

超配（维持）

DeepSeek 推动 AI 产业变革，上游算力、下游终端迎发展机遇

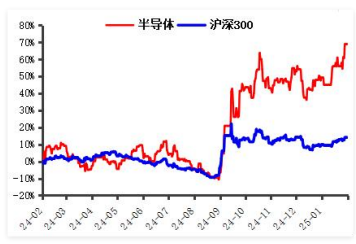
半导体行业专题报告

2025 年 2 月 25 日

投资要点：

- **DeepSeek推动AI产业变革，上线以来持续火爆。**DeepSeek（深度求索）公司自2003年成立以来专注于AI大模型的开发，并已完成多款大模型的迭代，规模、性能与复杂度持续提升。2025年1月，公司发布最新大模型DeepSeek-R1并同步开源权重，实际性能对标GPT 4o等海外领先大模型，在数学、代码和自然语言推理等任务中表现优异，并在多项第三方基准测试中表现出色。作为我国在开源AI领域的重大突破，DeepSeek-R1在发布后引发全球关注，上线一周即登顶中美等140多个国家应用商店下载榜首，日活用户突破4,000万，英伟达、微软、亚马逊等海外科技巨头先后上线支持部署。
- **DeepSeek加快应用场景落地，长期利好算力总需求增长。**根据Scaling Law理论，AI模型性能与参数量、数据量和计算量呈幂律关系，三者同步扩大时，模型性能会实现显著提升。作为AI发展的底层逻辑，Scaling Law是驱动技术进步的重要因素，推动各大模型参数量实现井喷式提升。DeepSeek在成本、性能、开源等方面具备显著优势，通过降本增效推动AI技术加速落地，带动中小企业、医疗、金融等传统行业的AI渗透，长期利好算力总需求增长。国产替代方面，海外出口管制持续升级，华为昇腾、海光等国产芯片企业通过积极适配DeepSeek，有望加速推进国产算力自主可控。
- **DeepSeek降低端侧部署难度，AI手机、AI耳机等智能终端有望加快渗透。**DeepSeek通过蒸馏技术和算法优化，显著降低模型的存储需求与计算量，并使得模型在端侧设备上的推理延迟大幅降低，同时保持较高的性能，降低端侧部署难度与成本，有望加速AI智能终端设备的普及与渗透。AI手机方面，国内多家手机厂商已接入DeepSeek模型，相比传统手机在处理器、存储等硬件迎来显著升级；AI可穿戴设备则凭借贴身性、自然交互与便携性等优势，成为端侧AI的最佳落地场景之一，以AI耳机为例，第三方机构预计到2035年AI+AR智能眼镜渗透率将达到70%，与智能手机规模相当，为相关SOC芯片和存储厂商带来新增长点。
- **投资建议：**DeepSeek-R1作为我国在开源AI领域的重大突破，具备高性能、低成本、开源三大优势，上线后引发全球关注，应用下载量高居榜首。DeepSeek通过技术创新与开源生态推动AI产业快速发展，并显著降低AI应用的门槛，长期利好算力总需求增长，而蒸馏技术与算法优化则显著降低模型的存储需求量与计算量，有效降低端侧

半导体（申万）指数走势



资料来源：Wind，东莞证券研究所

相关报告

部署难度与成本，有望加快AI智能终端的普及与渗透。建议关注算力芯片、存储、SOC、先进封装等相关受益环节。

■ **风险提示：**国产替代不及预期的风险、价格竞争加剧的风险等。

目录

1. DeepSeek 推动 AI 产业变革，上线以来持续火爆	5
2. DeepSeek 加速应用场景落地，长期利好算力总需求增长	8
3. DeepSeek 降低端侧部署难度，AI 手机、AI 可穿戴设备有望加速渗透	15
4. 投资建议	25
5. 风险提示	26

插图目录

图 1: DeepSeekR1 性能对齐 OpenAI-o1 正式版	6
图 2: DeepSeekR1 蒸馏小模型超越 OpenAIo1-mini	6
图 3: DeepSeek-R1 与 OpenAI 同类产品的 API 价格比较	7
图 4: DeepSeek 在全球多个国家下载榜排名第一	7
图 5: ChatgptvsDeepSeek 日活趋势（截至 2025/2/5）	7
图 6: 英伟达支持访问 DeepSeek-R1 模型	8
图 7: 亚马逊支持在 Amazon SageMaker Jumpstart 中快速部署 DeepSeek 模型	8
图 8: 英伟达 CEO 表态 Scaling Law 将会延续	10
图 9: 全球生成式人工智能和非生成式人工智能服务器市场规模预测（单位：百万美元）	10
图 10: 中国智能算力规模及预测（2020-2028，基于 FP16 计算，单位：EFLOPS）	11
图 11: 中国通用算力规模及预测（2020-2028，基于 FP64 计算，单位：EFLOPS）	11
图 12: 北美四大云厂商资本开支持续上行（亿美元，左轴）	13
图 13: AI 处理中心向边缘侧转移	16
图 14: 华为小艺接入 DeepSeek-R1	16
图 15: vivo 将深度融合 DeepSeek-R1	16
图 16: 苹果将于 3 月底在上海举行“利用 Apple 智能的力量”开发者活动	17
图 17: 三星 Galaxy AI 主要功能	17
图 18: 全球智能手机季度出货量	18
图 19: 国内智能手机月度出货情况	18
图 20: Apple Watch 可实现人体心电图检测	20
图 21: Vision Pro 可通过简易的手势操控训练机器人	20
图 22: Ola Friend 智能耳机参数	21
图 23: Ola Friend 智能耳机 AI 功能	21
图 24: AI 智能眼镜发展历程	22
图 25: AI 智能眼镜设计	22
图 26: CES2025 展出的 AI、AR 眼镜统计-1	23
图 27: CES2025 展出的 AI、AR 眼镜统计-2	23
图 28: 传统眼镜、AI 智能眼镜销量规模（2023-2035 年，含预测值）	23
图 29: MCU 级 SOC	24
图 30: 系统级 SOC	24
图 31: RayBanMeta 眼镜成本构成	25

表格目录

表 1: DeepSeek 自成立以来发布多款大模型	5
----------------------------------	---

表 2: GPT 参数量及预训练数据量	9
表 3: 西方国家或地区竞相加大算力基础设施布局	11
表 4: 国内部分 AI 芯片相关企业概览	14
表 5: 国产芯片企业积极适配 DeepSeek	14
表 6: 全球智能手机品牌出货品牌前 5 名	19
表 7: 多款智能耳机产品密集推出	20
表 8: 重点公司盈利预测及投资评级（截至 2025/02/24）	25

1. DeepSeek 推动 AI 产业变革，上线以来持续火爆

DeepSeek 自成立以来专注于 AI 大模型的开发。Deepseek(深度求索)公司成立于 2023 年 5 月，总部位于浙江杭州。公司由国内量化投资巨头幻方创立，自成立以来专注于大语言模型 (LLM) 及相关技术的开发，核心技术突破包括自主创新的 MLA (多头潜在注意力) 机制，在低算力环境下实现高效训练，并采用完全开源策略推动行业协作。公司创始人梁文峰，在量化投资、高性能计算领域具备丰富的研究背景和行业经验。

公司自成立以来已完成多款大模型迭代，规模、性能和复杂性持续提升。

2024 年 1 月，公司发布 DeepSeek Coder，主要用于代码生成领域，可提供智能代码补全、代码生成、调试优化等功能，该模型基于 2 万亿 token 的训练数据 (87% 为多语言代码，13% 为自然语言) 生成，参数规模涵盖 1B 到 33B 版本，支持包括 Python、Java、C++ 在内的 80+ 编程语言；

2024 年 2 月，发布 DeepSeek Math，基于一个包含 1,200 亿个数学标记的高质量预训练语料库进行训练，7B 版本在 GSM8K 数据集上达到 64.2% 的准确率，性能接近 Gemini-Ultra 和 GPT-4，能够快速解析复杂的数学表达式，通过逐步推理和计算，给出清晰、准确的结果；

2024 年 5 月，发布 DeepSeek V2，拥有 2,360 亿总参数，但在处理每个令牌时仅激活 210 亿参数，支持长达 128K 令牌的上下文长度。该模型采用 Transformer 架构，并在此基础上引入混合专家 (MoE) 架构，通过将任务分配给多个专家模型来处理，每个专家模型专注于特定的子任务，提高整体模型的性能和效率；2024 年 9 月，发布 DeepSeek V2.5，在 V2 基础上融合通用与代码能力，在写作任务、指令跟随等多个方面实现大幅提升；2024 年 12 月，发布 DeepSeek V3，生成速度提升到 60 TPS (每秒生成 60 个 token)，相比 V2.5 提升 3 倍。

2024 年 11 月，公司发布 DeepSeek-R1-Lite 预览版，在数学、编程和复杂逻辑推理任务上表现出色，推理过程透明，可实时展示 AI 的思考过程，提高模型的可解释性；2025 年 1 月，发布 DeepSeek-R1，在后训练阶段大规模使用强化学习技术，性能对标 OpenAI o1 正式版。

