头豹 LeadLeo

行业简报

2025年大模型云市场探析

如何重构企业智能化路径，开启大模型产业新浪潮？

企业标签：百度智能云、阿里云、华为云

大模型云行业创新发展

China Large Model Cloud Industry

中国大規模モデルクラウド 産業

头豹研究院

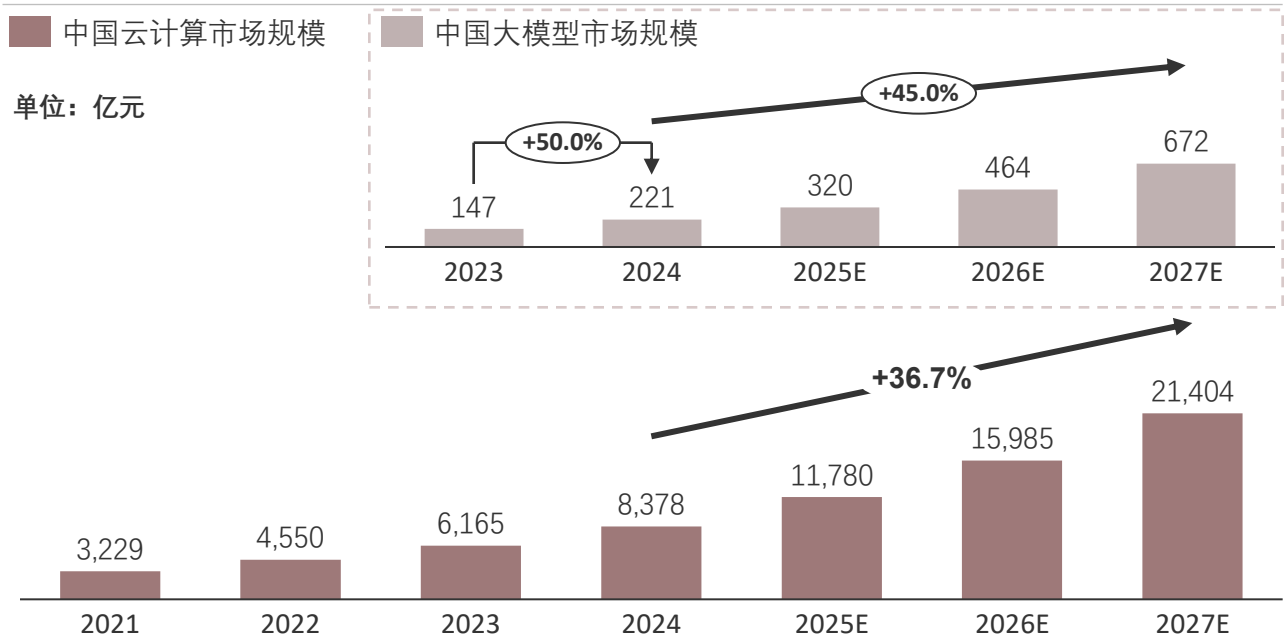
名词解释

- **大模型：**指参数数量巨大、在海量数据上进行训练的深度学习模型。这类模型通常具备强大的自然语言理解、生成、知识推理、代码生成等多种能力。
- **GPU：**图形处理器，最初用于图像和视频处理，现广泛用于深度学习训练和推理。GPU 具备大规模并行计算能力，适合执行矩阵乘法、卷积等复杂运算，代表厂商包括 NVIDIA、AMD。
- **TPU：**张量处理器，由 Google 专门为加速深度学习任务设计的 ASIC 芯片，针对张量操作进行优化，提供比 GPU 更高的能效比，主要用于大规模模型训练和推理。
- **大模型云：**指基于云计算基础设施，专门为大模型的训练、推理、部署、管理和应用开发提供支持的云服务平台或解决方案。
- **算力：**指用于支持大规模人工智能计算（特别是大模型训练和推理）所需的计算资源能力，通常以高性能计算集群（如GPU、TPU集群）的形式提供。
- **云端部署：**指将大模型及其应用部署在远程的云计算服务器上运行。用户通过网络访问服务，计算任务主要在云端完成。报告指出这是当前大模型项目的主流部署模式。
- **端侧部署：**指将经过优化或轻量化处理的大模型部署在终端设备（如智能手机、PC、汽车、摄像头等）或靠近数据源的边缘服务器上直接运行。
- **智算中心：**指专门为满足人工智能计算需求而设计和建设的大规模、高性能数据中心，提供强大的AI算力、数据处理和算法模型服务。
- **多模态：**指能够处理和理解多种不同类型数据（如文本、图像、音频、视频等）信息的人工智能技术或模型。

