

2025年中国大模型一体机行业研究

DeepSeek大模型一体机

如何破局行业发展

企业标签：华为、新华三、讯飞星火

中国产业创新发展

China Large Model All-in-One Machine Industry

中国大規模モデル一体型マシン産業

报告提供的任何内容（包括但不限于数据、文字、图表、图像等）均系头豹研究院独有的高度机密性文件（在报告中另行标明出处者除外）。未经头豹研究院事先书面许可，任何人不得以任何方式擅自复制、再造、传播、出版、引用、改编、汇编本报告内容，若有违反上述约定的行为发生，头豹研究院保留采取法律措施、追究相关人员责任的权利。头豹研究院开展的所有商业活动均使用“头豹研究院”或“头豹”的商号、商标，头豹研究院无任何前述名称之外的其他分支机构，也未授权或聘用其他任何第三方代表头豹研究院开展商业活动。

研究目标

研究背景

随着人工智能技术的快速发展，大模型（如GPT、BERT等）在自然语言处理、图像识别、智能推荐等领域展现出强大的能力。然而，大模型的训练和部署对计算资源、数据安全和系统运维提出了极高的要求。为了解决这些问题，大模型一体机作为一种集成化的人工智能基础设施应运而生，旨在为大模型的开发、训练、部署及运维提供端到端支持。中国大模型一体机行业在近年来得到了快速发展，各大科技企业纷纷推出自己的大模型一体机产品，市场竞争日益激烈。

研究目标

- 分析市场现状
- 探讨技术趋势
- 典型厂商分析

本报告的关键问题

- 大模型一体机的市场定位和用户需求是什么？
- 大模型一体机的技术架构和关键组件有哪些？

观点摘要

01 大模型一体机市场呈现多样化配置与功能定位：

- ◆ 2025年主流厂商的大模型一体机产品呈现多样化配置与功能定位，价格覆盖50万至500万元不等。
- ◆ 高端型号支持超大模型预训练与推理，中端型号适配行业主流大模型应用，基础型号赋能轻量级模型推理与边缘部署。这种多样化的产品布局满足了不同用户群体的需求，推动了市场的快速发展。

02 大模型一体机的价格由四大因素共同决定：

- ◆ 大模型一体机的价格由硬件性能基础、系统集成能力、行业定制化服务以及合规安全交付四大因素共同决定。
- ◆ 核心硬件配置是价格构成的基础，系统集成与软硬协同能力影响溢价空间，定制化服务与行业应用适配形成差异化定价模型。这种全链条价值溢价机制体现了大模型一体机从底层算力到场景落地的全面能力。

研究框架

◆ 中国大模型一体机行业研究	-----	5
• 定义与组成结构	-----	6
• 部署特点	-----	7
• 主要厂商及具体产品价格	-----	8
• 价格区间	-----	9
• 价格影响因素	-----	10
• 定价策略对比	-----	11
• 用户画像	-----	12
◆ 中国大模型一体机典型厂商分析	-----	13
• 华为	-----	14
• 新华三	-----	15
• 讯飞星火	-----	16
◆ 业务合作	-----	17
◆ 方法论及法律声明	-----	18

名词解释

- ◆ **AI加速芯片：**专门设计用于加速人工智能计算任务的芯片，能够提升深度学习模型的训练和推理速度，相比传统CPU具有更高的并行计算能力和能效比。
- ◆ **异构计算架构：**指在同一计算平台上集成不同类型的处理单元（如CPU、GPU、FPGA、AI芯片等），通过协同工作实现性能和能效的优化。
- ◆ **模块化设计：**系统设计理念之一，将复杂系统划分成相互独立、可单独升级或替换的模块，便于扩展、维护和灵活配置资源。
- ◆ **弹性扩展能力：**系统根据业务需求动态调整资源（如算力、存储等），支持快速扩展或缩减，保证资源的高效利用和系统的可持续发展。
- ◆ **全国产软硬件栈：**指软硬件体系结构均由中国自主研发，涵盖芯片设计、操作系统、中间件、应用软件等，实现自主可控。
- ◆ **安全隔离：**通过硬件和软件技术，将不同的计算环境或数据区域相互隔离，防止未经授权的访问或数据泄露，保障系统安全。
- ◆ **数据加密技术：**对数据进行编码处理，使其在传输或存储过程中无法被非法访问者读取，保障数据隐私和安全。

