中邮证券
CHINA POST SECURITIES

证券研究报告

底座算力跃迁到token工厂的新机会

行业投资评级：强于大市|维持

孙业亮/刘聪颖

中邮证券研究所　计算机团队

中邮证券

发布时间：2026-06-16

投资要点

- **算力需求爆发式增长，全球竞争日益激烈：**随着人工智能、大数据、工业互联网等数字化技术规模化应用，全球算力需求高速增长；据中国信通院数据，截至2024年底，全球通算规模达628EFlops，同比增加14.0%，智算规模达5693EFlops，同比增加64.7%；据IDC预测，2025年全球人工智能服务器市场规模为1587亿美元，2028年有望达到2227亿美元。
- **Token经济正在开展一场智能定价革命：**根据全球最大AI模型API聚合平台OpenRouter最新数据显示，3月16日至22日，全球AI大模型总Token调用量为20.4万亿，仅中国就达7.359万亿，占全球的36%，Token是其关键的主角之一。与传统算力租赁模式不同，“Token工厂”交付的不是算力时间，而是经深度优化后产出的智能单位Token，这不仅是技术升级，更是商业模式与价值分配的重构。
- **投资建议：**建议关注**1）AI基础设施：**海光信息、寒武纪、中科曙光、浪潮信息、禾盛新材、协创数据等；**2）AIDC厂商：**润泽科技、东阳光、光环新网、数据港、奥飞数据、大位科技、豫能控股、世纪互联等；**3）算力租赁厂商：**宏景科技、润建股份、东方国信、首都在线等；**4）运营商与云计算：**中国电信、中国移动、中国联通、网宿科技、彩讯股份、优刻得等。
- **风险提示：**行业竞争加剧风险；下游应用需求不及预期风险；Token价格大幅波动风险等。

