IDC | 浪潮信息

2026年

中国人工智能计算力发展评估报告

目录

核心观点

- **全球人工智能产业正从“能力验证”加速迈向“价值兑现”，智能体成为新的增长驱动力。**

  全球人工智能产业正从“能力验证”加速迈向“价值兑现”。IDC研究显示，到2031年，人工智能预计将累计创造22.5万亿美元的全球经济价值，规模远超此前任何一代通用技术。随着智能体技术走向成熟，人工智能正完成从内容生成到任务执行、从单轮问答到长链路协作的转变；作为下一阶段人工智能投入的重要增量，智能体预计将推动企业人工智能采购预算进一步增长约20%。

- **全球智能体发展呈现快速增长，并以乘数效应推动Token（词元）消耗加速跃升，中国智能体部署渗透率高于全球领先水平。**

  IDC预测，全球活跃智能体数量将从2026年的7,940万个增长至2030年的22.16亿个，复合增长率达129.8%。在任务执行次数增长、单任务Token消耗提升与多智能体协同的乘数效应作用下，同期全球Token消耗量的复合增长率达4,822.6%，约为活跃智能体数量增速的37倍。中国市场同样显著，IDC研究显示，2026年中国智能体部署渗透率达81.0%，高于全球领先水平。

- **算力跃升为国家战略基础设施，智能体拉动算力规模斜率大幅上调。**

  智能体的规模化运行正在带来持续、刚性的算力需求，推动算力跃升为与水利、电网、铁路同等层级的国家战略基础设施。IDC数据显示，2026年中国智能算力规模将达到2,576.5 EFLOPS，同比增长87.9%，并在2030年达到10,524.3 EFLOPS。2025-2030年中国智能算力规模和通用算力规模的复合增长率分别达50.3%和27.7%，较上一版本预测值46.2%和18.8%有显著提升。

- **物理世界的供给增速难以匹配数字世界的需求增速，供需缺口长期存在。**

  算力投资规模正以远超预期的速度扩张，但供给扩张仍难以同步：高端芯片、HBM（高带宽内存）、服务器CPU等核心部件的产能扩张周期普遍长于人工智能需求的爆发节奏，能源约束与之叠加。IDC预测数据显示，全球人工智能算力需求满足率从2024年的79%持续下滑，2027年降至71%的谷底，预计2030年回升至77%；缺口绝对值从2024年的389亿美元扩大到2030年的3,809亿美元，供需剪刀差正在持续扩大。

- **AI计算进入系统能力竞争时代，Capability AI Compute（能力型智算）突破智能上限，Capacity AI Compute（容量型智算）实现智能充分供给。**

  随着智能体从单次模型调用走向复杂任务执行，AI计算正从单芯片性能竞争转向系统能力竞争。能力型智算通过超节点、HBM（高带宽内存）、高速互连等系统级创新，提升单机模型承载能力和复杂任务处理能力，突破智能上限；容量型智算通过PD/AF分离、异构芯片协同和全链条优化，提高Token吞吐与资源利用效率，降低单位Token成本，实现智能规模化供给。

■ **智能体落地已形成明显发展鸿沟：Coding Agent（代码开发智能体）与Work Agent（办公智能体）遥遥领先，传统行业处于探索期但长期生产力影响大。**

代码开发智能体凭借高度数字化、任务边界清晰和结果可验证等特点，率先进入规模化应用；办公智能体则依托高频通用场景和较低落地门槛快速普及。相比之下，制造业、医疗健康、能源/电力等典型传统行业已经开始将人工智能应用嵌入具体业务流程，但普遍面临技术适配难、业务流程固化、组织变革阻力大等挑战，智能体应用仍停留探索试点阶段，远未达到实用化水平。

■ **智能体重构中国算力城市排名：北京、杭州领跑，深圳超越上海跻身前三。**

2026年中国人工智能算力基础设施进入全面提速期，城市人工智能算力发展格局正由单纯比拼算力规模，转向算力供给、模型能力、智能体与具身智能应用、产业生态和人才储备等综合能力的较量。北京凭借完整的算力供给产业链、实力雄厚的大模型与人工智能应用企业继续稳居首位；杭州依托阿里巴巴、DeepSeek等龙头企业以及持续扩大的算力和人工智能投入位居第二；深圳则凭借深厚的算力供给产业基础、腾讯持续加大的资本开支及智能体应用快速增长，超越上海跻身前三。

01.

全球及中国人工智能发展概述

智能体正在推动全球人工智能产业经历一次根本性的跃迁：从提升效率的工具，演进成为能够独立承担生产任务的运行单元。以可调度、可隔离、可运维的持续运行单元形态，智能体正嵌入企业的研发、办公、运营与决策流程，完成从内容生成到任务执行、从单轮问答到长链路协作的转变。这是继信息化、数字化之后，生产方式的又一次系统性更替；企业竞争的标尺，也随之从“是否部署人工智能”转向“能否规模化、安全化、可治理地运行智能体”。

这一变化的深层逻辑，在于智能体改变的不只是效率水平，而是价值创造的方式。当人工智能能够自主拆解任务、调用工具、协同多个智能体完成复杂流程时，企业的组织形态、成本结构和竞争力的定义都将随之改变——人工智能的产出不再是一次性的内容交付，而是持续的任务执行与结果反馈，Token成为成本核心，智能体成为价值核心。

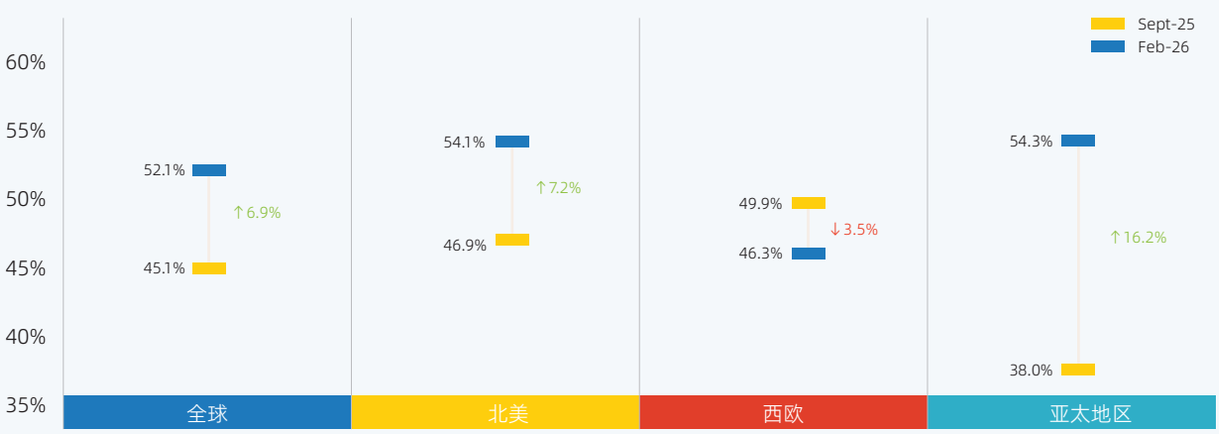
伴随智能体技术的持续演进和应用场景的不断丰富，全球人工智能算力发展正在呈现出新的趋势。

趋势一：全球人工智能价值兑现提速，智能体成为新增长驱动力

全球人工智能产业正从“能力验证”加速迈向“价值兑现”。IDC研究显示，到2031年，人工智能预计将累计创造22.5万亿美元的全球经济价值，这一规模远超此前任何一代通用技术所带来的经济增量。与此同时，人工智能的商业价值正从预期走向现实：2026年，全球已有52.1%的人工智能项目取得可衡量的业务成果。以微软为例，其2026财年人工智能相关收入已超过250亿美元，年增长率超过40%，Copilot等人工智能产品正成为其核心收入来源；谷歌母公司Alphabet 2026年第二季度财报显示，其云业务收入同比增长超过35%，其中人工智能基础设施和Vertex AI平台贡献了主要增量。

图1: 全球各区域企业AI项目取得可衡量业务成果成功率对比(2025.09-2026.02)

在过去两年中，您的人工智能相关项目中，有多少已经取得了可衡量的业务成果（例如，降低成本、提高流程效率和改善客户体验）？



n = 903（北美 n = 366，西欧 n = 228，亚太 n = 309）
注：展示的是成功率的平均变化百分比
来源：IDC's Future Enterprise Resiliency & Spending Survey Wave 1，2026年2月

人工智能价值兑现的加速正在持续强化企业的投入意愿。全球30.2%的企业表示，即使经济环境下行，人工智能仍是坚定投入的方向；金融、医疗、零售等领域的头部企业，正将人工智能从“锦上添花”的试点项目升级为“不可或缺”的核心业务系统。

智能体正在成为下一阶段人工智能投入的重要增量。随着智能体技术走向成熟，预计将推动企业人工智能采购预算进一步增长约20%。IT部门的角色也随之从“系统建设与维护者”转变为“智能生态系统的指挥者”——企业当下的行动选择，将直接影响其未来三到五年的增长韧性与竞争位势。

趋势二：全球智能体快速增长，中国智能体部署渗透率高于全球领先水平

全球智能体部署进入快速增长阶段。IDC预测，全球活跃智能体数量将从2026年的7,940万个增长至2030年的22.16亿个，复合增长率达129.8%；同期全球Token消耗量的复合增长率达4,822.6%，约为活跃智能体数量增速的37倍。中国增长同样显著，活跃智能体数量将从2026年的495万个增长至2030年的1.97亿个，复合增长率达151.2%；同期中国Token消耗量的复合增长率达3,499.3%。

图2:全球企业活跃智能体关键数据预测，2026-2030

全球智能体规模指标	2026	2027	2028	2029	2030	复合年均增长率 (26-30)
活跃企业智能体（单位：百万）	79.4	214.9	523.9	1,153.3	2,216.0	129.8%
年度Token消耗（单位：Peta）	0.026	1.2	48.6	10,877	152,667	4,822.6%

来源：IDC，2026

图3:中国企业活跃智能体关键数据预测，2026-2030

中国智能体规模指标	2026	2027	2028	2029	2030	复合年均增长率 (26–30)
活跃企业智能体（单位：百万）	4.95	14.90	39.10	95.13	197.18	151.2%
年度Token消耗（单位：万亿）	1.5	82.2	3,465.2	112,104	2,517,438	3,499.3%

来源：IDC，2026

注：Peta (P) 代表10^{15}，1 拍词元 = 1000万亿个词元；词元（Token）为AI大模型信息处理基本单元。

中国智能体部署渗透率高于全球领先水平。IDC研究显示，2026年中国智能体部署渗透率达81.0%，高于全球领先水平的78.1%。中国智能体的领先发展，源于三方面因素的共同作用：

- **开源生态降低应用门槛。**以DeepSeek（深度求索）为代表的开源模型在2025至2026年持续迭代，其技术普惠化效应显著降低了企业部署人工智能的门槛。开源框架和工具的普及使中小企业和个人开发者也能参与人工智能应用开发，极大地扩展了智能体创新的主体范围。开源社区的力量在中国尤为突出，Hugging Face平台上来自中国开发者的模型贡献量已位居全球第二，魔搭社区等本土平台汇聚了数千个开源模型和工具。

- **丰富业务场景加速价值验证。**中国拥有全球最丰富的数字化应用场景，制造、金融、电商、政务、医疗等领域已成为智能体落地的重点行业。在制造业，智能体驱动的智能质检已在多家头部企业的产线中实现规模化部署；在金融业，智能风控智能体已从单点试点走向全面推广；在电商领域，从商品选品到客服应答的端到端智能体闭环已成为行业标配。丰富的场景为智能体提供了持续的迭代数据和价值验证环境。

- **成熟工程化降低部署成本。**经过数年的技术积累，中国企业在模型微调、推理优化、系统集成等环节已形成完整工具链和人才储备，智能体从开发到部署的周期明显缩短，规模化落地成本显著下降。这使企业能够以更低门槛将智能体嵌入核心业务流，加速从试点验证向规模化应用推进。

趋势三：智能体驱动算力需求结构性跃升——看得见的智能体应用，看不见的算力基础设施

智能体规模化运行对算力需求的拉动，遵循一个清晰的三项乘数结构：

Token消耗总量 = 任务执行次数 × 单任务Token消耗 × 多智能体协同放大系数

- **任务执行次数：**全球智能体年执行任务数将从2026年的2,800亿次增至2030年的415.07万亿次，复合年均增长率达520.5%；中国将从2026年的178亿次增至2030年的39.41万亿次，复合年均增长率达586.0%。每一次任务执行——无论是一次信息检索、一次数据分析还是一次跨系统操作——都需要调用算力资源。

