下载日志已记录，仅供内部参考

广发研究 GF RESEARCH

计算机行业

从算力扩容到应用变现，继续把握 AI 全链机会

行业评级

买入

前次评级 买入

AI 基础设施板块：1.“国产算力+国产模型”的双向奔赴加速 AI 算力国产替代进程。2. CPU、超节点、AI 基础软件等受益于 Agent 应用带来的算力架构升级趋势。

模型端：模型层，建议重点关注具备前沿模型能力、高推理效率和商业化承接能力的厂商。国内模型能力持续提升，使国产模型具备进入企业工作流和开发者生态的基础。对于模型厂商，后续应重点观察真实调用、客户转化、收入兑现和盈利质量，核心指标包括 Token 调用量、企业客户数量、C 端付费转化、API/开放平台收入、ARR 增长以及单位 Token 成本等，而不应仅以单次模型评测排名作为判断依据。

AI 应用：无需非公开复杂数据、技术门槛较低、容错空间较大、功能性标准化程度高的软件公司将继续受到侵蚀和冲击，但具备强私域数据、基础通用工具类专业厂商如 Coding、database 等、以及卡位应用落地“最后一公里”的软件咨询实施商或脱颖而出。

EDA：行业兼具技术升级驱动与国产替代弹性，成长性强于周期性。当前国产厂商仍以点工具突破为主，与海外龙头在平台化、生态和先进工艺适配方面存在差距，但并购整合正在加速推进，行业有望从分散点工具向平台化、生态化演进。

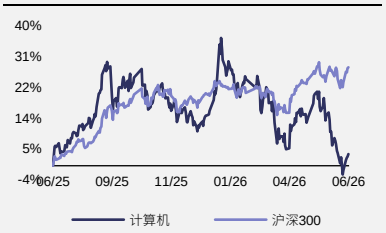
算力租赁与算力调度：在行业整体资本开支大幅抬升、算力需求持续高景气、算力基建加码投放的三重核心背景支撑下，算力租赁、算力调度赛道相关企业有望迎来明确成长机遇。

智能驾驶与机器人：智驾零部件环节主要投资机会集中在竞争格局较好的智驾芯片行业、规模效应明显的龙头 Tier1 以及充分受益于细分领域成长的个股。此外，智驾产业参与者立足现有技术积累，积极布局机器人新领域，有望打开新空间。

投资建议：1. AI 基础设施端：推荐寒武纪、浪潮信息、紫光股份、星环科技，同时建议关注中科曙光、范式智能。2. 模型端：建议关注智谱、月之暗面、DeepSeek、MiniMax，以及字节等厂商。3. AI 应用：推荐卓易信息、赛意信息、汉得信息、中控技术、晶泰控股、税友股份，建议关注中望软件、金蝶国际等。4. EDA 等工业软件：广立微、概伦电子、中控技术、中望软件、索辰科技等。5. 算力租赁与算力调度：算力租赁方面，我们建议关注杰创智能、智微智能、赛意信息、宏景科技。算力调度方面，我们建议关注卡莱特、并行科技、思特奇、云赛智联等。6. 智能驾驶与机器人：极智嘉、黑芝麻智能、锐明技术；同时建议关注：德赛西威、速腾聚创、小马智行、文远知行等。

风险提示。AI 应用落地不及预期风险；国产算力与模型生态发展不及预期风险；行业竞争加剧与盈利下滑风险；宏观环境与市场流动性波动风险。

相对市场表现



刘雪峰

SAC 执证号：S0260514030002
SFC CE No. BNX004
gfliuxuefeng@gf.com.cn

吴祖鹏

SAC 执证号：S0260521040003
wuzupeng@gf.com.cn

周源

SAC 执证号：S0260523040001
shzhouyuan@gf.com.cn

王钰翔

SAC 执证号：S0260526060001
SFC CE No. BXP717
wangyuxiang@gf.com.cn

请注意，吴祖鹏,周源并非香港证券及期货事务监察委员会的注册持牌人，不可在香港从事受监管活动。

相关研究

计算机行业 AI 模型系列（五）:Anthropic 暂停 Fable 5/Mythos 5 访问，国产大模型和算力迎发展机遇	2026-06-14
计算机行业 AI 模型系列（四）:大模型商业化加速落地，来自模型能力、Token 与 ARR 的多重验证	2026-06-10
计算机行业:模型和应用发展机会涌现，并继续推动算力需求总量和结构变化	2026-05-24

下载日志已记录，仅供内部参考

重点公司估值和财务分析表

股票简称	股票代码	货币	最新收盘价	最近报告日期	评级	合理价值（元/股）	EPS(元)		PE(x)		EV/EBITDA(x)		ROE(%)	
							2026E	2027E	2026E	2027E	2026E	2027E	2026E	2027E
寒武纪	688256.SH	CNY	1507.46	2026/05/31	买入	1648.46	8.15	23.55	184.96	64.01	176.77	63.12	30.10	47.20
浪潮信息	000977.SZ	CNY	65.66	2026/04/30	增持	80.91	2.70	3.40	24.32	19.31	20.95	17.48	15.60	16.60
紫光股份	000938.SZ	CNY	27.45	2026/04/29	买入	41.19	0.92	1.15	29.84	23.87	13.97	11.76	15.10	15.90
星环科技-U	688031.SH	CNY	127.35	2026/06/12	增持	170.94	-0.85	0.11	-	-	-	-	-13.20	1.60
智谱	02513.HK	HKD	2094.00	2026/04/02	增持	965.43	-8.49	-8.24	-	-	-	-		
卓易信息	688258.SH	CNY	128.63	2026/04/28	增持	156.60	1.57	2.48	81.93	51.87	62.58	42.46	18.40	24.20
赛意信息	300687.SZ	CNY	27.37	2026/04/29	买入	27.13	0.30	0.36	91.23	76.03	51.75	39.92	4.40	5.10
汉得信息	300170.SZ	CNY	17.88	2026/04/14	买入	28.04	0.30	0.37	59.60	48.32	39.25	32.52	5.10	6.10
中望软件	688083.SH	CNY	65.43	2026/04/23	买入	72.21	0.36	0.84	181.75	77.89	173.44	64.16	2.20	4.90
晶泰控股	02228.HK	HKD	6.61	2026/05/28	买入	15.16	0.03	0.05	191.58	114.95	89.29	68.13	1.40	2.10
税友股份	603171.SH	CNY	45.31	2026/04/23	买入	74.50	0.63	1.20	71.92	37.76	52.60	31.96	9.10	14.80
金蝶国际	00268.HK	HKD	6.45	2026/04/14	买入	17.94	0.10	0.20	56.08	28.04	21.24	14.40	4.10	7.60
极智嘉-W	02590.HK	HKD	12.00	2025/09/14	买入	31.51	0.31	0.60	33.66	17.39	24.22	14.01		
黑芝麻智能	02533.HK	HKD	11.94	2026/03/06	买入	28.46	-0.85	-	-	-	-	69.30	-29.00	0.10
锐明技术	002970.SZ	CNY	66.45	2026/04/27	买入	83.16	2.77	4.29	23.99	15.49	20.14	13.46	20.80	26.30
广立微	301095.SZ	CNY	102.50	2026/04/23	买入	105.60	0.68	1.14	150.74	89.91	143.56	88.87	4.10	6.40
概伦电子	688206.SH	CNY	38.40	2026/01/29	买入	44.31	0.11	0.22	349.09	174.55	232.84	118.90	2.20	4.50

数据来源：Wind、广发证券发展研究中心

备注：表中估值指标按照最新收盘价计算

下载日志已记录，仅供内部参考

目录索引

一、投资要点 7

（一）主要行业综合观点 7
（二）投资建议 8
（三）风险提示 8

二、AI 基础设施：国产替代进入快速放量的实质性阶段，Agent 推动算力架构升级 9

（一）“国产算力+国产模型”的双向奔赴加速 AI 算力国产替代进程 9
（二）CPU、超节点、AI 基础软件等受益于 AGENT 应用带来的算力架构升级趋势 13

三、模型端：大模型商业化加速落地，来自模型能力、Token 与 ARR 的多重验证 18

（一）模型能力：国产模型加速追赶，低成本推理打开 TOKEN 放量空间 18
（二）TOKEN 调用量：模型能力是否被真实使用？ 22
（三）ARR 兑现：模型使用能否兑现为商业化收入？ 25

四、AI 应用：大模型对各不同应用企业的影响因领域而异，具备强私域数据、基础通用工具类专业厂商、以及卡位应用落地“最后一公里”的软件咨询实施商或脱颖而出 30

（一）大模型对各个不同应用企业的影响因领域而异 30
（二）以上三种领域的典型代表公司 35

五、EDA：行业成长性大于周期性，国产厂商并购整合进行时 40

（一） EDA 行业在技术升级趋势下成长性较强，国产 EDA 成长弹性进一步放大 40
（二） 国产 EDA 并购整合提速，从点工具向平台化演进 43

六、算力租赁&算力调度：算力需求高增不减，算力租赁及算力调度公司迎来全新发展机遇 46

（一）应用新形态继续突破，算力需求增强态势不减 46
（二）算力租赁及算力调度公司有望迎来新的发展机遇 46

七、智能驾驶与机器人板块：高阶智驾渗透率加速抬升，油价抬升加速海外市场成长，产业玩家积极布局机器人新领域 60

（一）智能驾驶板块：智驾渗透率仍处抬升通道，但在 OEM 补贴退坡及价格竞争背景下，优选格局好的智驾芯片及出海赛道，建议关注从 0 到 1 突破的无人驾驶赛道 60
（二）机器人板块：看好智驾玩家的跨界突破，并重点推荐全球 AMR 龙头-极智嘉 62

下载日志已记录，仅供内部参考

八、风险提示 65
（一）AI 应用落地不及预期风险 65
（二）国产算力与模型生态发展不及预期风险 65
（三）行业竞争加剧与盈利下滑风险.....65
（四）宏观环境与市场流动性波动风险 65

下载日志已记录，仅供内部参考

图表索引

图 1：2024 年初-2026 年 3 月国内 AI 大模型 Tokens 日均使用量 9
图 2：2024 年 5 月-2026 年 3 月豆包大模型 Tokens 日均使用量 9
图 3：2018Q1-2026Q1 百度、腾讯和阿里巴巴资本开支（单位：亿元） 10
图 4：2019-2026Q1 英伟达在中国大陆地区营收及营收占比 11
图 5：2024 年 3 月-2026 年 3 月寒武纪存货（亿元） 12
图 6：2024 年 3 月-2026 年 3 月寒武纪预付款项（亿元） 12
图 7：硅基流动企业级 MaaS 平台架构 12
图 8：CPU 处理 AI agent 进程数的原理图 13
图 9：浪潮信息元脑 SD200 超节点 15
图 10：新华三高密 800G 智算 DCI 交换机 S12500R-64EP 16
图 11：星环科技认知数据库在 AI 算力不同层次存储系统中应用的架构图 16
图 12：星环科技认知记忆系统架构图 17
图 13：前沿大语言模型能力比较（按国别分类） 19
图 14：模型智能水平 vs. 成本 20
图 15：大语言模型推理价格随时间变化 21
图 16：模型分层价格（单位：美元/百万 tokens） 22
图 17：国内模型 API 永久降价 22
图 18：OpenRouter 周度 Token 调用（截至 2026 年 5 月 31 日，按厂商分类）（单位：十亿） 23
图 19：中国日均词元调用量（单位：万亿） 24
图 20：中国企业日均词元调用量（单位：万亿） 24
图 21：2025H2 中国企业级大模型调用市场 24
图 22：豆包日均词元调用量（单位：十亿） 25
图 23：典型模型厂商 ARR 收入（单位：百万美元） 26
图 24：OpenAI 与 Anthropic 最新营收结构 27
图 25：2023-2025 年智谱营收结构（单位：万元） 27
图 26：2023-2025 年 MiniMax 营收结构（单位：万美元） 27
图 27：Claude Code 与 Cursor 年化收入增长情况（单位：亿美元） 28
图 28：Codex 用户画像结构变化 28
图 29：赛意信息思方云协同平台“智能参数提取”功能界面 31
图 30：Anthropic 公司的金融分析解决方案的生态合作伙伴 33
图 31：AI 接入后，人效及成本获得显著优化 36
图 32：财税 SaaS 企业将 AI 技术创新融入产品中以保持竞争力 37
图 33：汉得信息 AI 能力架构 39
图 34：汉得信息 AI Agent 建设概览 39
图 35：全球 EDA 市场空间及增速（单位：亿美元） 40
图 36：国内 EDA 市场空间及增速（单位：亿元） 40

下载日志已记录，仅供内部参考

图 37：摩尔定律下每个微处理器上晶体管数量 41
图 38：不同制程芯片的设计成本 41
图 39：全球 ASIC 市场空间（单位：十亿美元） 42
图 40：美国对华的 EDA 管控加码 42
图 41：2024 年全球 EDA 市场竞争格局 43
图 42：2022 年国内 EDA 市场空间及增速（单位：亿元） 43
图 43：全球前九大 CSP 的 Capex 预测（2023-2026F） 47
图 44：算力租赁分类模式 48
图 45：算力租赁业务流 49
图 46：算租公司从采购安装-部署确收-稳定运行的各阶段利润率参考 50
图 47：中国智算能力租赁市场行业集中度 50
图 48：2025H2 中国企业级大模型调用市场 50
图 49：我国各省 IDC 上架率情况 52
图 50：“东数西算”八大算力枢纽、十大集群算力政策要点 53
图 51：算力互联网各节点和平台关系图 55
图 52：算力互联网体系相关主体图 56
图 53：电信运营商和第三方算力中心服务商特点比对 57
图 54：算力订购流程简析 58
图 55：我国智能驾驶渗透率 61
图 56：2023 年我国智驾芯片市占率 61
图 57：2024 年我国智驾芯片市占率 61
图 58：美国智驾渗透率 62
图 59：欧洲智驾渗透率 62
图 60：2024-2030 年机器人 3D 激光雷达市场规模（亿元） 63
图 61：极智嘉已形成商业壁垒正循环 64

表 1：英伟达 GB200、H200、A100 和华为昇腾 910C 性能对比 11
表 2：华为 Atlas 950、960 SuperPoD 和 Vera Rubin NVL144 性能对比 14
表 3：浪潮信息元脑 SD200 超节点性能指标 15
表 4：EazyDevelop 产品科亿知识库模板的软件架构 32
表 5：Anthropic 在金融分析解决方案中与生态合作伙伴的合作方式梳理 34
表 6：Anthropic 和 Infosys 合作中的应用场景梳理 34
表 7：大语言模型 Tokens 消耗量 46
表 8：agent 产品的 Tokens 消耗量 46
表 9：异构算力需要处理不同硬件架构 53

下载日志已记录，仅供内部参考

一、投资要点

（一）主要行业综合观点

算力端：

（1）“国产算力+国产模型”的双向奔赴加速 AI 算力国产替代进程：DeepSeek V4、Kimi K2.6 等国产模型与华为昇腾、寒武纪、壁仞等国产 AI 芯片的主动适配，有力推动了国产 AI 芯片的商业化进程和生态发展。再考虑到性能的提升、供应链的恢复以及政策的支持，AI 芯片的国产替代在互联网、模型公司等进入快速放量的实质性阶段。

（2）CPU、超节点、AI 基础软件等受益于 Agent 应用带来的算力架构升级趋势：Agent 已成为 Tokens 增长的主要驱动力，其不仅需要 GPU 的并行计算能力，对于 CPU 的核心数、芯片间的高速互联以及存储硬件间的数据同步提出了更高的要求。国产 CPU 公司、服务器公司以及数据基础软件公司针对 Agent 的应用已开发出相应算力产品，有望受益于算力架构升级而放量。

模型端：

（1）从模型能力看，中美头部模型差距仍在，但国产模型正快速追赶。从成本端看，硬件升级、模型架构优化、系统工程改进等共同推动推理成本下降。

（2）调用量高增反映模型能力正被真实使用。海外方面，全球开发者/API 生态中的模型调用需求持续释放。国内方面，2026 年 3 月我国日均 Token 调用量突破 140 万亿；2025 年下半年企业级大模型日均调用量达 37 万亿 tokens，豆包大模型日均 Token 使用量在 2026 年 3 月已突破 120 万亿，较年初实现翻倍。

（3）ARR 方面，海外头部率先形成大规模收入，国内厂商仍处于较小收入体量阶段，但增长弹性更高。OpenAI 2026 年 2 月 ARR 已超 US$25bn，Anthropic 2026 年 5 月 ARR 已超 US$47bn，模型能力正在转化为真实收入。

我们判断，后续 AI 投资判断不应只停留在模型榜单排名，而应更加关注模型能力能否转化为真实调用和付费。具备模型能力、应用入口、客户基础和数据沉淀的公司，有望在这一轮 AI 应用落地过程中率先受益。对于模型厂商，后续可重点观察真实调用、客户转化、收入兑现和盈利质量，核心指标包括 Token 调用量、企业客户数量、C 端付费转化、API/开放平台收入、ARR 增长以及单位 Token 成本等。

AI 应用端：

大模型对各个不同应用企业的影响因领域而异，对于功能标准化程度较高、技术门槛低的软件影响较大，但长期来看，部分具有行业壁垒和特定环节的软件公司仍有发展前景，具体来看，有以下三类：

（1）特定领域的数据专业性强，且非公开、非通用，相关公司如果能跟得上 AI 的发展，仍将会在未来有生存发展空间。

（2）基础通用工具类专业厂商如 Coding、database、IDE 开发工具厂商等所提供的底层基础工具是 AI 软件开发生态的核心基石，在 AI 新时代具备长期成长红利。

下载日志已记录，仅供内部参考

（3）即使在 AI 试图主导的新软件时代生态中，数据增值服务商和咨询实施开发集成商仍是不可或缺的产业链价值环节所在。

算力租赁和算力调度：

在行业整体资本开支大幅抬升、算力需求持续高景气、算力基建加码投放的三重核心背景支撑下，算力租赁、算力调度赛道相关企业有望迎来明确成长机遇。其一，头部云厂商大规模资本开支落地将持续扩充算力硬件供给池，为算力租赁服务商锁定稳定、可持续的货源基础，有力托底租赁业务营收规模与经营稳定性；其二，下游 AI 应用端旺盛且长期刚性的算力需求，叠加海量异构算力资源集中落地投产，市场对高效资源分配、负载均衡、跨池协同的算力调度配套服务需求同步释放，调度环节作为算力产业链运营提效的关键枢纽，价值与业务空间同步打开。

智能驾驶与机器人板块：

（1）智驾板块：在比亚迪等头部 OEM 加码智驾趋势下，国内智驾渗透率加速抬升，但在价格竞争仍未明显缓解，产业链仍处于市占率收敛的背景下，智驾零部件环节主要投资机会集中在竞争格局较好的智驾芯片行业、规模效应明显的龙头 Tier1 以及充分受益于细分领域成长的个股（如受益于欧标前装政策的锐明技术）。此外，智驾产业参与者立足现有技术积累，积极布局机器人新领域，有望打开新空间。

（2）机器人板块：看好智驾玩家的跨界突破，并重点推荐全球 AMR 龙头-极智嘉。

我们重点推荐 AI 国产算力、模型、工具和私域专业应用以及 EDA 与设备，建议关注智能驾驶变化。

（二）投资建议

算力端：推荐寒武纪、浪潮信息、紫光股份、星环科技，同时建议关注中科曙光、范式智能。

模型端：建议关注智谱、月之暗面、DeepSeek、MiniMax，以及字节等厂商。

AI 应用：推荐卓易信息、赛意信息、汉得信息、中控技术、晶泰控股、税友股份，建议关注中望软件、金蝶国际等。

算力租赁与算力调度：算力租赁方面，我们建议关注杰创智能、智微智能、赛意信息、宏景科技。算力调度方面，我们建议关注卡莱特、并行科技、思特奇、云赛智联等。

智能驾驶与机器人：极智嘉、黑芝麻智能、锐明技术；同时建议关注：德赛西威、速腾聚创、小马智行、文远知行等。

EDA 等工业软件：广立微、概伦电子、中控技术、中望软件、索辰科技等。

（三）风险提示

AI 应用落地不及预期风险；国产算力与模型生态发展不及预期风险；行业竞争加剧与盈利下滑风险；宏观环境与市场流动性波动风险。

下载日志已记录，仅供内部参考

二、AI 基础设施：国产替代进入快速放量的实质性阶段，Agent 推动算力架构升级

投资要点：

（1）“国产算力+国产模型”的双向奔赴加速 AI 算力国产替代进程：国产模型与国产 AI 芯片的主动适配，有力推动了国产 AI 芯片的商业化进程和生态发展。再考虑到性能的提升、供应链的恢复以及政策的支持，AI 芯片的国产替代在互联网、模型公司等进入快速放量的实质性阶段。国产 AI 芯片产业链恢复的节奏可以在 26 年 8 月各家公司中报披露的存货、预付款项中判断、验证。

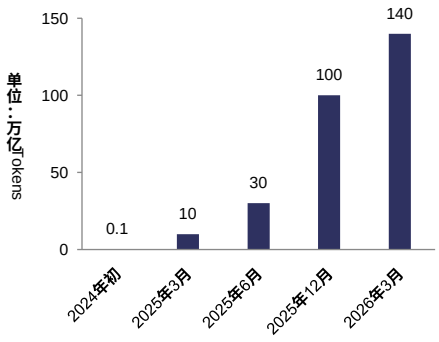
（2）CPU、超节点、AI 基础软件等受益于 Agent 应用带来的算力架构升级趋势：Agent 已成为 Tokens 增长的主要驱动力，其不仅需要 GPU 的并行计算能力，对于 CPU 的核心数、芯片间的高速互联以及存储硬件间的数据同步提出了更高的要求。国产 CPU 公司、服务器公司以及数据基础软件公司针对 Agent 的应用已开发出相应算力产品，有望受益于算力架构升级而放量。超节点放量的时间主要是在 26 年下半年，有望推动国产服务器公司盈利能力提升，从 26 年 10 月服务器厂商披露的毛利率等数据可以判断、验证。在 Agent 场景中应用的 CPU 和 AI 基础软件等产品仍需要升级迭代，2027 年上半年，相关产品有望放量。

（3）芯片领域，重点推荐寒武纪、国产 CPU 头部公司；AI 基础软件领域，推荐星环科技，建议关注范式智能；AI 服务器领域，推荐浪潮信息和紫光股份，建议关注中科曙光。

（一）“国产算力+国产模型”的双向奔赴加速 AI 算力国产替代进程

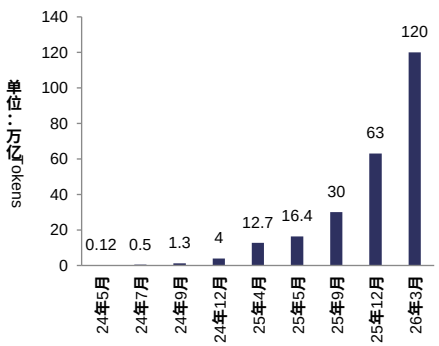
国产 AI 大模型 tokens 的增长驱动推理侧算力快速增长。根据国家数据局数据，截至 2026 年 3 月底，中国 AI 大模型日均 Tokens 消耗量已达 140 万亿，2025 年底该数值约为 100 万亿；根据火山引擎官微数据，2026 年 3 月，豆包大模型日均 tokens 使用量为 120 万亿，较 2025 年 12 月底的 63 万亿大幅增长。国产 AI 大模型 Tokens 的快速增长反映了其在编码、智能体及视频生成等多个场景的高频使用。展望 2026 年下半年，随着国产 AI 大模型在各行各业的应用落地过程中的商业化效果逐渐显现，互联网厂商与模型厂商对于推理侧算力的需求持续增长，相应算力的投入也更具持续性。

图1：2024年初-2026年3月国内AI大模型Tokens日均使用量



数据来源：国家数据局，广发证券发展研究中心

图2：2024年5月-2026年3月豆包大模型Tokens日均使用量

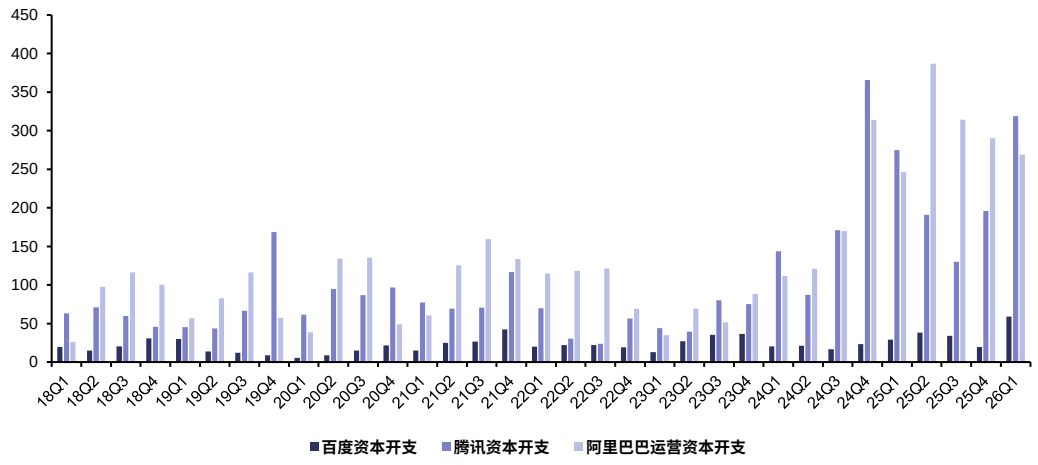


数据来源：火山引擎官方微信公众号，广发证券发展研究中心

下载日志已记录，仅供内部参考

AI 投资与互联网厂商核心业务已开始形成协同效应，AI 算力投入高增长可持续。阿里云通过百炼和人工智能平台（PAI）为企业提供用户友好的环境以构建和部署生成式 AI 应用。根据阿里巴巴财报电话会议，截至 26Q1 末，阿里云外部商业化收入增速加速至 40%，AI 相关产品收入连续十一个季度实现三位数增长，AI 相关产品收入已占阿里云外部商业化收入的 30%。腾讯方面，根据公司 26Q1 财报，广告智能投放产品 AIM+已在小游戏、短剧微信小店广告主中广泛应用；内容推荐模型的应用驱动视频号总用户使用时长增长超 20%。我们认为，AI 投资与腾讯、阿里等大厂的核心业务已开始形成协同效应，并逐步实现商业化落地。在 AI 应用持续渗透的趋势下，展望 2026 年下半年，AI 基础设施相关投入有望持续保持在较高水平。

