显的先进性。伏羲正将其技术试用至智能制造，包含无人装载机、工程机械以及家用机器人等多元领域，通过模型即服务（MaaS）的形式，向企业级用户提供服务，以解决实际生产问题。

生成式人工智能和相关模型训练提升了网易对于人工智能算力的需求。在技术研发和落地应用过程中，网易与浪潮信息合作，满足对大规模人工智能训练和推理算力的弹性需求。此外，基于浪潮信息对于计算业务的深入理解以及其在模型开发领域的实践经验，双方充分沟通，在数据资源治理、训练技术开发等方面深入合作，加速了网易一些算法业务的大规模启动的进程。

案例三

深势科技：以全栈技术能力加速人工智能与科学领域的融合发展

深势科技专注于AI for Science（AI4Sci）科学研究范式的探索和突破，通过运用人工智能和分子模多尺度的模拟仿真算法，结合高性能计算手段，加速科学问题的研究，为生物制药、能源、材料等领域的发展，提供从底层算法到工业软件及行业解决方案和场景落地的一揽子服务，以科学研究和应用的实际需求为导向，加速人工智能对科学领域的赋能。

传统上，无论是微观尺度的量子物理、分子动力学仿真，还是宏观的流体和气象仿真，都依赖于大量的算力。深势科技首席科学家张林峰博士及团队基于深度神经网络，对网络施加物理约束，并使用少量基本数据进行训练，从而开发出训练后的分子动力学软件。这种方法在相同的精度下，能够将训练时间降低十万倍，并可以进行更大尺度的模拟，该研究成果获得了2020年ACM戈登贝尔奖。

AI for Science 将为科研机构和企业节省时间和资金双重成本：以药物领域为例，在进行大规模虚拟筛选时，新算法可比旧算法快1600倍。这使得研究团队能够进行更大规模的筛选，从而有更高的几率找到合适的候选药物分子。通过这样的虚拟筛选，可以减少实验阶段所需的药物分子数量，从而降低成本。

目前，深势科技已经推出科研云平台、药物计算设计平台、难成药靶标研发平台及电池设计自动化平台工业设计与仿真基础设施。其中：

- Bohrium®科研云平台帮助研发人员进行光学、电学、磁学、力学等物理性质计算，并支持对材料微观结构组分与作用机理研究，满足多场景科学研究计算需求。

- Hermite®药物计算设计平台及RiDYMO®难成药靶标研发平台专注于临床前药物研发提供一站式计算解决方案，实现计算驱动的药物设计，覆盖从蛋白质结构建模、分子对接与虚拟筛选、结合亲和力评估，到基于大模型的分子性质预测等完整药物研发过程。

- Piloteye™电池设计自动化平台通过采用AI-Native和多尺度模拟方法，搭建从材料微观机理到宏观电芯电化学，到面向能源电池全生命周期研发的自动化设计平台（BDA），加速电池从材料到电芯的设计过程，提升了电池设计和研发的效率和创新性。

AI4Sci将为科学研究带来新的范式和机遇。它不仅可以帮助科研机构和企业解决现存的科学问题，还可以实现对新课题的启发，发现新的问题和方向。新算法的发展、大模型的涌现、数据的快速增长和计算能力的提升为AI4Sci的发展奠定基础，深势科技在算法、平台、应用、解决方案等多维度的突破将为更高效、准确地解决科学问题提供有力支持。

第三章
中国人工智能计算力发展评估

3.1 行业排名

回顾2018年到2022年的评估结果，可以看到互联网、金融、政府、电信、制造、服务、交通、医疗、能源和教育等领域一直在积极探索人工智能的应用。其中互联网、政府、金融、电信、制造行业在五年中一直保持着前五名的高渗透度。随着人工智能技术的不断发展与完善，企业决策者更深刻地体会到人工智能可以推动传统产业的升级和创新，促进产业的转型加速，同时也能催生新的产业和商业模式，为传统企业和初创企业带来更多的机会。因此，各行业对于人工智能技术的应用愈加重视，行业分布也愈加广泛。