- **单任务Token消耗：**单任务Token消耗由任务链路的长度与复杂度决定。一个完整的智能体任务不再是一次模型调用，而是包含任务规划、模型推理、工具调用、结果分析、再规划、再调用的循环往复过程，每一轮循环都产生新的Token消耗；随着大模型上下文窗口的扩展和推理能力的增强，单个任务所需的Token正从简单问答的数十个，增长到复杂文档分析的数万个，再到多轮对话与决策的数十万个，单任务Token消耗已增长数百倍。这意味着智能体的规模化运行带来的不是“更多用户使用人工智能”的线性增长，而是“每个任务变得更长、更复杂”的结构性变化。

- **多智能体协同：**多智能体协同已成为智能体发展的主导趋势，也是公式中放大系数持续走高的原因：一方面，协同将一个用户任务分解为多个智能体子任务，推高任务执行次数；另一方面，智能体之间的上下文传递与结果汇总延长任务链路，推高单任务Token消耗。

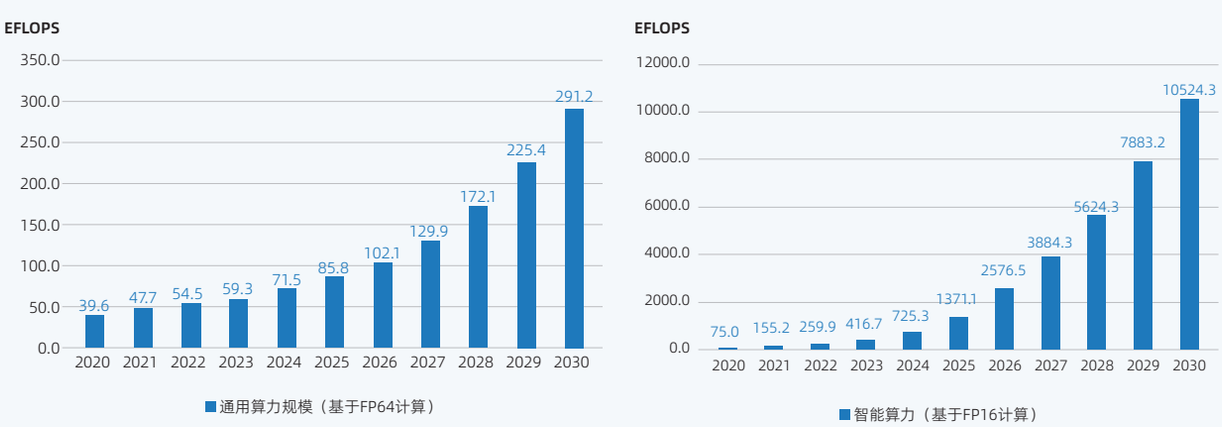
IDC预测，2026年平均每个企业计划部署19.65个智能体，70%的组织将采用融合生成式、处方式、预测式和智能体技术的Composite AI（复合人工智能），推动人工智能从单一模型调用走向多模型、多智能体的协同决策。Salesforce的Agentforce平台已支持企业部署多个专业智能体协同完成销售、服务、营销等跨部门任务，其客户Slack在此基础上的内部运营效率提升了40%。到2027年，全球G2000企业中智能体的使用量将增长10倍，调用量更将提升1,000倍。

趋势四：算力跃升为国家战略基础设施，智能体拉动算力规模斜率大幅上调

智能体的规模化运行正在带来持续、刚性的算力需求，推动算力跃升为与水利、电网、铁路同等层级的国家战略基础设施。“十五五”时期算力网建设将新增直接投资4万亿元，投资体量已达到与水利6万亿元、电网5万亿元、铁路4万亿元等传统重大基建相当的水平，算力作为国家战略基础设施的地位正在加速落地。

智能体拉动算力规模斜率大幅上调。IDC数据显示，2025年中国智能算力规模达到1,371.1 EFLOPS，2026年预计增至2,576.5 EFLOPS、同比增长87.9%，到2030年将达10,524.3 EFLOPS，2025–2030 年复合增长率达50.3%；同期，中国通用算力规模为85.8 EFLOPS，2026年预计增至102.1 EFLOPS、同比增长19.0%，到2030年将达291.2 EFLOPS，复合年均增长率达27.7%。智能算力和通用算力的复合增长率较上一版本预测值46.2%和18.8%有显著提升。

图4:中国智能算力和通用算力规模及预测，2020-2030



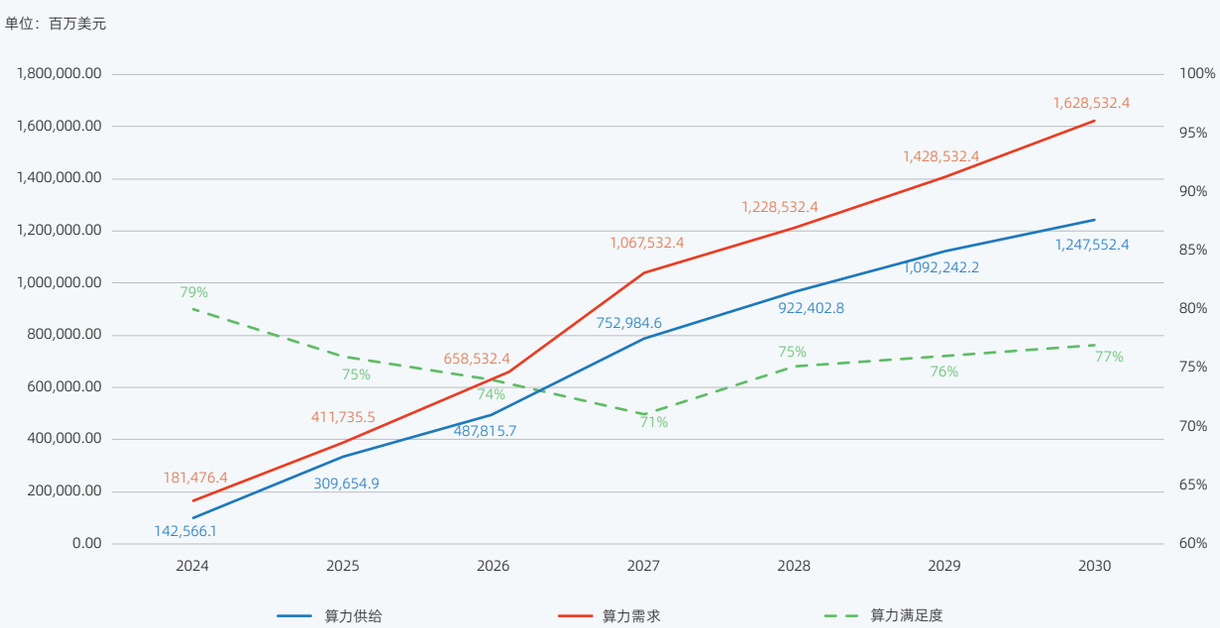
来源：IDC，2026

趋势五：物理世界的供给增速难以匹配数字世界的需求增速，供需缺口长期存在

从宏观市场看，算力投资规模正以远超预期的速度扩张。2030年全球人工智能算力市场规模预计达1.25万亿美元，是2025年的4倍；其中中国人工智能算力市场规模将在2026年达到515亿美元，同比增长37%，并在2030年达到1,501亿美元，2025-2030年复合增长率达31.8%。**微观层面，企业端投入意愿同样坚定：**97.3%的企业预计将持续扩大人工智能算力投入，其中37.7%需要明显增加推理需求，29.3%需要持续大规模资源支撑业务运行。**供需两端的同步升温直接反映在价格信号上**——2026年Token及云服务调价频率从一年一次加快至半年甚至更短，全年涨幅预计达30%至50%。

尽管投资力度空前，供给扩张仍难以同步。高端芯片、HBM（高带宽内存）、服务器CPU等核心部件的产能扩张周期普遍长于人工智能需求的爆发节奏，投资额的增长并未等比例转化为算力供给的增长；能源约束与之叠加——数据中心用电需求随智能体规模化运行激增，而电力供给受制于电网扩容与关键设备交付的长周期，扩容节奏显著慢于算力建设周期。IDC预测数据显示，全球人工智能算力需求满足率从2024年的79%持续下滑，2027年降至71%的谷底，预计2030年回升至77%；缺口绝对值从2024年的389亿美元扩大到2030年的3,809亿美元，供需剪刀差正在持续扩大。中国同样处于这一全球性供需短缺格局之中，智能体规模化运行推动国内算力需求斜率大幅上调，弥合缺口需要供给体系的长周期、系统性扩张。

图5:全球人工智能算力供给、需求与需求缺口（2024-2030）



来源：IDC，2026

02.

智能体重构
人工智能基础设施

智能体正在推动人工智能基础设施从单点性能优化转向任务级系统协同优化。智能体将人工智能计算从单次模型调用扩展为持续任务执行，使基础设施优化目标从提升单一性能转向提升完整任务执行效率。围绕智能体任务执行需求，人工智能基础设施各环节正在发生系统性重构：

- **需求特征：**由阶段性负载转变为生产型负载，具备长期稳定、实时响应与高并发能力。
- **竞争焦点：**从比拼芯片峰值算力，转向打造系统有效算力，实现能源、计算、存储、网络的低损耗协同。
- **运行重心：**从占有硬件资源转变为高效使用资源，重点做好算力调度、数据供给、可观测与Token运营。
- **部署形态：**跳出集中式智算中心模式，构建云-边-端协同体系，依据时延、成本与数据安全完成动态配置。
- **价值实现：**从资源供给售卖升级为输出智能能力服务，采购对象由服务器、GPU时长转向模型能力与Token产出。
- **行业配置：**摒弃统一标准，采用场景化组合模式，结合业务特征灵活匹配部署方案。

具体来讲，计算从单芯片峰值性能竞争转向能力型智算与容量型智算双线演进，能力型智算通过超节点、高速互连等系统级创新提升单机承载模型规模和复杂任务能力，突破智能上限；容量型智算通过异构资源协同、推理优化和全链条效率提升，提高Token产出效率，降低单位智能成本，实现智能规模化供给。与此同时，存储需要从数据保存向持续供给上下文、记忆和任务状态演进；网络需要支撑模型、数据、工具和业务系统之间的实时协同；数据中心需要保障高密度、持续运行的智能生产；终端推动云、边、端根据任务需求动态分工；模型与智能体框架负责将模型能力转化为可执行任务；调度与MaaS负责组织异构资源和智能服务；Token运营则推动基础设施评价从资源投入转向智能产出。

因此，人工智能基础设施不再只是支撑单次训练或推理的算力底座，而是围绕智能体开发、部署、运行和治理形成的综合基础设施体系，推动人工智能从模型能力突破走向规模化智能生产。

2.1 计算：AI计算进入系统能力竞争时代

智能体正在改变人工智能计算的负载形态，推动计算体系从单一算力优化向系统级协同演进。与传统人工智能主要面向模型训练和单次推理不同，智能体以完成任务为目标，需要在一个业务流程中连续执行模型推理、任务规划、状态管理、数据处理、工具调用、代码执行和多模态解析等多种计算任务，并通过多轮模型调用持续更新上下文。因此，智能体计算不再仅关注单次推理性能，而需要综合考虑任务完成效率和系统协同能力，推动CPU、GPU、NPU及其他加速资源根据任务特点进行协同调度。在这一趋势下，人工智能计算正在形成两个新的发展方向：一个是面向智能前沿的能力型智算，不断提升单机可承载模型规模与智能水平，支撑智能上限突破；另一个是面向智能普及的容量型智算，提升计算效率和资源供给能力，让更多智能以更低成本被触达，实现智能充分供给。

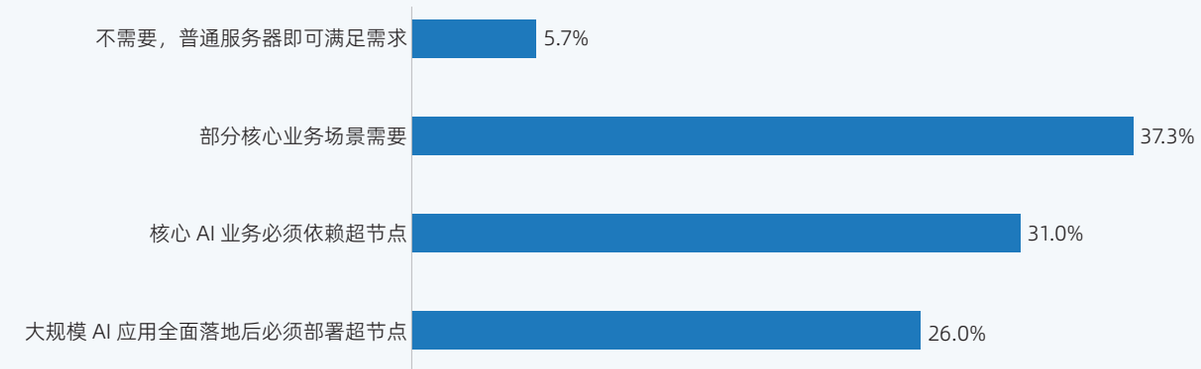
计算负载趋向多元化，CPU与GPU形成协同分工。随着智能体从单一模型调用走向复杂任务执行，一次完整任务可能涉及大语言模型推理、向量数据库检索、结构化数据查询、外部API调用以及多模态数据解析等多个环节。不同任务对计算资源需求存在差异，其中模型推理和大规模并行计算更依赖GPU，而任务规划、工作流编排、数据解析、工具调用、权限校验、状态管理和上下文维护等任务更适合由CPU承担。随着企业内部智能体数量增加，多个智能体将在统一平台并发运行，对CPU资源提出更高需求。相比传统人工智能服务器中CPU与GPU约1:8至1:4的资源配比，智能体场景下CPU承担的任务进一步增加，推动人工智能基础设施从以GPU为中心向CPU与GPU协同演进。