图 3：2018Q1-2026Q1百度、腾讯和阿里巴巴资本开支（单位：亿元）

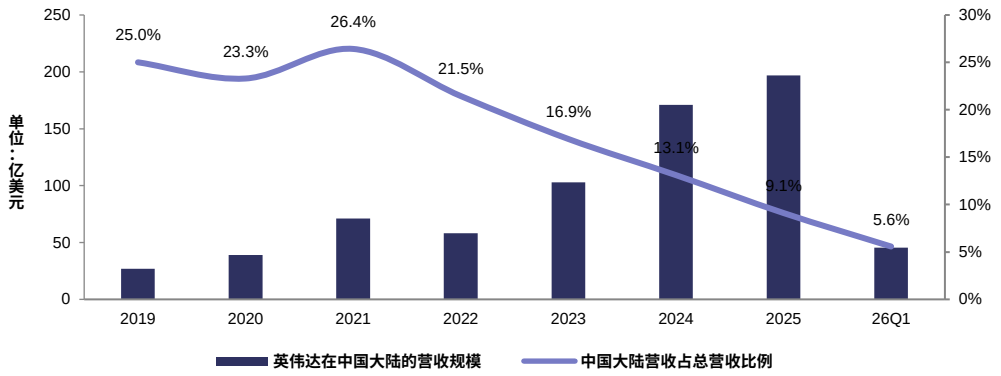


数据来源：iFinD，广发证券发展研究中心

国内 AI 芯片市场竞争格局进一步优化，国产 AI 芯片在国内互联网厂商的采购比例有望提升。根据英伟达财报，H200 芯片需要获得美方许可证才可向客户销售。在政策的影响下，26Q1，中国大陆市场给英伟达贡献营收的比例为 5.6%，持续下滑且以消费级游戏显卡为主。另一方面，在下游应用占比较大的推理场景中，以华为昇腾和寒武纪思元为代表的国产 AI 芯片的性能逐渐接近 H200。再考虑到国产 AI 芯片的性价比优势，2026 年，国产 AI 芯片产品在国内科技公司 AI 算力采购计划的比例有望大幅提升。2025 年，英伟达在中国大陆的销售额为 197 亿美元。展望 2026 年下半年，国产 AI 芯片有望填补英伟达限制销售带来的市场空间实现跨越式增长。

下载日志已记录，仅供内部参考

图 4：2019-2026Q1英伟达在中国大陆地区营收及营收占比



数据来源：英伟达财报，广发证券发展研究中心

推理场景中，AI 芯片国产替代稳步推进。尽管在模型训练场景，国产 AI 芯片与 H200 仍有一定差距，但已能满足推理场景的需求。我们通过对比国产 AI 芯片与全球头部 AI 芯片的性能可以发现，尽管在模型训练场景，国产 AI 芯片与 H200 仍有一定差距，但在强调高吞吐、高容量的推理场景中，国产 AI 芯片与 H200 性能接近。昇腾 910C 的内存容量是 128GB，与 H200 的 141GB 接近。而计划在 26Q4 推出的昇腾 950DT 的存储容量将进一步提升至 144GB。展望 2026 年下半年，Tokens 的增长主要与模型推理相关，国产 AI 芯片在推理场景的性能与 H200 等 AI 芯片接近的情况下，AI 芯片国产替代的应用场景和范围有望扩大。

表 1：英伟达 GB200、H200、A100 和华为昇腾 910C 性能对比

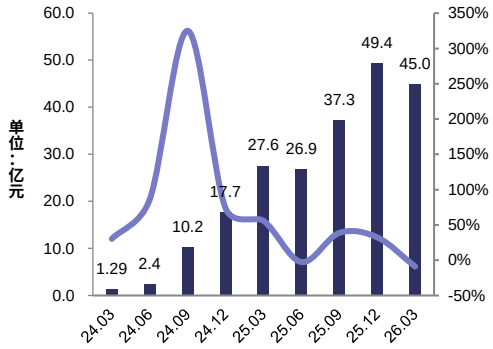
指标	Nvidia GB200	Nvidia H200	Nvidia A100	华为 Ascend 910C
BF16 密集计算性能	2,500 TFLOPS	1671 TFLOPS	624 TFLOPS	780 TFLOPS
HBM 容量	192 GB	141GB	40GB	128 GB
HBM 带宽	8 TB/s	4.8 TB/S	1.6 TB/S	3.2 TB/s
纵向扩展带宽（单向）	7,200 Gb/s	-	-	2,800 Gb/s
横向扩展带宽（单向）	400 Gb/s	-	-	400 Gb/s

数据来源：SemiAnalysis，英伟达官网，广发证券发展研究中心

寒武纪等国产 AI 芯片的供应链已有一定程度的恢复。寒武纪的存货由 2025 年一季度末的 27.6 亿元，增长至 2026 年一季度末的 45.0 亿元；公司的预付款项由 2025 年一季度末的 9.7 亿元，增长至 2026 年一季度末的 19.0 亿元。26Q1，公司预付款项增加了 11.5 亿元。寒武纪存货和预付款项的大量增加反映了国产 AI 芯片供应链稳步改善修复的趋势。在此背景下，寒武纪在于 2026 年 5 月 22 日发布公告拟向银行申请不超过人民币 120 亿元的综合授信额度，也反映了其对于供应链恢复和下游客户拓展方面的信心。展望 2026 年下半年，随着寒武纪等国产 AI 芯片供应链的恢复，其为下游互联网客户的产品交付做好了准备，收入规模有望上新的台阶。

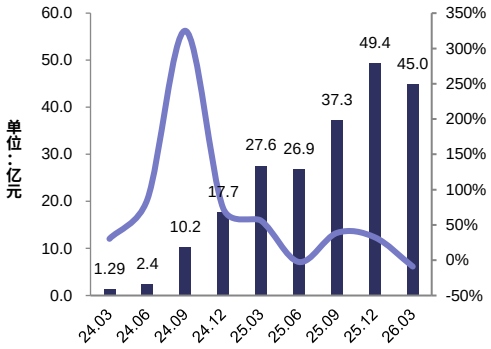
下载日志已记录，仅供内部参考

图5：2024年3月-2026年3月寒武纪存货（亿元）



数据来源：寒武纪财报，广发证券发展研究中心

图6：2024年3月-2026年3月寒武纪预付款项（亿元）



数据来源：寒武纪财报，广发证券发展研究中心

国产 AI 大模型加快适配国产 AI 芯片，生态建设节奏加快。DeepSeek V4 和 Kimi K2.6 在发布的第一时间即选择与华为昇腾、寒武纪思元和壁仞等国产 AI 芯片产品适配，开辟了国产 AI 大模型适配国产算力产品的新方向。此外，在政策层面，根据国家发改委官网，在 5 月份的新闻发布会上提到“我们始终坚持系统布局、分业施策、开放共享、安全可控，推动人工智能与经济社会各行业各领域广泛深度融合，指导国产大模型加大力度适配国产算力芯片。”国产 AI 产业自主可控建设正由国产 AI 芯片的单向适配向“国产算力+国产模型”双向奔赴转变。在此趋势下，26 年下半年，国产 AI 芯片软件生态拓展的节奏有望加快，为其后续应用场景和商业拓展奠定了良好基础。

图 7：硅基流动企业级 MaaS 平台架构



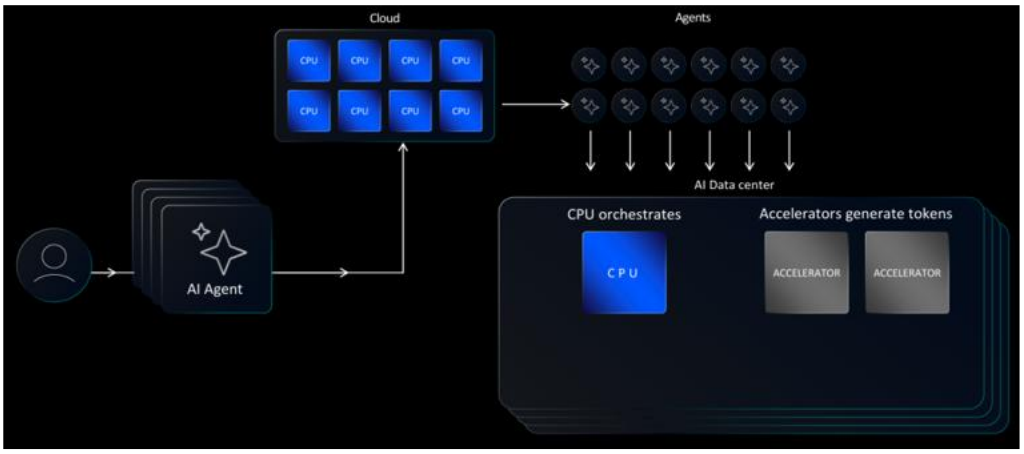
数据来源：硅基流动官网，广发证券发展研究中心

下载日志已记录，仅供内部参考

（二）CPU、超节点、AI 基础软件等受益于 Agent 应用带来的算力架构升级趋势

AI Agent 在各行业的渗透有望推动 CPU 呈现数量扩张与性能升级的双重趋势。AI Agent 的应用涉及资源调用、工具调用、多智能体协同等，较大程度依赖 CPU 进程数、核心数以及单核性能等硬件资源。为满足 AI agent 的应用，一方面，服务器 CPU 配置正从传统 8-16 核，向 32-64 核及更高规格演进；另一方面，在算力集群中配置更多数量的 CPU 或成为各厂商尝试的新路径。根据英特尔披露的投资者调研纪要，CPU 与 GPU 的配置比例从 1:8 向 1:1 靠拢，甚至可达 4:1。展望 2026 年下半年，CPU 和 GPU 的配置比例有望持续提升，国内 CPU 公司有望受益于技术架构升级带来的潜在市场规模扩大的趋势。

图 8：CPU 处理 AI agent 进程数的原理图



数据来源：ARM 官网，广发证券发展研究中心

国产超节点快速迭代，互联带宽与存储容量配置提升较大。根据 CSDN 官网，2025 年 9 月 18 日，华为在全联接大会上发布超节点产品 Atlas 950 SuperPoD 和 Atlas 960 SuperPoD 的技术路线图，预计分别于 2026 年第四季度和 2027 年第四季度上市。与英伟达的 Vera Rubin NVL144 超节点相比，Atlas 950、960 SuperPoD 超节点采用更多数量的芯片数量互联从而实现更高水平的算力，而互联带宽和存储容量等配置的提升是其算力提升的关键。

国产超节点与国产大模型生态适配，有望受益于国产 AI 大模型的应用落地。根据华为官方微信公众号，昇腾 950 通过融合 kernel 和多流并行技术降低 Attention 计算和访存开销，大幅提升推理性能，实现了高吞吐、低时延的 DeepSeek V4 模型推理部署。根据 DeepSeek 官方微信公众号，2026 年 4 月 24 日，DeepSeek V4 发布时提到“受限于高端算力，目前 Pro 的服务吞吐十分有限，预计下半年昇腾 950 超节点批量上市后，Pro 的价格会大幅下调”。而在 2026 年 5 月底，DeepSeek 官网上公布了 DeepSeek V4 Pro 价格将于 2026 年 5 月 31 日结束 2.5 折优惠，正式调整为原定价的四分之一。展望 2026 年下半年，在生态适配，软硬件协同优化的基础上，国产超节点产品有望随着国产 AI 大模型的应用落地而放量。

下载日志已记录，仅供内部参考

表 2：华为 Atlas 950、960 SuperPoD 和 Vera Rubin NVL144 性能对比

超节点	Atlas 950 SuperPoD	Atlas 960 SuperPoD	Vera Rubin NVL144	华为 Ascend 910C
推出时间	2026H2	2027Q4	2026H2	780 TFLOPS
芯片类型	昇腾950DT	昇腾960	Vera CPU+Rubin GPU	128 GB
AI芯片数量	8192	15488	72	3.2 TB/s
存储容量	1152 TB	4460 TB	75 TB	2,800 Gb/s
互联带宽	16.3 PB/s	-	260 TB/s	400 Gb/s
FP8算力	8 EFLOPS	30 EFLOPS	1.2 EFLOPS	
FP4算力	16 EFLOPS	60 EFLOPS	3.6 EFLOPS	

数据来源：convequity.substack.com，广发证券发展研究中心

超节点新架构有望推动服务器整机厂商的盈利能力提升。与服务器相比，超节点中新增了交换机等网络设备。而根据浪潮信息 2025 年年报，交换机等产品的毛利率高于服务器产品的毛利率。因此，若超节点产品放量，高毛利率的交换机产品占整体营收比例有望提升，从而推动浪潮信息整体盈利提升。除了硬件层面的优化以外，超节点产品还针对常用的模型算法进行了软件层面的优化，来提升模型的训推效率。例如，浪潮信息针对超节点产品 SD200 开发的通信库、并行推理框架、PD 分离策略以及动态负载均衡等技术，充分发挥了超节点的架构优势，实现 DeepSeek R1 最高 16.3 倍的超线性扩展率。展望 2026 年下半年，具备软硬件协同优化特点的超节点产品放量有望增强服务器整机厂商的产品竞争力和盈利能力。

下载日志已记录，仅供内部参考

图 9：浪潮信息元脑 SD200 超节点



数据来源：浪潮信息官网，广发证券发展研究中心

表 3：浪潮信息元脑 SD200 超节点性能指标

芯片配置	64 路 GPU
显存容量	4 TB
内存容量	64 TB
互联架构	3D Mesh
Intra-Node 延迟	0.69 微秒
Inter-Node 延迟	1.64 微秒
Tokens 生成速度（基于 DeepSeek-R（1）	112 tokens/s
单位 Token 生成时间 （基于 DeepSeek-R（1））	8.9 毫秒

数据来源：浪潮信息官网，广发证券发展研究中心

国产算力公司有望受益于算力向集群化方向发展的趋势。根据新华三官网，2026 年 5 月，新华三推出高密 800G 智算 DCI 交换机 S12500R-64EP 搭载超高密 800G OSFP 端口，构建超大带宽、高集约的硬件传输底座，单设备可支撑海量训练流量并行传输，大幅提升单纤带宽利用率，充分适配大模型持续训练、高频参数同步的带宽诉求，有效解决训练流量带宽瓶颈。新华三在交换机领域深耕多年，技术积累深厚，其 800G 智算 DCI 交换机产品性能行业领先。展望 2026 年下半年，新华三的交换机产品具备与以英伟达为代表的海外竞品竞争的技术基础，其 800G 智算 DCI 交换机产品有望在产品升级和国产替代的过程中放量。

下载日志已记录，仅供内部参考

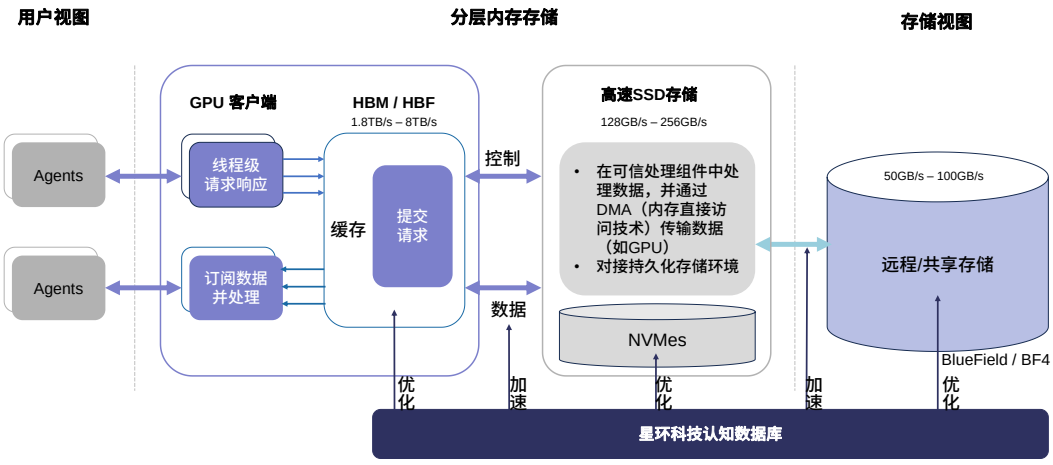
图 10：新华三高密 800G 智算 DCI 交换机 S12500R-64EP



数据来源：新华三官网，广发证券发展研究中心

AI 基础软件有望解决 AI 大模型推理任务的瓶颈而实现商业化突破。一方面，不同存储的物理介质不同导致数据读写的底层技术有差异；另一方面，不同存储介质间的数据传输速率不同，数据同步和对齐的难度较大。因此，优化不同存储介质上的数据读写以及加速不同存储硬件间的数据传输已成为突破 AI 大模型推理任务瓶颈的关键。而星环科技近期开发的认知数据库正是主要用于解决 AI 大模型推理过程中内存容量瓶颈的问题。根据星环科技披露的投资者调研纪要，截至 2026 年 4 月 1 日，星环科技认知数据库的单机版已开发完成，今年下半年将推出多卡版本，有望在 AI 推理应用的过程中逐步体现其商业价值。

图 11：星环科技认知数据库在 AI 算力不同层次存储系统中应用的架构图

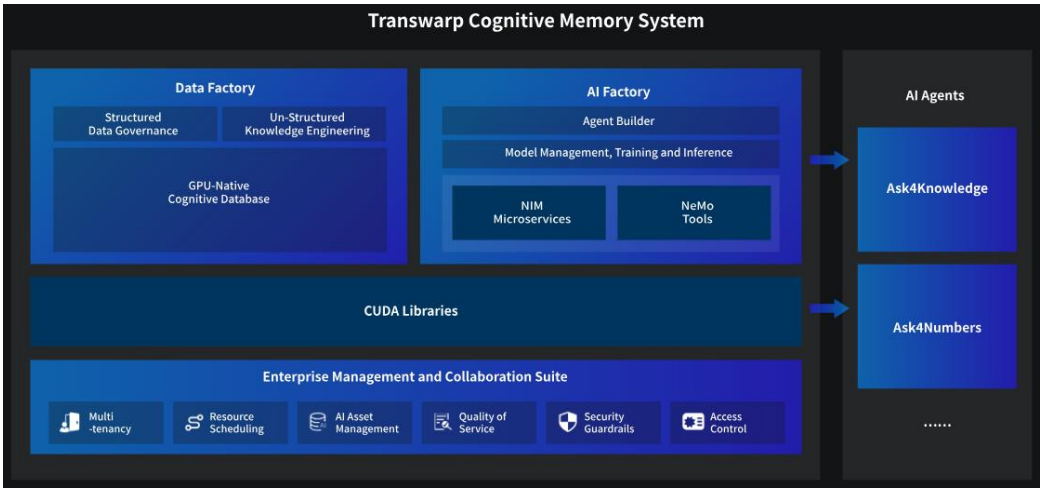


数据来源：广发证券发展研究中心

下载日志已记录，仅供内部参考

星环科技认知数据库预计上线云版本，商业模式转变在即。根据星环科技披露的投资者交流纪要，认知数据库在商业化策略上，预计 26 年下半年上线云版本，先进入海外市场，以高性价比吸引海外客户。公司目前在海外市场同步推进东南亚、中东及美洲区域的拓展，依托现有云资源搭建服务。相较于在国内市场以本地化部署为主的商业模式，基于云平台的收费模式更具弹性和可持续性，面对的潜在市场空间更大。与认知数据库对标的产品，Databricks 公司的 Photon Engine 和 Snowflake 公司的 Gen2 产品，即是以采用按照内存容量、CPU 数量等云资源使用量来进行收费。未来，随着认知数据库公有云版本的推出，其若按照 AI 推理任务相关的数据流量收费，则有望推动公司商业模式的转变，打开更大的潜在市场空间。

图 12：星环科技认知记忆系统架构图



数据来源：星环科技官网，广发证券发展研究中心

下载日志已记录，仅供内部参考

三、模型端：大模型商业化加速落地，来自模型能力、Token 与 ARR 的多重验证

投资要点：模型层，建议重点关注具备前沿模型能力、高推理效率和商业化承接能力的厂商。国内模型能力持续提升，使国产模型具备进入企业工作流和开发者生态的基础。对于模型厂商，后续应重点观察真实调用、客户转化、收入兑现和盈利质量，核心指标包括 Token 调用量、企业客户数量、C 端付费转化、API/开放平台收入、ARR 增长以及单位 Token 成本等，而不应仅以单次模型评测排名作为判断依据。推荐智谱，建议关注 DeepSeek、Kimi、 MiniMax、字节等厂商。

（一）模型能力：国产模型加速追赶，低成本推理打开 Token 放量空间

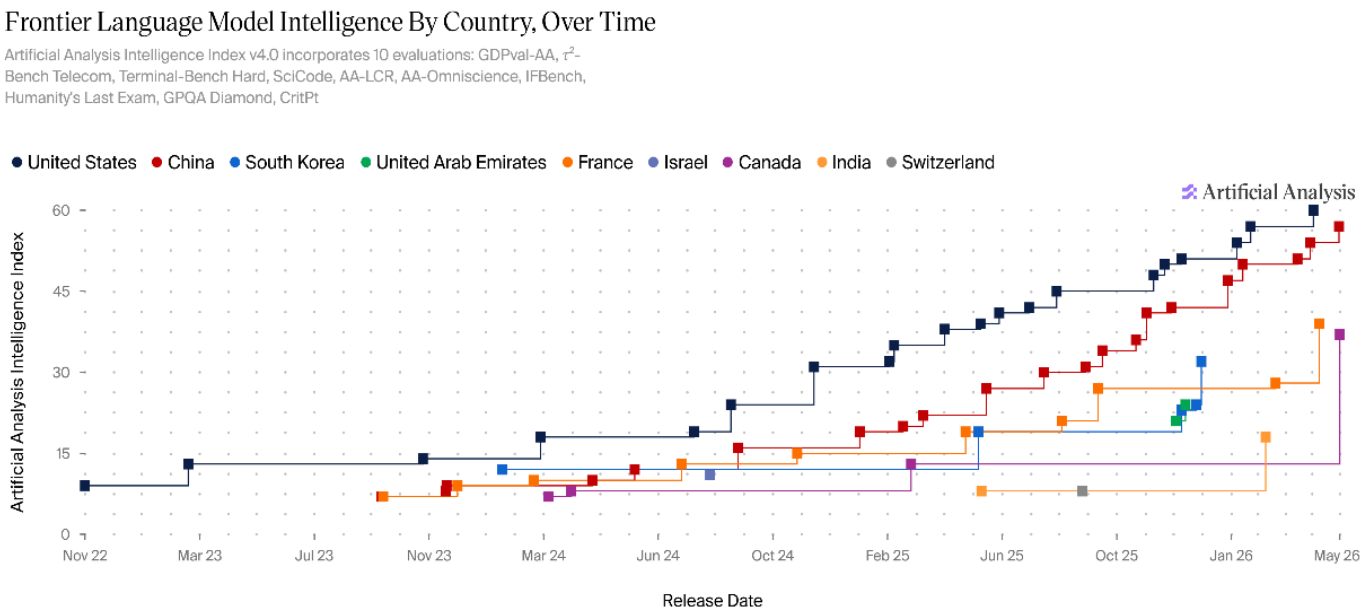
随着模型能力持续提升、推理成本快速下降，大模型正加速迈向规模化商业落地。从能力端看，今年以来，中美厂商模型能力差距持续收窄，国产大模型逐步跻身全球前沿梯队；从成本端看，受益于硬件性能提升、模型架构优化和系统工程改进，推理成本不断下探，进一步打开 Token 放量空间。**能力提升与成本下降形成正向循环，有望加速 Agent、AI Coding 及企业流程类应用走向规模化应用。**

1．国产模型追赶加速，已位居全球前沿梯队

从全球模型竞争格局看，中美头部模型差距仍在，国产模型加速追赶。DeepSeek 在 V4 技术报告中对自身前沿位置作出判断：DeepSeek-V4-Pro 在标准推理基准上已优于 GPT-5.2 和 Gemini-3.0-Pro，但仍落后于 GPT-5.4 和 Gemini-3.1-Pro，即大约落后最前沿模型 3-6 个月。根据 Stanford 2026 AI Index 数据，截至 2026 年 3 月，美国的顶级模型相对中国顶级模型的领先幅度仅为 2.7%。此外，从 Artificial Analysis 前沿语言模型能力指数看，2022–2023 年美国模型长期占据全球领先位置；但自 2024 年以来，中国模型能力曲线快速抬升，并在 2025 年后进一步逼近美国前沿模型区间。整体来看，**尽管美国模型仍处于全球最高能力带，但中国模型已成为全球前沿梯队中持续活跃、快速追赶且具备竞争力的重要力量。**

下载日志已记录，仅供内部参考

图 13：前沿大语言模型能力比较（按国别分类）



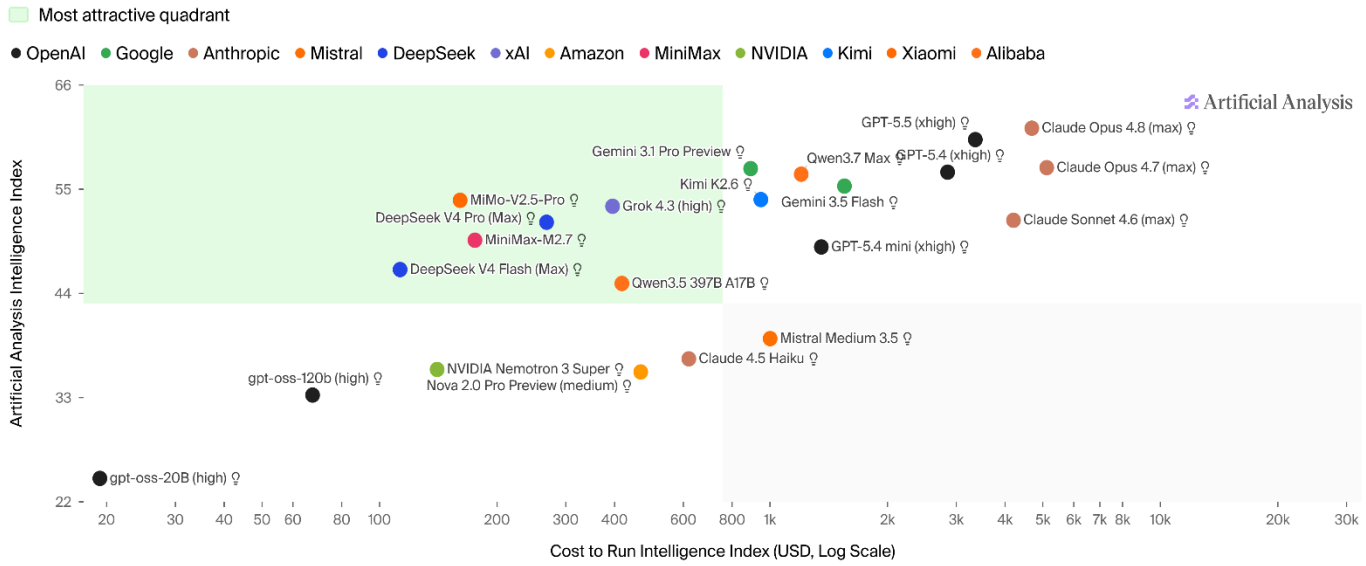
数据来源：Artificial Analysis、广发证券发展研究中心

美国头部厂商仍在闭源旗舰模型、产品化闭环和全球开发者工具链上具备优势。根据 Stanford 2026 AI Index，2025 年美国机构共推出 59 个 notable AI models，中国为 35 个；但从趋势看，领先倍数已由 2024 年的约 2.7 倍（40/1（5）收窄至 2025 年的约 1.7 倍（59/3（5），中国模型供给数量提升更快。除模型本身外，美国头部厂商的优势进一步延伸至产品闭环与开发者生态：OpenAI 围绕 ChatGPT、API 与企业级产品构建起覆盖个人用户、开发者和企业客户的商业化体系；Anthropic 通过 Claude Code 在 Agentic Coding 和企业级应用场景中形成较强产品化能力；Google 则凭借 Gemini API、AI Studio、Android、Workspace 和 Google Cloud 生态，将模型能力深度嵌入开发工具、操作系统和生产力软件之中。