表 1: DeepSeek 自成立以来发布多款大模型

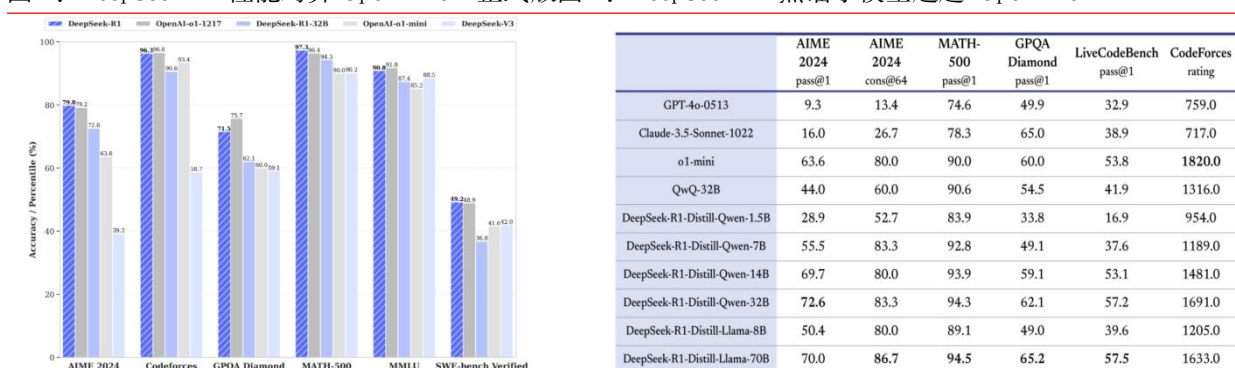
发布时间	大模型名称	模型介绍
2024.01	DeepSeek Coder	主要用于代码生成领域，目前仍是开源代码模型标杆
2024.02	DeepSeek Math	7B 版本性能接近 Gemini-Ultra 和 GPT-4
2024.05	DeepSeek V2	全球最强开源通用 MoE 模型，采用 Transformer 架构
2024.09	DeepSeek V2.5	融合通用与代码能力的全新开源模型，在写作任务、指令跟随等多个方面实现大幅提升
2024.11	DeepSeek-R1-Lite 预览版	能为用户展示完整的思考流程

2024. 12	DeepSeek-V2. 5-1210	通过 Post-Training 技术全面优化模型能力, 并新增联网搜索功能
2024. 12	DeepSeek V3	生成速度相比 V2. 5 提升 3 倍, 性能对齐海外领军闭源模型
2025. 01	DeepSeek R1	在后训练阶段大规模使用强化学习技术, 性能对标 OpenAI o1 正式版

资料来源: DeepSeek 官网, 东莞证券研究所

性能: DeepSeek-R1 性能比肩 OpenAI 正式版, 在多项基准测试中评分领先。2025 年 1 月 20 日, DeepSeek 发布旗下最新大模型 DeepSeek-R1, 并同步开源模型权重。该模型遵循 MIT License, 允许用户通过蒸馏技术借助 R1 训练其他模型。DeepSeek-R1 在后训练阶段大规模使用了强化学习技术, 在仅有极少标注数据的情况下, 极大提升了模型推理能力。根据第三方基准测试, DeepSeek-R1 实际性能出色, 表现优于 OpenAI、Anthropic 和 Meta 等海外领先 AI 企业, 在数学、代码编写和自然语言推理等任务中表现优异。其中, 在 AIME 2024 数学基准测试中, DeepSeek R1 得分为 79.8%, 超越 OpenAI 的 o1 推理模型; 在标准化编码测试中, DeepSeek 在 Codeforces 上获得 2029 Elo 的评级, 超越 96.3% 人类竞争对手。

图 1: DeepSeek R1 性能对齐 OpenAI-o1 正式版图 2: DeepSeek R1 蒸馏小模型超越 OpenAI o1-mini



资料来源: DeepSeek 官网, 东莞证券研究所

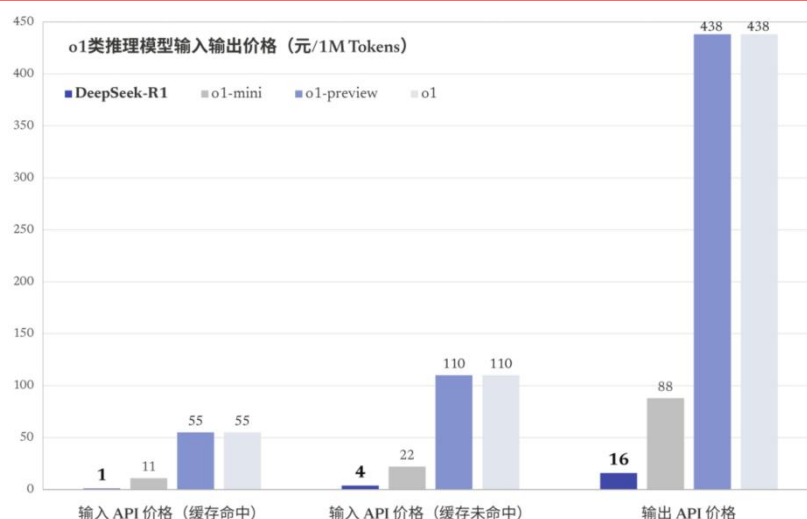
资料来源: DeepSeek 官网, 东莞证券研究所

成本: DeepSeek 通过软件与架构优化有效降低成本, 定价相比海外大模型极具优势。相比美国的领先大模型, DeepSeek-R 系列在技术路线上实现突破性创新, 采用混合专家 (MoE) 架构显著降低计算需求; 并大幅优化训练策略, 实现了更低的初始化与迭代成本。具体而言, DeepSeek 大模型首次摒弃监督微调环节并完全依赖强化学习训练, 对于小模型则通过知识蒸馏 (Knowledge Disillation) 技术成功获取大模型的推理能力, 实现在推理任务上的显著提升。成本方面, DeepSeek 训练成本仅为 OpenAI 同类模型的十分之一, 仅花费 557.6 万美元与 2048 块英伟达 H800 GPU 便完成了性能对标 GPT-4o 的模型训练; 推理方面, DeepSeek 的推理成本低至每百万 Token 0.14 美元, 而 OpenAI 推理成本为每百万 Token 7.5 美元。

定价方面, DeepSeek-V3 在 2 月 9 日结束了发布之初的优惠价格, 价格上调至每百万 tokens 输入价格为 0.5 元 (缓存命中)、2 元 (缓存未命中), 输出价格为 8 元, 价格上涨后定价仍远低于 GPT-4o 和 Claude-3.5-Sonnet 等海外顶尖闭源大模型;

DeepSeek-R1 输入价格为每百万 tokens 1 元（缓存命中）或 4 元（缓存未命中），输出价格为每百万 tokens 16 元，相比之下，OpenAI o1 模型为每百万输入 tokens 价格为 110 元，每百万输出 tokens 价格为 438 元，DeepSeek-R1 的定价仅为 OpenAI o1 的 3% 左右。

图 3: DeepSeek-R1 与 OpenAI 同类产品的 API 价格比较



资料来源：DeepSeek 官网，东莞证券研究所

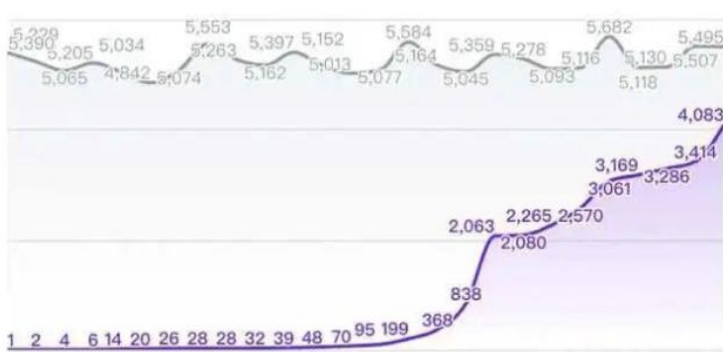
DeepSeek-R1 标志着我国在开源 AI 领域的重要突破，发布后迅速引发全球关注，在多家全球应用商店下载榜首。作为 DeepSeek 旗下的开源大型语言模型，DeepSeek-R1 专为复杂推理任务设计，在数学、代码和自然语言推理等领域的性能可对标 OpenAI 的 o1 正式版，其发布标志着中国在开源 AI 领域的重要突破，并迅速引发全球关注，上线一周即登顶中美等 140 多个国家应用商店下载榜首。根据 Appfigures 的数据，DeepSeek 于 1 月 26 日在苹果公司的 App Store 免费应用下载榜上排名第一，并一直保持全球第一的位置。数据显示，自发布以来，印度市场的下载量占各平台下载总量的 15.6%，占据新用户的最大比例，Sensor Tower 的研究显示，DeepSeek 还占据了谷歌安卓 Play 商店在美国的榜首位置，自 1 月 28 日以来一直保持着这一位置（TechWeb 2 月 2 日消息）。

日活方面，DeepSeek 自上线以来日活跃用户（DAU）增长迅速，据国内 AI 产品榜统计数据显示，DeepSeek 在上线仅 20 天后其日活用户迅速突破 2,000 万大关，达 2,215 万；此后，DeepSeek 日活用户持续攀升，根据 QuestMobile 数据显示，DeepSeek 在 1 月 28 日的日活跃用户数首次超越豆包，随后在 2 月 1 日突破 3,000 万大关，2 月 5 日突破 4,000 万大关，成为史上最快达成这一里程碑的应用。