目录

一 大模型驱动算力需求指数级增长

二 Token工厂重构算力商业模式

三 产业链与相关上市公司

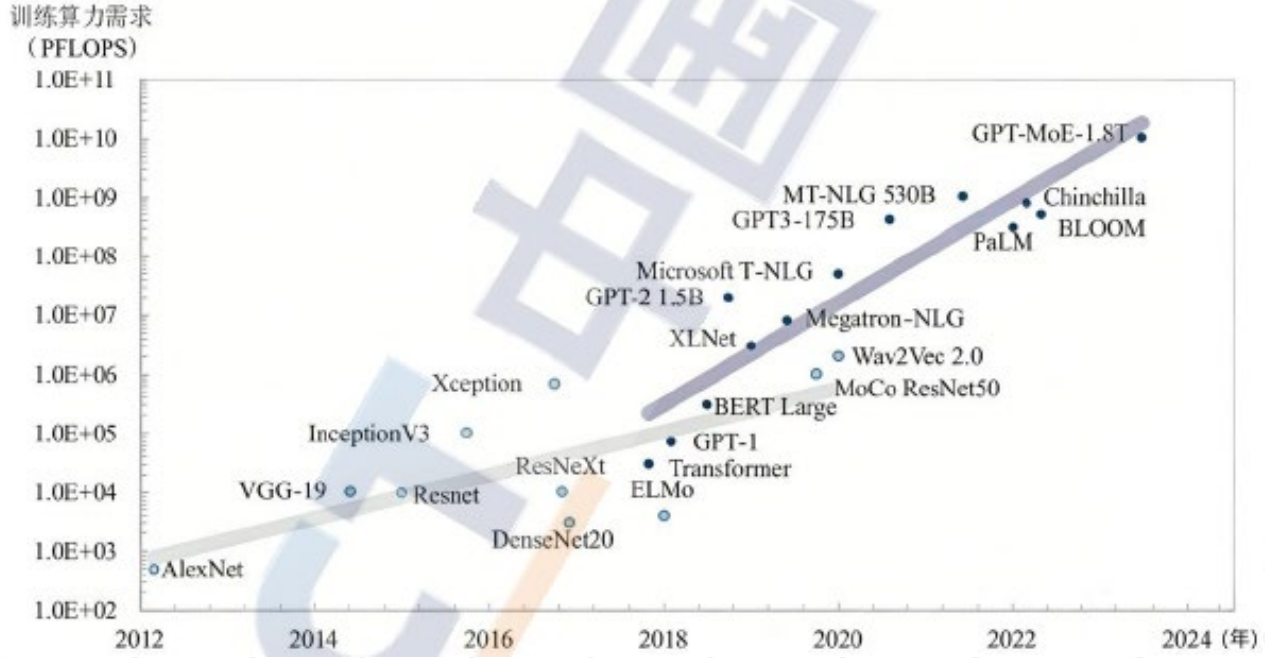
四 风险提示

一 大模型驱动算力需求指数级增长

大模型驱动算力需求指数级增长

- **AI高速发展推动各行业数智化转型，全球算力需求高速增长。**大模型正呈现由训练主导向训练与推理并重，由中心集聚向分布协同演进，训练与推理环节对算力双重刚需，显著推动智能算力基础设施踏上快速发展轨道；此外，用户对智能算力服务的诉求也由获取底层资源转向获取任务能力、结果交付与普惠化服务。据中国信通院数据，截至2024年底，全球通算规模达628 EFlops，同比增加14.0%；智算规模达5693 EFlops，同比增加64.7%。

图表1：人工智能模型算力需求持续攀升



资料来源：中国信通院《智能算力服务研究报告（2026年）》，中邮证券研究所

大模型驱动算力需求指数级增长

- **智能体带来AI应用革命，算力需求的重点正从训练侧逐步转向推理侧。**智能体需要持续感知环境、反复调用工具、不断生成结果，并与用户进行多轮交互，以Open Claw为代表的各类智能体应用加速涌现，推动AI产业迈向“应用落地”与“规模化服务”新阶段。据IDC数据，中国企业活跃智能体数量将在2031年突破3.5亿规模，年复合增长率达到135%以上，这一增速将领先全球主要市场；同时，由于智能体任务执行密度的增长和任务复杂度的提升，也将带来智能体Token消耗年均超30倍的指数级跃升。

图表2：到2031年中国企业将拥有3.5亿个活跃智能体

资料来源：IDC咨询公众号，中邮证券研究所

大模型驱动算力需求指数级增长

- **政策层面已明确路线图与时间表：**国务院印发的《关于深入实施人工智能+行动的意见》提出，到2027年新一代智能终端、智能体等应用普及率超过70%，人工智能在公共治理中的作用明显增强，到2030年提升至90%以上；工信部等八部门联合印发的《“人工智能+制造”专项行动实施意见》，明确到2027年将推动3-5个通用大模型在制造业深度应用，培育1000个高水平工业智能体，打造100个工业领域高质量数据集，推广500个典型应用场景，培育2-3家具有全球影响力的生态主导型企业和一批专精特新中小企业；地方政府同步跟进部署，面向制造业、金融、政务、医疗等重点领域，加快智能体产业布局。