Chapter 1

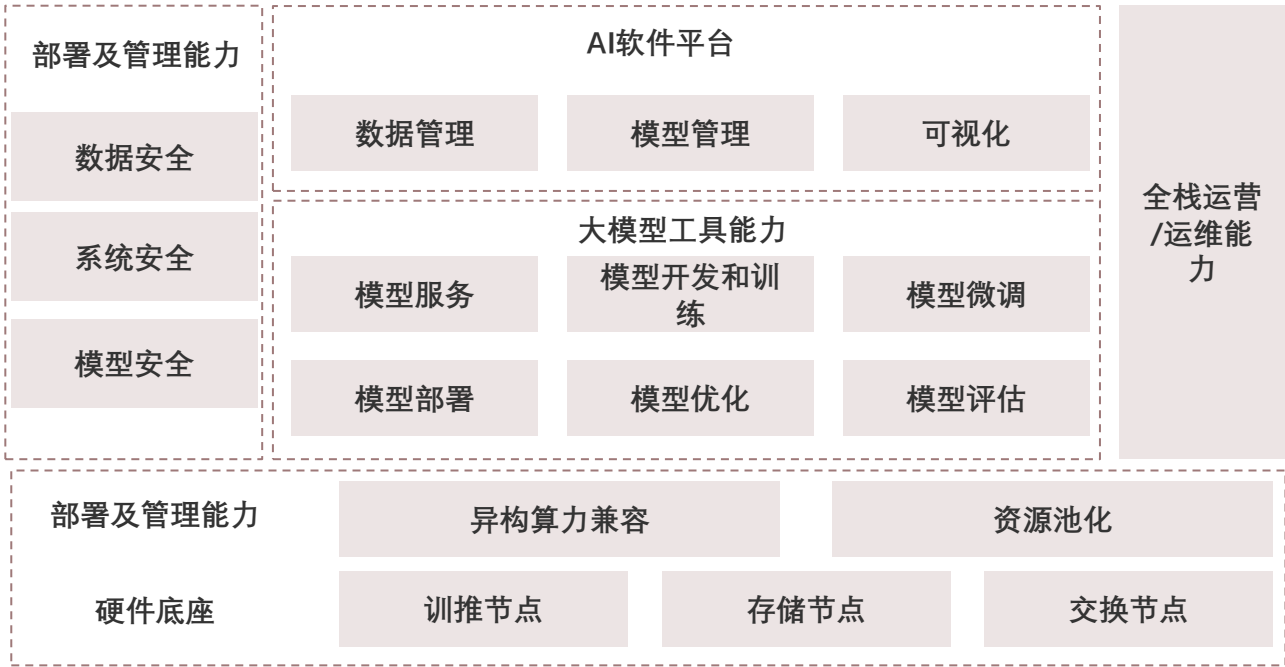
市场分析

- ☐ 定义与组成结构
- ☐ 部署特点
- ☐ 主要厂商及具体产品价格
- ☐ 价格区间
- ☐ 价格影响因素
- ☐ 定价策略对比
- ☐ 用户画像

中国大模型一体机行业研究——定义与组成结构

- 大模型一体机是一种集成化的人工智能基础设施，旨在为大模型的开发、训练、部署及运维提供端到端支持。通常以“软件+硬件+运维”三位一体的方式构建，覆盖从底层算力支持到上层AI服务的完整链条

大模型一体机的定义与组成结构



■ 大模型一体机架构可以被拆分为四个主要组成部分，部署及管理能力、AI软件平台、硬件底座，以及全栈运营/运维能力，具体表现为：

■ 多层安全体系，保障可信可控

在部署与管理层面，一体机构建了涵盖数据、系统、模型三大维度的安全机制：数据安全聚焦合规与隐私保护，系统安全防止底层被入侵或篡改，模型安全确保参数与推理结果不被泄露或操控，是保障大模型可信运行的关键。

