

2026年中国Token工厂 发展白皮书

从算力大国到产能强国：中国Token经济的机遇、格局与治理

2026年8月 · 中国深圳

中商产业研究院 编制



六个判断：中国正在成为全球最大的Token生产国，但产能质量决定最终座次

- 01

需求爆发全球罕见。中国大模型日均Token调用量从2024年初约1,000亿增至2026年3月超140万亿，两年增长超千倍；2026年全年MaaS调用量预计达40,000万亿。
- 02

供给规模全球第二。2026年6月中国智能算力达2,185 EFLOPS（FP16），在用标准机架超1,445万架，全国近300个智算中心项目覆盖29个省份。
- 03

利用率是最大短板。国内智算中心GPU平均利用率不足30%，2025年智算用算占比仅36.8%——“建多用少”倒逼产业从建设竞赛转向经营竞赛。
- 04

格局高度集中。火山引擎、阿里云、百度智能云占据公有云MaaS调用量近九成份额；第三方推理平台与运营商正在侧翼打开空间。
- 05

国产芯片跨过拐点。2025年国产AI加速卡出货165万张、份额41%；昇腾以81.2万张领跑，DeepSeek V4已将底座迁移至昇腾CANN框架。
- 06

西部成为Token成本洼地。八大枢纽承载全国八成智算，庆阳、乌兰察布、中卫凭绿电与气候优势承接主流大模型推理，“东数西算”升级为“东Token西产”。

目录

CONTENTS

01 概念与兴起背景

Token工厂定义 · 全球坐标 · 中国脉络 · 三大驱动力 · 政策牵引

02 市场现状：需求、价格与供给

千倍需求 · MaaS市场 · 价格体系 · 算力规模 · 建设热潮

03 区域格局：东数西算下的产能地图

区域分布 · 八大枢纽 · 庆阳样本 · 西部第二梯队

04 竞争格局：五类玩家与国产芯片底座

云厂商 · 运营商 · 第三方平台 · 芯片生态

05 问题与挑战

利用率 · 供需错配 · 盈利难题 · 风险矩阵

06 产业链、商业模式与政策体系

五层产业链 · 模式演进 · 算力券机制 · 政策清单

07 趋势研判与中商建议

规模预测 · 五大趋势 · 分主体行动建议

定义：Token工厂是把电力与芯片转化为"可结算智能"的生产设施

定义（中商产业研究院，2026）

Token工厂是以Token（词元）为标准产品组织生产的AI基础设施：它将电力、AI芯片、大模型与调度软件集成为连续流水线，批量产出Token并按其计量、计价与交付。Token是大模型处理信息的最小语义单元——一个汉字约等于1—2个Token，也是当前AI服务唯一的跨厂商结算单位。

理解Token工厂的四个关键词

计量

Token可计数、可计价、可核销，让算力第一次有了标准化"产品"

流水

7×24小时连续推理，像工厂排产一样追求稳定的产出曲线

转换

本质是"电力→算力→智能"的转换装置，效率决定成本

结算

按调用量付费，产出直接挂钩收入，产能即利润

为什么"工厂"这个比喻在中国格外贴切

中国是制造业大国，对"产能、良率、成本曲线"的产业语言最为熟悉。Token工厂把抽象的AI能力翻译成可经营的生产要素：绿电是原料、芯片是机床、模型是工艺、调度是产线管理、Token是出厂产品。在这一框架下，东部的需求、西部的绿电、国产芯片的突破与统一大市场的调度，恰好构成一条完整的"智能生产线"——这是中国发展Token工厂区别于欧美市场的结构性禀赋。

全球坐标：Token经济兴起，中国是最活跃的Token生产与消费双中心

8,000万亿/日

全球Token消耗展望 全球日均Token消耗预计从2025年约100万亿增至2030年8,000万亿，年复合增速约140%。

>1万亿美元

全球数据中心资本开支 2026年预计首次突破1万亿美元（Dell'Oro），北美四大云厂商指引上限合计近7,500亿美元。

-99.9%

Token成本三年降幅 同等能力模型调用成本从约60美元/百万Token降至约0.06美元，硬件、算法、系统三重效率相乘。

中国在全球Token经济中的三个"最"

最快的需求增速

日均调用量两年增长超千倍，2025年公有云MaaS调用量同比增长16倍，全球无出其右。

最激进的Token定价

国产主力模型输出均价已降至国际巨头的1/10—1/34，成为全球Token价格的"洼地"。

最完整的产业纵深

从绿电、国产芯片到模型、应用全链条齐备，是全球唯一具备全栈自建Token产能能力的经济体之一。

中商解读：全球Token竞争的上半场拼"谁有大模型"，下半场拼"谁能低成本量产Token"。中国的需求规模、电价与制造优势正对应下半场的胜负手。

中国脉络：从"东数西算"到Token工厂，五年完成三级跳



三个阶段的政策与市场重心

阶段	政策重心	市场重心	核心指标
布局期（2022-2024）	建枢纽、上项目、定目标	圈地、囤卡、抢节点	机架数、P算力、投资额
放量期（2025-2026）	算电协同、算力券、互联互通	抢调用量、打价格、争份额	Token调用量、MaaS份额
运营期（2027—）	盘活存量、标准结算、绿电考核	拼利用率、拼成本、拼服务	每瓦Token、单Token成本、毛利

形态辨析：从IDC到Token工厂，评价标尺换了三代

维度	传统数据中心（IDC）	智算中心（AIDC）	Token工厂
产品	机柜、带宽、电力	算力（卡时、算力包）	Token（可结算的智能产出）
核心指标	机架数、上架率、PUE	算力规模、卡数、利用率	Token吞吐、每瓦产出、单Token成本
计费方式	按机柜/带宽租赁	按卡时、按训练任务	按百万Token、按任务结果
客户买到	空间与资源	潜在产能（需自建平台）	现实产能（即调即用）
风险归属	几乎无经营风险	利用率风险在客户	利用率风险在厂商，倒逼效率

中商解读 三代形态不是替代关系，而是叠加关系：中国现有的近300个智算中心项目中，绝大多数仍停留在"卖资源"和"卖算力"阶段——收入依赖机柜租赁与卡时出租，真实Token产出能力不足。Token工厂的本质差别在于厂商自己承担利用率风险：客户按调用量付费，空转的成本由厂商消化。这意味着只有具备模型、调度与推理优化能力的厂商才敢进入这一业态，行业门槛从资金转向了工程能力。

需求引擎：智能体、推理模型与多模态，把Token消耗推向新量级

1

智能体（Agent）：单次任务Token消耗放大约两个数量级

对话场景输入短、输出长；Agent场景需要读取长上下文、调用搜索与数据库、拆解任务多轮推理，输入输出比例可达100:1，上下文从数万Token跃升至百万级。2026年Agent类应用规模化落地，是MaaS调用量预测再增20倍的首要原因。

×20

IDC预测2026年中国MaaS调用量相对2025年的增长倍数

2

推理模型："深度思考"成为可分级定价的商品

用户看到1个可见Token，模型内部可能已生成上百个思维链Token。国产推理模型（DeepSeek-R1系列等）把"深度思考"做成了付费档位，思考深度第一次成为Token定价的维度。

数百倍

复杂推理任务相对普通文本查询的能耗倍数（IEA）

3

多模态：视频与语音的Token密度高出一个数量级

视频生成能耗可达文本查询的数百倍（IEA）。国产多模态模型（豆包、Qwen-VL、可灵等）成熟，市场从"文本生成"扩展至"多模态理解与自动执行"，每次交互的Token量级大幅抬升。

