

行业简报

算力调度平台规模化发展

Deepseek带动算力需求井喷，
算力调度平台成最优解

企业标签：百度、华为、硅基流动

AI变革行业创新发展

China Large Model Industry

中国ビッグモデル産業

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。

内容目录

◆ 算力调度平台的规模化路径	-----	4
• 算力需求增长迅速	-----	5
• 算力调度平台的价值	-----	6
• 政府补贴是核心来源	-----	7
• 优质算力调度平台技术特征	-----	8
• 算力调度中心规模化的核心	-----	9
◆ 头豹业务合作介绍	-----	10
◆ 方法论与法律声明	-----	11

名词解释

- ◆ **算力调度平台**：用于统一管理、智能调度和优化来自不同提供商或地域的异构计算资源，如CPU、GPU等，旨在提升资源利用率、降低成本并简化用户操作。
- ◆ **vGPU**：一种将物理GPU的图形处理能力进行虚拟化的技术，允许多个虚拟机或用户共享同一块物理GPU卡，为虚拟桌面和应用提供硬件加速。
- ◆ **Kubernetes**：一个开源的容器编排系统，用于自动化部署、扩展和管理容器化应用，在算力平台中常用于管理承载计算任务的节点和容器资源。
- ◆ **异构算力**：指由多种不同类型处理器，如CPU、GPU、NPU、FPGA等组成的计算资源环境，能够为不同特性的计算任务提供最优的硬件支持。
- ◆ **任务调度**：算力调度平台的核心功能之一，指根据任务需求、资源状态、预设策略如优先级、成本等，将计算任务分配至最合适计算节点的过程。
- ◆ **资源池化**：将分散的物理或虚拟计算资源，如CPU、GPU、内存、存储，聚合成逻辑上的共享资源池，进行统一管理、分配与监控，提高利用率与效率。
- ◆ **弹性计算**：指计算资源如算力、存储等可根据业务负载或用户需求自动或手动扩展或缩减，常与按需付费模式结合，提升资源灵活性与成本效益。
- ◆ **智能调度**：区别于固定规则调度，指使用复杂算法如AI、机器学习分析任务特性与资源状态，实现最优化的资源分配与任务调度决策。
- ◆ **资源利用率**：衡量计算资源在一定时间内被有效使用的程度，通常以百分比表示，是评估算力调度平台效率的重要指标。
- ◆ **细粒度切片**：指将计算资源尤其是GPU精细拆分为可独立调度的小单元，更精准地匹配不同任务需求，提高资源共享与并发处理能力。

Chapter 1

算力调度平台的规模化路径

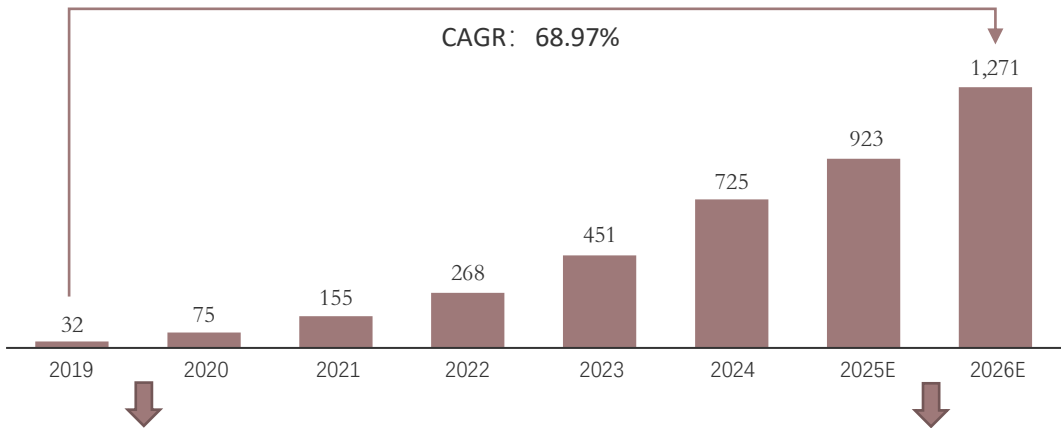
- 在AI大模型驱动下，中国智能算力需求与供给同步高速攀升，超大规模、高性能算力中心加速布局，带动整体算力基础设施快速扩容；在此背景下，具备高效资源整合与调度能力的平台，将具备突出的规模化发展价值
- 算力调度中心的盈利高度依赖政府补贴，该补贴旨在通过结构化机制确保本地资源充分利用并控制风险。成功的关键在于企业需战略性地早期介入以影响决策，同时其方案必须精准契合政府经济、社会及政绩等多维核心价值
- 平台必须具备高效整合和调度各地异构算力的技术能力，这为实现低成本、规模化和市场化变现奠定了基础。平台应具备的核心技术能力包括资源虚拟化、切片、实时监控和潮汐调度等，确保同一资源能够低成本、多次高效利用
- 算力平台规模化价值变现的核心在于庞大且多元的客户基础，这决定了盈利速度与定价潜力。同时，结合高效的生态合作与增值服务策略是进一步提升毛利、控制风险并最大化客户价值的关键手段