国产模型厂商则在高性价比推理和开源生态扩散上形成差异化优势。中国头部模型在能力快速追赶的同时，正通过更低推理成本提升开发者调用和应用落地的经济性。缓存未命中情况下，DeepSeek-V4-Pro 当前每百万输入 / 输出 Token 价格分别为 0.435 / 0.87 美元，显著低于 OpenAI GPT-5.5 的 5 / 30 美元，以及 Anthropic Claude Opus 4.8 的 5 / 25 美元。考虑到海外厂商同样布局低价模型，国内厂商的优势并非简单体现为低价，而是在高能力模型的成本效率维度更为突出。**开源扩散方面，**Hugging Face 2026 年春季开源生态报告显示，2025 年 Hugging Face 用户数达到 1300 万，公开模型数量超过 200 万个；其中，中国模型在过去一年快速成为下载量主力，约占总下载量的 41%，并在月度下载量和累计下载量上超过美国模型。此外，根据阿里云官方公众号，2026 年 1 月 21 日，Qwen 在 Hugging Face 上的衍生模型数突破 20 万个，成为全球首个达成此目标的开源大模型，同时下载量突破 10 亿次，反映出国产开源模型在全球开发者生态中的影响力正在快速提升。

下载日志已记录，仅供内部参考

图 14：模型智能水平 vs. 成本



数据来源：Artificial Analysis、广发证券发展研究中心

2．推理 COGS ：每百万 Token 成本的拆解

低成本推理是 Token 大规模放量的前提。以每百万 Token 的推理 COGS 衡量，单位成本可以近似表示为：

$$\text{COGS}_{(1\text{M tokens})} = \frac{C_{\text{cluster/hour}}}{T_{\text{delivered}}} \times \frac{10^6}{3600}$$

其中，$C_{\text{cluster/hour}}$**表示推理服务集群在单位时间内消耗的总服务成本。**若仅考虑 GPU 成本，则可以进一步写为：$C_{\text{cluster/hour}} = n_{\text{GPU}} p_{\text{GPU-hour}}$，$n_{\text{GPU}}$ 表示服务所需 GPU 数量，$p_{\text{GPU-hour}}$ 表示单张 GPU 的小时成本。若从完整 COGS 角度衡量，还应进一步纳入 CPU、内存、网络、存储、电力、机房、调度冗余和运维等配套成本。

分母中 $T_{\text{delivered}}$ **表示系统在给定 Workload 和 SLA 约束下，实际稳定交付的总吞吐量，即** aggregate tokens/sec。这里的吞吐并非理论峰值吞吐或单卡吞吐（TPS），而是在真实线上服务条件下，扣除调度损耗、batch 不满、负载波动、显存约束和延迟约束之后，系统实际能够交付给用户的 Token 产出速度。

因此，推理降本本质上可以从两个方向理解：一方面，降低单位时间内的算力成本，即降低 $C_{\text{cluster/hour}}$；另一方面，提高单位时间内可交付的 Token 数量，即提升 $T_{\text{delivered}}$。进一步看，真正决定单位经济性的并不是绝对 GPU 数量，而是每个 Token 消耗的有效 GPU 时间，也就是 GPU-seconds/token。若 GPU 数量增加的同时吞吐量同比提升，单位 Token 成本未必上升；反之，若受显存、通信、调度、batching 或 SLA 约束影响，吞吐无法随 GPU 数量线性扩展，则单位成本会恶化。

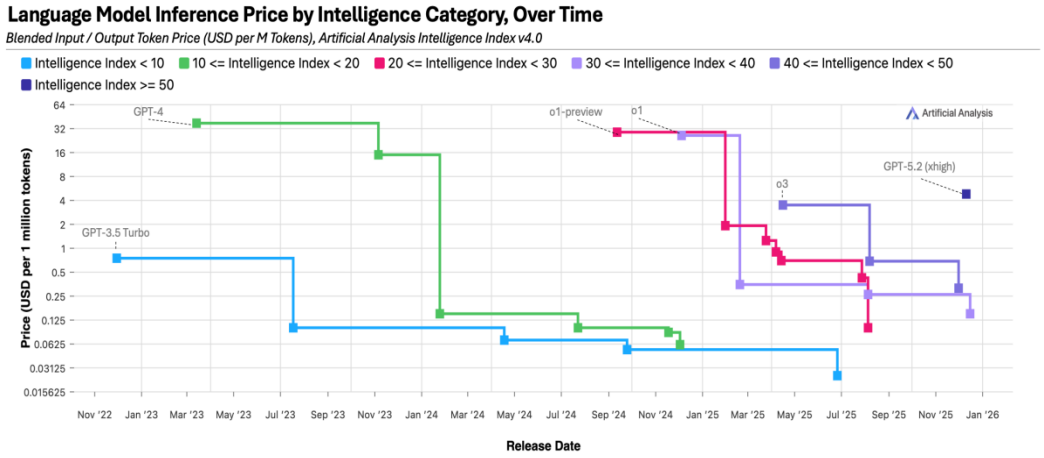
沿着“降低分子、放大分母”的思路继续拆解，影响这一公式的变量可以归纳为四类：硬件、模型架构、系统

下载日志已记录，仅供内部参考

工程，以及 Workload/SLA 约束。其中前三者更多决定供给侧效率，Workload/SLA 则决定真实服务场景下有效吞吐能否接近理论吞吐。

当前，推理降本正在硬件、模型架构、系统工程和负载适配四个层面同时推进，并已在模型 API 价格中有所体现。Artificial Analysis 2025 年末报告指出，模型能力与调用成本之间的关系持续改善——企业既可以在相同预算下调用更强模型，也可以以更低成本获得接近此前前沿模型的能力。以 2025 年初 o1 级智能水平为参照，对应模型的 blended token 价格已下降约 128 倍。其最新实时榜单进一步显示，这一趋势仍在延续。o1 在 Artificial Analysis Intelligence Index 上约为 31 分，输入和输出价格分别为 15 美元/百万 Token 和 60 美元/百万 Token；而截至 2026 年 5 月，得分处于 31–33 区间、接近或略高于 o1 水平的部分模型，其第三方调用 blended token 价格已降至约 0.11–0.20 美元/百万 Token 区间，表明在同等或相近能力水平下，模型调用成本正在快速下行。

图 15：大语言模型推理价格随时间变化



数据来源：Artificial Analysis、广发证券发展研究中心

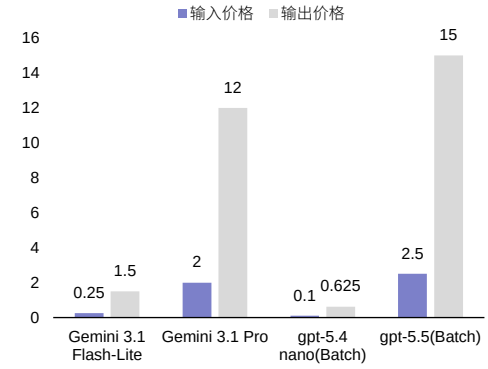
头部模型厂商构建了多层次模型体系，通过不同模型档位覆盖差异化场景降低平均调用成本。以 Google 为例，Gemini 3.1 Flash-Lite 定价为 0.25 美元/百万输入 Token、1.50 美元/百万输出 Token，明显低于 Gemini 3.1 Pro 在短上下文下的 2.00 美元/百万输入 Token、12.00 美元/百万输出 Token，官方定位也更偏向高吞吐、成本敏感场景。OpenAI 的价格梯度同样明显：对于成本敏感型、多步骤、可异步拆解任务，在 Batch/Flex 模式下，gpt-5.5 短上下文输入和输出价格分别为 2.50 美元和 15.00 美元/百万 Token，而 gpt-5.4-nano 分别仅为 0.10 美元和 0.625 美元/百万 Token。模型分层后，使用者不必在所有环节都调用最高价模型，而可以通过模型路由，将复杂推理、代码生成、长链路规划等高难度任务交给旗舰模型，将分类、抽取、改写、简单问答、结构化生成等标准化或重复性环节交给轻量模型处理，从而在保证效果的同时显著降低整体调用成本。

此外，近期国内模型厂商的 API 调价也体现出成本效率优化和调用门槛下降的趋势。2026 年 5 月，DeepSeek 宣布 DeepSeek-V4-Pro 在限时优惠结束后将永久调整为原定价的 1/4；随后小米 MiMo-V2.5 系列也宣布 API 永久降价，最高降幅达 99%，并取消按上下文窗口长度区分定价的方式，同等付费价格下 Token 用量可提升至原来的 5–8 倍。随着模型效率、推理工程和国产算力生态持续优化，国产模型有望以更

下载日志已记录，仅供内部参考

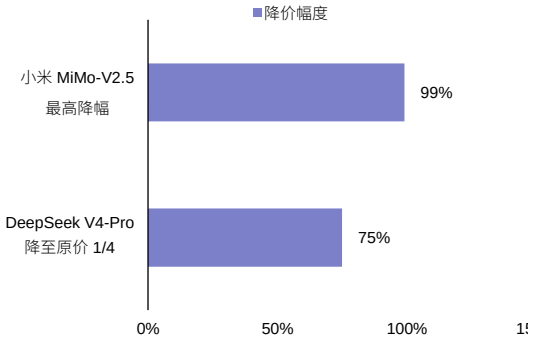
低成本承接高频、标准化和多轮调用任务，进一步降低应用侧调用门槛。

图 16：模型分层价格（单位：美元/百万 tokens）



数据来源：Google 官网、OpenAI 官网、广发证券发展研究中心

图 17：国内模型 API 永久降价



数据来源：小米官方微博、DeepSeek 官网、广发证券发展研究中心

（二）Token 调用量：模型能力是否被真实使用？

Token 调用量是观察模型能力从技术指标走向真实应用、并进一步转化为商业收入的重要中间变量。对于模型厂商而言，只有当模型能力从展示走向真实调用，才有可能进一步沉淀为用户粘性，并转化为持续收入增长。海外头部厂商方面，OpenAI 披露其 API 每分钟处理量已突破 150 亿 tokens，Google 则披露其 API 每分钟处理约 190 亿 tokens，表明模型调用正进入高频需求阶段。但需注意的是，Token 调用量本身并不等同于收入质量，仍需结合调用价格、免费流量占比、内部调用占比、缓存命中率、推理成本等指标综合判断。

1．全球：以 OpenRouter 等第三方 API 聚合平台为窗口观察开发者/API 生态

API 调用情况能够相对直接地反映模型是否被嵌入真实应用和工作流。OpenRouter 作为第三方模型聚合平台，接入 OpenAI、Anthropic、Google、DeepSeek、Qwen 等多家模型厂商，开发者可以在同一平台上选择不同模型，因此其周度 Token 数据适合观察全球开发者/API 生态中的模型偏好、价格敏感度和调用迁移情况。其优势在于频率高、口径相对可比，能够捕捉模型发布、能力提升或价格调整后的边际变化；但局限也较明显，即不覆盖头部厂商自有应用、官方 API、企业私有化和本地部署等完整流量。因此，OpenRouter 更适合用于观察第三方开发者生态中的真实调用变化，而不宜简单等同于各模型的全渠道调用规模或市场份额。

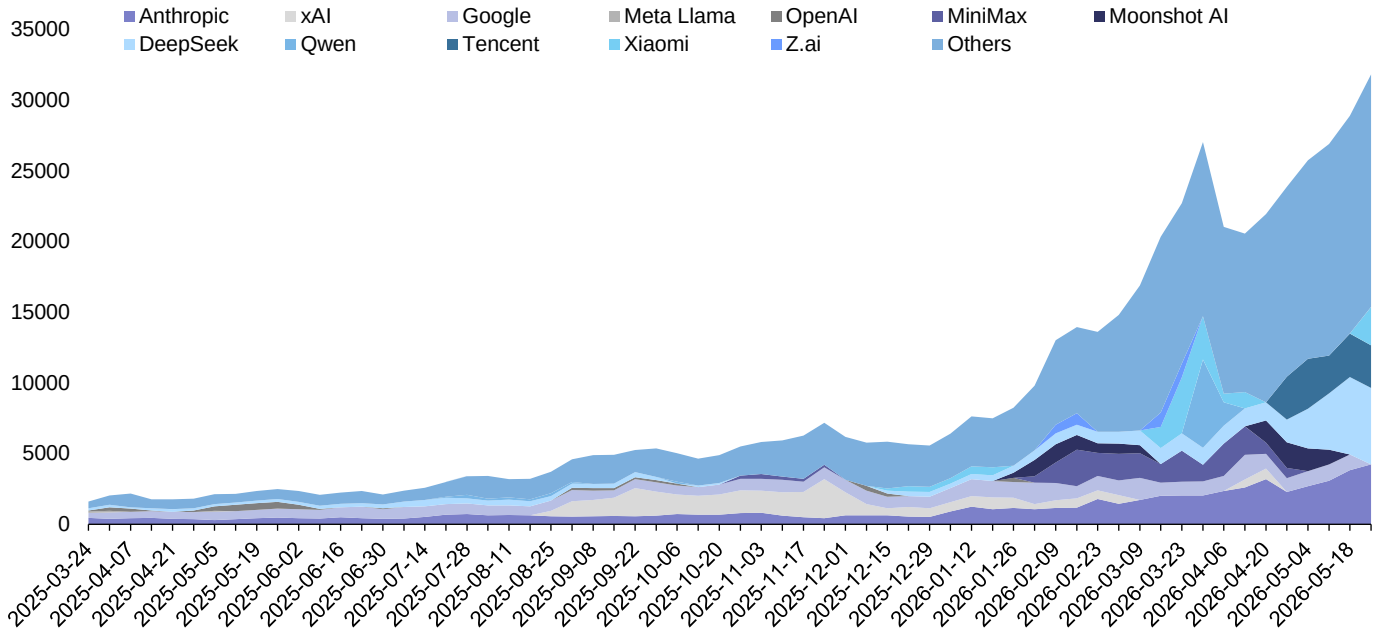
过去一年中，OpenRouter **平台周度 Token 调用量已经实现数量级增长。**2025 年 5 月平台周度调用量均值约为 2.29T tokens，2026 年 5 月已提升至 28.3T，规模达去年同期十倍以上。

从厂商结构看，OpenRouter **调用结构正在由海外头部集中走向多元化。**2025 年 5 月，平台调用主要集中在 Google、OpenAI、Anthropic，三者合计占比超过 60%；到 2026 年 5 月，Google、OpenAI、Anthropic 合计调用占比降至约 15%。OpenRouter 数据作为第三方 API 平台样本，虽然不能直接等同于全球模型市场份额，但能够较好反映开发者生态中模型调用迁移和多模型共存的趋势。在这一变化中，国内模型已成为调用增

下载日志已记录，仅供内部参考

长的重要力量：DeepSeek、Kimi、Qwen、GLM、MiniMax 等模型陆续进入高调用区间，说明国产模型不仅在能力上持续追赶，也开始在全球开发者 API 生态中获得真实调用。

图 18：OpenRouter 周度 Token 调用（截至 2026 年 5 月 31 日，按厂商分类）（单位：十亿）



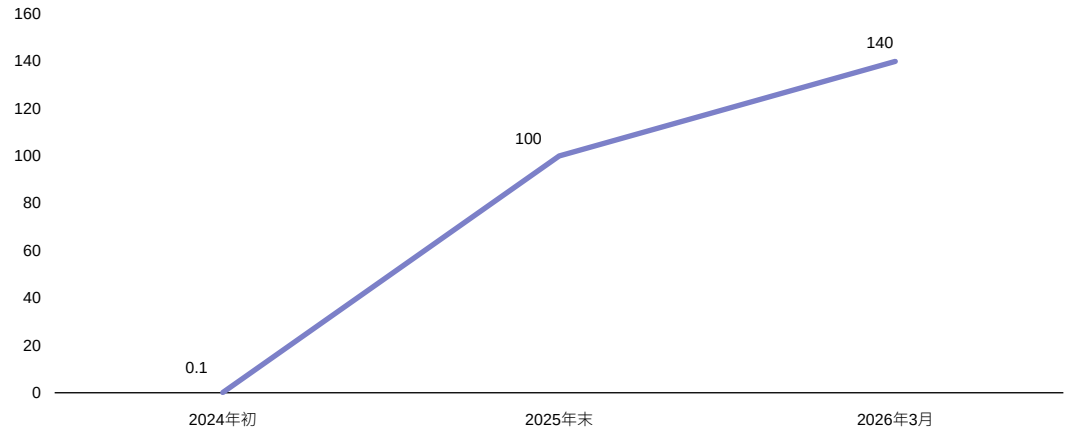
数据来源：OpenRouter、广发证券发展研究中心

2．中国：日均 Token 与典型应用数据反映国内推理需求提升

国内大模型推理需求已进入快速放量阶段，最直接的验证来自全国日均 Token 调用量的持续攀升。根据国家数据局披露，2025 年我国全年 Token 调用量约为 21,100 万亿。从日度调用频次看，2024 年初我国日均 Token 调用量约为 1,000 亿，至 2025 年底已提升至 100 万亿，并在 2026 年 3 月进一步突破 140 万亿。两年时间内，日均 Token 调用量增长超过千倍；仅 2026 年一季度，较 2025 年底的增幅即超过 40%，显示出国内大模型推理需求正在加速释放。

下载日志已记录，仅供内部参考

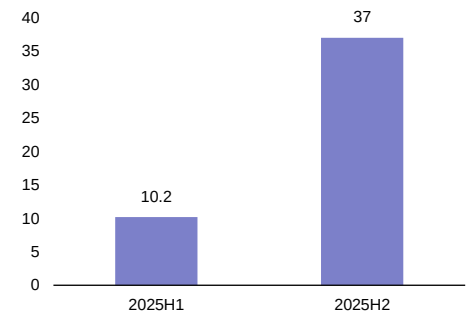
图 19：中国日均词元调用量（单位：万亿）



数据来源：国家数据局、广发证券发展研究中心

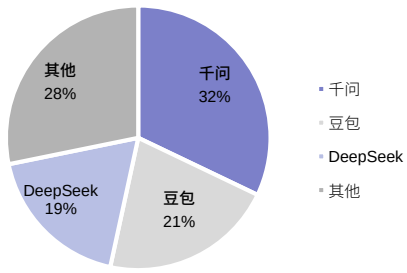
企业级调用是国内 Token 增长的重要来源。沙利文数据显示，2025 年上半年，中国企业级大模型日均调用量已超过 10 万亿 tokens，较 2024 年下半年增长 363%；2025 年下半年进一步提升至 37.0 万亿 tokens，较上半年增长 263%，企业侧推理需求呈现加速释放态势。分厂商看，2025 年下半年中国企业级大模型日均 tokens 消耗前三名分别为千问、豆包和 DeepSeek，占比分别为 32.1%、21.3% 和 18.4%，合计占比超过七成，头部模型在企业级调用市场中的集中度进一步凸显。

图 20：中国企业日均词元调用量（单位：万亿）



数据来源：沙利文、头豹研究院、广发证券发展研究中心

图 21：2025H2 中国企业级大模型调用市场

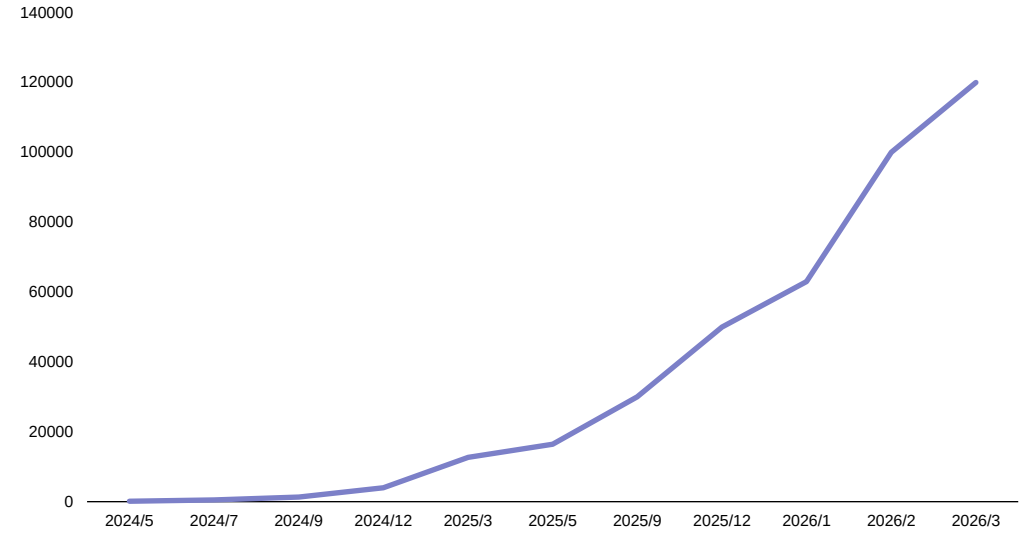


数据来源：沙利文、头豹研究院、广发证券发展研究中心

从具体厂商表现看，豆包 Token 使用量的快速增长，进一步提供了微观数据验证。豆包大模型日均 Token 使用量在 2026 年 3 月已突破 120 万亿，较年初水平继续翻倍，较 2024 年 5 月首次发布时增长约 1,000 倍。同期，火山引擎上累计 Token 使用量超过一万亿的企业客户数量，也由 2025 年底的 100 家增至 140 家。以上数据表明了豆包 Token 调用增长并非仅来自单一 C 端入口，也与企业侧和平台侧调用扩张密切相关，反映出应用侧、企业侧与云平台侧需求共同推动模型调用量持续提升。

下载日志已记录，仅供内部参考

图 22：豆包日均词元调用量（单位：十亿）



数据来源：火山引擎、新华网、证券时报、广发证券发展研究中心

总结来看，伴随模型能力提升和推理成本下降，国内大模型推理需求逐步进入规模化调用阶段。Token 放量已成为观察国内 AI 应用渗透和模型商业化兑现的重要指标，全国日均 Token 调用量持续攀升，验证了需求总量的快速扩张；企业级调用量高增，则表明大模型正在加速向生产环境和业务系统渗透。豆包等典型厂商 Token 使用量快速增长，进一步说明当前调用扩张已不再依赖单一聊天入口，而是由应用嵌入、企业生产系统接入和平台型 API 调用共同推动。后续随着模型路由、Agent 工作流和多模态应用进一步落地，国内大模型调用量仍有望保持较快增长。

（三）ARR 兑现：模型使用能否兑现为商业化收入？

Token 调用量说明模型能力正在被实际使用，而 ARR 则进一步验证这种使用是否能够沉淀为持续收入，体现客户付费意愿、留存复购能力和商业化质量。当前，收入兑现率先出现在两类主体中：一类是 OpenAI、Anthropic、智谱、MiniMax 等模型公司，通过 C 端订阅、API 调用和企业客户形成公司级 ARR；另一类是 Cursor 等 AI 原生应用，以及 Claude Code、Codex 等模型厂商自有代码/Agent 产品，将模型能力嵌入具体工作流，率先在代码、Agent 等生产力场景中实现商业化。

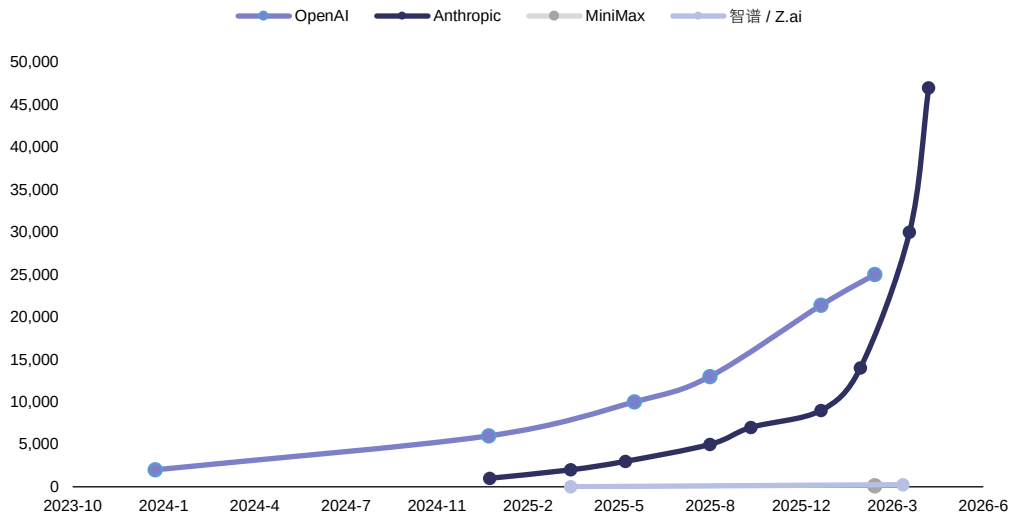
此外，近期部分模型厂商降价，说明国内大模型调用价格仍处在动态调整阶段；而金山云上调 AI 算力及文件存储等相关产品服务价格，则反映了另一层矛盾：底层高质量算力、存储资源、稳定云服务和企业级交付能力仍然稀缺。模型价格下降有助于降低应用侧使用门槛，推动 Token 放量和场景渗透，但也意味着单纯依赖模型调用收费获取高利润的商业模式将面临压力；相反，具备稳定算力供给、推理成本优化、企业客户交付、私有化部署和应用闭环能力的厂商，更有可能在价格竞争中维持商业化韧性。后续判断行业空间，不能只看模型能力和调用量增长，还要看降价能否带来足够大的需求弹性，以及厂商能否通过算力效率、客户粘性、应用入口和增值服务来承接。

1．模型端：海外头部已率先兑现，国内厂商 ARR 进入快速上升期

下载日志已记录，仅供内部参考

ARR / 年化收入是观察大模型商业化兑现的重要指标。相比 Token 调用量更侧重反映真实使用强度，ARR 进一步验证用户是否愿意为模型能力持续付费，以及模型调用能否沉淀为订阅、API、企业席位或平台服务收入。当前，OpenAI、Anthropic 等海外头部厂商率先形成大规模年化收入，国内模型公司则仍处于从项目制收入、应用收入向 API 调用、MaaS 平台和企业订阅等可持续收入模式迁移的验证阶段，ARR 正在快速上升。