图 4: DeepSeek 在全球多个国家下载榜排名图 5: Chatgpt vs DeepSeek 日活趋势（截至 2025/2/5）
第一



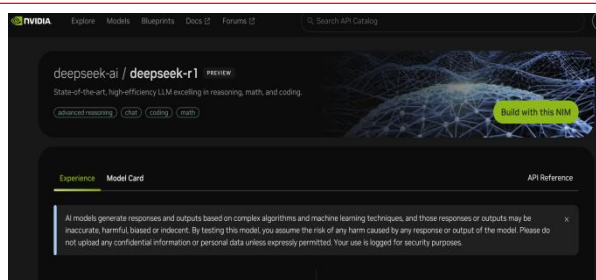
资料来源：TechWeb，东莞证券研究所



资料来源：非凡产研，东莞证券研究所

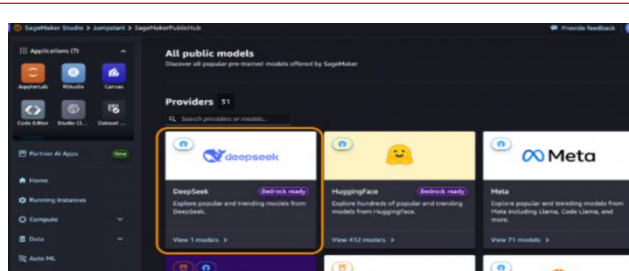
海外英伟达、微软、亚马逊等公司先后上线部署，支持用户访问 DeepSeek-R1。1 月 30 日，微软宣布已通过 Azure AI Foundry 和 Github 提供 DeepSeek-R1 模型，并宣布将该模型应用于其 AI 电脑 Copilot + PC，使开发者能够构建本地运行的 AI 应用程序；1 月 31 日，英伟达官宣上线搭载 DeepSeek-R1 的软件服务，据 NVIDIA 官网显示，DeepSeek-R1 模型已作为 NVIDIA ANIM 微服务预览版提供，为开发者开启了测试和体验该 API 的通道；1 月 31 日，亚马逊云科技（AWS）宣布已将 DeepSeek-R1 系列模型集成至其核心 AI 服务平台 Amazon Bedrock 和 Amazon SageMaker，支持用户通过公有云在线部署及私有化实例部署，此外，亚马逊通过线上直播和技术文档指导开发者部署 DeepSeek 模型，进一步扩展其应用范围。

图 6：英伟达支持访问 DeepSeek-R1 模型



资料来源：英伟达官网，东莞证券研究所

图 7：亚马逊支持在 Amazon SageMaker Jumpstart 中快速部署 DeepSeek 模型



资料来源：亚马逊，东莞证券研究所

2. DeepSeek 加速应用场景落地，长期利好算力总需求增长

Scaling Law 指引 AI 产业发展。Scaling Law（缩放定律）是 AI 领域的核心理论，最早由 OpenAI 于 2020 年在论文《Scaling Laws for Neural Language Models》中提出。其核心内容为：AI 模型的性能（如对话质量、图像生成精度）与模型参数量、数据集大小和计算量三者呈幂律关系，即：三者同步扩大时，模型性能会实现显著提升。近年来，Scaling Law 为大模型训练提供理论基础并主导 AI 研发方向，直接引发了各企业的算力军备竞赛，各大企业通过“堆算力、扩参数、增数据”提升模型性能，驱动大模型

参数从千亿级迈向万亿级。英伟达等硬件厂商则通过高性能 GPU 集群的迭代（如 Blackwell 架构），支撑了模型规模的扩展。

以 GPT4 为例，GPT 系列模型作为生成式预训练模型的典范，其参数量实现井喷式提升。生成式 AI 主要依赖于 AI 大模型，如 GPT 系列、BERT、Transformer 等，这些模型通常包含数十亿甚至数万亿参数，需要庞大的数据集进行训练。以 GPT 系列模型为例，从 2018 年的 GPT-1 到 2023 年推出的 GPT-4，每一代模型都在规模、复杂性和性能上有了显著提升。大模型的训练成本通常包括 GPU 等算力芯片成本、服务器成本、标准机柜成本、电力成本、人力投入费用等。以 GPT-4 为例，训练参数量达到 1.8 万亿个、在训练过程中使用了约 13 万亿个 token 的数据，预训练成本估计达到 6,300 万美元。高昂的训练成本提高了行业的准入门槛，因此在 DeepSeek 问世之前，多模态大模型的产业发展进程通常由科技巨头主导。

表 2：GPT 参数量及预训练数据量

模型	发布时间	参数量	预训练数据量
GPT-1	2018/06	1.17 亿	约 5GB
GPT-2	2019/02	15 亿	40GB
GPT-3	2020/05	1750 亿	45TB
Chatgpt	2022/11	千亿级别	百 T 级
GPT-4	2023/04	1.8 万亿个	约 13 万亿个 tokens

数据来源：Semianalysis，CSDN，东莞证券研究所

英伟达 CEO 表态 Scaling Law 将会延续。在 2025 年 1 月的 CES 大会上，英伟达 CEO 黄仁勋发表了主题演讲，演讲聚焦于 AI 技术的发展及其在机器人、智能驾驶等前沿领域的应用。黄仁勋指出，在接下来的几年中，人类产生的数据量将超过自古以来所有人类产生的数据总和，这将为 Scaling law 的延续提供了数据基础。他指出，除了传统的预训练扩展定律（Pre-Training Scaling Law）和后训练扩展定律（Post-Training Scaling Law）外，新的测试时间扩展定律（Test-Time Scaling Law）也已被提出，这意味着 AI 在资源分配上可以更加灵活，而不仅仅依赖于参数的改善。黄仁勋强调，作为 AI 发展的底层逻辑，Scaling Law 依然是推动技术进步的重要因素。随着数据和算力的增加，模型的能力将不断增强，即使在数据达到瓶颈时，通过智能化的算力管理依然可以推动 AI 产业发展。

图 8：英伟达 CEO 表态 Scaling Law 将会延续



资料来源：CES 2025，东莞证券研究所

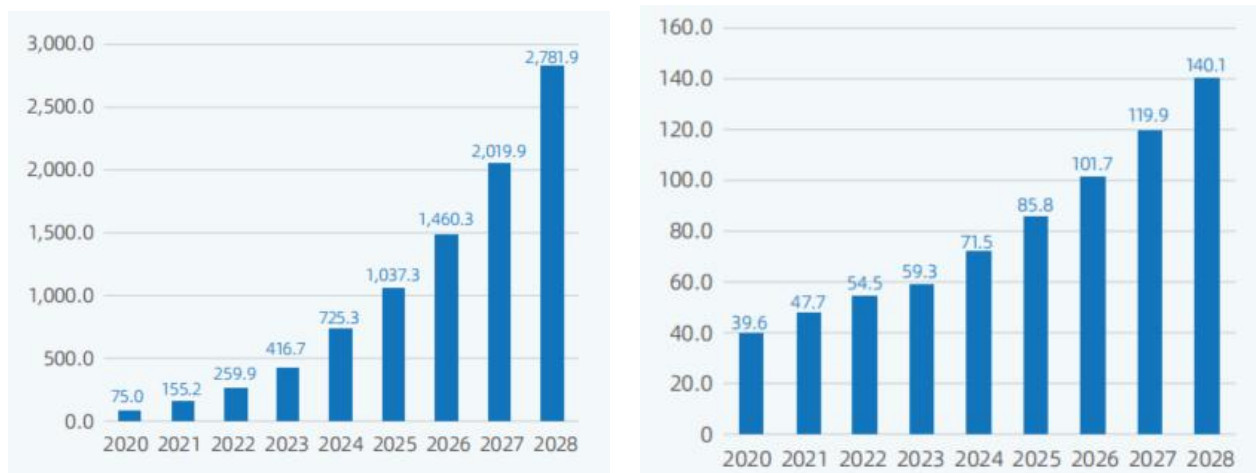
AI 大模型参数量快速增长，提升算力需求。大模型在训练过程中，通常需要高性能计算资源，如 GPU 和 TPU 等加速器。从 GPT 系列到 PaLM 和 Gemini 等模型，参数量和复杂性的不断增加，导致对算力需求的急剧上升。根据 IDC 数据显示，2024 年全球人工智能服务器市场规模预计为 1,251 亿美元，2025 年将增至 1,587 亿美元，2028 年有望达到 2,227 亿美元，其中生成式人工智能服务器占比将从 2025 年的 29.6% 提升至 2028 年的 37.7%；国内方面，IDC 预计 2025 年中国智能算力规模将达到 1,037.3 EFLOPS，到 2028 年将达到 2,781.9 EFLOPS，2023—2028 年期间智能算力复合增长率预计达到 46.2%，通用算力规模增速预计达到 18.8%。

图 9：全球生成式人工智能和非生成式人工智能服务器市场规模预测（单位：百万美元）



资料来源：《IDC&浪潮信息：2025 年中国人工智能算力发展评估报告》，东莞证券研究所

图 10:中国智能算力规模及预测(2020-2028,基于 FP16图 11: 中国通用算力规模及预测 (2020-2028, 基于 FP64
计算, 单位: EFLOPS) 计算, 单位: EFLOPS)



资料来源:《IDC&浪潮信息: 2025 年中国人工智能算力发资料来源:《IDC&浪潮信息: 2025 年中国人工智能算力发展评估报告》, 东莞证券研究所 评估报告》, 东莞证券研究所

海外竞相加大算力基础设施布局投入, 深化关键领域政策措施。随着新一轮科技革命加速演进, 算力已成为数字时代的关键生产力, 以美国为代表的西方国家或地区竞相加大算力基础设施的布局投入, 深化关键领域相关政策, 为算力产业发展奠定坚实基础。