大模型云市场探析——大模型云市场发展现状

- 大模型云不仅是“算力承载平台”，更是企业迈向智能时代的“技术中枢”与“创新引擎”。其价值不仅在于提供AI能力，更在于构建从模型训练、数据治理、应用开发到业务落地的智能基础设施闭环

中国大模型与云计算市场协同发展，2023-2027年



- **中国大模型与云计算市场正呈现深度协同发展态势，大模型云已超越“算力承载平台”的定位，成为企业智能化转型的核心基础设施**

从市场规模看，中国云计算市场自2021年的3,229亿元起步，预计以36.7%的年复合增长率扩张至2027年的21404亿元，其中2023-2025年增速分别达50.0%与45.0%，显示市场已进入爆发式增长阶段。同期，大模型市场规模从2023年的147亿元增至2027年的672亿元，两者增长曲线高度同步，印证了“大模型驱动云需求、云支撑大模型落地”的双向赋能关系。

这一协同效应的深层逻辑在于：大模型对算力的极致需求（如GPT-4训练消耗超百万GPU小时）直接拉动云计算的异构算力供给，而云计算的弹性资源池、模型优化工具链（如TensorRT-LLM）及MaaS商业模式，又大幅降低大模型落地门槛，形成“训练-推理-应用”的商业闭环。例如，企业通过云端大模型可实现从智能客服到供应链优化的全场景升级，其ROI较传统IT架构提升3-5倍。

展望未来，市场将呈现三大趋势：一是“模型即服务”（MaaS）渗透率持续提升，预计2025年超60%的企业将通过云平台调用大模型能力；**二是行业垂直模型爆发**，云计算厂商将深化与医疗、制造等领域合作，构建定制化模型生态；**三是边缘计算与大模型融合**，云-边-端协同架构将支撑实时性要求更高的场景（如自动驾驶）。

挑战亦不容忽视：算力成本占大模型TCO的60%-70%，云服务商需通过芯片定制（如AWS Trainium）、存算一体等技术进一步降本；同时，数据隐私、模型可解释性等合规风险，要求云平台构建从联邦学习到模型审计的全链路安全体系。

来源：头豹研究院

大模型云市场探析——大模型云服务模式

- 随着大模型产业的纵深发展，企业对AI能力的获取方式正从“算法驱动”转向“模型即服务”范式转型。大模型云作为其主要承载平台，正逐步形成从底层算力服务到上层行业应用的全栈商业模式闭环

服务模式演进：从底层资源到模型服务的全栈闭环

服务模式	服务类型	核心内容	主要价值
IaaS	弹性AI算力服务AI存储与网络服务	提供GPU/AI芯片集群、容器化部署、分布式训练支持等底层资源服务	降低大模型运行门槛，按需获取弹性资源，优化成本结构
PaaS	一站式模型开发与管理平台	支持训练、微调、部署、数据管理、评估等全生命周期功能	降低AI开发门槛，提升模型开发效率与治理能力
MaaS	API调用、托管推理服务	提供大模型的通用能力，如文本生成、多轮问答、图像理解等	快速集成AI能力，无需模型开发与部署
SaaS	AI应用产品	将大模型能力嵌入政务、办公、金融、医疗等场景中，打包为AI应用工具	面向业务用户，推动AI应用落地和价值变现

- **大模型云并非传统云计算在AI领域的简单延展，而是围绕“大规模模型生命周期管理”所构建的高耦合、高垂直的一体化基础设施体系**

IaaS层不再是面向通用业务的虚拟计算平台，而是为大模型并行训练与超大规模推理任务量身构建的AI原生计算底座。该层需支持异构加速芯片的统一调度，优化分布式训练通信拓扑，动态分配千卡级GPU集群资源，并引入参数快照、断点续训、冷热数据分层管理等机制，以保障预训练与推理阶段的高吞吐与低成本运行。

IaaS的核心价值在于将算力与模型调度逻辑深度耦合，支撑高频率参数交互与PB级数据带宽的持续供给。PaaS层围绕大模型的工程化需求展开，构建从训练、微调、部署到评估、审计、更新的全生命周期管理平台。

与传统AI平台不同，PaaS需支持大模型跨规模多阶段训练、企业私有数据适配、RAG结构集成、Agent构建能力，并提供细粒度权限与隐私保护机制。模型治理能力尤为关键，包括训练数据对齐、毒性过滤、偏见识别、响应可控等内容安全流程，构成企业自建模型能力的关键屏障。