180万亿/日

豆包大模型日均Token调用量（2026年6月，一年多增长超10倍）

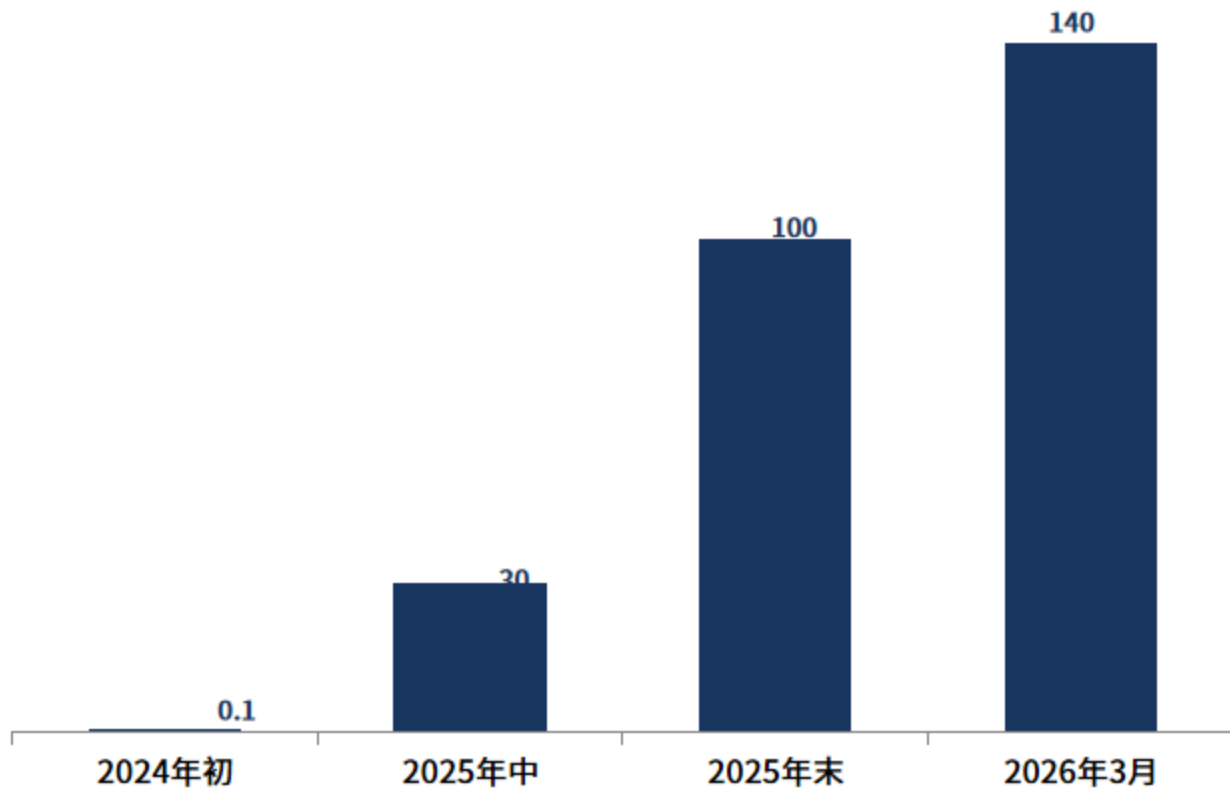
政策牵引：从"鼓励建设"到"建用并举"，政策工具箱持续迭代

时间	政策/事件	要点与影响
2022.02	"东数西算"工程全面启动	八大枢纽十大集群布局落地，算力资源跨区域配置成为国家战略
2023.10	六部门《算力基础设施高质量发展行动计划》	提出2025年算力超300 EFLOPS、智算占比35%——两项指标均大幅超额完成
2025	各地算力券/训力券密集落地	深圳首发近2亿元训力券；上海年发10亿元算力券；需求侧补贴成为新工具
2026	"算电协同"写入"十五五"规划纲要	能源与算力规划首次并轨，枢纽节点新建算力设施绿电占比要求80%以上
2026	工信部《算力互联互通行动计划》推进	2026年底建成互联互通标准体系，破解异构算力"各自为政"难题
2026	国产化比例刚性要求加码	上海2025年新建智算中心国产芯片占比超50%；北京2027年具备100%自主可控建设能力

中商解读 政策主线经历三次切换：2022-2024年"给地、给电、给指标"鼓励建设；2025年起"发券、补贴"激活需求；2026年"并轨、互联、国产化"治理存量。对Token工厂而言，绿电指标与互联互通接入资格正在取代土地与税收优惠，成为新的政策资源。

需求总览：日均Token调用量两年增长超千倍，斜率仍在变陡

中国大模型日均Token调用量（万亿）



注：2024年初口径约1,000亿（0.1万亿）。资料来源：[国家数据局，转引自虎嗅（2026-07）](#)，中商产业研究院整理

× 1,000+

两年千倍 从2024年初日均1,000亿到2026年3月突破140万亿，Token消耗成为中国AI产业最直观的扩张刻度。

6倍/年

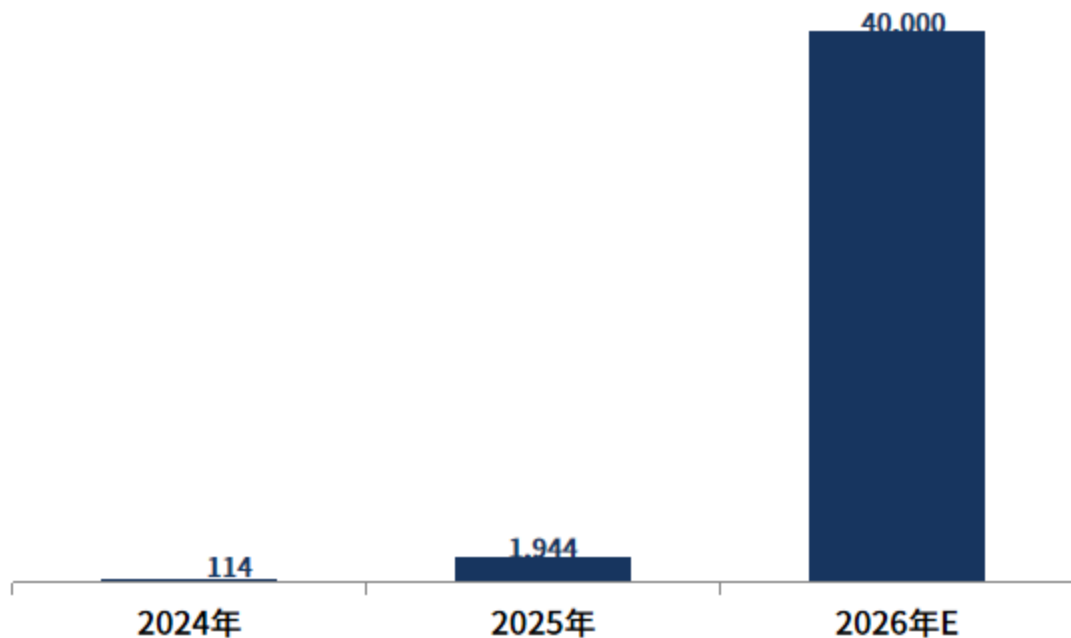
年内仍在加速 2025年1月全市场日均消耗约1.6万亿，12月底已攀升至9.6万亿，年内增长6倍（IDC）。