表 3: 西方国家或地区竞相加大算力基础设施布局

国家或地区	时间	法案或政策	主要内容
美国	2022 年	《芯片与科学法案》	通过大量资金支持来加强美国半导体生产和研发能力, 减少对国外供应链的依赖
	2023 年	《国家量子计划重新授权法案》	将《国家量子计划法案》的支持期限延长至 2028 财年, 并计划在十年内追加超过 48.75 亿美元, 用于量子信息领域的基础研究、人才培养和产业联盟建设
	——	一系列出口管制规则	限制先进计算集成电路、超级计算机以及半导体制造设备出口, 以保持美国在前沿技术领域的全球领先地位
日本	2022 年	《人工智能战略 2022》	将人工智能置于科技创新和经济增长战略的核心位置, 并确立一体化的人工智能技术体系, 以提升日本科技产业竞争力
	2023 年	《半导体和数字产业战略》	修订《半导体和数字产业战略》, 旨在加快半导体基础设施建设、研发颠覆性半导体技术以及促进产业链协同发展
欧盟	2023 年	《芯片与科学法案》	欧洲《芯片法案》正式生效, 通过欧洲芯片计划促进关键技术产业化, 鼓励公共和私营企业对芯片制造商及其供应商的制造设施进行投资, 以推动为欧洲半导体领域工业基地的发展加强研究和创新, 并为欧洲应对未来芯片供应危机做好准备。
	2023 年	《2030 年数字十年政策方案》	《2030 年数字十年政策方案》正式生效, 提出“加强欧盟范围内传输、计算和数据基础设施建设”, 以实现欧盟 2030 年数字化转型的共同目标。
	2023 年	《2023—2024 年数字	概述了未来几年有关键信息技术的政策重点, 并表示将投入

		欧洲工作计划》	1. 13 亿欧元用于改善云服务安全性、创设人工智能实验及测试设施以及提升各个领域的数据共享水平。
	2024 年	《人工智能法案》	根据风险的严重程度，将 AI 系统分为不可接受风险、高风险、特定透明度和最低风险四个级别，其中高风险 AI 系统受到严格监管
英国	2024 年	《数字发展战略 2024-2030》	提出将努力实现数字化转型、数字包容、数字责任、数字可持续性四大目标，优先发展数字公共基础设施和人工智能。此外，英国将进行新一轮投资，约 3.88 亿英镑将用于支持科研与创新基础设施建设，以保持英国在未来科研与创新领域的关键地位。

数据来源：《中国算力发展报告》（2024年），互联网资料整理，东莞证券研究所

DeepSeek 降本增效推动 AI 应用场景加速落地，长期利好算力总需求增长。根据杰文斯悖论，当技术进步提高资源使用效率时，成本下降反而会刺激需求激增，导致资源总消耗量上升而非减少。在 AI 算力领域，这一现象体现为：尽管 DeepSeek 等技术创新降低了模型训练和推理的单次成本，但由于 AI 应用门槛降低、场景加速普及（如中小企业部署、医疗/金融等传统行业渗透），带动数据中心、边缘及端侧算力建设，反而推动算力需求持续膨胀。例如，开源生态促进 AI 应用爆发，推理算力需求显著增长；同时边缘计算和物联网设备扩展进一步拉动了算力分层需求，形成“训练集中化、推理边缘化”的双轨格局。因此，长期来看，AI 大模型效率提升将通过技术普惠化、场景纵深化和算力泛在化三重路径，推动大模型的普及与应用落地，驱动算力需求增长。

北美四大云厂商资本开支持续上行，且对 2025 年资本开支展望乐观。受 AI 旺盛需求驱动，亚马逊、谷歌、微软与 META 四大北美云厂商 2024 年资本开支实现同比大幅增长。具体而言，四家企业 2024 年资本开支合计达 2,273.06 亿美元，同比增长 54.16%，主要用于支持数据中心的建设与发展。展望方面，四大厂商均明确上调 2025 年对 AI 基础设施的持续投入，包括数据中心建设、服务器升级（如部署新一代 GPU）以及生成式 AI 的商业化需求等，预计四家企业 2025 年合计资本开支有望超过 3,000 亿美元。由此可见，即便面临算法优化的技术路径，高端算力储备仍被大型科技公司视为实现模型突破的基础条件。

亚马逊：亚马逊计划在 2025 年将资本开支增加到 1,000 亿美元，较 2024 年的约 830 亿美元实现显著增长，主要用于支持其云计算部门 AWS 的扩展。亚马逊 CEO 安迪·贾西（Andy Jassy）表示，公司正在加速投资数据中心、网络设备和硬件，以满足对生成式 AI 的庞大需求；

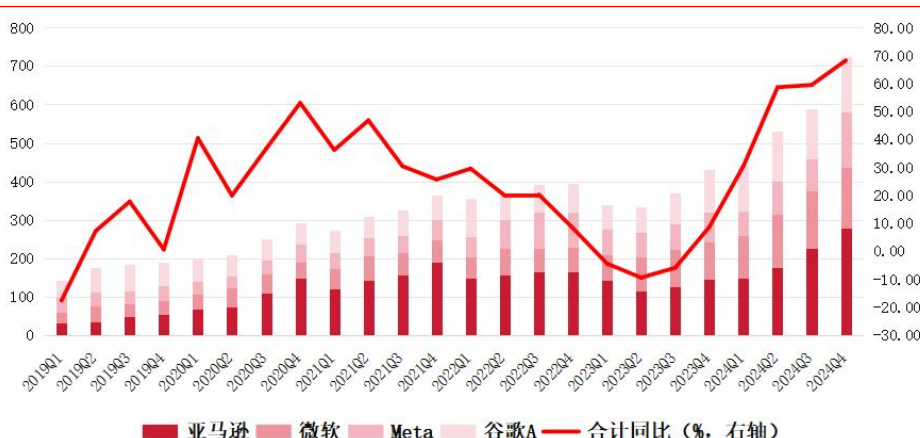
谷歌：谷歌母公司 Alphabet 计划在 2025 年投入 750 亿美元用于资本开支，较 2024 年的 500 亿美元增长 42%，其中，约 60% 资金将用于 GPU、TPU 等 AI 基础硬件采购，剩余部分则用于数据中心和网络设备的升级；

微软：微软于 1 月 7 日宣布，计划在 2025 财年（截至 2025 年 6 月）投入 800 亿美元用于基础设施建设，较 2024 财年的 444.77 亿美元大幅增长近 80%。其中，超过一半资金将用于美国本土的 AI 智算中心建设，支持 Azure 云服务中 AI 模型的训练与部署。

据 Omdia 报道，微软在 2024 年购买了接近 50 万片 H 系列 AI 加速芯片，带动 AI 业务年收入超过 130 亿美元；

Meta: Meta 计划在 2025 年投入 600 亿至 650 亿美元用于资本开支，较 2024 年的 380 亿至 400 亿美元大幅增长。资金将主要用于生成式 AI 研发、核心 AI 算法优化以及数据中心扩建。Meta 计划在路易斯安那州投资 100 亿美元建设其最大的 AI 智算中心，并持续推动 Reality Labs 的硬件开发。

图 12：北美四大云厂商资本开支持续上行（亿美元，左轴）



资料来源：iFind，东莞证券研究所

注：季度均为日历季度

国产替代视角：半导体领域出口管制措施不断升级，倒逼自主可控提速。2022 年 8 月，美国总统拜登签署《2022 年芯片和科学法案》，该法案授权对美国本土芯片产业提供巨额补贴和减税优惠，并要求任何接受美方补贴的公司必须在美国本土制造芯片，旨在增强美国本土芯片产业的竞争力，同时限制芯片制造产业向中国等国家的转移；同年 12 月，将长江存储、上海微电子、寒武纪等 36 家半导体相关企业列入实体清单，并采用“外国直接产品规则”，包含美国技术/产品/软件的物项，都将受到约束；2023 年 10 月，美国商务部公布的对华半导体出口管制最终规则，计划停止向中国出口由英伟达等公司设计的先进人工智能芯片，并限制将更广泛的先进芯片和芯片制造工具出口至包括伊朗和俄罗斯在内的更多国家；2024 年 11 月，美国政府下令台积电停止向中国大陆供应 7 纳米及以下先进制程芯片，主要针对 GPU 等高性能计算芯片，企图遏制我国在人工智能领域的发展；同年 12 月，美国工业和安全局（BIS）修订了《出口管理条例》（EAR），将 140 个中国半导体行业相关实体添加到“实体清单”，针对 24 种半导体制造设备和 3 种软件工具实施新的管制，并对 HBM 实施新的管制。总的来看，美国出口管制力度逐步升级，覆盖面逐步扩大，倒逼国内企业自主可控提速。

国产 AI 芯片性能显著提升，产业竞争力有所增强。据中国信通院《中国算力发展报告》（2024 年），近年来国内 AI 芯片相关企业发展迅速，海光信息、景嘉微等上市公司凭借卓越的技术实力成为国内 AI 芯片行业的引领者，天数智芯、燧原科技、壁仞科技等非上市企业在产品研发方面也取得突破性进展，为我国 AI 芯片发展注入动力。据 Bernstein Research 显示，目前国内 AI 芯片方面，华为自研的昇腾 910B、寒

