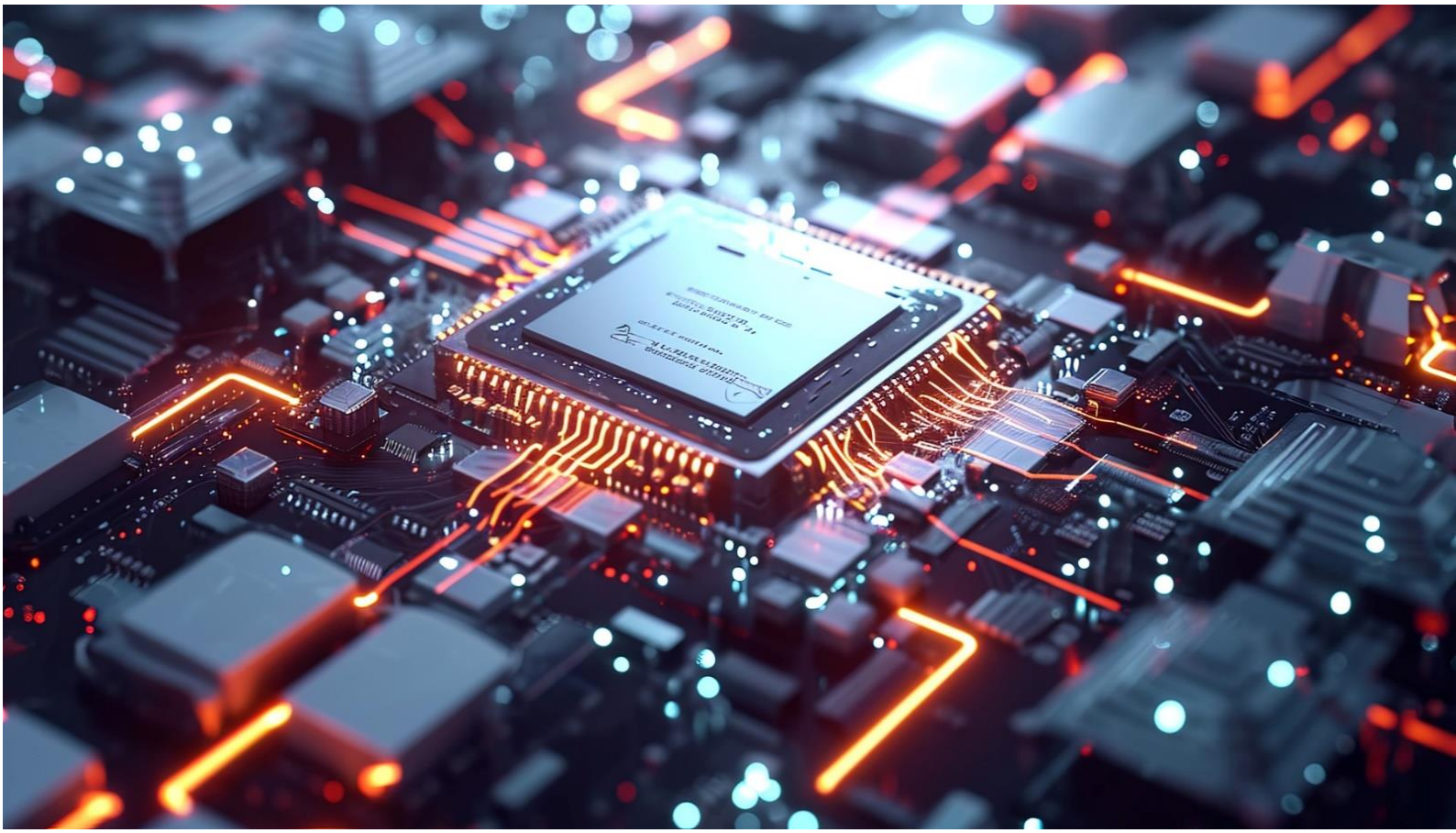
中国银河证券|CGS



重构科技生态，引领智能革命

电子行业首席分析师：高峰

电子行业分析师：王子路　　电子行业分析师：钱德胜

重构科技生态，引领智能革命

2025 年 2 月 18 日

核心观点

- 大模型商业生态推动端侧场景落地。过去 ChatGPT 引领了全球 AI 产业，国内外公司纷纷布局 AI 赛道。国内多家厂商探索商业化路径，在激烈竞争下大模型成本不断降低，推动技术普及，Deepseek-V3 及 R1 大模型凭借性能出色、成本低且开源，获国内厂商适配，加速商业化进程，并为产业发展探索新路径。在端侧应用方面，端侧 AI 能有效赋能终端产品。其中，手机引入 AI 功能升级，手机、PC 有望承载个人大模型需求，提供个性化服务。同时，AI 大模型助力 AR 眼镜升级为智能眼镜，端侧 AI 产业链上下游积极推动 AI 在端侧部署，在 SoC、存储、散热、电池等方面有升级需求。

- 全球云厂商 capex 投入稳步增长，Deepseek 让端侧模型落地成为可能。海外四大 CSP 厂商（亚马逊、微软、谷歌、Meta）2023-2024 年资本开支增速上升，2024 年 Q4 资本开支分别同比增长 29.6%、55.1%、34.4%、88.2%。2024 年全球 AI 服务器约 2050 亿美元，2025 年预计升至 2980 亿美元，占总服务器出货金额比超 70%。2021-2025 年推理负载占比从 55.5%升至 60.8%。端侧 AI 借 NPU 拓展应用场景，2023-2032 年全球设备边缘人工智能市场规模预计以 25.9%复合年增长率增长，随着蒸馏模型能力提升，端侧 SoC 发展将推动 AI 从云端向本地转移。

- AI 引领硬件创新，催化换机需求。消费电子终端换机需求受刚性、诱导性、隐形需求驱动，近年因设备耐用等因素，换机周期放缓。AI 的出现有望逆转这一趋势，智能手机近年同质化严重，2023 年全球换机周期达 51 个月，换机率 23.8%，但 2024-2028 年 AI 手机市场规模预计以 40.5%年均复合增长率扩容，2028 年渗透率达 54%。PC 因强大性能等优势，借 AI 从“工具”升级为“生产力伙伴”，2024-2028 年 AI PC 渗透率预计从 2%提至 65%，出货量从 500 万台增至 1.75 亿台。AI 玩具等借 AI 提升趣味性，2022-2030 年全球市场规模将从 87 亿美元提至 351.1 亿美元，各领域均因 AI 与 Deepseek 迎来新发展契机。

- 投资建议：我们乐观看待 DeepSeek 创新对电子行业的改变。它虽未打破 scaling laws 模式，却加速 AI 应用与硬件普及，加速了边缘部署 LLM 需兼顾性能与效率，看好 AI 推动消费电子换机及相关终端硬件发展，看好蓝思科技、鹏鼎控股、瑞芯微、海光信息等产业链相关公司。

- 风险提示：AI 应用与智能硬件进展不达预期的风险，全球经济疲软需求不及预期的风险，科技自立自强进展不及预期的风险，国际政治环境变动不确定性的风险。

重点公司盈利预测与估值

股票代码	股票名称	EPS			PE			投资评级
		2024E	2025E	2026E	2024E	2025E	2026E	
300433.SZ	蓝思科技	0.8	1.1	1.37	32.61	23.61	18.94	推荐
002938.SZ	鹏鼎控股	1.55	1.94	2.21	26.04	20.82	18.35	推荐
603893.SH	瑞芯微	1.22	1.8	2.48	134.37	91.57	66.25	推荐
688041.SH	海光信息	0.82	1.21	1.64	155.3	106.1	78.2	推荐

资料来源：Wind、中国银河证券研究院

电子行业

推荐 维持

分析师

高峰
☎：010-80927671
✉：gaofeng_yj @chinastock.com.cn
分析师登记编码：S0130522040001

王子路
☎：010-80927632
✉：wangzilu_yj @chinastock.com.cn
分析师登记编码：S0130522050001

钱德胜
☎：021-20252665
✉：qiandesheng_yj @chinastock.com.cn
分析师登记编码：S0130521070001

相对沪深 300 表现图 2025-2-17

资料来源：Wind，中国银河证券研究院

相关研究

目录
Catalog

一、大模型商业生态推动端侧场景落地

（一）AI 大模型的商业生态与格局——价格下探，应用百花齐放

Chat GPT 引领 AI 产业发展进入新阶段。Chat GPT 是 Open AI 在 2022 年底推出的人机对话模型，上线 5 天内收获 100 万用户，不到 2 个月日活量突破 1000 万，成为史上增长最快的消费级应用，引发国内外公司竞相推出 AI 大模型，并在多模态融合、推理与逻辑能力强化和上下文理解与长序列处理优化等多个方向上不断向前迭代。截至 2024 年 12 月 31 日，中国共有 302 款生成式人工智能服务在国家网信办完成备案，其中 2024 年新增 238 款备案。AI 产业发展较此前明显加速，进入繁荣发展新阶段。这一阶段全球多数公司对 AI 的投入集中在大模型的训练，从而拉动相关硬件需求。

图1：OpenAI 发展时间线

资料来源：中国电信网，中国银河证券研究院

各参与方积极探索 AI 大模型商业化路径。中国 AI 通用大模型参与方主要包括云厂商、初创公司、高校研究院和传统 AI 企业，在过去两年中都在积极探索 AI 通用大模型商业化路径。AI 通用大模型的盈利方式分为 B 端和 C 端，在实际应用中，AI 通用大模型的行业参与者通常混合使用多种商业模式。目前，国内 AI 通用大模型在 B 端的商业化模式已日渐清晰，主要以行业定制化解决方案和 Maas 模式为主。C 端的商业化，如订阅付费，尚未达成预期的稳定收益并形成清晰的发展路径。

大模型降价助推商业化进程。短期内 AI 大模型领域涌入众多参与者，行业竞争异常激烈。以阿里云为例，2024 年 5 月阿里云率先调整降价，此后在 9 月、12 月再次降价，行业内其他参与方如百度、字节跳动、智普也选择降价。大模型价格降低推动了 AI 技术在各行业的普及和应用，降低了企业和个人使用 AI 技术的门槛 。

表1：国内大模型降价情况

大模型名称	降价情况
阿里通义千问	阿里云在 2024 年多次调降通义千问相关大模型价格。5 月 21 日，将 Qwen-Long 的 API 输入价格降至 0.0005 元/千 tokens，直降 97%。9 月 19 日，Qwen-Turbo 价格直降 85%，低至每百万输入 tokens 0.3 元，Qwen-Plus 和 Qwen-Max 的输入价格分别再降价 80%和 50%。12 月 31 日，通义千问视觉理解模型 Qwen-VL-Plus 输入价格每千 tokens 从 0.008 元降至 0.0015 元，降幅达 81.3%；Qwen-VL-Max 输入价格每千 tokens 从 0.02 元降至 0.003 元，降幅高达 85%。
百度文心大模型	7 月 5 日，百度智能云宣布文心大模型 4.0 Turbo（ERNIE 4.0 Turbo）面向企业客户全面开放，文心旗舰款模型 ERNIE 4.0 和 ERNIE 3.5 宣布大幅降价。ERNIE 4.0 Turbo 输入输出价格分别低至 0.03 元/千 Tokens、0.06 元/千 Tokens。此外，ERNIE 4.0、ERNIE 3.5 两款旗舰模型大幅降价，ERNIE Speed、ERNIE Lite 两款主力模型持续免费。
智谱大模型	6 月 5 日，在智谱 AI OpenDay 上，智谱 AI 宣布全模型矩阵降价，其中 GLM-3-Turbo 降价后价格为 0.6 元/百万 tokens。2025 年 1 月 23 日，智谱又推出包括 GLM-Realtime、GLM-4-Air、GLM-4V-Plus 在内的三大新模型，并宣布降价 50%。

腾讯混元大模型	5 月 22 日，腾讯云加入大模型降价阵营，宣布混元大模型全面降价。主力模型之一混元-lite 模型价格从 0.008 元/千 tokens 调整为全面免费，同时，API 输入输出总长度计划从目前的 4k 升级到 256k。混元- standard API 输入价格从 0.01 元/千 tokens 降至 0.0045 元/千 tokens，下降 55%，API 输出价格从 0.01 元/千 tokens 降至 0.005 元/千 tokens，下降 50%。新上线的混元-standard-256k API 输入价格下调至 0.015 元/千 tokens，下降 87.5%，API 输出价格下降至 0.06 元/千 tokens，下降 50%。腾讯最高配置的万亿参数模型混元-pro，API 输入价格从 0.1 元/千 tokens 降至 0.03 元/千 tokens，降幅达 70%。