算力调度平台的规模化路径——算力需求增长迅速

- 在AI大模型驱动下，中国智能算力需求与供给同步高速攀升，超大规模、高性能算力中心加速布局，带动整体算力基础设施快速扩容；在此背景下，具备高效资源整合与调度能力的平台，将具备突出的规模化发展价值

中国智能算力发展及预测（2019-2026E）

单位：百亿亿次浮点运算/秒 (EFLOPS)



城市 | 智能算力规模 (EFLOPS)

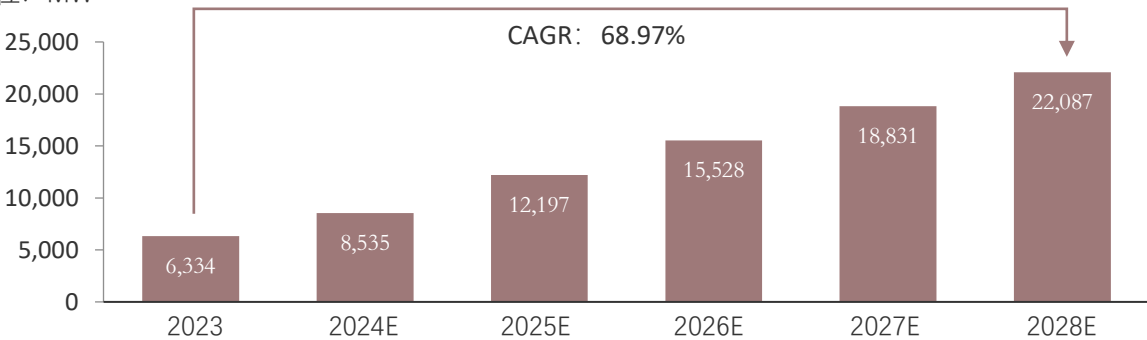
北京	45	广东	38	上海	30	深圳	25	贵州	80	河北	35
----	----	----	----	----	----	----	----	----	----	----	----

- 智能算力需求的激增，特别是在大模型应用领域，显示了AI技术快速发展和未来算力市场的巨大潜力

伴随着人工智能大模型及其相关应用的爆发式发展，智能算力的需求正在经历前所未有的激增。当前，智能计算已经成为中国算力结构中增速最快的类型，尤其在大规模语言模型、图像识别、自动驾驶等前沿领域，对高性能计算资源的需求尤为迫切。据统计，大模型作为智能算力的最大需求方，其需求占比接近60%，这不仅反映了AI技术的快速发展，也揭示了未来算力市场的巨大潜力。