40,000万亿

2026年全年预测 IDC预计2026年中国MaaS市场Token消耗量约40,000万亿，较2025年再增约20倍。

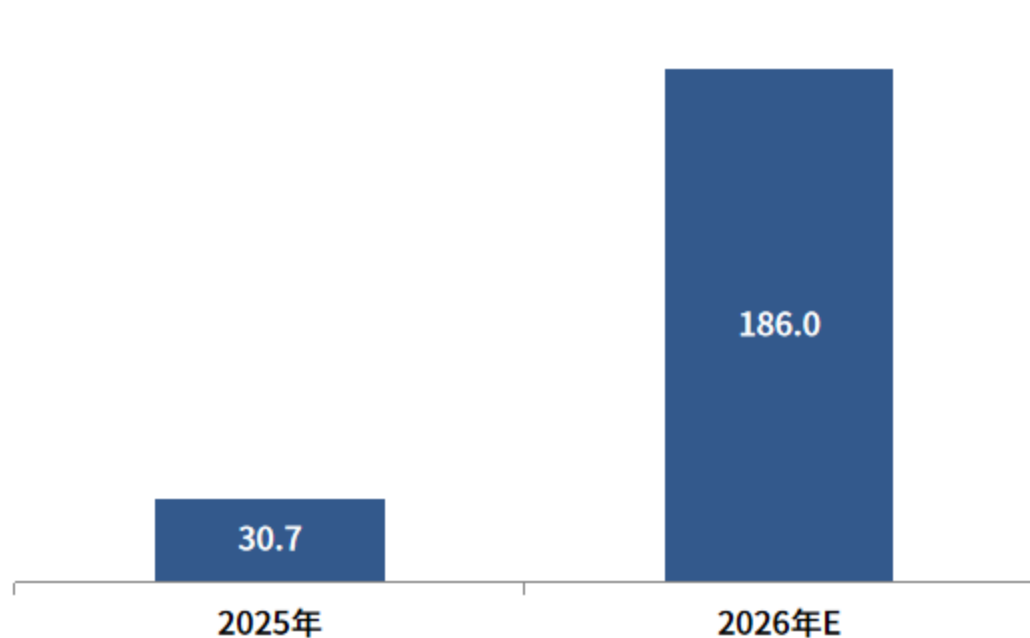
MaaS市场：调用量一年16倍，营收一年5倍，从试点走向规模商用

公有云MaaS大模型调用量（万亿Token）



资料来源：IDC《中国企业级MaaS市场最新格局分析报告》（2026-05）

公有云MaaS市场营收（亿元）



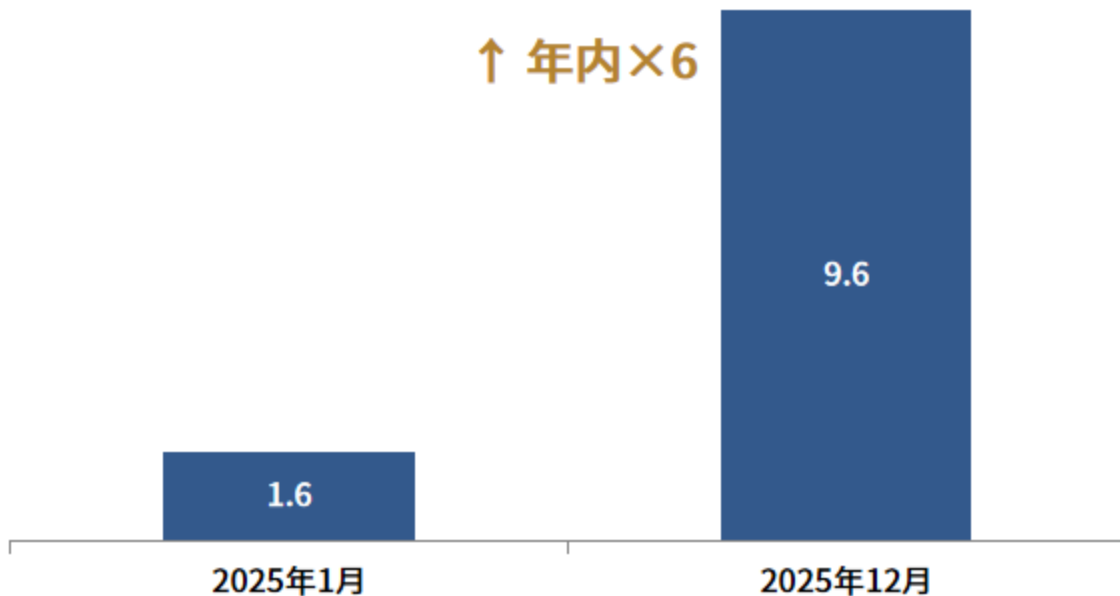
2026年营收预测约186亿元，同比增长约5倍。资料来源：IDC（2026-05），中商产业研究院整理

中商解读

调用量增速（16倍→20倍）持续跑赢营收增速（约5倍），说明Token单价仍在下行通道，市场以“量”换“价”的扩张逻辑未变。营收口径2025年仅30.7亿元——基数虽小，但意味着MaaS仍处于变现早期，谁能把调用量转化为高毛利收入，谁就能在下一阶段拉开身位。竞争焦点正从单纯价格比拼转向“价格、性能与工具链”综合能力（IDC）。

结构透视：调用量年内再翻6倍，需求主力仍在泛互联网

2025年公有云MaaS日均Token消耗（万亿）



资料来源：IDC（2026-05），中商产业研究院整理

需求结构：公有云与私有化两条线

公有云MaaS

主力客群集中在泛互联网（游戏、娱乐、教育）、智能办公、智能硬件与大消费赛道——与C端流量入口强相关，火山引擎凭豆包生态占优。

私有化部署

传统政企项目集中在智能办公、数据分析、市场营销方向；头部厂商为百度智能云、商汤、电信AI、中关村科金、创新奇智等——格局与公有云完全不同。

结构判断

当前Token消耗仍以“流量型”需求为主，金融、制造、政务等行业级纵深需求刚刚起步——这是Token工厂下一轮增长的储备市场。

价格演进：从"一元百万Token"的价格战，走向K型分化

第一阶段：价格战（2024.05—2025）

DeepSeek-V2以极低价改锚，豆包低至0.0008元/千Token，通义主力模型降价97%，智谱GLM-4-Plus降价90%——"一元钱买百万Token"。

第二阶段：K型分化（2026—）

智谱GLM-5三次涨价累计83%；DeepSeek-V4-Pro反向再降75%，输出降至0.87美元/百万Token，并试点峰谷分时定价。

第三阶段：价值定价（探索中）

豆包C端试水订阅制（68—500元/月）；头部厂商在编程、办公场景探索按任务与效果结算。

三个值得注意的市场信号

① **涨价代价立竿见影** GLM-5三次涨价后，OpenRouter全球排名从第3跌至第17——用户对Token价格高度敏感，差异化壁垒尚不足以支撑独立提价。

② **缓存价逼近"厘"级** DeepSeek缓存命中输入价低至0.025元/百万Token，比GPT-5.5便宜约34倍——高命中场景的边际成本趋近于零。

③ **分时定价引入电力逻辑** DeepSeek拟在高峰时段（9—12点、14—18点）价格翻倍——Token定价开始像电价一样反映产能的时空稀缺性。

④ **国际价差成为出海杠杆** 国产五强输出均价为国际巨头1/10—1/34，价格优势正在转化为海外开发者生态与推理出海订单。

中商解读

Token定价正在复刻大宗商品市场的成熟路径：拼不过成本的退场，活下来的人赚微利。低价不是终点而是筛选器——只有具备芯片适配、推理引擎与调度优化能力的Token工厂，才能在"厘级"价格下保住毛利。

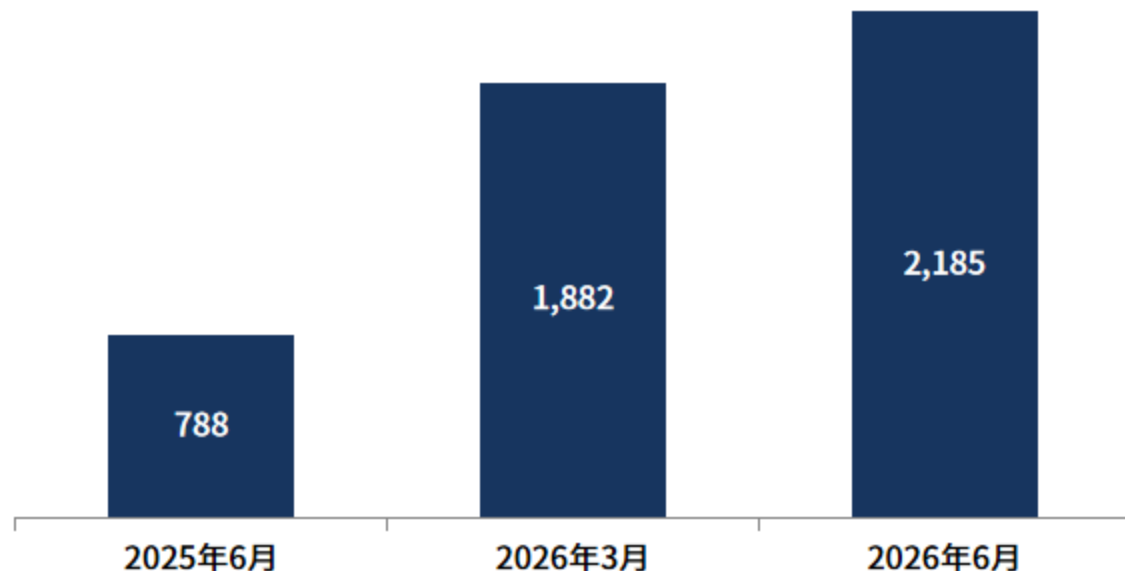
定价全景：国产主力模型输出价已进入"个位数元/百万Token"时代

厂商/模型	输入 (元/百万)	输出 (元/百万)	上下文	定价特征
DeepSeek V4-Pro	0.025 (缓存命中) / 3 (未命中)	6	128K	2026年5月永久降价75%，拟引入峰谷分时价
豆包 Pro (火山引擎)	0.8	2	128K	温和定价+C端订阅制 (68—500元/月)
通义千问 Qwen3-Plus	2	4	128K	跟随策略，百万Token免费额度
文心 ERNIE 4.0 Turbo	4	8	128K	中端定价，绑定百度云生态
智谱 GLM-4-Plus	5	15	128K	2026年GLM-5逆势提价67%-100%
Kimi K2.6 (月之暗面)	约1.1-14	约17.5	128K	分层定价，缓存命中价行业最低之一
国际对照：GPT-5.5	—	约210 (\$30)	—	国产输出均价仅为国际巨头1/10-1/34