资料来源：腾讯网、华尔街见闻、搜狐网，中国银河证券研究院

Deepseek-V3 及 R1 大模型问世有望加速 AI 大模型商业化。2024 年 12 月，DeepSeek 发布大语言模型 V3，同时宣布开源，测试结果显示，V3 的多项评测成绩超越 Qwen2.5-72B 和 Llama-3.1-405B 等其他开源模型，甚至可以与 GPT-4o、Claude 3.5-Sonnet 等闭源模型相媲美。

DeepSeek 的模型定价显著低于 Open AI。DeepSeek-V3 及 R1 问世以来，国内 GPU 厂商如华为昇腾、海光信息、沐曦、摩尔线程和天数智芯等国内 GPU 厂商适配其大模型，云厂商如阿里云、腾讯云、华为云、京东云均上线 DeepSeek 模型，支持公有云和私有化部署，国内相关生态呈现百花齐放，有望加速 AI 大模型商业化进程。

图2：DeepSeek-V3 多项评测成绩领先

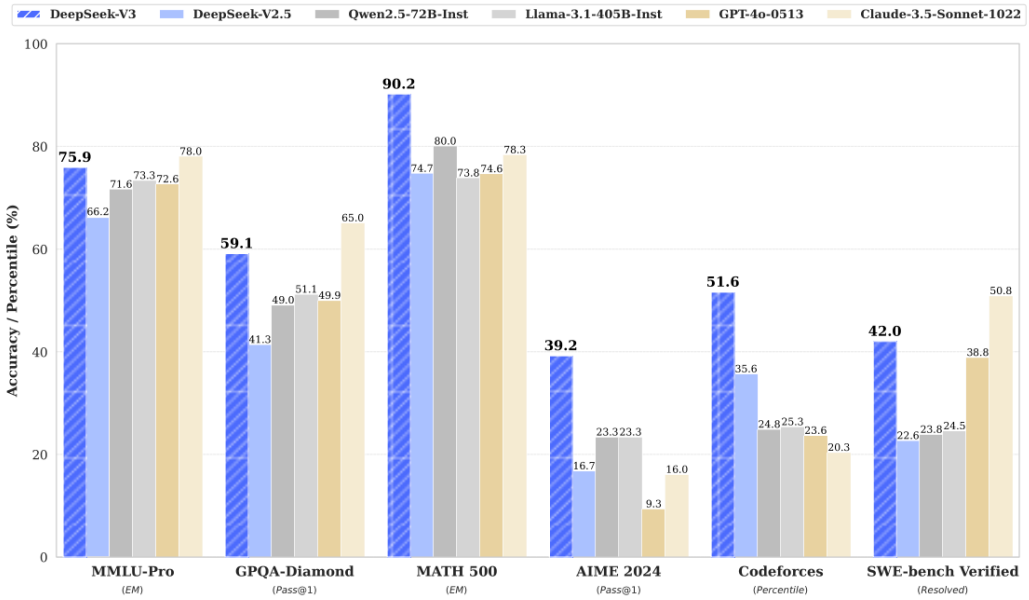


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

资料来源：《DeepSeek-V3 Technical Report》DeepSeek，中国银河证券研究院

2022 年底 ChatGPT 问世至今，AI 产业发展明显加速，各参与方积极投资硬件，同时探索商业化路径。期间，由于美国对中国技术封锁，国内 AI 产业发展略微滞后。Deepseek-V3 及 R1 大模型的出现尽管没有打破 scaling law，但是为产业发展探索出新的路径，即在有限算力的前提下，通过算法优化和工程创新可以得到相同效果的大模型。DeepSeek 选择开源有望加速大模型商业化进程。

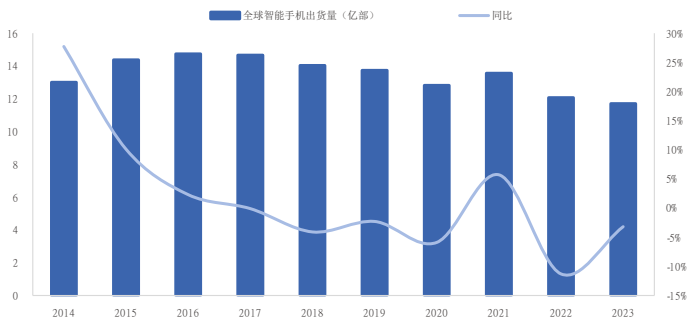
（二）端侧垂直应用落地与场景需求

端侧 AI 是指在终端设备上运行 AI 模型，相较于云端大模型，端侧模型在资源有效的设备上高效运行，需要进一步多模型进行压缩、推理加速和能耗优化。将大模型部署在端侧设备，可高效赋能智能终端，如降低延迟，更快响应用户请求；隐私保护，减少数据传输，从而降低隐私泄露的风险；减轻云端服务器的计算负担，降低对中心化计算资源的依赖，从而降低成本；根据用户的具体设备和使用习惯进行定制化优化，提供更加个性化的服务；部分场景下不需要连接网络，提高了应用的灵活性。

AI 加持下，手机将由工具升级为助手。手机的发展从早期的功能机过渡现阶段的智能机，目前各品牌智能手机高度同质化，缺乏创新也导致消费者换机周期拉长。各品牌厂商将 AI 视为下一轮手机升级的主要方向，苹果在 iOS 18 中引入全新的 AI 功能，全新的 Siri 采用了先进的大型语言模型，新版本的 Siri 能够更准确理解用户的意图。华为在 HDC 2024 上提出 Harmony Intelligence，

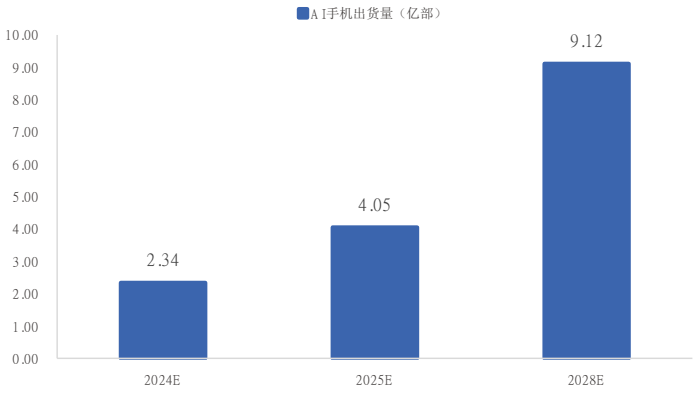
搭载了大模型的小艺，在鸿蒙原生智能框架下升级为系统级智能体，其交互和意图理解能力增强。IDC 预计 2024 年全球生成式人工智能手机出货量为 2.34 亿部，在全球智能手机出货量占比为 19%，2024 年至 2028 年全球 AI 手机复合增速将达到 40.5%。

图3：全球智能手机出货量保持稳定



资料来源：IDC、中国银河证券研究院

图4：AI 手机出货量有望快速增长



资料来源：IDC、中国银河证券研究院

AI PC 有望承载个人大模型需求。普通用户不仅需要公共的大模型服务，也需要专属的个人大模型。个人大模型的普及，必然带来用户对大模型的专属化需求的提高，而云端公共大模型无法满足消费者多样化的需求，专属化的成本相对较高。在众多消费电子产品中，PC 有望成为适合承载大模型的理想平台，其主要优势包括：

PC 具备全模态的人机自然交互条件。PC 拥有最多样化交互方式的终端设备，既包括相对直接触控交互、语音交互、手势控制，也具备更加复杂的键鼠交互、数字笔交互等。多元化的交互方式使得 PC 在承载创新的 AI 交互方式方面具备巨大潜力。

PC 是承载最多场景的个人通用设备。PC 作为一种通用生产力平台，既能够承载以消费内容为主的生活娱乐场景，也能够承载以创作内容为主的工作、学习等场景。

PC 是迄今为止最强的个人计算平台。PC 自诞生以来始终代表了个人计算平台的能力巅峰，PC 的通用计算能力强劲，并得到长期优化，在性能、成本、体验方面达到最佳配置，是个人计算设备中拥有最强性能的通用计算平台。

PC 是存储容量最大、最受信赖的安全终端。随着用户使用 AI 应用的频次提高，个人交互数据量快速增加，数据安全性和隐私保护的重要性日益凸显。PC 拥有大容量的本地安全存储，用户在本地终端设备上进行数据分析、模型推理与计算，个人数据不再需要存储在云端或远程服务器，可以安全地保留在用户的设备，提高了数据的安全性。

图5：联想 AI PC 五大特性

资料来源：联想官网，中国银河证券研究院

AI PC 能够为用户提供通用场景下的个性化服务，提供即时、可靠的服务响应，更低的大模型使用成本以及可信、安全的个人数据和隐私保障。针对工作、学习等场景，AI PC 可提供个性化创作服务、设备管家服务等在内的个性化服务。响应速度慢、反馈时间长是用户对 AIGC 平台使用的主要负面反馈，AI PC 以本地推理为主，边缘和云端推理为辅，能够在混合算力、混合模型之间智能、合理调配任务，有效缩减响应时间，离线状态下的可操作性也是 AI PC 的优势。AI PC 的消费者一

次性购买后即可享受全生命周期的本地免费推理服务，再加上有限的云端订阅，可显著降低用户使用 AI 大模型的成本，同时也降低了带宽成本。AI PC 有专门用于存储用户的特定类型文件与数据的安全空间，确保个人用户与企业用户的隐私与涉密信息能在本地实现安全隔离，仅在受信任的环境下才可以被调用。

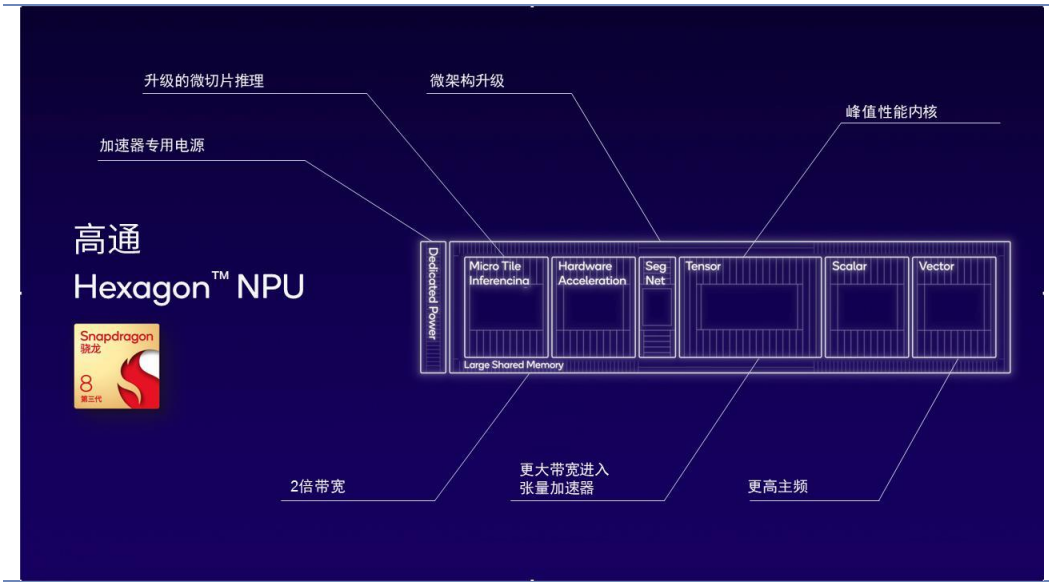
AI 大模型加持，AR 眼镜升级为智能眼镜。据 VR 陀螺不完全统计，2024 年全球已公开亮相、正式发布的智能眼镜系列产品达到了 46 款。其中，聚焦 AI 音频、拍摄的眼镜为 9 款，AI+AR 眼镜、AR 信息提示眼镜为 22 款，分体式 XR 眼镜、观影眼镜为 15 款。2023 年已有部分厂商为 AR 眼镜提供基础的 AI 语音交互，2024 年部分厂商开始在 AR 眼镜上增加摄像头，以满足多模态交互的可能性。对于传统的 AR 厂商而言，AI 大模型加持让 AR 眼镜升级为智能眼镜。以 Meta 和雷朋合作的眼镜为例，产品虽然在 2023 年发布，但多模态的 AI 功能，从 2024 年才开始陆续推送到用户侧。Ray-Ban Meta 的多模态 AI 功能，通过眼镜上的摄像头，可以让用户实现如翻译眼前的法语菜单、通过语音指令查询实时信息、识别街道建筑等。