图 23：典型模型厂商 ARR 收入（单位：百万美元）



数据来源：OpenAI、Anthropic、MiniMax、Financial Times、Business Times、证券时报、广发证券发展研究中心

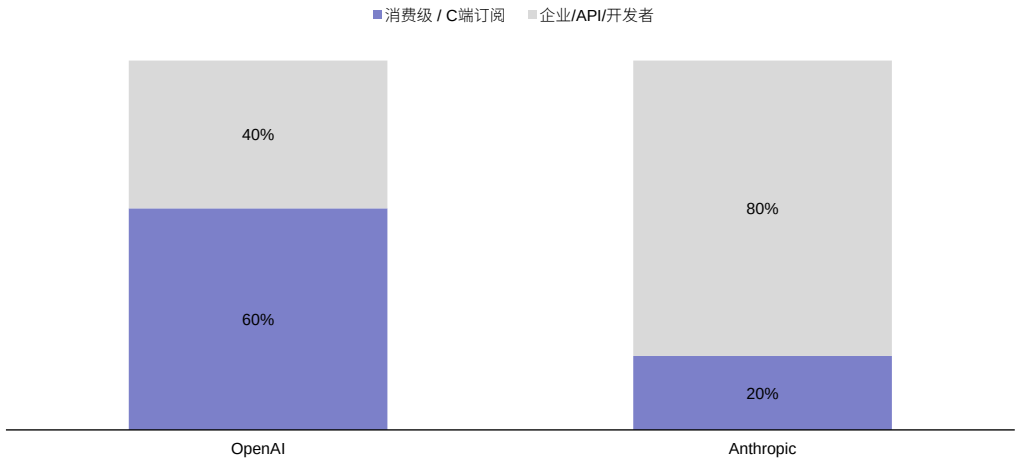
注：智谱披露口径为开放平台及 API 业务 ARR，并非全公司

OpenAI 的收入增长首先来自 ChatGPT 建立的高频用户入口，企业级市场的增长势头同样较强。ChatGPT 形成了全球级用户入口和付费心智，API 与企业版产品进一步承接开发者和企业客户需求，使模型能力提升能够较直接地转化为订阅、API 调用和企业服务收入。企业级市场目前已贡献了公司 40% 以上的营收，并有望在 2026 年底前实现与消费级业务的并驾齐驱。根据公司披露，OpenAI 年化收入已从 2023 年约 20 亿美元提升至 2024 年约 60 亿美元、2025 年超过 200 亿美元，并在 2026 年继续增长；Business Times 报道称，截至 2026 年 2 月底，OpenAI 年化收入已超过 250 亿美元。

Anthropic 的收入更集中在企业客户、开发者和代码工作流。Claude 模型在复杂推理、长文本处理和代码任务中形成了较强用户心智，Claude Code 的快速增长进一步强化了其在 AI Coding 场景中的商业化能力。根据证券时报报道，Anthropic 年化收入在 2025 年 10 月已接近 70 亿美元；而在 2026 年 5 月，公司年化收入已达 470 亿美元，主要由企业产品、API 调用和 Claude Code 等开发者工具拉动。相比 OpenAI 更强的 C 端入口，Anthropic 的收入结构更偏企业客户、API 和开发者工作流，专业场景粘性更强。

下载日志已记录，仅供内部参考

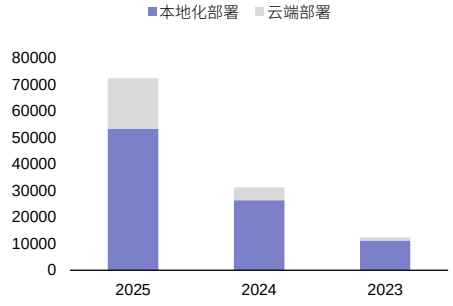
图 24：OpenAI 与 Anthropic 最新营收结构



数据来源：OpenAI、福布斯官网、广发证券发展研究中心

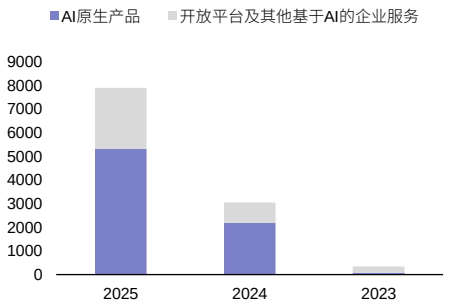
国内厂商中，智谱和 MiniMax 正在形成不同的 ARR 验证路径。智谱更偏 MaaS / API 和企业服务，其 MaaS API 平台 ARR 已达到约 17 亿元人民币，过去 12 个月提升约 60 倍，反映国内企业客户对模型调用、私有化部署和行业 Agent 的需求正在形成可持续收入。MiniMax 则更偏 AI 原生应用、多模态产品和开放平台并行推进，C 端产品贡献用户入口和付费转化，API 业务 ARR 也逐步提升，带动公司级 ARR 在 2026 年初超过 1.5 亿美元。整体来看，智谱路径更偏企业级 MaaS/API 和行业服务，MiniMax 路径则更偏 C 端应用、开放平台与多模态能力的协同变现。

图 25：2023-2025 年智谱营收结构（单位：万元）



数据来源：iFind、广发证券发展研究中心

图 26：2023-2025 年 MiniMax 营收结构（单位：万美元）



数据来源：iFind、广发证券发展研究中心

整体来看，国内模型公司与海外头部相比仍有明显规模差距，但在中文场景、低成本推理、开源生态和本土企业需求上具备差异化机会。随着 Token 调用量持续放大，国内厂商能否将使用强度进一步转化为稳定向上的 ARR 将成为验证模型商业化质量的关键。

2．应用端：从 AI Coding 到办公与多模态，高价值工作流场景率先兑现

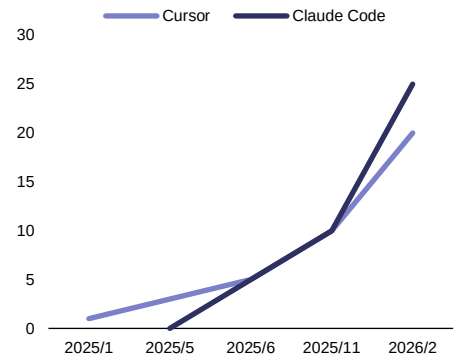
下载日志已记录，仅供内部参考

应用端商业化正在向高频工作流入口扩散，AI Coding 是首个跑出规模化 ARR 的场景。与通用聊天、内容生成等场景相比，代码任务的结果更容易验证，用户能够直接感知效率提升，企业也更容易通过研发效率、交付周期和工程质量衡量 ROI。此外，AI Coding 工具既是应用层收入入口，也是底层模型调用量增长的重要放大器，因为传统代码补全更多停留在单点辅助，而 Agent 化开发会将一次需求拆解为任务规划、代码生成、文件读写、调试、测试和提交等连续步骤，显著放大了 Token 消耗。

Claude Code 和 Cursor 是当前 AI Coding 商业化的典型案例。Anthropic 披露，Claude Code 年化收入已超过 25 亿美元，而公司 2026 年 2 月年化收入为 140 亿美元，Claude Code 占比为约 18%；同时，企业使用贡献了 Claude Code 超过一半的收入，说明代码 Agent 已经从个人开发者工具进入企业软件工程流程。Cursor 则代表独立 AI IDE 路线，据 Bloomberg 报道，其 2026 年 2 月年化收入已超过 20 亿美元，并在三个月内实现翻倍，其中企业客户贡献约 60%，表明 AI Coding 的商业化正在从个人订阅进一步扩展至企业席位和团队级采购。

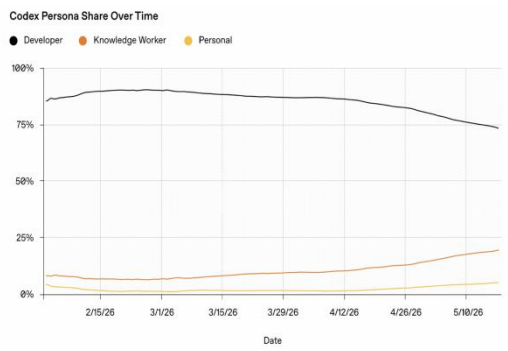
Codex 则体现了 OpenAI 对开发者和 Agent 工作流入口的强化，且覆盖场景正从软件开发向知识工作扩展。2026 年以来，Codex 用户规模快速放量：3 月底周活用户已超过 200 万，过去三个月增长 5 倍，使用量月环比增速保持在 70% 以上；4 月初，Codex 周活跃用户进一步提升至 300 万；OpenAI 后续披露称，4 月下旬每周使用开发者已超过 400 万。到 2026 年 6 月，OpenAI 最新报告进一步披露，Codex 周活跃用户已超过 500 万；除开发者外，知识工作者已贡献约 20% 用户，且增速超过开发者的 3 倍。由此看，Codex 正在从代码生成、测试、代码审查等软件工程场景，扩展至文档处理、数据分析、邮件、跨应用协作等办公流程，成为 OpenAI 切入企业软件工程、内部数据环境和知识工作流的新入口。

图 27：Claude Code 与 Cursor 年化收入增长情况（单位：亿美元）



数据来源：Anthropic、Cursor、广发证券发展研究中心等

图 28：Codex 用户画像结构变化



数据来源：OpenAI、广发证券发展研究中心

办公和多模态等泛生产力场景有望接力成为代码之后重要的应用落地方向。办公场景覆盖文档、表格、会议、邮件、知识库和企业协同，用户基数更大、使用频次更高，且天然嵌入企业日常生产流程。以 Microsoft 365 Copilot 为例，微软披露其已拥有超过 2000 万个付费席位，90% 的《财富》500 强企业正在使用 Microsoft 365 Copilot，说明 AI 办公已逐步进入企业级规模部署阶段；国内亦已有 WPS AI 等办公产品形成较大用户基础，截至 2025 年末 WPS AI 国内月活用户超过 8013 万，同比增长 307%，WPS Office 全球月活设备

下载日志已记录，仅供内部参考

数达 6.78 亿，亦能体现办公软件作为 AI 入口的高渗透潜力。

（数据来源：Microsoft 官网，金山办公《2025 年年度报告》）

多模态场景则将模型能力从文本扩展至图像、语音、视频和实时交互，有望在内容生产、营销创意、教育培训、客服、陪伴和生产力工具中打开新的应用空间。视频生成已率先展现商业化弹性：快手公司披露，其旗下可灵 AI 2026 年一季度收入超过 6.5 亿元人民币，同比增长超 300%，2026 年 3 月 ARR 已接近 5 亿美元，较 2025 年 3 月的 1 亿美元实现约 4 倍增长。另外，据《智能涌现》，字节跳动 Seedance 2.0 单模型月收入已超过 10 亿元人民币，并推动火山引擎将 2026 年 MaaS 收入目标上调至 150 亿元。

下载日志已记录，仅供内部参考

四、AI 应用：大模型对各不同应用企业的影响因领域而异，具备强私域数据、基础通用工具类专业厂商、以及卡位应用落地“最后一公里”的软件咨询实施商或脱颖而出

投资要点：无需非公开复杂数据、技术门槛较低、容错空间较大、功能性标准化程度高的软件公司将继续受到侵蚀和冲击，但具备强私域数据、基础通用工具类专业厂商如 Coding、database 等、以及卡位应用落地“最后一公里”的软件咨询实施商或脱颖而出。从这三类厂商出发，我们建议关注卓易信息、赛意信息、汉得信息、中控技术、中望软件、晶泰控股、税友股份、金蝶国际等。

（一）大模型对各个不同应用企业的影响因领域而异

无需非公开复杂数据、技术门槛较低、容错空间较大、功能性标准化程度高的软件公司将继续受到侵蚀和冲击。在 AI 辅助编程功能持续提升的趋势下，软件开发效率持续提升，软件开发门槛在降低，对于软件行业有一定影响。但软件产品根据功能复杂程度、应用场景以及行业领域不同，受 AI 大模型影响的程度各不相同。当前 AI 大模型对于功能标准化程度较高、技术门槛低的软件影响较大，具体来看：

① **对于逻辑简单、场景单一的软件，AI 大模型可以通过简单指令即可完成开发任务：**例如，智谱推出的 Z Code 产品，可实现通过自然语言指令驱动 AI Agent 完成从编码、调试到预览的全流程开发任务。在此基础上，在给 AI 大模型编写提示词的过程中，用户可以摆脱对于复杂算法的描述，而直接描述软件功能，即可完成开发工作。在 Z Code 中，用户输入提示词“给我开发一个贪食蛇的小游戏”，Z Code 即可自动实现对于前端软件界面的设计，后端数据组织方式的设计，并将整体开发工程分为若干个模块，再针对每一个模块来生成代码。

② **涉及复杂流程、模块较多、用户需求更加细节的大型软件项目，AI 大模型需要结合行业模板来完成开发任务。**大型软件开发项目较难实现一个提示词即实现整体大型软件开发流程的自动化。部分 AI 代码辅助工具产品往往针对常用场景，设置好固定软件开发功能的模板，供用户调用。以 EazyDevelop 产品中预置的科亿知识库模板为例，其提供了知识管理、权限管理、知识库配置等 8 大功能，并基于此提供了超过 30 个更为细分的功能模块。在此基础上，用户的定制化需求以提示词的形式向 AI 大模型提出，即可实现知识库软件的快速开发。卓易信息的 EasyDevelop 产品针对金融风控、零售仓储、制造业设备巡检等场景提供了超过 20 个行业模板。（数据来源：卓易信息 2025 年半年度报告）将常用软件的功能模块预置在 AI 辅助编程工具中大大提升了 EasyDevelop 针对不同行业大型软件的开发效率，有效增强其差异化竞争力。

尽管 AI 大模型 agent 能力的提升，对于功能标准化程度较高、技术门槛低的软件影响较大，但长期来看，部分具有行业壁垒和特定环节的软件公司仍有发展前景。具体而言，有以下三类厂商：

1. **特定领域的数据专业性强，且非公开、非通用，相关公司如果能跟得上 AI 的发展，仍将会在未来有生存发展空间。**

不同应用场景的技术栈、开发标准及流程规范存在显著差异。AI 大模型需要在深刻理解场景的基础上，才能自动筛选合适的开发工具、优化开发流程。反之，缺乏场景理解的 AI 大模型，其自动化开发功能只会流于形

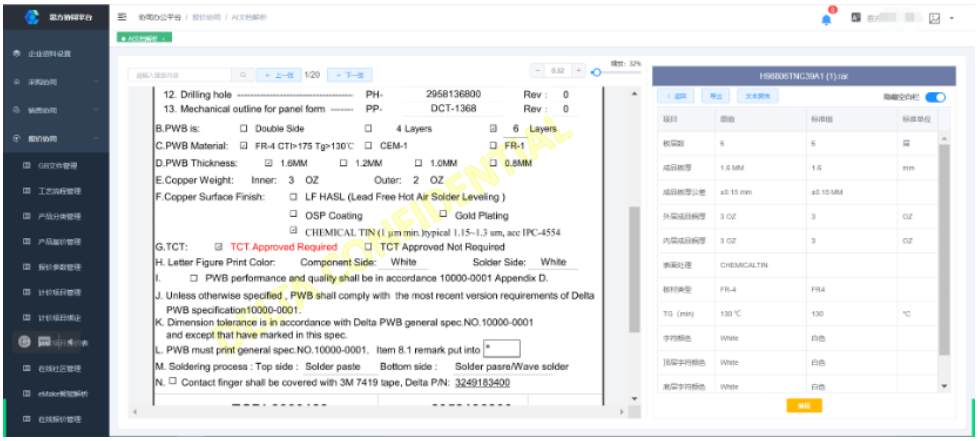
下载日志已记录，仅供内部参考

式，无法真正贴合实际开发需求，难以发挥其在提升开发效率、降低开发门槛、减少人为失误等方面的核心价值，最终无法实现程序开发流程的真正自动化。

尤其是各垂直领域的数据往往沉淀了行业专属的业务逻辑、经验规则等，并非通用 AI 模型可轻易获取和解析，比如财税领域的不同地方的底层财税长文本数据、医疗领域的病例隐私数据和诊疗经验数据，工业领域的生产参数数据、设备运维数据等。而深耕各垂直领域的软件公司，是深刻理解相关数据和业务流程的，其将自身掌握的行业 Know-how 和专属数据与 AI 技术深度融合，打造出的 AI 解决方案具备很强的差异化竞争力。

例如，赛意信息开发的 PCB 行业大模型能准确识别和提取 PCB 工程图纸中的关键的工程参数，如尺寸、材料、层数等，将原本需要 4~6 小时的参数提取工作缩短至几分钟，同时确保了 95%以上的准确率，大大提升了从订单接收到生产准备的效率。因此，具备行业壁垒，理解垂直领域业务流程的公司在 AI 赋能行业升级的浪潮中，仍具有获得较好的发展前景。

图 29：赛意信息思方云协同平台“智能参数提取”功能界面



数据来源：赛意信息官方微信公众号，广发证券发展研究中心

2. **基础通用工具类专业厂商如 Coding、database、IDE 开发工具厂商等所提供的底层基础工具是 AI 软件开发生态的核心基石，在 AI 新时代具备长期成长红利。**

从产业链底层逻辑来看，AI 大模型仅为软件开发提供智能化能力加持，暂时无法独立完成全链路软件研发工作。而编码开发、数据库管理、程序编译部署、项目运维等底层基础工具，是连接 AI 模型与终端软件研发场景的关键载体，所有 AI 辅助开发功能均需依托 IDE、数据库、代码托管平台等基础工具落地。与此同时，通用大模型本身的训练、迭代、部署与运维，同样离不开数据库、开发框架、代码管理工具等配套基础设施，这也让基础通用工具厂商跳出低端标准化软件的内卷赛道，免受 AI 技术的直接替代冲击。

当前国内 AI 代码辅助工具持续迭代，用户生态快速拓展。以 Cursor 和 GitHub Copilot 为代表的 AI 代码辅助工具商业化取得较好效果的情况下，国内字节跳动、阿里巴巴等厂商也分别开发了 TRAE 和通义灵码等产品，其功能也逐渐由代码生成、补全、调试等基础功能，向 AI 自主完成从需求分析、算法设计、程序部署的全流

下载日志已记录，仅供内部参考

程自动化方向发展。根据阿里云官网，一汽集团、蔚来汽车、中华财险等上万家企业已接入通义灵码商业化拓展持续发力。

EazyDevelop 产品自 25 年 9 月上线以来，商业化落地进展超预期。根据卓易信息官方微信公众号，截至 2025 年 12 月 31 日，EazyDevelop 平台用户已突破 1.3 万，订单金额突破 1800 万元。进入 2026 年，商业化迎来显著提速：2026 年第一季度，EazyDevelop 进一步实现收入 2583.7 万元，占公司当季总营收的 30%，单日 Token 消耗量突破 200 亿，同时累计订单金额已突破 8200 万元，其中 4 月单月订单超过 3200 万元，呈现量价齐升态势。EazyDevelop 产品订单金额的快速增长不仅反映了在 AI 赋能之下 IDE 工具给企业级客户带来降本增效效果显著，从而拉动客户付费意愿有明显提升，还体现了 EazyDevelop 产品在 AI 辅助编程工具领域的差异化竞争优势，具体而言：

（1）独立第三方的中立性具有独特价值：EasyDevelop 产品接入了包括 GLM、DeepSeek、Qwen、Kimi 等多个 AI 大模型，供用户自由选择。EasyDevelop 产品提供的较为丰富的模型选择推动其用户覆盖面快速的拓展。在各模型厂商的代码编程能力处于激烈竞争的状态下，作为第三方 IDE 公司，卓易信息较好的享受了 AI 大模型快速迭代的技术红利，在生态拓展方面具有一定优势。

（2）全流程开发经验赋能的 Multi-Agent 架构优势：公司基于对软件开发全流程的理解，针对需求调研、架构设计、测试运维等多个环节，设计了 Multi-Agent 架构，可实现各智能体间的协作既能分工明确，又能无缝衔接，是 EasyDevelop 产品高效赋能软件开发的底层技术优势。

（3）信创领域先发优势构建的自主可控壁垒：公司正在积极推进 EasyDevelop 对仓颉语言、鸿蒙原生开发的支持。公司推动 EasyDevelop 产品与国产芯片、操作系统、数据库等信创产业链关键环节深度协同，并已具备一定的客户基础，在信创相关客户的商业拓展上具有一定优势。

（4）深入垂直领域，赋能不同领域专业软件开发的行业模板：根据卓易信息 2025 年年报，EazyDevelop 以“AI+IDE”为基底，聚焦 AI 编程、多智能体协作与国产化适配，内置 300 余种行业场景模板，覆盖政务、金融、教育、物联网等多个领域。卓易信息将常用软件的功能模块预置在 AI 辅助编程工具中大大提升了 EasyDevelop 针对不同行业大型软件的开发效率，有效增强其差异化竞争力。

表 4：EazyDevelop 产品科亿知识库模板的软件架构

主要功能	细分功能模块
检索首页	普通检索、高级检索、精准检索、结果中二次检索、文件预览、知识版本控制、知识详情、目录分类浏览、个人工作区
个人中心	我的知识、我的收藏、我的分享、我的消息、修改密码
知识管理	知识创建、知识编辑、待审核文件、已审核文件、工作流代办、文件统计
知识库配置	系统参数、分类标签定义、密级定义、目录分类定义、文件脱敏规则、轮播图管理
知识权限管理	附加授权设置、密级放行设置
系统管理	用户管理、角色管理、菜单管理、权限设置、表单权限、部门管理

数据来源：EazyDevelop 官网，广发证券发展研究中心

、

下载日志已记录，仅供内部参考

3. **即使在 AI 试图主导的新软件时代生态中，数据增值服务商和咨询实施开发集成商仍是不可或缺的产业链价值环节所在。**

由于企业软件并不是简单的技术部署，而是涉及业务流程重构、组织协同和管理方式调整的系统工程，而不同企业在行业背景、发展阶段、管理成熟度上存在较大差异，因此 通用 AI 或标准化软件难以“一键适配”各项真实的业务场景。

在 AI 主导的新软件生态中，数据增值服务商和咨询实施集成商承担着将企业业务语言转化为系统逻辑的关键角色，这一职能并不会因为 AI 的进步而消失，反而会被重新放大其价值。一方面，AI 需要高质量、强语义的数据才能发挥效果，数据清洗、建模和指标体系设计仍然高度依赖行业经验与业务理解；另一方面，企业为系统结果和业务效果负责，而非为模型本身负责，因此需要有人能够对系统整体方案、实施路径和最终成效负责。

从 Anthropic 与 Infosys 深度合作的角度出发，也进一步证明了数据增值服务商和咨询实施开发集成商的价值潜力与机会。Anthropic 于 2026 年 2 月 17 日发布公告，其与 Infosys 宣布达成合作，将为电信、金融服务、制造及软件开发领域开发并交付企业级 AI 解决方案。早在 25 年 7 月 Anthropic 推出金融分析解决方案（Financial Analysis Solution）时就已经和两类公司（数据服务商和软件咨询实施商）有所合作，而今合作的趋势变得更加紧密。金融分析解决方案能够实现三种功能，包括数据整合与验证功能、金融分析与建模自动化、流程自动化与系统升级。而实现这三种功能的基础是建立在和核心的数据服务商以及专业的软件服务咨询商合作的基础上来进行实现的：

① 通过和数据服务商如 Daloopa、FactSet、Morningstar、Palantir、PitchBook 和 S&P Global 等来提供高质量、可交叉验证的实时数据，显著降低因单一信息来源导致的分析错误风险，并提升金融决策的可靠性。

② 同时，德勤、毕马威和普华永道等软件服务咨询商能够通过其在数字化转型、风险管理、审计与合规方面的专业能力，帮助金融机构在引入 AI 技术的同时，满足监管、安全与治理要求。

图 30：Anthropic 公司的金融分析解决方案的生态合作伙伴



数据来源：Tribe.ai，广发证券发展研究中心

下载日志已记录，仅供内部参考

表 5：Anthropic 在金融分析解决方案中与生态合作伙伴的合作方式梳理

	类型	典型公司	所实现效益
Anthropic 在金融分析解决方案中的合作伙伴	数据服务商	Daloopa FactSet PitchBook 等	提供高质量、可交叉验证的实时数据，显著降低因单一信息来源导致的分析错误风险，并提升金融决策的可靠性
	软件咨询服务商	德勤 毕马威 普华永道	具备深厚的咨询、风险管理与财务审计领域经验,其专业支持能帮助金融机构应对 AI 落地过程中的复杂问题，为系统整合与合规流程提供高价值洞察。同时，也能够遵循严苛的审计追踪要求，进而提升数据管理策略的可信度与透明度。

数据来源：Opentools 官网，广发证券发展研究中心

这说明软件咨询服务商正成为技术转型的推动者，凭借其在特定领域的专业能力和壁垒，帮助通用 AI 在相关特定性领域中取得更好的落地。这一行业趋势正逐步清晰。通俗的来说，就是通过软件咨询服务商的咨询、风险管理、财务审计等方面的经验来实现 AI 在特定行业落地的同时符合合规、安全、监管等一系列的要求。

具体来看，这次的合作是将 Anthropic 的 Claude 大模型、Claude 代码工具与 Infosys 的 Topaz 平台进行深度融合，主要落地的应用场景包括电信行业、金融服务业、制造业、软件开发业以及企业运营领域。根据 Anthropic 官网，之所以选择 Infosys 作为合作伙伴，有两点原因，一是 Infosys 总部在印度，而印度是 Claude.AI 在全球范围内的第二大市场，和 Infosys 的合作有助于 Anthropic 快速打开印度市场，二是在 demo 中表现较好的 AI 大模型与在受监管行业中实际落地的模型相比仍存在较大差距，而 Infosys 在电信、金融服务、制造等这类受到强监管且数据高度敏感的行业拥有丰富的经验壁垒，能够助力 Anthropic 的 AI 应用快速落地的同时，符合当地政策，做到合法合规。

表 6：Anthropic 和 Infosys 合作中的应用场景梳理

	应用场景	AI 所能够实现的赋能
Anthropic 和 Infosys 的合作	电信行业	AI Agent 助力运营商实现网络运营现代化，简化客户生命周期管理，提升服务交付质量
	金融服务行业	AI Agent 将帮助金融机构更快地识别和评估风险，实现合规报告自动化，同时提供更个性化的客户交互体验，例如根据客户的完整账户历史与市场行情定制专属金融建议
	制造与工程行业	Claude 大模型将助力企业加快产品设计与仿真验证进程，缩短研发周期，让工程师在产品量产前能完成更多次的设计迭代与测试
	软件开发行业	开发团队可借助 Claude 代码工具完成代码的编写、测试与调试，推动开发流程从设计阶段到量产阶段的高效落地
	企业运营领域	Claude 智能协作助手（Claude Cowork）将帮助各团队实现文档摘要、状态汇报、审核流程等日常工作的自动化处理

数据来源：Anthropic 官网，广发证券发展研究中心

因此，我们认为，AI 更多是降低交付成本和提升效率，而不是消灭数据服务和咨询实施在产业链中的地位，在复杂企业环境中，数据增值与咨询实施等专业服务仍然是成功的关键一环。

伴随软件咨询服务商成为新技术转型的推动者的这一行业趋势逐渐清晰，在特定行业有专业壁垒的软件咨询实施商有望和模型公司一同提升，迎来属于自己的“蜕变”。

下载日志已记录，仅供内部参考

（二）以上三种领域的典型代表公司

1．晶泰控股

晶泰控股，AI4S 领航者，平台进入规模化兑现阶段。(1)2026 年 6 月 9 日，晶泰控股公告拟于公开市场进行股份回购，回购金额最高达 1 亿美元，回购一方面体现公司对自身长期业务发展及资本市场价值的信心，另一方面公司现有财务资源足以支持该计划实施，同时维持稳健财务状况。(2)公司业绩增速强劲，商业化能力持续验证。2025 年公司收入同比增长 201.2%至 8.03 亿元，并首次实现全年盈利。(3)与 DoveTree 的合作继此前收到首付款后，已进一步收到第二笔付款 1900 万美元，验证公司 AI 制药平台在外部合作中的持续兑现能力。(4)公司三位创始人自愿延长股份禁售期，叠加本次最高 1 亿美元回购计划，体现核心团队对公司中长期价值的坚定信心。(5)公司口服降糖多肽已获得美国 Self-GRAS 认证，即将进入美国食品及膳食补充剂市场。继此前生发固发分子取得进展后，本次口服降糖多肽的推进，是公司在消费健康领域的产品化和商业化能力进一步验证。