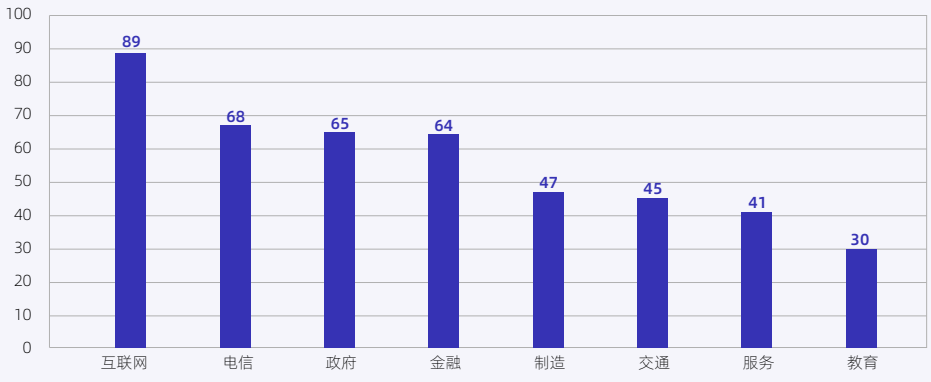
图19 中国人工智能行业应用渗透度，2018-2023

2018	2019	2020	2021	2022	2023	排名
互联网	互联网	互联网	互联网	互联网	互联网	1
政府	政府	政府	金融	金融	电信	2
金融	金融	金融	政府	政府	政府	3
制造	制造	电信	电信	电信	金融	4
服务	电信	制造	制造	制造	制造	5
电信	服务	服务	服务	服务	交通	6
教育	教育	医疗	交通	交通	服务	7
医疗	医疗	教育	医疗	医疗	教育	8

来源：IDC，2023

2023年人工智能的行业渗透度排名Top5的行业依次为：互联网、电信、政府、金融和制造。此外，交通、服务、教育等行业在人工智能领域的投资力度也可圈可点。

图20 2023年中国人工智能行业应用渗透度



来源：IDC，2023

互联网

在过去的几年中，尽管人工智能在互联网的投资增速有所放缓，但在中国，互联网仍是生成式人工智能技术应用和研发的主战场。从自然语言处理到图像识别，再到智能客服，人工智能的技术不仅提高了用户的体验，也极大地提高了服务质量。如今，互联网行业更为重视人工智能技术。自然语言处理，如语音识别、文本生成和机器翻译等，已被广泛应用。在图像识别方面，深度学习技术使得机器能够识别和理解图像中的各种元素，如人脸识别在安全和身份验证中的应用，以及在智能驾驶中的障碍物识别技术。

互联网企业将各种人工智能技术广泛应用于其业务模块，以优化用户体验并提升商业运营效率：智能客服借助自然语言处理和机器学习技术，为用户带来更为高效和流畅的服务，同时使企业服务效率更高、成本更低；大数据分析与智能决策也逐渐成为互联网企业的核心能力，通过基于智能搜索技术提取有价值的信息，从而助力企业做出更加智慧的决策；在营销领域，智能营销和个性化推荐技术，使得互联网企业能够更准确地洞察用户的兴趣和需求，增加客户粘性，向客户提供更有针对性的帮助。可以说，中国的互联网行业正在经历一个深度的智能化革命，由人工智能、大数据、自动化和其他先进技术共同推动，旨在更好地满足市场和用户的多样化需求。

电信

人工智能技术正深入地渗透到电信行业，各大运营商都在积极探索其潜在价值。凭借海量、多样且真实的用户数据，这些运营商已经构建了高价值的人工智能训练数据集，为电信行业未来的智能化发展打下了坚实的基础。一方面，中国运营商加速云数据中心建设，积极部署云上智能化能力，加速数据中心业务发展；另一方面，加速基础设施的智能化建设，支持电信网络的建设和优化。例如，网络智能化已使得资源管理、网络优化和故障诊断更为高效，能够根据实时数据动态调整，以提升网络的性能和用户体验。此外，运营商越来越多地依赖人工智能技术改进自身服务，如在智慧营业厅的建设中，采集的消费者行为数据如停留时间等，都会被智能化地分析并及时做出相应的反应。此外，防诈骗也得到了加强，电信企业利用大数据和人工智能技术大大降低了电信诈骗的可能性。同时，运营商也投入资金到提供智能化算力的平台，以获得针对各种高技术场景的解决方案。至于数据智能化，借助对通话记录和网络行为的深度分析，企业能够更深入地了解用户需求，并为他们提供更加个性化的服务。总的来说，人工智能在电信行业的广泛应用，不仅提升了客户体验，还大幅增加了运营商的业务营收、降低了运营成本，并推动了整体效率的提升。