■ 软件平台与硬件底座协同发力

AI软件平台是系统的智能核心，上层实现数据与模型管理、可视化操作，下层支撑训练、微调、部署与评估等完整开发流程。硬件底座则提供高性能异构算力支持，具备资源池化管理、训推分离、快速互联等能力，满足多样化AI计算需求。

■ 全栈运维体系，支撑稳定运行

运维模块贯穿部署、监控、告警、恢复与优化全过程，确保系统在多用户、高并发环境下长期稳定运行，具备企业级可运维性与可扩展性，是大模型一体机持续运营的核心保障。

综上所述，大模型一体机通过结构清晰、功能完备的模块化设计，实现了AI模型开发和部署的集约化与工程化，不仅提升了模型研发效率，也增强了系统的安全性、可维护性与资源利用率，成为支撑新一代人工智能发展的关键基础设施。

来源：专家访谈，企业公告，头豹研究院

中国大模型一体机行业研究——部署特点

- 大模型一体机部署以极致性能、本地数据闭环与高度可控性为核心优势，区别于私有云的灵活定制与公有云的弹性共享，更适用于对实时性和数据安全要求极高的专有场景

大模型一体机的部署特点

对比维度	大模型一体机部署	私有云/混合云部署	公有云部署
部署架构	模型与算力、存储高度集成，软硬件一体，部署在本地	模型部署在企业自建或托管的数据中心，通过虚拟化和容器编排实现资源管理	模型部署在云服务商平台，通过多租户架构实现资源共享
算力资源	专属本地物理算力，独占、无虚拟化损耗	物理或虚拟算力可控，资源调度自主管理	云平台提供的虚拟算力，资源由云厂商统一调度
资源调度	固定物理资源，无调度延迟	基于本地K8s等系统调度，具备自主权	多租户资源调度，受限平台优先级策略
性能表现（推理/训练）	推理延迟极低，吞吐率高，适合实时场景	性能稳定，网络和存储可控	性能易受网络、虚拟化和多租户影响
数据安全性	数据不出本地，完全本地闭环处理	数据在自主管理网络中流转，可实施本地隔离	数据需上传云端，合规和隐私控制依赖厂商能力
定制化能力	支持软硬件协同优化，灵活适配模型结构和推理方式	可定制平台、框架和训练流程	通常为标准化服务，定制能力有限
模型更新与微调	支持本地模型微调、参数更新，更新可控	支持完整模型训练与微调流程	微调能力受限，部分平台仅支持API级模型调用
扩展能力	横向扩展需新增设备，部署粒度为整机	可通过横向扩展集群节点实现动态扩容	支持弹性扩容，可在分钟级部署大规模任务
运维复杂度	一体交付，运维简化，部分具备远程监控功能	需构建完整的MLOps体系，具备较高技术门槛	云平台统一运维，用户主要关注模型层和服务配置
成本结构	固定资产投入高，长期使用成本稳定	建设成本高，长期运维成本与使用规模相关	按量计费，前期成本低，但长期费用随用量增长

