

电子行业深度报告

端侧 AI 重构终端生态：硬件升级驱动换机潮，多设备协同催生商业模式跃迁

增持（维持）

2025 年 06 月 26 日

证券分析师 陈海进

执业证书：S0600525020001

chenhj@dwzq.com.cn

证券分析师 陈妙杨

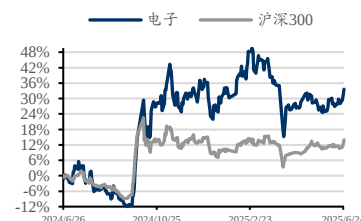
执业证书：S0600525020002

chenmy@dwzq.com.cn

投资要点

- **智能体构建全新交互中枢，终端流量入口重塑：**AI 智能体通过“决策（LLM）+记忆+规划+工具”构建智能闭环，正逐步重塑终端交互中枢，成为新一代超级入口核心。其发展沿着“数字助理型”与“具身/社会型”两条技术路径演化，并加速落地于 PC、手机与可穿戴设备等多元终端。各大厂商围绕端侧部署与云端协同展开技术布局，推动智能体从功能演示走向系统集成。手机、PC 等终端正采用“端优先”策略强化本地推理能力，而可穿戴设备则通过“端-近端-云”架构解决算力瓶颈，构建多模态数据驱动的个性化体验闭环。
- **手机端侧 AI 发展驱动硬件升级，有望成为新一轮换机潮核心推力：**随着端侧大模型能力的持续演进，AI 正逐步从辅助交互工具向“具身智能体”转变，应用场景正从语音助手延展至多模态理解、内容生成和个性化生产力等高价值任务。这一演化路径带动了对更高参数规模模型的需求，进而对手机 SoC、内存、散热和电池等核心部件提出更高配置要求，也带动声学传感部件和结构件的升级，构成推动手机硬件系统升级的重要力量。
- **AI 终端重塑价值链：**多设备协同正重塑终端生态，可穿戴设备成为差异化关键入口，推动手机厂商向 AI 终端平台转型。订阅变现模式兴起，使终端厂商从一次性销售走向“硬件+服务”并重，带动云推理、模型部署等产业链环节需求增长，具备高订阅贡献与留存率的厂商有望获得互联网估值溢价。聚焦消费电子终端厂商竞争优势的重塑路径看，随着端侧 AI 引入，“硬件+订阅”模式将成为主流，订阅收入比重持续上升，推动终端盈利结构向 SaaS 逻辑演化。各大厂商加速布局多设备智能生态，强化跨平台协同，可穿戴设备作为关键数据入口和智能体节点价值显著提升。本地化适配能力由“加分项”转变为核心壁垒，AI 能力深度融合本地语境与行为特征，有望提升区域市场渗透率与用户粘性。AI 推动渠道模式由硬件销售向场景体验转型，运营商等新兴参与者正在拓展终端分发新范式。全产业链协同助力产品快速迭代并优化成本结构。
- **投资建议：**整机组装：立讯精密、华勤技术、歌尔股份、国光电器、龙旗科技、天键股份等；端侧算力：瑞芯微、恒玄科技、乐鑫科技、炬芯科技、泰凌微等；存储：兆易创新、江波龙、德明利、普冉股份、佰维存储、香农芯创等；外观件/功能件：蓝思科技、鹏鼎控股、东山精密、领益智造、比亚迪电子；充电/电池/散热：奥海科技、德赛电池、珠海冠宇、中石科技、思泉新材等
- **风险提示：**底层创新不及预期、技术发展不及预期、竞争加剧风险

行业走势



相关研究

《可穿戴 SoC，从大厂自研窥见战略纵深——大厂自研三两事系列》

2025-06-25

《Marvell 上调数据中心 TAM，关注 ASIC 趋势对铜连接市场的驱动-算力周报》

2025-06-22

表 1：部分重点公司估值

代码	公司	总市值 (亿元)	收盘价 (元)	EPS (元)			PE			投资评级
				2024A	2025E	2026E	2024A	2025E	2026E	
002475	立讯精密	2,409.53	33.24	1.84	2.36	2.83	18.03	14.08	11.75	买入
603296	华勤技术	756.74	74.50	2.88	3.70	4.48	25.86	20.14	16.63	买入
002241	歌尔股份	828.18	23.72	0.76	1.12	1.42	31.08	21.18	16.70	买入
688608	恒玄科技	431.29	359.27	3.84	7.72	10.96	93.66	46.54	32.78	买入
688591	泰凌微	115.94	48.16	0.40	0.81	1.24	119.02	59.46	38.84	买入

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

603986	兆易创新	833.39	125.50	1.66	2.48	3.36	75.59	50.60	37.35	买入
688525	佰维存储	307.48	66.66	0.35	1.34	1.69	190.70	49.75	39.44	买入
300475	香农芯创	169.74	36.60	0.57	1.47	1.87	64.28	24.90	19.57	买入
300433	蓝思科技	1,098.70	22.05	0.73	1.12	1.45	30.32	19.69	15.21	买入
002938	鹏鼎控股	708.86	30.58	1.56	1.92	2.18	19.58	15.93	14.03	买入
002384	东山精密	667.62	36.45	0.59	1.58	2.07	61.50	23.07	17.61	买入
002600	领益智造	597.80	8.53	0.25	0.40	0.53	34.09	21.32	16.09	买入
300684	中石科技	70.53	23.55	0.67	0.94	1.18	35.02	25.05	19.96	买入

数据来源：，东吴证券研究所

内容目录

1. 智能体全新交互中枢，终端流量入口重塑	5
1.1. 智能体定义：决策+记忆+规划+工具	5
1.2. 智能体落地终端，超级入口迎来重塑窗口期	6
1.3. 端侧 AI 部署加速落地，功能形态与技术路径初步成型	7
1.3.1. PC/手机端承接系统枢纽，可穿戴设备探索形态突破	7
1.3.2. 技术路径：“端优先”成趋势，可穿戴侧重协同算力	10
2. 手机端侧 AI 发展驱动硬件升级，有望成为新一轮换机潮核心推力	13
2.1. 手机厂商端侧大模型布局对比及核心技术路径梳理	13
2.2. 推演未来：多模态与跨应用驱动模型扩容，参数规模与压缩算法面临双重挑战	15
2.3. 算力、存力、电池和其它结构件协同进化，硬件成为大模型落地关键	17
2.4. AI 成为新一轮智能手机换机潮的核心驱动	20
3. AI 终端重塑价值链：从硬件销售到智能生态	23
3.1. 商业模式重塑：订阅变现+产品协同，AI 终端引领价值体系重构	23
3.1.1. 多设备协同驱动 AI 终端生态成型，可穿戴产品成差异化关键入口	23
3.1.2. 订阅收入成为增长引擎，推动终端厂商盈利模式转型	23
3.2. 端侧厂商的传统商业模式在智能体时代会如何变化	24
3.2.1. 订阅收入占比增大	24
3.2.2. 跨平台协同与多设备智能体生态加强	24
3.2.3. 本地化适配能力有望从“加分项”演变为厂商核心护城河	25
3.2.4. AI 驱动下的渠道模式转型：从售卖硬件到构建场景体验	25
3.2.5. 全产业链协同加速终端产品迭代与成本优化	26
4. 投资建议	27
5. 风险提示	28

图表目录

图 1:	OpenAI 定义的智能体架构.....	5
图 2:	Google 定义的智能体架构.....	5
图 3:	AI 时代入口形态的变化	6
图 4:	不同品牌手机代劳工作完成情况.....	8
图 5:	端云协同智能计算的基本要素.....	11
图 6:	可穿戴端侧的“端-边-云”架构.....	12
图 7:	国内 2025 年 3 月移动端 AI 原生 App 活跃用户规模	14
图 8:	AI 综合助手类应用活跃用户数量最多	14
图 9:	终端 AI 设备向常驻式、无屏化具身智能体迈进	15
图 10:	端侧大模型业务场景分析.....	16
图 11:	读取内存数据功耗远高于计算.....	18
图 12:	高通 Snapdragon 8 Gen3 AI 任务性能每瓦提升 40%	18
图 13:	端侧大模型推理演进驱动硬件配置全面升级.....	20
图 14:	iPhone 侧键唤醒 GPT 功能或带动结构件新需求.....	20
图 15:	2023-2027E 国内 AI 手机出货量及占比（亿只，%）	20
图 16:	全球智能手机换机率在 2024 年迎来拐点.....	20
图 17:	多模态感知赋能可穿戴设备个性化智能体验.....	23
图 18:	华为可穿戴新品将全部搭载原生鸿蒙系统.....	25
图 19:	小米 Hypermind 使设备能够主动了解用户需求并采取相应措施	25
表 1:	智能体的定义与结构比较.....	5
表 2:	AI 手机产品总结	7
表 3:	AI PC 产品总结	9
表 4:	可穿戴端侧产品总结.....	10
表 5:	重点厂商端云协同架构.....	11
表 6:	主要智能手机厂商端侧大模型部署情况对比.....	13
表 7:	不同模态对应 token 换算.....	17
表 8:	端侧 AI 推理负载下的手机硬件资源需求测算	19
表 9:	端侧 AI 推理负载下的手机使用寿命测算	21

1. 智能体全新交互中枢，终端流量入口重塑

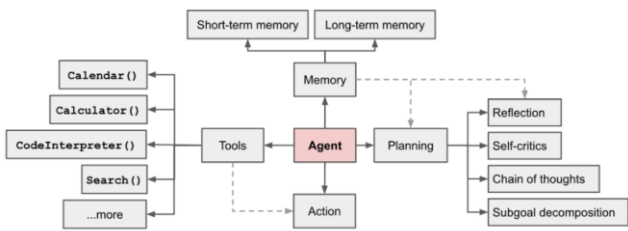
1.1. 智能体定义：决策+记忆+规划+工具

AI 智能体是一种具备环境交互、自主决策与任务执行能力的软件程序，由人类设定目标，智能体独立选择最优行动路径。虽然业界在模块命名和架构安排上各有差异，例如 OpenAI 更偏向工程实现，复旦 NLP 实验室则立足于未来 AI 产业趋势构想，但在底层结构上高度一致，体现出一定程度的行业共识：