政府

在现代数字化时代，人工智能技术在政府领域的应用日益显著，极大地推动了公共服务和城市治理的进步。例如，数字政府已经利用自然语言处理和机器学习算法，推出了智能客服系统，使公众能够获得快速、准确的在线信息查询与答疑服务。数据挖掘和预测技术也使数字政府能够深入分析海量政府数据，找到潜在的模式，进而为决策提供科学依据。在城市管理方面，结合物联网、大数据和人工智能技术，政府正朝着智能城市管理的方向发展，优化城市基础设施、交通、环境的监测与管理，确保城市的安全、便捷和可持续性。数字政府还在积极推进办公流程与服务的自动化，通过机器人流程自动化（RPA）等工具，极大提升了工作效率和公众的满意度。值得注意的是，这些应用只是目前数字政府采纳的一部分，随着人工智能技术的不断演进，未来无疑会有更多的创新应用涌现。总的来说，通过这些人工智能应用，数字政府不仅提供了更便捷、高效和智能化的公共服务，还显著提升了公众满意度和城市治理水平。

金融

人工智能在金融行业的应用迅猛增长，已经渗透到诸如银行、投资机构及保险和证券等各个领域中，主要应用包括智能客服、实体机器人、智慧网点和云上网点等，为各机构提供了更好的客户体验和高度的便利性。基于人工智能，金融企业利用机器学习算法进行数据分析，根据客户的风险状况进行准确评估，实现了贷款审批流程的自动化，确保贷款的准确性和效率。金融科技应用程序的开发者也正在积极地将更多功能——如EMI计算器和贷款资格自我评估等，集成到人工智能和机器学习技术中。此外，通过引入人工智能技术，金融机构现在可以实时监测交易和市场波动，从而及时制定策略。大量的数据，如交易、用户行为和市场数据，现在都可以通过大数据分析和人工智能技术进行深入挖掘，为风险评估、投资决策和市场预测等提供更强大的支持。为了进一步提供个性化服务，金融机构采用机器学习和推荐系统，为客户提供更符合其需求的投资方案和贷款产品。再者，金融机构也开始利用区块链技术实现更快、更安全、更透明的交易和结算，大大提高了交易效率并减少了中间环节。此外，金融企业与科技公司加速合作，通过技术合作、数据共享和创新孵化等方式，推进智能化进程，推动金融行业向智能化、自动化和个性化方向发展，为整个行业带来了创新和增长的机会。

制造

人工智能正在改变制造业，为其带来工业4.0和工业互联网时代的技术变革。通过结合大数据、物联网和其他先进技术，制造业实现了生产的自动化、智能化和灵活化，进而提高了生产效率。人工智能在制造业中应用包括但不限于：交互界面的智能化，质量管理，维修与生产检测的自动化，以及供应链管理的智能化。无人驾驶技术已进入汽车制造业，人工智能的感知、决策和控制能力促进了车辆间的智能协作，可以提高道路安全和交通效率；工业机器人也在改变汽车制造业，带来智能制造的革新，实现更高效的生产线布局和生产流程。机器视觉技术被广泛应用于产品的智能质检，对产品进行分类和检测；例如，在手机制造中利用图像识别技术，通过亮度、均匀度以及像素纹理等指标对屏幕素质进行检测。对于汽车的维修与保养，智能化系统能够自动检测故障并生成维修方案，以及通过数据分析来预测零部件的使用寿命。制造业也利用大数据分析和人工智能技术来优化生产过程，实现生产的实时监测和分析，以此来优化生产过程，预测设备故障，提高产品质量等。利用物联网和大数据技术，制造业更进一步实现了供应链的智能管理，通过可视化、协同化、智能化，提高了供应链的响应速度，降低了库存成本，大大增加物流管理的灵活度。