容量型智算成为支撑人工智能规模化应用的关键路径。随着人工智能从模型能力突破走向规模化应用，计算需求正在从追求单点性能提升转向支撑更大规模、更低成本的智能服务供给。容量型智算面向智能普及，目标是实现智能能力的充分供给和低成本触达。针对训练和推理等不同阶段的负载特征，计算架构需要开展贯穿式创新，通过软硬件协同提升整体效率。例如，针对注意力计算等核心负载开发专用加速单元，使系统效率优于通用芯片的简单组合，而非单纯追求峰值算力。在推理阶段，通过Prefill-Decode Disaggregation（预填充-解码分离）优化预填充和解码过程，使预填充专注计算吞吐、解码降低时延，并进一步结合Attention-FFN Disaggregation（注意力-前馈网络分离）、Encoder-Prefill Disaggregation（编码器-预填充分离）、Draft-Target Disaggregation（草稿-目标分离）等策略，针对不同场景优化时延和吞吐，在既定硬件条件下实现更多、更快、更低成本的Token产出。DeepSeek通过系统级优化实现高吞吐、低成本Token供给，体现了贯穿式优化提升计算效率的价值。同时，贯穿式创新还体现在多元芯片生态支持，通过开放硬件架构和统一异构计算软件系统，使不同厂商、不同代际的计算资源能够按需升级和高效复用，满足不同业务场景需求，在保障模型能力上限的同时实现智能能力普惠供给。

能力型智算推动人工智能向更高智能上限演进。随着模型规模扩大、多模态融合以及复杂任务需求提升，人工智能计算面临突破单芯片性能边界的需求，计算竞争正从芯片峰值性能提升转向系统级能力构建。能力型智算面向智能前沿，目标是不断突破智能上限，超节点、Chiplet（芯粒）、HBM（高带宽内存）和先进封装等技术，代表了一条自下而上的系统能力路径，对多模型并发、长上下文处理和复杂任务稳定运行具有关键支撑作用。面向多模融合场景，开放超节点人工智能服务器采用多主机低时延内存语义通信架构，可在单机内实现多路GPU统一寻址与高速互连，最大可承载4万亿参数模型，也可支持多个万亿参数模型协同运行，使更大参数规模的模型能够在单节点内高效运行，持续推高智能上限，为K3级乃至更高性能的模型提供系统级支撑。在实际部署中，美国实验室的前沿模型（万亿至十万亿参数）在超节点上完成训练后，可蒸馏到小模型进行推理分流，形成“大模型训练、小模型推理”的分层算力格局。这种系统级设计使计算产业从单芯片峰值性能的竞争，扩展到多芯片协同的系统计算能力，其关键在于芯片间通信带宽、内存统一寻址和任务调度的协同优化。

IDC调研显示，94.3%的企业认为超节点是人工智能基础设施的重要能力，其中31.0%表示核心人工智能业务必须依赖超节点，37.3%表示部分核心业务场景需要，26.0%表示大规模人工智能应用全面落地后必须部署。

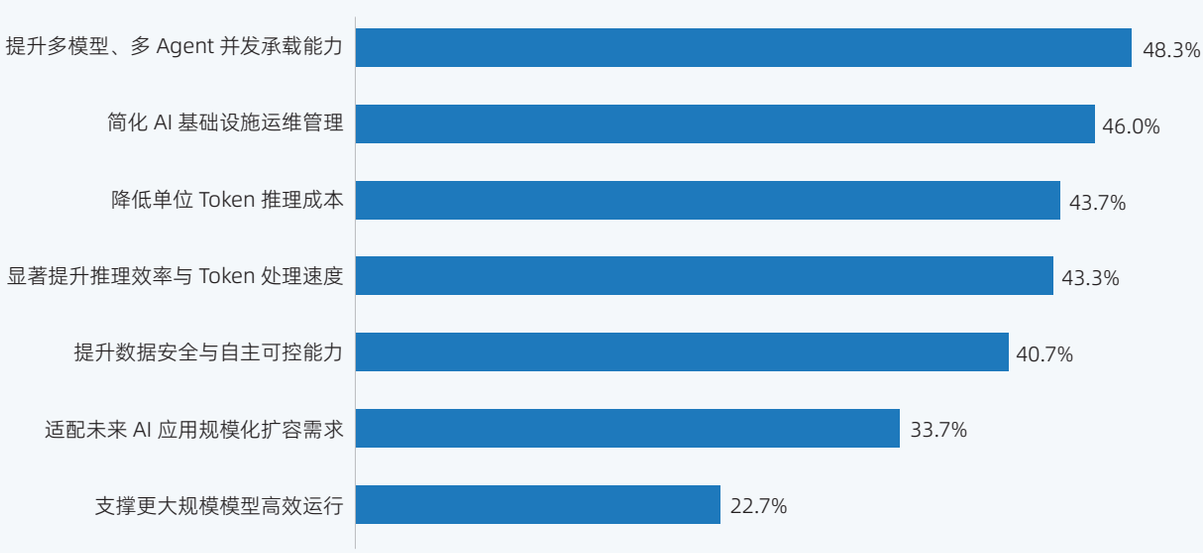
图6:企业对超节点作为 AI 基础设施关键能力的认可度



来源：IDC，2026

企业部署超节点的首要驱动因素是提升多模型、多智能体并发承载能力（48.3%），其次是简化人工智能基础设施运维管理（46.0%）、降低单位Token推理成本（43.7%）和提升推理效率与Token处理速度（43.3%）。各行业的超节点诉求差异显著：医疗健康以多模型多智能体并发为核心驱动（66.7%），电信运营商最看重简化运维（60.0%），能源电力聚焦推理效率提升（60.0%），信息技术与软件行业侧重未来规模化扩容（63.3%），机器人企业则最关注数据安全与自主可控（48.9%），整体呈现出与各行业人工智能应用场景深度绑定的差异化驱动格局。

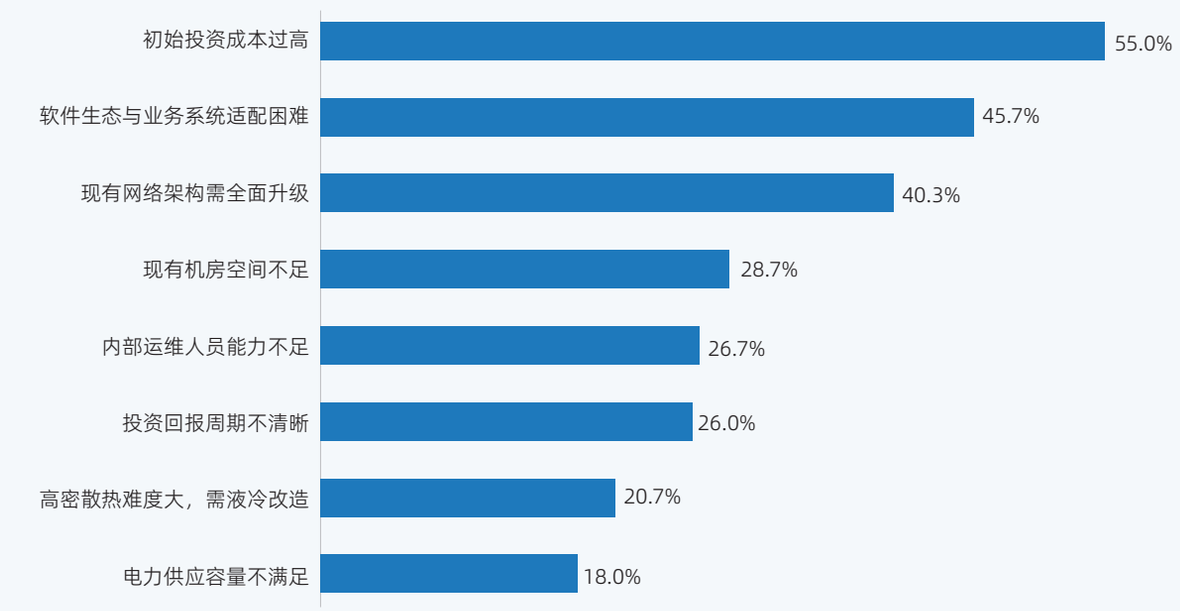
图7:企业部署 AI 超节点的主要驱动因素



来源：IDC，2026

企业部署超节点的痛点是初始投资成本过高（55.0%），其次为软件生态与业务系统适配困难（45.7%）和现有网络架构需全面升级（40.3%）；其中，制造业成本敏感度最高（72.2%），信息技术与软件行业受系统适配制约最突出（66.7%），量化交易与能源电力则面临网络架构升级压力（53.8%/53.3%），电信运营商独受电力容量瓶颈制约（43.3%），整体呈现出“成本普适、技术痛点分化”的行业格局。

图8:企业部署 AI 超节点面临的主要挑战



来源：IDC，2026

2.2 存储：从数据保存走向智能数据供给

智能体正在改变存储的数据组织和访问方式。智能体运行需要持续调用企业知识库、向量数据库、用户历史、会话上下文、执行状态和工具结果等多类型数据。与传统人工智能训练阶段主要关注数据集中管理不同，智能体需要根据任务状态持续读取、更新和复用数据，推动存储从数据保存向上下文供给和任务支撑演进，对访问时延、并发能力和跨系统连接能力提出更高要求。

数据访问模式的变化推动存储架构调整。传统存储架构以文件或块为单位组织数据，面向批量读写优化，而智能体场景下的数据访问模式更偏向随机、高频和低延迟。每次智能体调用都涉及知识检索、上下文加载和状态读取，这些操作需要在毫秒级内完成。随着智能体逐步进入更多业务场景，实时数据访问、上下文关联和广泛数据连接已成为基本要求，数据治理、可观测性、身份与访问控制也需要纳入存储架构的设计考量。

多层级记忆体系对存储性能提出差异化要求。智能体记忆体系可分为四个层次：工作记忆承载当前任务上下文，需要低时延和高带宽；短期记忆管理近期会话和执行状态，需要快速读写和动态更新；长期记忆存储用户偏好和历史经验，需要持久化和可检索；组织记忆管理企业知识库和业务资产，需要权限治理和版本管理。四层记忆对存储介质的性能要求各不相同，工作记忆和短期记忆要求微秒级延迟，长期记忆和组织记忆则更侧重容量和可靠性。存储系统需要根据数据访问频率和延迟要求，在不同存储层级间自动完成数据迁移和缓存调度。

图9:Agent记忆体系

记忆类型	主要内容	基础设施要求
工作记忆	当前任务上下文	低时延和高带宽
短期记忆	近期会话和执行状态	快速读写和动态更新
长期记忆	用户偏好和历史经验	持久化和可检索
组织记忆	企业知识库和业务资产	权限治理和版本管理

来源：IDC，2026

KV Cache（键值缓存）管理从单请求扩展到跨会话。长上下文和多轮交互使KV Cache成为影响推理效率的重要资源。缓存复用、显存与内存分层、跨节点缓存和前缀缓存可减少重复计算，但需要配套一致性、迁移和隔离机制。在智能体场景下，KV Cache的生命周期不再局限于单次推理请求，而是跨越多次对话轮次和智能体调用。如何在不同请求间共享KV Cache、如何在智能体切换时保持缓存有效性，成为新的技术挑战。面向超大规模推理集群，上下文记忆存储支持KV语义直通或采用专用DPU进行语义卸载，可扩展为PB级共享KV缓存池。

统一数据空间支撑智能体跨系统访问。智能体需要在数据、模型、工具和权限之间持续交互，传统的数据孤岛架构难以满足这种跨系统、跨模态的访问需求。统一数据空间通过标准化的数据访问协议和权限模型，使智能体能够透明地访问分布在不同系统中的数据，同时满足数据治理和合规要求。

2.3 网络：低时延高可靠连接智能体执行链

智能体任务执行涉及终端、模型服务、知识库、企业系统和外部工具等多个环节，网络需要支撑不同组件之间的持续交互与协同调用。相比传统网络主要关注数据连接和传输效率，智能体场景下网络需要同时保障调用链路的低时延、可靠性、安全性和可观测性。随着智能体逐步进入核心业务流程，网络正在从连接基础设施演进为支撑模型、数据、工具和业务系统协同运行的任务执行基础设施。