- 1. 以 LLM 为大脑，具备推理、知识整合与语言交互能力。
- 2. 记忆系统支撑上下文连贯与经验积累。
- 3. 工具接口打破模型“只读”边界，实现动作执行
- 4. 反馈闭环支持状态保存与动态调整，保障任务的持续感知-决策-行动循环。

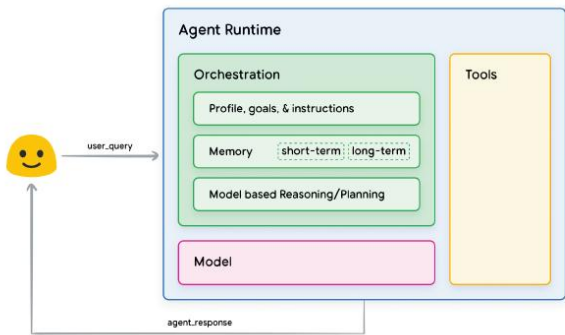
总体来看，“LLM+记忆+工具+闭环”构成智能体的核心架构要素。

图1: OpenAI 定义的智能体架构



数据来源：OpenAI，东吴证券研究所

图2: Google 定义的智能体架构



数据来源：Google，东吴证券研究所

未来智能体有望沿着“数字助理型”与“具身/社会型”两大技术路径演进。根据研究机构对智能体的定义不同，我们判断未来其生态系统有望沿着两条技术路径同步演化：一类是以提升个人效率为目标的“数字助理”，聚焦于自然语言交互与多工具协同，广泛应用于办公自动化、信息检索、知识管理等场景；另一类则是具备感知与行动能力的“具身/社会型”智能体，强调多模态感知、实体控制与多智能体协作，在机器人、智能制造、智慧交通等领域具备落地潜力。尽管两者在功能侧重与硬件要求上存在显著差异，但其底层均依托于大语言模型（LLM）为核心，辅以记忆系统、任务规划与工具调用构成“智能闭环”。

表1: 智能体的定义与结构比较

项目	OpenAI	复旦大学 NLP 实验室	Google
定义	能够代表用户执行任务的系	能感知环境、做出决策并采	一种能自主观察世界、使用

	统	取行动的人工智能体	工具采取行动、以实现既定目标的智能应用程序，在人类不提供明确指令的情况下也能自我决策。
模块	规划模块	根据用户目标自主划分任务，安排执行顺序	智能大脑：由 LLM 构成，负责任务理解、记忆调用、知识应用、推理与规划
	记忆模块	持续记录任务完成过程中的状态、历史和上下文，确保行为的连贯性	感知模块：支持从文本、图像、音频甚至 3D 感应器获取信息，使智能体能全面理解外部环境
	工具系统	支持调用网页浏览器、Python 运行环境、插件系统等，智能体可借助这些工具完成用户任务	行动模块：包括文本输出、工具使用（API、代码、插件等）和具身动作（如控制机器人）
	决策模块	LLM 是控制器——用大模型做任务理解、推理、子任务拆分与决策	模型层：负责理解指令、推理、生成响应
共识	记忆模块	智能体需要可读写的外部 / 持久记忆，保证上下文连贯、经验累积	控制中枢：负责管理智能体的整个执行循环（信息输入→推理决策→工具调用→状态更新），实现状态管理、记忆保持、上下文传递和多轮任务规划
	行动模块	单靠语言输出不够，智能体必须能“调用外部能力”（数字工具或物理执行）	工具层：提供智能体与外部世界交互的能力
	闭环反馈	任务执行是循环而非一次性——持续感知-决策-行动迭代	

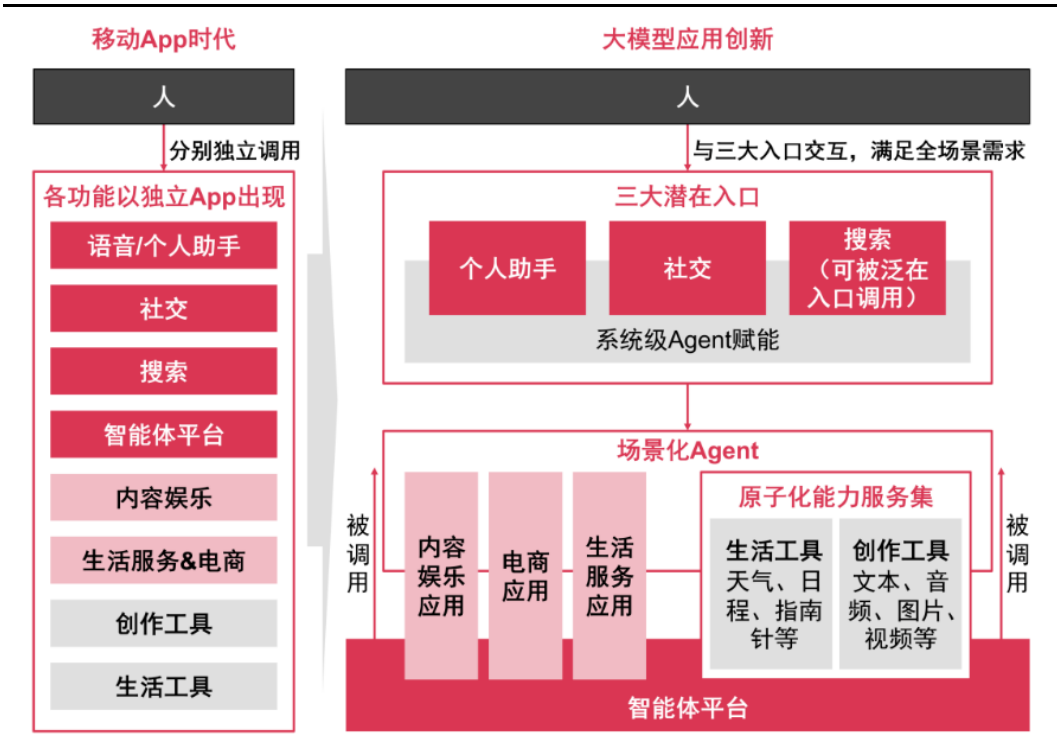
数据来源：《The Rise and Potential of Large Language Model Based Agents: A Survey》，OpenAI，Google，东吴证券研究所

大语言模型作为智能体的认知中枢,承担理解、决策与交互的核心职责,是通向 AGI 的关键起点。根据南洋理工大学研究，智能体技术已历经三代演进路径：第一代算法驱动阶段（2006~），以博弈论与优化算法为主，解决结构化策略问题，智能水平有限；第二代为强化学习驱动阶段（2017~），强化对复杂动态环境的建模能力，应用于推荐、交易、博弈等场景，但在泛化与迁移上仍存瓶颈；第三代为大语言模型驱动阶段（2023~），以 ChatGPT、Gemini 等为代表的 LLM 赋予其自然语言理解、任务规划与工具调用等能力，通过 ReAct、CoT 等机制实现“感知—认知—执行”闭环，首次使智能体具备自主决策与完成复杂任务的能力，标志其从“智能工具”向“通用生产力”跃迁。

1.2. 智能体落地终端，超级入口迎来重塑窗口期

AI 时代超级入口强调跨应用场景智能协作，终端厂商生态控制力有望加强。超级入口的本质在于高频用户需求的聚合与流量变现能力。回顾 PC 与移动互联时代，终端设备和操作系统通过整合多种功能和服务，成为用户获取商品、服务和内容的主要入口。进入 AI 时代，超级入口的核心逻辑从“功能驱动”转向“任务驱动”。用户无需逐一操作应用程序，仅通过自然语言提出任务，AI 助手即可跨应用调取服务完成如报告生成、日程管理等复杂操作。该类跨应用协同依赖底层系统级调用，天然利好掌握操作系统与终端硬件的厂商。根据普华永道，当前手机端涌现出三类超级入口雏形：个人助手、社交与搜索。终端厂商正加快 AI 生态系统构建，围绕自身能力圈深耕多样化场景，抢占 AI 时代流量入口重构的战略窗口。

图3: AI 时代入口形态的变化



数据来源：普华永道，东吴证券研究所

1.3. 端侧 AI 部署加速落地，功能形态与技术路径初步成型

1.3.1. PC/手机端承接系统枢纽，可穿戴设备探索形态突破

当前，智能体在消费电子终端的集成正加速推进。本节将从 PC、手机与可穿戴设备三大典型终端出发，梳理端侧 AI 落地的最新进展与功能演化路径。

手机方面，厂商正把智能体从“语音助手”升级为深度嵌入操作系统的执行引擎。智能体不再停留在问答层面，而是接管联系人、相册、定位、支付等高权限接口，能够跨应用解析用户意图、自动完成多步操作。苹果计划让新版 Siri 直接在端侧调用日程、照片与邮件等私有数据，并在必要时通过加密云推理补齐复杂逻辑，目标是把填写表格之类的事务完全交给系统处理；三星的 Galaxy AI 已在 S25 系列统一 30 个本地应用和十余款高频第三方应用；安卓阵营的 OPPO、vivo、小米等也在各自定制的系统框架里加入“星环”“PhoneGPT”“超级小爱”等智能体，重点打通外卖、出行、社交消息等高频场景，以展示跨 App 动态调度能力。