来源：专家访谈，企业公告，头豹研究院

中国大模型一体机行业研究——主要厂商及具体产品价格

- 2025年主流厂商大模型一体机产品呈现多样化配置与功能定位，价格覆盖50万至500万元不等，华为、智谱AI等在性价比与高性能推理方面具备显著优势

2025年主流厂商的大模型一体机产品价格与配置对比

厂商	典型产品	核心配置	功能定位	价格区间	备注
华为	FusionCube A3000	昇腾910芯片，适配DeepSeek V3/R1模型，支持私有化部署	训推一体	200万-500万元	与80余家伙伴合作推出昇腾一体机，性价比优势显著
浪潮信息	元脑RI推理服务器	多卡GPU（A100/H100），支持单机推理及小集群组网	推理优化	50万-200万元	硬件占比80%，软件占比20%，强调初期投入性价比
紫光股份（新华三）	灵犀Cube	覆盖14B-671B参数模型，支持UniCube训推一体服务	全场景覆盖	100万-400万元	支持国产化芯片（昇腾/海光），2025年Q1政企业务收入增长37.49%
中科曙光	曙光DeepSeek人工智能一体机	SothisAI3.0管理平台，兼容10B-1,000B参数模型，支持国内外主流开源大模型	全场景AI支持	150万-350万元	硬件扩展性强，适配多种GPU算力卡
云从科技	从容大模型智用一体机	昇腾AI基础软硬件平台，面向审计、政务等垂直场景	行业定制化	80万-250万元	黄金级昇腾生态合作伙伴，强调数据安全与本地化部署
广电运通	智审大模型一体机	S500K2/S800K2服务器集群，支持7B-671B模型推理	行业垂直应用	入门级50万-100万元；旗舰级300万-500万元	审计场景专用，硬件与行业模型深度适配
恒为科技	昇腾DeepSeek一体机	样机阶段，适配国产AI加速芯片	推理优化	未公开（样机阶段）	尚未形成销售订单，市场竞争力待验证
智谱AI	GLM昇腾大模型一体机	集成GLM-4-Plus等高精度模型，支持代码计算、数据分析等复杂任务	高性能推理	300万-500万元	2025年4月价格下调90%，GLM-4-Plus降至5元/百万tokens，性价比大幅提升

来源：专家访谈，企业公告，头豹研究院

中国大模型一体机行业研究——价格区间

- AI一体机呈现高、中、基础三个层级，分别覆盖超大模型预训练、主流行业落地与边缘推理等多样化场景，体现出硬件市场对分层算力需求的精细匹配能力

大模型一体机价格区间



高端型号150万-500万元

面向超大规模模型（如671B参数）的预训练与推理，硬件配置包括多卡并行计算（如NVIDIA A100/H100集群）、大内存及分布式存储。
例如广电运通旗舰级解决方案GRG-WTIM001，支持S900K3/S800K3算力中心集群部署。



中端型号50万-150万元

支持100亿-300亿参数模型的训练与推理，适用于中型企业及行业垂直场景。
例如浪潮信息推出的训推一体机，单台价格在50万-200万元之间，可支持10亿-300亿参数模型的训练。



基础型号10万-50万元

适用于小型企业或边缘计算场景，支持10亿-70亿参数模型的推理任务。
例如航天联志Z4713-T4（搭载英伟达RTX4090 GPU）成交价17.38万元。

■ 高端型号（150万-500万元）：支持超大模型预训练与推理

此类大模型一体机面向如671B参数级别的大模型训练与推理任务，硬件配置通常包括多卡并行架构（如NVIDIA A100或H100集群）、大容量内存与高速分布式存储系统，具备极高的数据吞吐与计算能力。典型如广电运通旗舰级GRG-WTIM001，支持部署在S900K3/S800K3等高等级算力中心，适用于政府、央企及国家级实验平台等超大规模AI基础设施建设场景。这类产品强调系统级集成、高稳定性与模型并发调度能力，属于算力资源池建设的核心设备。

■ 中端型号（50万-150万元）：适配行业主流大模型应用

中端设备可支持100亿-300亿参数模型的训练与高效推理，适合中大型企业部署行业大模型、专有知识库微调等任务。以浪潮信息的训推一体机为例，其售价区间为50万-200万元之间，覆盖单台多卡（如A800、H20）架构，具备一定扩展性和能源效率，广泛应用于金融、制造、医疗等行业落地的AI业务系统。该价位段成为“软硬协同+行业落地”的主力产品，强调算力性价比与行业适配性。

■ 基础型号（10万-50万元）：赋能轻量级模型推理与边缘部署

该类别主要服务中小企业、科研单位或边缘计算需求，支持10亿-70亿参数级模型的本地推理，适合轻量级生成式AI应用，如客服问答、营销内容生成、知识检索等。典型如航天联志Z4713-T4，搭载NVIDIA RTX4090 GPU，售价为17.38万元，兼具推理性能与成本优势。此类产品强调低部署门槛、快速上线及本地私有化能力，满足资源受限环境下的模型运行需求。