图表3：到2027年智能体等应用普及率超过70%

资料来源：中国政府网，中邮证券研究所

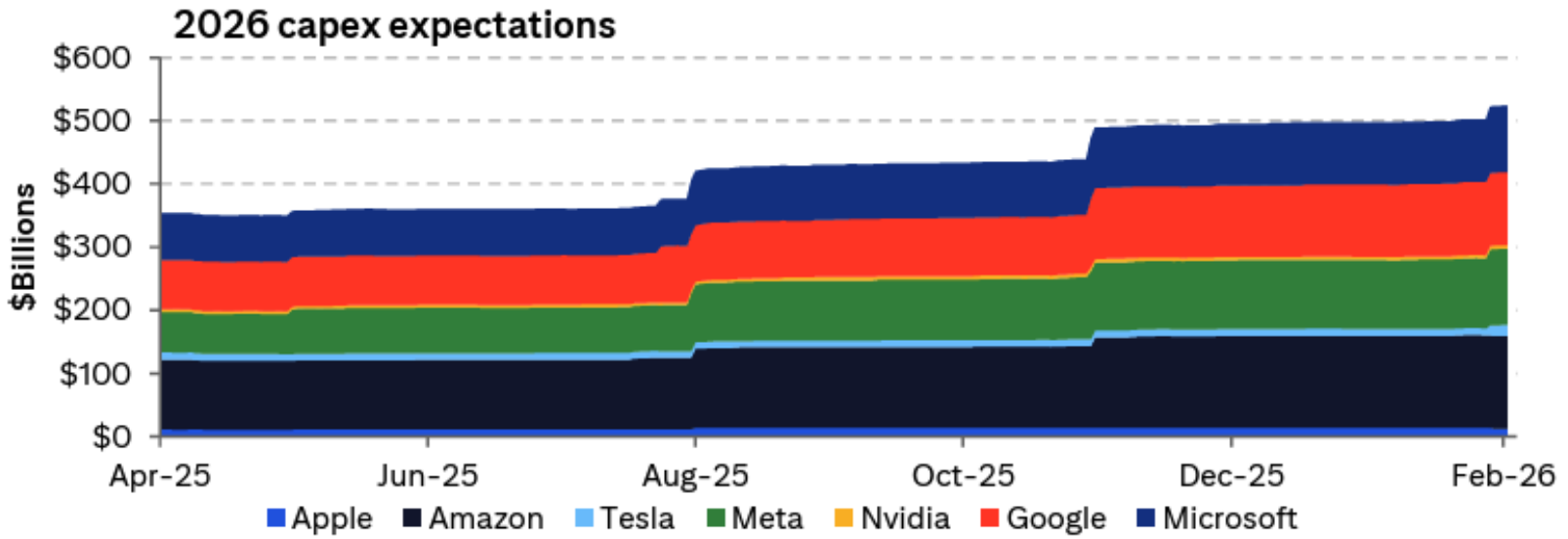
图表4：2027年推动3-5个通用大模型在制造业深度应用

资料来源：工信部官网，中邮证券研究所

大模型驱动算力需求指数级增长

- **海外AI资本开支快速增长：**微软、谷歌等美国科技巨头纷纷加大资本开支，特别是在GPU采购、数据中心、电力上的投入，大模型训练成本极高，导致只有头部云服务商和AI公司能负担，早期积累的模型质量、用户数据和算力规模会形成“马太效应”，谁在大模型和应用生态中占领先机，就会获得巨大战略优势。
- **中国正快速实现追赶：**腾讯表示，今年下半年AI相关的资本支出会进一步增加；阿里表示，未来AI基建相关投入资金会远远超过3800亿；字节计划2026年资本支出将超过2000亿元人民币，较此前的初步计划增加了25%。