东西向和南北向流量同步增长，推动网络从集群互联向全链路协同演进。GPU、节点和智能体之间的数据交换增加了东西向流量，用户、终端、模型服务、企业系统和外部工具之间的调用增加了南北向流量，两类流量同时影响智能体任务的完成时间。从大规模训练集群到企业边缘推理的人工智能工作负载广泛部署，推动了对高速、低延迟数据中心网络基础设施的持续需求。QUIC/UDP等低延迟传输协议已被部分新智能体用于优化传输，但很多企业防火墙默认屏蔽UDP，导致丢包率激增，需要网络架构进行系统性适配。

多轮任务执行推动网络从平均性能优化向尾时延保障演进。智能体任务通常包含多次模型和工具调用，任何环节的延迟都会累积。网络评价需要覆盖P95和P99尾时延、抖动、调用成功率、跨域通信效率和故障恢复能力。企业数据中心短期内将采用混合模式，同时运行传统三层架构与现代化叶脊网络架构，并逐步向叶脊架构迁移。尾时延的控制比平均时延更为重要，在智能体的多步执行链路中，一次尾时延尖峰就可能导致整个任务超时失败。

跨域协同扩大算力调度范围。Scale-out（横向扩展）支持集群内部扩展，Scale-across（跨集群扩展）支持不同集群、数据中心和算力域之间的协同，跨域网络使分散部署的资源形成可统一调度的算力供给。智能体启动时会向网络发送探测包，自动选择当前延迟最低、丢包率低于阈值的服务器节点来运行推理。跨地域的智能体协同还需考虑数据主权和合规要求，网络层需要支持数据本地化路由和跨境流量控制。

多智能体并发运行推动网络向精细化治理演进。企业部署多个智能体后，网络还需提供流量隔离、身份认证、访问控制、服务质量保障和工具调用审计，避免不同业务之间相互干扰。在多租户环境下，不同部门的智能体需要运行在独立的网络命名空间中，确保流量隔离和安全策略的独立配置。

2.4 数据中心：AI负载演进驱动高密度计算与运营模式重构

智能体推理具有全天候运行、流量波动、并发不确定和服务连续性要求高等特征，数据中心需要同时支持高密度部署、弹性供给和稳定运营。与模型训练具有较明确任务周期不同，智能体服务随着业务活动持续产生请求，容量规划需要兼顾基线负载、突发并发、任务优先级和服务恢复。

持续性推理负载推动数据中心向高弹性和高密度演进。模型训练通常具有相对明确的任务周期，而智能体服务面向持续在线的业务请求，需要根据实时负载变化快速响应，单纯依靠请求排队难以满足低时延服务需求。数据中心需要具备面向峰值负载的容量保障能力，同时支持资源快速扩展和动态调度。随着智能体基础设施向GW（吉瓦）级AIDC（智算中心）演进，单机柜功率从几十千瓦级向百千瓦级甚至MW（兆瓦）级发展，对算力密度、空间利用效率和能源效率提出更高要求。

高密度计算需求推动散热体系向原生液冷演进。随着智能体持续运行和大模型推理规模扩大，人工智能基础设施对算力密度和持续运行能力提出更高要求，单机柜功率正从几十千瓦级向百千瓦级甚至MW（兆瓦）级演进。以传统风冷为主的散热方式，其单机柜功率通常受限于约40至50千瓦级，而面向高密度人工智能计算的机柜功率预计将在2026年底突破300千瓦级，传统散热方式难以满足持续提升的算力密度需求。高密度部署带来的挑战已不再是单颗芯片散热问题，而是计算、互连、散热和空间布局高度耦合后的系统性问题。传统“先定计算、再配散热”的后置适配模式，容易造成散热结构占用计算空间、液路设计与高速互连相互制约等问题。原生液冷将热设计前置到系统架构定义阶段，与计算、互连和空间布局协同设计，从单一散热部件优化转向整机柜系统级优化。其价值不仅在于提升散热能力，更在

于通过热、算、连一体化设计，在有限空间内实现算力密度、空间集成密度和互连密度同步提升，为更高密度的Scale-up（纵向扩展）计算域和MW级整机柜提供系统级支撑。对于面向Token生产和智能体运行的人工智能基础设施而言，原生液冷最终实现的并非简单降低芯片温度，而是将有限空间转化为更高的有效算力和更大的单位机柜Token产能。

数据中心评价体系向智能生产效率演进。传统数据中心主要通过PUE（电能利用效率）等指标衡量能源使用效率，但在人工智能时代，仅关注能源效率无法全面反映计算资源是否有效转化为智能生产能力。面向智能体和Token生产场景，评价体系需要从基础设施能效进一步延伸至算力利用效率、Token产出能力以及单位智能任务成本等维度，衡量数据中心投入的能源、空间和计算资源能够产生多少有效智能服务。以一座50MW数据中心为例，液冷方案每年可节省超过400万美元的能源和水费，但更重要的是评估这些投入是否转化为更高的有效算力和Token产出能力。随着能源、数据和计算资源通过基础设施协同转化为Token、模型服务和任务结果，数据中心正在从资源承载平台向智能生产系统演进。

2.5 终端：云边端协同扩展智能边界，终端正成为Agent运行节点

个人AI和行业现场智能应用推动算力向用户侧延伸，终端开始承担本地模型推理、私有知识检索、智能体持续运行、数据处理和设备控制等任务。传统终端主要负责输入和显示，人工智能终端加入本地计算、记忆和行动能力后，可以在断网、低时延或隐私敏感条件下独立完成部分任务。云–边–端协同的分布式计算架构，以及全时感知与自然交互、端侧算力与个人记忆、安全可信与隐私保护、连续计算与跨端协同等能力，正在成为支撑个人AI发展的重要基础。

终端正在成为智能体运行节点。人工智能终端具备本地计算、记忆和行动能力后，可在断网、低时延或隐私敏感条件下独立完成部分任务。终端智能体与云端智能体的分工正在从“云端执行、终端展示”向“端云协同执行”演进，终端智能体负责实时响应、隐私数据处理和离线任务，云端智能体负责大规模推理、跨域知识检索和复杂任务编排。计算架构将从云侧集中逐步走向端云协同，并进一步向端、边、云协同方向演进。

本地智能体设备呈现多形态发展。个人AI工作站包括AI PC和小型GPU工作站，面向开发者、专业用户和小型团队，可在本地运行轻量级智能体并处理隐私敏感数据。桌面级人工智能计算设备在桌面环境提供模型开发和本地运行能力。行业边缘设备部署在工厂、汽车、机器人、医疗设备和门店现场，承担实时推理和控制任务，这类设备需在复杂环境下保持7×24小时稳定运行，对功耗、散热和可靠性有严格约束。

本地与云端形成明确分工。本地终端适合处理隐私数据、低时延、离线和高频轻量任务；边缘节点承担行业现场推理、设备控制和区域协同；云端负责大模型推理、高并发和跨区域服务。三类节点根据任务特征各司其职，共同构成云边端协同的计算体系。

2.6 模型与智能体框架：从模型能力竞争走向智能执行协同

模型提供智能能力，决定智能体能够理解和解决问题的上限；智能体框架则将模型能力嵌入任务执行流程，通过上下文管理、工具调用、状态维护和任务编排，将一次性的模型输出转化为持续、稳定、可控的业务执行能力。随着智能体从简单问答走向复杂任务闭环，竞争焦点正在从单一模型能力扩展到模型、框架与业务流程的系统协同，二者共同构成智能体落地的核心基础。

模型选择决定能力边界。不同模型在推理深度、上下文长度、多模态处理、工具使用和生成成本上存在显著差异。智能体系统需根据任务复杂度、时延、数据安全和成本要求，在通用模型、专业模型和小型模型之间做出选择。单一模型难以同时满足所有任务需求，复杂推理需要大参数模型，简单分类任务用小模型即可，代码执行则需要专门的代码模型。模型路由层根据任务类型、用户意图和当前负载动态选择最优模型，并在多个模型之间进行协同调度，以实现能力与成本的平衡。

智能体框架将模型输出转化为可执行流程。智能体框架位于模型和业务任务之间，负责组装上下文、管理提示词、连接工具、保存状态、控制执行步骤，并处理超时、重试和异常。它使模型能够在明确的流程和权限范围内持续完成任务。智能体框架采用插件化架构将模型、工具、技能、会话、沙箱、存储、智能体循环、任务调度、用户界面等能力拆分为独立组件，支持灵活替换和重组，以适应不同业务场景的需求。

智能体框架的核心组件共同构成控制平面。上下文管理负责选择并压缩任务所需信息，控制上下文长度和数据来源；模型路由根据任务难度、时延和成本选择模型，并支持大小模型协作；工具连接管理API、企业系统、代码环境和其他智能体的调用；状态与记忆管理记录任务进度、中间结果、会话历史和长期记忆；执行控制负责规划、循环、并行、超时、重试、中断和人工接管；安全治理负责身份、权限、沙箱、内容检查、审计和结果验证；可观测与评测记录调用链、模型输出、工具结果、成本和任务成功率。这七个组件相互配合，任何一个组件的短板都会影响整体系统的可靠性。

评估需兼顾模型能力和系统可靠性。模型能力提升能够改善单步推理质量，但智能体任务的稳定完成不仅取决于模型能力，还依赖上下文准确性、工具可用性、状态一致性以及异常恢复能力。因此，智能体评估体系需要从单一模型指标扩展到系统级执行指标：模型侧关注准确率、推理能力和Token成本，智能体框架侧关注任务成功率、重试率、人工接管率和恢复时间，整体系统则关注任务完成时间、结果质量、合规性和综合成本。随着开源模型能力持续提升，Kimi K3、Qwen 3.8等不断缩小与闭源模型的能力差距，LangChain Deep Agents、DeepSeek Harness等智能体框架也在上下文管理、工具编排和任务调度方向形成代表性方案，为智能体稳定运行提供执行基础。

2.7 调度与MaaS：从资源调度走向任务级智能服务编排

算力调度位于基础资源和智能体应用之间，负责将分散的异构计算资源转化为可调用、可组合的智能服务。随着智能体从单一模型调用走向复杂任务执行，调度对象正在从单一计算设备扩展到模型、数据、工具和任务流程，调度能力也从资源分配进一步向任务编排、服务协同和跨域资源管理演进。在这一过程中，MaaS平台逐步从模型托管和调用能力扩展为支撑智能体运行的服务底座。

任务级智能调度成为算力服务核心能力。传统调度主要决定任务使用哪块GPU及分配多少资源。智能体调度还需要考虑模型选择、运行位置、芯片类型、智能体协作方式以及失败后的迁移和重试策略。与传统批处理调度不同，智能体任务具有实时性和交互性特征，执行过程中可能根据中间结果动态调整后续策略，对调度器的自适应能力提出了更高要求。

异构资源统一编排成为规模化运行基础。调度范围涵盖算力资源、模型服务、数据服务、工具服务和治理策略等多个层面。算力资源包括CPU、GPU、NPU和不同地域的集群；模型服务包括通用模型、专业模型和不同规模的模型；数据服务包括知识库、向量数据库、上下文和缓存；工具服务包括企业系统、外部API和执行环境；治理策略包括身份、权限、安全、预算和服务等级。统一编排的目标是让上层应用无需关注底层资源的物理分布和异构特性，通过声明式API描述任务需求，由平台自动完成资源分配和服务组合。

MaaS从模型服务向智能体服务演进。MaaS早期主要提供模型接口，智能体应用则要求其进一步提供模型路由、推理加速、智能体开发环境、工作流编排、知识与工具连接、评测监控和成本控制等能力。MaaS的竞争焦点已从单一的价格比较，扩展至价格、性能、工具链和合规能力的综合较量。

跨芯片、跨云和跨地域调度成为常态。异构芯片、混合云和全国算力网使资源调度范围持续扩大。统一调度需要处理兼容性、数据位置、网络时延、能源成本和服务连续性等多重约束，在保证服务质量的前提下实现资源的最优配置。

2.8 Token运营：以Token产出衡量智能生产效率

Token是连接算力投入、推理服务和智能体任务的核心度量单位，但Token数量本身并不等同于业务价值。基础设施需要同时管理Token生成效率、服务质量和任务结果，算力产业的竞争逻辑正在从算力规模的简单扩张，转向以Token生产效率为核心的智能产能竞争。

Token成为衡量智能生产效率的核心计量单位。能源和设备投入首先形成有效计算，再转化为Token生成、智能体任务执行和业务结果。不同模型的Token口径和信息密度存在差异，跨模型比较需要保留模型、场景和任务条件。Token作为衡量推理工作量的统一单位，为不同模型、不同场景下的算力消耗提供了可比口径，但也需要关注同等Token数量在不同模型下信息密度和推理质量可能存在的显著差异。

Token服务以“交互性–吞吐量”为双维度评价体系。横轴为Interactivity（交互性），反映单个用户获得有效输出的响应速度，决定智能服务的实时体验；纵轴为Throughput（吞吐量），反映系统单位时间内可处理的最大Token数量，决定整体服务规模。两者存在天然权衡：提升交互性需要预留更多计算资源以降低排队时延，但会降低并发服务能

力；提升吞吐量则需要扩大批处理规模，可能增加单用户响应等待时间。因此，Token服务优化并非追求单一指标最大化，而是在具体业务场景和服务等级约束下，在交互性与吞吐量之间寻找帕累托最优解，实现资源利用效率与用户体验的综合优化。

Token度量体系推动智能服务精细化运营。当Token成为衡量智能生产效率的重要指标后，企业需要进一步建立围绕Token的运营管理体系，对模型调用、资源消耗和服务质量进行统一管理。Token运营应覆盖统一模型入口、用户与智能体身份、调用配额、模型路由、成本归集、预算管理、服务质量监控、安全审计和内外部结算。其核心价值在于将分散的Token消耗转化为可管理、可分析、可优化的运营数据。Token支出应从软件采购中分离，作为与人力、供应链同等重要的核心生产资源进行独立核算。

03.