此外，在**交通**领域，人工智能技术也正在带来深刻的变革。以物流为例，该行业已经大量采用了智能技术来提高效率和用户体验。智能物流已经成为了发展的趋势，包括智能仓储和智能配送。在智能仓储中，货物的自动存储、拣选和库存管理的自动化，都极大地提高了仓库的运营效率，同时降低了人工成本和管理的难度。智能配送方面，则采用了多模式运输规划，无人机和机器人技术，以实现物流运输路线的优化，确保货物能够在最短的时间内，以最低的成本送达目的地。此外，汽车制造商为了增强驾驶安全和便捷性，已经开始在其产品中实施半自动驾驶功能，例如先进的驾驶辅助系统（ADAS），从而帮助用户提高使用体验，如智慧入库停车、恶劣天气控制车辆避免碰撞等。除了私人用车方面的应用，人工智能技术在公共交通管理和道路监控方面也扮演了关键的角色，确保道路交通得到智能化管理，从而大幅提升整体交通效率。

在**教育**领域，科研院所和高等学府正在深度融入智能化技术以提升其教育与管理质量，提高科研创新能力，通过机器学习、深度学习、大模型等人工智能技术，基于对模拟和真实科研数据的分析，加速计算推演，帮助基础学科和应用学科更快、更准确地进行假设的验证和应用。

在**医疗**行业，尽管当前只有少数医疗人员参与人工智能的开发与应用，这主要是由于相关标准与规范尚处在完善阶段，但随着中国政府对人工智能法规的日益完善，未来五年中，其应用势必将得到快速的扩展。医疗数字化转型中，传统模式向互联网医疗模式的转型趋势显而易见。如人工智能和大数据等新技术，已经促使疾病的诊断与治疗模式从单一领域扩展到多个领域，从而推动了医疗信息化的整体升级。预计在未来，受到政策支持和市场需求的双重推动，生物方向的人工智能应用将展现出巨大的发展潜力。

3.2 地域排名

在过去五年中（2018-2022），北京、杭州、上海、深圳、广州、合肥、苏州、重庆等城市在人工智能领域具有较为突出的表现。可以看到，北京在过去几年中一直稳居人工智能算力发展城市排名的前列，拥有大量的人才和成熟的企业，同时政策扶持也非常有力；上海、深圳、杭州等城市不断加速技术积累，拓展应用场景，构建具有特色的人工智能发展路线。其他城市的排名有所波动，但可以看到，越来越多的城市正加入到人工智能发展的浪潮中。

图21 中国人工智能计算力发展评估——城市排行，2018-2023

2018	2019	2020	2021	2022	2023	排名
杭州	北京	北京	北京	北京	北京	1
北京	杭州	深圳	杭州	杭州	杭州	2
深圳	深圳	杭州	深圳	深圳	深圳	3
上海	上海	上海	南京	上海	上海	4
合肥	广州	重庆	上海	广州	苏州	5
成都	合肥	广州	苏州	苏州	广州	6
重庆	苏州	合肥	广州	合肥	济南	7
武汉	重庆	苏州	济南	济南	合肥	8
广州	南京	西安	成都	天津	重庆	9
贵阳	西安	南京	合肥	重庆	成都	10

来源：IDC，2023

2023年，中国人工智能城市评估排行榜中，北京依然位居首位，杭州和深圳分别位列第二位和第三位。其中，北京在大模型领域表现突出，聚集了大批大模型企业，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，位居TOP10的城市还有上海，苏州，广州，济南，合肥、重庆和成都。

整体来说，排名靠前的城市因具有更好的政策、资金和技术支持，可以稳定吸引更多的人才和企业聚集，从而形成更强大的人工智能产业集群，而排名相对靠后的城市依然保持着对人工智能产业的热忱，不断地推进人工智能产业的发展，挖掘出具有地方特色的发展路径。在这个过程中可以看到，智算中心的建设是拉动地区实现人工智能发展的重要驱动力，既可以提升基础设施建设水平，也为吸引更多企业共谋发展起到积极的推动作用。过去几年来，南京、天津、西安等城市正是通过智算中心建设的浪潮跻身前十名。