图表5：MAG7资本支出持续超预期

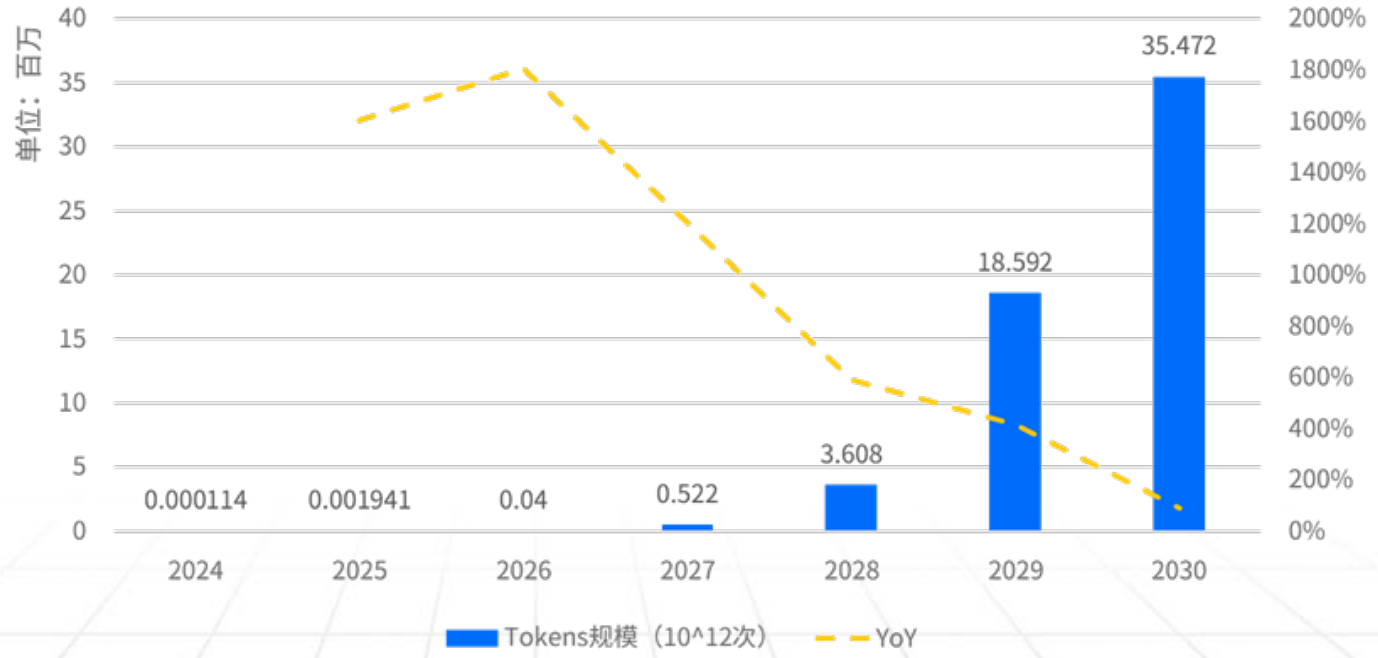


资料来源：Factset，中邮证券研究所

大模型驱动算力需求指数级增长

- **Token消耗快速增长：**Token（词元）是连接技术供给与商业需求的“结算单位”，到2026年3月，我国日均Token调用量已超过140万亿，较2024年初增长了1000多倍，标志着AI发展已进入以推理和应用为核心的快速增长阶段；智能时代Token成为核心生产要素，伴随着智能体应用爆发，其需求呈指数级增长，预计2030年中国日均消耗Token量将达万万亿级，AI基础设施变得愈加重要。据IDC预测，2026年中国MaaS市场的Token消耗量将达到约4万万亿次，营收规模约186亿元，2024-2030年的年复合增长率约为1154.9%。

图表6：预计2026年中国MaaS市场的Token消耗量将达到约4万万亿次



资料来源：IDC咨询公众号，中邮证券研究所

二

Token工厂重构算力商业模式

Token工厂重构算力商业模式

- **“Token工厂”重塑价值产出：**作为大模型处理信息的最小单元，Token具备可计量、可定价、可交易的特性，既是AI服务的基本结算单元，也正成为数字经济的核心“能源”，“Token工厂”是高效连接起算力、模型与应用的枢纽；与传统算力租赁模式不同，“Token工厂”交付的不是算力时间，而是经深度优化后产出的智能单位Token，这不仅是技术升级，更是商业模式与价值分配的重构。未来智算中心竞争不只是芯片之争，更是软件、网络、能源、运营和生态的综合竞争。Token工厂正在把数据中心从基础设施行业推向制造业，而制造业的核心从来不是规模，而是效率。

图表7：软通动力韶关Token工厂

资料来源：软通动力公众号，中邮证券研究所

Token工厂重构算力商业模式

- **Token的经济成本可拆解为资本性开支和运营性开支两部分：**训练阶段以CAPEX主导，是一次性的重资本密集投入；推理阶段以OPEX主导，是随Token规模扩张的持续性边际成本。
- **训练阶段CAPEX主导固定成本：**训练一个千亿参数级别的通用大模型，全流程需驱动数万张高端GPU连续运转数月，总成本约1-5亿美元；成本结构中，硬件折旧与高端研发人力薪酬合计占比超95%，电力消耗不足5%；单位Token摊销额随Token总产出增加而持续递减，构成Token规模经济的根本来源。
- **推理阶段OPEX主导边际成本：**AI大模型一旦部署，每产生新Token，都需消耗固定的算力与电力，形成与Token调用量高度正相关的可变成本，其中电力成本通常占到OPEX总额的60%-70%，是最大单项开支，电力作为Token推理阶段主体边际成本的地位凸显。