（三）端侧产业链全景图

目前，端侧 AI 芯片厂商、中游模组及软件厂商、下游终端厂商的产业链上下游正积极推动 AI 在端侧部署落地。其中，AI 手机的核心变化是 SoC 升级。

手机 SoC 厂商重点布局 AI。在功耗和散热受限的终端上使用通用 CPU 和 GPU 服务平台，难以满足 AI 用例严苛且多样化的计算需求；同时，AI 用例在不断演进，在功能完全固定的硬件上部署这些用例不切实际。因此，支持处理多样性的异构计算架构能够发挥每个处理器的优势，例如以 AI 为中心的定制设计的 NPU，以及 CPU 和 GPU。每个处理器擅长处理不同的任务：CPU 擅长顺序控制和即时性，GPU 适合并行数据流处理，NPU 擅长标量、向量和张量数学运算，可用于核心 AI 工作负载。高通在 2015 年正式推出的骁龙 820 处理器集成首个高通 AI 引擎，支持成像、音频和传感器运算，2018 年高通在骁龙 855 中为 Hexagon NPU 增加了 Hexagon 张量加速器。2019 年高通在骁龙 865 上扩展了终端侧 AI 用例，包括 AI 成像、AI 视频、AI 语音和始终在线的感知功能。2020 年高通凭借 Hexagon NPU 变革型的架构更新，融合标量、向量和张量加速器，同时还为加速器打造了专用大共享内存。2021 年推出的骁龙 8 Gen 1 AI 算力为 9 INT8 TOPS，骁龙 8 Gen2 AI 算力提升 4.35 倍；骁龙 8 Gen3 的 NPU 算力进一步提升 98%。2017 年华为发布的麒麟 970 首次集成人工智能专用 NPU 神经网络单元，同年苹果推出的 A11 处理器首次引入神经网络引擎（Neural Engine）。

图6：第三代骁龙 8 的 Hexagon NPU 升级以低功耗实现领先的生成式 AI 性能



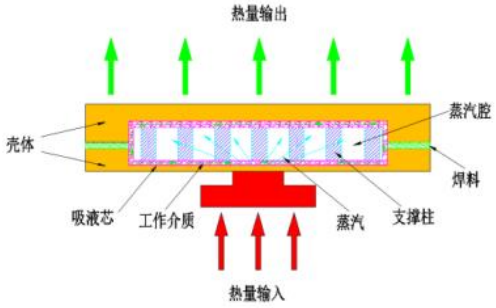
资料来源：《通过 NPU 和异构计算开启终端侧生成式 AI》高通，中国银河证券研究院

满足 AI 需求需要升级手机内存。在手机上运行 10 亿参数的大模型至少需要 1GB 内存，运行 70 亿参数大模型需 4GB 内存，运行 130 亿参数大模型需要超过 7GB 内存。在处理数据时，需要在手机内存中进行临时的存储和运算。AI 手机可能支持多个 AI 任务运行，需要更多内存支持多个任务同时运行。模型存储需求、数据处理需求和多任务处理需求驱动 AI 手机内存容量提升，安卓旗舰手机从 2023 年开始逐步淘汰 8GB 内存，以 12GB 及以上内存配置为新标配，部分高端机型已配备 24GB 内存。此外，AI 运算需要快速读取和存储数据，对内存的读写速度要求更高；更大的内存带宽能够确保 CPU、GPU 和 NPU 等硬件和内存之间的数据传输更加顺畅。

AI 手机对散热系统要求更高。为了保证手机在高负荷运行 AI 任务时的性能和稳定性，需要散热系统具备更强的散热能力。传统的散热方式可能无法及时有效地将热量散发出去，因此 AI 手机通常需要采用更先进的散热技术和材料，如更大面积的散热石墨片、散热铜管、均热板等，以提高散热效率。

图7：均温板工作原理

均温板



均温板工作原理

资料来源：苏州天脉招股说明书，中国银河证券研究院

提升电池容量和发展快充技术以满足 AI 手机需求。高计算需求、多任务处理和数据传输会加快 AI 手机电量消耗，为了满足 AI 手机对电量的高需求，手机厂商会倾向于采用更大容量的电池，以延长手机的续航时间。例如，部分高端 AI 手机的电池容量已经从过去的 3000mAh-4000mAh 提升到 4500mAh-5000mAh。碳硅电池比传统的石墨电池能实现更高的能量密度，从而实现长时间续航。同时，为了弥补 AI 手机电量消耗快的问题，快充技术得到了更广泛的应用和发展。

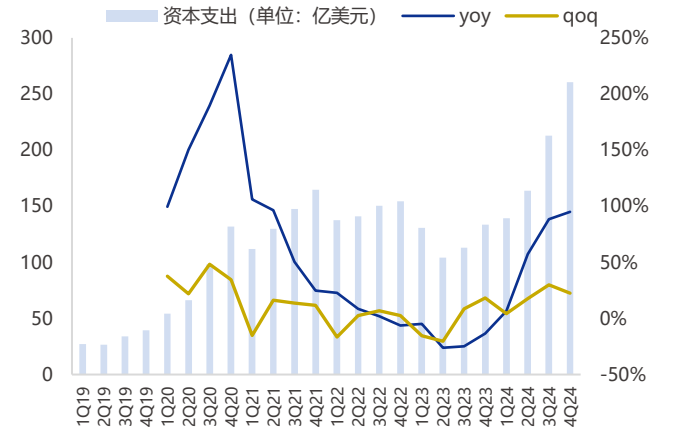
综上，端侧 AI 最大的变化在于在 SoC 升级，其次存储容量的提升以及带宽、传输速率的提升也至关重要，散热方案的升级、电池容量的提升需要同步进行，推荐恒玄科技、全志科技、晶晨股份、兆易创新、苏州天脉、领益智造、欣旺达、德赛电池。

二、全球云厂商 capex 投入稳步增长，Deepseek 让端侧模型落地成为可能

（一）全球云厂商 Capex 投入持续攀升，算力需求呈现高速增长

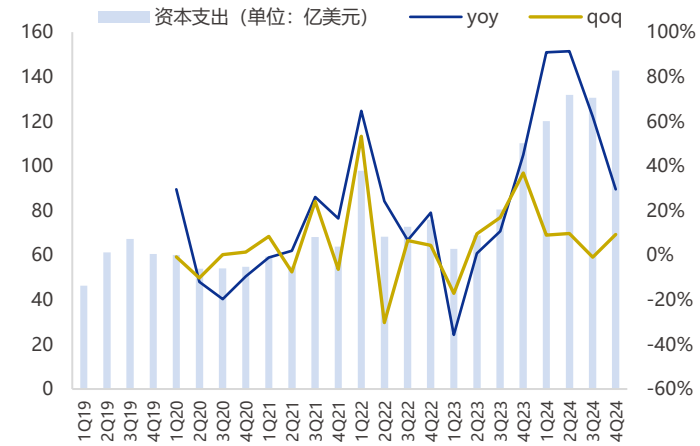
国外大型云服务厂商近两年资本开支快速增长。从最近几个季度的数据能明显看出，海外四大 CSP 厂商亚马逊、微软、谷歌、Meta 的资本开支增长强劲，尤其是从 2023 年到 2024 年期间，资本开支增速呈现出明显的上升趋势，AI 算力需求的持续紧缺是重要的驱动因素，这些公司都在加大在 AI 领域的投入，加快部署 AI 数据中心以抢占市场，对服务器、数据中心和网络基础设施等方面的投资成为资本开支的重点。从具体数据来看，亚马逊、微软、谷歌、Meta 的 4Q24 的资本开支分别为 260.52、142.76、138.73、144.25 亿美元，同比分别增长 29.6%、55.1%、34.4%、88.2%，环比分别增长 9.3%、26.7%、35.2%、74.7%。海外对于算力投入均在高速增长。

图8：亚马逊近 5 年资本支出情况



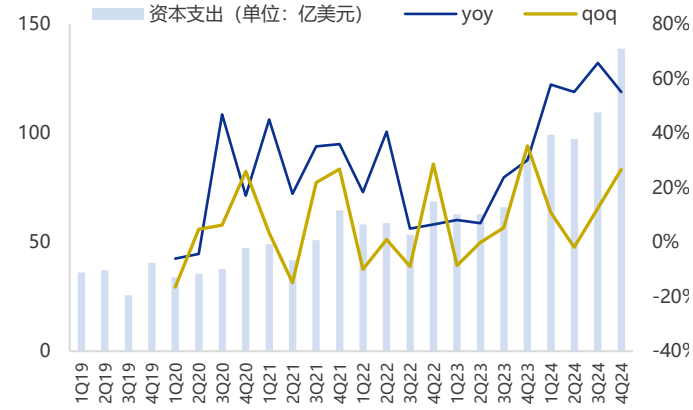
资料来源：Wind，中国银河证券研究院

图9：谷歌近 5 年资本支出情况



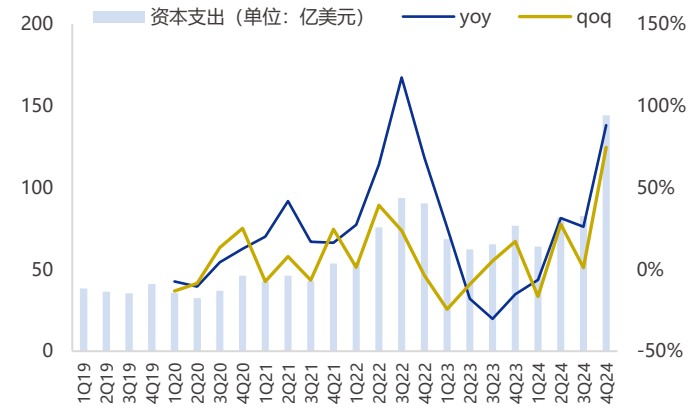
资料来源：Wind，中国银河证券研究院

图10：微软近 5 年资本支出情况



资料来源：Wind，中国银河证券研究院

图11：META 近 5 年资本支出情况



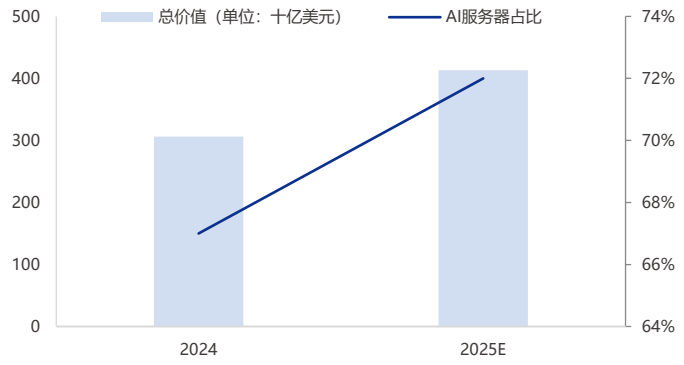
资料来源：Wind，中国银河证券研究院

全球 AI 服务器市场规模呈现出迅猛增长的态势。根据 TrendForce 的最新研究，预计 2024 年整个服务器行业的总价值将达到 3060 亿美元。其中，与 AI 服务器相关的行业价值估计约为 2050 亿美元，与标准服务器相关的行业价值相比，增长更为强劲。展望 2025 年，由于需求持续旺盛且产品平均售价较高，预计 AI 服务器细分市场的价值将升至 2980 亿美元。此外，预计到 2025 年，AI 服务器将占整个服务器行业总价值的 70% 以上。

从行业规模来看，2023 年中国智能算力总规模已经达到了 389.1EFlops，预计到 2027 年，规模将会增长至 1694.9EFlops。从算力结构的视角来看，虽然目前训练算力占比仍然比较多，但随着大模型应用的不断加深和推广，推理算力的需求正在逐步增加。智能芯片架构的快速迭代升级，单张芯片可承载的算力快速增长；智能芯片厂商加紧增加产能，保证产量的供给，但高端算力的供给缺口仍然严重。应用端的人工智能需求迅速爆发并向下传导，影响了各企业对于自身数智业务布局。