表2: AI 手机产品总结

厂商	端侧 AI 功能
Apple	苹果工程师团队正在开发“具备更强个性化与场景感知能力的新版 Siri”，该版本将实现跨应用程序的任务处理功能。具体应用场景包括：根据设备本地存储信息自动调取好友推荐的播客节目、即时追踪亲属航班动态等需要多维度数据联动的复杂操作；此外，Siri 还能根据照片中获取的驾照号码，代用户填写表格等。
三星	通过 Galaxy AI 在 S25 系列手机中全面升级 Bixby 助手，具备图文理解、多轮对话、情绪感知和智能提醒等能力，跨应用执行能力打通了 30 多个本地应用和 10 多个高

	频的第三方应用
华为	华为将小艺升级为系统级智能体，能够实现跨多应用执行任务。经测试能够完成“导航”、“订机票”等场景
荣耀	发布搭载智能体的 AI 手机 Magic7 Pro，能够理解用户意图，并实现跨应用完成一系列复杂任务，如“一键点饮品”、“一键旅行规划与订票”等。
vivo	推出了名为“PhoneGPT”（手机智能体）的多模态助理，能够独立完成预定餐厅等任务。
OPPO	集成阶跃星辰大模型，能实现跨 App 操作的“一键全能搜”、图像理解与问答的“一键问屏”等智能助理功能，主动感知用户意图并完成复杂任务。
小米	“超级小爱”是其系统级 AI 智能体，具备系统级全局执行能力，可以基于用户习惯自动完成个人信息填写、文件搜寻、信息记录等功能

数据来源：和讯，第一财经，阶跃星辰，公司官网，东吴证券研究所

实测结果显示，这一代手机智能体在“导航”和“订机票”这类结构化流程中的成功率已达到可用水平，而在外卖下单、社交消息等任务上仍响应不稳定。总体来看，我们认为手机端 AI 已从“功能演示”过渡到“系统能力”，为未来进一步向生产力工具演进奠定了平台基础。

图4：不同品牌手机代劳工作完成情况

手机	导航	点外卖	订机票	发微信消息
小米 14ultra	✓	✓	✓	✗
荣耀 Magic 7	✓	✓	✓	✓
vivo x2000	✓	✗	✓	✗
OPPO Find X7	✓	✓	✓	✓
华为 Mate70	✓	✗	✗	✗
三星 Galaxy S25	✓	✗	✓	✗

数据来源：南财合规科技研究院，Tech4Good，东吴证券研究所

PC 方面，重点体现办公创意效率提升与系统级自动优化。功能上，我们认为各厂商端侧 AI 已从简单对话助手演进为复杂任务协同系统，涵盖办公与创意和系统级智能优化两大核心方向：

- 办公与创意：包括文档润色、内容生成、跨语种翻译、图像直搜、结构化摘要等。
- 系统级智能优化：涵盖亮度、键盘背光调节、噪音降除、屏幕聚焦、语言控制系统设置等。

表3: AI PC 产品总结

厂商	代表机型	端侧 AI 功能
Dell	XPS13/Latitude 7455/Inspiron 14+	预装 Recall 有了 Recall, 你就能记录下你电脑过去的记忆。只需描述一下你对某件事的记忆, 它就能检索到你看到它的那一刻。 实时体验任何语言的内容 在播客录制过程中自动消除噪音并改善照明
Lenovo	ThinkPad X1 2-in1 Gen 10 Aura Edition	Lenovo AI Now: 本机 Llama 3 模型+个人知识库; 支持自然语言调系统设置; 支持第三方 AI 应用生态系统
EdgeRunner × Intel	搭载 Intel 处理器的 AIPC	军工级离线智能体, 功能包括本地化离线文档分析、代码重构、实时翻译和转录、问答、总结和创意生成等 AI 硬件调校 (AI 算力调度、AI 屏幕亮度和键盘亮度调节等)
小米	REDMI Book Pro 16 2025	AI 办公及创作 (AI 文档处理、截图直搜、文生图、文件搜索等) AI 搜索 (包括本地视频/图片/文档和互联网信息) AI 阅读 (支持知识问答、中英文及方言交互、文档总结) AI 创作 (代码生成、文档创作、一键生成纪要) AI 工作台 (自动识别人声并降噪、文档和通知多终端同步投屏隐私保护等)
荣耀	MagicBook Pro 14	书写工具 (文字校对与改写)
Apple	MacBook Air	影像视图 (根据描述或图库中的某个人, 创作原创影像) Siri (语言理解能力更强, 集成 ChatGPT 带来更多解答)

数据来源: 公司官网, Businesswire, 东吴证券研究所

可穿戴硬件方面, 我们认为厂商“先看硬件天生擅长的感知维度, 再向 AI 延伸”。

回到三大典型形态——眼镜、TWS 耳机、智能手表——它们各自的 AI 侧重点与用户原生使用动机基本吻合, 但又在向下一阶段的“多模协同”递进。整体看, 可穿戴硬件细分产品各自都在探索如何利用自身形态优势放大端侧 AI 的常用价值, 行业仍处百花齐放的早期阶段。

智能眼镜: 听觉+视觉的双通道。眼镜天生占据视线与耳道, 从接听电话、语音助手这类“免手占耳”落地场景, 逐步拓展到实时字幕、场景识别和导航提示等场景。最新一代 Ray-Ban Meta 与 OPPO Air Glass 3 已可在端侧完成语音问答、拍照解说, 并把翻译、导航等重计算任务分级上云。往前看, 一旦光机与电池再减薄、算力再下沉, AI 眼镜有潜力替代手机的部分功能。

TWS 耳机: 听觉入口的深耕。耳机原本的卖点是音质和降噪, 如今 AI 在优化这些基本功能后, 还在实时语音助手、同声传译、会议转写等场景深耕。接下来耳机还可能切入声学健康, 或将全天候的听觉代理。

智能手表: 持续感知与生理闭环。目前集成 AI 功能的智能手表落地产品仅见 Phancy 等少数厂牌, 功能上与耳机、眼镜的语音交互并无本质区分; 然而苹果已公开把下一代 Watch 系列定位为“AI 驱动的腕上医疗监控”, 腕表赛道的差异化最终或将落在对多模态

生理信号的融合分析与长期模型训练。

表4：可穿戴端侧产品总结

端侧产品类别	品牌及型号	端侧 AI 功能
智能眼镜	Ray-Ban Meta（第二代）	集成 Meta AI 助手，支持语音命令、接听电话、视频直播和环境识别等功能 未来将推出实时翻译、导航等应用
	OPPO Air Glass 3	搭载 OPPO 自研的 AndesGPT 大模型，支持语音助手、音乐播放、接听电话、信息展示等功能
	Halliday AI 眼镜	支持备忘录加载、实时翻译、对话纪要、对话理解等功能
	星际魅族 StarV Air2	将支持线下支付功能，用户无需掏出手机，仅需语音确认即可完成支付
	字节跳动 Ola Friend	接入字节豆包大模型 Seed-ASR(语音识别)技术模型，提供实时翻译、个性化推荐等功能
TWS 耳机	华为 FreeBuds Pro 4	支持通过语音或手势唤醒手机端“小艺智能体”直接进行交互，提供实时翻译、听力健康监测、AI 音频、鸿蒙生态设备协同等功能
	iKKO ActiveBuds	自带 AI 语音助手、45 种语言同声翻译、支持 WIFI/4G/蓝牙 5.3。AI 助手聊天可在本机实现，耳机实时翻译，会议录音(带会议纪要摘录)、语音转文本等功能。充电盒配备 1.8 英寸 AMOLED 屏幕，可作为智能音乐播放器
	Cleer ARC 5 音弧	搭载小度 DuerOS 的对话式操作系统，能够用于会议/录音要点生成会议笔记等 AI 办公场景、AI 体感操控、AI 抗风噪双向高清通话、音频降噪
	讯飞会议耳机	通话录音转写、音视频录音转写转译、现场录音转写，环境降噪
	iFLYBUDS Nano+	开放式耳机，语音识别准确率 96%，能实时处理和翻译语音信号，具有 AI 内容纪要整理功能
智能手表	时空壶 W4 Pro 同声传译耳机	致力于将 Apple Watch 打造成“手腕上的医疗实验室”和追踪用户活动的健康应用程序。
	Apple Watch	Phancy AI 智能体模组可本地运行端侧大模型，无需调用云端，提供天气、股票、资讯等信息查询、翻译、同声传译、会议速记及纪要整理、智能导航、运动助手等功能
	Phancy 兰博基尼、李小龙 AI 手表	将整合谷歌最新的 Gemini AI 技术，成为智能手表的核心语音助手。
	三星 One UI 7 Watch	

数据来源：Investors Business Daily, Kickstarter, The Sun, 四方维, 与非研究院, 深智联融媒, 电子发烧友网, 公司官网, 东吴证券研究所