中国智能算力需求及预测（2023-2028E）

单位：MW



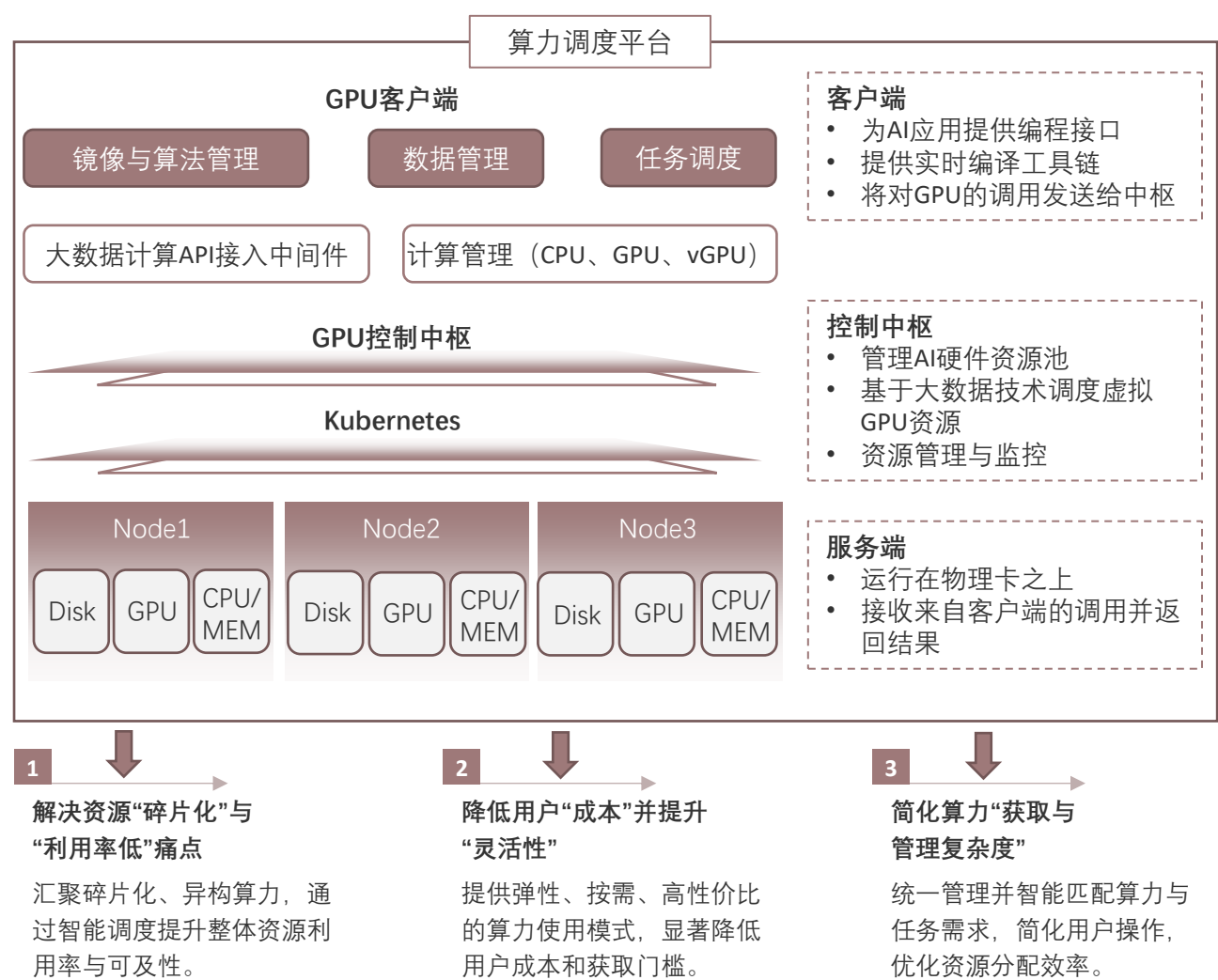
- 为了抢占AI市场份额，互联网巨头和超大规模云服务提供商持续扩展其超大规模、高性能的算力中心。预计至2028年，中国智能算力整体需求将超20,000MW。