图22 中国人工智能计算力发展评估——城市排行，2023

来源：IDC，2023

北京

北京拥有多家世界知名的云计算和大数据公司，在该领域有着雄厚的技术实力和庞大的计算资源。这些公司不断投资和扩展其数据中心和服务器集群，以满足快速增长的人工智能计算需求。北京是中国人才聚集的地方，拥有众多优秀的人工智能研发机构和高校，在人工智能领域具有较高的研究水平和人才培养能力。此外，北京还吸引了大量的国内外优秀人才，他们为人工智能计算力的发展做出了重要贡献。同时，北京市政府投入大量资金支持人工智能相关项目，推动创新研发，并为初创公司和研发项目提供早期到成熟阶段的资金支持。政府出台了多项鼓励人工智能发展的政策，加速了产业应用、知识产权保护等方面的进展。北京本地大型企业在人工智能研发上进行巨额投入，促进了前沿技术的快速发展。

杭州

杭州，作为中国的技术和创新中心区，一直在人工智能领域展现出极大的活力和潜力。从投资角度来看，杭州吸引了大量的风险资本和投资机构，为人工智能初创公司和项目提供资金支持。政府也在政策层面给予大力扶持，推出了一系列优惠政策和措施，以鼓励人工智能技术的研究、开发与应用。在人才储备方面，杭州的高校早已在人工智能领域设立专业，为行业培养了大量的技术和管理人才。在创新平台和创业生态方面，杭州建设了一系列创新平台和创业生态系统来支持人工智能的研发和应用。例如，杭州云栖小镇是一个集聚了众多人工智能、大数据和云计算公司的创新园区，为创业者提供了良好的创业环境和资源支持。此外，杭州还成立了人工智能发展引领小组，推动人工智能发展的战略规划和政策支持。

深圳

深圳作为全球重要的科技创新中心，在人工智能技术的运用和发展方面展现出显著的领导地位。首先，从投资的角度看，深圳吸引了大量的资本注入人工智能领域，为各类企业和初创公司提供了雄厚的资金支持。政策方面，深圳政府提供了多项扶持和激励政策，如建立前海蛇口自贸区的人工智能创新中心，不仅为企业提供资金支援，还为创业者提供了技术咨询和服务。在人才储备上，深圳与多所高校和研究机构建立了紧密的合作关系，培养并吸引了大批人工智能专业人才。此外，深圳人工智能领域的企业规模也不容小觑，拥有众多的人工智能大企业和初创公司，在智能交通、智慧城市、无人零售和人工智能芯片等领域进行了大量的创新和应用探索。产业聚集与应用创新为深圳在人工智能领域确立了领先地位，政府与企业的合作更是推动了人工智能技术在公共服务领域的广泛应用，显著提升了城市治理和公共安全水平。

上海

上海，作为中国的经济与科技中心，正以其无与伦比的人才集聚力展现出人工智能领域的独特魅力。得益于一流的高等教育和科研机构，上海为人工智能领域培养出了大量的顶尖人才。上海政府及企业组织的各种人才培养计划和比赛，进一步加强了人工智能的人才储备。上海涉足的人工智能应用范围广泛，从人脸识别到物联网，再到智能制造，不断推动人工智能技术与传统产业的深度融合，助力传统行业转型升级。无论是智慧交通还是智能医疗，上海都在积极开展创新应用和示范项目，以期为人工智能领域铺设一条宽广的发展大道。在投资方面，上海受益于其庞大的企业规模和政府的政策支持，持续吸引着全球视野下的资本和技术，成为推动全国人工智能领域发展的重要引擎。

苏州

苏州作为国内早期布局人工智能的地区，已经形成了覆盖整个城市的“人工智能+”应用创新区，也因此成为“中国十大人工智能创新城市”之一。该城市不仅引进了大型企业，还培养了本土企业巨头。苏州也正积极推动人工智能在各种场景中的应用：在交通领域，自动驾驶技术已经应用于公共交通和特定的车辆，如售货车和观光车；在教育领域，171所学校已经成为人工智能教育实验学校；在文旅领域，3D超写实人工智能数字主播和人工智能驱动的艺术创作平台也吸引了众多目光。而在医疗、金融、农业、安全以及日常生活中，人工智能的应用也越来越普遍。与国家层面的政策支持、企业投资以及其本身的优势相结合，苏州在人工智能领域的未来可期。