图表8：AI大模型的Token成本分为CAPEX和OPEX两部分

成本类型	核心内容	摊销方式
CAPEX 资本性支出	大模型预训练（数据清洗、架构搜索、全量训练）、AI算力硬件（GPU/NPU）、智算中心基建（土建、供电+液冷系统）	模型资产按2-3年摊销； 硬件按3-5年直线折旧
OPEX 运营性支出	推理服务算力消耗、电力消耗、系统散热、人力成本（算法工程、运维）、网络宽带	随Token调用量线性发生的持续性开支

资料来源：21世纪跨国企业观察公众号，中邮证券研究所

Token工厂重构算力商业模式

- **Token的定价锚逐渐从GPU→能源→人才方向转移：**Token的生产成本由芯片、电力、数据与人才四大要素构成，在不同阶段，决定Token价格底线的关键要素并不相同，定价权会沿着技术演进的节奏在四大要素之间逐步让渡。
- **当前阶段芯片是主锚：**高端GPU供不应求，芯片的可获取性直接决定了Token的供给量与价格；
- **中期来看电力将成为硬性约束：**电力受物理定律限制，随着AI数据中心能耗激增，能源成本将成为不可压缩的底线；
- **长期来看人才与知识密度将主导定价：**大模型能力越强、在高端场景中为用户带来的边际收益越高，Token的溢价空间就越大。

图表9：Token的定价体系在不同阶段存在差异

资料来源：交大安泰经管学院公众号，中邮证券研究所

Token工厂重构算力商业模式

- 海外三巨头OpenAI、Google、Anthropic均采用**“输入/输出双轨Token计费”**，核心依靠产品分层与生态协同盈利。
- **OpenAI用分层价格锁定全层级客户，依靠生态壁垒竞争：**GPT系列采用阶梯化价值定价，高端推理模型保持高溢价，轻量微型模型小幅让利吸引开发者；其中，旗舰GPT-5.5每百万输入Token 5美元、输出30美元，面向企业复杂智能体、长周期推理任务，赚取核心利润；GPT-4o mini轻量化模型大幅下放价格，用于吸纳C端应用、小型开发者，构建全球最大开发者生态；收入结构不单一依赖Token计费，还叠加ChatGPT订阅、企业席位、智能体会话时长、文件存储等增值服务。
- **Google进行云生态捆绑，将Token作为GCP的导流工具：**Google不单独售卖大模型API，而是将Gemini全套Token能力与Google Cloud深度打包，Token定价拥有全行业最宽价格区间，从Flash-Lite超低价版本到3.1 Pro高端模型全覆盖；低价轻量模型补贴云业务，高端推理算力绑定政企云订单，依托搜索增强、多模态原生能力，把Token消耗转化为云存储、算力集群、容器服务增值收入；Token是云计算的增值配套，AI业务目标是拉动GCP整体市占提升。
- **Anthropic坚持品质定价：**依靠超长上下文、企业级数据安全、合规能力建立品牌溢价；其中，主力Sonnet 4.6每百万输入3美元、输出15美元，旗舰Opus 4.6则定价翻倍，瞄准金融、法律、政务高敏感企业客户，不比拼调用总量，只为对数据安全、推理精度有刚性需求的企业提供服务，天然避开同质化低价竞争；此外还推出智能体会话按时长计费、分层缓存定价，跳出单纯按Token收费的单一模式。