来源：头豹研究院

算力调度平台的规模化路径——算力调度平台的价值

- 算力调度平台通过统一的技术架构智能整合异构且碎片化的资源，其核心价值在于显著提升资源利用率，降低用户获取算力的成本与门槛，并简化管理与匹配流程，实现高效、经济、便捷的算力服务供给

算力调度平台的价值



- 算力调度平台通过统一架构智能整合异构资源，核心价值在于提升利用率、降低使用门槛，并简化算力获取流程，实现高效便捷的服务供给

算力调度平台的核心价值，在于其扮演了一个智能化的资源整合与调度中枢的角色。它通过运用虚拟化（如vGPU）、容器化（如图中Kubernetes所示）、任务调度、数据及镜像管理等一系列技术手段，将底层分散、异构的计算资源（CPU、GPU等）进行池化管理和细粒度切分。这种模式旨在根本性地解决传统算力使用中存在的资源碎片化导致利用率低下、用户（尤其是GPU需求者）获取成本高且缺乏弹性、以及跨多种资源类型进行统一管理和高效匹配任务极其复杂这三大核心痛点。最终，该平台致力于为用户提供按需、弹性、高性价比且易于管理的算力服务，同时最大化底层硬件资源的整体利用效益。

来源：头豹研究院

算力调度平台的规模化路径——政府补贴是核心来源

• 算力调度中心的盈利高度依赖政府补贴，该补贴旨在通过结构化机制确保本地资源充分利用并控制风险。成功的关键在于企业需战略性地早期介入以影响决策，同时其方案必须精准契合政府经济、社会及政绩等多维核心价值

政府补贴是算力调度中心盈利的核心来源

- 1

提前铺垫干预政府决策

赢得政府订单重在投标前的早期介入与战略布局。通过对接关键决策者，企业能洞察真实需求并影响标准制定，从而塑造有利竞争格局，提高中标概率。
- 2

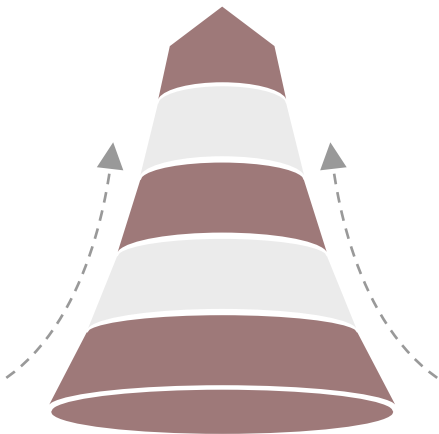
切实满足政府的核心价值点

政府决策高度关注经济、社会及政绩等多维核心价值，超越技术与价格考量。企业方案需精准回应这些诉求，清晰展示对本地发展目标的具体贡献，以此赢得政府信任。



赢下政府订单的关键

政府客户的核心需求



1.

政府的首要目标是确保本地已建算力资源得到充分消纳

地方政府为保障本地算力投资回报，常通过引入头部云厂商并配套财税、能源优惠。核心在于实施“算力券”等用户侧退坡式补贴，显著降低初期使用成本，提升本地算力相对公有云的定价优势，加速资源消纳进程。
2.

补贴机制设计规避项目风险

政府倾向于“先承诺补贴、后启动建设”模式以锁定政策支持，降低平台投资风险。补贴兑现机制明确，或通过用户直接抵扣，或基于平台达成算力消纳KPI后的政府结算支付，确保财政激励与使用效果强关联。
3.

政府多元目标塑造合作模式

政府推动算力建设的目标超越单纯营收，涵盖本地就业、产业升级及数字基建政绩等多重维度。因此，要求平台本地化运营成常态，政府亦会其资源网络导入本地客户，形成互利合作加速目标达成。
4.