1.3.2. 技术路径：“端优先”成趋势，可穿戴侧重协同算力

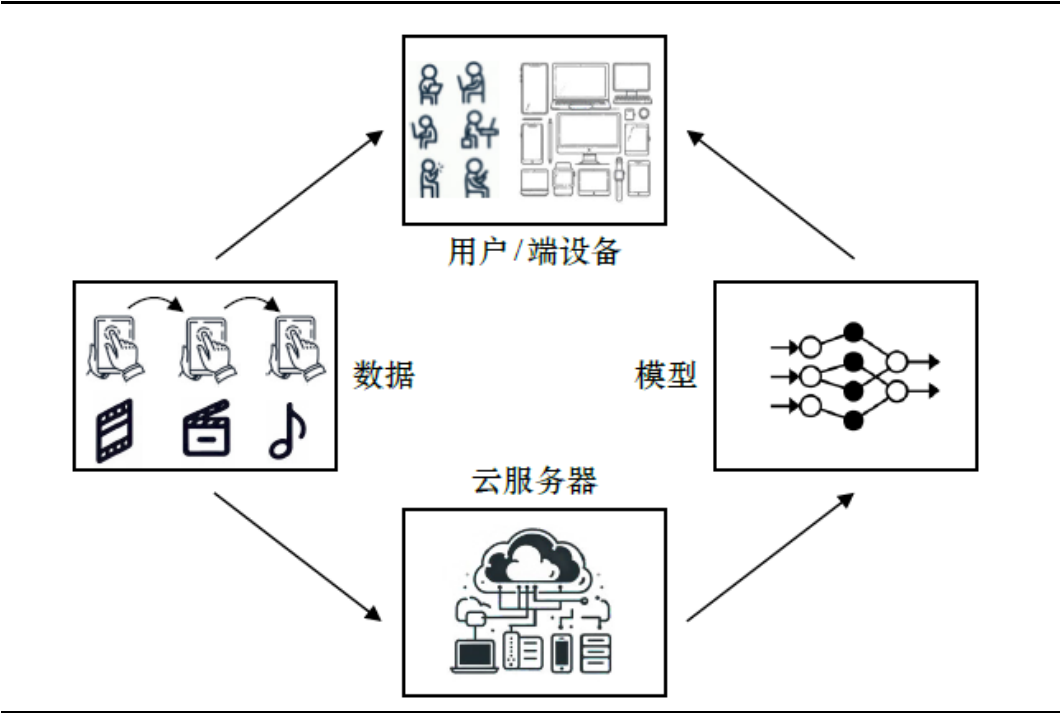
端云协同：云端大模型提供通用知识能力，端侧小模型提供个性化和快速响应。随着大语言模型在智能终端的落地需求上升，端侧集成能力成为技术焦点。传统的中心化云端机器学习架构虽具计算优势，但在实际应用中逐渐暴露出用户隐私数据风险、中心服务器带宽压力和离线不可用瓶颈。端云协同模式应运而生：云端负责大模型的集中训练与能力输出，端侧则部署经过裁剪优化的小模型以实现本地实时推理，并将关键反馈

返回云端用于进一步优化，形成“有机循环”的能力增强体系。

当前端云协同智能计算主要分为五类路径：

- 1) 高效计算硬件：在端侧部署高性能、低功耗芯片，支撑本地推理任务
- 2) 以端为中心的协同计算：终端主导计算，云端辅助。适用于隐私保护与个性化训练。
- 3) 以云为中心的协同计算：云端主导全局训练，终端仅保留本地数据并回传模型更新。解决隐私保护与模型泛化兼容问题。
- 4) 端云双向协同计算：端云模型双向优化、协同推理与训练。
- 5) 可信端云协同计算：结合因果推断、去偏学习等方法，提升多端异构环境下的稳定性与可信性。

图5：端云协同智能计算的基本要素



数据来源：《端云协同智能计算的关键问题、方法和应用》，东吴证券研究所

PC 和手机端：端云协同向“端优先”策略收敛。我们认为产业已经在“大模型云端、轻模型端侧”这一基本范式上形成高度共识：多数厂商都将轻量级的语言模型下沉至终端，用以实现实时推理、离线可用与隐私本地化；而更大规模的通用模型仍托管在云侧，通过加密或可信执行环境等机制来保证远程调用的安全可信。共同诉求背后，对应的正是前文提到的三大痛点——隐私保护、中心服务器算力压力和离线可用性——这几乎驱动了各家向“端优先”策略收敛。

表5：重点厂商端云协同架构

厂商	大模型	端云协同模式
----	-----	--------

Apple	Apple Intelligence	端侧部署轻量化语言模型（~3B 参数），在端侧进行推理；云端部署大模型，通过私有云计算确保安全访问。
OPPO	AndesGPT	推出多种不同参数规模的端侧模型规格，端侧用于任务响应、降级应答等场景；云端结合可信计算技术提供可信的大模型服务
小米	Xiaomi HyperAI	端侧模型基于动态稀疏化压缩； 为终端设备生态构建动态能力矩阵，动态分配 AI 任务，实现跨设备算力与数据的无缝调度； 基于可信执行环境构建“向量化处理+端云数据加密传输”“端侧优先、端云协同”的技术策略，能在端侧处理的数据绝不上云。在面对复杂的 AI 任务处理时，在保障用户隐私和数据安全的前提下调用云侧大模型能力
vivo	蓝心智能	着重提升端侧 AI 能力，与合作伙伴协作开发多样化的端侧 AI 应用场景
三星	Galaxy AI	边缘端借助 GPU 加速硬件在本地进行 AI 推理；云端采用多云架构支持灵活资源调度，可将部分任务回传到边缘执行。
DELL	NativeEdge	在端侧部署 DeepSeek-R1 7B 模型，完成 AI 总结、翻译、写作等功能；将发布可信计算环境下的联想个人云，升级到端云一体技术架构
联想	天禧 AS	

数据来源：公司官网，DOIT，腾讯云，证券时报，一起联想，东吴证券研究所

可穿戴端侧：“端-近端-云”三层协同，云端推理重要性提升。以 Ray-Ban × Meta 智能眼镜为例，设备本体承担唤醒词识别等轻量实时推理；手机作为近端算力负责任务转发与复杂任务处理；云端托管大规模多模态模型，为复杂推理与知识检索提供高精度输出。未来，“端—云—近端”动态算力调度将成主流，手机等近端设备将成为协同核心。同时，可穿戴设备的多模态感知能力（如 ECG）为个性化 AI 模型提供高价值数据支撑，助力厂商构建生态闭环，提升用户粘性与产品附加值。

图6：可穿戴端侧的“端-边-云”架构



数据来源：芯原股份，Google，东吴证券研究所

2. 手机端侧 AI 发展驱动硬件升级，有望成为新一轮换机潮核心
推力

2.1. 手机厂商端侧大模型布局对比及核心技术路径梳理

目前，各大智能手机厂商已相继推出可部署于终端设备的 AI 大模型，其发布的端侧大模型大多根据参数规模划分为多个版本，整体参数规模尚处于相对轻量化阶段，普遍低于 7B 参数。其中，实际在终端设备上实现大规模应用的模型多集中在 3B 参数级别，兼顾性能与算力资源限制。

表6: 主要智能手机厂商端侧大模型部署情况对比

厂商	时间线	端侧参数规模/算法演化	新增/升级能力
苹果	2024 年 iOS 18	搭载 Apple Foundation Model 3B，采用低比特量化与 LoRA 微调，支持 30 tokens/s 推理	文本润色、通知摘要、AIIens 识图
	2025 年 WWDC25		引入 AIKit 开放接口
Google	2023 年	Gemini Nano 1.8B / 3.25B，量化压缩部署于端侧，搭载于 Pixel 8 Pro	聊天、即录即译、代码补全
三星	2024 年 1 月	Galaxy S24 系列搭载 Gemini Nano，与云端 Gemini Pro 协同运行	在备忘录、录音、输入法等原生 App 中实现文字摘要、多模态处理
	2024 年开发者大会	发布 Gauss 2 Compact，多模态结构，将搭载于 Galaxy S24 之后的机型	可以处理图像、文本和计算机代码
小米	2023 年 8 月	自研 1.3B 模型部署于端侧	集成于小爱助手，支持系统级联动操作（如一键跳转小米商城）
	2024 年 5 月	MiLM-6B 模型注册完成，支持端侧部署	
	2024 年 11 月	推出 MiLM2 4B 模型，基于 6B 模型剪枝压缩	已在澎湃 OS、小爱同学、智能座舱、智能客服中开始应用落地
华为	2024 年 6 月	盘古端侧大模型 1.5B/7B	支持手机和 PC 等端侧绝大部分任务
	2025 年	基于 CBQ 算法优化的盘古大模型 1.5/7B 已应用于端侧部署	
荣耀	2023 年	发布 7B 模型，搭载于 Magic 6 机型，	
	2024 年 10 月	推出 3B 大模型，算力更强，搭载于 Magic 7 系列	搭载 YOYO 智能体，实现语音+行为链自动执行（如一键点咖啡）
vivo	2023 年 VDC	在端侧跑通 1B/7B/13B 模型	推出蓝心小 V 和蓝心千询 APP
	2024 年 VDC	发布 3B 端侧大模型，性能对比 7B 提升 300%	
	2024 年 11 月	发布 BlueLM-V-3B，多模态端侧模型	
OPPO	2024 年 1 月	Find X7 机型搭载 7B 参数大模型，量化压缩后可部署于手机	文字生成图片、图片生成文字、AI 文章摘要等
	2024 年 8 月	AndesGPT-2.0 发布，强化对话策略、安全训练和多模态能力	引入多轮对话建模、偏好对齐优化与 RLFH，支持语言+视觉+触控

数据来源：公司官网，IT之家，搜狐，中国科普网，TechRadar，GIZMOCHINA，东吴证券研究所

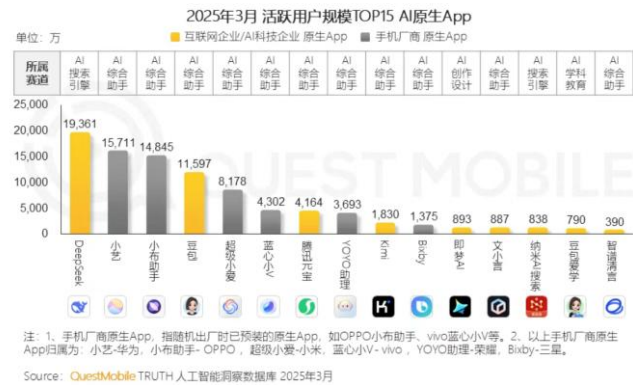
模型压缩主要通过量化、稀疏化和蒸馏三种方法，实现大模型在终端侧的高效部署。当前主流厂商在将大模型部署至终端设备时，普遍面临算力与存储限制，因此在模型达到 7B 规模后，均采用算法优化以实现轻量化部署。荣耀、vivo 在 7B 模型基础上推出 3B 版本，小米则在 6B 模型后发布了剪枝版 4B 模型，华为、OPPO 等也采取了算法优化策略。这些优化主要包括量化、稀疏化和蒸馏三类：