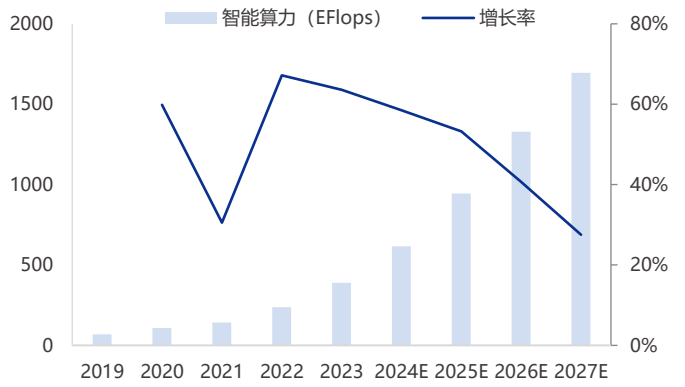
各行业的企业都积极探索人工智能潜在应用场景。对算力资源的采购上也有所倾斜，整体算力仍处于供不应求的状态。

图12：2024-2025 年 AI 服务器总价值和占比情况



资料来源：Trendforce，中国银河证券研究院

图13：2019-2027 年中国智能算力规模及增速(浮点运算口径)



资料来源：艾瑞咨询，中国银河证券研究院

在数字经济蓬勃发展的当下，政策扶持是推动中国 AI 服务器市场规模扩张的关键。国家接连出台政策支持 AI 产业，为 AI 服务器行业的高速增长注入动力。

相关企业纷纷抢占先机，加速在 AI 服务器领域布局，从技术研发到产能扩充全方位发力。同时，人工智能在智能安防、智慧医疗等领域的应用场景逐步落地，这使得 AI 服务器在整体服务器市场中的占比稳步提升。

为满足不断攀升的计算和数据处理需求，中国企业与研究机构积极投身技术研发创新。在高性能处理器、大容量内存、高速存储器和高效冷却系统等领域，全力突破技术瓶颈，探索新架构、新材料，提升自主创新能力与核心竞争力，为产业长远发展奠定基础。

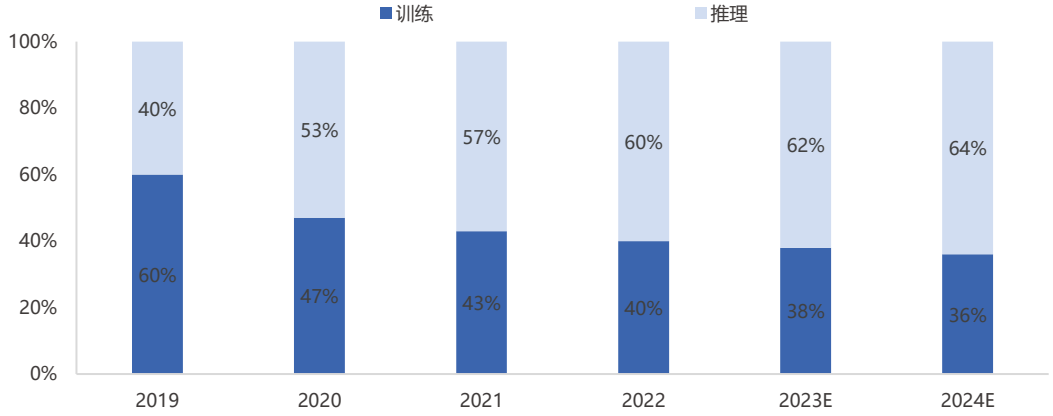
表2：人工智能相关政策

发文机关	发布时间	主要内容
全国网络安全标准化技术委员会	2024/9/9	框架以鼓励人工智能创新发展为第一要务，以有效防范化解人工智能安全风险为出发点和落脚点，提出了包容审慎、确保安全，风险导向、敏捷治理，技管结合、协同应对，开放合作、共治共享等人工智能安全治理的原则。
第 78 届联合国大会	2024/7/1	第 78 届联合国大会上，协商一致通过中国主提的加强人工智能能力建设国际合作决议，140 多国参加决议联署。该决议强调人工智能发展应坚持以人为本、智能向善、造福人类的原则，鼓励通过国际合作和实际行动帮助各国特别是发展中国家加强人工智能能力建设，增强发展中国家在人工智能全球治理中的代表性和发言权，倡导开放、公平、非歧视的商业环境，支持联合国在国际合作中发挥中心作用，实现人工智能包容普惠可持续发展，助力实现联合国 2030 年可持续发展议程。
国家药监局	2024/6/13	《清单》列出了 15 个具有引领示范性的、有发展潜力的、针对工作痛点的、需求较为迫切的应用场景，旨在推动人工智能技术在药品监管领域的研究探索，以促进人工智能与药品监管深度融合为主线。
工业和信息化部、中央网络安全和信息化委员会办公室等	2024/6/5	到 2026 年，我国人工智能产业标准与产业科技创新的联动水平持续提升，新制定国家标准和行业标准 50 项以上，引领人工智能产业高质量发展的标准体系加快形成。
市场监管总局、中央网信办、国家发展改革委等	2024/3/18	在集成电路、半导体材料、生物技术、种质资源、特种橡胶，以及人工智能、智能网联汽车、北斗规模应用等关键领域集中攻关，加快研制一批重要技术标准。

资料来源：全国网络安全标准化技术委员会，国家药监局，工业和信息化部、中央网络安全和信息化委员会办公室等，中国银河证券研究院

从我国 AI 服务器推理和训练工作负载情况来看，据统计，2021 年我国 AI 服务器推理负载占比约 55.5%，未来有望持续提高，预计到 2025 年我国 AI 服务器推理负载占比达 60.8%。

图14：2019-2025 年中国 AI 服务器推理和训练工作负载情况



资料来源：华经产业研究，中国银河证券研究院

（二）端侧 AI 快速发展，DeepSeek 让端侧模型低成本成为可能

AI 计算主要依赖于云端，但云端计算存在延迟和数据隐私的问题。而端侧 AI 可以在本地处理数据和任务，实现快速响应，无需将数据传输到云端，从而提升用户体验并保护隐私。随着 NPU 的广泛应用，端侧设备逐渐具备了处理 AI 任务的能力。NPU 是专为神经网络计算设计的加速器，与传统 CPU 和 GPU 相比，它在执行 AI 模型时效率更高、功耗更低，适合资源受限的设备。NPU 技术的进步推动了终端设备在语音识别、图像处理和自然语言处理等多模态任务上的性能提升。例如，现代智能手机可以利用内置 NPU 实现实时物体识别，帮助用户管理和分类照片。

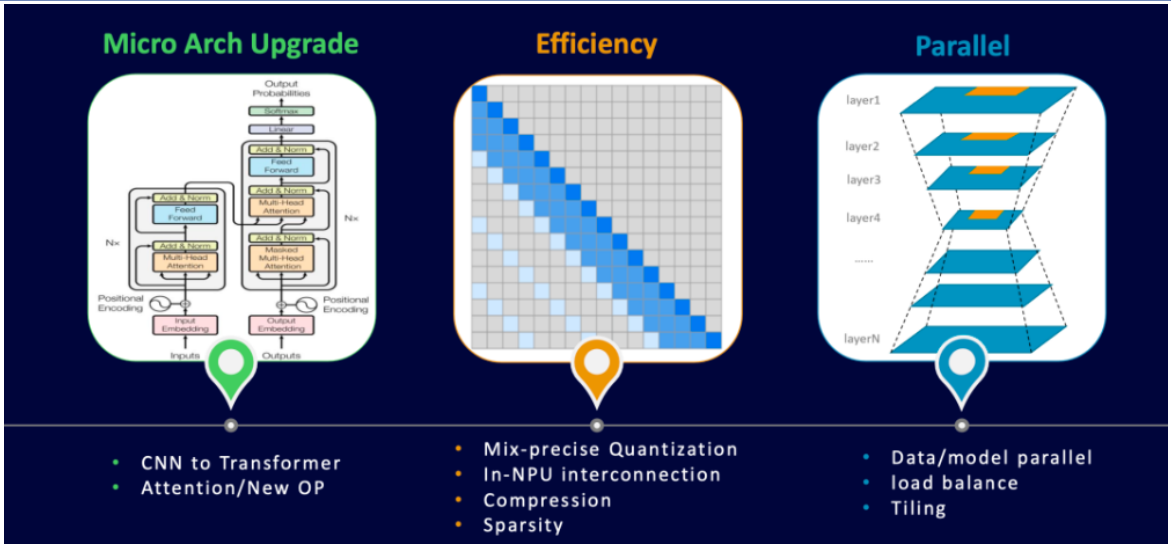
图15：端侧模型的目前发展情况

资料来源：ARM China，Canalys，中国银河证券研究院

目前，端侧大模型主要发展语言模型和文生图模型，未来将向多模态领域拓展。多模态模型能处理语言、图像、音频、视频等多种数据，满足多样应用需求。比如用户能用自然语言和图片与设备交互，获得更智能体验。这不仅推动人机交互进步，还让终端更懂用户需求。端侧 AI 的个性化也是未来重点。通过在设备上个性化训练，它能适应不同用户习惯，提供定制服务。像手机 AI 学习用户拍照偏好后，能自动调参数、推荐拍摄模式。这既能提升用户依赖度，也为终端厂商带来商业机遇。

端侧 AI 的应用场景从智能手机、PC 等拓展到可穿戴设备。随着 NPU 技术提升，可穿戴设备虽体积小巧，却能实现语音识别、健康监测和图像处理等功能。比如搭载 NPU 的智能手表，能本地处理语音命令，使用更便捷。新兴的智能眼镜等可穿戴设备，借助端侧 AI 实现实时翻译、情绪识别等功能，拓宽了应用边界。在可穿戴设备中引入 AI，厂商能打造更具吸引力的产品，在市场竞争中占据优势。像兼具实时语音翻译和心率监测功能的智能手表，在出国旅行、健康管理等场景实用性强，提升了用户生活质量。

图16：端侧 AI 赋能可穿戴等新兴设备



资料来源：ARM China，中国银河证券研究院

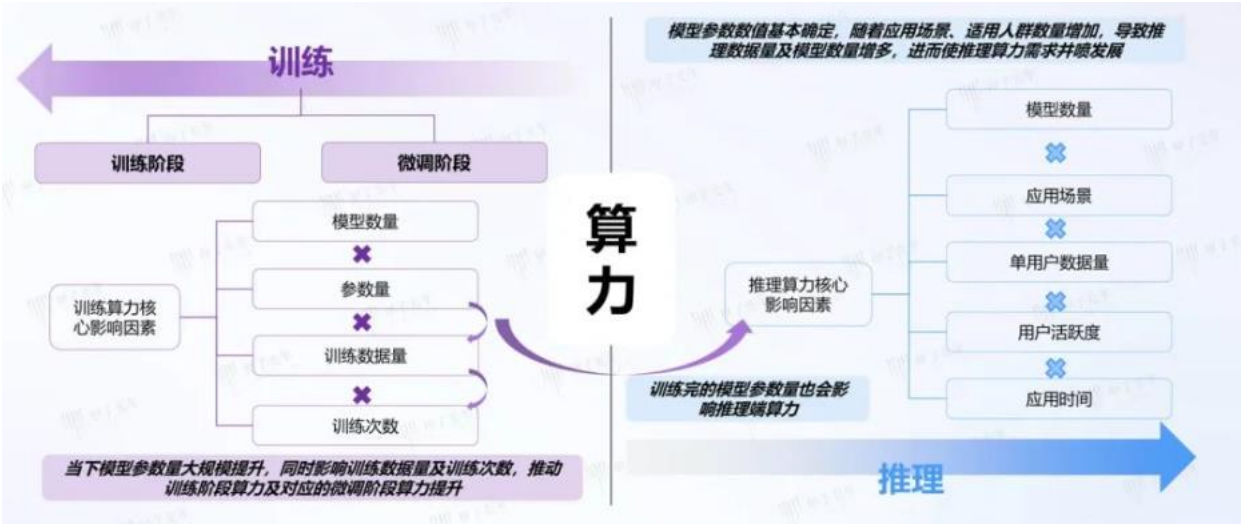
智算在 AI 深度学习领域发挥着举足轻重的作用，主要应用体现在以下关键方面：

加速模型训练：深度学习模型结构复杂，参数数量庞大，训练过程对计算资源需求极高。智能算力借助 GPU（图形处理器）、TPU（张量处理单元）这类高性能计算设备，显著加快了模型训练进程。以 GPT-4 等大型语言模型为例，训练时需要处理海量文本数据，智能算力的强大运算能力可快速完成复杂的矩阵运算，大幅缩短训练时间。