战略性政府关系是合作前提

企业欲参与政府算力项目，需进行前瞻性的战略布局。构建专业的政府关系团队，主动对接决策层以洞察政策深层意图至关重要，旨在通过早期介入影响规划，实现与政府目标的战略协同。

■ 政府补贴是驱动算力调度中心商业模式及盈利的关键。政府的核心诉求在于确保本地已建算力资源得到充分利用（消纳），并通过结构化的补贴机制（如算力券、退坡机制、KPI挂钩）来激励用户、降低平台风险。要成功赢得政府订单，企业不仅需要提供技术上可行的方案，更关键的是要进行战略性的早期介入以影响决策，并精准把握、满足政府在经济、社会效益及政绩等多维度的核心价值诉求。

来源：头豹研究院

算力调度平台的规模化路径——优质算力调度平台技术特征

- 平台必须具备高效整合和调度各地异构算力的技术能力，这为实现低成本、规模化和市场化变现奠定了基础。平台应具备的核心技术能力包括资源虚拟化、切片、实时监控和潮汐调度等，确保同一资源能够低成本、多次高效利用

优质算力平台需具备精细化资源切片/异构兼容/动态调度能力



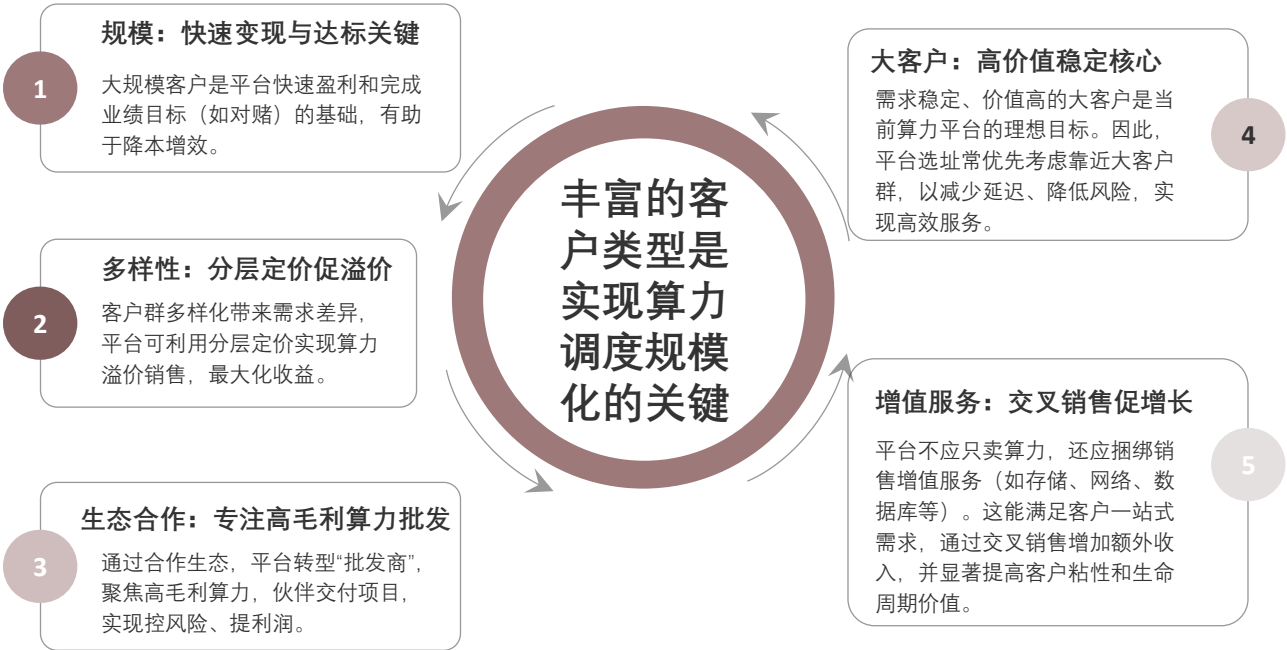
- 平台需要具备精细化资源切片、异构算力兼容、跨地区调度、实时监控与动态调度等能力，以实现算力资源的高效复用和低成本变现
- a) **细粒度切片：**平台需要将整块算力资源精准拆分为细小单元，以满足不同客户的多样化需求，实现闲置资源的充分复用。就像一套房子整租租5,000，将其切成4个空间，每个空间可以租2,000块一样，实现增值。
- b) **异构算力兼容能力：**面对英伟达GPU、华为昇腾以及国产天数智芯等不同芯片架构，平台必须具备兼容性，通过资源虚拟化技术将这些异构硬件统一调度，确保用户体验不受硬件差异影响。这种能力能够使平台构建更大规模的算力池，为不同应用场景提供定制化服务。
- c) **跨地区调度能力：**平台应能整合各地区的算力资源，实现资源跨区域调度，降低单一地区资源不足或故障的风险。例如，当某一地区（如上海）的算力需求激增时，平台可以快速调用邻近城市（如苏州、内蒙古等）的闲置资源，确保整体供给稳定、成本均衡。
- d) **实时监控与动态调度能力：**平台须建立完善的监控系统，实时跟踪各客户的算力使用状态，精准捕捉闲置或低利用率的资源。通过动态调整和打包复用这些闲置算力，平台可以将同一客户的冗余资二次打包出售，提升资源利用率和收入。
- e) **潮汐调度能力：**平台应根据客户需求的波动和使用高峰低谷，合理安排任务分配，将部分离线或对实时性要求不高的任务放在夜间等低峰时段执行，进一步实现资源高效变现。