来源：专家访谈，企业公告，头豹研究院

中国大模型一体机行业研究——价格影响因素

- 大模型一体机价格由“硬件性能基础 + 系统集成能力 + 行业定制化服务 + 合规安全交付”四大因素共同决定，体现了从底层算力到场景落地的全链条价值溢价机制

大模型一体机价格影响因素

01

核心硬件配置是价格构成的基础

- GPU类型与数量是决定成本的首要因素。当前主流使用如NVIDIA A100、H100、A800、H20等高端AI训练卡，不同算力芯片价格差异巨大，
- 例如：单块H100价格可达20万元以上；国产替代如昇腾910B、寒武纪思元系列在同等性能下价格较低，但生态兼容性可能影响总成本。
- 高速互联结构（如NVLink、InfiniBand、PCIe Gen4/5）会显著增加成本，尤其在多卡并行场景中是性能瓶颈决定因素。
- 内存与存储架构：大模型训练要求高带宽、高并发读写，需配置高端HBM、DDR5内存、NVMe SSD、RAID阵列或分布式文件系统（如Lustre、Ceph），此部分可占总价15%-25%。

02

系统集成与软硬协同能力影响溢价空间

- 整机设计与系统优化：涵盖电源管理、散热设计、机柜封装、GPU调度与功耗控制策略。高密度异构计算集成（如8-GPU单机）需大量工程调优与验证，提升整机售价。兼
- 容的软件生态与框架优化：是否支持主流AI框架（如PyTorch、TensorFlow、MindSpore）和高效的模型并行调度工具（如Megatron-LM、Deepspeed、Colossal-AI）影响使用效率和客户选择。
- 国产化适配需求：政企客户对“全国产”软硬件适配要求增加，可能额外投入适配工作，提升交付价格。

03

定制化服务与行业应用适配形成差异化定价

- 模型微调与任务适配：如嵌入金融、医疗、政务等特定知识图谱、语料库，需模型工程师深度介入，构成定价的重要增量项。
- 行业场景落地能力：具备垂直行业Know-how的一体机（如审计智能机、司法问答系统）能在硬件规格相近的前提下形成高溢价。
- 全生命周期服务支持：涵盖远程运维、模型升级、版本管理、安全加固等，尤其在政企采购中常被作为长期技术服务合同打包进采购预算。

来源：专家访谈，企业公告，头豹研究院

中国大模型一体机行业研究——定价策略对比

- 不同类型厂商通过差异化的定价策略构建竞争优势，硬件型靠规模降本，软件型靠模型与服务溢价，场景型则依赖行业Know-how实现高附加值

大模型一体机定价策略对比

硬件主导型厂商

- **成本结构与定价逻辑：**
 - 硬件成本占比高达80%-90%，主要集中在GPU、存储、散热、网络互联等关键部件。
 - 规模化生产带来边际成本下降，依托自建工厂或高度垂直整合的供应链体系（如华为自研芯片+整机整合能力），具备“以价换量”能力。
 - 定价策略趋于“硬件贴近成本、服务带动附加值”，以提升市占率和生态粘性为首要目标。
- **典型案例：**
 - 华为昇腾一体机通过昇腾+MindSpore+昇思生态联盟合作模式，实现R&D成本共享。

软件服务溢价型厂商

- **成本结构与定价逻辑：** 相较硬件主导型，软件与服务占比提升至30%-40%
 - 行业大模型微调服务；
 - 长周期的交付与技术支持；
 - 订阅式运维与定制接口开发。
 - 定价策略强调“性能+服务”双重溢价，以模型优化能力和持续迭代为卖点。
- **典型案例：**
 - 智谱GLM昇腾一体机虽硬件平台基于国产昇腾，性能并非行业顶尖，但GLM模型在中文任务、行业知识问答等任务中表现优越，形成“模型即品牌”的溢价空间。