武纪思元 590、百度昆仑芯 II-R200、腾讯紫霄 C100 等在一定程度上表现出性能优势，产业竞争力有所增强。

表 4：国内部分 AI 芯片相关企业概览

公司名称	成立年份	地点	注册资本	其他信息
景嘉微	2006	长沙	45,863 万元	已上市
昆仑芯	2011	北京	1,835 万元	未上市 C 轮融资完成
海光信息	2014	天津	232,434 万元	已上市
地平线	2015	北京	150,000 万美元	港股上市
天数智芯	2015	上海	19,210 万元	未上市 C+轮及 C+轮融资超 10 亿
寒武纪	2016	北京	41,659 万元	已上市
燧原科技	2018	上海	33,522 万元	未上市 D 轮融资 20 亿元
海思	2018	上海	8,000 万元	未上市
壁仞科技	2019	上海	3,292 万元	未上市 B 轮融资完成
摩尔线程	2020	北京	2,441 万元	未上市，已进入上市辅导阶段
沐曦	2020	上海	778 万元	未上市 Pre-B 轮融资 10 亿元
黑芝麻智能	2021	武汉	20,200 万美元	港股上市

数据来源：《中国算力发展报告》（2024年），东莞证券研究所

国产芯片企业积极适配 DeepSeek，推进国产算力生态自主可控。据国际电子商情，截至 2025 年 2 月 7 日，DeepSeek 已与包括华为昇腾、沐曦及天数智芯等在内的 16 家国产 AI 芯片公司完成了适配。其中，华为 2025 年 2 月 4 日宣布 DeepSeek-R1、DeepSeek-V3、DeepSeek-V2、Janus-Pro 上线昇腾社区，支持昇腾硬件平台上开箱即用，带来更快、更高效便捷的 AI 开发和应用体验；海光信息成功完成了 DeepSeek V3、R1 以及 DeepSeek-Janus-Pro 多模态大模型与海光 DCU（深度计算单元）的国产化适配工作，涉及底层指令集优化、软件生态兼容等，确保 DeepSeek 模型能在海光 DCU 平台上高效运行。这些合作不仅为 DeepSeek 技术的生态构建奠定了基础，也意味着行业内部将形成良性的技术竞争与相互促进，尤其是华为昇腾凭借其自主研发的推理加速引擎，成功使 DeepSeek 模型在昇腾硬件上的表现达到了与高端 GPU 相媲美的效果，为国产算力弯道超车提供思路。

表 5：国产芯片企业积极适配 DeepSeek

芯片企业	适配进展
华为	华为云宣布与硅基流动联合首发并上线基于华为云昇腾云服务的 DeepSeek R1/V3 推理服务；

	DeepSeek-R1、DeepSeek-V3、DeepSeek-V2、Janus-Pro 正式上线昇腾社区；华为 DCS AI 全栈解决方案中的重要产品—ModelEngine，全面支持 DeepSeek 大模型 R1&V3 和蒸馏系列模型的本地部署与优化，加速客户 AI 应用快速落地
沐曦	Gitee AI 联合沐曦首发全套 DeepSeek R1 千问蒸馏模型，全免费体验；DeepSeek-V3 满血版在国产沐曦 GPU 首发体验上线
天数智芯	成功完成与 DeepSeek R1 的适配工作，并且已正式上线多款大模型服务，其中包括 DeepSeek R1-Distill-Qwen-1.5B、DeepSeek R1-Distill-Qwen-7B、DeepSeek R1-Distill-Qwen-14B 等
摩尔线程	基于 Ollama 开源框架，完成了对 DeepSeek-R1-Distill-Qwen-7B 蒸馏模型的部署，并在多种中文任务中展现了优异的性能
海光信息	DeepSeek V3 和 R1 模型完成海光 DCU 适配并正式上线；海光 DCU 成功适配 DeepSeek-Janus-Pro 多模态大模型
壁仞科技	DeepSeek R1 在壁仞国产 AI 算力平台发布，全系列模型一站式赋能开发者创新
太初元基	基于太初 T100 加速卡 2 小时适配 DeepSeek-R1 系列模型，一键体验，免费 API 服务
云天励飞	完成 DeepEdge10 “算力积木”芯片平台与 DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Llama-8B 大模型的适配，可以交付客户使用
燧原科技	完成对 DeepSeek 全量模型的高效适配，包括 DeepSeek-R1/V3 671B 原生模型、DeepSeek-R1-Distill-Qwen-1.5B/7B/14B/32B、DeepSeek R1-Distill-Llama-8B/70B 等蒸馏模型。截至目前，DeepSeek 的全量模型已在庆阳、无锡、成都等智算中心完成了数万卡的快速部署
昆仑芯	完成全版本模型适配，这其中包括 DeepSeek MoE 模型及其蒸馏的 Llama/Qwen 等小规模 dense 模型
灵汐芯片	完成了 DeepSeek-R1 系列模型在灵汐 KA200 芯片及相关智算卡的适配，助力国产大模型与类脑智能硬件系统的深度融合
鲲云科技	全新一代的可重构数据流 AI 芯片 CAISA 430 成功适配 DeepSeek R1 蒸馏模型推理
希姆计算	仅用数小时就将 DeepSeek-R1 全系列蒸馏模型快速适配到自研 RISC-V 开源指令集的推理加速卡系列之上，并落地全国多个千卡级以上智算中心
算能科技	算能自研 RISC-V 开源指令集融合服务器 SRM1-20，成功适配并本地部署 DeepSeek-R1-Distill-Qwen-7B/1.5B 模型
清微智能	可重构计算架构 RPU 芯片已完成 DeepSeek-R1 系列模型的适配和部署运行
芯动力	在 24 小时内完成了与 DeepSeek-R1 大模型的适配，AzureBlade K340L M.2 加速卡是关键
龙芯中科	搭载龙芯 3 号 CPU 的设备成功启动运行 DeepSeek R1 7B 模型，实现本地化部署
瀚博半导体	已完成 DeepSeek-V3 与 R1 全系列模型训推适配，单机可支持 V3 与 R1 671B 全量满血版模型部署

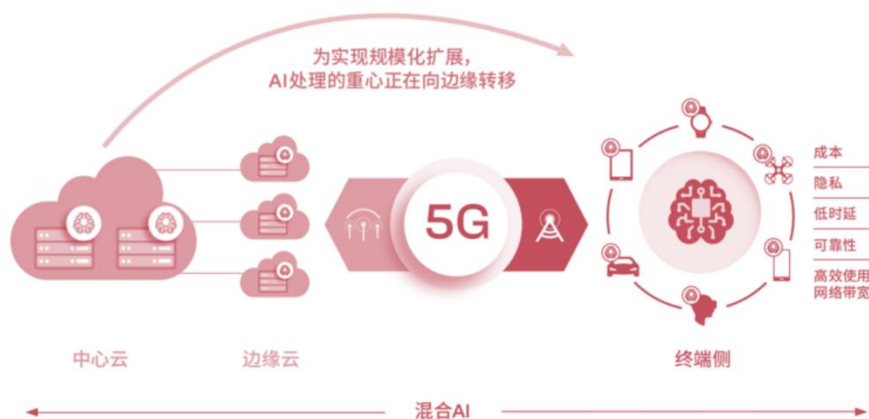
数据来源：国际电子商情，东莞证券研究所

3. DeepSeek 降低端侧部署难度，AI 手机、AI 可穿戴设备有望加速渗透

隐私保护+成本考量，驱动 AI 处理中心向边缘侧转移。近年来，用户对于数据隐私愈发关注，希望终端数据能够留在本地，确保数据可靠并且能够获取实时结果。边缘计算可以在终端设备上处理数据，无需将数据上传至云端，从而更好地保护用户隐私。从成本效益角度考虑，生成式 AI 模型的推理规模远大于训练规模，导致推理成本极高，将 AI 处理从云端转移到边缘侧可以减轻云端基础设施的压力，有效降低运营成本。

据《2024 高通 AI 白皮书》，终端侧 AI 能力是赋能混合 AI 并让生成式 AI 实现全球规模化扩展的关键。在生成式 AI 出现之前，越来越多的 AI 推理工作负载就已在手机、笔记本电脑、XR 头显、汽车和其他边缘终端上运行，如手机利用终端侧 AI 支持许多日常功能，如暗光拍摄、降噪和人脸解锁。

图 13：AI 处理中心向边缘侧转移



资料来源：《2024 高通 AI 白皮书》，东莞证券研究所

DeepSeek 模型参数量大幅减少且推理延迟降低，降低端侧部署难度。DeepSeek 通过蒸馏技术，将大模型压缩为小模型，例如从 671B 参数的原始模型压缩到 1.5B-70B 参数的版本，显著降低模型的存储需求和计算量；通过优化计算路径和算法，使得模型在端侧设备上的推理延迟大幅降低，同时保持了较高的性能；此外，DeepSeek 的开源模式与生态使用户的本地部署更加灵活，开源社区的支持也为用户提供了丰富的技术资源和解决方案，降低部署过程中的难度。综上所述，DeepSeek 可在不显著降低性能的前提下，降低端侧部署的难度与成本，对 AI 手机、AI PC 与 AI 可穿戴设备等构成利好。

AI 手机：国内智能手机厂商积极接入 DeepSeek。AI 与智能手机的融合是必然趋势，DeepSeek 上线以来，国内包括华为、荣耀、OPPO、魅族、努比亚和 vivo 在内的众多厂商都已宣布接入 DeepSeek 模型。目前大多数手机厂商的合作方案是把 DeepSeek“嫁接”原有的 AI 界面，而努比亚宣布将会把 DeepSeek 嵌入到整个系统中，实现 6,710 亿参数量的全尺寸嵌入。据 IDC 指出，目前国内厂商的“接入”只是增加一个 DeepSeek 的入口，本质上与下载一个 DeepSeek 的 APP 无太大区别，若后续进行后台 API（应用程序编程接口）的打通，方可使得 DeepSeek 的推理能力真正融入手机的 AI 中。