中商解读 比价不能只看单价：输出价通常是输入的2-4倍，缓存命中率、分层档位与重试成本都会改变实际账单。对Token工厂而言，缓存架构、长上下文压缩与推理引擎效率已成为定价权的工程基础——能把缓存价打到"厘级"的厂商，实际上是在用系统效率补贴获客。

供给总览：智能算力一年近三倍，全国上架率升至71.4%

中国智能算力规模（EFLOPS，FP16口径）



资料来源：[中国算力平台监测数据](#)、国家数据局、[工信部（2026-07）](#)，中商产业研究院整理

962 EFLOPS

算力总规模（2025年6月，FP32口径），全球占比约21%，同比增速73%；智能算力占比81%

1,445万架

在用算力中心标准机架（2026年3月）；2025年6月为1,085万架，九个月净增360万架

71.4%

全国算力设施整体上架率（2026年6月，工信部）；围绕枢纽节点建成超70条算力大通道，节点间网络性能提升10%

里程碑：信通院预计2025年底智能算力突破1,000 EFlops，中国算力进入ZFlops量级。

建设热潮：一季度128个项目投资超4,300亿元，投产者不足两成

近300个

智算中心项目总数 覆盖全国29个省份；截至2024年11月已投运近150个，在建及规划近400个。

4,300亿元

单季度投资额 2025年一季度128个披露投资额的智算项目合计投资超4,300亿元。

仅16个

已投产/试运行 128个项目中仅16个投产，绝大多数处于筹建或在建状态。

谁在建设：三类主导力量

地方政府与国资平台

2024年投运项目中60%以上由政府、国资平台及运营商主导；约1/3项目规划算力超500 PFlops，普遍按区域算力目标倒推规模。

电信运营商

三大运营商智算规模合计超228 EFLOPS（2025年底），央企构成智算底座主力，哈尔滨、临港等万卡级集群相继投运。

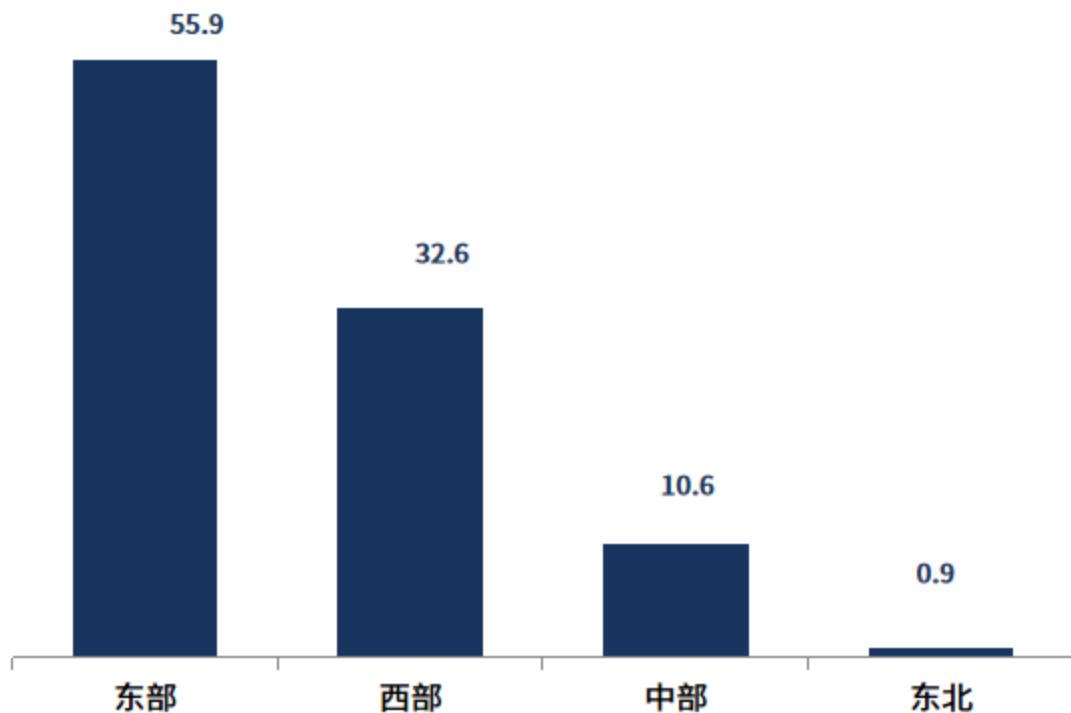
互联网与转型企业

字节、阿里等按自身模型需求自建；弘信电子等制造业转型企业绑定绿电与地方资源切入区域Token产能。

中商解读：招标热度仍在上升（2024年一季度智算相关招标172条，同比近4倍），但“重建设、轻运营”的路径依赖明显。项目从开工到形成稳定Token产能平均需要18—24个月，当前在建项目的真实投产率将决定2027年的有效供给。

区域分布：东部占56%主导需求侧，西部33%快速承接产能侧

智算规模区域占比（2026年6月，%）



资料来源：[工信部国新办发布会（2026-07）](#)，中商产业研究院整理

东部：需求与低时延产能的根据地

京津冀、长三角、粤港澳大湾区聚集了大模型厂商、Agent应用与金融级客户，承担对时延敏感的推理与训推一体需求。上海2027年智算规模目标200 EFLOPS，北京适度超前布局智算中心。

西部：绿电成本洼地，推理产能西进

甘肃、宁夏、内蒙古、贵州、成渝五大西部枢纽凭低电价、低温气候与土地空间承接大模型推理外溢。推理对时延容忍度高于训练，"东部的请求、西部的Token"成为主流供给模式。

中商解读：东部56%与西部33%的占比结构，恰是"需求在东、产能在西"的镜像。随着算力大通道网络性能提升，西部份额仍将上行；未来五年区域竞争的主线不是"谁建得多"，而是谁能把枢纽节点的绿电优势转化为可结算的Token成本优势。

八大枢纽：承载全国八成智算，成为Token产能的国家级布局

80%+

八大国家枢纽（含十大集群）已建成智算规模138.8万PFlops，占全国智算总规模（159万PFlops，2025年底）的比重超过80%

八大枢纽中五个位于西部——成渝、贵州、内蒙古、甘肃、宁夏；三个位于东部——京津冀、长三角、粤港澳大湾区。

围绕枢纽节点已建成超70条算力大通道，跨节点调度能力持续增强。

布局逻辑：训练与重算力任务向西部绿电富集区集中，低时延推理留在东部——枢纽体系实质是全国Token产能的"总排产表"。

枢纽节点定位速览

枢纽	Token产能定位
甘肃（庆阳）	增速最快、智算占比99%，主流大模型推理重镇
内蒙古（乌兰察布/和林格尔）	绿电直供+大型AIDC，大摩预测的主要受益集群
宁夏（中卫）	西部云基地，绿电框架下承接推理外溢
贵州（贵安）	老牌数据中心集群向智算升级
成渝	就地热电+边缘智算，服务西南需求
京津冀	需求核心区，张家口承接北京外溢
长三角	上海200 EFLOPS目标，金融级低时延产能
粤港澳	深圳训力券先行，应用与产能双密集

庆阳样本：一座黄土高原小城，如何长出"Token工厂"