优化推断推理：在模型推断阶段，智能算力通过高性能计算设备与专门的推理芯片，显著加速了深度学习模型的推断过程。这使得模型能够在更短时间内处理输入数据并输出结果，大幅提升了模型的实时性和稳定性。在智能安防领域，实时视频监控的目标检测任务需要快速响应，智能算力可确保检测结果及时准确输出。

助力模型优化：借助智能算力，能够对模型进行自动化的超参数调优，精准找到最适配的参数组合；开展网络结构搜索，探寻更高效的模型架构；实施模型剪枝，去除冗余连接，从而进一步提升模型的精度和效率。

图17：模型的算力需求主要来自于训练和推理两大类



资料来源：甲子光年，中国银河证券研究院

人工智能芯片的架构与应用类型存在较大区别。架构上，传统计算机架构含 GPU、FPGA、ASIC。GPU 用于图像处理，适合数据密集计算；FPGA 可无限次编程，开发时间短；ASIC 依特定需求定制，能效高。类脑计算架构的 NPU 模仿大脑神经结构，多为研究型芯片，商业化程度低。应用方面，训练算力对芯片算力要求高，多为 FP32 与 FP16 精度；推理算力对芯片算力要求低，多为 FP32 与 FP64 精度，强调低延时等特性。

Deepseek 浪潮席卷 AI 领域，端侧成本逐步下降。2025 年 1 月 20 日，DeepSeek 正式发布其 DeepSeek-R1，最大的亮点：1，性能比肩 OpenAI o1 正式版。2，价格为每百万输入 tokens 1 元（缓存命中）/4 元（缓存未命中），每百万输出 tokens 16 元。而且在开源 DeepSeek-R1-Zero 和 DeepSeek-R1 两个 660B 模型的同时，通过 DeepSeek-R1 的输出，蒸馏了 6 个小模型开源给社

区，其中 32B 和 70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果。更强的性能，更低的训练与推理成本，将加速推动 AI 应用与硬件的普及和落地。

图18：DeepSeek-V3 在不同测试集的表现

测试集		DeepSeek-V3	Qwen2.5 72B-Inst.	Llama3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513
	模型架构	MoE	Dense	Dense	-	-
	# 激活参数	37B	72B	405B	-	-
	# 总参数	671B	72B	405B	-	-
英文	MMLU (EM)	88.5	85.3	88.6	88.3	87.2
	MMLU-Redux (EM)	89.1	85.6	86.2	88.9	88
	MMLU-Pro (EM)	75.9	71.6	73.3	78	72.6
	DROP (3-shot F1)	91.6	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	84.1	86	86.5	84.3
	GPQA-Diamond (Pass@1)	59.1	49	51.1	65	49.9
	SimpleQA (Correct)	24.9	9.1	17.1	28.4	38.2
	FRAMES (Acc.)	73.3	69.8	70	72.5	80.5
	LongBench v2 (Acc.)	48.7	39.4	36.1	41	48.1
代码	HumanEval-Mul (Pass@1)	82.6	77.3	77.2	81.7	80.5
	LiveCodeBench(Pass@1-COT)	40.5	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	37.6	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	51.6	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42	23.8	24.5	50.8	38.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	84.2	72.9
	Aider-Polyglot (Acc.)	49.6	7.6	5.8	45.3	16
数学	AIME 2024 (Pass@1)	39.2	23.3	23.3	16	9.3
	MATH-500 (EM)	90.2	80	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	43.2	15.9	6.8	13.1	10.8
中文	CLUEWSC (EM)	90.9	91.4	84.7	85.4	87.9
	C-Eval (EM)	86.5	86.1	61.5	76.7	76
	C-SimpleQA (Correct)	64.1	48.4	50.4	51.3	59.3

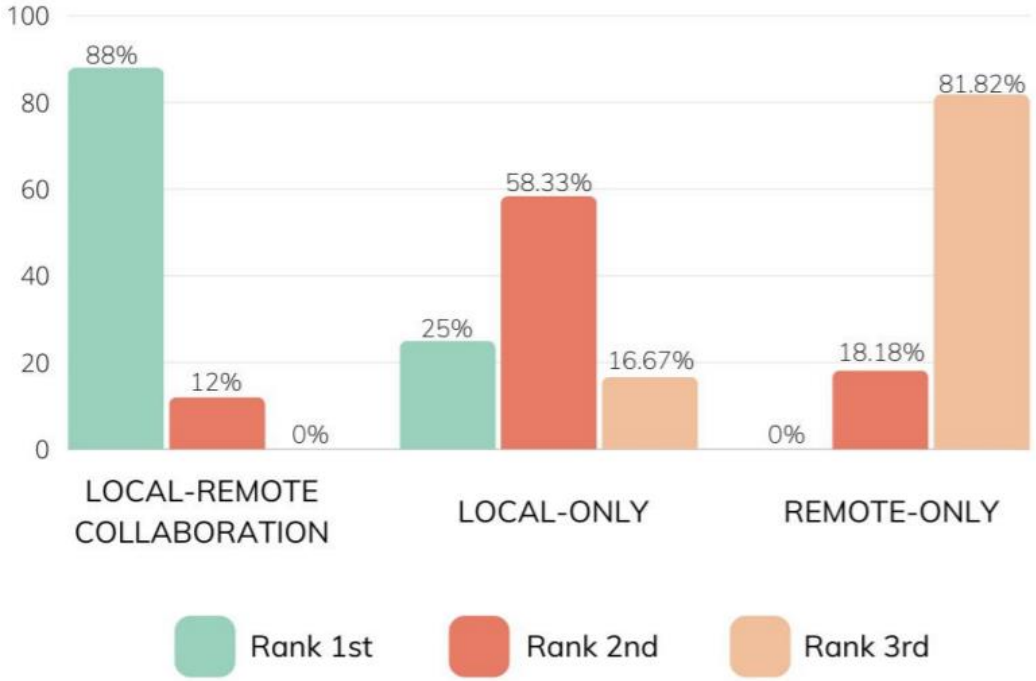
资料来源：DeepSeek 官微，中国银河证券研究院

借助算法能力，DeepSeek 降低 AI 普及成本。借助算法创新，如 FP8 低精度训练，大幅降低大模型训练成本，减少内存占用与通信开销，让国产 GPU 在算力不足时也能高效训练。推动开源，吸引全球开发者，降低模型使用门槛，助力 AI 技术向产业渗透。端侧应用爆发推高算力需求，其推理端低成本优势或推动市场格局转变，让端侧模型低成本成为可能。

（三）端侧模型快速发展，端侧 SoC 未来发展呈现新趋势

LLM 单纯云端部署（例如 ChatGPT）并不广泛接受。如下图统计所示，88%的参与者倾向于边缘-云协作架构，其中 58.33%支持本地部署，81.82%对现有的仅云端解决方案不满意。他们的主要担忧是:1)远程大型语言模型服务的高延迟，2)将个人数据传输到云端的风险，3)云端大型语言模型服务的成本。

图19：个人对不同 LLM 部署策略的投票分布

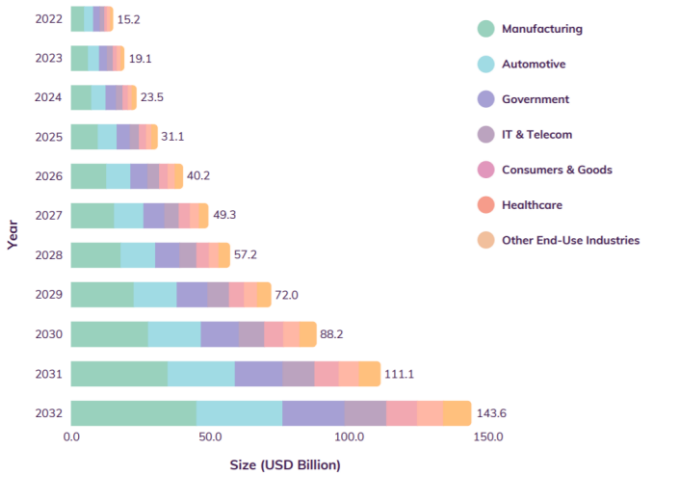


资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

2023 年边缘大型语言模型开始陆续爆发，当时出现了几个参数量低于 10B 的模型，使其能在边缘设备上运行，包括 meta 的 LLaMA 系列，微软的 Phi 系列，智谱的 ChatGLM，阿里巴巴的 Qwen 等。进入 2024 年创新步伐加快，边缘端部署的优势是能够缩短响应时间，并直接应用在如手机、汽车、可穿戴设备上。2022 年至 2032 年，按终端用户划分的全球设备边缘人工智能市场规模。市场将以 25.9%的复合年增长率增长，预计 2032 年的市场规模为 1436 亿美元。

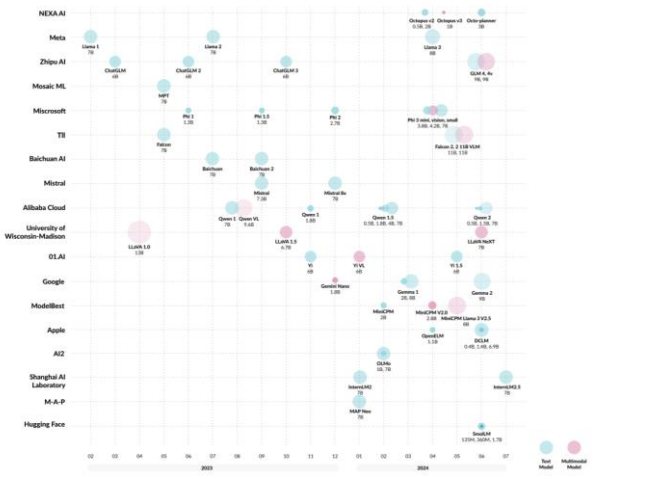
尽管在边缘端部署大模型有诸多优势，但考虑到端侧有限的计算能力、存储能力和能源限制等，使得直接部署基于云端的 LLM 困难重重。再评估设备端大型语言模型的性能时，有几个关键指标需要考虑：延迟、推理速度、内存使用、存储和能耗。通过优化这些性能指标，设备端大型语言模型能够在更广泛的场景中高效运行，提供更好的用户体验。同时针对边缘设备的部署，在保持性能的同时提高计算效率至关重要，通过量化、剪枝、知识蒸馏和低秩分解，这些方法通过平衡性能、内存占用和推理速度来提高大语言模型的运行效率，确保其在设备端应用中的可行性。

图20：边缘 AI 的市场规模（十亿美金）



资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

图21：端侧大语言模型的演变



资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

近年来，人工智能技术的迅猛发展和移动设备硬件的不断升级，使得在边缘设备上部署大型语言模型成为可能。作为人们日常生活中最常用的设备，智能手机上的语言模型引人注目。目前，全球主要手机品牌已开发并发布了多款先进的模型，这些模型采用设备端部署或设备-云协同策略。

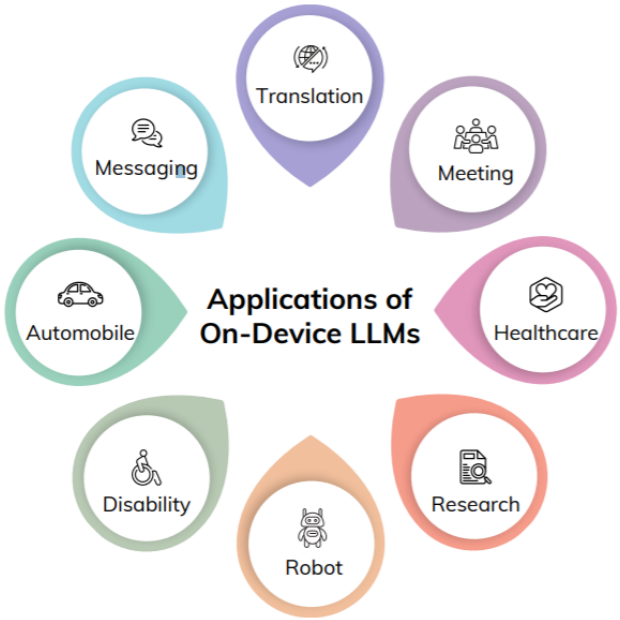
图22：手机厂商发布的设备端 LLM

Year	MODEL NAME	Model Size	Edge	Cloud
2023	Google Gemini Nano	7B	√	
2023	OPPO AndesGPT	7B	√	√
2024	Honor MagicLM	7B	√	
2024	VIVO BlueLM	7B	√	√
2024	XiaoMi MiLM	6B	√	
2024	Apple OpenELM	1.1B	√	√