广州

广州在人工智能领域的发展日益受到关注。从投资的角度来看，广州持续增加对人工智能产业的资金支持，推动技术的快速发展，从而产生较强的聚集效应；众多人工智能企业和创新实体在广州发展壮大，涉及多个领域，如智能制造、智慧城市、医疗健康等。政策层面，广州为人工智能企业提供了一系列的优惠政策，以鼓励创新和发展。通过启动了“智慧广州”建设，广州不仅借助人工智能和大数据技术提高了城市管理与公共服务水平，还积极推进了智慧交通、智慧医疗等示范项目，为人工智能的实际应用积累了丰富经验。教育方面，广州的高等学府为人工智能产业培育了大量的专业人才，为未来的技术研究与发展提供了坚实的后盾。在企业规模上，广州拥有众多人工智能企业和创新实体，它们在智能制造、智慧城市、医疗健康等多个领域发展壮大，展现出强大的聚集效应。

济南

济南正在努力成为人工智能创新的中心，尤其在创新平台和科技园区的建设上显现出其坚定决心。济南高新技术产业开发区和济南大学科技园等创新中心为人工智能技术的研发、转化与商业化提供了宝贵的支撑。尽管济南在人工智能领域的起点相对较晚，但城市在人才引进、政策支持及与各方合作上都展现出了积极态势。当地高校及研究机构也为人工智能产业提供了丰富的人才储备。政府为企业和创新团队创造了有利的投资环境，同时也大力推进政策扶持，力图让济南在人工智能领域逐渐崭露头角。在不久的将来，济南将在山东省甚至全国的人工智能领域展现出更加引人注目的发展势头，对整个产业的壮大起到关键作用。

合肥

合肥正迅速崭露头角，成为中国人工智能发展的重要城市。这座城市凭借其强大的创新平台，如合肥高新区和技术转移中心等，为众多创新企业和研发团队创造了一个充满活力的创新环境，不仅获得了丰富的资源支持，也助推了合肥在智能制造、智慧城市、无人驾驶等领域的产业发展。同时，合肥在智慧交通、智能安防及医疗健康领域，已经开展了多个应用创新与示范项目，旨在更好地将人工智能技术落地实际应用。从投资策略、政策支持，到高校的人才储备，合肥都展现出对人工智能产业的坚定支持和深厚期望，因此本土企业和创新团队在人工智能技术研发与应用上都取得显著的进展。更为重要的是，这座城市深刻认识到人工智能的社会影响，正在将其技术与城市的整体发展战略紧密结合，确保人工智能在服务城市、推动创新和助力发展中扮演核心角色。

重庆

重庆致力于在2025年前，推动人工智能产业迈向新的高峰。该地区不仅设定了打造10个标杆场景项目的目标，还计划孵化10家亿级人工智能企业和集结1000家相关企业，以形成3-5个产业集聚区。重庆的“场景驱动”策略涵盖了从制造业智能化升级到沉浸式虚拟展馆等多个应用落地场景，与其产业结合十分紧密。为此，该地区不仅提出了“5大方向16条任务”，如推进成渝的国家算力节点和加强人工智能算法研究等，更鼓励企业联手高校和科研机构，共同攻克关键技术。此外，重庆也将交通、医疗和教育等行业的智能化发展纳入考量。以交通为例，围绕西部（重庆）科学城等地区，智慧交通的构想已经成为实际的规划，涉及智慧收费和联程导航等应用场景。而在医疗领域，智慧医院的概念也逐渐得到推广，智能预约和诊断等技术开始进入人们的视野。重庆还积极引进国内外人工智能领军企业，期望通过与本地优质企业的合作，形成一个健康、融合的创新生态。通过支持企业与高校及科研机构的紧密合作，不仅加速了技术的研发，还能培养出一批具备实战经验的高端人才。重庆还积极关注硬件的研发，例如图像处理芯片、语音处理芯片等，意在提升本地智能传感器的供给能力，确保人工智能技术在各个层面都能得到完善和发展。总体而言，重庆正通过全面、多方位的策略，迅速崭露头角，成为中国西部的人工智能产业中心。