图 14：华为小艺接入 DeepSeek-R1

图 15：vivo 将深度融合 DeepSeek-R1



资料来源：华为，东莞证券研究所

资料来源：vivo，东莞证券研究所

海外厂商:苹果国行宣布与阿里合作,三星则搭载智谱 AI。2月11日,据 The Information 消息,苹果已与阿里达成合作协议,由通义大模型为大陆用户提供 AI 功能。结合苹果将于 3 月底在上海举行“利用 Apple 智能的力量”开发者活动,预计国行苹果 AI 功能渐行渐近。近期苹果 FY25Q1 业绩说明会上,管理层表示,在推出 Apple Intelligence 的地区, iPhone16 系列销售的同比表现要好于未推出 AI 功能的市场,并且表示将进一步拓展 AI 功能,在 4 月份推广包括法语、德语、简体中文等版本,以及更新更强大的智能化 Siri;三星国行 S25 系列则于 2 月 11 日亮相,全系搭载智谱 Agentic GLM,具备多模态能力,能够支持基于 AI 的实时语音和视频通话,以及基于实现视觉理解和系统功能调用、AI 搜索、文案写作等功能,具备跨应用操作能力。

图 16: 苹果将于 3 月底在上海举行“利用 Apple 智能图 17: 三星 Galaxy AI 主要功能的力量”开发者活动

与 Apple 会面交流

利用 Apple 智能的力量

想要了解 Apple 智能可以为你的 App 做些什么？最新的线下和线上开发者活动等等你来参加。

利用 Apple 智能和 App Intents 优化你的 App

当地时间 2 月 13 日，艾哈迈达巴德 >

当地时间 2 月 20 日，古尔冈 >

深入探索 Apple 智能和机器学习

当地时间 3 月 24 日，上海 >



www.it-home.com

Galaxy AI 主要功能³

生产力

跨应用执行链
即时简报
即圈即搜
笔记助手
转录助手

交流

同传
通话助手
写作助手

创造力

照片助手
绘图助手

公众号·量子位

资料来源：IT 之家，东莞证券研究所

资料来源：量子位公众号，东莞证券研究所

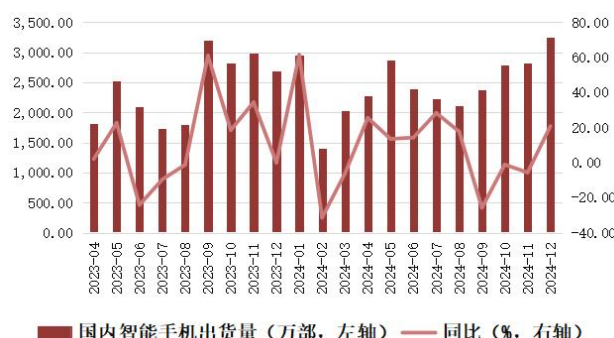
受益 AI 驱动和“国补”推广，国内智能手机销量实现复苏。受益于 AI 手机逐步渗透，叠加全球宏观经济回暖，全球智能手机 2024 年销售情况有所回暖，带动供应链企业业绩回升。据 IDC，2024 年全球智能手机出货量为 12.23 亿台，同比增长 5.23%；国内方面，受益于“国补”等消费电子补贴逐步落地，国内智能手机换机需求得到充分释放。根据工信部数据，2024 年全年国内智能手机出货量为 2.94 亿台，同比增长 6.48%。而根据第三方机构数据，春节前一周，我国手机销售量、销售额同比分别增长 30%和 34%。

图 18：全球智能手机季度出货量



资料来源：IDC，东莞证券研究所

图 19：国内智能手机月度出货情况



资料来源：工信部，东莞证券研究所

国产终端品牌话语权提高，利好国产供应链份额提升。据 TechInsights，2024 年第三季度全球智能手机出货量同比增长 3.8%，达到 3.077 亿部，连续四个季度保持复苏态势，其中非洲中东地区、中拉丁美洲和亚太地区等新兴市场成为复苏的主要动力。而全球前十大供应商中，有八家企业来自中国大陆或为中国大陆所控制的企业（小米、

OPPO（含一加）、vivo、传音（包含 Tecno、Infinix 和 itel）、联想－摩托罗拉、荣耀、realme 和华为），国产品牌话语权继续提升。近年来，国产品牌在技术研发方面投入不断加大，取得了一系列的创新成果。如在折叠屏技术、影像技术、快充技术、芯片研发等方面，国产品牌已经处于行业领先地位，引领了行业的发展趋势。以华为为代表的国产智能手机话语权提高，有望带动国产供应链企业份额提升。

表 6：全球智能手机品牌出货品牌前 5 名

手机品牌	24Q3 出货量（百万台）	24Q3 市场份额	24Q3 同比
三星	57.7	18.8%	-3.0%
苹果	52.4	17.0%	12.4%
小米	42.8	13.9%	3.1%
OPPO	27.7	9.0%	0.0%
vivo	27.6	9.0%	24.9%
其他	99.5	32.3%	0.6%
合计	307.7	100.0%	3.8%

资料来源：TechInsights，东莞证券研究所

AI 手机在处理器、存储等方面硬件迎来显著提升。处理器方面，AI 手机通常配备更高的系统级芯片（SOC），其算力要求显著提高。据 IDC 定义，新一代 AI 手机指 NPUs 算力大于 30 TOPS、搭载能够支持更快速高效端侧 Gen AI 模型的 SOC、支持包括 Stable Diffusion 和各种大语言模型在内的 Gen AI 模型在端侧运行的智能手机，能够处理复杂的 AI 任务。传统手机的处理器性能较低，无法满足这些需求。

存储方面，运行 AI 大模型通常需要更大的内存来处理复杂的计算和数据分析，AI 运行 130 亿参数的大模型通常需要至少 13GB 的内存，而传统智能手机通常内存较小，无法支持此要求。以苹果搭载的 AI 系统为例，运行 Apple Intelligence 的功能要求设备配备至少 8GB 的 RAM，包括 iPhone、iPad 和 Mac。为支持苹果 AI 相关功能，苹果于 2024 年 9 月新发布的 iPhone 16 全系列手机均配备 8GB 以上存储，其中 iPhone 16 标准版的运存相比前代提升 33%，而 10 月发布的新一代 iPad mini（A17 Pro 版）则配备了 8GB 内存，相较前一代的 iPad mini 6 的 4GB 内存规格提升幅度达到 100%，以更好地支持 AI 相关功能。

AI 可穿戴有望成为 AI 在端侧最佳落地场景之一。相比 AI 手机、AI PC 等，AI 可穿戴设备（如 AI 耳机、AI 手表、AI 眼镜等），具备如下优势，因此有望成为 AI 在硬件端的最佳落地场景之一。

（1）**贴身性与数据采集优势：**手表、眼镜等可穿戴设备与人体皮肤高度贴合，能够实时采集运动、健康等多维数据，为 AI 算法提供动态的输入源，方便对人体提供实时监测，实现精准分析与个性化服务；

（2）**自然交互与多模态能力：**可穿戴设备通常集成摄像头、麦克风等传感器，结合

语音、手势等交互方式，使 AI 能够“看用户所看、听用户所听”，提供更直观的伴随式服务；

（3）**便携性与场景适用性**：相比手机和电脑，AI 眼镜、手表和耳机等更加轻便易携，突破传统终端的时空限制，在运动、医疗等垂直场景中实现无缝替代或补充；

（4）**技术迭代与生态整合**：近年来国产 SOC 芯片取得一定进步，带动端侧 AI 算力提升，叠加 DeepSeek 降本增效利好端侧轻量化部署，云端协同的不断发展增强了实时 AI 能力，形成“端－云”联动的服务生态。

图 20：Apple Watch 可实现人体心电图检测



资料来源：Apple，东莞证券研究所

图 21：Vision Pro 可通过简易的手势操控训练机器人



资料来源：VR 陀螺公众号，东莞证券研究所

AI 耳机：多款智能耳机密集发布，有望催化行业增长。2024 年 4 月，国际智能声学品牌 Cleer 发布 ARC3 开放式耳机，搭载高通 S5 Gen2 芯片，单次续航 10 小时，搭配充电盒续航时间达 50 小时，能够通过 AI 语音控制实现切换歌曲、接听电话等操作，具备 AI 双向高清通话、AI 声境恒听等功能，起售价为 1,399 元。2024 年 5 月，科大讯飞推出面向会议场景的 iFLYBUDS Pro2 真无线耳机，单次续航 9 小时，搭配充电盒续航时间 36 小时，搭载 viaim AI 会议助理，能够智能分析耳机实时转写的内容，提取关键信息、生成会议摘要及待办事项，此外亦具备录音转文字、多种语言及方言实时同传翻译的功能，首发价为 1,399 元。