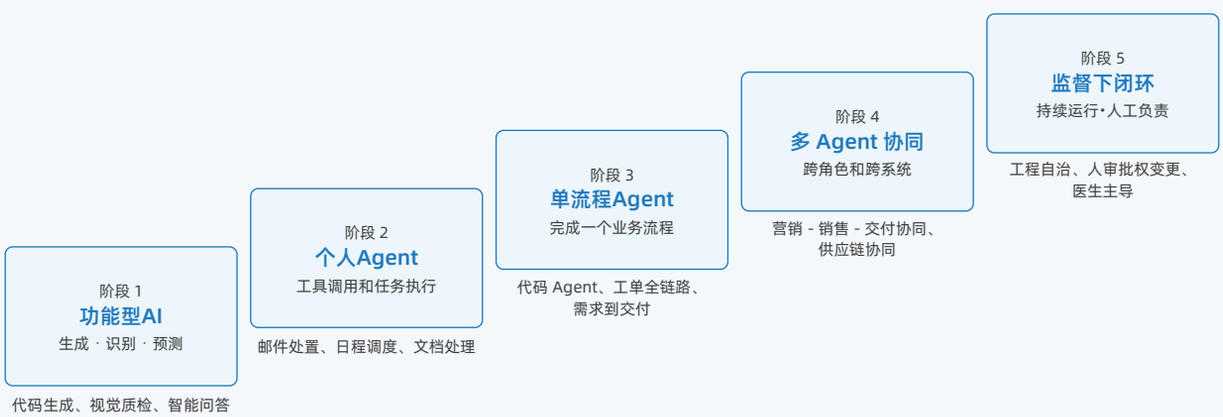
中国人工智能
行业应用发展评估

3.1 中国人工智能行业应用正在进入从单点能力部署向业务流程融合的新阶段

人工智能正在形成“功能型AI→个人智能体→单流程智能体→多智能体协同→监督下闭环”的渐进式演进阶梯。这不仅是模型能力的升级，更反映人工智能介入企业生产活动的深度持续提升：从生成、识别、预测等离散单点任务，到通过工具调用承接个人连续性工作，再到执行完整业务流程、实现跨角色跨系统协同，最终在人工监督与权责约束下形成可持续运行的业务闭环。IDC调研显示，国内企业智能体普遍遵循“优先切入高频、边界清晰、结果可验证场景”的落地规律，能力成熟需分层迭代推进。相应地，企业衡量人工智能价值的标准也在转变，从关注模型能力、部署数量和单点效率，转向智能体能否稳定进入核心业务、持续完成任务并形成可衡量的业务结果。

图10:中国人工智能行业应用发展评估

Agent五级演进阶梯：AI 介入生产活动的深度持续提升



来源：IDC，2026

企业生产方式和组织分工也正在因智能体发生改变。当智能体能够持续承接原本由人完成的部分工作后，人工智能带来的变化就不再只是单个岗位效率提升，而是进一步影响“哪些工作由人完成、哪些工作由智能体承担，以及人机如何协同完成任务”。以OPC（一人公司，One Person Company）等新型组织形态，以及FDE（前线部署工程师，Forward Deployed Engineer）等新型岗位为代表，企业正在探索以更精简的人力配置、更高的智能体协同密度和更强的专业技术能力，承接过去需要多人协作完成的业务任务，推动组织由传统岗位分工逐步向“人定义目标、智能体执行任务、工程师负责部署与优化”的协作模式演进。

这一变化也进一步改变人工智能基础设施和算力的价值衡量方式。过去，算力主要用于支撑模型训练和推理；随着智能体逐步进入业务执行环节，算力开始直接支撑持续性的任务处理和业务运行。由此，算力投入的价值将不再只体现为模型性能或资源规模，而更多通过智能体承接的任务数量、业务流程覆盖范围以及单位人力产出提升，转化为可衡量的业务成果。

3.2 代码开发智能体与办公智能体成熟度领先，多数传统行业仍处于探索阶段

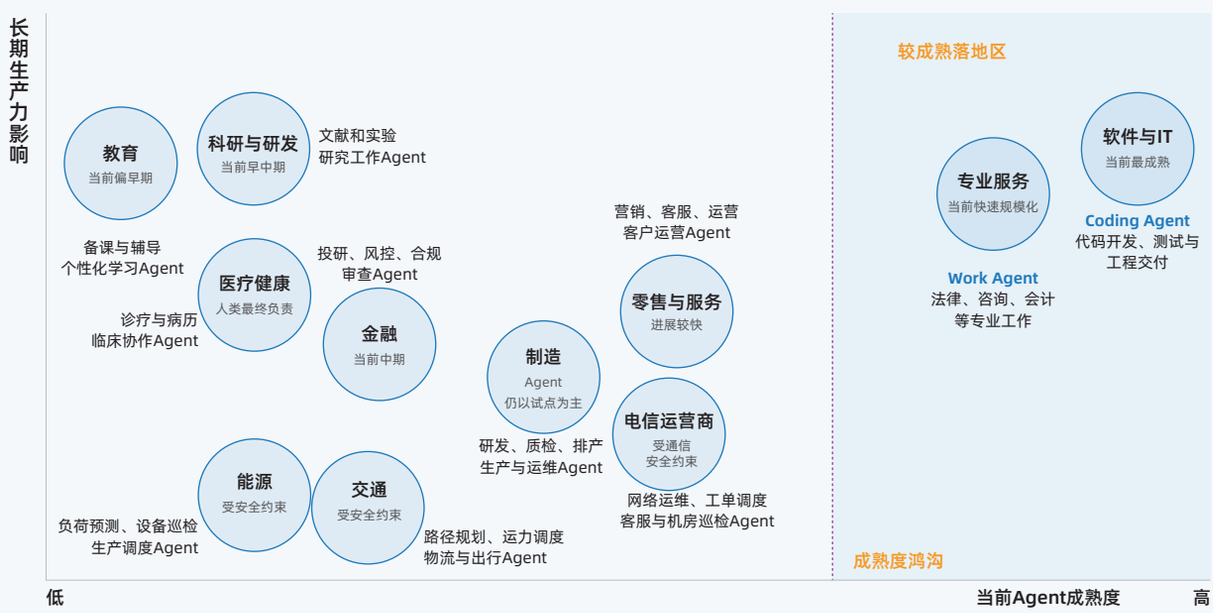
从实际落地情况来看，不同行业/领域受数字化基础、业务流程标准化程度、专业知识密度、专业判断占比以及可被人工智能重构的工作环节等因素影响，智能体应用成熟度与长期产业影响并非完全同步。

具体来看：

- 数字化程度高、任务边界清晰的软件与IT、专业服务行业的智能体成熟度遥遥领先，其中代码开发智能体和办公智能体已进入规模化应用阶段，与其他传统行业形成鲜明对比；
- 零售、电信运营商、制造等行业加速向核心业务流程渗透；
- 金融、医疗等高风险行业则在强化人机协同的同时逐步扩大智能体应用边界；
- 医疗健康、教育以及科研与研发虽然当前成熟度相对有限，但长期生产力影响较为突出；
- 能源、交通则受到物理环境、安全要求等因素约束，整体仍处于较早阶段。

图11:各行业智能体落地成熟度与长期生产力影响分布矩阵

Coding与Work场景成熟度领先，传统行业受业务复杂度与安全约束影响，多数仍处于探索阶段。



来源：IDC，2026

图中以**智能体成熟度和长期生产力影响**两个维度刻画不同行业的发展状态：

横轴代表智能体当前进入业务流程的成熟程度，核心由**数字化基础、业务流程标准化程度、结果可验证性**三要素来决定，越向右说明智能体应用越深入；

纵轴代表人工智能对行业长期生产力的潜在影响，核心由**专业知识密度、专业判断占比、可被人工智能重构的工作环节**三要素来决定，越向上说明人工智能对行业工作方式、专业分工和生产组织方式的重构潜力越大。

（1）代码开发智能体和办公智能体应用成熟度遥遥领先

代码开发智能体：具备高度数字化、流程相对标准化、输出结果可验证等先天条件，是国内智能体落地成熟度最高的业务领域，已进入规模化应用阶段。

- **数字化基础高。**代码开发任务高度数字化，代码、开发环境、代码仓库、测试结果和项目流程具有较强结构化特征，为智能体进行自主规划、工具调用和跨系统执行提供了良好基础。
- **业务流程标准化程度高。**代码开发通常围绕需求分析、代码开发、研发测试和工程交付等相对清晰的流程展开，任务边界和执行步骤较为明确，为智能体从单点功能向连续任务执行演进提供了良好条件。
- **结果可验证性高。**代码能否运行、测试是否通过、功能是否达到预期，都可以通过运行结果和测试体系进行验证，为智能体持续执行任务、验证结果并进行后续调整提供了基础。

当前行业已经由代码生成、代码补全等功能型AI向代码检索、修改、运行和测试等任务型智能体发展，并逐步覆盖需求分析、研发测试和工程交付。未来，随着智能体与开发环境、代码仓库、项目管理和测试体系进一步融合，代码开发将率先形成由智能体持续承接工程任务、人类进行目标定义和关键监督的新型生产模式。这一模式的核心变化在于，智能体正在从提高人员效率的工具，转变为软件生产过程本身的执行主体。其产业意义不仅限于研发效率的提升，更在于为其他行业的智能体产业化提供了从“辅助工具”走向“生产系统”的参考范式。**目前代码开发智能体已在大模型与AI行业率先部署，融入核心业务流程，形成业务执行、反馈和优化闭环。**

办公智能体： 法律、咨询、会计、审计等专业服务是典型的知识密集型行业，业务高度依赖专业知识、行业经验和复杂判断。人工智能应用早期主要承担资料检索、文档生成和信息整理等基础工作，智能体则开始进入合同审阅、尽职调查、审计和研究分析等专业流程。随着专业知识库、行业案例和企业数据进一步结构化，智能体将从“辅助专业人员”逐步转向承担标准化专业任务。长期来看，专业服务行业的变化将更多体现为专业分工重构：由智能体承担资料处理、分析和标准化执行，专业人员进一步聚焦复杂判断、客户沟通和最终责任。个人办公成为智能体规模化普及的核心入口。办公场景通用性强、落地门槛相对更低，人工智能能力已经由写作生成、信息检索、会议纪要等单点功能，拓展到邮件处置、日程调度、文档处理等长序列任务。个人办公智能体同样开始规模化应用，具备了调用办公套件、处理多源信息、持续执行任务的能力，人机交互模式由传统“一问一答交互式问答”转向持续性任务承接。未来随着企业知识库、内部办公系统、业务工具深度打通，个人办公智能体将从个体效率工具，进一步延伸至团队协作与组织级知识运营层面。

（2）传统行业：智能体应用整体仍处于早期，制造、医疗、能源以探索试点为主，电信运营商进展相对更快

当前，制造业、医疗健康、能源/电力等典型传统行业已经开始将人工智能应用嵌入具体业务流程，但智能体整体仍处于**探索试点和局部流程应用阶段**，真正进入核心业务并形成“执行–反馈–优化”闭环的比例均为**0%**，相比之下，大模型与AI行业已有**21.7%进入这一阶段**。这说明传统行业虽然已经开始部署人工智能，但智能体真正成为核心业务执行主体的案例仍较少。

电信运营商在传统行业中进展相对较快。其“智能体已接入部分业务流程”的比例已经达到70.0%，显著高于其他行业，说明其智能体应用已经从探索试点较快进入任务链路自主执行阶段；但深入核心业务并形成闭环的比例仍只有3.3%，表明从局部自主执行进一步走向核心生产流程仍处于早期。

整体来看，**传统行业普遍面临技术适配难、业务流程固化、组织变革阻力大等挑战，导致智能体应用仍停留探索试点阶段，距离规模化、深度化落地仍有显著差距，远未达到实用化水平。**