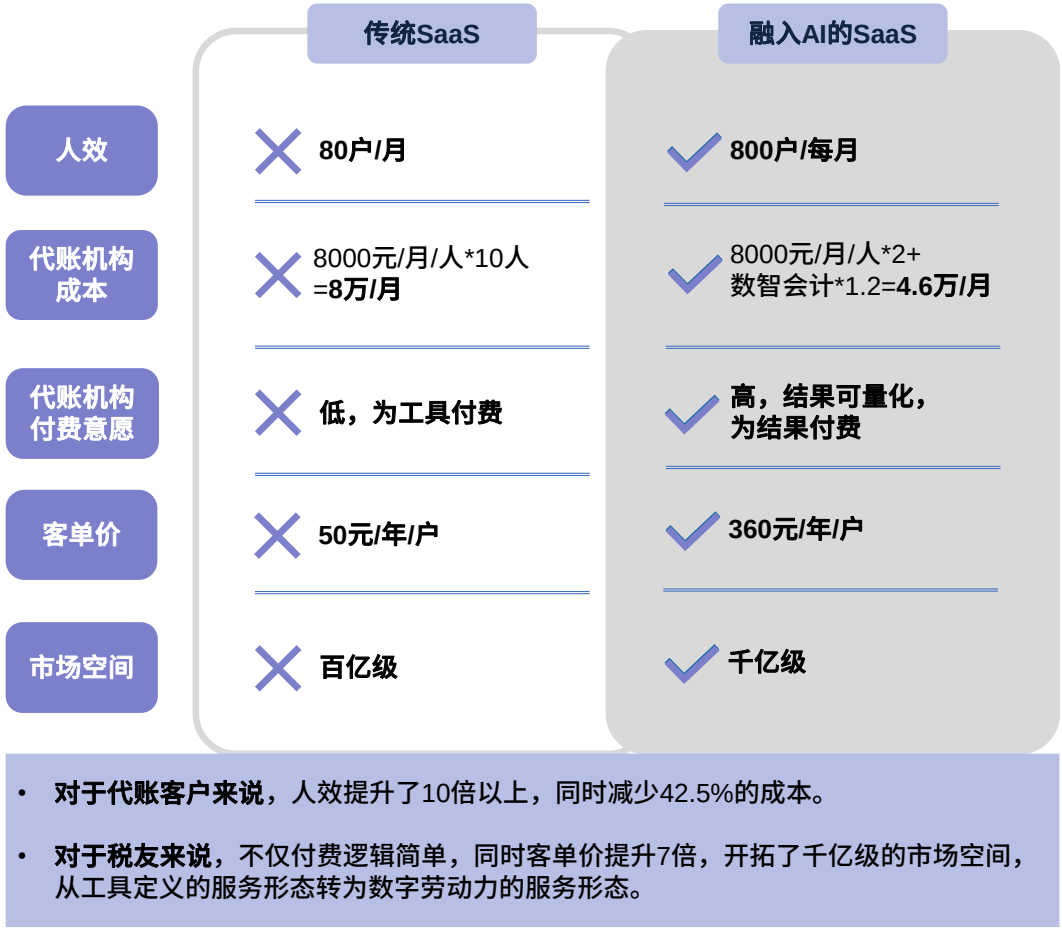
2．税友股份

AI 加入大幅提高了财税处理效率。通过人机协同，AI Agent 能够精简大量的繁琐重复性工作，显著降低企业财税人员的工作量。同时，企业只需要购买附带 AI Agent 的高价值产品，即可用更少的成本达到原先人员数量的产出，进一步降低营业成本。

以税友的数值开票员 Agent 和数智会计 Agent 为例，根据税友股份投资者交流 PPT，传统的代账机构开票员平均每月要开 700 账票，工资约 6000 元每月；而 AI 开票员能够提升开票效率 330%，降低开票用工成本 40%，在提升开票准确性与服务体验的同时，还能够提供全天候 7*24 小时服务。数智会计 Agent 方面，行业传统平均会计人效为 80 账套/月，如果是规模为 1000 户的代账机构则需要 13 个会计；而在数智会计 Agent 的辅助下，会计人效提高至 800 户/月，即只需要 2 名会计和 2 个 AI Agent 完成先前的工作量，用人成本大幅减少。

下载日志已记录，仅供内部参考

图 31：AI 接入后，人效及成本获得显著优化

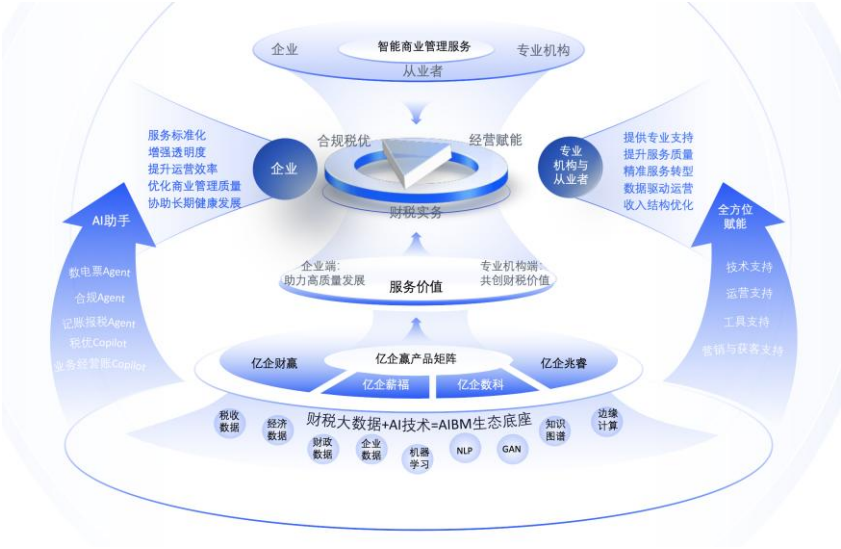


数据来源：公司投资者交流 PPT，广发证券发展研究中心广发证券发展研究中心

此外，AI 的加入能够降低高价值财税服务的业务成本，使得附加值高的财税 SaaS 市场接受度增加。传统模式下，税务诊断、合规分析等深度服务高度依赖于专家经验，往往伴随着高昂的付费成本，而中小微企业大多对价格敏感，因此难以向中小微企业推广。而财税 SaaS 企业能够通过自身厂家大模型，深度融合相关业务的场景与痛点，推动 AI Agent 从工具向决策助手进化，帮助企业在相应场景找到异常点并给出建议，从而在某种程度上取代传统的财税专家。

下载日志已记录，仅供内部参考

图 32：财税 SaaS 企业将 AI 技术创新融入产品中以保持竞争力



数据来源：沙利文，广发证券发展研究中心

公司 B 端进展顺利，财税 SaaS 空间可观，市场格局分散且有收敛迹象。企业数量快速上升的同时，合规要求进一步提高，两者驱动财税 SaaS 市场空间持续扩容。同时，伴随合规逐渐趋严、企业自身对财税信息化的要求愈高且 AI 等各类信息化技术发展迅猛等因素下，难以支撑较高研发强度迭代产品的供应商将逐步退出，行业竞争格局有望呈现逐步优化的态势，具备丰富产品布局和深刻行业 Know-how 的企业有望受益。

税友股份以亿企赢品牌作为依托，提供包括亿企财税、亿企代账、创新产品系列等产品，实现大中小客群全方位覆盖，同时积极推动 AI 应用落地，公司依托“犀友”垂直大模型推出财赢 Agentic 数智工场，平台进一步巩固了竞争壁垒与生态用户黏性，2025 年公司 AI 产品及业务实现回款 4.21 亿元，占数智财税业务 2025 年度回款 28%，较上年同期增长 60%，全年财税业务合同负债 8.05 亿元，较上年期末增长 21%，未来收入可持续性持续提升。

G 端业务保持稳健发展，2025 年毛利率显著修复。公司先后承接了金税三期、自然人税费治理、金税四期等国家重大工程，已成为国家税务信息化建设的关键参与者，政务系统优化与维护的需求长存将为公司 G 端业务贡献稳定驱动力，政务业务营收同比下滑 4.3%至 7.3 亿元，毛利率同比增长 7.2pct 至 36.3%。

3．赛意信息

赛意信息拓展算力租赁新的业务版图。2026 年 5 月 14 日，赛意信息公告拟开展以高性能算力服务器为标的的融资租赁业务，总额预计不超过人民币 200 亿元。2026 年 5 月 20 日，赛意信息公告拟采购高性能算力服务器，并签署相关采购合同，合同总金额预计不超过人民币 8.3 亿元。26 年 6 月，公司与广东七号智算技术有限公司签署战略投资协议，依托七号智算的一体化算力底座能力，进一步提升算力底座的交付效率和服务能力。公司在善谋 GPT、PCB 行业大模型的基础上拓展 AI 算力业务，有望形成较好的业务协同。通过 Agent+模型+算力的全栈布局，进一步增强公司在 AI 领域的竞争力，打开新的市场空间。

下载日志已记录，仅供内部参考

4．卓易信息

EazyDevelop 产品自 25 年 9 月上线以来，商业化落地进展超预期。根据卓易信息官方微信公众号，截至 2026 年 4 月 30 日，AI Coding 业务的核心产品 EazyDevelop 订单突破 8200 万元，其中 4 月的订单金额超过 3200 万元，超过了 26 年 1-3 月订单金额总和。公司 EazyDevelop 产品于 2026 年 3 月推出 EazyBot 等创新功能，大幅提升产品力，用户活跃度表现亮眼。26Q1，EazyDevelop 产品单日 Token 消耗量突破 200 亿，相关项目收入贡献超 50%。26 年 3 月 25 日，公司宣布 EazyDevelop 产品会员订阅套餐价格提升。从 26 年 4 月的表现来看，EazyDevelop 产品的用户和订单的增长趋势并没有因为套餐价格提升而变慢，反而出现了量价齐升的趋势。

公司聚焦于 AI 大模型在 IDE 工具的功能实现，在 AI Coding 领域具有差异化优势。在产业链地位上，公司与 Cursor 公司类似，与 AI 大模型公司是合作关系，受益于 AI 大模型快速迭代的技术红利。其 EazyDevelop 产品接入了包括 GLM、MiniMax、Qwen、Kimi 等多个 AI 大模型，供用户自由选择。26 年 4 月 26 日，EazyDevelop 接入 DeepSeekV4，代码生成能力进一步增强。在产品功能上，EazyDevelop 产品具有三大差异化优势：①全流程开发经验赋能的 Multi-Agent 架构优势；②信创领域先发优势构建的自主可控壁垒；③深入垂直领域，赋能不同领域专业软件开发的行业模板。未来，EazyDevelop 产品的差异化优势有望在 AI 辅助编程商业化落地过程中凸显。

5．汉得信息

得·灵产品/服务系列全面助力企业构建 AI 能力体系，25 年 AI 应用实现 3.1 亿元收入。汉得的 AI PaaS 融合平台，通过预置成型应用组件+丰富的 AI 编排功能+满足 B 端管控要求的管理工具，帮助企业快速搭建 AI 应用。基于 AI PaaS 融合平台，汉得在营销、供应链、制造和财务等领域研发的智慧导购、智能客服、智能物流调度、财务共享精灵等智能体已率先在头部客户实际场景中落地。

经过数年投入打磨，汉得在 2025 年汉得用户大会正式发布得·灵的 B 端 AI 应用产品/服务系列，包括覆盖“模型+平台+应用+服务”四层一体的 AI 产品体系：

① 应用层：“灵手”业务智能体，各大业务领域数十个场景 AI 助手/智能专家

② 工具层：“灵猿”大圣 AI 中台，B 端 AI PaaS 平台+开箱即用基础应用

③ 模型层：“灵睿”垂直模型，多个行业模型+飞码代码大模型

④ 服务层：“灵策”配套服务，AI 咨询规划、企业私有模型训练、AI 应用定制开发

下载日志已记录，仅供内部参考

图 33：汉得信息 AI 能力架构



数据来源：汉得信息投资者交流资料，广发证券发展研究中心

基于 AI 中台，汉得逐步研发多个不同类型的 AI Agent，高效赋能企业 AI 应用实践最后一公里。汉得智能体覆盖业务运营与 IT 建设两大方面，在业务层面，汉得在制造域、营销域、财务域、供应链域等多场景下实现 AI Agent 的研发应用；在 IT 建设层面，汉得实现代码智能辅助、ERP 替换、程序逻辑智能分析等业务应用。汉得通过 AI 中台赋能智能体，逐步从传统的“人工触发操作”升级为“Agent 主动引导、干预”，重构企业运营的核心逻辑。

图 34：汉得信息 AI Agent 建设概览



数据来源：汉得信息投资者交流资料，广发证券发展研究中心

下载日志已记录，仅供内部参考

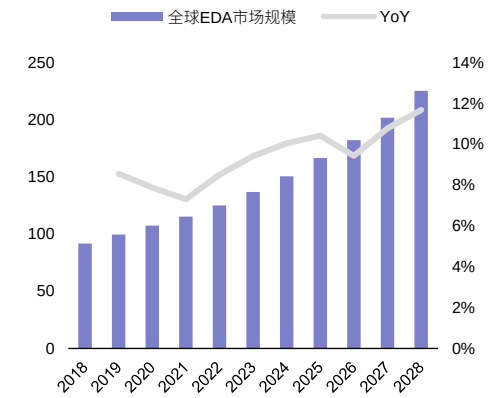
五、EDA：行业成长性大于周期性，国产厂商并购整合进行时

投资要点：EDA 行业兼具技术升级驱动与国产替代弹性，成长性强于周期性。芯片复杂度提升、先进制程演进和定制 ASIC 扩张持续抬升 EDA 工具需求，而外部管制加码和本土先进制程推进进一步强化国产 EDA 的战略价值。当前国产厂商仍以点工具突破为主，与海外龙头在平台化、生态和先进工艺适配方面存在差距，但并购整合正在加速推进，行业有望从分散点工具向平台化、生态化演进。建议关注具备外延整合能力和细分场景卡位优势的国产 EDA 公司，重点包括广立微、概伦电子等。

（一）EDA 行业在技术升级趋势下成长性较强，国产 EDA 成长弹性进一步放大

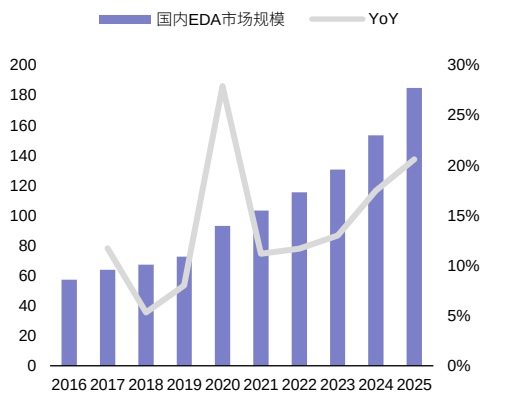
从市场空间看，EDA 行业整体保持稳步扩张态势。根据 technavio，全球 EDA 市场规模由 2018 年不足 100 亿美元增长至目前 150 亿美元以上，年均复合增速维持在高个位数水平。过去几年国产 EDA 整体保持较快增长，但早期受产品成熟度与生态完善度限制，国产替代节奏相对温和。随着近年来技术积累、产品线完善以及行业整合加速，国产 EDA 已进入替代加速阶段，中长期扩容空间清晰。因此，国内 EDA 市场体量虽小但增长更为迅速，根据头豹研究院，市场规模由 2016 年约 50 亿元提升至 2025 年的近 200 亿元，年均复合增速维持在双位数以上，显著高于全球水平。

图 35：全球 EDA 市场空间及增速（单位：亿美元）



数据来源：technavio.com，广发证券发展研究中心

图 36：国内 EDA 市场空间及增速（单位：亿元）



数据来源：ESD Alliance，CSIA，头豹研究院，广发证券发展研究中心

芯片复杂度提升、先进制程演进以及定制 ASIC 市场快速扩张，共同构成了 EDA 行业需求增长的三大核心驱动。一方面，算力规模、系统集成度与架构复杂度的提升，使芯片设计与验证难度显著加大，对 EDA 工具在功能深度、计算精度与效率上的要求提高；另一方面，随着芯片制程持续向先进节点演进，芯片设计成本呈指数级上升，EDA 相关支出也将同步增加；与此同时，云计算巨头及大型系统厂商加速自研定制 ASIC，芯片设计参与者逐渐多元化、项目数量与复杂度也同步上行，进一步放大了 EDA 的使用强度与预算投入。三重因素叠加推动了 EDA 行业需求长期增长。

1. 芯片复杂度提升，对 EDA 工具要求将明显提升

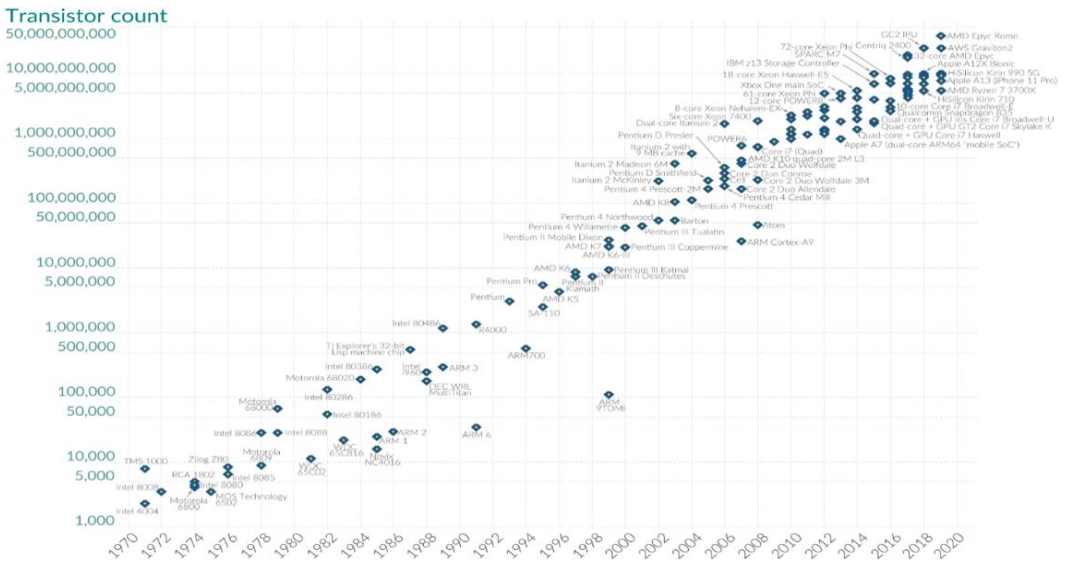
根据摩尔定律，集成电路上可以容纳的晶体管数目在大约每 18 个月到 24 个月就会增加一倍，而单位面积上晶体管数量的增加会对 EDA 软件的功能与性能有明显更高的要求，例如更密集分布的晶体管之间各类物理效应的计算。

因此，随着下一代大规模数字芯片在算力规模、系统集成度和架构复杂度上的持续提升，芯片设计和验证难度

下载日志已记录，仅供内部参考

显著增加，架构设计、功能验证及系统级验证等环节对 EDA 工具要求也将明显提升。

图 37：摩尔定律下每个微处理器上晶体管数量

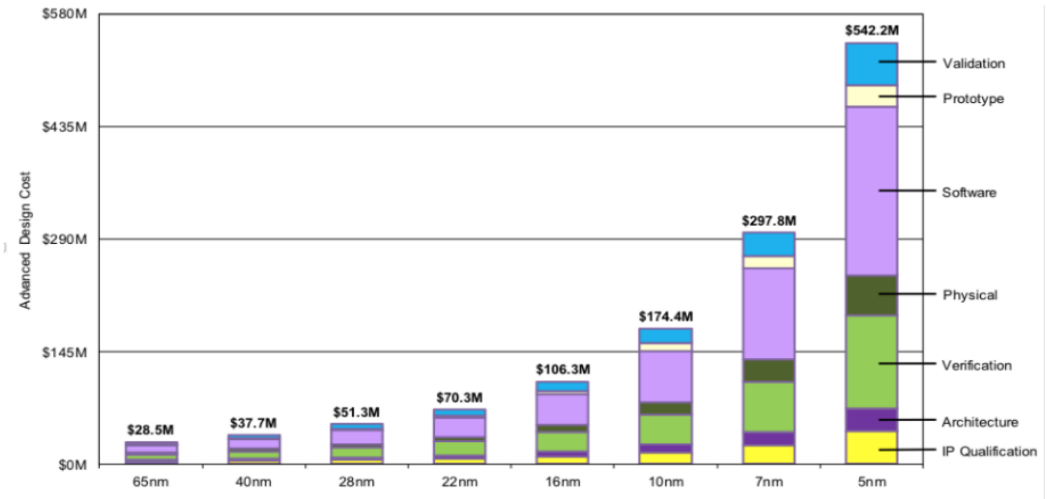


数据来源：SIMA 官网，广发证券发展研究中心

2. 先进制程演进推动芯片设计成本上升，推动 EDA 支出增长

芯片设计成本主要涵盖架构设计、IP 授权与验证、功能验证以及物理实现等多个环节，而这些关键流程几乎全部高度依赖 EDA 工具的支持。每推进一个先进制程节点，芯片的设计复杂度和成本都会大幅提升。例如，相较于 65nm 制程，5nm 工艺下的芯片设计成本增长近二十倍。因此，随着芯片制程持续向先进节点演进，芯片设计成本呈指数级上升，EDA 相关支出也将同步增加。

图 38：不同制程芯片的设计成本



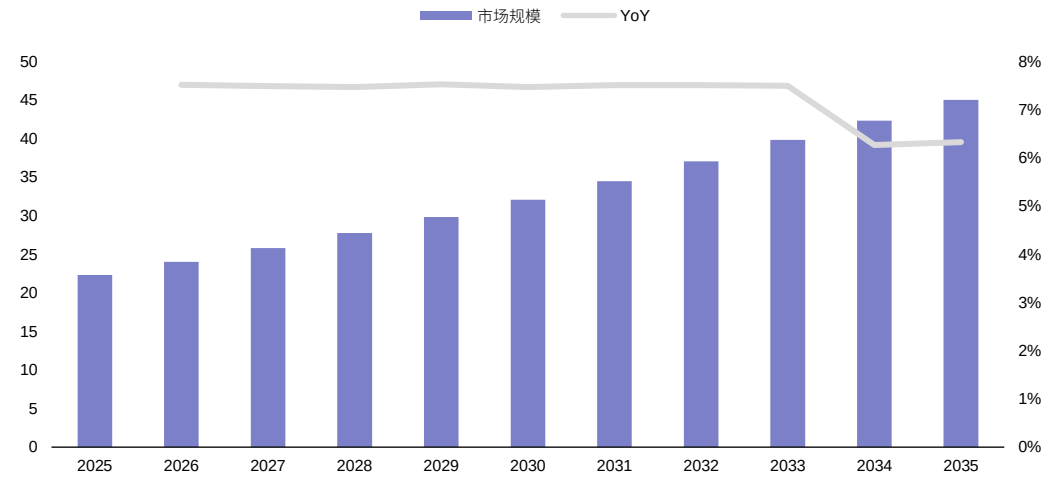
数据来源：semi engineering，广发证券发展研究中心

3. 定制 ASIC 芯片市场高速扩张，成为 EDA 需求的重要增量

下载日志已记录，仅供内部参考

除上述行业共性趋势外，云计算巨头及大型系统厂商加速推进自研 ASIC 芯片，正成为 EDA 行业需求增长的重要增量。一方面，ASIC 项目数量及单项目复杂度显著提升；另一方面，自研 ASIC 通常采用先进制程与高度定制化架构，对设计、验证与物理实现环节依赖度更高，从而抬升单颗芯片对应的 EDA 工具使用强度与预算投入。因此，相较于传统模式下少数专业芯片厂商主导设计，自研 ASIC 的普及扩大了芯片设计的参与者数量，有望直接带动 EDA 需求提升。

图 39：全球 ASIC 市场空间（单位：十亿美元）



数据来源：Precedence RESEARCH，广发证券发展研究中心

4. 国产 EDA 叠加国产替代逻辑，成长弹性进一步放大

伴随外部环境的持续变化，EDA 领域的自主可控需求愈发迫切。继 2018 年 4 月美国商务部对中兴通讯重启禁令后，2019 年 4 月，美国商务部将华为列入实体清单，三大 EDA 巨头暂停软件更新与技术支持；2022 年 8 月，美国首次将 GAAFET 架构 EDA 工具纳入管制；2024 年 12 月，新增四大 ECCN 分类，覆盖计算光刻、多重曝光等制造辅助工具，扩大对于 EDA 软件的限制范围；2025 年 5 月，BIS 限制多家 EDA 公司对中国提供 EDA 软件工具。

图 40：美国对华的 EDA 管控加码



数据来源：Federal Register，Reuters，The New York Times，BIS，广发证券发展研究中心

下载日志已记录，仅供内部参考

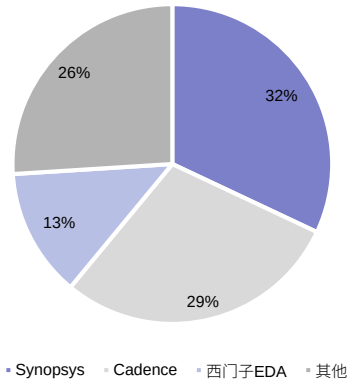
此外，从生态协同角度来看，随着国内先进制程研发进入关键攻坚阶段，制造工艺与国产 EDA 工具之间的绑定趋势将愈发显著。国产设计/制造厂商在推进先进制程过程中，由于工艺参数、器件模型及 PDK 文件等具备高度敏感性，对工艺安全性与协同效率提出了更高要求，海外 EDA 厂商的参与程度有所受限。因此需要与国产 EDA 厂商开展紧密协作，逐步构建面向本土工艺平台的 EDA 流程体系。

总结看，伴随着外部环境变化以及国内先进制程技术的持续推进与阶段性成熟，国产 EDA 工具在本土工艺生态中的绑定效应有望不断增强，相关设计类、制造类等 EDA 需求将随之逐步释放，推动国产 EDA 行业加速发展。