表 7：多款智能耳机产品密集推出

产品	发 布 时 间	价 格 (元)	AI 功能
iKKO ActiveBuds	2024 年 2 月	1,799	接入 OpenAI 的 GPT-4o 模型，能够回答用户提出问题；支持 62 种语言同声传译，57 种语言录音转笔记，提炼内容要点、优化语法、精准翻译等。续航时间 7.5 小时，搭配充电仓可达 30 小时
Cleer ARC3	2024 年 4 月	1,399	通过 AI 语音控制实现切换歌曲、接听电话等操作，具备 AI 双向高清通话、AI 声境恒听等功能。续航时间 10 小时，搭配充电仓达 50 小时
科大讯飞 iFLYBUDS Pro2	2024 年 5 月	1,399	搭载 viaim AI 会议助理，智能分析耳机实时转写的内容，提取关键信息、生成会议摘要及待办事项，具备录音转文字、多种语言及方言实时同传翻译的功能。单次续航 9 小时，搭配充电仓达 36 小时
SoundAI FairyClip C1	2024 年 7 月	199	接入 AzeroGPT 壹元大模型，支持实时转写、文件转写、多语翻译、智能摘要等功能。续航时间 7 小时，搭配充电仓可达 30 小时
时空壶 W4 Pro	2024 年 9 月	2,999	支持双向同传技术，外语电话及会议实现 AI 双语转写、智能总结，在线支持 40 种常见语言、93 种常见口音。续航时间 6 小时，搭配充电仓可达 18 小时
FIIL GS Links	2024 年 11 月	399	搭载字节豆包 AI，提供会议录音摘要、通话录音摘要、同声传译三大功能，可识别四川话、粤语、上海话等方言。续航时间 7 小时，搭配充电仓可达 22 小时

数据来源：京东，SoundAI官网，东莞证券研究所

字节推出 Ola Friend 智能耳机，可调用豆包大模型，实现智能化操作。Oladance 由前 Bose 高管联合 Bose 工程师在 2019 年成立，公司在全全球首次提升 OWS 专利技术，并且在 2021 年推出 OWS 系列耳机，具备无佩戴感、不入耳、音响级声效、超长续航等特点。2024 年 9 月，字节以 5,000 万美元对价收购 Oladance，并且在 10 月正式推出 Ola Friend 智能耳机。参数方面，Ola Friend 搭载恒玄科技蓝牙主控芯片，单只耳机重量为 6.6g，8 小时续航时间，搭配充电盒整体续航时间达 28 小时，售价为 1,199 元。功能方面，Ola Friend 通过蓝牙与手机端的豆包 APP 连接，用户只需要通过语音指令或耳机触控即可调用豆包，实现智能问答、英语陪练、播放音乐等操作。

图 22：Ola Friend 智能耳机参数

图 23：Ola Friend 智能耳机 AI 功能



资料来源：Apple，东莞证券研究所

资料来源：VR 陀螺公众号，东莞证券研究所

AI 眼镜在传统眼镜基础上逐步迭代发展，通过搭载 AI 大模型实现多种功能。智能眼镜是在传统眼镜基础上迭代发展，除了具备基础的视力矫正、遮光、装饰功能外，在嵌入摄像头、麦克风、耳机、蓝牙、Wi-Fi 等电子模块后，从而具备了拍摄、音频、通讯等功能，能够给用户多媒体体验。随着多模态大模型性能持续提升，终端厂商开始将 AI 大模型搭载至智能眼镜，用户能够通过语音或者触控板与大模型进行交互，最终实现智能问答、一键拍摄、实时翻译等一系列功能。相较于 AR 眼镜，AI 眼镜不具备光学显示系统，重量和佩戴舒适感与普通眼镜相近，成本相对较低，且在 AI 加持下功能丰富，具备多模态交互能力，后续有望在海内外终端积极推动下加速渗透。

图 24：AI 智能眼镜发展历程

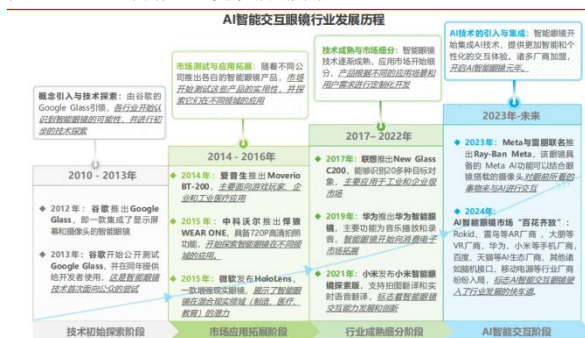


图 25：AI 智能眼镜设计



资料来源：《AI 智能交互眼镜产业洞察报告》，东莞证券研资料来源：《AI 智能交互眼镜产业洞察报告》，东莞证券研究所

AI 及 AR 眼镜成 CES 2025 重要焦点。1 月 7 日-1 月 10 日，国际消费电子展 CES 在美国拉斯维加斯正式拉开序幕，AI 及 AR 眼镜成为展会焦点。据 VR 陀螺统计，CES 现场展示的 AI 眼镜、AR 眼镜的数量达到 47 款，主要包括纯音频眼镜、音频+拍照 AI 眼镜、带显示 AR 眼镜等产品。从产品来看，基本都搭载 AI 大模型，具备多模态能力，能够

实现智能问答、翻译、识别物体等多个功能。从展出商来看，基本集中在国内品牌，如雷鸟创新、Rokid、雷神科技、李未可科技、星纪魅族等，国内厂商引领力进一步提升。

图 26: CES 2025 展出的 AI、AR 眼镜统计-1

品牌	产品	摄像头	光学	重量	其他亮点
1MORE 万魔	Q70	500万像素	×	不詳	280mAh 电池、运动追踪、睡眠充电、设计灵感中心、Q3 上市
Gaio 捷多摩	AI 智能眼镜	×	×	不詳	售价 30 美元，支持 ChatGPT 交互
	AI 智能眼镜	1200 万像素	×	不詳	搭载 5.8 英寸 OLED 屏，IPX4 防水，支持 200 万像素，Q3 上市
Gyges Labs/Hoodly	Halliday	×	×	35g	集成集成触控手
RAIR	RAIR 2 Pro 智能眼镜	×	Micro-OLED+BirdBath	不詳	RAIR 2 Pro 智能眼镜支持 3840*1920 分辨率，120Hz 高刷，电量变色
	RAIR AI 智能眼镜	×	×	不詳	支持外接摄像头，AI 智能眼镜
RayNeo 雷鸟创新	雷鸟 X1 AI 智能眼镜	×	×	39g	透光率提升 2 倍，支持 AI 智能眼镜，支持 1179Hz
	雷鸟 Air 3	×	Micro-OLED+BirdBath	36g	支持外接摄像头，集成摄像头
	雷鸟 X1 Pro	×	Micro-OLED+BirdBath	不詳	0.66 英寸，Q3 上市
Rokid	Rokid Glasses	×	Micro-OLED+BirdBath	49g	轻量化设计，支持 AI 智能眼镜
	Rokid Air Lite	×	Micro-OLED+BirdBath	75g	轻薄、大屏显示、低功耗
	Rokid Air Studio	×	Micro-OLED+BirdBath	不詳	支持 AI 智能眼镜
Solidot	Solidot Vision	×	×	不詳	针对开发者优化的 AI 智能眼镜
星纪魅族	Starry Air2	×	Micro-LED+BirdBath	44g	与 Capillary 合作，定位为高端人士服务，提供沉浸式体验
ThinkAR	Altium	×	Micro-LED+BirdBath	37g	Fov 12.3°，搭载 Ambio Apollo4 处理器
Vuzix	Ultimate 智能眼镜参考设计	800 万像素	光波导	78.3g	30° Fov，AR1 平台，Android 12
XREAL	XREAL Air 2 Ultra	×	Micro-OLED+BirdBath	80g	XR 智能眼镜，集成摄像头
	XREAL One 智能眼镜	×	Micro-OLED+BirdBath	84g	X1 芯片，104° 超广视野，Pico 芯片为备用设备
大朋 VR	DPVR AI Glasses 系列	×	×	不詳	配置具有较大提升，支持 AI 智能眼镜，支持 1179Hz
	DPVR AI Glasses 系列	×	×	不詳	配置具有较大提升，支持 AI 智能眼镜，支持 1179Hz
AR 眼镜参考设计 HoloLens 2	AR 眼镜参考设计 HoloLens 2	不詳	Micro-LED+BirdBath	36g	超薄设计，支持 AI 智能眼镜，支持 1179Hz
雷神科技	雷神 2	×	×	48g	超薄设计，支持 AI 智能眼镜，支持 1179Hz
	雷神 AURA 智能眼镜 AR-AI 版	×	Micro-OLED+BirdBath	79g	2025 年 3 月上市
	雷神 AURA 智能眼镜 AI 版	×	×	不詳	搭载骁龙 AR1 平台，支持 180° 超广视野，支持 AI 智能眼镜
李未可	Meta Lens View	1200 万像素	×	37g	支持 AI 智能眼镜，支持 1179Hz