对于尚处较早阶段的传统行业，智能体的规模化落地难以依靠行业自身自发完成，标杆案例的示范和引导作用尤为关键。小马智行、美的等具备前瞻意识的企业已在各自领域形成可参考的转型路径，为传统行业验证了智能体落地的价值与方向。

零售与服务：从内容与交互智能向业务运营智能深化

零售与服务行业业务线上化程度较高、用户数据丰富、经营反馈周期较短，是智能体规模化落地较快的领域。人工智能已经广泛应用于内容生成、推荐和客户服务，智能体进一步向营销、销售、客户运营和服务交付等连续任务渗透。在电商、直播、游戏等数字化程度较高的业务中，智能体正在从内容和服务工具向业务运营执行者转变，逐步将用户洞察、内容生产、营销决策和执行反馈连接起来，部分场景已经形成较为完整的业务闭环。IDC调研显示，**36.7%的相关企业已经实现智能体接入业务流程，21.7%的企业形成人工智能与核心业务的闭环优化，**表明零售与服务正成为人工智能由单点应用向业务闭环演进的重要先行领域。

电信运营商：从单点辅助向通信服务全链路智能化演进

电信运营商是典型的运营与资产密集型行业，智能体成熟度处于中等偏上水平。作为通信基础设施的运营主体，电信运营商在安全合规方面具有特殊行业约束，当前人工智能应用正从客服问答、网络告警等单点辅助，向故障根因分析、工单自动派发与闭环处置、机房无人巡检等流程型智能体延伸，网络运维、工单调度、客服与机房巡检是当前应用的核心场景。IDC调研显示，**70.0%的电信运营商已将接入智能体业务流程，63.3%实现初步效率提升。**随着网络运维数据、客服系统与业务平台的进一步打通，电信运营商将形成覆盖运维、调度、服务与巡检的全链路智能体协同体系，在通信安全约束下持续提升服务效率与网络可靠性。

制造：从局部智能优化向生产流程智能化演进

制造业是智能体进入实体经济的核心领域，智能体整体处于由局部应用向流程级应用加速推进的阶段。制造业此前已经较大规模应用视觉质检、故障预测等传统人工智能技术，当前正在向生产排产、设备运维、工艺调度、供应链协同等流程型智能体演进。但与纯数字化场景相比，工业智能体需要连接ERP（企业资源计划）、MES（制造执行系统）、SCADA（数据采集与监视控制系统）、工业机器人等异构系统，并满足实时性、可靠性和生产安全等要求，因此整体演进速度相对较慢。随着生产设备、业务系统和边缘节点进一步互联，智能体将由数字空间中的信息处理和任务执行逐步延伸至实体生产过程。未来，智能体将进一步连接订单、采购、生产、质量和设备等环节，根据实时生产状态进行任务拆解和资源调度，推动制造业由“局部智能优化”向“全流程动态协同”演进。IDC调研显示，**47.2%的制造企业正在试点智能体，制造智能体优先级达到83.3%，**表明制造业已经成为智能体向实体产业纵深渗透的重要方向。

与此同时，具身智能正推动人工智能从数字空间走向物理世界。机器人企业是该赛道的核心力量，IDC调研显示，64.4%的机器人企业正在试点智能体应用，22.2%将多智能体协同列为优先方向。搭载异构智能体的机械臂、人形机器人、多足机器人与轮式机器人已在探索中形成初步的自主协作成果，端到端感知-决策-控制链路自动调优智能体与工业作业场景技能学习智能体正在推动机器人向环境自适应、任务自主决策进化，并在工厂巡检、仓储分拣等领域完成多个规模化试点。

最佳实践

小马智行：以世界模型2.0加速L4规模化落地

小马智行成立于2016年，是一家专注于L4级自动驾驶技术研发与商业化的科技企业，业务涵盖Robotaxi、Robotruck和智能解决方案三大板块。目前，小马智行已在北京、上海、广州、深圳开展全无人Robotaxi商业化运营，并在广州和深圳实现城市级单车盈利；业务同时拓展至欧洲、中东及亚洲多个国家和地区，全球车队规模接近2000台。在Robotruck领域，第四代无人重卡已进入量产阶段且完成车规级测试验证，现与招商港口在深圳妈湾港展开商业化部署。

同时，小马智行依托L4级自动驾驶技术积淀打造的车规级智能域控制器小马方载，可赋能无人配送、环卫清扫、港口、矿区等多元场景，产品已获多家行业头部公司量产定点，客户覆盖全球数十个国家。

■ **挑战：规模化落地对世界模型提出更高要求**

L4级自动驾驶对安全性要求极高，模型训练目标不是“像人开得一样好”而是“比人开得好”，这意味着研发范式必须从模仿学习转向强化学习。为此，小马智行自主研发了PonyWorld世界模型，让人工智能在“虚拟驾校”中反复训练车端模型的驾驶能力。然而随着运营规模持续扩大、场景复杂度不断攀升，世界模型1.0在实际应用中遇到新的课题：

首先是规模化效率瓶颈。仿真验证环节每周产生数十亿条量级的数据，规模远超人工可分析的范围。在运营范围扩大的同时，以有限的人力处理高速增长的数据，成为PonyWorld迭代升级的核心方向之一。

其次是Sim-to-Real Gap（仿真与真实的差距）。只有世界模型的精度足够高，人工智能司机才能在其中取得正向训练结果，否则模型能力可能越学越错。随着人工智能司机能力远超人类，进一步提升精度、主动发现自身不足并实现定向进化，成为新的技术挑战。

第三是长尾场景与海外数据合规的双重考验。“开门杀”等极端场景在十年运营中仅遇几次，传统方法以季度乃至更长时间为单位才能完成收集与验证。海外业务拓展进一步加剧了这一挑战——多个国家和地区要求当地采集的运营数据不得出境，公司既无法将海外数据传回国内集中处理，也难以在每个运营国家都建立独立的数据中心。

■ **方案：PonyWorld世界模型2.0——一个会自我进化的世界模型**

针对上述挑战，小马智行推出PonyWorld世界模型2.0，核心突破在于系统具备了人工智能驱动的自我诊断与定向进化能力。在研发环节，小马智行构建了覆盖数据分析、代码建议、模型训练与效果评估的自动迭代闭环。在算力基础设施层面，通过高性能计算集群与并行存储系统支撑世界模型的大规模分布式训练与仿真验证，在PB级数据规模下实现高聚合带宽，有效支撑千卡级GPU集群高效运转。

随着PonyWorld世界模型2.0的上线，在效率方面，系统可在仿真完成后自动识别模型薄弱环节并给出根因分析，将分析从人工抽样升级为全量自动评估，并根据车端模型薄弱环节自动生成针对性训练场景，大幅减少无效训练开销，分析周期从“周级”压缩至“天级”。在自主进化方面，系统能够自主判断模型表现不足的场景，主动生成定向数据采集任务，例如推送指令要求采集特定时段、路口的特定场景数据，形成人工智能引导下的迭代闭环。在长尾场景与海外合规方面，系统结合互联网素材与生成式仿真手段主动合成训练数据——以中东校车为例，从互联网获取校车外观、信号灯含义后直接生成仿真场景，在车辆抵达前完成模型适配，将corner case验证周期从季度级缩短至周级，同时规避数据出境合规风险。

■ **项目收益**

PonyWorld 2.0落地为小马智行带来显著的效率提升。模型迭代周期从“周级”压缩至“天级”，corner case的收集与验证周期从“季度级”缩短至“周级”，研发迭代速度实现数量级跃升。远程监控人车比已具备从约1:3提升至1:30至1:50的技术条件，单车运营成本持续优化。同时，通过世界模型的仿真与生成能力，可在车辆实际部署前完成海外特殊场景适配验证，将海外部署周期大幅压缩。随着小马智行向更大规模车队和更多海外市场拓展，世界模型2.0构筑的“精度飞轮”将持续转动——大规模的商业运营产生高价值数据，驱动世界模型精度提升，进而增强车端模型能力，支撑更大规模无人车队部署，由此形成越跑越快的正向循环。

美的集团：以数字员工体系驱动制造业全链路智能化升级

美的集团是一家成立于 1968 年的全球化科技集团，业务覆盖智能家居、工业技术、智能建筑科技、机器人与自动化等板块。2024年7月将人工智能正式确立为集团级战略，目前算力规模约1500至1600 PFLOPS（每秒千万亿次浮点运算），日均Token消耗达1500亿。围绕“自上而下”端到端流程重构与“自下而上”全员创新两条主线，美的构建了统一的智能平台，推动人工智能能力从单点工具逐步演变为可调度、可管理、可组织的数字员工体系，已在HR系统中为数字员工设立正式工号、标准岗位和管理上级。

■ **挑战：端到端业务流程重构的多重阻力**

随着数字员工从单点辅助向端到端业务流程延伸，美的面临多重挑战。首先是业务变革阻力，端到端重构意味着要打破原有流程节点、组织架构乃至岗位设置，这是推进中最大的障碍。其次是知识准确性的挑战，企业级应用对准确率要求普遍在90%以上，知识的一致性和标注数据的统一管理至关重要。在算力与成本方面，随着模型精度与长文本处理需求提升，叠加近两年存储硬件成本上涨，高算力供给缺口持续存在。异构算力管理同样构成工程挑战，不同厂商、不同架构的算力如何纳入统一调度体系，是算力平台建设的核心问题。

■ **解决方案：自上而下与自下而上相结合的数字员工体系**

在自上而下层面，美的由各领域负责人牵头，业务专家与算法工程师组成联合小组，对核心业务流程进行系统性重构。以供方引入为例，传统流程涉及研发、品质、供应链、工厂等多部门评价，周期约60天；现由数字员工自动发起供方评价、完成寻源与数据核验，仅在异常时转人工处理。新物料新增流程从7天缩短至最快半小时，并根据历史准确率逐步放开人工审核。

在自下而上层面，美的依托“小美”平台让每位员工都能低代码创建智能体，将个人经验沉淀为可复用的Skill。目前平台上智能体总量已超过4万个，可自主完成端到端任务的智能体约190个。

在算力基础设施层面，采取自建与外采并重的混合模式，本地与云端算力比例约为1:1，通过统一模型网关屏蔽底层算力差异，使前端开发人员能够无感调用各类算力资源。

■ **项目收益**

截至2026年8月，数字员工体系已为美的带来显著经营回报。系统测算提效金额接近10亿元，实际体现财务收益4到5亿，累计节省工时接近1500万小时。供方引入流程从60天缩短至数小时，新物料新增流程从7天缩短至最快半小时。美的已建立与工信部联合制定的三级智能体等级认证体系，超过12000名员工通过认证。随着端到端流程变革的持续推进，数字员工与人类员工协作共存将成为美的长期运营的主流形态。

金融：专业智能体持续深化，自主执行保持审慎

金融行业智能体成熟度处于中等水平，但专业化程度较高。金融业务高度依赖数据分析、风险识别和专业判断，人工智能应用正在由客服、营销等外围环节进一步向投研、风控、合规和保险等核心业务渗透。当前智能体的价值主要体现在复杂信息处理和专业分析效率提升，包括资料检索、因子挖掘、风险监测和合规审查等。随着行业知识、实时数据和业务系统进一步融合，智能体将逐步从“提供分析建议”向“执行标准化专业任务”发展。调研显示金融行业专业领域智能体的优先级已达到79.5%，反映出行业对专业化智能体的强烈需求。但金融行业涉及重大风险和责任承担，智能体自主执行仍存在明确边界，未来更可能形成“高自主执行、强人工监督”的应用模式，而非完全无人化。金融行业正在构建统一人工智能开发与驾驭平台，目标是让所有智能体实现可调用、可追溯、可干预、可回滚、可评估，确保智能体嵌入核心业务流程时完全符合合规要求。

值得注意的是，尽管银行、保险等传统金融业态在人工智能落地上面临与制造、能源类似的合规与流程制约，**量化交易却凭借其数据驱动、算法密集、决策链路短的行业特性表现出相对更快的智能体渗透，虽然核心业务闭环比例仍为0%，但已有35.9%进入自主完成特定任务链路阶段。**

医疗健康：专业化深度高于应用广度，坚持人类最终负责

医疗健康行业当前智能体成熟度相对有限，但专业化程度较高。诊疗、药物研发和患者管理等业务具有较高专业门槛，同时涉及医疗安全、隐私保护和责任归属，因此行业整体仍以“人工智能辅助、医生负责”为主要发展路径。当前智能体已经从影像分析、病历处理等单点应用向药物研发、临床辅助和患者管理等专业流程延伸。IDC调研显示，**56.7%的医疗企业已进入智能体试点阶段，与此同时，36.7%的医疗企业已主动规划多智能体协同，比例居各行业首位，**这与诊疗流程天然涉及问诊、检查、诊断、治疗、随访等多专业协作密切相关。