- **低比特量化：**通过将原本使用的高精度浮点数（如 16-bit）转换为低位宽整数或浮点数（如 8-bit 甚至 4-bit），以降低存储占用、加快推理速度、减少功耗。例如，从 16-bit 压缩到 8-bit 理论上可减半内存和功耗，同时实现约 1 倍的推理加速。目前 4-bit 量化已逐步成为手机端主流方案。
- **稀疏化（剪枝/MoE）：**通过移除模型中“冗余”或“贡献较小”的权重参数，实现模型压缩与计算简化。以剪枝为代表的稀疏化方法能显著减少模型计算量，同时尽可能保留原有精度；Mixture of Experts（MoE）等结构也在端侧应用逐渐展开。
- **模型蒸馏：**通过结构更大、精度更高的教师模型指导体积小、结构简化的学生模型进行训练，最终在保证性能接近的前提下显著缩减模型体量。

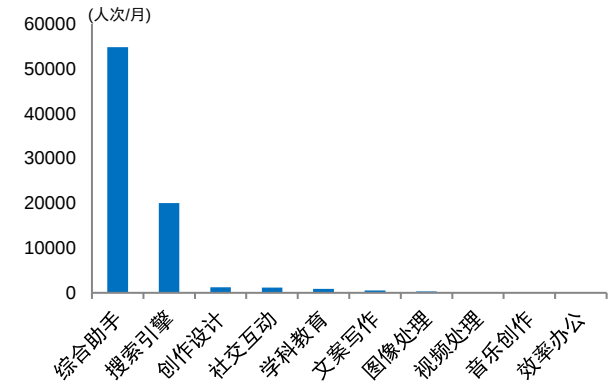
端侧大模型加速落地智能语音助手等核心场景，用户认可度与生态规模同步提升。目前端侧大模型的应用主要集中在智能手机语音助手的升级，推动其从传统指令识别向具备自然语言理解与任务执行能力的智能对话系统演进，支持如文本摘要、多语种翻译等轻量级生产力场景。以华为为例，其基于盘古 L0 开发的 L1 对话模型已在小艺语音助手中落地，具备自主识别复杂语义并执行任务的能力，显著提升用户交互体验，尤其适配老龄化需求。随着用户对端侧 AI 功能认可度的提升，整体生态亦快速扩展。海外方面，谷歌 Gemini 月活跃用户已达 4 亿，开发者超 700 万，形成成熟闭环；国内六大手机厂商均跻身 AI 原生应用活跃榜前列，其中华为与 OPPO 智能助手月活超亿。端侧 AI 正加速融入用户日常，具备成为下一代主流人机交互入口的潜力。

图7：国内 2025 年 3 月移动端 AI 原生 App 活跃用户规模

图8：AI 综合助手类应用活跃用户数量最多



数据来源：QuestMobile，东吴证券研究所

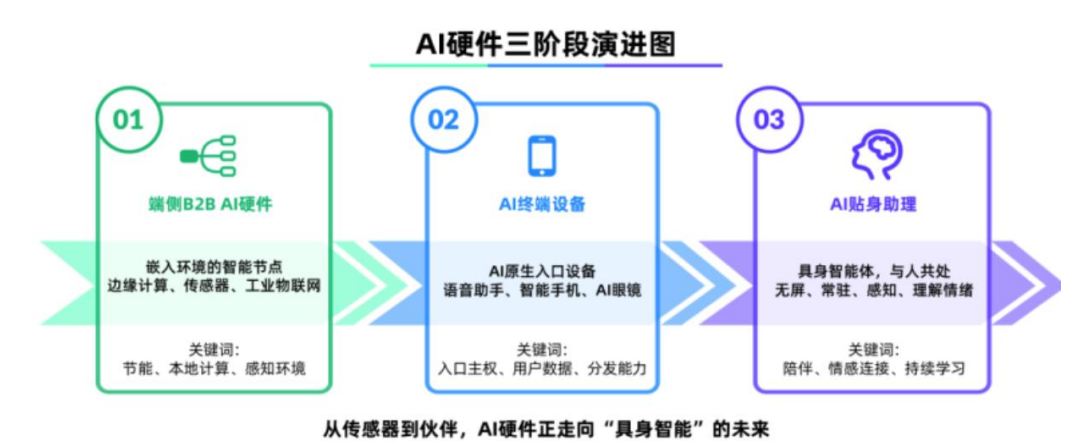


数据来源：QuestMobile，东吴证券研究所

2.2. 推演未来：多模态与跨应用驱动模型扩容，参数规模与压缩算法面临双重挑战

行业共识聚焦“端侧具身智能”，端侧大模型需构建多模态与跨应用调度两大关键能力。OpenAI 首席执行官 Sam Altman 指出“未来 AI 的潜力要想真正释放，必须让用户以自然的方式日常使用 AI”。该判断强调 AI 的价值不应局限于通过传统 App 方式调用，而应实现随时随地、无感交互的“具身智能体”形态。智能手机凭借低时延、离线可用与持续交互优势，正成为具身智能体早期形态，例如无需唤醒词、实时语音响应等功能，显著提升了用户体验和陪伴属性。面向下一阶段发展，端侧大模型需构建多模态感知与内容理解和跨应用执行两项关键能力。其中，多模态能力优先落地，代表厂商如华为、OPPO、谷歌已实现本地图像识别、语音翻译与场景解读等功能；跨应用执行能力也正成为核心竞争点，使模型能统一调动多个 App 完成任务，推动从“工具集合”向“任务助手”演进。

图9：终端 AI 设备向常驻式、无屏化具身智能体迈进

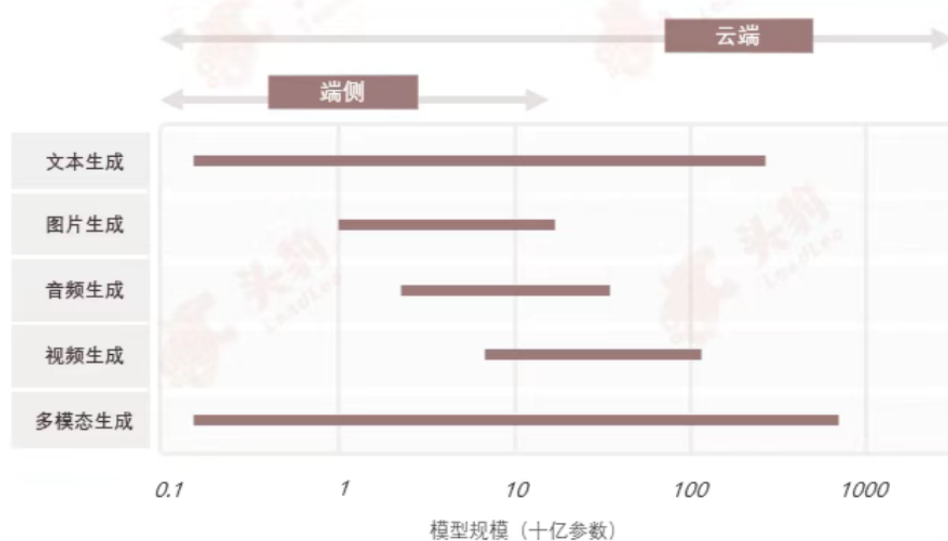


数据来源：物联网智库，东吴证券研究所

为支撑多模态感知、跨应用执行等关键能力的落地，端侧大模型仍需在模型规模上持续演进。目前行业普遍认同，大模型的“涌现能力”依赖于参数堆叠到一定规模后才能显现；反之，若参数压缩过度，则存在能力坍缩风险。以小米 MiLM2 为例，其模型矩阵划分显示，6B 与 13B 参数规模是具备零样本能力与上下文学习潜力的关键起点，能够在多任务场景中接近甚至达到百亿级开源模型的表现。

端侧大模型在多模态感知与跨应用执行两大核心能力上的突破，均依赖于参数规模驱动的能力涌现。具体来看，多模态感知能力需借助图像、音频与文本等模态在同一语义空间中的精准对齐。当前主流做法是通过专门的编码器生成各模态向量，并借助 Cross-Attention 或 Perceiver-Resampler 机制实现融合。若模型容量不足，语义映射空间维度受限，容易导致各模态“各说各话”，推理失真，因而对参数规模提出刚性要求。而跨应用执行能力则依赖模型对函数/工具调用接口的理解与调度能力，通常以结构化 JSON Schema 格式触发外部 API 调用，并将结果写入上下文持续推理。ReAct 框架的“思考—行动”交替范式也需要模型具备更强的规划能力与记忆能力，对注意力机制宽度和激活带宽提出更高要求。Meta 官方 LLaMA 文档明确指出 8B 模型无法稳定完成复杂工具调用任务，推荐使用 70B 以上规模以保障交互稳定性。

图10：端侧大模型业务场景分析



数据来源：头豹研究院，东吴证券研究所

尽管模型压缩技术不断演进，但这些方法本身也存在技术瓶颈。首先，低比特量化会显著增加推理误差，尤其在多模态模型中更为敏感。据 Xie et al. (2024) 研究表明，多模态模型对量化噪声极为敏感，4-bit 以下精度会导致误差成倍增长，影响语义对齐精度。其次，稀疏化虽可降低计算开销，但对模型的跨应用调用成功率、长链任务执行能力以及模态融合表现均存在负面影响。知识蒸馏方面，若教师模型所学习的跨模态语义结构在学生模型的低维空间中缺乏映射通道，知识迁移效果将受限，难以保留复杂任务能力。

在上述限制下，头部厂商并未放弃提升端侧模型参数规模。据彭博社，苹果目前已在内部测试 7B、33B 等不同规模的大模型版本，表明其正为未来在端侧部署更大模型能力进行中长期储备。