资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

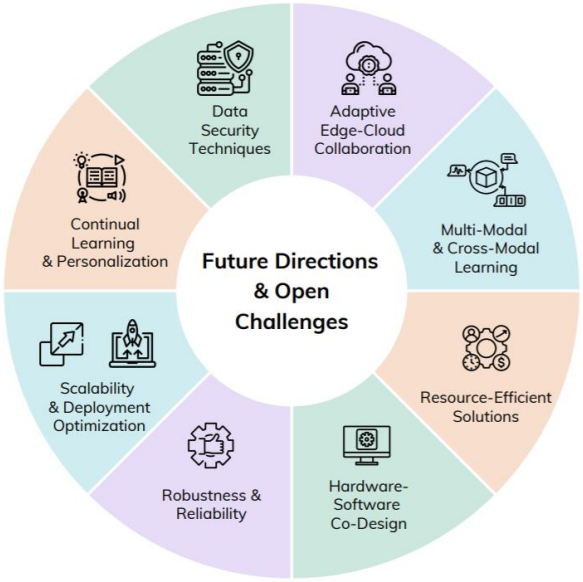
设备端语言模型正开启一个智能、响应迅速、个性化应用的新时代。通过将先进的自然语言处理能力直接引入用户设备，这些模型正在改变人们与技术互动的方式。从即时消息建议到实时语言翻译，从保密医疗咨询到尖端自动驾驶汽车。在资源受限设备上部署 LLM 面临独特挑战，这些挑战与传统的基于云的实施有显著不同。这些挑战涉及多个领域，包括模型压缩、高效推理、安全性、能源效率，以及与多样化硬件平台的无缝集成等。

图23：端侧 LLM 的应用



资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

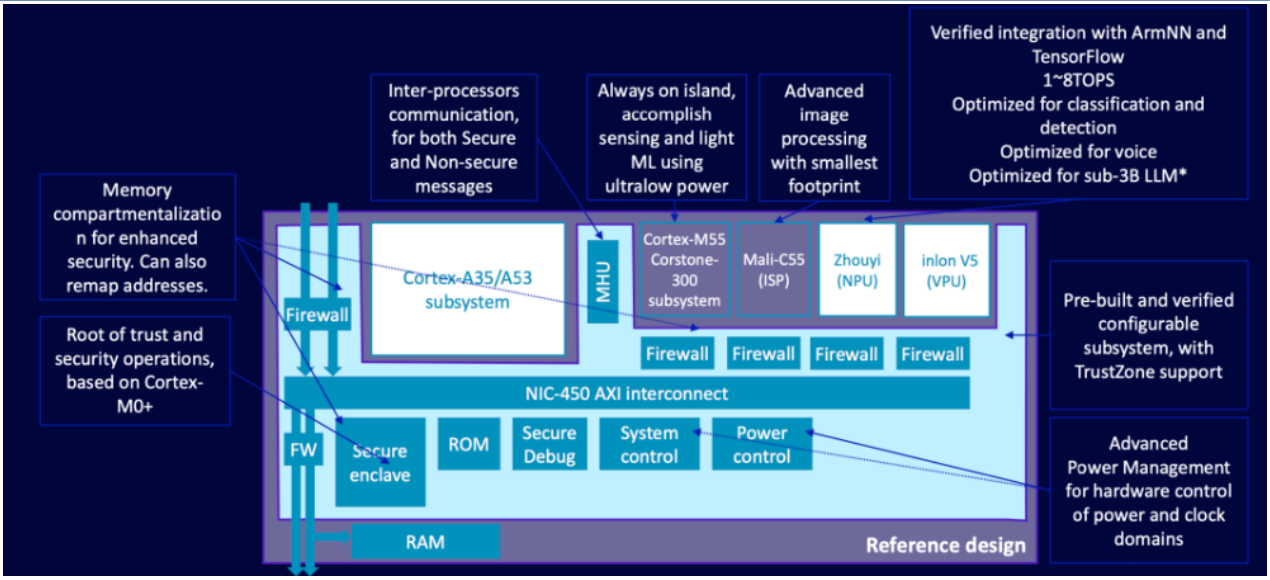
图24：端侧 LLM 的挑战与未来方向



资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

随着蒸馏模型能力提升，端侧 SoC 未来发展呈现新趋势。蒸馏模型可将大型模型知识迁移至小型模型，大幅降低对硬件资源要求，让端侧设备运行复杂 AI 模型成为现实。以 DeepSeek R1 等模型为例，其运用强化学习和模型蒸馏技术，能在本地设备高效推理，且训练只需极少量标注数据，既增强了模型适应性与灵活性，又降低了部署成本。在硬件方面，未来 NPU 等专用硬件加速器进一步发展，将使端侧设备处理复杂 AI 任务的能力更高效。这些趋势将推动 AI 从云端向本地设备转移。一方面，设备能够在离线状态下独立完成复杂 AI 任务，数据无需上传至云端处理，提高了数据隐私保护水平，避免数据在传输与存储过程中可能面临的泄露风险。另一方面，减少了对网络连接的依赖，无需等待云端响应，直接在本地快速处理，提升了处理效率，使 AI 应用在更多场景下得以快速、安全地实现，为用户带来更便捷、高效且安全的体验，也为 AI 在更多领域的深入应用拓展了空间。

图25：端侧 SOC 的解决方案



资料来源：ARM China，中国银河证券研究院

三、AI 引领硬件创新，催化换机需求

（一）传统终端：催化换机需求，缩短换机周期

消费电子终端换机需求通常由三重因素催动：1）刚性需求：硬件性能瓶颈、物理损坏；2）诱导性需求：技术革新和生态变化、功能升级；3）隐形需求：运营商补贴、以旧换新等政策刺激，消费观念与生活方式的改变。近年来，受设备耐用性和可靠性提升、产品性能提升速度放缓、宏观经济环境较差等因素影响，换机周期逐步放缓。我们认为，AI 的出现或将带来颠覆性的性能革命，从而催动新一轮的换机周期加速到来。

Deepseek 的出现将再次推进这一进程。我们认为，Deepseek 的关键在于降本增效，通过更低的训练与推理成本，实现更强的性能。此前，端侧 AI 受限于计算能力、存储能力、能源限制等，而云端 LLM 的部署受限于高延迟、隐私风险、服务成本等，因此 AI 应用的落地仿佛镜花水月。Deepseek 通过对资源的高效利用，降低了对高端硬件的依赖，使得本地设备上中小规模 LLM 的性能得到提升，从而推动 AI 应用从云端向端侧迁移。同时，AI 应用向端侧迁移很好的解决了用户关于隐私风险的顾虑，进一步推动 AI 应用在端侧的落地。

Deepseek 不仅证明了通过数据蒸馏可以使大模型迁移至小型高效模型并部署在本地，同时证明了其性能不受影响。根据 Deepseek 官方公布的了 DeepSeek-R1-Zero 与 OpenAI 的 o1-0912 模型在各种推理相关基准上的比较分析，可以看出 DeepSeek-R1-Zero 基本达到了与 OpenAI 的 o1-0912 相当的性能水平。这种“小而美”的模型的出现，为开发者、内容创作者乃至小型初创者提供了更多的选择，端侧小模型或将遍地开花，从而带动 AI 消费终端的快速渗透。

图26：DeepSeek-R1-Zero 与 OpenAI o1 模型在推理相关基准上的比较。

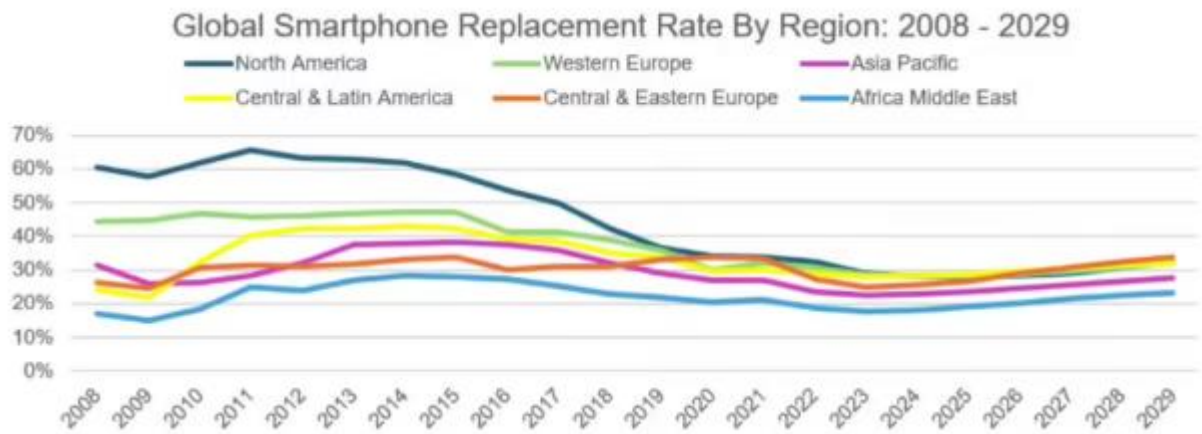
Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

资料来源：《On-Device Language Models: A Comprehensive Review》Meta，中国银河证券研究院

1） 手机：从“有限的性能提升”到“AI 的无限可能”

近年，智能手机同质化严重，功能创新匮乏，一众厂商主要通过堆叠手机的硬件性能来进行迭代升级。然而，这种升级方式对于消费者的吸引力逐渐下降，用户换手机的欲望越来越低。2023 年，全球消费者的换机周期延长至自 2008 年以来历史最长的 51 个月（换机率为 23.5%）。2024 年受经济复苏和 5g 迁移影响，换机率略有回升但未见明显改善，为 23.8%。新一轮换机潮的来临，需要颠覆性功能创新的推动。

图27：2008-2029 年智能手机换新率



资料来源：《DeeSeek-R1：通过强化学习激励 LLMs 中的推理能力》deepseek，中国银河证券研究院

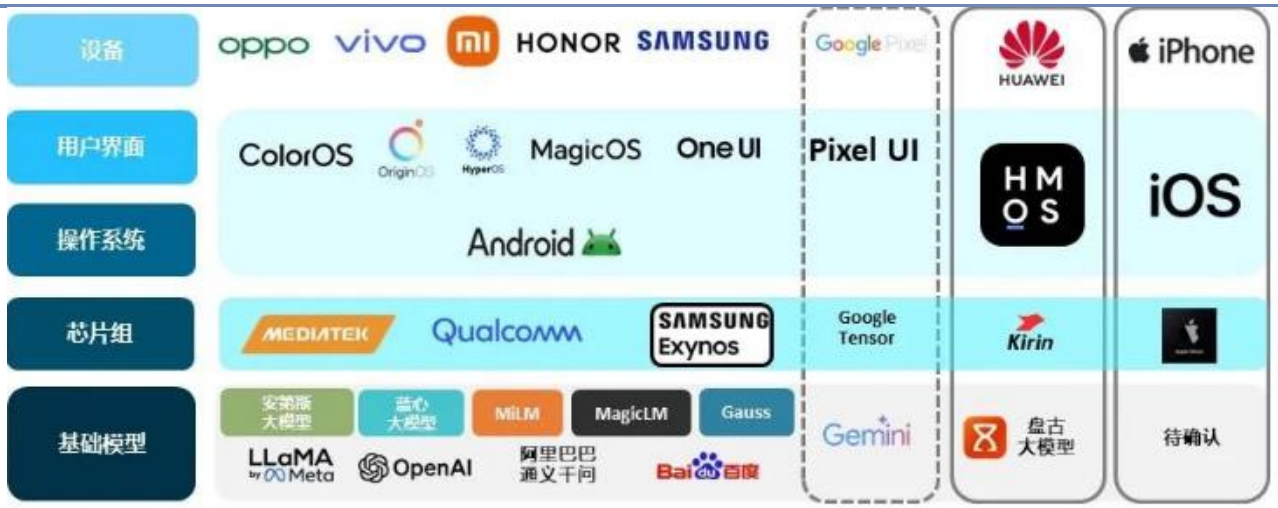
我们认为在 AI 的催化下，手机即将迎来新一轮换机潮，主要有以下两个原因：1）以 AI 技术驱