在科研创新方面，医疗行业走在前列，**40.0%的企业将科研创新与仿真计算列为优先人工智能方向，研发设计智能体优先级达到90.0%，**人工智能驱动的靶点发现全流程赋能已成为行业标配，人工智能辅助的药物筛选、蛋白质结构预测与分子动力学仿真正在将新药研发周期从传统的“十年十亿美金”模式大幅压缩。未来，医疗智能体的核心突破并非单独提高自主程度，而是提升专业判断的准确性、可解释性和可验证性，在明确责任边界的前提下承接更多专业任务。

最佳实践

数力聚：人工智能推动心理健康服务从被动筛查走向持续监测与人机协同干预

青少年心理健康问题已成为公共卫生与社会治理的重要议题。当前，我国青少年心理健康服务需求持续增长，但服务体系仍面临筛查预警机制不健全、专业力量供给不足、家校医社协同不畅等结构性短板。数据显示，中国儿童青少年精神疾病整体患病率约为8.9%，而全国精神科医师每10万人口仅约3.2名，超过50%的中小学未配备专职心理教师。供需之间的巨大落差，迫切要求心理健康服务模式进行系统性升级。

■ **挑战：传统筛查体系的效率瓶颈与模式局限**

传统心理健康服务高度依赖PHQ-9等主观量表，患者可能因病耻感隐瞒真实情况，量表结果也易受瞬时情绪干扰，存在滞后和失真问题。在大规模筛查场景中，一位学校心理老师往往同时承担备课和行政工作，千人规模学校的心理筛查需占用两到三周时间，流程繁琐且效率有限。更根本的问题在于，传统筛查以事件驱动为主，缺乏持续监测能力，深夜等非工作时段成为服务真空，危机风险难以及时发现。心理健康服务亟需从“出了问题再干预”的被动模式，转向“持续监测、主动预警”的主动模式。

■ **方案：以多智能体协同推动心理健康服务模式重构**

数力聚（北京）科技有限公司成立于2021年，聚焦人工智能算法与数据治理领域，并将心理健康作为重点应用方向，自主研发了人工智能心理健康平台，面向青少年、高校学生及医疗机构人群，提供心理筛查、评估、疏导与危机干预服务。围绕心理健康服务的全流程，数力聚以多智能体协同架构为核心，从评估客观化、交互感知和服务连续性三个维度推动服务模式升级。

在评估客观化方面，数力聚基于近30万条真实临床数据打造了人工智能血液检测抑郁症模型，通过分析常规血液体检报告中的相关生物标志物生成风险提示，为传统量表筛查提供额外的客观信息参考。该模型与MDD LLM深度评估模型形成双模型协同路径，前者负责全量初筛，后者对高风险人群进行深度评估。

在交互感知方面，平台采用结构化CBT认知行为疗法的对话骨架，融合语音语调与文本语义的双模态情绪识别，将心理测评融入自然对话过程，降低传统问卷式交互带来的压力和抵触感。

在服务连续性方面，平台设计了多智能体协同架构。陪伴、咨询、测评和情绪分析智能体在后台并行运行，调度器智能体负责统一协调，避免交互突兀。当系统识别到危机相关表达或异常情绪信号时，相关智能体进行风险提示，并按照预设机制升级至人工专业人员处理。平台提供7×24小时的交互支持；经用户授权产生的相关记录按照数据安全与权限管理要求进行本地化存储，为后续连续服务提供参考。

■ **项目收益**

人工智能介入后，学校心理筛查从两到三周压缩至几个工作日，个案访谈有效时长提升近一倍，7×24小时服务覆盖了传统模式无法触及的深夜时段。在权责边界方面，数力聚坚持人工智能负责信息支持与辅助评估，不替代临床诊断，完整量表分数仅提供给教师和医生，危机场景下，人工智能进行风险信号识别、提示和辅助支持，并按照预设机制及时转交专业人员进行判断和干预。这一实践的意义不仅在于效率提升，更在于服务模式的转变：当人工智能能够持续在场、主动预警并协同专业力量时，心理健康服务正在从一次性的量表筛查，走向覆盖日常监测、风险识别与分级干预的持续管理闭环，为缓解专业力量不足与青少年心理服务需求之间的结构性矛盾提供了一条可复制的技术路径。

教育：当前成熟度有限，长期生产力重构潜力突出

教育行业当前智能体成熟度相对较低，但长期生产力影响较大。当前人工智能应用主要集中于备课、内容辅助、智能问答和个性化学习，智能体正在进一步探索学习计划制定、学习过程跟踪和教学评价等连续任务。未来，教育智能体有望逐步改变教师与教学资源之间的分工方式，使教师从大量重复性的备课、辅导和评价工作中释放，更加聚焦教学设计、学生沟通和复杂教育判断。因此，**教育行业当前的智能体成熟度并不能完全代表其长期产业价值，其生产力影响将随着智能体能力增强而逐步释放。**

科研与研发：从知识获取向知识生产协同演进

科研与研发属于高知识密度领域，当前智能体仍处于探索阶段，但其长期生产力影响较为突出。人工智能已经广泛进入文献检索、知识整理和数据处理，并开始向实验设计、仿真分析和研究任务管理等连续流程延伸。

AI for Science正在重塑基础科研与工程创新的方法论，随着科研数据库、实验设备和仿真工具进一步连接，智能体有望从辅助科研人员获取知识，进一步参与知识发现和实验探索，推动科研生产方式由**“人提出问题、工具辅助验证”向“人定义研究目标、智能体持续探索”**演进。

2026年9月，OpenAI宣布仅用88小时、调度约1万个智能体协同运行，便给出了纳维-斯托克斯方程存在性与光滑性问题的证明——这道克雷数学研究所七大“千禧年难题”之一的题目，此前已悬置约九十年。在这一过程中，人类研究人员仅设定研究目标，上万智能体分组并行探索、交叉汇总中间结果，累计消耗约1,300亿个输出Token；核心证明形成后，GPT-6 Astra再用约17小时将166页证明全文转写为Lean形式化代码，使证明的每一步都可由计算机核验。

能源：安全约束下稳步推进，向运行协同演进

能源行业受到物理系统复杂性、实时运行要求以及安全责任边界等因素制约，当前智能体成熟度相对较低，应用仍以预测、巡检和调度辅助为主。随着实时数据、数字孪生、边缘智能和业务系统进一步融合，智能体将逐步具备持续感知、预测和动态调度能力。该领域的核心发展方向并非追求完全无人化，而是在安全边界内**逐步扩大智能体自主决策和执行范围，推动能源系统由被动响应向主动运营转变。**电网调度领域已开始应用于调度策略推演，以计算实验替代部分高成本现场调试；在流域梯级电站联合调度中，多智能体协同已实现水文预测、水库调度与生态流量管控等多环节联动；虚拟电厂场景中，多智能体已参与分布式资源的实时聚合与调度响应。

交通：从单一路径规划走向多模式出行协同

交通行业涉及人、车、路、货多要素动态匹配，实时性和安全要求较高。当前人工智能应用仍以路径规划、运力调度、物流优化等单环节辅助为主。随着实时路况数据、车路协同和边缘计算能力持续提升，智能体将进一步向跨方式出行衔接、动态运力匹配和物流全链路编排延伸。在智能驾驶场景中，车端智能体负责实时环境感知与决策执行，路侧边缘节点完成车路协同与局部交通优化，云端则承担模型训练、仿真验证与全局路径规划。其演进方向是在安全合规框架下，让智能体承担出行方案生成、运力实时匹配和运输路径动态优化等任务，提升交通网络的整体运行效率与出行服务体验。

04.

中国人工智能算力地域发展评估

2026年，中国人工智能算力基础设施建设进入全面提速期。本报告的评估框架从投资规模、政策支持、算力供给、技术成熟度、人才供给五个维度出发，对各城市人工智能算力发展水平进行综合评估。鉴于智能体与具身智能技术的战略重要性日益凸显，本年度的城市评估框架将各城市在智能体应用落地、具身智能产业布局、大模型开发及推理算力基础设施建设等领域的投资力度、建设进度、政策支持、应用水平和规划布局纳入关键评价指标体系。

从整体排名来看，北京、杭州、深圳构成第一梯队。北京凭借完整的算力供给产业链、实力雄厚的模型与人工智能应用企业稳居首位；杭州依托阿里巴巴、DeepSeek等龙头企业，叠加阿里云算力投资与All in AI的整体投入紧随其后；深圳抓住智能体爆发机遇，腾讯算力资本开支巨大，“算力饥渴”程度远超其他头部云厂商，WorkBuddy、混元等模型与应用快速放量，排名超越上海跻身前三。上海、广州、成都、南京、武汉、济南、苏州等城市紧随第一梯队，公共算力资源和人工智能人才储备持续提升。领先城市与追赶城市之间的差距有所拉大，后续城市需在算力供给、应用落地和生态建设上持续发力。

图12 中国人工智能计算力发展评估——城市排行

来源：IDC，2026

北京

北京继续领跑全国人工智能算力发展，稳居城市排名首位。2025年人工智能硬件投入高达176.84亿美元，相关企业超2,500家，人工智能领域融资达到全国四成以上。算力供给方面已形成从芯片设计、服务器制造到智算服务的完整产业链，海淀区集聚400余家人工智能企业，形成要素高效聚合的“十分钟创新圈”。今年发布《北京人工智能创新高地建设行动计划》，力争两年左右推动人工智能核心产业规模突破万亿元；亦庄同步推出全市首个“Token经济”专项政策，对开展推理效能跃升的企业给予最高2,000万元资金支持，依托一系列举措在“十五五”期间打造“京内筑基、京外拓源”算力格局。目前已完成备案大模型259款，在科学智能、具身智能等前沿领域保持全国领先。人才方面，集聚全国超40%的人工智能顶尖人才，清华、北大等14所高校设立人工智能学院，入围AI 2000全球最具影响力学者榜单148人，在全球人工智能最具创新力城市中排名第二。

杭州

杭州作为排名第二的城市，凭借在具身智能、大模型、脑机接口等前沿领域的持续突破，以及从“六小龙”到“新八骏”不断涌现的创新企业集群，进一步巩固了其全国人工智能技术创新与产业应用高地的地位。2025年浙江省人工智能硬件投入48.08亿美元，其中杭州占比最高。算力供给方面，依托阿里巴巴等龙头企业，形成从基础模型研发到智算云服务的完整算力生态，全国最大液冷万卡集群已投用。开源生态方面，魔搭社区汇聚超500家贡献企业、服务全球逾2,900万开发者，成为全国最重要的开源模型与算力服务枢纽之一。当地“十五五”规划提出打造人工智能、视觉智能2个万亿级集群，明确支持阿里千问、DeepSeek迭代升级，加快培育第三个开源基础大模型。具身智能机器人“强链补链”行动方案同步推进，国家级具身智能中试基地已落子滨江。人才方面，人工智能岗位渗透率达28.54%、位列全国第二，全市引进35岁以下大学生累计超过250万名。

深圳

深圳排名第三，作为全国人工智能服务器与算力芯片制造的重要基地，在算力供给侧的产业厚度和硬件生态位居全国前列，并把握住今年智能体爆发带来的需求增长机遇，凭借算力需求和应用落地的突出表现，排名超越上海。2025年人工智能硬件投入约37.43亿美元，规上企业超2600家。算力供给方面汇聚了昇腾生态以及腾讯云等核心力量，其中腾讯围绕全模态数据治理与算力统筹持续投入，WorkBuddy、混元模型与深圳百亿算力平台形成协同，在头部云厂商中形成了兼具技术厚度与地域特色的实践路径。当地出台《深圳市推动人工智能与应用发展行动计划（2026–2028年）》，提出新一代智能终端、智能体等应用普及率超90%，支持工业智能体研发，加快智能体在参数寻优、代理仿真等场景的落地。人才方面，2026年上半年人工智能等新兴产业吸纳就业5.48万人，人工智能岗位投递占比17.43%、居全国第三。

上海

上海排名第四，展现出全栈生态优势，但在智能体浪潮中面临追赶压力，需要进一步加速应用落地节奏，将产业基础转化为更快的市场响应能力。2025年全市394家规上人工智能企业产业规模同比增长39.5%，人工智能硬件投入33.16亿美元。算力供给方面，壁仞科技、沐曦科技等一批算力企业集中发布最新智算产品，187款大模型完成备案，全国首款具身智能交互大模型完成备案。“十五五”规划新增打造人工智能万亿级产业集群，聚焦“芯–模–云”生态体系建设，重点发展大模型和智能体、科学智能、具身智能、智算产业体

系，同时大力推进智能体即服务等新商业模式。目前多款开源基座模型能力跻身全球前列，办公智能体用户量稳居全国第二，引领长三角智能体应用深度嵌入真实工作流。人才方面，人工智能人才规模已接近30万人，约占全国三分之一，14所高校成立人工智能创新学院。