2.3. 算力、存力、电池和其它结构件协同进化，硬件成为大模型落地关键

算力：预计 SoC 需实现 50 TFLOPS 以上 AI 算力以支撑复杂推理与多模态交互。

根据 OpenAI 在《Scaling Laws for Neural Language Models》中的测算公式：算力需求 $\approx 2 \times \text{参数量} \times \text{Token 数}$ ，可粗略估算出主流模型在端侧运行所需的计算资源。例如，对一个包含 1000 个 Token（约合 750 个英文单词或 500 个汉字）的输入进行推理，当前主流 7B 模型大致需要约 14TFLOPS 的浮点运算能力。随着厂商尝试在端侧运行 7B 乃至 10B 级别大模型，对 AI 计算资源的要求将进一步攀升，未来高端 SoC 需至少具备 50TFLOPS 以上的 AI 算力，以实现多模态处理、复杂推理与持续交互等能力的实时响应。

表7：不同模态对应 token 换算

模态	计量单位	所需平均 token 值
英文文本	1 word	$\approx 1.3 \text{ tokens per word}$
简体中文字符	1 汉字	2 – 3 tokens 每字
图像 (ViT, 224×224 , patch16)	1 张	197 tokens = $14 \times 14 \text{ patches} + [\text{CLS}]$ ，可以结合算法降低 token 数
图像 (ViT, 512×512 , patch16)	1 张	1025 tokens = $32 \times 32 + [\text{CLS}]$ ，可以结合算法降低 token 数
音频 (AudioLM 16 kHz)	1 秒	Token 数 = [音频时长 (秒) / 帧步长 (秒)] $\times$ 每帧 Token 数
视频	1 段	Token 数 = 视频时长 (秒) $\times$ 帧率 (fps) $\times$ 每帧 Token 数。

注：不同的模型对 Token 的理解和处理方式可能有所不同，具体的换算规则和 Token 数量会因模型而异。

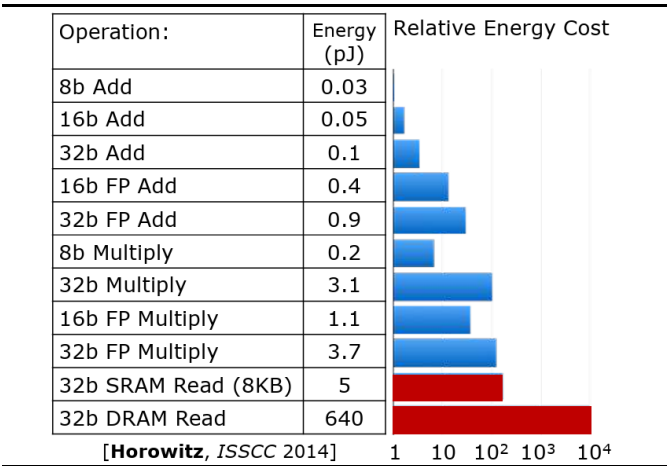
数据来源：OpenAI, Keras, Google, CSDN, 东吴证券研究所

内存：未来设备预计需配备 32GB 以上 RAM。模型参数和中间激活会消耗大量内存，不同精度的参数消耗内存不同。例如 FP32 格式每个参数占用 4 字节，而 FP16 则为 2 字节，当前主流端侧部署多采用 4-bit 或 8-bit 低精度表示，以压缩模型体积。以 7B 模型为例，4-bit 精度下参数存储约为 3GB；若采用 8-bit 精度，14B 模型参数则约为 12GB。目前高端智能手机普遍配备 12GB 至 16GB RAM，考虑到系统运行及多任务场景中其他 App 需占据一定内存空间，厂商通常将端侧大模型通过低比特量化等方式进行模型压缩与优化。综合评估模型推理所需的参数加载及系统安全冗余，我们预计未来支持 14B 及以上端侧大模型的设备预计需配置 32GB 以上 RAM。内存规格的提升将与算力一道，成为未来高端终端硬件升级的关键方向。

功耗管理：内存访问成主要能耗瓶颈。推理所产生的能耗主要由两部分构成：一是

乘加操作（MAC）本身的计算功耗，二是数据在内存之间搬运所引发的访问功耗。前者功耗极低，例如在 8-bit 精度下，单次乘法和加法操作仅消耗约 0.2 pJ 与 0.03 pJ，即便是 7B 模型、以 20 tokens/s 速度运行，整体计算功耗仅约 0.06W，即便模型参数进一步扩大，计算层面的功耗增长仍相对可控。真正的能耗瓶颈来自于内存访问，尤其是外部 DRAM。根据 MIT 研究，单次 DRAM 访问功耗高达 640 pJ，远超计算本身，因此提升 AI 计算能效比的核心在于优化内存架构与访问策略。芯片厂商已在此方向加快投入，如 OPPO 在 Find X7 中通过优化并行计算和内存调度策略，显著压缩了整体功耗。

图11：读取内存数据功耗远高于计算



数据来源：MIT，东吴证券研究所

图12：高通 Snapdragon 8 Gen3 AI 任务性能每瓦提升 40%



数据来源：TheShortCut，东吴证券研究所

电池：未来机型或需配备 6200mAh 以上容量。当前主流智能手机电池容量普遍在 5000mAh 左右，在 3.7V 电压下约合 18.5Wh。随着模型参数规模扩大、推理速度提升以及 AI 助手逐步向全天候待机、常驻交互方向演进，推理耗电将显著上升。据测算，未来典型推理功耗(下表 8 中 ⑧+⑨)有望达到现阶段的 4 倍左右，叠加运行时间的增长，对电池提出更高要求。为保障高强度 AI 应用下的续航体验，我们测算得到下一代支持本地大模型的智能终端，其电池容量需进一步扩展至 **6200mAh 以上**，以支撑全天候 AI 辅助的使用场景。

为评估未来端侧大模型在智能手机上的部署压力，我们参考 OpenAI 提出的模型推理计算规则，并结合主流芯片功耗与内存结构，构建如下测算框架，对未来 3 年端侧推理对算力、内存、电池等关键硬件资源的要求进行量化预测。核心测算逻辑与假设如下：

- **端侧大模型推理负载假设：**以当前主流 7B 参数规模、INT4 运算精度模型为基础，预测未来模型将扩展至 14B 并采用 INT8 精度，以提升模型表现力；同时，单轮推理 token 数由当前的 1000 提升至 2000，吞吐速率从 20 tokens/s 提升至 40 tokens/s，以满足未来语义任务在响应速度与上下文长度方面的要求。同时日均使用时长从 0.5h 增长到 1h。
- **内存访问能耗假设：**7B 参数模型经 INT4 量化后约需 3.5 GB 内存，显著超过

手机 SoC 内部 eSRAM 容量，因此模型推理过程必然依赖 DRAM 与 SRAM 的联合访问路径。参考 MIT 能效模型，FP32 精度下 SRAM 访问为 5 pJ，DRAM 访问达 640 pJ。考虑当前主流模型多采用低位宽量化，同时配合内存复用与片上缓存管理优化 DRAM 访问占比，综合设定单位访问能耗为 **3 pJ**。未来 3 年随着运算精度提升（至 INT8）及 SoC SRAM 容量扩展、访存调度能力增强，整体访存能效将基本持平，故继续采用 **3 pJ 假设不变**。

- **整机功耗构成假设：**整机功耗按三部分组成：核心乘加计算功耗+存储访问功耗+系统/屏幕运行功耗。其中，屏幕亮度维持常规亮屏状态（约 300nits），根据 ProductScience，其静态亮屏功耗约为 0.8W，本测算据此设定系统基础功耗 1W，包括屏幕、电源管理等模块损耗，大模型推理功耗在此基础上叠加。
- **电池相关假设：**目标放电深度为 80%，即用户在日常使用中平均在剩余 20%电量时开始充电；电池标称电压为 3.7V。

表8：端侧 AI 推理负载下的手机硬件资源需求测算

步骤	指标项	当前配置假设	3 年后演进假设	计算公式
①	大模型参数规模（Billion）	7	14	
②	运算精度	INT 4	INT 8	
③	单轮推理 token 数	1000	2000	
④	推理吞吐率（tokens/s）	20	40	
⑤	AI 日均连续运行时间（h）	0.5	1	
⑥	理论算力需求(TFLOPS)	14	56	$=2*①*③/1e12$
⑦	内存需求（GB）	3	13	$=②*①/8$
⑧	核心乘加操作功耗(W)	0.06	0.22	$=2*①*④*0.2/1e12$
⑨	存储访问带来的能耗（W）	0.84	3.36	$=2*①*④*3/1e12$
⑩	整机功耗（W）	1.90	4.58	$=⑧+⑨+1$
⑪	所需额外电池容量（mAh）	303	1211	$=(⑧+⑨)/0.8*1000/3.7$

数据来源：OpenAI，MIT，东吴证券研究所测算

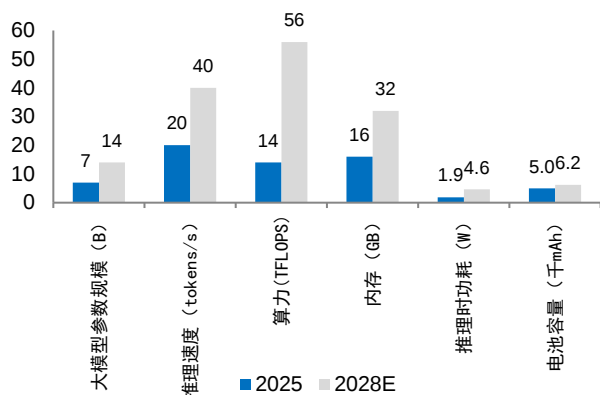
在端侧大模型推动人机交互形态重构的背景下，智能终端对硬件感知系统提出更高要求，麦克风与交互按键等模块正加速演进：