（二）国产 EDA 并购整合提速，从点工具向平台化演进

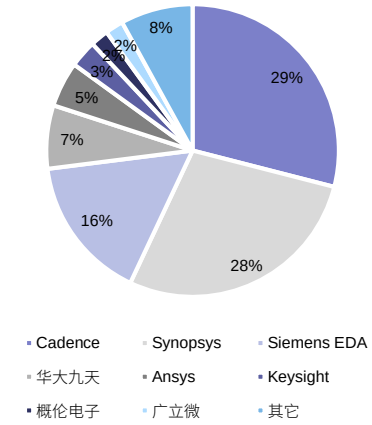
从市场格局看，EDA 行业仍呈现高度集中，国产厂商整体处于二三梯队。①第一梯队由 Synopsys、Cadence 和 Siemens EDA 三大巨头构成，凭借长期高强度的研发投入以及持续的并购整合，已建立较高行业壁垒和用户粘性，确立了行业垄断地位；②第二梯队包括 Keysight、PDF Solutions、华大九天等企业，在部分细分领域具备较为完整的全流程工具，或在特定关键环节拥有领先技术，从而形成了一定竞争优势；③第三梯队则以 JEDAT、芯华章、立芯科技等厂商为代表，主要聚焦于特定应用场景或细分领域的点工具，整体规模和产品体系完整性方面与前两大梯队存在一定差距。

图 41：2024 年全球 EDA 市场竞争格局



数据来源：TrendForce，广发证券发展研究中心

图 42：2022 年国内 EDA 市场空间及增速（单位：亿元）



数据来源：集微咨询，广发证券发展研究中心

国内 EDA 厂商普遍以点工具突破为主，全流程平台较为稀缺。尽管本土厂商在部分环节取得实质性进展，但与国际 EDA 三大巨头相比，在技术完整性、平台化程度和先进工艺适配性方面仍存在显著差距。目前，仅少数头部企业在特定领域实现了局部流程闭环。此外，其他国产 EDA 企业主要聚焦于各自细分赛道，开发具有差异化优势的点工具。

因此国产 EDA 厂商则借助并购整合补全技术短板，推进替代进程。相比国际巨头以巩固全球领先地位为主的并购策略，国产厂商的并购更多是为了补齐关键环节、强化核心能力，形成从点到面的突破，从而缩短差距。例如，2025 年，广立微收购硅光芯片设计自动化领域全球领军企业 LUCEDA，实现从 EDA 到 PDA 的拓展；概伦电子计划收购锐成芯微及纳能微，获得其过去十多年积累的、覆盖数十个工艺平台的上千套各类物理 IP 库，加速实现从 EDA 工具提供商向一站式芯片设计解决方案平台转型；华大九天投资入股思尔芯，基于

下载日志已记录，仅供内部参考

其在数字仿真验证工具链、高效硬件辅助验证技术领域的专业优势，有助于其补全数字 EDA 平台短板。

表 1：国产 EDA 公司投资并购情况梳理

公司	时间	标的	交易金额	持股比例	概述
	2022.10	芯達科技	1000 万美元	100%	补齐数字设计和晶圆制造 EDA 工具短板
华大九天	2024.6	阿卡思微	1.49 亿元	49.75%	补强数字前端形式化验证能力，完善数字验证工具链，提高设计验证效率与覆盖度。
	2025.12	思尔芯	1.10 亿元	7.78%	强化在数字芯片设计与验证环节的综合服务能力
	2023.9	亿瑞芯	0.35 亿元	43%	提升公司集成电路良率提升整体解决方案的核心竞争力
广立微	2025.8	LUCEDA NV	4000 万欧元	100%	作为公司在硅光产业布局的锚点，实现从传统 EDA 到 PDA 的拓展。
	2019.12	博达微科技	0.72 亿元	80%	补强器件建模与 PDK 相关 EDA 与器件特性测试仪器、工程服务能力，完善制造类 EDA 生态
	2021.6	Entasys	约 800 万美元	100%	加强在重点客户的业务拓展深度和广度。
概伦电子	2023.5	福州芯智联	-	100%	填补公司在板级和封装级设计的空缺
	2023.8	Magwel	-	100%	拓展功率半导体及汽车电子领域的上下游客户
	2025.4	锐成芯微	19 亿元	100%	补齐高速接口 IP 与模拟 IP 等关键模块

数据来源：天眼查，iFind，各公司年报，各公司收购公告，合见工软官方公众号，广发证券发展研究中心

在行业理性回归与政策支持推动下，并购整合有望加速。自 2023 年以来，融资事件数量与金额均有所回落，一级市场趋于冷静，行业逐步回归理性发展。在政策端，国家鼓励关键工业软件领域的自主创新与国产替代，设立产业基金和专项项目以支持技术攻关；同时，资本渠道的畅通也为企业并购与扩张提供了资金保障。整体看，国家政策与资本市场的支持为国产 EDA 厂商的并购整合提供了良好条件，未来行业集中度有望提升，格局将逐步优化。

总结看，未来国产 EDA 行业的收并购趋势将呈现出资本背景交叉化、区域集聚化、客户需求及平台能力牵引化的特征。并购不再只是规模扩张的手段，而将逐步演变为推动国产 EDA 从点工具向平台化、生态化演进的核心路径。

建议关注广立微、概伦电子等国产 EDA 公司。

下载日志已记录，仅供内部参考

(1) 广立微

广立微是国内领先的集成电路 EDA 软件与晶圆级电性测试设备供应商，长期深耕集成电路良率提升、电性测试及数据分析领域，面向集成电路制造、设计企业提供 EDA 软件、测试芯片设计服务、晶圆级电性测试设备、数据分析系统等产品与服务，已形成覆盖设计、测试、分析环节的良率提升闭环。

2025 年，公司实现营业收入 7.35 亿元，同比增长 34.40%；归母净利润 8868.80 万元，同比增长 10.49%。2026Q1，公司实现营业收入 1.05 亿元，同比增长 58.47%，收入端延续较快增长，但归母净利润为 -1071.67 万元，主要仍受费用投入、研发投入和季度收入确认节奏影响。

硬件业务方面，公司继续受益于晶圆厂扩产、国产替代以及先进制程/特色工艺对高效率电性测试需求提升。公司在 WAT 测试设备基础上持续向 WLR、WLBI 等可靠性测试场景延伸，硬件品类从单一 WAT 向晶圆级测试和可靠性平台扩张。

软件业务是公司目前的重点投入与布局方向。一方面，公司传统良率分析 EDA 产品和半导体大数据分析系统持续迭代；另一方面，公司通过 DFT、DFM、PDA/PDK 等方向扩展产品边界，推动从测试设备公司向良率提升平台型 EDA 公司升级。

硅光是更大的估值弹性。CPO 量产趋势下，硅光测试和良率管控的重要性提升，公司通过 Luceda 补齐 PDA 能力，有望与自身电性测试设备、良率分析和数据平台形成协同，逐步延伸至硅光测试及良率管控场景。

(2) 概伦电子

概伦电子面向集成电路设计和制造企业，深耕器件建模和电路仿真领域，以“设计-工艺协同优化（DTCO）”理念为核心，产品及服务包括制造类 EDA、设计类 EDA、半导体器件特性测试系统和技术开发解决方案等。

国资背景基金入股提供资源支持。2025 年 7 月，公司发布公告，多位股东与上海芯合创一号私募基金签署了《股份转让协议书》，拟将其合计持有的 5% 股份转让给上海芯合创。上海芯合创是一家具有深厚国资背景的基金，上海国有资本投资母基金有限公司作为有限合伙人认缴出资 58.69%。2025 年 7 月，公司与上海国投、上海芯合创签订《战略合作框架协议》，各方将以 EDA 生态为牵引，聚焦 EDA 行业，推动产业投资合作。此外，2026 年 3 月，上海科创集团也受让公司 5% 股份并完成过户。

收并购与产业投资同步推进，平台化扩张打开长期空间。拟通过发行股份及支付现金方式收购锐成芯微 100% 股权及纳能微 45.64% 股权，并募集配套资金；本次交易已于 2026 年 6 月获上交所并购重组审核委员会审核通过。若交易完成，锐成芯微及纳能微将成为公司全资子公司，公司有望获得覆盖多工艺平台的半导体 IP 资源，加速从“EDA 工具提供商”向“EDA+IP 一站式综合解决方案平台”转型。与此同时，公司持续推进 EDA 产业链投资布局。公司作为有限合伙人对临科芯伦认缴出资 2.5 亿元，临科芯伦认缴出资总额已变更为 16.23 亿元；截至目前，临科芯伦受让国家集成电路产业投资基金所持鸿芯微纳 38.74% 股权已完成工商变更登记。此外，临科芯伦已签署协议，拟以约 2.66 亿元受让国微集团持有的鸿芯微纳 7.8125% 股权，该交易仍待先决条件满足及工商变更完成。整体看，公司通过内生研发、外延并购和产业基金投资同步推进，有望持续补齐 EDA/IP 产品链条，强化在国产 EDA 生态中的平台化卡位。

下载日志已记录，仅供内部参考

六、算力租赁&算力调度：算力需求高增不减，算力租赁及算力调度公司迎来全新发展机遇

投资要点：在行业整体资本开支大幅抬升、算力需求持续高景气、算力基建加码投放的三重核心背景支撑下，算力租赁、算力调度赛道相关企业有望迎来明确成长机遇。其一，头部云厂商大规模资本开支落地将持续扩充算力硬件供给池，为算力租赁服务商锁定稳定、可持续的货源基础，有力托底租赁业务营收规模与经营稳定性；其二，下游 AI 应用端旺盛且长期刚性的算力需求，叠加海量异构算力资源集中落地投产，市场对高效资源分配、负载均衡、跨池协同的算力调度配套服务需求同步释放，调度环节作为算力产业链运营提效的关键枢纽，价值与业务空间同步打开。

算力租赁方面，我们建议关注杰创智能、智微智能、赛意信息、宏景科技。算力调度方面，我们建议关注卡莱特、并行科技、思特奇、云赛智联等。

（一）应用新形态继续突破，算力需求增强态势不减

应用不断出现新形态，带动算力需求持续增长。以 OpenClaw 为例，其带来的丰富功能和便捷化操作大大推动了相关 Agent 应用的落地。AI Agent 对于 Tokens 的消耗远高于 AI 大模型文字生成任务。与传统 Chatbot 仅需单轮、短文本交互的低 Token 消耗模式不同，AI Agent 因具备自主规划、多步推理、循环思考、工具调用及长上下文读写能力，其 Token 消耗并非线性增加，而是呈现指数级、数量级跃升：单智能体执行任务时的 Token 消耗通常为大语言模型文字生成任务的 10 倍以上。例如，在社交媒体平台上调用 Agent，生成并发布一个带有文本和图片的帖子，其消耗量约 1500-2000 Tokens；而 GPT 5.2 模型在博客和邮件场景中生成 300 字的文字内容，仅消耗 110 Tokens。而多 Agent 协同、复杂业务流程自动化等场景下更是可达百倍甚至更高。以 OpenClaw 为代表的 AI Agent 在执行任务过程中自主规划任务以及调用工具产生的大量 Tokens 有望进一步推动 AI 算力需求的快速增长。

表 7：大语言模型 Tokens 消耗量

模型	简短回复（100字）	博客开头/邮件（300字）	长文/深度回答（800–1000字）
Gemini 2.5 Flash	约6 tokens	约20 tokens	约60 tokens
Gemini 3 Flash	约8 tokens	约24 tokens	约70 tokens
GPT-5.2	约36 tokens	约110 tokens	约320 tokens
Claude Sonnet 4.5	约40 tokens	约120 tokens	约350 tokens

数据来源：SiteGround.com，广发证券发展研究中心

表 8：agent 产品的 Tokens 消耗量

Agent	行为	消耗tokens
社交媒体Agents	生成并发布一篇帖子（文本+图片）	约1500–2000 tokens
电子邮件营销Agents	起草一个活动和主题	约1000–1800 tokens
WordPress Agents	创建一篇格式规范、分类、标签以及由大型语言模型生成的特色图片的博客文章	约2000–3000 tokens

数据来源：SiteGround.com，广发证券发展研究中心

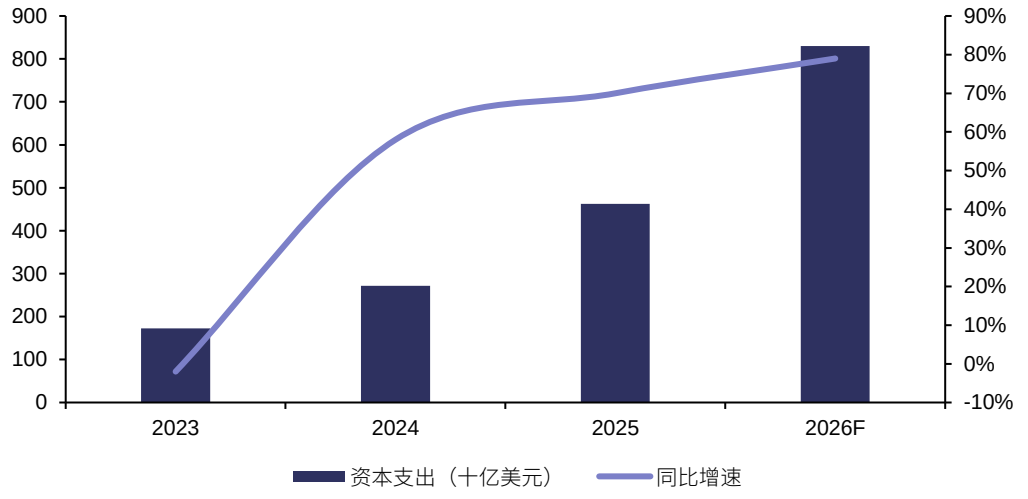
下载日志已记录，仅供内部参考

我们预期伴随应用的持续发展，以及新形态的不断突破，算力需求当维持高位。此外，从相关龙头的财报数据中，也可以印证这一点。以 Oracle 为例，26Q3Oracle 剩余履约义务（RPO）达到 5530 亿美元，同比增长 325%，显著超出市场预期，订单增长主要来自与 OpenAI、Meta Platform 等公司在 AI 数据中心与算力基础设施方面的合作订单。

从供给角度出发，算力基本建设投入维持高位。根据 IDC 的数据，我国算力规模将于 2028 年达到 2781.9EFLOPS，24-28 年 CAGR 为 39.9%；通用算力于 2028 年达 140EFLOPS，24-28 年 CAGR 为 18.3%。根据 IDC 的数据，2025-2027 年期间，中国人工智能服务器市场规模增速预计分别为 36.2%、30.2%和 28.8%。

根据 TrendForce，美国四大 CSP 集体上修 CapEx。微软将其资本支出展望上调至 1900 亿美元，同比增长 130%。Google 将其指引从 1750-1850 亿美元上调至 1800-1900 亿美元。Meta 将其资本支出区间从 1150-1350 亿美元向上修正为 1250-1450 亿美元。AWS 今年的资本支出受 AI 云服务需求推动，预计将超过 2300 亿美元。因此，TrendForce 也修订了全球前九大 CSP——Google、AWS、Meta、Microsoft、Oracle、字节跳动、腾讯、阿里巴巴和百度——2026 年合计资本支出的预测，上调至约 8300 亿美元，年增长率从 61%提高至 79%。

图 43：全球前九大 CSP 的 Capex 预测（2023-2026F）



数据来源：TrendForce、广发证券发展研究中心

下载日志已记录，仅供内部参考

（二）算力租赁及算力调度公司有望迎来新的发展机遇

1．算力租赁

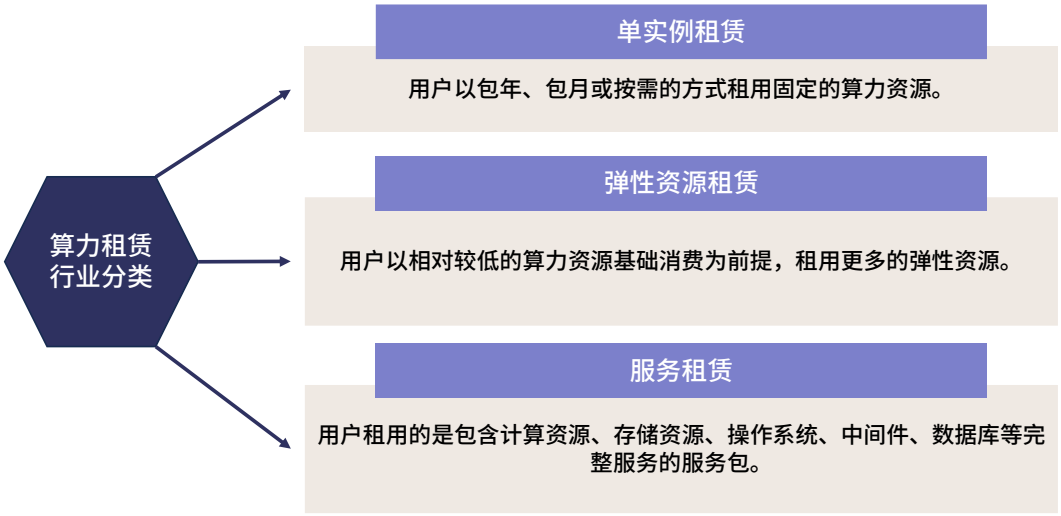
算力租赁是将计算能力或云计算资源以租赁的方式提供给用户使用，这一过程中用户可按需选择使用资源类型和时间，且无需承担运维、升级等工作及相应成本。算力租赁一方面使得下游客户灵活使用算力的门槛降低，更利于众多 AI 初创公司快速实现产品服务的开发与应用，另一方面也充分利用闲置的算力资源，提升算力资源的利用效率。目前，我国算力租赁的模式主要包括单实例租赁、弹性资源租赁与服务租赁。

① **单实例租赁：**用户可通过包年、包月或是按需结算的形式，租用固定规格的算力资源，资源配额相对稳定，适配需要持续稳定算力支撑的业务场景。

② **弹性资源租赁：**该模式以基础算力资源消费作为前提，用户在此之上可额外调用更多弹性算力资源，算力使用规模能够跟随业务需求灵活调整。

③ **服务租赁：**用户租赁的并非单一计算硬件，而是整合了计算、存储资源，搭配操作系统、中间件、数据库在内的全套完整服务包，整体交付形态一体化。

图 44：算力租赁分类模式

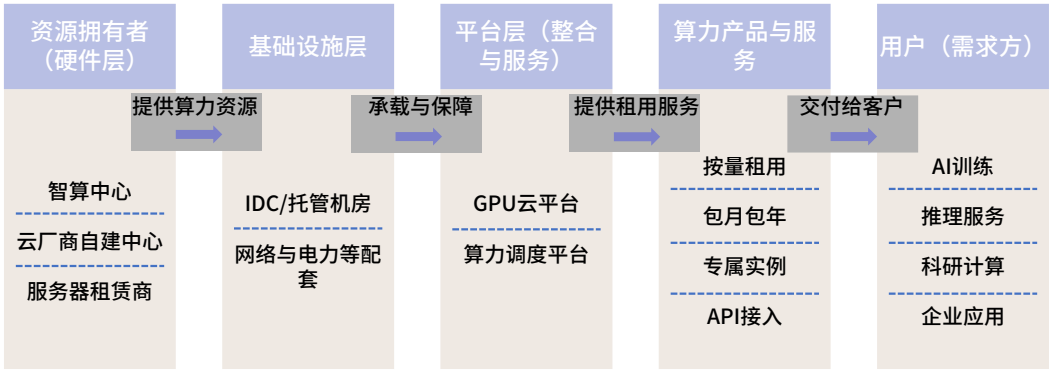


数据来源：华经产业研究院、广发证券发展研究中心

从业务流出发，算力租赁本质上是将底层硬件资源转化为标准化服务进行输出的过程。产业链上游主要包括智算中心、云厂商及服务器资源拥有者，负责提供 GPU 服务器及相关算力资源；中游依托 IDC 机房、电力网络及运维体系完成算力资源部署和运行保障，并通过 GPU 云平台、算力调度平台等系统实现资源池化管理和统一调度；在此基础上，运营商将底层算力封装为标准化产品，通过按量计费、包月包年、专属实例或 API 调用等方式向客户提供服务；下游客户则根据自身需求调用算力资源，广泛应用于人工智能训练与推理、科研计算以及企业数字化转型等场景。整体来看，算力租赁行业通过“资源整合—基础设施承载—平台调度—服务交付”的运营模式，实现算力资源的规模化运营和持续性变现，而资源获取能力、平台调度能力及客户拓展能力则构成企业核心竞争力。

下载日志已记录，仅供内部参考

图 45：算力租赁业务流

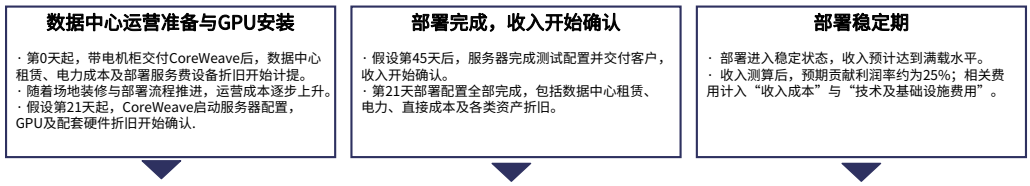


数据来源：广发证券发展研究中心

典型算租从采购安装-部署确收-稳定运行的各阶段利润率呈现“前低后高”的特征。根据 CoreWeave 官网提供的各阶段利润率来看，在项目启动阶段，运营商完成 GPU 服务器采购、上架及数据中心部署，由于收入尚未确认，而需承担设备折旧、机房租赁、电力及运维等成本，因此项目通常处于亏损状态。以单卡月收入 100 美元为例，第 1 个月仅完成部署，收入尚未确认，而收入成本及技术基础设施费用分别为 27 美元和 12 美元，对应贡献利润为-39 美元。随着服务器完成配置并交付客户使用，项目进入收入确认阶段。假设部署周期约 45 天，第 2 个月开始部分确认收入，月收入约 50 美元，但仍需承担设备折旧、机房租赁、电力及技术服务费用，对应贡献利润约-25 美元，利润率约为-50%。当项目进入稳定运营阶段后，GPU 利用率逐步提升，客户持续租用算力资源，收入确认趋于稳定。以第 3 个月及以后为例，月收入达到 100 美元，在收入成本 27 美元和技术基础设施费用 48 美元的基础上，项目贡献利润约 25 美元，对应贡献利润率约 25%。此后随着上架率和利用率进一步提升，以及电力、运维等固定成本被摊薄，项目盈利能力仍有继续改善空间。

下载日志已记录，仅供内部参考

图 46：算租公司从采购安装-部署确收-稳定运行的各阶段利润率参考

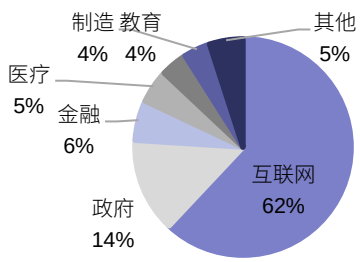


项目	第1个月	第2个月	第3-60个月
示意性部署费用收入	-	$50	$100
(-) 收入成本	-27	-27	-27
(+) 技术及基础设施费用	-12	-48	-48
示意性部署费用贡献利润	($39)	($25)	$25
示意性部署费用贡献利润率	无意义	-50%	25%

数据来源： CoreWeave 官网、广发证券发展研究中心

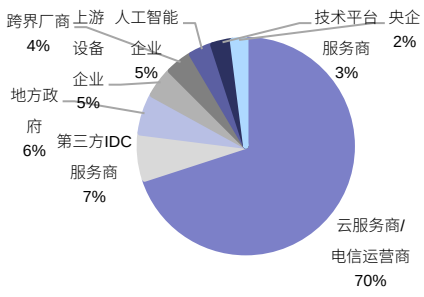
从当前的市场主流参与者以及行业格局出发，当前我国智能算力租赁市场呈现较高集中度特征。根据科智咨询，供给端主要由云服务商及电信运营商主导，市场份额合计约 70%，凭借资金实力、数据中心资源及客户渠道优势占据行业主导地位；第三方 IDC 服务商、地方政府主导平台分别占据约 7%和 6%的市场份额；上游设备厂商、跨界厂商、人工智能企业、技术平台服务商及央国企等主体合计占据约 17%的市场份额，行业整体呈现“头部集中、多元参与”的竞争格局。需求端方面，互联网行业是智能算力租赁最主要的应用领域，占整体市场需求的 62%，主要用于大模型训练、AI 推理、推荐算法及云计算等场景；政府部门占比约 14%，金融、医疗、教育及制造行业占比分别为 6%、5%、4%和 4%，广泛应用于智慧政务、金融科技、辅助诊疗、智能教育及工业仿真等领域。整体来看，当前智能算力租赁市场呈现供给侧向云厂商和运营商集中，需求侧向互联网行业集中的双重集中化特征，随着大模型应用从训练逐步向推理侧扩展，以及“AI+”政策持续推进，金融、医疗、制造等行业有望成为未来算力需求增长的重要来源。

图 47：中国智算能力租赁市场行业集中度



数据来源：科智咨询《2024-2025 年中国智能算力租赁行业市场研究报告》，广发证券发展研究中心

图 48：2025H2 中国企业级大模型调用市场



数据来源：科智咨询《2024-2025 年中国智能算力租赁行业市场研究报告》，广发证券发展研究中心

下载日志已记录，仅供内部参考

我们认为，对于算力租赁公司而言，最核心的能力有三项，分别为稳定的卡资源获取能力、全链路交付运维落地能力以及稳健的资金周转调度能力。

① **稳定的卡资源获取能力：**互联网等行业客户算力采购体量偏大，单笔订单对应服务器部署规模较高，对服务商硬件批量交付水平形成较高门槛；供货履约存在时效约束，若交付进度不及约定标准，企业或将面临相应履约赔付约束，稳定长效的芯片拿货渠道是承接大额订单的基础支撑条件。

② **全链路交付运维落地能力：**AI 算力服务器具备严苛的运行环境标准，硬件到货后需经过设备调试、集群组网、性能调优一系列流程，达标后方可对外提供算力服务；设备上线后还要依靠专业技术团队开展长期运维值守。高效的交付上线节奏、更低的硬件损耗与故障频次，有助于优化业务运转效率、摊薄整体运营开支，改善经营盈利水平。

③ **稳健的资金周转调度能力：**AI 算力服务订单普遍具备大额资金体量特征，经营主体通常需要依托多元融资工具匹配订单垫资需求；资金周转调度水平，直接影响企业订单按期交付、设备投产上线全流程的推进节奏。

我们建议关注杰创智能、智微智能、赛意信息、宏景科技等具备算力租赁业务的相关公司。

2．算力调度

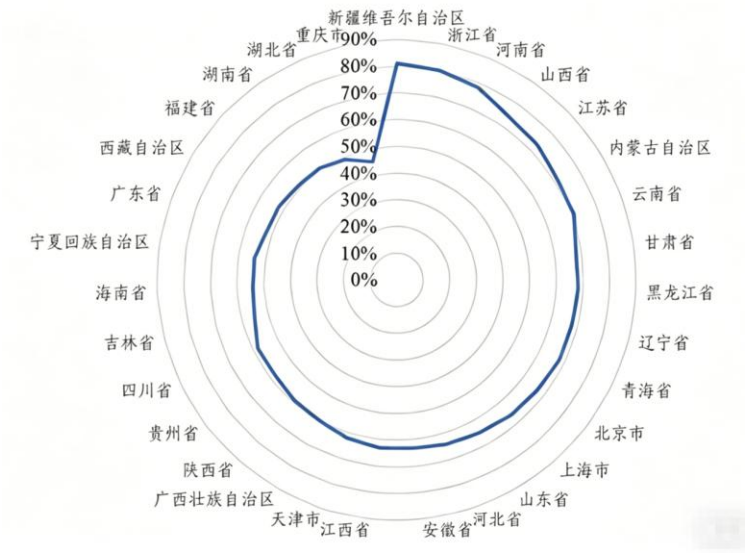
尽管我国算力相关资源建设加速，投入维持高水平，但是有着算力资源利用率低、算力的弹性需求高、算力中心所处区域错配下导致供需不平衡、以及异构资源较多导致管控难度较高的特点。