成都

成都正在快速崭露头角，不仅因为其先进的创新平台和科技园区，如天府软件园和高新西区，更得益于这些园区给予人工智能企业及创新团队的良好环境和无微不至的支持。这座城市深知“人才为根本”，所以始终注重众多高等教育机构和丰富的人才储备。在政府的大力支持下，成都积极吸引了众多国内外知名企业和创新机构落户，从而促进了人工智能领域的交流与合作。可以预见，随着各方的紧密协同和不断努力，成都的人工智能产业创新能力和国际竞争力必将进一步增强，为整个西部地区的科技崛起和人工智能发展做出更为显著的贡献。

此外，南京、武汉、长沙、厦门、石家庄、郑州、天津等诸多城市的表现也可圈可点。

- **南京**智能计算中心，被列为全国首批且江苏省唯一的国家新一代人工智能公共算力开放创新平台，形成了一个综合算力、算法、数据与运营的完备服务生态，进一步确立了其作为华东地区主导的算力生产与供应中心的地位。

- **武汉**，旨在构建国内领先的人工智能高地，包括科技策源、算法创新、产业集聚、应用场景和人才培养等多个领域。武汉正专注于深度学习、大模型基础架构、图计算和模拟计算等基础技术，并积极攻关核心技术难题，同时推进人工智能标准化研究。

- **长沙**正积极打造成为全球研发中心城市。长沙已建立数据采集、存储、挖掘和信息感知等领域的稳定基础，长沙工业云等平台为人工智能提供云计算和数据支持。

- 福建省正在稳步推进其人工智能产业的发展，其中**厦门**园区为核心引擎。依托于周边一流的研究机构和高等学府，厦门园区旨在建立一个集资源、创新、应用、服务于一体的丰富而又活跃的人工智能产业生态。目前，这里已经具备从人工智能基础设施，到技术，再到应用的完整产业链布局。厦门将充分利用现有的市政和区域政策，将政策红利转化为真正的发展动力。

- **石家庄**，通过与本地数字化明星企业的战略合作，打造生态联盟，把不同的业务场景与新技术进行有机融合，使石家庄形成一支富有特色的数字经济企业“雁阵”。与此同时，孵化和培育大量的“专精特新”科创企业。

- **天津**积极布局人工智能产业，兴建了一系列创新平台和科技园区，在智能制造、智慧城市、人工智能芯片等关键领域积极布局。通过主动与国内外的科研机构和企业达成合作，深化研发合作与技术转移，进一步推动人工智能技术的创新与广泛应用。

总体来说，中国的人工智能市场呈现出快速发展的态势，越来越多的城市和省份都在不断加大在人工智能算力方面的投资力度，重视政策扶持、人才储备、推进相关企业的发展。

第四章
行动建议

考虑到中国人工智能市场的现状，IDC针对行业用户和技术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。

4.1 对行业用户的建议
跟随人工智能技术的进步，第一时间赋能业务场景

生成式人工智能将引领人工智能发展的重大变革，行业用户应尽快从中受益，这一过程充满挑战，需根据自身现状分步骤进行：

1. 评估企业生成式人工智能基础准备情况。其中包括整体人工智能战略和路线图、人工智能基础架构和人员技能的提升等——无论这一能力获得是企业内部还是通过战略合作伙伴。企业的人工智能基础现状和获得的技能是成功的关键因素。

2. 根据业务转型路线图调整人工智能投资组合。企业应意识到，根据本企业的业务需求、技术能力和预算制定合理的人工智能投资组合至关重要，包括供应商选择、采购节奏和模式、后期管理等。

3. 为未来而建设。人工智能必须满足不断发展的业务需求，而不仅仅局限于当前。企业应当评估技术栈各层的能力：基础设施、计算、数据、模型和应用，以实现高度可扩展、模块化和开放的解决方案。

4. 考虑与生成式人工智能解决方案供应商建立战略合作伙伴关系，其中包括人工智能平台和基础模型提供商。利用战略合伙伙伴的专业知识和工具为特定功能和行业专用用例实施概念验证（POC）方案，以获得所需的底层功能；此外还应全面评估供应商的能力，以应用的实际场景和目标为核心导向，考虑计算、存储、网络性能和稳定性，重视基础设施平台的协调、开放、适配等能力，以实现工作负载的可移植性、安全性和持续的优化，同时，控制成本，提高可持续性和安全性。