麦克风：厂商正致力于实现全天候、深度降噪的语音输入能力。Syntiant 斥资 1.5 亿美元收购 Knowles 消费级 MEMS 业务，意图将 AI 算法深度集成至扬声器模组中，提升语音识别前端的智能化水平；Aizip 与博世（Bosch）联合开发深度降噪模型，即使在高噪声环境下也能保证语音输入的清晰稳定。我们认为，这类创新有望加速智能手机向“24 小时常驻感知体”演进，成为用户的“耳朵和喉咙”。

交互方式：结构件正在被重新定义。iPhone 16 已将侧边相机键与 ChatGPT 的视觉识别功能相结合，实现通过长按按钮调出 AI 识别镜头所见内容的能力。由此推测，未来侧边键有望扩展出更多 AI 快捷功能，如一键呼出语音助手、AI 写作或翻译等，强化

手机作为具身智能体的主动交互入口。新型硬件的持续升级将为端侧大模型提供更强感知与响应基础，成为智能终端差异化竞争的重要方向。

图13: 端侧大模型推理演进驱动硬件配置全面升级



数据来源: 东吴证券研究所测算

图14: iPhone 侧键唤醒 GPT 功能或带动结构件新需求



数据来源: 哎咆科技, 东吴证券研究所

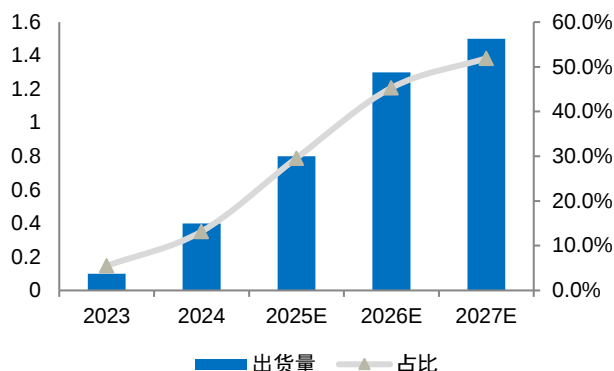
2.4. AI 成为新一轮智能手机换机潮的核心驱动

AI 能力成为驱动智能手机硬件升级与换机需求上行的核心变量。随着端侧大模型能力的持续演进, AI 正逐步从辅助交互工具向“具身智能体”转变, 应用场景正从语音助手延展至多模态理解、内容生成和个性化生产力等高价值任务。这一演化路径带动了对更高参数规模模型的需求, 进而对手机 SoC、内存、NPU、存储与散热等核心部件提出更高配置要求, 构成推动手机硬件系统升级的重要力量。

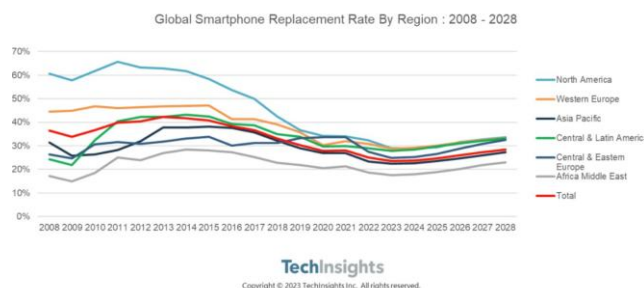
产业层面, 各大手机厂商已密集推出定位于“AI 手机”的新一代旗舰机型, 加快产品更新节奏。根据 IDC 数据, 2024 年中国 AI 手机渗透率由 2023 年的 5.5% 升至 13.2%, 预计 2025 年将达到 0.8 亿台。从长期趋势来看, IDC 预测到 2027 年, 中国 AI 手机出货量将进一步增长至 1.5 亿台, 渗透率达到 51.9%, 成为智能手机市场的重要组成部分。AI 能否真正驱动用户换机, 核心取决于应用层是否能精准捕捉用户高频、刚性需求, 拉动消费侧热情。

图15: 2023-2027E 国内 AI 手机出货量及占比 (亿只, %)

图16: 全球智能手机换机率在 2024 年迎来拐点



数据来源：IDC，东吴证券研究所



数据来源：TechInsights，东吴证券研究所

手机换机周期有望反转，AI 功能演进驱动硬件迭代提速。据 Counterpoint，全球智能手机换新周期近年来持续拉长，2023 年达 43 个月。然而，AI 手机的兴起正有望改变这一趋势。全球最大运营商回收平台 Assurant 表示 4Q24 全球被回收的 iOS 与 Android 设备平均使用年限均出现下降，该变化与 AI 手机的集中发布时间高度重合，或表明 AI 手机的早期消费者正在淘汰旧设备，加速转向新一代终端。

随着手机 SoC/NPU 能力提升，端侧部署大语言模型（LLM）成为新一代高端智能终端的主要竞争方向。大模型参数规模、推理频率与整机功耗同步抬升，逐渐对电池模组带来连续性负载压力，进而可能加快电芯老化并压缩整机使用寿命，成为推动换机周期变化的新变量。为量化该趋势，我们以当前主流 7B 参数规模、INT4 精度模型为基准，结合典型使用情景（日均推理运行 30 分钟），预测中期（3 年后）演进至 14B、INT8 模型时，用户日均推理时长提升至 1 小时，并测算在两类情景下电芯所需充放循环频率与预期寿命变化，结论如表 9 所示。核心测算逻辑与假设如下：

- **电芯寿命阈值：**依据 Apple 官网披露，iPhone 14 及以前机型所用电池支持 500 次循环至剩余 80%容量，iPhone 15 起提升至 1000 次。本测算中取中值 800 次循环作为评估基准。
- **电池相关假设：**目标放电深度为 80%，即用户在日常使用中平均在剩余 20%电量时开始充电；电池标称电压为 3.7V。

表9：端侧 AI 推理负载下的手机使用寿命测算

步骤	指标项	当前（7B INT4）	3 年后(14B INT8)	计算公式
①	额外整机功耗 (W)	0.9	3.58	见表 8
②	AI 日均连续运行时间 (h)	0.5	1	
③	电池容量 (mAh)	5000	6200	见表 8
④	电芯寿命 80%所对应用循环次数 (次)	800	800	
⑤	日均额外耗电 (Wh)	0.45	3.58	=①*②
⑥	电池总能量 (Wh)	18.5	22.94	=③*3.7/1000
⑦	AI 负载带来的额外年循环次数 (次)	9	57	=⑤*365/⑥

⑧	基础年循环次数（次）	292	292	=0.8*365
⑨	年总充放循环次数（次）	301	349	=⑦+⑧
⑩	电芯寿命（80%）所对应可用年限（年）	2.7	2.3	=④/⑨

数据来源：苹果官网，东吴证券研究所测算

用户对电池续航高度敏感，大模型运行带来的高功耗将加速电芯老化，成为推动换机周期提前的重要驱动因素。在当前配置下（7B 模型、INT4 精度、AI 日均运行 0.5 小时），电芯至剩余 80% 容量所对应的可用年限约为 2.7 年。若未来推理模型迭代至 14B、INT8 精度，且 AI 使用时长提升至日均 1 小时，测算显示年均充放循环次数将上升至 349 次，电池寿命相应压缩至 2.3 年，减少约 0.4 年。考虑到电量快速衰减在实际使用中容易引发“续航焦虑”，电池老化构成驱动换机的重要心理因素。在 AI 功耗持续上行、单日使用时长增加的趋势下，电芯寿命成为智能手机用户主动换机的主因之一，从而推动换机周期提前。

3. AI 终端重塑价值链：从硬件销售到智能生态

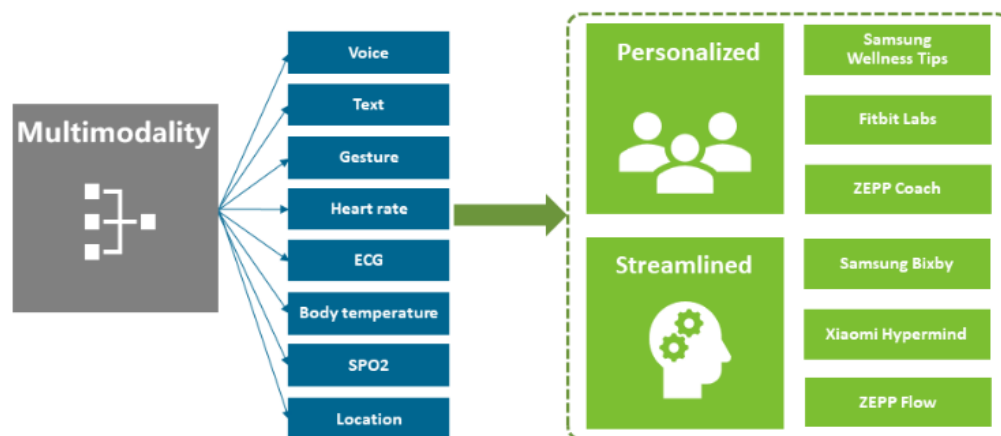
3.1. 商业模式重塑：订阅变现+产品协同，AI 终端引领价值体系重构

3.1.1. 多设备协同驱动 AI 终端生态成型，可穿戴产品成差异化关键入口

智能终端正从单一设备形态演化为多设备协同、跨场景融合的智能服务节点，驱动手机厂商加快向 AI 终端生态平台转型。荣耀于 MWC 2025 发布“阿尔法战略”，宣布从手机制造商向全球领先的 AI 终端生态公司转型；中兴推进“AI for All”战略，推动 AI 能力在全品类终端中的融合落地。小米“人车家全生态”实现设备实时动态组网，所有设备协同如一个整体。华为的新版 HarmonyOS 以统一 AI 底座、系统级智能体、原生智能应用和开放生态为支撑，推动“小艺”向具备全场景感知与多模态交互能力的智能体升级。