① 算力资源利用率低

我国算力资源整体利用率较低，存量盘活空间大。受前期“新基建”驱动及地方政策激励影响，我国数据中心建设节奏整体偏快，行业供给在短期内集中释放，但实际需求消化相对滞后，导致资源利用效率整体偏低。分省份来看，浙江省、江苏省、山西省、内蒙古自治区等省份依托本地需求或周边地区外溢去求，实现算力资源供需高效对接，上架率平均超 70%；广东省、河北省等省份上架率约为 50%，仍有提升空间。根据科智咨询数据，2023 年 IDC 机架平均上架率在 58%左右，同时其也预测该数据在未来几年将保持平稳，而我国在用算力中心机架总规模已达 1085 万标准机架（截至 25 年 6 月底），以此数据为基础进行测算，我国仍有约 456 万标准机架的存量算力空间有待盘活与释放。

下载日志已记录，仅供内部参考

图 49：我国各省 IDC 上架率情况



数据来源：中国信通院，广发证券发展研究中心

② 算力的弹性需求较高

AI 训练、云渲染及科学计算等核心业务场景的算力需求呈现出显著的波动性特征，从而驱动企业对弹性算力的需求持续提升。具体来看，在模型训练、项目交付等阶段性任务推动下，算力需求在短期内快速攀升，而在非高峰阶段又明显回落，整体利用率存在较大不确定性，导致算力资源易出现阶段性紧缺与长期闲置并存的现象。在此背景下，传统以一次性大规模资本开支为主导的算力基础设施建设模式难以有效适配企业需求：一方面，其前期投入高、回收周期长，显著加重企业资本负担；另一方面，固定化部署缺乏灵活调整能力，既难以应对业务高峰期的算力需求激增，也易在需求回落阶段造成资源闲置，整体投入产出效率承压。因此，能够实现按需获取、动态扩展与灵活调配的弹性算力逐步成为企业刚性需求，而算力调度平台通过整合分散资源并优化配置效率，正成为支撑弹性算力供给、提升资源利用率的关键基础设施。

③ 算力中心所处区域分散且错配导致供需不平衡

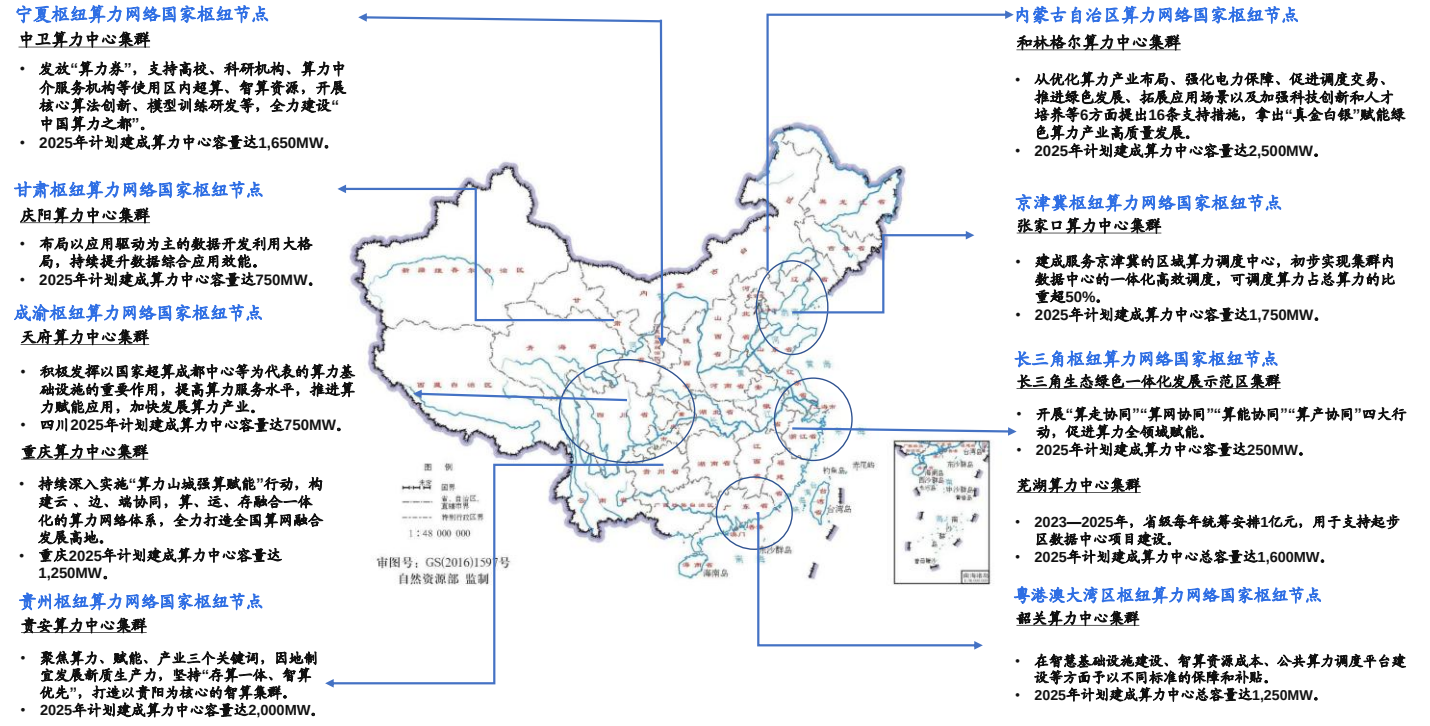
我国算力资源分布呈现显著的区域错配，供需矛盾突出。一方面，东部沿海地区数字经济发达，人工智能、大模型等新兴应用驱动算力需求持续旺盛，但当地面临土地资源紧张、能耗指标收紧等硬约束，电力成本高昂且供给紧缺，一线城市及周边算力中心供给已现稀缺性。另一方面，西部地区风光水电等清洁能源丰沛，气候条件适宜自然冷却，土地资源广阔，且电价优势明显，但本地数字产业化基础相对薄弱，算力需求有限。这种结构性错配，使得单纯依靠本地建设难以满足算力的高效、普惠供给，跨区域协同成为必然选择。

为破解区域错配难题，国家层面大力推进“东数西算”工程与全国一体化算力网建设，通过顶层设计引导算力资源向西部集聚。目前，我国已布局京津冀、长三角、粤港澳大湾区、成渝、贵州、内蒙古、甘肃、宁夏等 8 大国家算力枢纽节点及 10 大数据中心集群，构建了“东数西算”的骨架。同时，跨区域算力协同网络加速成型，国家枢纽节点间传输时延不超过 20ms，集群到周边主要城市传输时延不超过 5ms，区域内算力节点间时延达 1ms，显著降低了东西部数据传输壁垒。通过网络与调度的协同，东部的人工智能训练推理、视频渲染、

下载日志已记录，仅供内部参考

存储备份等业务正有序向西部迁移，逐步实现“东数西算”、“东数西训”、“东数西渲”的跨域供需匹配与资源互补。

图 50：“东数西算”八大算力枢纽、十大集群算力政策要点



数据来源：灼识咨询，广发证券发展研究中心

④ 异构资源较多导致管控难度较大

算力资源涵盖 CPU、GPU、FPGA、ASIC 等多种架构，每种硬件的性能特点（如并行计算能力、能效比）和指令集差异显著。例如，GPU 擅长并行计算（如深度学习训练），而 CPU 更适合逻辑控制密集型任务。同时，不同任务依赖特定的软件框架（如 TensorFlow、PyTorch）和库（如 CUDA、OpenCL），调度系统需确保软件环境兼容性。若任务代码未适配目标硬件的编程接口（如 CUDA 代码无法直接在 HIP 生态中运行），则需进行代码迁移或重构，显著增加调度成本。同时，由于当前算力调度缺乏统一的计量标准和接口规范，导致跨架构资源难以互通，管控难度较大。

表 9：异构算力需要处理不同硬件架构

	CPU	GPU	FPGA	DSA	ASIC
架构特点	通用处理器，擅长复杂指令和串行任务	大规模并行计算单元，适合高吞吐量任务	可编程硬件，支持自定义逻辑电路	针对特定领域优化的专用硬件	针对单一任务完全定制的专用芯片
性能	单线程性能强，但并行能力有限	并行能力强，适合大规模数据处 理	针对特定领域优化的专用硬件	针对单一任务完全定制的专用芯片	专为特定任务设计，性能最优
灵活性	高度灵活，可运行各种软件	灵活性中等，需依赖并行算法	针对单一任务完全定制的专用芯片	在特定领域性能极高	完全无灵活性，只能执行固定功能
功耗	功耗较高，尤其在	功耗较高，但单位	性能介于 CPU 和	专为特定任务设	功耗最低，专为效

下载日志已记录，仅供内部参考

	高性能计算时	计算能耗优于 CPU	ASIC 之间，延迟低	计，性能最优	率优化
开发成本	开发成本低，成熟工具链	开发成本中等，需掌握并行编程技术	在特定领域性能极高	灵活性低，仅适用于特定领域	开发成本最高，需要大量前期投入
升级与迭代	易于升级，只需更换软件	需要重新设计算法，但硬件兼容性较好	专为特定任务设计，性能最优	完全无灵活性，只能执行固定功能	无法升级，需重新制造

数据来源：头豹研究院，广发证券发展研究中心

因此，在面对上文中我们所提到的四点现状下，如何更好的对算力进行调度运营则更显得重要，而算力调度平台则是实现这一点的基础。总的来说，算力调度运营平台能够通过统一架构整合异构资源，并简化算力获取流程，实现高效便捷的服务供给。因此也能够在一定程度上解决我国算力需求与资源分布不均的矛盾,快速满足上层应用多样化的算力需求,助推我国数字经济进入普惠共享的新阶段。

算力互联网体系架构解析：

从整个算力互联网体系出发，根据中国信通院，最终的组网大致可分为三层，由下至上分别为算网设施层、互联资源层、应用服务层。

① **算网设施层：**算网设施层由计算与传输两类物理基础设施组成，是算力互联网的底层资源载体。①计算设施包括通用计算、智能计算、超级计算等数据中心以及云、边、端等算力资源。②传输设施包括公共互联网、企业算力专网等，承载算力节点间数据高效、可靠传输的承载功能。

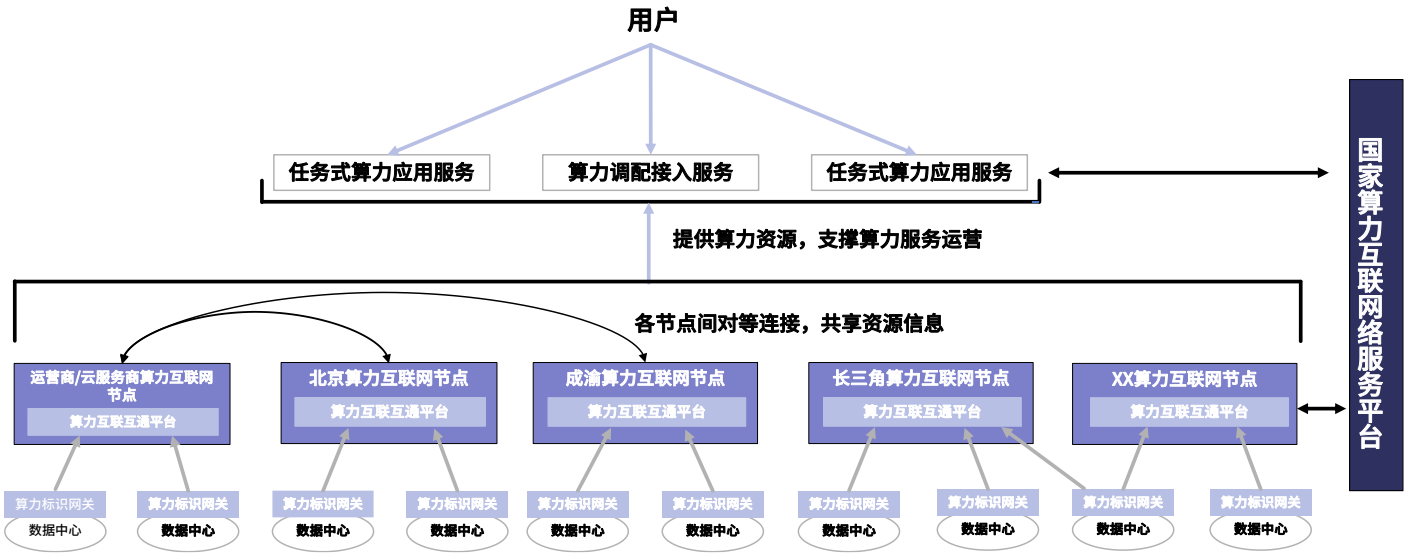
② **互联资源层：**互联资源层由国家算力互联网服务平台、区域和行业算力互联互通平台组成，位于算网设施层和业务应用层之间，是算力互联网的核心节点。①国家算力互联网服务平台，承担算力资源标识标准化分配、运行监测、信息发布等综合管理职能，旨在实现全国算力资源的统一调配与管理。②国家算力互联网（区域）核心节点，整合统筹区域内跨服务商的多元异构算力资源。③国家算力互联网（行业）核心节点，当前主要面向基础电信运营商、云服务商等具备全国算力资源布局且正在开展算力并网的企业，结合算力专网或云专网形成算网一体化调度平台，依据算力互联网建设相关标准规范升级已有平台，接入国家算力互联网平台体系，承担加速算力供需匹配、培育算力服务业态等功能。

互联资源层中，其中的算力互联网各级节点具有对等互联关系，也就是区域及行业核心节点通过算力互联互通平台接入公共算力资源，并基于对等交换机制实现算力资源信息共享与协同。

③ **应用服务层：**面向用户提供多元化的算力服务，主要包括传统算力资源式服务、算力调配接入服务以及算力应用服务等不同算力服务业务。该层通过统一的服务接口与服务目录，实现算力能力的封装与开放，为用户提供建立在统一调度体系之上的便捷化、普惠化算力使用模式，助力实现“像用电一样使用算力”的服务愿景。

下载日志已记录，仅供内部参考

图 51：算力互联网各节点和平台关系图

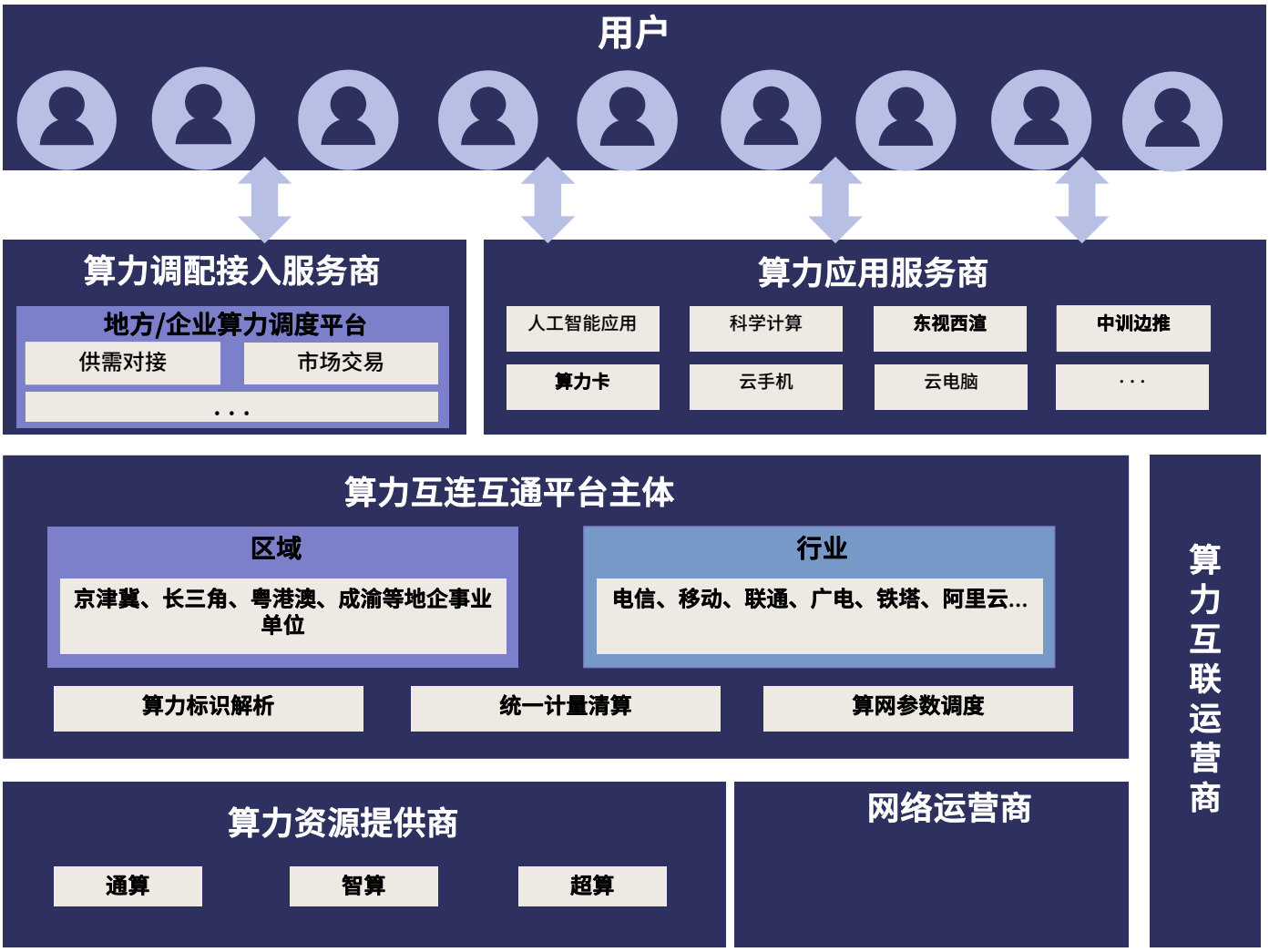


数据来源：中国信息通信研究院，广发证券发展研究中心

因此，算力互联网生态体系是由多元主体协同构成，涵盖从算力资源供给、网络传输、平台支撑到应用服务的全链条参与方。主要生态主体包括：算力资源提供商（提供算力）、网络运营商（提供传输服务，保障算力资源高效调度）、算力互联互通平台主体（负责建设和运营算力互联互通平台的机构）、算力调配接入服务商（面向算力任务需求，整合多源异构算力资源，并提供供需对接）、算力应用服务商（提供基于算力资源封装的各类应用层服务，如算力卡、云电脑等）以及算力互联运营商（同时具备算、网服务的主体）等，各主体在生态中承担不同角色，共同推动算力资源的高效整合与服务化演进。

下载日志已记录，仅供内部参考

图 52：算力互联网体系相关主体图



数据来源：中国信息通信研究院，广发证券发展研究中心

从产业链的角度出发，总共涉及供给方、需求方以及中间方。

需求方主要由政府及企业构成。大模型计算需求场景主要包括训练、微调以及推理，模型参数规模与对算力的消耗成正比，参数规模越大，对智能算力的需求越大。不同体量的智算中心支撑不同的大模型计算场景。根据中国信通院，推进基础预训练大模型（千亿级以上参数）的训练，需要具备 E 级（EFlops）计算能力的高端万卡集群中心支撑。推进小模型 (百亿级及以下参数) 的训练、微调，或推进模型的推理，依托百 P 级（100PFlops）计算能力的中小型智算中心即可支撑。

供给方则是上图中提及的算力资源提供商。该主体主要为用户提供通算、智算、超算等软硬件计算能力的主体，核心业务包括裸机租赁、云服务等。主体类型包括运营商（如中国移动、中国联通等，多用于满足标准化需求，便于企业快速部署和接入）、第三方算力中心服务商（如万国数据、世纪互联等，兼具传统零售业务和大规模批发定制业务，其中集群化的算力中心定制业务主要服务于对算力中心有高度定制化要求的大型企业）、地方

下载日志已记录，仅供内部参考

政府自建（服务于本地经济发展，促进产业能升级和转型）。根据灼识咨询，以算力中心行业供给 IT 容量计算，2023 年第三方服务商供给 IT 容量占比接近 50%。同时，分通用算力和智能算力来看，智能算力占绝对体量，以灼识咨询所预计的 2026 年数据基础为计算，中国计算设备算力规模中，智能算力占比达到 89%。

中间方则是算力运营调度平台主体，主要对接供需，并完成最终的算力互通。从分类上看，按行政层级分类可以划分为国家级算力调度平台、省级平台以及市级平台。国家级算力调度平台多由政府主导或运营商/头部企业建设，强调跨区域协同与市场化交易；省级平台覆盖长三角、成渝、京津冀等重点区域，结合本地产业需求；市级平台则聚焦本地 AI、智能制造等场景。

以上海市智能算力资源统筹调度服务平台为例，根据中国信通院，该平台接入上海本地及市外各类空闲算力资源，可实现跨市域、跨资源池、跨厂商的异构算力资源交易和调度，截至 2025 年 7 月，平台已接入上海仪电、上海电信、上海移动、上海联通以及阿里、百度等互联网云厂商、商汤、算丰等第三方算力企业的算力，并实现新疆克拉玛依、湖南长沙的优质算力接入，上架总算力规模已超过 1.8 万 P。

图 53：电信运营商和第三方算力中心服务商特点比对

	传统电信运营商	第三方算力中心服务商
服务灵活性	· 低，以标准化产品为主	· 高，提供定制化服务，从设计、建设、服务 · 内容均可按照客户需求定制
服务内容	· 单一网络接入，带宽租赁服务为主 · 云资源池&机柜托管 · 标准化服务	· 网络接入丰富 · 多元化服务，包含带宽租赁、机柜托管、异地镜像、VPN、数据备份等 · 定制化服务解决方案
运维能力	· 较弱，以标准化的运维服务方案为主	· 较强，提供7×24小时专人运维服务，具备高 · 灵活专业性，可提供定制化运维方案
主要客户类型	· 政府、央企、国企及小部分互联网厂商	· 以大型互联网厂商为主
主要承接业务模式	· 以零售业务为主，含小部分定制批发业务	· 以定制批发业务为主
地域布局	· 以完善全国布局为主，在中西部地区投建 · 超大规模算力中心	· 算力中心资源布局在一线城市及其周边地区 · 响应“东数西算”政策，向西部地区算力枢 · 纽与节点积极布局
代表性服务商	中国电信 CHINA TELECOM、中国移动 China Mobile、中国联通 China unicom	GDS万国数据、世纪互联 VNET、CHINDATA GROUP 秦淮数据集团

数据来源：中国信通院，灼识咨询，广发证券发展研究中心

从算力订购流程来看，由用户、运营层、调度层、监测层、资源层（并网）五大主体协同完成，整体分为订购开通与使用计费两大阶段。

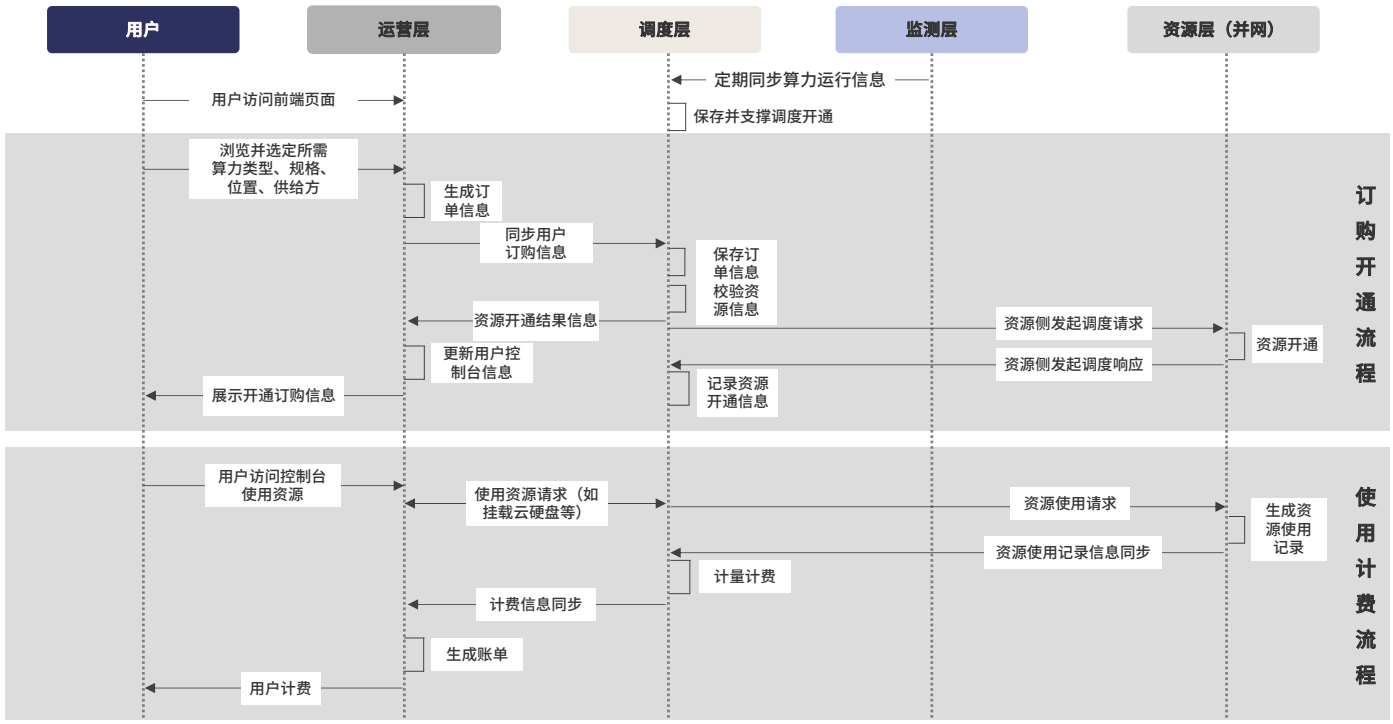
① 在订购开通阶段，用户首先访问运营层前端页面，浏览并选定所需算力的类型、规格、位置、供给方等信息；运营层生成订单信息并同步至调度层，调度层完成订单信息保存、资源信息校验后，向资源层发起资源调度开通请求，资源层完成资源开通并返回响应；调度层记录资源开通信息，由运营层更新用户控制台

下载日志已记录，仅供内部参考

信息，并向用户展示开通订购信息，完成算力资源的开通。同时，监测层定期同步算力运行信息，由调度层保存并支撑调度开通全流程。

② 在使用计费阶段，用户访问控制台使用资源，运营层向调度层发起资源使用请求，调度层同步至资源层生成资源使用记录；资源层同步资源使用记录信息至调度层，调度层完成计量计费后，由运营层同步计量计费信息、生成账单，最终向用户完成计费，实现算力资源从订购到使用、计费的全流程闭环管理。