MaaS层将大模型能力以API或Agent形式封装为标准化调用接口，实现能力即服务。区别于传统AI API，MaaS服务需支持多模态统一封装、Prompt工程资产化管理、上下文保持、模型选择与路由机制，并具备服务过程的可观测性、稳定性与安全响应能力。该层强调无需开发、即开即用，是大模型从底层能力向通用业务系统渗透的桥梁。

SaaS层则是大模型与行业知识深度融合的最终形态，通过模型驱动的智能系统赋能政务、金融、医疗、制造等垂直场景。与传统SaaS产品不同，这类系统强调语言理解与知识调用能力的自主进化，结合企业知识库与交互式Agent形成可成长、可协同的智能体体系。其核心在于将模型能力转化为可执行、可对话、可决策的应用智能，实现AI能力的行业级商业落地。

来源：头豹研究院

大模型云市场探析——大模型高度依赖云计算

- 参数规模迈入万亿时代，模型结构趋于复杂，单次训练成本动辄数百万美元以上，仅A100租用在主流云厂商上月租达1,000美元/GPU以上，推理高并发需求推动云平台服务

技术本质决定大模型高度依赖云计算

参数规模

模型名称	发布机构	参数规模	发布时间
GPT-3	OpenAI	1750亿	2020年
GPT-4	OpenAI	未公开（估超1万亿）	2023年
Gemini 1.5	Google DeepMind	万亿级（混合专家架构）	2024年

➢ 进入2024年后，大多数主流模型均向“万亿参数+多模态+专家混合”方向发展，计算图复杂度指数级增长。

训练部署

模型名称	训练GPU数量	训练时长	估算成本
GPT-3	~10,000个NVIDIA A100	数周	~$460万 USD（2020年估）
GPT-4	推测使用>20,000个A100/H100	数月（并行）	超$1亿美元
Gemini 1.5	Google TPU v5e 芯片集群	未公开（预计数月）	Google称为“有史以来最大AI训练任务”之一

➢ GPT-4 在训练期间消耗的算力总量超百万GPU小时，其对电力、冷却、调度系统的依赖远超传统模型。

➢ Meta 在2024年初宣布将采购 35万张NVIDIA H100 GPU 用于训练Llama 3、Llama 4，这是全球最大训练部署之一。

■ **大模型参数规模爆炸式增长与技术本质共同决定了其对云计算的深度依赖，且这一趋势将随着模型复杂度的指数级上升而持续强化**

从数据层面看，GPT-3到GPT-4的参数规模跃升，以及2024年后主流模型普遍采用的“万亿参数+多模态+专家混合”架构，直接推高了计算图复杂度，使得算力需求呈指数级增长。例如，GPT-4训练消耗超百万GPU小时，这种非线性增长的需求是传统IDC模式无法承载的。从成本结构分析，大模型训练成本高昂，硬件成本占比超70%，而云计算通过弹性扩缩容、竞价实例等模式，可显著降低闲置率，实现成本优化。

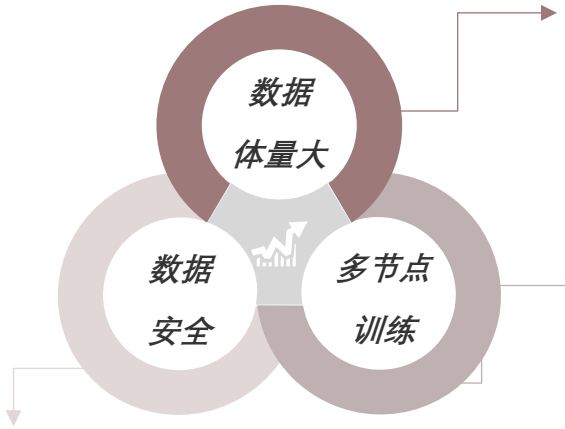
在推理环节，云计算通过预填充-解码分离、KV缓存分布式路由管理等技术架构创新，同时满足低延迟与高吞吐的需求，并通过MaaS模式将推理成本转化为按需付费，极大降低了企业AI应用门槛。技术本质上，大模型对计算资源、存储资源和网络资源的海量需求，以及云服务商提供的异构算力融合、全链路工具链和生态壁垒，共同构成了大模型与云计算不可分割的共生关系。

来源：头豹研究院

大模型云市场探析——大模型数据密集型特征

- 在数据密集特性驱动下，决定了大模型“天然上云”的技术路径。云平台不仅能够解决大模型在数据处理规模、速度、安全、治理等方面的核心挑战，更成为其从训练、推理到迭代优化的关键运行底座