15万PFlops

庆阳数据中心集群算力规模（截至2026年5月），
标准机架超10万架

99%

智算占比——八大枢纽中算力增速最快、智算占比
最高、消纳最充分的集群

366亿元

华庆智算产业园投资，将建国内首个国产十万卡训
推一体智算集群

谁在庆阳生产Token

阿里千问、智谱清言、Kimi、DeepSeek等主流大模型的推理请求，相当一部分在庆阳的服务器上运转；DeepSeek、MiniMax、美图等企业的海外业务也依托庆阳算力运行——庆阳产出的Token已在服务北美、欧洲与亚太用户。

为什么是庆阳：Token成本的经济账

电价优势

电力是Token成本的第一构成，甘肃绿电资源
富集、电价显著低于东部

政策组合

"算力+数据"双枢纽城市，园区整体纳入国家首
批零碳园区

气候与土地

低温降低散热成本，土地空间支撑十万卡级
集群铺开

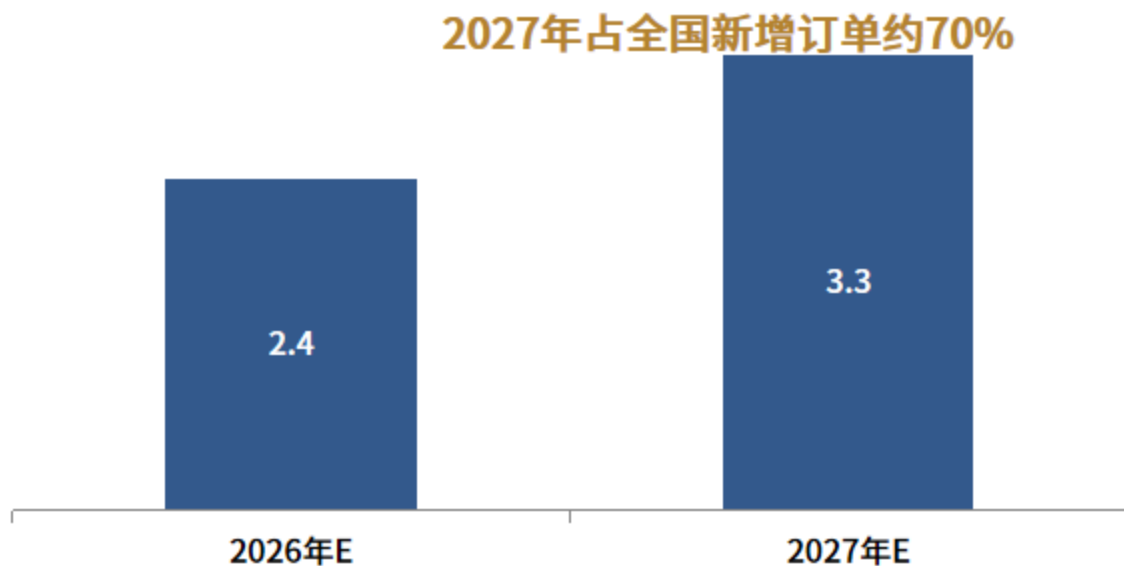
产业纵深

数字经济千亿级全产业链，从机柜到Token
运营一体布局

中商解读：庆阳的启示在于：Token工厂的选址逻辑已经从"靠近用户"变成"靠近能源与政策"。谁能锁定低价绿电与并网指标，谁就握有Token成本曲线的起点——这一逻辑正在乌兰察布、中卫、贵安复制。

产能西进：偏远地区将贡献2027年全国七成新增数据中心订单

偏远地区数据中心新增订单预测（GW）



2025—2028年偏远地区年度新增订单CAGR约60%，远高于一线城市的19%。资料来源：[摩根士丹利](#)，转引自[腾讯新闻](#)（2026-04）

乌兰察布（内蒙古枢纽）

绿电使用率超84.5%，风光存量2GW；和林格尔首期数千P智算2026年内开工、2027年投产，承接京津冀推理外溢。

中卫（宁夏枢纽）

老牌"西部云基地"，以绿电供货框架绑定本地云基地，与乌兰察布并列为大摩预测的两大受益集群。

贵安与成渝（贵州、成渝枢纽）

贵安从传统IDC集群向智算升级；成渝以"就地热电+边缘智算"服务西南本地需求，形成"源网荷储算"一体化探索。

中商解读

西部Token产能的胜负手在"消纳"而非"建设"：庆阳之所以领跑，是因为先有大模型推理订单落地。对西部节点而言，绑定头部模型厂商的推理产能，比再建十万卡更重要。

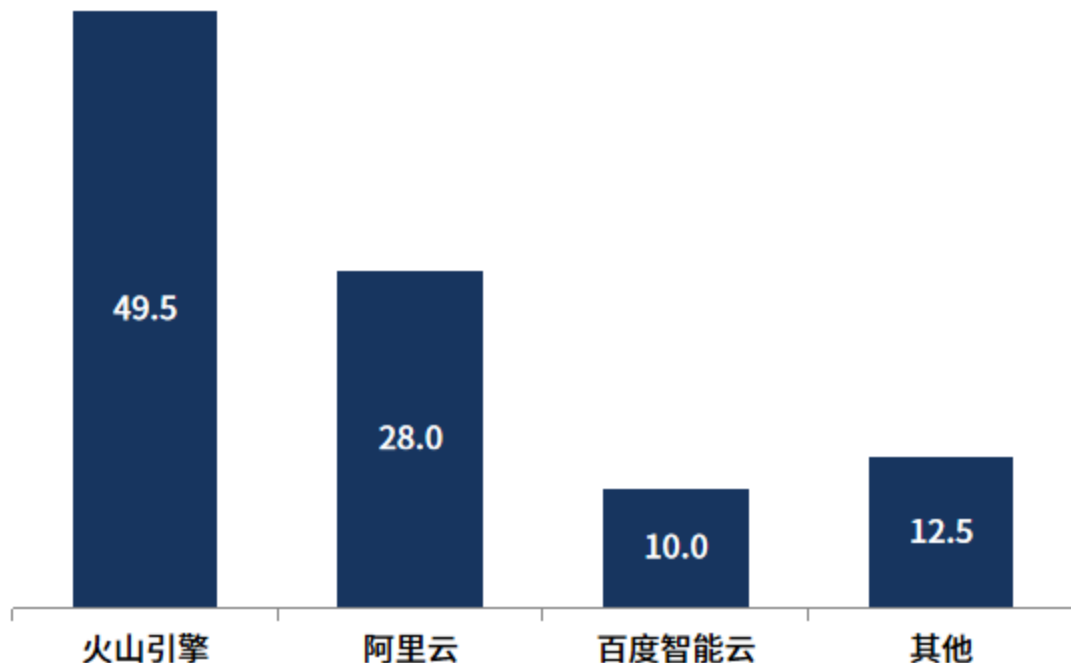
五类玩家：从"卖资源"到"卖Token"，站位决定生死

玩家类型	代表厂商	核心打法	2026年关键坐标
云厂商MaaS	火山引擎、阿里云、百度智能云	自研模型+云基础设施一体，按Token计费主平台	三家合计占公有云MaaS调用量近九成
电信运营商	中国移动、中国电信、中国联通	央企智算底座，网算一体、属地化交付	智算规模合计超228 EFLOPS（2025年底）
第三方推理平台	硅基流动、无问芯穹、九章云极	芯片适配+引擎调优封装为标准API	硅基流动跻身MaaS调用量前五
芯片厂商生态	华为昇腾、平头哥、寒武纪、海光	以超节点与集群方案绑定下游产能	国产加速卡份额41%，昇腾出货81.2万张
算力运营商	弘信电子、云赛智联等转型企业	绑定绿电与地方资源，建设区域Token产能	庆阳等西部枢纽规划数万P级绿色算力项目

中商解读 五类玩家掌握的资源禀赋不同：云厂商有模型与流量，运营商有网络与政企渠道，第三方平台有调度与适配技术，芯片厂商有硬件底座，算力运营商有绿电与地方资源。Token工厂的竞争，本质是五类禀赋的组合效率竞争。

云厂商：火山引擎一家占半，"模型+流量"模式暂时领跑

公有云MaaS调用量份额（2025年，%）



前三家合计近九成；"其他"含硅基流动、移动云等。营收口径前五为火山、阿里、百度、智谱、移动云。资料来源：[IDC \(2026-05\)](#)