动的手机功能变革是新一轮手机迭代的核心推动力。AI 手机通过集成和应用人工智能技术，可以提供更加智能化、个性化的用户体验，并在摄影、性能优化、安全、健康管理等方面展现出显著的优势。随着 AI 技术的不断发展，这些差异可能会进一步扩大，人们的生产力和创造力将进一步释放。2）手机厂商积极入局是手机快速革新的催化剂。今年以来，有 10 余款 AI 手机面市，年初发布的三星 Galaxy S24 为用户带来通话实时翻译、即圈即搜、转录助手和笔记助手、浏览助手以及生成式编辑等创新 AI 应用；荣耀 Magic6 带来包括任意门交互、智慧成片功能以及灵动胶囊等多种基于意图识别的全新人机交互体验；华为 Pura 70 也内置盘古大模型，具备 AI 隔空操作和智感支付等功能。

图28：AI 手机的用户价值

资料来源：《AI 白皮书》IDC，中国银河证券研究院

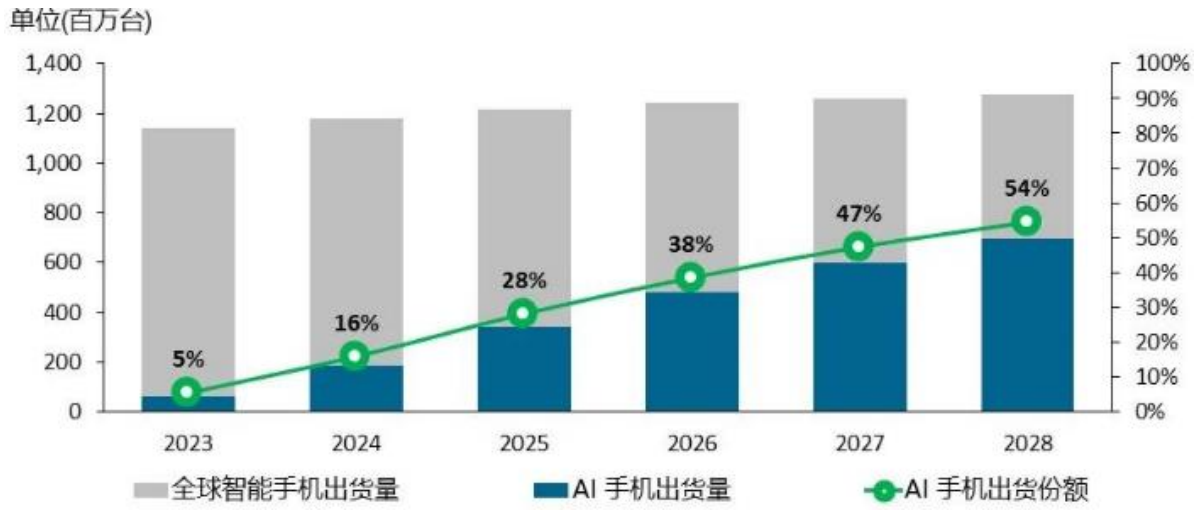
图29：AI 手机系统生态及主要参与者



资料来源：《AI 白皮书》IDC，中国银河证券研究院

手机虽然不是 AI 普惠化的唯一载体，但是其在高频、实时、个性化的 AI 服务中具有不可替代性，是 AI 普惠化的关键载体。根据 Canalys 的数据，2024 年 AI 手机的渗透率为 16%，同比增加 11pct，到 2028 年这一比例将大幅提升至 54%。2024-2028 年间，AI 手机的市场规模将以 63%的年均复合增长率快速扩容，手机的新一轮换机周期蓄势待发。

图30：2023-2028 年 AI 手机的出货份额

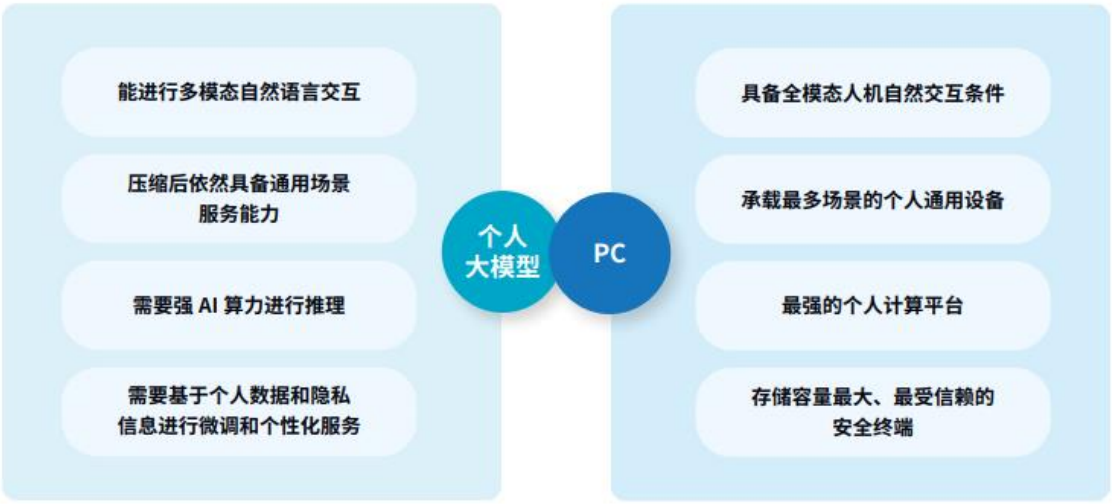


资料来源：canalys，中国银河证券研究院

2）PC：从“工具”到“生产力伙伴”

在 AI 出现之前，PC 市场也堪称“无聊”，其外观形态已较为固定，厂商主要围绕性能对 PC 进行升级。AI 的出现，也将使得 PC 具有被重新定义的机会。不同于平板、手机等智能终端设备，PC 具有强大的计算和存储能力、丰富的交互方式、广泛的应用场景、可进行多任务处理，同时还可以通过增配硬件和预装支持 AI 开发、运行的软件环境实现定制化及功能拓展。因此，AI PC 是承载大模型的理想载体。

图31：PC 与 AI 大模型的天然匹配



资料来源：《AI PC 产业（中国）白皮书》IDC，联想，中国银河证券研究院

通过搭载大模型，AI PC 可以实现多种形式的生产力跃迁，如：1）内容创作：通过 Notion AI 自动生成文章大纲、修正语法，提升写作者效率；2）数据分析：从手动核实分析数据到通过 AI 自动生成可视化报告；3）编程开发：AI 辅助编码工具 GitHub Copilot 可以自动生成部分代码，帮助程序员快速编写高质量代码等。

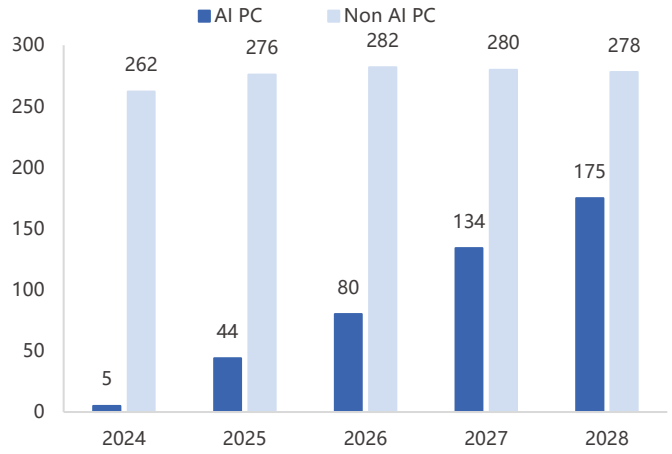
通过生产力的跃迁，PC 将不再是单纯的硬件设备，而是具备了训练、运营的价值，从“工具”升级为“生产力伙伴”。因此，用户在购买 PC 产品时将不仅仅考虑硬件性能，AI PC 生态的构建和应用都将影响用户的购买决策。

图32：AI PC 在多种场景下的价值

资料来源：《AI PC 产业（中国）白皮书》IDC、联想，中国银河证券研究院

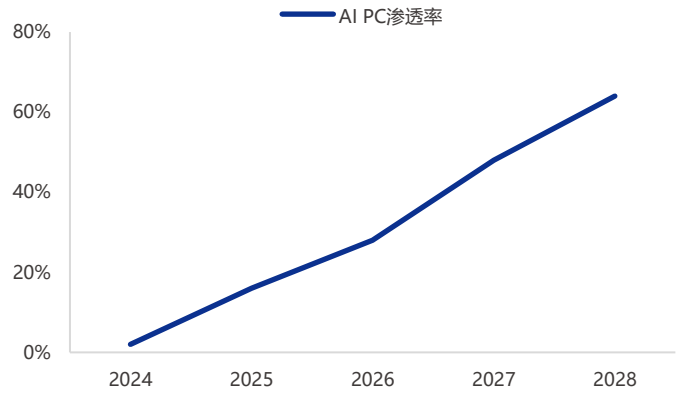
Deepseek 在端侧 AI 体现出来的优势无疑会加速 AI PC 的迭代和优化，从而加速用户的决策进程，缩短用户换机周期。恰逢 Win10 也即将终止支持，PC 市场或将迎来今年最强换机潮。摩根士丹利预测，AI PC 市场的渗透率将从 2024 年的 2%提升至 2028 年的 64%，出乎量将从 2024 年的 500 万台增加至 2028 年的 1.75 亿台，CAGR 高达 36.78%。

图33：PC 和 Non AI PC 的出货量（单位：百万台）



资料来源：Fortune Business Insights，中国银河证券研究院

图34：AI PC 渗透率



资料来源：IDC，中国银河证券研究院

（二）新型终端：创新产品高发区，AI 个人化的重要拼图

1）可穿戴设备：极适合 AI 个人化的产品形态

与手机、PC 等传统的计算设备通过按键、触碰等物理接触进行交互不同，可穿戴设备可以通过语音、手势、图像、心率等方式进行交互，可以创造更直观、更自然且身临其境的用户体验。可穿戴设备在贴身性、数据连续性、多模态感知等方面具备突出优势，因此，是极适合 AI 个人化的产品形态。同时，新型的可穿戴设备呈现隐形、轻量、柔性的趋势，也是产品形态创新的高发区。在 2025 的 CES 大会上，联想展示了可穿戴设备 AI Travel Set，其中包含的 AI 吊坠作为新的产品形态搭载了 AI 实时处理功能，可实现场景识别、重点事件抓取以及即拍即分享的功能，为可穿戴 AI 产品的多样化打开了想象空间。

图35：AI 硬件产品创新图谱

资料来源：定见咨询，中国银河证券研究院

此前，可穿戴设备多作为手机的延伸产品出现。2024 年的 MWC 大会上，AI PIN 带着“为用户带来没有智能手机的世界的愿景”横空出世，虽然其实际表现不及预期，存在响应速度慢、电池续航短、设备过热、激光投影亮度不足以及 AI 功能表现不佳等多种问题，但是为可穿戴设备向着独立 AI 硬件发展提供了方向。

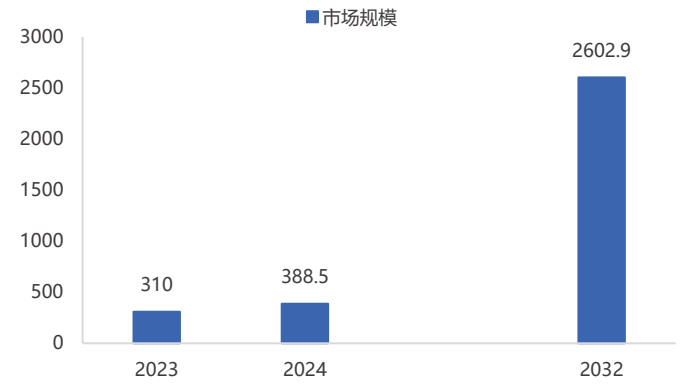
AI PIN 的核心功能依赖于云端大模型，通过将指令文本上传至 GPT-4 等云端模型进行推理，推理的结果再通过设备反馈给用户，从而也带来了连接不稳定和延迟的问题。近日，Deepseek 的出现展现了端侧大模型落地的可能，为解决此类问题带来了希望，可穿戴设备的迭代进程或将加速。