数据密集型特征要求大模型依赖云的存储与调度能力



■ **训练数据体量庞大，超出本地基础设施承载上限**

以OpenAI的GPT-3为例，其训练数据量超过45TB文本数据，而如Google的Gemini、Meta的LLaMA-3、百度文心等新一代多模态模型，常训练于PB级别的图文、视频等异构数据。

■ **数据安全合规推动“数据不动，模型来动”的云上范式**

在金融、医疗、政务等数据高敏行业，数据出域受到严格监管。比如《中国个人信息保护法》《金融数据安全管理办法》均限制大规模数据外传。

云端提供就近部署大模型能力，使模型在“数据原地”训练成为可能。

腾讯政务云支持政务大模型“域内训练+本地推理”，阿里金融云也落地了“合规沙箱+模型开发”。

■ **多节点训练带来数据高频读写与高速交换需求**

大模型训练多采用分布式并行策略（如Data Parallel、Model Parallel）。

在千卡、万卡训练环境下，模型参数和训练数据需在多个节点间高频交互。

这种训练结构对I/O性能、数据调度效率提出极高要求。

阿里云2023年对大模型训练的实测数据显示：1PB训练数据需支持10GB/s+持续带宽才能保证每轮迭代稳定训练；

普通本地集群网络I/O在GB级别，容易成为瓶颈。

■ **在大模型时代，计算资源仍是基础性投入，但对领先企业而言，其挑战已逐步由“能否获得足够算力”转向“能否高效、安全地调度海量异构数据”。大模型的核心瓶颈正在从“算力争夺”转向“数据调度”，而这一能力高度依赖云平台提供的系统性能力支持**

一方面，训练所需数据规模从TB跃升至PB量级，本地基础设施难以提供所需的存储扩展性与高并发I/O能力；另一方面，云平台通过对象存储、分布式文件系统、RDMA与高速互联网络，为大模型训练建立起端到端的数据高速通路，成为支撑分布式并行训练的唯一可行方案。

同时，在数据合规性约束愈发严格的环境下，云上的“模型就地训练”成为满足合规的唯一路径。从政务到金融，从医疗到能源，越来越多的行业采用“数据不动，模型来动”的策略，强化了云平台在大模型时代的数据中枢角色。

因此，**大模型对云的依赖不只是性能选择，而是能力边界的决定因素。云平台已从“弹性算力池”演进为“可信数据基础设施”，是大模型可持续演进的必要前提。**

来源：头豹研究院

大模型云市场探析——从模型演进到业务重构

- 随着开源模型推动技术民主化、闭源模型主导服务一体化，企业“上云”正从技术部署行为演进为组织智能的系统性重塑，谁能率先建立模型与云的协同优势，谁将在未来智能经济中占据主动

开源与闭源双轮驱动下的企业“全面用云”趋势

模型类型	企业选用动因	推动上云的核心机制	企业获得的价值
开源模型	高性价比、可定制、安全可控	云端推理部署、调用API、接入RAG平台	降低成本、灵活开发业务AI能力
闭源模型	强大通用能力、零门槛使用	MaaS服务、Agent平台、企业模型托管	快速集成、加速业务重构、提升效率

■ **开源模型视角：重构业务架构的工具箱 + 云推理的性价比优势**

➢ **模型能力释放 + 可控性提升 = 企业主动拥抱开源模型**

近年来，以 DeepSeek-VL、Qwen、LLaMA2/3、Baichuan 等为代表的开源模型，在如下几个方面对企业极具吸引力：

- **性能逼近闭源模型：**在MMLU、CMMLU、CodeEval、AgentEval等基准测试中，一线开源模型已经能达到 GPT-3.5 或 GPT-4-Turbo 近似水平，特别是在中文、代码等细分任务中有超越表现；
- **更灵活的定制能力：**企业可以根据自身场景进行指令微调、低秩适配（LoRA）、RAG增强，实现业务专属模型；
- **模型权重开放，数据不出域：**便于合规性和数据安全控制，特别是金融、医疗、政务等行业；
- **避免闭源锁死风险：**不依赖某一家模型供应商，可自主掌控技术路线和成本。

➢ **云推理的方式成为开源模型落地的现实选择**

虽然开源模型权重开放，理论上可以私有部署，但在实践中，真正跑得动、调得优的企业仍少数，背后存在三大现实瓶颈：

- **高性能GPU成本高昂：**尤其是A100/H100资源在2023~2025年期间供需紧张，企业自建成本极高；
- **基础设施复杂度高：**如多卡推理、KV缓存、模型量化、请求调度等需专业团队维护；
- **企业业务负载波动大：**本地部署资源利用率低，存在资源闲置浪费问题。
- 因此，企业倾向于选择“开源模型 + 云推理 API”或“轻微本地化 + 云异构推理”组合模式，由阿里云、腾讯云、火山引擎、百度智能云等云厂商托管部署高性能推理服务，再由企业通过API调用或对接RAG框架，实现场景化应用。