广州

广州排名第五，算力基础设施持续扩容。2025年广东省（不含深圳）人工智能硬件投入约21.49亿美元，广州占比最高。广州市智算运行服务平台已整合全市21家智算中心资源，94款大模型通过备案，数量居全国第三，聚集人工智能相关企业超2200户，覆盖“基础层–技术层–应用层”全链条。2026年出台《广州市人工智能产业2026年工作要点》，以市场为主导开展“算力规模跃升”行动，聚焦人工智能、智能网联新能源汽车、金融、虚拟现实等业务场景，布局城市智算中心和分布式边缘计算中心。人才方面，海珠区已集聚近10万名人工智能与数字经济人才，推动上下游产业链集聚产业人才超30万人。

成都

成都排名第六，排位连续两年稳步提升，在西部人工智能产业版图中占据了不可替代的核心位置。凭借14个大型算力设施以及全球首个具身智能工程机器人多模态大模型的发布，成都在算力基础设施建设和具身智能赛道上加速发力，2025年四川省人工智能硬件投入约2.28亿美元，成都作为核心城市占比最高，集聚上下游企业超1200家，26款大模型通过国家备案。2026年9月，全国首个人工智能地方政府规章《成都市促进人工智能产业发展办法》正式施行，明确推动新一代智能终端、具身智能、智能体等前沿领域高质量发展。四川省“十五五”规划进一步明确提升其国家新一代人工智能创新发展试验区和应用先导区发展能级。人才方面，当地发布“卧龙九条”，面向顶尖人才创设最高10亿元综合支持机制。

南京

南京排名第七。全市集聚人工智能核心企业突破500家、相关企业3600余家，已投用智算中心6个，41款大模型通过国家备案，从业人员数十万人，形成中国（南京）软件谷、南京软件园、江苏软件园“一谷两园”核心产业格局，为人工智能发展筑牢产业底座。当地“十五五”规划明确加快建设具有全国影响力的“人工智能+”先导城市，围绕“强协同、给场景、建生态”发力，聚焦软硬协同联动布局“人工智能+软件”，打造算力集群中心城市和Token之城。

武汉

武汉排名第八。作为新上榜城市，依托“光谷智能体引力计划”全域布局，凭借从芯片到应用的全链能力实现快速进位。2025年湖北省人工智能硬件投入约6.82亿美元，武汉占比最高，已建成8个智算中心、1个超算中心，“通算 + 智算 + 超算 + 边缘算力”四算协同的算力矩阵初具规模。在算力供给方面，当地已集聚一批服务器制造、光模块、芯片设计等领域龙头企业，2026年上半年光通信与存储器件量价齐升，电子设备制造业增加值增长90.8%，对规上工业贡献率达100.4%。汉江湾人工智能产业园已落地百度云千帆大模型创新中心，蚂蚁数科、零一万物等头部企业聚集。人才方面，武汉拥有90余所高校，其中33所已设立人工智能学院；在全球高产出、高被引人工智能科学家数量排名中，武汉位列第六。

济南

济南排名第九。依托“中国算谷”产业生态建设，当地智能算力规模持续突破，山东首个移动云Token中心已落地，打通Token生成、传输、应用全链条。2025年山东省人工智能硬件投入约3.28亿美元，济南占比最高，人工智能企业近600家，29个大模型通过国家备案。当地同步推进“人工智能+制造”行动方案与“十五五”规划，以“数据要素×”和“人工智能+”为双核驱动，力争到2028年打造40个以上特色化工业大模型与智能体。2026年初济南将人工智能纳入高层次人才分类认定体系，形成覆盖15个细分领域的“1+N”人才认定政策架构，已发布31项人工智能专项认定目录。

苏州

苏州排名第十，是江苏省第二个跻身榜单的城市，其优势在于不可替代的硬件制造底座，中际旭创、天孚通信、东山精密等企业已成为人工智能算力互联产业链的核心力量，精密制造能力能够将研发成果高效转化为可量产的高品质产品。算力供给方面，依托苏州工业园区公共算力服务平台整合智算资源，通过算力补贴降低企业用算成本。政策层面，出台“人工智能+”城市专项措施，在企业创新、产业培育等方面提供多维支持。技术成熟度方面，已有十余款大模型通过国家备案，昆山建成全国首个“超算、智算、通算”三算齐全的县级市。人才方面，中科大苏州高等研究院、苏州大学等高校形成产教融合集聚，构建起从基础研究到应用转化的人才培育链条。

从区域发展格局来看，2026年中国人工智能算力城市发展呈现三大特征：京津冀集群以北京为创新核心，算力规模全国领先，区域协同效应持续增强；长三角与粤港澳集群中，杭州、深圳、上海、广州、南京、苏州等地智算中心投资和大模型应用场景数量持续提升，产业融合度高；中西部及东部新兴城市中，成都依托西南算力枢纽地位加速建设，武汉、济南则分别作为长江中游和环渤海经济圈的重要节点，凭借差异化的产业定位快速崛起。随着“十五五”规划的全面落地，各城市围绕算力基础设施建设、大模型孵化、智能体与具身智能布局形成了差异化竞争路径，中国人工智能算力的城市格局进一步走向多极化发展。

05.

行动建议

智能体驱动人工智能规模化应用，Token消耗持续增长、智能体基础设施重构、行业落地效果分化、算力供需缺口持续扩大。在此背景下，IDC分别针对行业用户、解决方案提供商和产业规划者提出行动建议，助力国内智能体产业与算力基础设施协同成熟，释放智能体的生产力价值。

对行业用户的建议

■ **立足业务价值制定智能体长期演进路线，分阶段推进落地**

企业应以业务价值作为出发点，遵循“功能型AI–个人智能体–单流程智能体–多智能体协同–监督下闭环”的演进阶梯规划建设路径，优先选择高频、边界清晰、结果可验证的场景开展试点，避免盲目追求多智能体协同、完全自主闭环等高难度目标。金融、医疗等高风险行业坚持“人工智能增强、人类负责”原则，明确智能体的自主执行边界，保留关键节点人工终审；制造、零售等行业优先打通单业务流程闭环，再逐步向跨部门协同演进。评估指标应聚焦任务完成率、业务提效、成本节约、人工接管率等业务结果，避免陷入技术指标导向。

■ **构建适配智能体负载的混合算力架构，平衡成本、供给与安全**

企业应优先采用“自建Token工厂+公有MaaS+边缘算力”的混合部署模式。涉及核心商业机密、敏感客户数据或关键业务逻辑的智能体应用，应优先私有化部署，保证数据自主、Token自控；通用性、探索性强的应用宜借助公有MaaS快速上线；时延敏感的实时场景则依托边缘算力就近承载。三者按需组合、动态调配，构成安全、灵活、实时的智能体算力底座。

■ **完善智能体全栈治理体系，将安全与可观测前置到设计阶段**

随着企业内部智能体数量快速增长，安全、权限、审计、可观测不再是后期补充能力，而是规模化运行的前置条件。企业需要建立智能体注册管理机制，统一身份、权限隔离，明确每个智能体可调用工具、可访问数据、可执行操作的边界；构建完整链路审计日志，做到行为可追溯、异常可回滚。搭建可观测平台，同时监控模型侧指标与智能体框架执行层指标，及时发现幻觉、任务死循环、越权访问等风险。随着Token使用规模扩大，有条件的企业可建设企业级Token管理平台，统一模型入口、调用配额、模型路由、成本归集、预算管理和安全审计，实现Token资源的统一管理与运营。在数据与记忆底座方面，构建分层记忆体系，打通内部知识库、业务系统、向量数据库，搭建统一数据访问空间。同步完善数据治理，做好数据权限分级，控制智能体的数据访问范围，避免核心敏感数据无限制向大模型输入。在选型时重点考察存储系统对跨会话KV Cache、高频随机访问、权限管控的支持能力，而非仅关注存储容量。

对解决方案提供商的建议

■ **强化系统创新与智能体基础设施全栈能力，提升算力转化效率**

智能体系统由模型能力与智能体框架执行治理体系共同构成，厂商应发挥我国在应用、模型和整机系统上的综合优势，推动应用、算法、系统与芯片跨层贯穿式协同设计优化，提升算力资源利用率和Token生产效率。同时补齐上下文管理、工具编排、状态记忆管理等智能体框架平台能力。面向智能体的混合负载，优化软硬协同方案，兼顾GPU推理加速与CPU工作流编排；优化存储产品适配多层记忆、跨会话KV缓存池；网络产品重点优化尾时延、东西向流量、跨域协同能力。面向高密度智能体部署场景，积极推进原生液冷、超节点架构、高密整机柜方案，支撑大规模并发智能体运行。MaaS平台需升级为支持工作流编排、智能体Runtime、多模型路由、Skill市场、成本计量的任务级服务平台。同时将安全、可观测、可治理能力内建于产品底层，实现智能体身份管理、最小权限管控、工具调用沙箱隔离、执行链路全链路日志、异常行为检测和任务回滚，满足不同行业的安全与合规要求。

■ **面向行业场景做深度打磨，提供分级落地路径并构建开放产业生态**

不同行业智能体成熟度、风险约束差异巨大，服务商需要针对知识密集型行业、实体制造行业、互联网办公行业输出差异化解决方案。对于高风险行业，重点强化人机协同、审计留痕、权限管控；对于制造业、能源等实体行业，做好与生产系统的对接。提供从试点、单流程智能体到多智能体协同的分级实施路径，降低客户试错门槛；沉淀行业Skill、行业知识库模板，减少客户从零开发的工作量。同时，智能体完整落地需要多方协同，服务商应坚持开放，积极与芯片厂商、智算中心、行业软件厂商开展生态合作，推动云–边–端协同能力建设，支持智能体任务在云端、边缘、终端之间合理拆分，并积极参与行业标准建设，推动智能体 Skill、记忆格式、调用协议、审计日志格式的标准化。

■ **构建以Token效率为核心的运营计量体系，创新商业模式适配客户付费诉求**

“交互性–吞吐量”是Token服务的关键度量维度，二者权衡构成帕累托前沿曲线，服务商应帮助客户理解这一关系，在给定业务场景和服务等级约束下协助客户靠近帕累托前沿，实现交互性与吞吐量的最优平衡。在Token运营方面，提供覆盖统一模型入口、用户与智能体身份、调用配额、模型路由、成本归集、预算管理、服务质量监控、安全审计和内外部结算的精细化治理工具，将分散的Token消耗转化为可管理、可分析、可优化的运营数据，帮助客户将Token支出作为独立生产成本进行核算。在商业模式上，拓展Token按量付费、按任务结果计费、订阅式数字员工服务等模式，清晰披露总体拥有成本，适配智能体时代客户付费诉求。

对产业规划者的建议

■ **推动智算基础设施建设机制与评价体系同步改革，缩短建设周期、预留弹性扩容空间，引导从“建设导向”转向“落地导向”**

传统数据中心建设周期长，难以匹配人工智能算力快速迭代需求，面临“建成即落后”的扩容窘境。建议在项目审批、能源及土地指标配置等环节，探索面向智算基础设施的差异化机制，缩短从立项到投产的周期，并为建成后的弹性扩容预留制度空间，使基础设施供给节奏能够更好适配智能体负载从“阶段性配置”向“持续性生产”转变的要求。同时产业统计口径同步跟进，将任务成功率、智能体应用渗透率等反映实际产出的指标，逐步纳入官方评价体系。

■ **针对传统行业智能体落地建立定向支持机制，帮助其跨越障碍**

医疗健康、金融、教育、科研等行业呈现当前智能体成熟度相对有限、但长期生产力重构潜力突出的特征，同时受合规和数据敏感性制约明显。建议针对此类高意愿、高潜力但落地速度偏慢的行业，通过建立合规数据共享机制、开展政府主导的试点示范项目等方式，帮助其跨越结构性障碍，避免长期生产力重构价值更大的领域因外部制约而发展滞后。

■ **以标准先行、示范牵引推动超节点、原生液冷产业化落地，支撑智算中心向更高功率密度演进**

超节点等系统级设计使计算产业的竞争从单芯片峰值性能扩展到多芯片协同的系统计算能力，但初始投资成本过高是企业部署超节点的首要痛点。建议在评价和支持超节点等系统级算力方案时，避免单纯以芯片堆叠规模作为衡量标准，转而将单位Token成本、任务成功率等实际运行效率指标纳入考核体系，并探索通过专项资金、绿色金融等工具，引导超节点能力向中小企业、教育、医疗等对成本更敏感的场景下沉。到2026年底，高密度人工智能算力机柜单机柜功率将突破300千瓦级，原生液冷正从单一散热方案演进为整机柜乃至数据中心的系统设计理念。建议在新建和改扩建高密智算中心项目中，将原生液冷纳入前瞻技术方向和示范工程，推动芯片、服务器、液冷部件、数据中心设计与运维企业协同完善接口标准和可靠性验证规范，以标准先行、示范先试的方式，支撑单柜功率密度持续向兆瓦级演进。

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