火山引擎：MaaS时代最大赢家

份额从2024H2的16%跃升至49.5%。豆包日均调用量180万亿（2026年6月），C端订阅制试水付费。字节内部对云的战略定位从一开始就是"AI云"——先锁定Token赛道。

阿里云与百度智能云：基本盘稳固

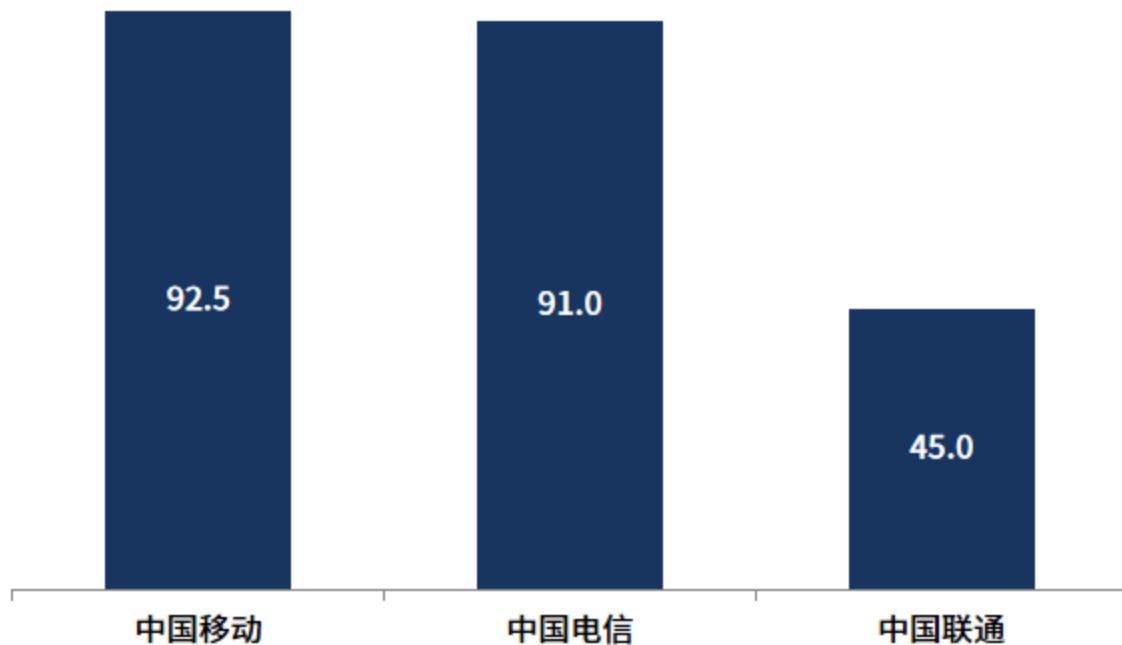
阿里云凭Qwen开源生态与企业客户存量守住28%；百度智能云在私有化部署市场居头部，文心绑定云生态。两家均在加大AI原生基础设施投入。

腾讯云的警示

MaaS调用量与营收双榜均跌出前三，公有云IaaS份额降至8%排名第五——传统云的资源优势若不能转化为模型与Token运营能力，在MaaS时代会被快速边缘化。

电信运营商：228 EFLOPS央企底座，从"管道"转向"产能运营商"

三大运营商智算规模（EFLOPS，2025年底）



三家合计超228 EFLOPS。资料来源：运营商公告（2026-03），中商产业研究院整理

中国移动

哈尔滨智算中心部署1.8万张国产AI加速卡，实现100%国产化；移动云跻身MaaS调用量与营收双榜前五。

中国电信

上海临港投运全国首个国产单池万卡液冷算力集群，"魔方"架构实现万卡高速互联，支撑万亿参数模型训练。

中国联通

呼和浩特超大规模智算中心落地首期国产算力集群，聚焦北方枢纽节点布局。

中商解读：运营商的独特价值在于网络+算力+属地化政企渠道三位一体。在互联互通标准落地后，运营商最有可能成为全国Token产能的"总调度"之一；短板是模型与推理优化能力，需通过合作补齐。

第三方推理平台：把"调优能力"做成生意，跻身调用量前五

硅基流动 (SiliconFlow)

以芯片适配与推理引擎调优为核心，把数十种主流模型封装为标准API按量付费；跻身2025年公有云MaaS调用量前五（IDC），是唯一进入头部阵营的独立第三方平台。2026年6月完成超20亿元B轮融资。

无问芯穹 (Infinigence AI)

定位"大模型操作系统"，主打异构算力池化与跨芯片混部调度，把不同厂商的芯片变成统一算力池对外服务——直击国产算力"碎片化"痛点。

九章云极 (DataCanvas)

提出"1度算力=312 TFLOPS运行1小时"的标准化计量单位，推动算力像电力一样统一计价；从AI平台软件延伸至智算中心运营。

为什么第三方平台能活下来

① **芯片中立** 不被任何芯片或模型厂商绑定，能做跨厂商异构调度——这是云厂商和芯片厂都做不到的站位。

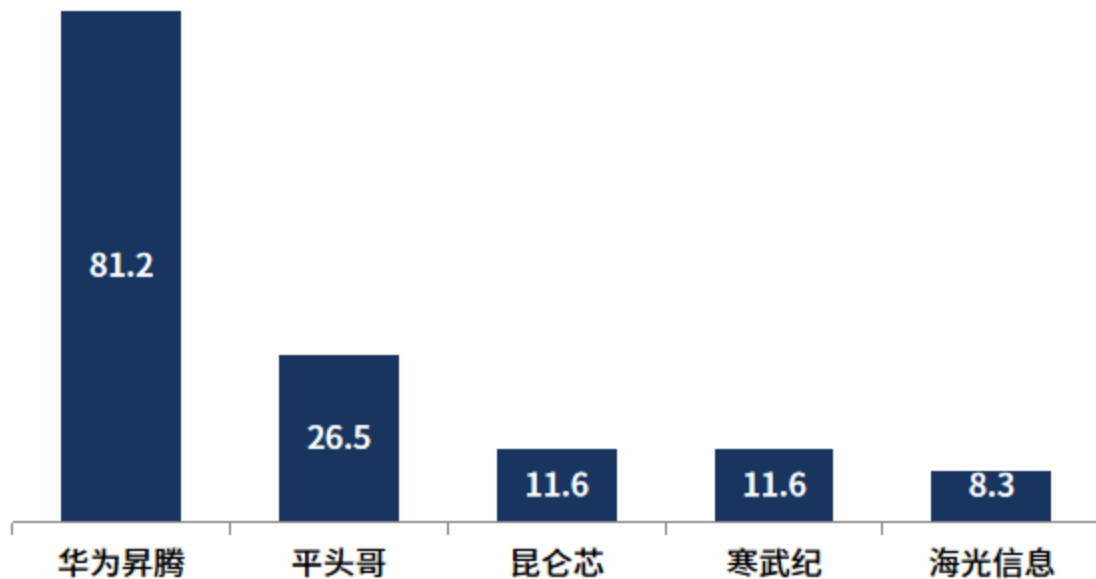
② **效率差价** 同一批卡，经过引擎调优后Token产出可提升30%以上，效率提升直接转化为毛利。

③ **盘活存量** 大量地方智算中心缺运营能力，第三方平台以托管、联营方式切入，轻资产扩张。

中商解读：2026-2028年是第三方平台的窗口期；长期看，其归宿要么被云厂商/运营商收编，要么成为独立的"Token交易所"。

国产AI芯片：份额从不足15%升至41%，昇腾一家拿走半壁江山

中国云端AI加速卡出货量·国产厂商前五（2025年，万张）



2025年中国云端AI加速卡总出货约400万张：国产165万张（41%）、英伟达约220万张（55%）。资料来源：[IDC《2025年度中国云端AI加速器市场报告》](#)，中商产业研究院整理

三年跃迁：41% vs 不足15%

2023年国产加速卡份额尚不足15%，2025年已达41%。昇腾以81.2万张独占国产出货约49%，等于第2至第5名之和。

英伟达份额从约七成退至55%

出口管制下合规供给受限，H20等特供版难以支撑需求；退出的份额几乎全部被国产芯片承接，国产替代由"可选项"变为"必选项"。

格局判断：一超多强，场景分层

昇腾凭全栈生态领跑训练与政企市场；平头哥依托阿里云推理场景放量；昆仑芯绑定百度、寒武纪卡位互联网推理、海光深耕信创。Token工厂的芯片选型已进入"场景定芯片"阶段。