垂直场景定制化厂商

- **成本结构与定价逻辑：**
 - 强调软硬一体+行业Know-how集成能力，溢价来源于深耕行业的数据积累、模型调优、标准适配等。
 - 在硬件规格上趋于中等配置，但能通过算法+数据优化实现任务表现领先。
 - 定价策略体现为“场景定制+算法封装+接口适配”的解决方案价值，而非单纯按硬件算力标价。
- **典型案例：**
 - 广电运通智审一体机审计领域模型以多年政企项目中积累的大量结构化财务数据、法规语料为基础，具备行业壁垒；审计智能问答、多表比对、风险图谱等能力非通用大模型所能替代，因此具备较强定制溢价能力。

来源：专家访谈，企业公告，头豹研究院

中国大模型一体机行业研究——用户画像

- 大模型一体机适用于无专业AI基础设施和团队、以模型推理或轻量训练为主、预算有限、对上线效率要求高的企业或机构，尤其是中小企业、本地化需求行业或初期AI探索阶段的用户

大模型一体机vs私有化——用户画像

维度	用户特征说明	一体机推荐原因
部署需求	需本地部署，不考虑或不适合使用公有云	网络隔离、数据隐私要求高
机房环境	无标准IDC机房，仅具备普通办公环境、弱电间	无需改造机房，适配部署条件
运维能力	缺乏专业AI/GPU运维团队	开箱即用、免维护、运维友好
使用场景	以模型使用（推理）为主，或轻量化微调（如LoRA）	可直接加载预训练模型或轻量微调
训练需求	非大规模训练需求，或仅进行小规模个性化训练	性能与易用性兼顾
预算水平	IT预算 < 100 万元	成本可控、部署快，性价比高
用户规模	单业务线、单团队使用	适合一体机轻量部署，无需复杂权限管理
管理复杂度	无需多租户管理、权限划分简单	无需搭建完整平台，降低软件平台运维复杂性
上线周期要求	希望快速部署上线	集成度高、上线周期短
行业适配场景（典型）	教育、医疗、政务、制造业中小企业、科研机构、小型AI应用场景	数据需本地处理，缺乏运维能力，部署门槛低

- 大模型一体机的目标用户群体，主要集中在具备本地部署需求但缺乏完整AI基础设施建设与运维能力的中小型组织。这类用户通常不具备标准IDC级别的机房环境，仅有普通办公室或弱电间，难以支撑传统AI集群的建设与散热/供电要求，因此对一体机“空间适配性”与“即插即用性”有较高依赖。
- 此外，该类用户多缺少专业的GPU运维团队，无法承担裸金属部署或分布式训练环境的长期维护责任，因此偏好“低运维”甚至“零运维”的交付模式。
- 因此，大模型一体机的典型用户画像为：具备本地部署刚需，但缺乏专业AI/GPU基础设施与运维团队，聚焦于模型使用与轻量调优，预算有限、部署环境受限、希望快速上线并简化运维流程的中小型组织。代表性行业包括政务、教育、医疗、制造等对数据安全要求较高、IT能力有限的领域，以及处于AI能力导入初期阶段、正在验证大模型业务价值的企业用户。

来源：专家访谈，企业公告，头豹研究院

Chapter 2

典型厂商分析

- ❑ 华为
- ❑ 新华三
- ❑ 讯飞星火

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



商务咨询与深度合作

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

- ◆ 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。
- ◆ 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。
- ◆ 本报告所涉及的观点或信息仅供参考，不构成任何投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。本报告所指的公司或投资标的的价值、价格及投资收入可升可跌。
- ◆ 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本文所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本文所载资料、意见及推测不一致的报告和文章。头豹不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。



FROST & SULLIVAN

沙利文

诚邀

2025沙利文新投资大会

第十九届沙利文全球增长、科创与领导力峰会
暨第四届新投资大会

THE 19TH FROST & SULLIVAN GROWTH, INNOVATION AND
LEADERSHIP SUMMIT AND THE 4TH NEW INVESTMENT EVENT

2025年8月27日-28日 中国·上海
August 27th-28th, 2025, Shanghai · China

2025年9月2日 中国·成都
September 2nd, 2025, Chengdu · China

开幕倒计时

期待与您再度携手
共赴增长之旅、共创美好明天

大会咨询热线：021-3209-6800 转 8672

大会咨询邮箱：gil@frostchina.com