Token工厂重构算力商业模式

- 随着MaaS与Token成为行业共识，头部厂商悉数重兵入场，竞争焦点也从规模化消耗Token，转向高质量、高效率、高价值消耗Token。
- **阿里Token Foundry：**今年3月阿里已推出Alibaba Token Hub事业群，搭建“创造-分发-应用”词元基础框架；时隔三月升级为Token Foundry事业部，定位完成全面跃迁，不再单纯标准化输出通用词元，而是根据电商、零售、本地生活、企业办公等不同场景，定制推理链路、优化词元结构、搭建专属Agent词元工作流；同样一组词元，在通用对话仅产生流量，在阿里本地生活智能体可完成商家运营、订单调度、用户服务全流程，直接带来营收增量，以此重构词元定价逻辑，对冲行业低价内卷带来的毛利压力。
- **字节：**一方面，Seed团队持续攻坚技术极限，打造视频生成、图像创作、代码编程、文本理解等全领域SOTA标杆模型，不断刷新模型能力上限；另一方面，火山引擎也将技术能力加速工具化、产品化，高效推向市场；Token价格必须与模型能力、产出价值绑定，即使单Token理论成本可能更高，但创造的经济价值要同步提升，要提升Token能力，并确保定价优势。
- **华为数字能源Token Factory：**与阿里聚焦上层模型、场景应用的Token Foundry形成鲜明对照，华为数字能源从AIDC算力基础设施维度，定义产业底层的Token Factory；传统数据中心是存储数据的电子仓库，而新一代智算中心是工业化量产词元的专属工厂，电力作为生产原料，昇腾算力集群、液冷供电、储能系统作为生产设备，持续稳定输出标准化、可计量的推理词元；华为Token Factory不绑定单一厂商大模型，是中立的算力基础设施底座，无论阿里Token Foundry、华为云盘古大模型，还是第三方企业MaaS平台，均可接入华为AIDC词元工厂获取规模化算力供给。
- **腾讯：**将原隶属于腾讯云的MaaS大模型服务平台升级为“TokenHub”，升级后的TokenHub支持通过API调用混元、DeepSeek、MiniMax等主流大模型，并提供Token Plan统一计费。

Token工厂重构算力商业模式

- **五象云谷词元（Token）工厂：**基础层，五象云谷智算中心总算力规划40000P，为Token生产提供坚实的算力底座；平台层，算力调度+模型优化+算法加速，三重技术协同发力，通过异构算卡细粒度虚拟化、队列抢占机制与自研算法加速引擎，大幅提升Token吞吐量，将单位生产成本极致压榨，降低Token单位成本；服务层，Token管理平台集成API接口、用量监控、账单结算、跨境服务等全功能模块，无缝对接国内及东盟市场需求；能力层，算力承载、集成、调度、维保再到算力贸易全栈式服务——五象云谷以完整闭环能力，让Token像水电一样按需使用、按量计费、据实结算。

图表10：五象云谷Token工厂四层架构环环相扣

资料来源：五象云谷有限公司公众号，中邮证券研究所

Token工厂重构算力商业模式

- **软通动力点亮“北京壹号Token工厂”：**6月9日，软通动力“词元（Token）工厂计划”的首个标杆示范项目——“北京壹号词元工厂”在京点亮，并同步向全球开源“词元工厂性能基准”，这是行业首次针对智能体长时运行特征建立统一的性能度量标准，标志着大模型算力供给从粗放式吞吐比拼，进入标准化、工业化的Token流水线时代。北京壹号词元工厂聚焦Agentic Serving（智能体服务）场景，通过极限工程化手段压榨硬件性能，集成前沿算力调度与KV Cache极致复用算法，以确定性的服务质量与极致的性价比，为智算时代提供确定性、高弹性的供应保障。

图表11：开源性能基准，打破“指标迷雾”

资料来源：软通动力公众号，中邮证券研究所

Token工厂重构算力商业模式

- **Token出海的比较优势：**1）中国大模型通过架构创新，用更少的资源达到了同级别的性能，以开源模式让全球开发者零门槛获取模型权重；2）中国拥有全球最大、最复杂的工业现场数据资产，工业垂类模型是别国无法复制的壁垒；3）国产芯片替代将开启中国Token产业的供应链安全自主与长期成本下降曲线；4）电力成本的结构性优势。
- **Token出海路径：**1）API是大模型的标准化调用接口，API直接出口门槛最低、速度最快，但可能面临美国制裁；2）本地化部署，在目标国境内建设或租用算力基础设施，将模型运算放在当地完成；3）主权AI合作，不依赖外国科技巨头的技术垄断，实现发展中国家Token基础设施级渗透；4）产业链嵌入，将Token服务作为中国海外工厂数字化方案的标准选项嵌入，随中国制造业同步出海。