图36：AI Pin 的图形显示方案

资料来源：雷科技，中国银河证券研究院

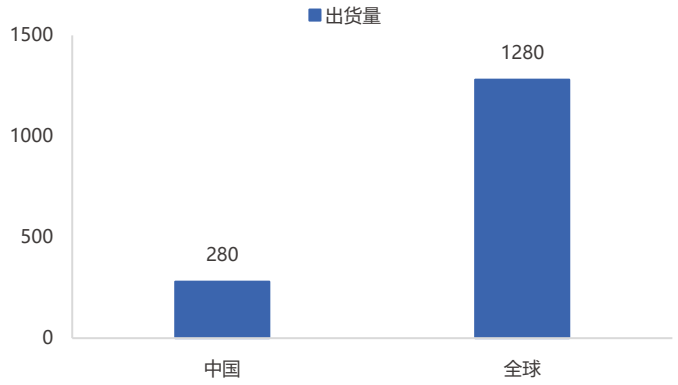
根据 Fortune Business Insights 数据，2023 年全球可穿戴人工智能市场规模为 310 亿美元。随着人工智能的快速崛起，预计该市场将从 2024 年的 388.5 亿美元增长到 2032 年的 2602.9 亿美元，预测期内复合年增长率为 26.8%。最有希望快速落地的 AI 眼镜出货量将在 2025 年达到 1280 万副，同比增长 26%，其中我国 AI 眼镜出货量将同比增长 107%达到 280 万副。

图37：2023-2028 年可穿戴人工智能市场规模（单位：亿美元）



资料来源：Fortune Business Insights，中国银河证券研究院

图38：2025 年 AI 眼镜出货量（单位：万副）



资料来源：IDC，中国银河证券研究院

2）AI 玩具等：创新产品高发区，交互体验显著提升

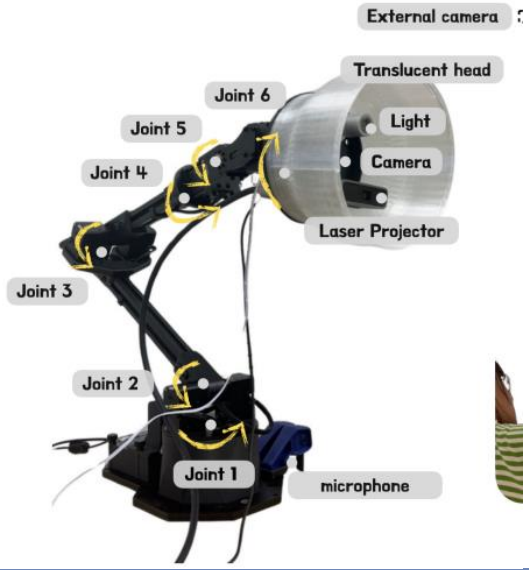
在 2025 年的 CES 展上，除了较为常规的手机、PC、可穿戴产品外，AI 玩具也成为一大亮点。玩具本就是具有情绪价值属性的产品类型，AI 技术的加持使得玩具的趣味性、陪伴性进一步加强，同时提升了用户的参与感。玩具的边界逐渐被打破，从被动的陪伴工具变成了智能伙伴，有望成为连接现实与数字世界的桥梁。

目前，已有多家公司推出相关产品，除较早布局的 Folo Toy 基于“大模型+故事机”推出了火火兔等 IP 合作玩具外，字节跳动、苹果等海内外大厂也纷纷入局。近日，苹果推出了一款具有表现型的 AI 伴侣台灯，再次为 AI 伴侣型机器人的发展打开了无限可能。

图39：Folo Toy 火山兔

资料来源：Folo Toy 官网，中国银河证券研究院

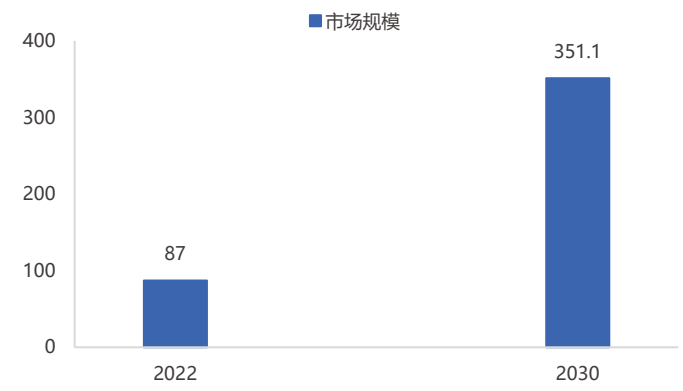
图40：苹果台灯机器人的示意图



资料来源：苹果，中国银河证券研究院

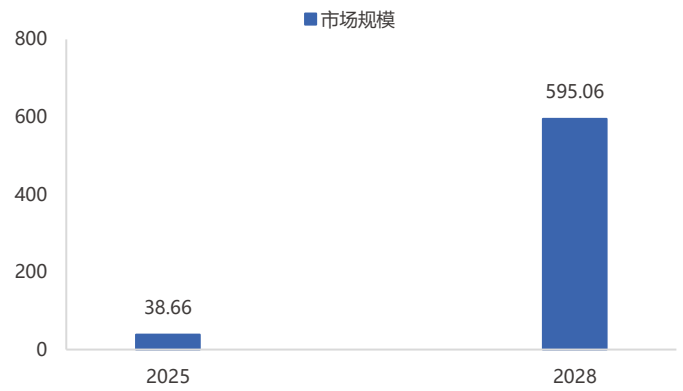
相比于其他类型的产品，AI 玩具的硬件限制本就较少，成本限制是 AI 玩具发展过程中的一大问题。DeepSeek 的出现很大程度上的解决了这一问题，更将加速 AI 玩具的快速迭代和落地。根据 Research and Markets 数据，2022-2030 年全球 AI 玩具的市场规模将从 87 亿美元提升至 351.1 亿美元。随着承担情感需求的 AI 玩具逐渐量产，我国 AI 情感陪伴行业市场规模也将从 38.66 亿美元增长至 595.06 亿美元，CAGR 达到 148.74%。

图41：2022-2030 年全球 AI 玩具市场规模（单位：亿美元）



资料来源：Research and Markets，中国银河证券研究院

图42：2025-2028 中国 AI 情感陪伴行业市场规模



资料来源：共研产业咨询，中国银河证券研究院

四、投资建议

我们乐观看待 DeepSeek 创新对电子行业带来的改变。我们总结以下几条结论：

1，我们认为 DeepSeek 的创新并没有实质性打破 scaling laws 。DeepSeek 模型具有更强的性能，更低的训练与推理成本，将加速推动 AI 应用与硬件的普及和落地。虽然更低的训练与推理成本减少了当前的算力需求，但是并不意味着 AI 的未来发展对半导体整体需求的减少，相反由于其模型架构、基础设施数据等方面的优化，以及更低的成本，使得其更加容易布置在端侧，从而加速 AI 的普及。AI 能力边际的扩张依然需要依赖更大的模型和强大的算力，DeepSeek 在算法和架构上的创新给 AI 的发展增加了一条新的道路。

2，Scaling laws 正在从 pre-training 转向 post-training 和推理，通过增加模型规模、扩展训练数据、提高计算资源以及合理的任务设计，可以加速模型学习更复杂的推理能力，这一过程遵循 scaling law。随着模型规模、数据量和计算资源的增加，模型能够更好地进行推理。

3，针对边缘设备的 LLM 部署，在保持性能的同时提高计算效率至关重要，通过量化、剪枝、知识蒸馏和低秩分解，这些方法通过平衡性能、内存占用和推理速度来提高大语言模型的运行效率，有利于 AI 硬件端的落地与普及。我们看好 AI 应用持续落地带来的传统消费电子的换机周期，苹果产业链值得关注，同时看好 AI 终端硬件如耳机、眼镜、桌面机器人、小家电、周边硬件等。建议关注：寒武纪、海光信息、蓝思科技、鹏鼎控股、领益智造、水晶光电、蓝特光学、恒玄科技、中科蓝讯、乐鑫科技、瑞芯微、全志科技、翱捷科技、敏芯股份、兆易创新、普冉股份、艾为电子。

表3：建议关注相关标的盈利预测情况（截至 2025 年 2 月 10 日）

代码	标的名称	总市值（亿元）	EPS（元）			P/E		
			2024E	2025E	2026E	2024E	2025E	2026E
300433.SZ	蓝思科技	1296.05	0.80	1.10	1.37	32.61	23.61	18.94
002938.SZ	鹏鼎控股	938.79	1.55	1.94	2.21	26.04	20.82	18.35
002600.SZ	领益智造	596.40	0.29	0.42	0.55	29.37	20.05	15.37
002273.SZ	水晶光电	304.41	0.74	0.92	1.11	29.45	23.73	19.76
688127.SH	蓝特光学	109.63	0.67	0.92	1.13	40.67	29.47	24.00
688608.SH	恒玄科技	476.59	3.24	4.86	6.58	122.44	81.74	60.31
688332.SH	中科蓝讯	172.04	2.52	3.32	4.14	56.80	43.13	34.50
688018.SH	乐鑫科技	303.21	3.09	4.07	5.32	87.43	66.33	50.81
603893.SH	瑞芯微	689.09	1.22	1.80	2.48	134.37	91.57	66.25
300458.SZ	全志科技	298.58	0.38	0.57	0.78	124.83	82.90	60.81
688220.SH	翱捷科技	290.80	-1.34	-0.70	0.37	-51.85	-98.92	190.05
688286.SH	敏芯股份	39.40	-0.54	0.53	1.37	-130.55	133.09	51.24
603986.SH	兆易创新	846.01	1.68	2.47	3.17	75.80	51.48	40.20
688766.SH	普冉股份	114.36	2.66	3.40	4.16	40.76	31.89	26.02
688256.SH	寒武纪	2388	-1.10	-0.06	1.15	-520.7	8893.3	499.6
688041.SH	海光信息	2975	0.82	1.21	1.64	155.3	106.1	78.2
688798.SH	艾为电子	168.01	0.92	1.65	2.42	78.42	43.68	29.81

资料来源：Wind 一致预期，中国银河证券研究院

五、风险提示

（1）**AI 应用与智能硬件落地进展不及预期的风险**：若 AI 大模型及相关应用开发进展缓慢，相关硬件产品功能无法满足市场预期，可能导致 AI 下游应用与硬件产业链下修增长预期。

（2）**全球经济疲软需求不及预期的风险**：随着全球贸易保护主义的盛行，再通胀的压力之下可能导致全球经济恢复进展不达预期，从而导致整体需求的萎缩。

（3）**科技自立自强进展不及预期的风险**：由于半导体上游环节技术壁垒高，且技术封锁更加严格，可能导致部分核心卡脖子环节的进展落后于市场预期。

（4）**国际政治环境变动不确定性的风险等**：随着美国新一任总统特朗普的上台，以及俄乌冲突的外部环境改变，均可能会导致国际政治环境的变化，引发连锁反应，从而影响出口。

图表目录

分析师承诺及简介

本人承诺以勤勉的执业态度，独立、客观地出具本报告，本报告清晰准确地反映本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告的具体推荐或观点直接或间接相关。

高峰：北京邮电大学电子与通信工程硕士，吉林大学工学学士。2 年电子实业工作经验，6 年证券从业经验，曾就职于渤海证券、国信证券、北京信托证券部。2022 年加入中国银河证券研究院，担任电子团队组长，主要从事硬科技方向研究。

王子路：英国布里斯托大学金融与投资学硕士，山东大学经济学学士。2020 年加入中国银河证券研究院，主要从事科技产业研究。

钱德胜：电子行业分析师，硕士学历，曾就职于国元证券研究所，5 年行业研究经验。

免责声明

评级标准

评级标准		评级	说明
评级标准为报告发布日后的 6 到 12 个月行业指数（或公司股价）相对市场表现，其中：A 股市场以沪深 300 指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准，北交所市场以北证 50 指数为基准，香港市场以恒生指数为基准。	行业评级	推荐：	相对基准指数涨幅 10%以上
		中性：	相对基准指数涨幅在-5%～10%之间
		回避：	相对基准指数跌幅 5%以上
	公司评级	推荐：	相对基准指数涨幅 20%以上
		谨慎推荐：	相对基准指数涨幅在 5%～20%之间
		中性：	相对基准指数涨幅在-5%～5%之间
		回避：	相对基准指数跌幅 5%以上

联系

中国银河证券股份有限公司 研究院

深圳市福田区金田路 3088 号中洲大厦 20 层

上海浦东新区富城路 99 号震旦大厦 31 层

北京市丰台区西营街 8 号院 1 号楼青海金融大厦

公司网址：www.chinastock.com.cn

机构请致电：

深广地区：程 曦 0755-83471683 chengxi_yj@chinastock.com.cn

苏一耘 0755-83479312 suyiyun_yj@chinastock.com.cn

上海地区：陆韵如 021-60387901 luyunru_yj@chinastock.com.cn

李洋洋 021-20252671 liyangyang_yj@chinastock.com.cn

北京地区：田 薇 010-80927721 tianwei@chinastock.com.cn

褚 颖 010-80927755 chuying_yj@chinastock.com.cn