5. 优先实施与相关专业知识交叉的用例，以获得早期优势。部分企业已经将生成式人工智能模式与其专有的数据相交叉，使员工能够更快、更有效地利用机构知识。人工智能协作是企业在短期内快速创造价值的有效途径。因此，在充分评估后，企业应尽快实施生成式人工智能的应用，并在实施阶段继续进行评估和优化，形成闭环，实现对创新应用的持续改进。

6. 做好数据准备工作，持续关注数据安全与隐私。数据是模型训练的基础要素，也决定了模型的性能和效果，大模型的应用离不开高质量的数据。企业应重视优化数据治理、提升数据质量，通过确保数据完整性来推进人工智能系统的进阶。同时，企业需要保护数据的安全性和正确性，以合规为前提，控制风险，与值得信赖的伙伴合作，不要将敏感或机密信息贸然上传和分享。

4.2 对技术供应商的建议
围绕前瞻性、定制化、系统化、工程化和安全化、打造核心竞争力

处在生成式AI爆发的节点，人工智能技术供应商应尽快打造核心竞争力，尤其重视以下几个方面：

1. 展示创新和前瞻性。生成式人工智能给各行各业带来很多希望，技术供应商需要证明他们不仅看到了价值，并且已经采取措施、开始探索具体用例，以充分利用这一技术。能够在这个方面为用户提供更高用例清晰度的供应商，将被视为富有创新意识和前瞻性的供应商。

2. 提供定制化解决方案并予以验证。了解目标受众的需求、痛点和组织目标。定制专门的人工智能应用以满足这些特定需求，并邀请潜在客户通过概念验证（POC）项目测试解决方案提供商的人工智能应用，以便在完全投入之前了解方案如何为企业带来好处。

3. 重视数据安全与业务健康发展。人工智能技术供应商除了关注技术能力和商业模式之外，还应具有坚固的数据安全性并承担起数据来源合法性的责任，推动人工智能在安全、可信任的环境中实现最大化赋能。同时，应与最终用户建立信任，重点应关注提高透明度、履行承诺、支持客户业务发展等方面。

4. 构建开放多元的算力平台生态，系统化角度解决算力挑战。技术供应商需要不断投入资源，加强算力基础设施的建设，构建开放、兼容、灵活、高效、全栈的算力软件基础设施平台，加强多种算力资源的融合发展，能够根据业务需求和用户需求，动态分配和调度算力资源。通过更好地整合和利用算力资源，提高算力服务的效率和质量。

5. 及时、可量化地衡量投资回报率（ROI）。技术供应商应通过为最终用户提供工具或分析仪表盘来检测关键性能指标，实现自动化数据管理和分析，从而帮助用户及时量化他们在实施人工智能应用中所获得的价值，并缩短洞察和决策时间。

6. 根据用户反馈不断增强功能。算力基础设施供应商应从行业用户具体的需求出发，定期收集用户反馈，并将他们的建议纳入更新中，重点加速底层算力设施的升级，以系统化以需求为中心，系统为抓手，应用为导向，推进人工智能算力的规模化、泛在化发展，助力智能涌现。

关于浪潮信息

浪潮信息是全球领先的IT基础设施产品、方案和服务提供商，业务涵盖服务器、存储和网络三大领域，有8个研发中心、10个生产基地、26个分支机构，业务遍及全球120多个国家和地区。

人工智能是浪潮信息战略重点业务之一，浪潮信息是全球第二大服务器供应商，也是全球第一的 AI 服务器提供商。在中国 AI 加速计算市场占有率连续第六年超过 50%，可提供从计算平台、算法模型、管理套件、框架优化到应用加速的完整方案。浪潮信息推动 AI 领域开放计算的发展，参与制定了 OCP社区的 OAM 规范以及 ODCC 社区的 GPU 服务器规范，为不同的 AI 技术提供统一的技术标准。浪潮信息坚持“伙伴第一”的原则，不断发展元脑生态，聚合具备AI开发核心能力的左手伙伴和具备行业AI整体方案交付能力的右手伙伴，加速行业智能的构建，最终帮助用户完成业务智能转型升级。

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
idc-community.com
www.idc.com