我们认为主流手机厂商加速向 AI 终端生态平台演化，核心驱动力在于通过多模态 AI 体验构建差异化优势，提升用户粘性与品牌护城河。前文提到，可穿戴设备是多模态输入的入口。可穿戴设备在此过程中不仅作为多模态交互入口，更承担关键的数据输入与生态协同角色，将带动整个 AI 终端产品线的销售增长与生态价值重估。例如，华为强调端云协同下的“场景智能”，小米构建人车家一体的全场景协同体验，均体现出头部厂商对 AI 终端战略的系统性投入与落地节奏。

图17：多模态感知赋能可穿戴设备个性化智能体验



数据来源：Canalys，东吴证券研究所

3.1.2. 订阅收入成为增长引擎，推动终端厂商盈利模式转型

订阅收入在 AI 硬件厂商中的比重持续提升，推动其利润结构从一次性销售向“硬件+服务”并重转型。以 Oura Ring 为例，其售价 349 美元，订阅费 5.99 美元/月，订阅业务自 2022 年推出后已贡献 20% 收入，毛利率高达 80%，显著优于硬件销售的 20-30%。Plaud Note 则通过 159 美元硬件+9.9 美元/月订阅的模式，年化收入已达 1 亿美元，连续两年实现 10 倍增长。订阅模式不仅抬高产品均价，也增强用户粘性和收入可持续性。我们认为，该模式变革将在产业链上带动云推理、模型部署、订阅管理等基础设施需求，

利好云计算相关厂商。同时，对终端企业的算力、模型能力与隐私保护提出更高要求，提升其云端生态构建的重要性。估值体系也将由出货导向转向 ARPU、付费渗透率与留存率等 SaaS 指标，具备高订阅贡献与用户粘性的硬件公司有望获得类似互联网企业的估值溢价。

3.2. 端侧厂商的传统商业模式在智能体时代会如何变化

在前一节中，我们系统分析了消费电子端侧商业模式在未来可能面临的结构性变革与商业模式演进趋势，核心包括：1) 智能体逐步取代传统 App，成为智能交互的中枢，端侧 AI 能力从工具向生产力伙伴演进；2) 多模态终端矩阵助力端侧 AI 实现个性化建模与人机共生，可穿戴设备拉动整个终端产品矩阵；3) 订阅收入推动盈利模式变革，增强估值弹性；4) “私有语境”成为差异化竞争核心，数据构建护城河。

下一步我们将探讨当前消费电子端侧厂商长期积累的传统竞争优势——包括平台制订阅收入、生态布局、性价比定位、本地化运营能力及全产业链协同优势——将如何在 AI 原生浪潮中被重塑或重估。

3.2.1. 订阅收入占比增大

我们判断端侧 AI 将进一步强化“硬件+订阅”的商业模式，推动终端厂商订阅收入占比提升，从而增强盈利结构的稳健性与可预测性，并有望带动整体估值体系向 SaaS 逻辑靠拢。Apple FY25 Q1 服务收入 263 亿美元、付费用户破 10 亿；在已上线 Apple Intelligence 的市场，iPhone 16 同比增速更快，印证 AI 功能对需求的拉动。三星方面，目前其 Galaxy AI 功能对用户开放“免费至 2025 年底”，我们认为其意在培养使用习惯，为订阅收费埋下价格锚点。小米方面，2024 年小米互联网服务收入达到 341 亿元人民币，同比增长 13.3%，创历史新高。毛利率 76.6%，显著高于集团总体毛利率。2024 年 12 月，MIUI 全球月活跃用户达 7.02 亿，同比增长 9.5%，为持续 ARPU 提升奠定基础。华为方面，2024 年其消费者业务（含 HMS、终端软件与数字内容）收入 3390 亿元人民币，同比增长 38%。

3.2.2. 跨平台协同与多设备智能体生态加强

在终端智能体能力逐步提升的背景下，主流厂商普遍强化跨平台协同和多设备智能生态的布局。小米 HyperOS 构建“人-车-家”统一操作系统底座，具备在 200+处理器平台间灵活部署能力，其 Hypermind 系统可基于用户行为预测需求并自动响应，实现智能体跨设备协作。华为宣布自 2025 年起，所有新发布的可穿戴设备将标配原生 HarmonyOS。目前，80 余款头部应用已完成 HarmonyOS NEXT 适配，实现了音频播放控制与离线支付等多场景操作，显著提升可穿戴设备的独立智能性。

端侧设备在生理状态感知、行为追踪等方面具备连续、精细的数据采集能力。根据谷歌研究，穿戴设备可提供如心率变异性、睡眠时长、步数等纵向健康数据，结合生成式 AI 将为用户提供更个性化的健康洞察与行为建议。考虑到端侧设备受限于算力与能

耗瓶颈，我们判断厂商构建跨设备计算协同的 AI 架构的目的在于，将数据采集功能下沉至可穿戴设备等前端硬件，而将模型推理任务转移至手机、PC 等近端设备。本地部署多模态模型与大语言模型将有助于提升响应时效、强化隐私保护，进一步巩固终端厂商在“数字分身”时代的用户锁定能力。

图18：华为可穿戴新品将全部搭载原生鸿蒙系统



数据来源：鸿蒙开发者官网，东吴证券研究所

图19：小米 Hypermind 使设备能够主动了解用户需求并采取相应措施



数据来源：小米官网，东吴证券研究所

3.2.3. 本地化适配能力有望从“加分项”演变为厂商核心护城河

我们认为在以端侧 AI 主导的新产业范式下，消费电子厂商的本地化能力正从传统意义上的服务优化升级为差异化竞争壁垒。具体而言，能否实现算法、语料、传感系统与本地文化语境的深度融合，决定了厂商在区域市场的用户黏性与定价能力。相比全球化方案，本地化 AI 模型在多模态感知与自然语言交互中更具适配性，尤其在语言多样、网络基础设施欠发达的地区具有明显优势。以传音为例，其在非洲市场积累大量用户数据，并推出支持多语言与口音识别的 Ella 2.0 AI 助手，初步形成本地化 AI 能力。随着功能迭代，有望进一步增强品牌粘性与区域市场竞争力。

3.2.4. AI 驱动下的渠道模式转型：从售卖硬件到构建场景体验

AI 的普及推动终端由单点功能产品向“场景化+生态化”体系演进。智能手机、智能电视等产品正逐步演化为家庭和个人的 AI 交互中枢，企业竞争重心也由“卖硬件”向“卖场景、卖生态”转移。在此过程中，AI 能力如何与智能硬件深度融合，成为厂商开拓新增量空间的核心命题。我们认为，随着“硬件+服务+AI”模式的推进，终端厂商亟需构建可视化、可体验的产品生态展示体系，这将利好小米等已建立线下体验式门店网络的厂商，形成渠道与体验协同的差异化优势。

运营商介入 AI 生态，有望推动渠道进一步下沉和融合。除终端厂商外，电信运营商也在加速布局 AI 能力，将服务触角从连接与资费扩展至智能体运营。韩国 SK 电讯将 AI 助手嵌入 TPhone，支持通话摘要等功能；Vodafone 则与微软合作，面向全球用户提供生成式 AI 服务。我们认为未来终端销售有望从“存话费送手机”演化为“订阅 AI 服

务赠终端”，运营商将在终端分发中扮演更重要的协同角色。

3.2.5. 全产业链协同加速终端产品迭代与成本优化

我们认为垂直整合产业链各环节可以有效缩短了 AI 终端从设计到量产的周期，提升硬件迭代效率，降低制造成本。三星在 SFF 2024 中正式提出“代工+存储+先进封装”一体化 AI 解决方案路径，在终端侧，这种高度协同模式可实现以季度为单位进行 OTA 新模型的快速部署，并率先应用于自有 AI 终端产品，构建“芯片-模型-终端”快速迭代闭环，加速产品能力进化与市场响应。此外，当前 AI 功能集成对算力硬件提出更高要求，尤其是在集成 NPU（神经网络处理单元）等高成本部件的背景下，我们认为全产业链一体化所带来的规模与协同效应，有望在制造、封装及系统集成等环节形成明显成本优势，进一步增强终端厂商在 AI 硬件竞争中的盈利能力与产品性价比。

4. 投资建议

随着端侧大模型参数持续增长（从 3B 向 7B、10B+演进），对本地算力、存储、功耗管理及散热等硬件能力提出更高要求，带来智能手机底层硬件架构的系统性升级。建议关注如下核心环节标的：

端侧算力：瑞芯微、恒玄科技、乐鑫科技、炬芯科技、泰凌微等

存储：兆易创新、江波龙、德明利、普冉股份、佰维存储、香农芯创等

充电/电池/散热：奥海科技、德赛电池、珠海冠宇、中石科技、思泉新材

端侧 AI 的大规模落地加速智能手机进入新一轮技术迭代周期，有望推动消费者换机节奏恢复，从而带动整机出货回升。建议关注整机设计制造及外观件配套环节的优质企业：

整机组装：立讯精密、华勤技术、歌尔股份、国光电器、龙旗科技、天键股份等

外观件/功能件：蓝思科技、鹏鼎控股、东山精密、领益智造、比亚迪电子

5. 风险提示

底层创新不及预期：端侧 AI 模型的发展需要底层技术架构和模型优化技术、硬件技术的创新，若创新不及预期，模型体验较差会影响用户体验。

技术发展不及预期：除了底层创新，新技术如新存储方式等的落地也需要后端制造环节的技术落地和成本优化，若技术发展不及预期，产品成本或过高无法普及。

竞争加剧风险：当前互联网厂商/消费电子品牌厂商等均在抢占入口环节，竞争或加剧。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。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口头承诺均为无效。

在法律许可的情况下，东吴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并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持：预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性：预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持：预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021

传真：（0512）62938527

公司网址：<http://www.dwzq.com.cn>