生态突围：从单卡性能到"超节点+软件栈"的体系竞争

昇腾：全栈对标英伟达

昇腾950PR：FP4算力1.56 PFLOPS，搭载112GB自研HBM；Atlas 350超节点推理性能达H20的2.87倍。CloudMatrix 384超节点已部署超300套；DeepSeek V4训练从CUDA迁移至CANN，标志国产软件栈通过最严苛验证。生态伙伴超3000家，2026年出货目标150-200万颗。

上市公司：业绩兑现拐点

寒武纪2025年营收64.97亿元（+453%），首次年度盈利；2026Q1营收28.85亿元（+160%）。同期海光信息40.34亿元（+68%）、摩尔线程7.38亿元（+155%）。芯片公司从"讲故事"进入"交报表"阶段。

资本与空间：3440亿入场

国家集成电路产业投资基金三期注册资本3440亿元，重点投向算力芯片与先进制造。摩根士丹利预计：2030年中国AI芯片市场达670亿美元，国产份额约76%——芯片自主是Token工厂成本主权的前提。

对Token工厂的三点含义

采购成本下行

多供应商竞争使算力硬件进入降价通道，Token单位成本随芯片BOM成本同步走低。

适配能力成壁垒

异构芯片池化与模型迁移工程能力，决定Token工厂能否吃到国产芯片的成本红利。

政策采购绑定

多地新建智算中心设定国产化比例门槛，国产芯片成为拿项目、拿补贴的入场券。

利用率之困：标称算力与实际产出之间隔着两个"打折系数"

算力生产力的形成公式（中商产业研究院测算框架）



腾讯云副总裁张晋（MWC26上海）：行业GPU平均利用率不足30%。效能为集群工程水平、满载为商业运营水平——前者靠技术，后者靠订单。

四个口径的利用率实测（2025年）

统计口径	利用率	来源
行业GPU平均利用率	不足30%	腾讯云（MWC26上海）
信通院：用算力/共算力	36.8%	中国信通院（2025）
机架级平均利用率	20%-30%	行业调研
企业自建GPU利用率	不足10%	行业调研

信通院口径的对照

2025年全国智算中心对外"共算力"约38亿卡时，实际"用算力"约14亿卡时——近三分之二的供给能力没有找到买家。

利用率不足直接摊薄毛利：一座万卡集群年固定成本以亿元计，负载每降10个百分点，Token单位成本显著抬升。

结构性错配：为训练峰值建的厂，撞上潮汐式的推理需求

错配一：训练供给 × 推理需求

2023-2024年建设的智算中心多按大模型训练的峰值需求设计（万卡互联、高带宽组网）；而2025年以来的真实增量是推理——潮汐式、碎片化、对时延敏感。训练集群改跑推理，组网与显存配置并不经济。

错配二：通用供给 × 专用需求

部分国产芯片因软件生态不成熟、模型适配周期长，实际可用性大打折扣——据36氪调查，个别国产智算集群芯片闲置率高达80%。“有卡无栈、有栈无模”的链条断裂普遍存在。

错配三：区域供给 × 时延需求

西部算力成本最优，但推理业务要求贴近用户——东部需求无法全部西迁，“东数西算”在训练端顺畅、在推理端打折。

建设进度的另一面（2025年调研）

16 / 128

已统计智算中心项目中，已投运数量 / 总数

80%

部分国产智算集群的芯片闲置率（36氪调查）

错配的财务后果正在显化：

- 大量2024年开工的项目至今未投运，形成“半拉子”智算中心；
- 已投运项目为抢订单低价出租裸算力，陷入“越忙越亏”循环；
- 地方政府主导的园区型项目，招商承诺的“大模型企业入驻”多未兑现。

中商产业研究院判断：错配的根本原因是把Token工厂当“房地产”建——先盖楼、后招商；而正确的顺序是先锁定Token买方（模型厂/行业客户），再按需定建。未来两年，“以销定产”将成为新项目审批的隐形门槛。

盈利分化：涨价保不住排名，降价反而赢下规模

涨价派：智谱GLM的三次提价实验

2025年下半年至2026年，智谱对GLM系列API三次提价，累计幅度约83%。结果是OpenRouter全球调用量排名从第3位滑落至第17位——开发者用脚投票，Token消费对价格高度敏感。

教训：在模型能力趋同的阶段，涨价等于把份额拱手让人。

降价派：DeepSeek的逆向操作

DeepSeek V4-Pro错峰时段价格再降75%，夜间缓存命中价低至"厘级"。配合峰谷分时定价，把闲置时段的算力变成纯增量收入——本质是把航空业的收益管理引入Token工厂。

结果：以价换量锁定开发者生态，日均调用量持续居国产模型首位。

结论：Token定价正在"大宗商品化"

标准化Token没有定价权——拼不过成本的退场，活下来的赚微利。

厂商策略	价格动作	市场结果
智谱（涨价）	累计提价约83%	排名3→17
DeepSeek（降价）	错峰再降75%	调用量居首
火山引擎（低价）	豆包0.8元/百万	份额49.5%
百度（免费）	ERNIE部分免费	守私有化市场

资料来源：OpenRouter（2026-05）、各厂商定价页，中商产业研究院整理

利润从哪里来？三个例外

① 缓存/错峰运营：把电力峰谷差转成利润；② 私有化交付：按项目制收费，避开公开市场价格战；③ 出海定价：同一Token卖给海外客户可获3-5倍价差。

中国Token工厂产业链：五层结构，价值向上下游两端集中



价值链判断：两端厚、中间薄

上游能源与芯片掌握成本主权，下游模型与应用掌握定价主权，中间的基础设施层在产能过剩中利润最薄。中国市场的特殊性在于：能源层（绿电成本）与模型层（开源生态）两端均为全球最强，芯片层正在快速补齐——这决定了中国Token工厂的全球成本竞争力。

各环节代表主体

能源：国家能源集团/华能/三峡 | 芯片系统：华为/寒武纪/海光/中兴 | 基础设施：三大运营商/云厂商/第三方智算 | 模型：DeepSeek/字节/阿里/智谱 | 应用：各行业Agent与SaaS

商业模式进化：从卖机柜到卖结果，四级跳正在进行



阶段	计费单位	毛利特征	中国代表
1.0 卖资源	元/机柜·月	稳定但天花板低	传统IDC运营商
2.0 卖算力	元/卡时	价格战重灾区	多数智算中心
3.0 卖Token	元/百万Token	规模定生死	火山/阿里/硅基流动
4.0 卖结果	元/任务·分成	与客户价值绑定	头部Agent厂商

中商产业研究院判断

中国Token工厂的主体收入将在2026-2027年完成2.0→3.0切换：裸算力租赁占比持续萎缩，MaaS与模型订阅成为主要现金流。

4.0"卖结果"已在客服、编程、营销等场景出现按单计费雏形，但大规模落地要等Agent可靠性跨过商用阈值。

机制创新：算力券与计量革命，正在修补供需断裂

地方"券类"补贴政策一览（2025-2026）

地区/机构	政策内容	规模
深圳	训力券+语料券+模型券三券齐发	训力券近2亿元；语料券5000万/年；模型券1亿/年
上海	算力券补贴智能算力租用	10亿元/年
中国银行	券贷联动·算力贷	80亿元授信·覆盖38家企业
杭州/成都等	算力券+首购首用支持	多城跟进，单城千万至亿级

资料来源：深圳市科技创新局（2025-03）、上海市经信委、中国银行（2025-12）、各地政府公开文件，中商产业研究院整理

券类补贴的共同指向：把补贴从"建设端"移到"使用端"——按真实Token调用量与训推任务兑付，让低效产能拿不到钱，让用算企业降成本。

计量革命：1度算力

九章云极提出"1度算力=312 TFLOPS·h"的标准化计量单位，把异构算力折算为可交易、可结算的标准品——Token工厂需要自己的"千瓦时"。

定价革命：峰谷分时

DeepSeek在低谷时段降价75%，缓存命中价低至厘级——把电力市场的峰谷定价引入Token销售，闲置产能变增量收入，推动行业从"一口价"走向"分时电价式"定价。