来源：头豹研究院

算力调度平台的规模化路径——算力调度中心规模化的核心

算力平台规模化价值变现的核心在于庞大且多元的客户基础，这决定了盈利速度与定价潜力。同时，结合高效的生态合作与增值服务策略是进一步提升毛利、控制风险并最大化客户价值的关键手段

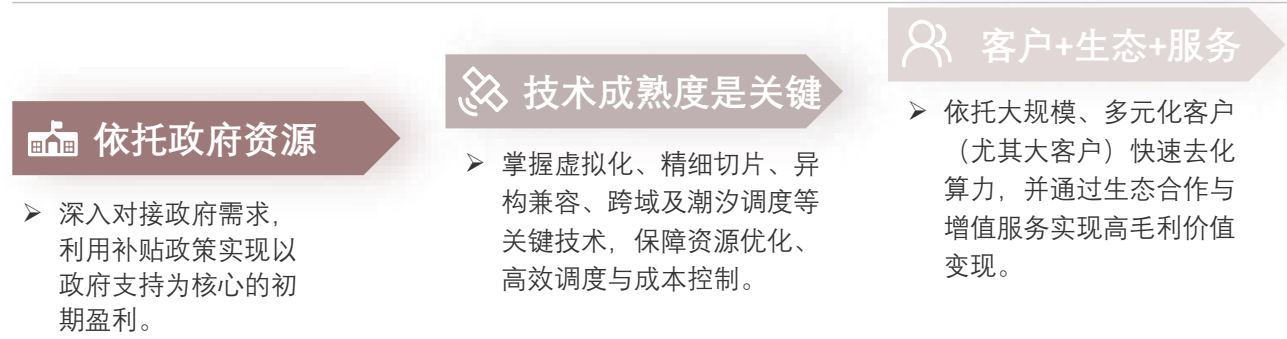
客户类型的规模与丰富性是算力调度规模化关键



■ 算力平台需依托庞大且多元的客户基础，并结合高效的生态合作与增值服务策略，方能实现算力的规模化价值变现

为实现算力价值的规模化，平台必须拥有足够大的客户规模作为快速盈利和达成目标的基础。其中，高价值、需求稳定的大客户是核心目标，促使平台就近部署以优化服务；而客户群的丰富性则允许通过分层定价实现算力溢价，最大化整体收益。运营层面，通过合作生态转型为“批发商”，聚焦高毛利算力销售并由伙伴完成低毛利交付，可以有效控制风险、提升利润。此外，在基础算力之上提供存储、网络等增值服务进行交叉销售，能进一步增加收入，并显著提升客户粘性 with 终身价值。

算力调度中心规模化的三大支柱



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



常乔雨
行业分析师
charles.chang@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

- ◆ 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。