图 54：算力订购流程简析



数据来源：全国数据标准化技术委员会，广发证券发展研究中心

算力调度运营的商业模式，核心是通过算力资源池化、智能调度与市场化服务实现价值变现，主要涵盖算力银行、算力超市两大创新形态，辅以算力租赁、调度佣金、增值服务等多元盈利路径。2026 年 4 月 2 日，工信部发布《关于开展普惠算力赋能中小企业发展专项行动的通知》，明确提出探索“算力银行”“算力超市”等新业态，支持闲置算力存入、跨区域跨周期调度与灵活取用，为算力调度运营划定政策方向与合规框架，推动算力从资源租赁向资产化、服务化、普惠化转型，构建供需高效匹配、价值持续增值的产业生态。

① 算力银行借鉴金融存贷逻辑，以算力存管、错峰调度、价值变现为核心盈利模式。运营商搭建统一平台，吸纳企业、数据中心的闲置 GPU/CPU 算力“存入”，按性能、时长给予供给方算力利息或现金收益；面向 AI 初创、中小企业等需求方，提供算力 借贷服务，按卡时、核时或 Token 进行灵活计费。其获利路径清晰，①赚利差，以较低成本归集闲置算力，高价出租获取差价；②抽佣金，每笔算力交易抽取一定比例的服务费；③金融增值，开展算力质押、算力贷、资产证券化等服务，赚取年化利差与服务费，实现算

下载日志已记录，仅供内部参考

力资源跨时空配置与价值最大化。

② 算力超市是算力领域的“电商平台”，以资源聚合、按需选购、交易撮合为核心盈利逻辑。运营商整合多家算力服务商的通算、智算、超算等异构资源，标准化封装为算力产品上架国家数据局，企业可像网购一样自由对比性能、价格，按需选择短时租用、弹性扩容、包月包年等服务，一键开通、即用即付。其盈利主要来自三方面：①交易佣金，向入驻服务商收取交易额一定比例的平台费；②流量分成，按算力调用量、Token 消耗量获取阶梯分成；③增值服务，提供模型部署、运维托管、算力券对接、技术支持等，叠加高毛利服务收入，降低企业用算门槛的同时，实现规模化、普惠化盈利。

我们建议关注卡莱特、并行科技、思特奇、云赛智联等。

下载日志已记录，仅供内部参考

七、智能驾驶与机器人板块：高阶智驾渗透率加速抬升，油价抬升加速海外市场成长，产业玩家积极布局机器人新领域

投资要点：在比亚迪等头部 OEM 加码智驾趋势下，国内智驾渗透率加速抬升，但在价格竞争仍未明显缓解，产业链仍处于市占率收敛的背景下，智驾零部件环节主要投资机会集中在竞争格局较好的智驾芯片行业、规模效应明显的龙头 Tier1 以及充分受益于细分领域成长的个股（如受益于欧标前装政策的锐明技术）。此外，智驾产业参与者立足现有技术积累，积极布局机器人新领域，有望打开新空间。

（一）智能驾驶板块：智驾渗透率仍处抬升通道，但在 OEM 补贴退坡及价格竞争背景下，优选格局好的智驾芯片及出海赛道，建议关注从 0 到 1 突破的无人驾驶赛道

在技术发展、政策红利、OEM 差异化竞争、消费者需求逐步攀升等多重因素的驱动下，自动驾驶的渗透率持续提升。根据佐思汽研数据，目前国内乘用车 ADAS 系统功能（L2-L2.9）装配量持续攀升，相较于 2023 年，2024 年 L2、L2.5、L2.9 配置渗透率同比增长 5.14pct、1.42pct、3.38pct。

目前 L2.5 正逐渐替代 L2 成为乘用车入门级标配功能。中国乘用车高阶智能辅助驾驶渗透率将大幅提升，进入“全民智驾”时代。同时，部分 OEM 主机厂已提出，将减少、或不再开发 L2.5 级以下新车型，推动高速 NOA 逐渐成为新车入门标配。根据佐思汽研数据，预计 2030 年 L2、L2.5、L2.9 配置渗透率分别为 26.3%、35.6%、24.5%，相较于 2024 年同比增长-14.0pct、+33.4pct、19.6pct。从微观层面看，伴随 2025 年初比亚迪官宣大规模上智驾后，进一步倒逼智驾渗透率抬升，尤其是中低端车型。

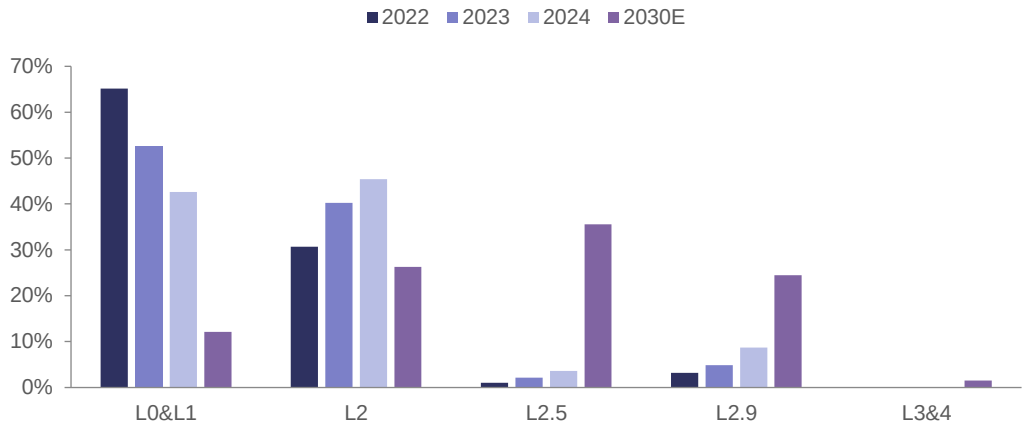
结合 2026 年上半年进展来看，OEM 价格竞争未明显缓解，且面临补贴退坡，上半年国内汽车销量承压，但在油价高企背景下，海外电车市场及智驾市场迎来新机遇，我们建议重点关注如下方向：

（1）竞争格局好的细分赛道：智驾芯片赛道，目前具备量产能力的国产厂商数量仅有 3 家（地平线、黑芝麻智能、华为），且国产化抬升空间可观；

（2）出海方向：2026-2027 年海外主机厂亦将更为规模化上智驾系统，受益方向包括海外业务规模敞口大的 tier1，以及激光雷达。

下载日志已记录，仅供内部参考

图 55：我国智能驾驶渗透率



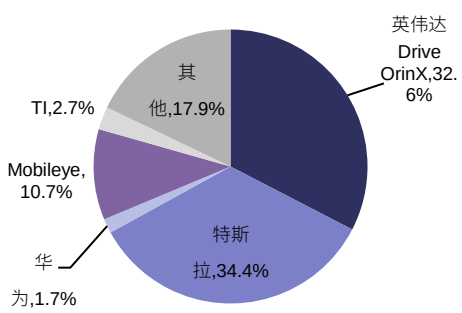
数据来源：佐思汽车，广发证券发展研究中心

推荐方向 1：智驾芯片行业。智驾芯片放量趋势受益于智驾渗透率的持续抬升，结合主要计算平台的软件开发便利性、功能安全、生态、价格等维度来看，本土智驾芯片参与者综合能力已逐步壮大。

尤其值得注意的是，地缘政治格局变动之下，智驾芯片自主可控是大趋势。2013-2015 年以来汽车智能化潮流日盛，而在 2018-2019 年之后，汽车智能化的国产化方案越发受 OEM 重视，仅以智能驾驶芯片为例，2020 年之前产业主要玩家皆是海外芯片商，如 Mobileye、TI、英伟达，从 2021 年至今，地平线、华为等国产智能驾驶芯片商强势崛起，在国内份额日益扩张，但国产智驾芯片总体市占率仍偏低。

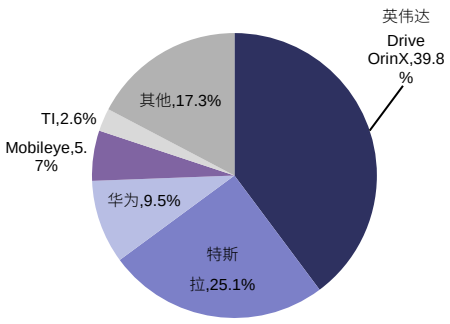
2025 年下半年尤其是 2026 年以来，国产高阶智驾芯片进展迅速，黑芝麻 A2000 进展加速，地平线 J6 系列持续量产，在国内高阶制程有望提速的背景下，国产高阶智驾芯片的供给侧压力弱化，而竞争格局亦或出现一定新变化。

图 56：2023 年我国智驾芯片市占率



数据来源：盖世汽车，广发证券发展研究中心
注：不包括进口和选配

图 57：2024 年我国智驾芯片市占率



数据来源：盖世汽车，广发证券发展研究中心

下载日志已记录，仅供内部参考

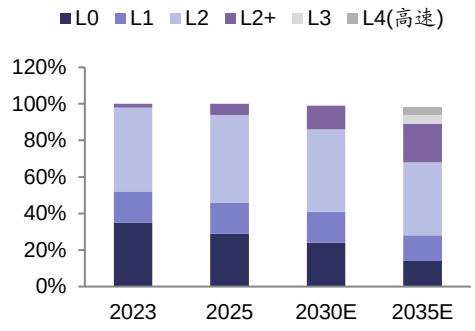
推荐方向 2：海外智驾渗透率处于抬升中早期阶段，且竞争烈度优于国内，看好海外敞口高兼具产品竞争力的智驾供应商。

乘用车领域，在电动化转型加速、中国 OEM 竞争等多重因素驱动之下，海外 OEM 智驾有望加速推进。从供给端来看，目前中国汽车智能化产业链总体处于全球领先位置，产业规模持续壮大，智能驾驶行业的中国本土供应商已经具备了国际竞争力，以德赛西威、速腾聚创、禾赛科技等为代表的本土智驾供应商在海外市场进展迅速；从需求端来看，近年来海外新能源汽车渗透率呈现快速提升趋势，尤其随着近年来丰田、大众等海外车企加速智能化转型，但海外智驾渗透率仍处于显著低位。

中东地缘政治变化驱动油价抬升，加速海外电车市场渗透率抬升，利好智驾发展前景。根据中汽协、佐思汽研数据，2026 年 1-4 月新能源出口销量累计 120 万辆，渗透率达 44.1%，累计同比增长 102.6%。在全球加速脱碳、多地禁售燃油车时间表提前、国际油价因地缘冲突居高不下的大环境下，中国新能源汽车以“高性价比+低使用成本+智能化体验”三重优势，强势叩开欧洲、东南亚、中东乃至拉美市场的大门。

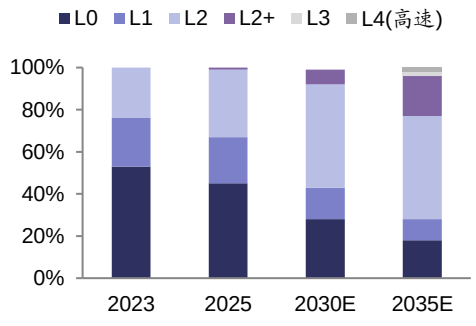
商用车领域，以锐明技术为代表智驾细分领域集成商，充分受益于欧标前装的新机遇。根据欧盟 GSR 法规 EU 2019/2144 的要求，自 2024 年 7 月 7 日起，所有新注册上牌的 M2、M3 和 N2、N3 类车辆都应装备相应的安全系统，覆盖的功能主要是车道保持（ELKS、LDWS）、盲区监测（BSIS、倒车监测、MOIS）、前向预警与紧急制动（AEBS）和智能车速辅助（ISA）等 4 种功能。欧标实施后，有望催生 10-15 亿美元新增前装需求市场，锐明技术在欧标客车方面已取得先发优势，2024 年开始批量化出货，后续有望在空间更大的货车领域取得突破。

图 58：美国智驾渗透率



数据来源：波士顿咨询，广发证券发展研究中心

图 59：欧洲智驾渗透率



数据来源：波士顿咨询，广发证券发展研究中心

（二）机器人板块：看好智驾玩家的跨界突破，并重点推荐全球 AMR 龙头-极智嘉

从产品形态和技术路线来看，自动驾驶汽车本质上就是将人从 A 点运输至到 B 点的机器人，因此应用于智驾领域的主控芯片、传感器等核心零部件亦可运用于人形机器人领域。

从技术框架来看，两者也均遵循“感知-决策-控制”的自主系统架构。在技术架构以及核心算法都存在较多重合领域的情况下，智驾芯片、传感器等零部件是可以被复用在机器人上的。目前诸多芯片厂商如黑芝麻智能、

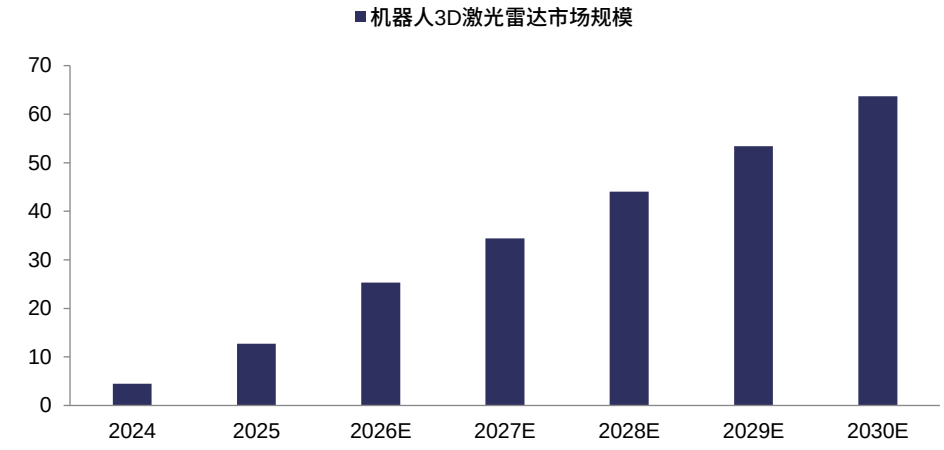
下载日志已记录，仅供内部参考

地平线机器人以及英伟达等已开始将自研发的智驾芯片置于 IOT 或是机器人之上。

以黑芝麻智能为例，公司在开拓机器人领域方面具备可观的产业化潜能。公司基于智驾领域积累的深厚实力，推出的机器人算力平台应用场景覆盖低速轮式机器人、多足机器人、具身智能等主流品类。基于 SesameX 计算平台，按照算力及可支持的软硬系统配置，划分为 Kalos、Aura、Liora 三款模组，由低到高，可支持低速轮式机器人（如送餐机器人/迎宾机器人/巡检机器人/清洁机器人/物流机器人/教育机器人等）、多足机器人（如遥操机器人/工业协作机器人、巡检维护机器人、智能机械臂、协作机械臂等）、具身智能（如人形机器人、多机器人协作系统等）。

在传感器领域，激光雷达在机器人领域的跨界进展亦在加速。依托自动驾驶技术的长期发展，为激光雷达提供了成熟的技术路径和规模化降本基础。从技术侧来看，机器人相对低速的自主移动是汽车高速自动驾驶的延伸，3D 激光雷达的精准、全天候测距功能，助其实现精准定位、自主导航、安全避障，是机器人动态感知物理世界、进行安全交互的三维感知入口。目前，3D 激光雷达已规模应用于智能割草机器人、具身智能、低速无人配送等领域，未来，伴随规模化成本的进一步下降，3D 激光雷达应用场景将持续扩容。

图 60：2024-2030 年机器人 3D 激光雷达市场规模（亿元）



数据来源：高工，广发证券发展研究中心

针对具备细分应用场景的机器人，我们重点推荐具备全球竞争力的 AMR 产业龙头极智嘉。相较于 AGV、机械臂等传统仓储自动化技术，AMR 解决方案回报周期显著缩短，后续 AMR 渗透率有望加速抬升。

AMR VS **机械臂及传送带：AMR 解决方案具备更强柔性和可复制性。**对于电商零售、快消、物流等下游客户而言，由于货物吞吐有明显的季节性，如果采用机械臂及传送带等传统仓储自动化解决方案，往往需要锚定货物吞吐峰值要求进行自动化设备投入，而 AMR 解决方案的硬件载体为单个机器人，可以根据货物吞吐量的波动来动态调整 AMR 机器人布置数量，具备更强的柔性和可复制性。

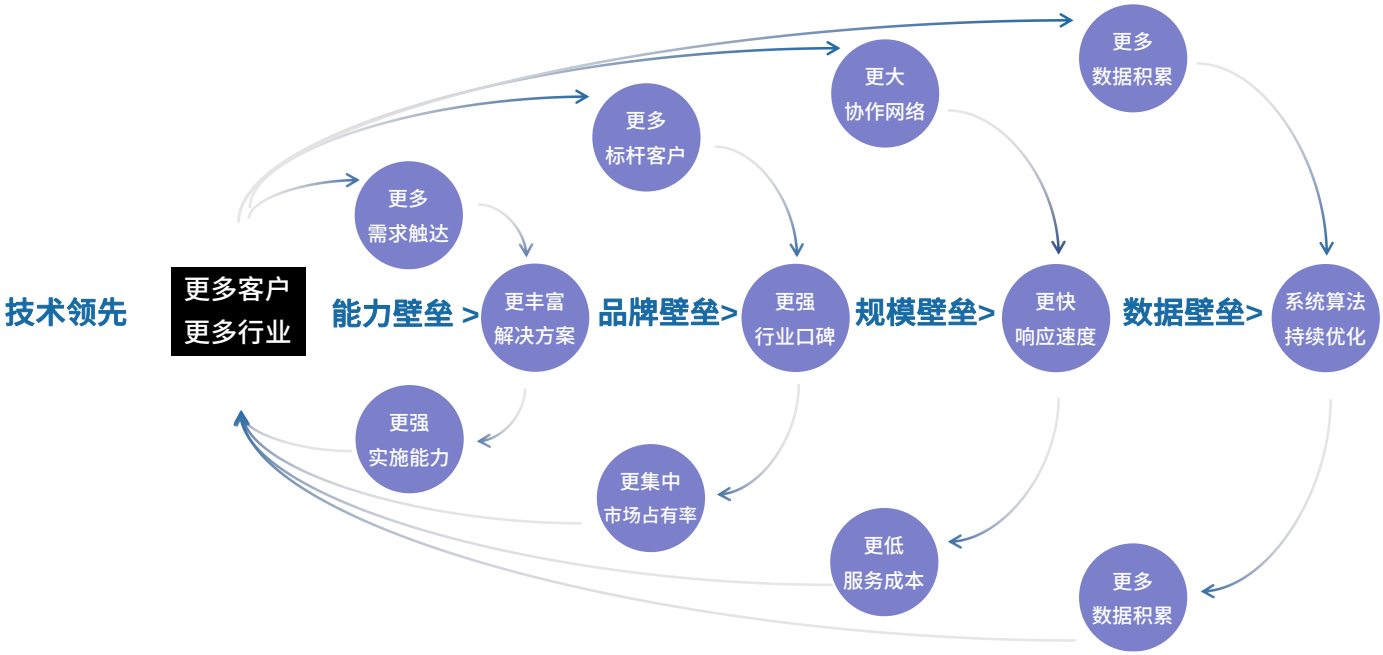
极智嘉已形成商业壁垒的正循环，充分受益于仓储自动化产业红利。自 2015 年公司成立以来，经过逾 10 年的商业实践，公司已在软件算法及硬件层面确立了行业领先性，以此为基础，公司形成了产品、品牌、成本与

下载日志已记录，仅供内部参考

软件算法层面竞争优势的正循环。

尤其是在硬件产业链已相对成熟的背景下，软件算法作为 AMR 产业的核心差异竞争点，需要更多行业客户的场景打磨，作为拥有全球最大客户基础的仓储履约 AMR 解决方案提供商，极智嘉深刻掌握行业 know-how。

图 61：极智嘉已形成商业壁垒正循环



数据来源：广发证券发展研究中心

下载日志已记录，仅供内部参考

八、风险提示

（一）AI 应用落地不及预期风险

当前国内 AI 应用场景较为分散，整体业务体量有限，尚未形成全行业级别的增长主线。若大模型、AI Agent 等技术在各行业落地进度放缓、商业化效果未达预期，将拖累整条 AI 产业链的估值与业绩兑现。

（二）国产算力与模型生态发展不及预期风险

“国产算力 + 国产模型” 适配仍处于推进阶段，若国产 AI 芯片性能迭代、供应链恢复节奏放缓，或是大模型适配优化、生态建设进度滞后，会延缓 AI 算力国产替代进程，相关企业业绩增长承压。

（三）行业竞争加剧与盈利下滑风险

AI 模型、算力租赁、智能驾驶等赛道参与者持续增多，市场价格战频发。模型推理价格持续下行、算力租赁行业前期投入大回报周期长、智驾领域车企降价竞争传导至上下游，将压缩产业链整体盈利空间。

（四）宏观环境与市场流动性波动风险

市场风险偏好、资金流动性存在不确定性，尤其港股等市场受情绪影响显著。同时公共部门与企业端 IT、AI 相关支出修复力度不足，叠加外部政策、地缘等因素扰动，会压制板块整体表现。

下载日志已记录，仅供内部参考

广发计算机行业研究小组

刘雪峰 ：首席分析师，东南大学工学士，中国人民大学经济学硕士，1997 年起先后在数家 IT 行业跨国公司从事技术、运营与全球项目管理工作。2010 年就职于招商证券研究发展中心负责计算机组行业研究工作，2014 年加入广发证券发展研究中心。

吴祖鹏 ：资深分析师，中南大学材料工程学士，复旦大学经济学硕士，曾先后任职于华泰证券、华西证券，2021 年加入广发证券发展研究中心。

周 源 ：资深分析师，慕尼黑工业大学硕士，2021 年加入广发证券，曾任职于 TUMCREATE 自动驾驶科技公司，负责大数据相关工作。

王钰翔 ：高级分析师，哥伦比亚大学运筹学硕士，2024 年加入广发证券发展研究中心。

戴亚敏 ：高级研究员，北京大学金融硕士，2024 年加入广发证券发展研究中心。

分析师及联系人



刘雪峰

SAC 执证号：S0260514030002
SFC CE No. BNX004
gfliuxuefeng@gf.com.cn



吴祖鹏

SAC 执证号：S0260521040003
wuzupeng@gf.com.cn



周源

SAC 执证号：S0260523040001
shzhouyuan@gf.com.cn



王钰翔

SAC 执证号：S0260526060001
SFC CE No. BXP717
wangyuxiang@gf.com.cn

请注意，吴祖鹏,周源并非香港证券及期货事务监察委员会的注册持牌人，不可在香港从事受监管活动。

联系人：
戴亚敏 021-38003697
daiyamin@gf.com.cn

下载日志已记录，仅供内部参考

广发证券
行业投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 10%以上。
持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。
卖出：预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券
公司投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 15%以上。
增持：预期未来 12 个月内，股价表现强于大盘 5%-15%。
持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-5%～+5%。
卖出：预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

地址	邮政编码	客服邮箱
广州市： 广州市天河区马场路 26 号广发证券大厦 47 楼	510627	gfzqyf@gf.com.cn
深圳市： 深圳市福田区益田路 6001 号太平金融大厦 31 层	518026	
北京市： 北京市西城区月坛北街 2 号月坛大厦 18 层	100045	
上海市： 上海市浦东新区南泉北路 429 号泰康保险大厦 37 楼	200120	
香港： 香港湾仔骆克道 81 号广发大厦 27 楼	-	

法律主体声明

本报告由广发证券股份有限公司或其关联机构制作，广发证券股份有限公司及其关联机构以下统称为“广发证券”。本报告的分销依据不同国家、地区的法律、法规和监管要求由广发证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。

广发证券股份有限公司具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管，负责本报告于中国（港澳台地区除外）的分销。

广发证券（香港）经纪有限公司具备香港证监会批复的就证券提供意见（4 号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的分销。

本报告署名研究人员所持中国证券业协会注册分析师资质信息和香港证监会批复的牌照信息已于署名研究人员姓名处披露。

重要声明

广发证券股份有限公司及其关联机构可能与本报告中提及的公司寻求或正在建立业务关系，因此，投资者应当考虑广发证券股份有限公司及其关联机构因可能存在的潜在利益冲突而对本报告的独立性产生影响。投资者不应仅依据本报告内容作出任何投资决策。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或者口头承诺均为无效。

本报告署名研究人员、联系人（以下均简称“研究人员”）针对本报告中相关公司或证券的研究分析内容，在此声明：（1）本报告的全部分析结论、研究观点均精确反映研究人员于本报告发出当日的关于相关公司或证券的所有个人观点，并不代表广发证券的立场；（2）研究人员的部分或全部的报酬无论在过去、现在还是将来均不会与本报告所述特定分析结论、研究观点具有直接或间接的联系。

研究人员制作本报告的报酬标准依据研究质量、客户评价、工作量等多种因素确定，其影响因素亦包括广发证券的整体经营收入，该等经营收入部分来源于广发证券的投资银行类业务。

本报告仅面向经广发证券授权使用的客户/特定合作机构发送，不对外公开发布，只有接收人才可以使用，且对于接收人而言具有保密义务。广发证券并不因相关人员通过其他途径收到或阅读本报告而视其为广发证券的客户。在特定国家或地区传播或者发布本报告可能违反当地法律，广发证券并未采取任何行动以允许于该等国家或地区传播或者分销本报告。

本报告所提及证券可能不被允许在某些国家或地区内出售。请注意，投资涉及风险，证券价格可能会波动，因此投资回报可能会有所变化，过去的业绩并不保证未来的表现。本报告的内容、观点或建议并未考虑任何个别客户的具体投资目标、财务状况和特殊需求，不应被视为对特定客户关于特定证券或金融工具的投资建议。本报告发送给某客户是基于该客户被认为有能力独立评估投资风险、独立行使投资决策并独立承担相应风险。

本报告所载资料的来源及观点的出处皆被广发证券认为可靠，但广发证券不对其准确性、完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策，如有需要，应先咨询专业意见。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券的立场。广发证券的销售人员、交易员或其他专业人士可能以书面或口头形式，向其客户或自营交易部门提供与本报告观点相反的市场评论或交易策略，广发证券的自营交易部门亦可能会有与本报告观点不一致，甚至相反的投资策略。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且无需另行通告。广发证券或其证券研究报告业务的相关董事、高级职员、分析师和员工可能拥有本报告所提及证券的权益。在阅读本报告时，收件人应了解相关的权益披露（若有）。

本研究报告可能包括和/或描述/呈列期货合约价格的事实历史信息（“信息”）。请注意此信息仅供用作组成我们的研究方法/分析中的部分论点/依据/证据，以支持我们对所述相关行业/公司的观点的结论。在任何情况下，它并不（明示或暗示）与香港证监会第5类受规管活动（就期货合约提供意见）有关联或构成此活动。

下载日志已记录，仅供内部参考

权益披露

(1)广发证券（香港）在过去12个月内与极智嘉-W(02590)公司有投资银行业务关系。

版权声明

未经广发证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、刊登、转载和引用，否则由此造成的一切不良后果及法律责任由私自翻版、复制、刊登、转载和引用者承担。