交易革命：算力互联互通

全国一体化算力网试点推进跨主体调度与结算，信通院牵头算力互联互通标准（2026年底成形）——算力交易所从概念走向实操。

地方竞赛：国产化比例成为硬指标，绿电与自主可控双达标

重点省市智算政策目标（2025-2030）

地区	时间目标	国产化/自主可控要求	产业规模目标
上海	2025年起	新建智算中心国产芯片占比超50%	2027年智算云产业规模2000亿元、智算规模200 EFLOPS、自主可控占比70%
北京	2027年	智算中心100%自主可控	建设全国领先的自主可控智算集群
广东	2025年起	新增智算设施国产化率70%	打造粤港澳大湾区智能算力枢纽
江苏	2030年	国产芯片占比70%、国产存储占比90%	建成全国一体化算力网核心节点
成都	2027年	优先采用国产软硬件	人工智能产业规模1300亿元

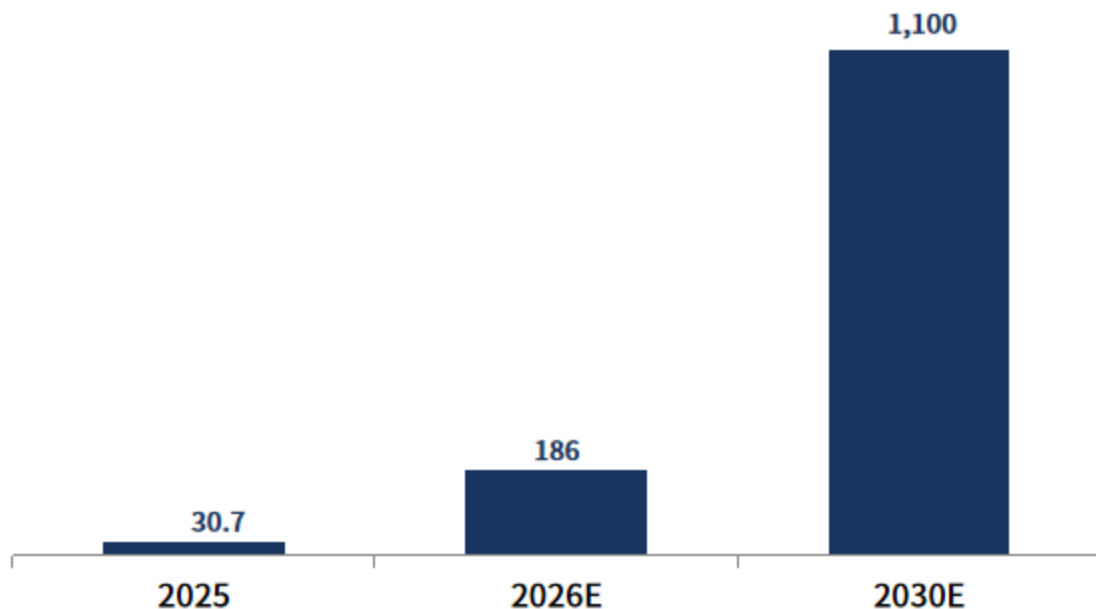
资料来源：上海市《智算云产业创新发展实施意见》、北京市《加快建设具有全球影响力的人工智能创新策源地实施方案》、广东省及江苏省相关政策文件（2024-2026），中商产业研究院整理

国家层面的两条主线

算电协同："十五五"规划方向要求枢纽节点绿电占比达80%，算力布局与新能源基地绑定；互联互通：全国一体化算力网标准体系2026年底基本成形，跨主体、跨区域的算力调度与结算将有统一规则。

市场预测：MaaS收入两年六倍，2030年前后迈向千亿级

中国公有云MaaS市场规模（亿元）



2025/2026E为IDC统计与预测；2030E为中商产业研究院测算（“千亿级”区间中值），假设：Token单价年降幅30%–40%、调用量年增3–5倍。资料来源：IDC（2026-05）、中商产业研究院测算

三个关键预测

2886 EFLOPS

2030年中国智能算力规模预测

1:3.5

信通院测算：算力每投入1元，带动3–4元经济产出

中商产业研究院核心测算

基准判断：2026年中国MaaS市场约186亿元，其中Token按量计费占比过半；2028年前后整体迈向千亿级。

结构判断：推理Token收入占比将从当前约六成升至2030年的八成以上；出海Token（服务海外客户的推理产能）成为第二增长曲线，贡献增量收入的15%–20%。

边界提示：若模型能力或Agent落地不及预期，价格降幅无法被调用量增长对冲，市场规模可能下修三至四成。

五大趋势：未来三年中国Token工厂的确定性走向

01 推理主导：Token工厂的定义权移交

训练算力建设见顶回落，推理成为绝对主力负载；新建产能按推理SLA（时延、并发、成本）而非训练峰值设计，“训推一体”让位于“推理优先”。

02 每瓦Token成为行业KPI

考核指标从“P数/机架数”切换为“每瓦Token产出、每元Token成本”；绿电占比80%的硬约束使能效直接等于成本竞争力。

03 Token产能出海：从卖设备到卖“电”式服务

推理出海订单快速增长——海外客户调用部署在中国西部绿电集群上的国产模型，Token像电力一样跨境“输送”；3-5倍国际价差是核心驱动，数据合规规则将决定其天花板。

04 计费跃迁：按价值分成的苗头出现

Agent按任务计费、按效果分成的合约在客服、编程场景落地；Token从“计量商品”升级为“价值载体”，头部模型厂开始抽取应用层收入分成。

05 整合出清：2026-2027年并购窗口开启

利用率不足30%的存量产能将被头部运营商低价整合；地方国资平台从“建设者”转向“持股者”，行业集中度快速提升。

行动建议：四类主体各自的卡位与避险清单

对产业投资者

避开同质化重资产，沿三条线布局：① 绿电+算电协同资源（西部并网指标）；② 异构调度与推理优化软件（把30%利用率做上去的公司）；③ 出海通道（海外合规实体+跨境专线）。估值锚：按Token吞吐量而非P数定价。

对地方政府

把招商筹码从土地优惠换成“绿电指标+Token产能+券类补贴”组合；补贴只兑付真实调用量，防止骗补式空转；存量低效项目尽早引入专业运营商托管；以国产化比例门槛换取芯片企业落地。

对算力运营商

从“建厂思维”转向“电厂思维”：签长期Token购售协议锁定基荷，用峰谷分时消化波动；接入全国互联互通调度体系提升负载；国产化适配团队是拿单前提；停止为无明确买方的训练峰值扩容。

对用算企业

当前是买方市场：用长约锁定低价Token产能，用多云/多模型架构避免单一供应商绑定；优先使用缓存、错峰、小模型蒸馏三件套降本；把“每百元Token产出”纳入AI项目的立项核算。

结语：把绿电与模型优势，炼成Token时代的中国产能

Token工厂是继高铁、光伏、新能源车之后，中国又一个具备"全产业链+规模成本"双重优势的赛道：西部绿电提供全球最便宜的算力能源，开源模型生态提供全球最便宜的Token供给，41%的国产芯片份额正在补上最后一块短板。

但产能过剩的阴影同样真实——利用率不足30%、大量项目未投运、价格战压缩毛利。未来三年是"以销定产、以效定价"的出清期。

本报告的六个关键数字

- 140万亿：2026年中国日均Token调用量量级（中商测算）
- 2185 EFLOPS：2026年中国智能算力总规模
- <30%：行业GPU平均利用率
- 49.5%：火山引擎公有云MaaS调用份额
- 41%：国产AI加速卡出货份额（2025年）
- 186亿元：2026年中国MaaS市场规模预测

关于中商产业研究院

中商产业研究院是中国产业研究与咨询机构，长期跟踪数字经济、人工智能与新型基础设施产业，为政府部门、产业园区与企业提供产业规划、市场研究与投资咨询服务。

本白皮书基于公开数据与独立研究撰写，版权归中商产业研究院所有。

免责声明

本报告所引用数据均来自国家发改委、工业和信息化部、中国信通院、IDC、摩根士丹利、OpenRouter等机构的公开发布及上市公司披露，中商产业研究院力求但不保证信息的准确性与完整性；预测性内容仅代表研究观点，不构成任何投资建议。任何机构或个人使用本报告内容，应自行核实并承担相应责任。