资料来源：VR 陀螺公众号，东莞证券研究所

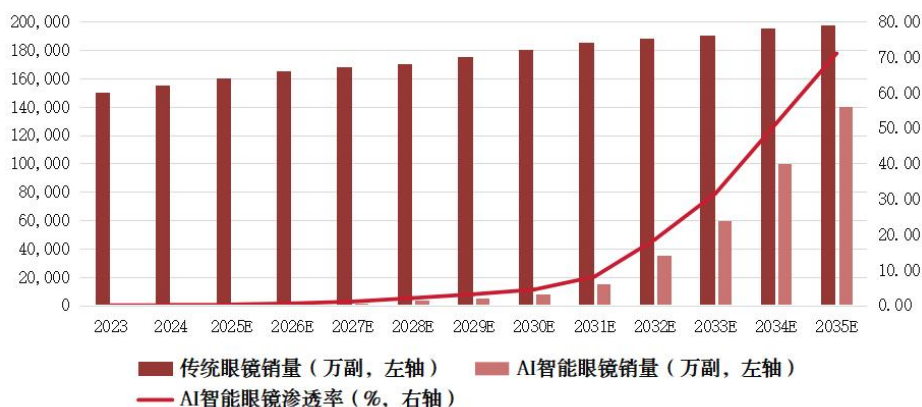
图 27: CES 2025 展出的 AI、AR 眼镜统计-2

品牌	Legion Glasses 2	×	Micro-OLED+BirdBath	45g	Fov 43.3°，最大输入延迟可低至 0.0001 秒
雷鸟科技	AI+AR 智能眼镜参考设计	×	Micro-OLED+BirdBath	50g	支持 AI 智能眼镜，支持 AI 智能眼镜
	AI+AR 智能眼镜参考设计	×	Micro-OLED+BirdBath	38g	支持 AI 智能眼镜，支持 AI 智能眼镜
雷神科技	AR 智能眼镜参考设计	×	×	不詳	支持 AI 智能眼镜
闪极	Stormes AR 眼镜	×	×	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
雷鸟	雷鸟智能眼镜参考设计	×	×	45g	支持 AI 智能眼镜，支持 AI 智能眼镜
雷神科技	Migait Omnicast 智能眼镜	×	Micro-OLED+BirdBath	不詳	Fov 12°，搭载骁龙 8020，最大输入延迟可低至 0.0001 秒
亿通思创 EndoorVR	AI 智能眼镜参考设计	×	×	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
雷鸟	AR 智能眼镜 QAT109	×	Micro-OLED+BirdBath	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
	SA1401	×	Micro-OLED+BirdBath	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
	SA1501	×	×	55g	支持 AI 智能眼镜，支持 AI 智能眼镜
Cyber	参考设计	×	×	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
雷鸟科技	AI 智能眼镜参考设计	×	×	40g	支持 AI 智能眼镜，支持 AI 智能眼镜
AAC	AI 智能眼镜参考设计	×	×	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
	MUTRICS G8-30	×	×	不詳	支持 AI 智能眼镜
MUTRICS	MUTRICS X	×	×	不詳	支持 AI 智能眼镜
	AI 智能眼镜参考设计	×	×	不詳	支持 AI 智能眼镜
Esilux/Luxtica	Nuance Audio	×	×	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜
Meta	Ray-Ban Meta	×	×	50g	支持 AI 智能眼镜，支持 AI 智能眼镜
品牌	Pisen Smart Goggles	×	×	不詳	支持 AI 智能眼镜，支持 AI 智能眼镜

资料来源：VR 陀螺公众号，东莞证券研究所

预计 AI 眼镜渗透率将持续提升。目前 AI 智能眼镜仍处于探索期，传统手机厂商、互联网厂商和初创公司等纷纷布局探索 AI 智能眼镜方案，并密集发布相关新品。根据 wellsenn XR 预测，预计 2025 年将有更多大厂进场竞争，共同推动 AI 眼镜趋向成熟；而到 2035 年，AI+AR 智能眼镜将实现对传统智能眼镜的替代，渗透率达到 70%，全球 AI+AR 智能眼镜销量将达到 14 亿台规模，与智能手机规模相当，成为下一代通用计算平台和终端。

图 28: 传统眼镜、AI 智能眼镜销量规模（2023—2035 年，含预测值）



资料来源：wellsenn XR，东莞证券研究所

SOC 为带摄像头 AI 智能眼镜的硬件核心，按照技术路径可分为两种。SOC，即片上

系统芯片，即将多个电子组件集成到单一芯片上，以提供完整系统所需的所有电子电路。按照构成划分，SOC 可分为 MCU 级 SOC 和系统级 SOC。

MCU 级 SOC：在 MCU 基础上发展而来，主要以 MCU 内核为控制中心，添加特定的功能模块如蓝牙模块、音频模块等形成特定类型 SOC，如恒玄 BES2500YP 芯片。该类 SOC 注重低功耗下的功能应用，多采用 arm 的 Cortex-M 系列核心或其他低功耗处理器核心，专注于低功耗、紧凑尺寸和实时控制，适用于小型嵌入式系统；

系统级 SOC：在 CPU 基础上发展而成，以 CPU 为中央控制单元，根据功能需求集成相应的硬件模块，如 GPU、ISP、DSP、Wi-Fi 蓝牙模块、视频编解码系统、音频系统等，代表性芯片有高通 AR1 Gen1 芯片、紫光展锐 W517 SOC 等。相比 MCU 级 SOC，系统级 SOC 所提供的算力更高，可支持的集成度更高，且大多集中了 ISP 模块，因此不需要外接 ISP 芯片即可实现拍摄功能。但是，系统级 SOC 方案具有更高的功耗及成本，因此对 AI 眼镜的定价和续航提出挑战。

图 29：MCU 级 SOC



资料来源：wellsenn XR，东莞证券研究所

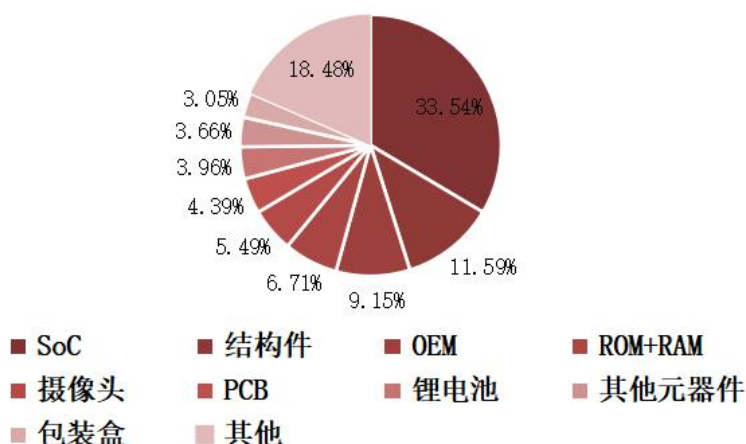
图 30：系统级 SOC



资料来源：wellsenn XR，东莞证券研究所

AI 眼镜逐步普及，利好相关 SOC 厂商和存储企业。无论采用哪种方案，AI 眼镜要实现相关功能，均以 SOC 作为硬件核心，需具备低功耗、高集成度及 AI 加速能力等特性，构成端侧 AI 硬件的算力核心。以 2024 年下半年 Meta 发布的 Ray Ban 智能眼镜为例，据 wellsenn XR 数据，Ray Ban Meta 的不含税综合硬件成本为 164 美元，其中 SOC 成本达 55 美元，占比 33.54%，为成本占比最高的硬件构成；而 ROM+RAM 成本为 11 美元，占比约 6.71%。随着多模态 AI 大模型加速落地，以 AI 眼镜为代表的端侧 AI 产品所需的算力、存力也将实现同步增长，有望为国内 SOC 厂商、存储厂商提供新的增长点。

图 31：Ray Ban Meta 眼镜成本构成



资料来源：wellsenn XR，东莞证券研究所

4.投资建议

投资建议：DeepSeek-R1 作为我国在开源 AI 领域的重大突破，具备高性能、低成本、开源三大优势，上线后引发全球关注，应用下载量高居榜首。DeepSeek 通过技术创新与开源生态推动 AI 产业快速发展，并显著降低 AI 应用的门槛，长期利好算力总需求增长，而蒸馏技术与算法优化则显著降低模型的存储需求量与计算量，有效降低端侧部署难度与成本，有望加快 AI 智能终端的普及与渗透。建议关注算力芯片、存储、SOC、先进封装等相关受益环节。

建议关注标的：海光信息（688041）、兆易创新（603986）、澜起科技（688008）、东芯股份（688110）、佰维存储（688525）、普冉股份（688766）、全志科技（300458）、长电科技（600584）、通富微电（002156）。

表 8：重点公司盈利预测及投资评级（截至 2025/02/24）

股票代码	股票名称	股价（元）	EPS（元）			PE（倍）			评级	评级变动
			2023A	2024E	2025E	2023A	2024E	2025E		
688041	海光信息	154.06	0.54	0.84	1.24	134.64	183.49	124.46	买入	维持
603986	兆易创新	143.00	0.24	1.65	2.65	135.64	86.67	53.96	买入	维持
688008	澜起科技	77.67	0.39	1.19	1.95	136.64	65.27	39.83	买入	维持
688110	东芯股份	30.97	-0.69	-0.18	0.22	137.64	-172.06	144.05	买入	维持
688525	佰维存储	70.60	-1.45	0.95	1.51	138.64	74.00	46.82	买入	维持
688766	普冉股份	124.00	-0.64	2.66	3.40	140.64	46.55	36.44	买入	首次
300458	全志科技	54.78	0.04	0.35	0.57	141.64	156.51	96.78	买入	维持
600584	长电科技	40.83	0.82	1.13	1.68	142.64	36.13	24.30	买入	维持
002156	通富微电	31.62	0.11	0.6	0.87	143.64	52.70	36.34	买入	维持

资料来源：同花顺 iFind，东莞证券研究所

注：兆易创新，澜起科技，长电科技，通富微电 2025 年 EPS 为东莞证券研究所预测，其他均采用同花顺一致预测

5.风险提示

国产替代不及预期的风险：若业内企业技术突破或产品导入不及预期导致国产替代进程受阻，则可能面临业绩增速放缓的风险；

价格竞争加剧的风险：若业内上市企业进行大量产能扩张，则行业未来可能面临产能过剩的局面，带来价格竞争导致盈利能力下滑的风险。