图表12：四条差异化出海路径

出海路径	目标市场	核心策略
路径一： API直接出口	北美、欧洲、东南亚	开源建生态，API做变现； 高性价比切入，垂直场景突破
路径二： 本地化部署	东南亚、中东	海外建立或租赁算力节点，Token本地化生产与合规运营
路径三： 主权AI合作	发展中国家	政府间框架协议； 联合研发一体化
路径四： 产业链嵌入	东南亚、一带一路沿线	Token服务作为数字化标配，随中国制造业供应链同步落地海外

资料来源：21世纪跨国企业观察公众号，中邮证券研究所

三

产业链与相关上市公司

- Token工厂的产业链以国产AI芯片与算力基础设施为底层供给，其能力直接决定Token生产的速度、质量和成本。

图表13：算力中心产业链

资料来源：灼识咨询《中国算力中心行业白皮书》，中邮证券研究所

相关上市公司

- **建议关注：**
- **1）AI基础设施：**海光信息、寒武纪、中科曙光、浪潮信息、禾盛新材、协创数据等；
- **2）AIDC厂商：**润泽科技、东阳光、光环新网、数据港、奥飞数据、大位科技、豫能控股、世纪互联等；
- **3）算力租赁厂商：**宏景科技、润建股份、东方国信、首都在线等；
- **4）运营商与云计算：**中国电信、中国移动、中国联通、网宿科技、彩讯股份、优刻得等。

四 风险提示

风险提示

- 行业竞争加剧风险；
- 下游应用需求不及预期风险；
- Token价格大幅波动风险等。

感谢您的信任与支持！

THANK YOU

孙业亮（首席分析师）
SAC编号：S1340522110002
邮箱：sunyeliang@cnpsec.com

刘聪颖（分析师）
SAC编号：S1340525100001
邮箱：liucongying@cnpsec.com

免责声明

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。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，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。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。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，中邮证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

中邮证券及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供中邮证券签约客户使用，若您非中邮证券签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为签约客户。

中邮证券对于本申明具有最终解释权。

免责声明

公司简介

中邮证券有限责任公司于2002年9月经中国证券监督管理委员会批准设立，公司注册资本61.68亿元人民币，是中国邮政集团有限公司绝对控股的证券类金融子公司，公司是中邮创业基金管理股份有限公司的第二大股东。

公司经营范围包括:证券经纪，证券自营，证券投资咨询，证券资产管理，融资融券，证券投资基金销售，证券承销与保荐，代理销售金融产品，与证券交易、证券投资活动有关的财务顾问，具备展业的各项资格。截至2025年10月底，公司在全国设有58家分支机构(含29家分公司、29家营业部)，1家资产管理分公司和1家另类投资子公司。

中邮证券紧密依托中国邮政集团有限公司的雄厚实力，通过强化“自营+协同”发展模式，实现快速发展，当前服务的经纪客户已超过260万人。公司始终坚持诚信经营、践行金融为民，为社会大众提供全方位专业化的证券投融资服务，努力成为员工自豪、股东放心、客户信赖、社会尊重的优秀企业，打造契合中国邮政资源禀赋和市场地位的特色精品券商。

投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准：报告发布日后的6个月内的相对市场表现，即报告发布日后的6个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A股市场以沪深300指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在20%以上
		增持	预期个股相对同期基准指数涨幅在10%与20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在5%与10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

中邮证券研究所

北京
邮箱：yanjiusuo@cnpsec.com
地址：北京市东城区前门街道珠市口东大街17号
邮编：100050

上海
邮箱：yanjiusuo@cnpsec.com
地址：上海市虹口区东大名路1080号大厦3楼
邮编：200000

深圳
邮箱：yanjiusuo@cnpsec.com
地址：深圳市福田区滨河大道9023号国通大厦二楼
邮编：518048


中邮证券
CHINA POST SECURITIES