➢ **这一趋势本质上意味着：即便采用开源模型，云服务仍是不可替代的承载平台。**

来源：头豹研究院

从模型演进到业务重构（续上页）

闭源模型视角：以MaaS为核心的全栈“即服务”体系推动企业全面上云

闭源模型的技术壁垒+产品封装决定其天然依赖云端

闭源模型如 GPT-4、Claude 3、Gemini 1.5、通义千问、百度文心、腾讯混元等，具备以下特点：

- **规模超大：**参数规模 100B 级别以上，需高密集算力，无法本地部署；
- **能力强泛化广：**具备复杂指令理解、代码生成、多模态交互能力，可作为企业大脑；
- **产品打包为服务：**封装为API、Agent、RAG引擎、SaaS插件等形式，无需了解底层模型结构即可直接调用；
- **接入门槛极低：**企业只需调用API，即可在业务系统（如客服、CRM、办公套件）中集成AI能力。

闭源模型厂商提供的 MaaS 服务，本质是“模型即服务 + 知识工程 + 工具链即插即用”的综合云服务。

MaaS 模式绑定企业核心系统，增强对云的结构性依赖

以阿里云通义千问为例，其提供的“千问 API + 百炼平台 + RAG 工具链”可用于：

- 智能问答客服系统对接
- 企业知识库AI化
- 流程自动化（表单填报、任务调度）
- 智能办公助手（如钉钉 AI、文档生成）

这些模块高度依赖 MaaS 能力，且通常绑定云平台提供的算力、数据存储、API调用管理、身份认证、权限控制等服务，导致企业即使不主动迁移核心系统，也被动把AI能力接入工作负载搬上了云。

此外，一些闭源模型还开放了“企业私域大模型”方案（如阿里“企业通义”、百度“文心千帆私有化部署”），但其部署架构仍基于云原生设计，依赖：

- 弹性K8s调度（如ACK、EKS）
- 云对象存储（OSS、COS）
- 服务网格与微服务治理

即使是私有化，也很难脱离云环境运行，因此仍然推动了企业基础架构的云化升级。

来源：头豹研究院

大模型云市场探析——大模型重塑企业智能化路径

- 大模型正驱动企业在客户体验与运营效率两端实现深度革新，打通业务与职能的智能协同路径，并通过从工具化接入到平台化治理的演进，重塑企业智能化架构与组织形态，开启全域智能时代的新篇章

企业智能化转型的阶段性演进

大模型典型应用场景：驱动业务与职能双轮转型

■ **在大模型技术广泛渗透的背景下，企业智能化已从以往“业务为主”的单轮驱动，演化为业务与职能双轮协同转型的新格局**

一方面，大模型通过智能客服、营销助手、内容生成等手段，助力企业提升用户体验、释放增长潜力；另一方面，在财务、法务、人力等职能部门，大模型则推动“数字员工”体系落地，实现流程重构与管理效能跃升。

从行业维度来看，大模型已在金融、制造、能源、政务等关键领域形成了具有代表性的典型应用场景：包括金融领域的智能投研与文档合规、制造业的知识问答与文档处理、能源行业的专家系统与调度生成、政务场景下的公文起草与政策问答。

企业正逐步从局部试点、工具化接入，迈向平台化治理、组织级嵌入的深水区。在此过程中，大模型不仅是一项新技术，更是一种重塑企业智能化路径的关键力量。

未来企业的智能演进，将呈现出从“自动化→智能协同→自主决策”的渐进路线，最终构建起以大模型为核心驱动的智能中枢系统。

来源：头豹研究院

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

行研训练营

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

报告作者

袁栩聪
首席分析师
oliver.yuan@leadleo.com

王利华
行业分析师
lihua.wang@leadleo.com

业务咨询

- **客服电话：400-072-5588**
- **官方网站：www.leadleo.com**

深圳办公室

广东省深圳市南山区粤海街道华润置地大厦E座4105室

邮编：518057

上海办公室

上海市静安区南京西1717号会德丰国际广场 2701室

邮编：200040

南京办公室

江苏省南京市栖霞区经济开发区兴智科技园B栋401

邮编：210046

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。

- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。

- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。

- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。

- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明