《中国端侧 AI 全景图谱报告》

编写单位：智次方研究院、广东智用人工智能应用研究院

- **智次方研究院**作为中国 AIoT 产业研究的引领者，致力于为产业输出深度洞察观点，为企业提供高价值研究服务，共同参与和见证智能物联行业的成长与发展。智次方研究院近十年来专注于智慧园区、智能制造、工业互联网、AI、车联网、5G、物联网、端侧 AI 等 AIoT 相关领域研究，为企业和政府提供 AIoT 产业相关的市场调研、数据洞察、业务/战略规划、投研尽调、行业分析、产业规划、园区规划、政策分析等咨询服务，助力客户洞察行业趋势、科学布局发展、实现价值增长。

- **广东智用人工智能应用研究院**成立于 2023 年 6 月，由多名前微软首席技术专家团队发起，致力于推动人工智能等前沿技术的产业化应用，是一家以人工智能技术与数据科学驱动产业变革的创新机构。研究院围绕人工智能技术的商业化场景展开深入研究，打造了自主可控的 AI Agent OS 产品——AI Agent Foundry，通过模块化工程设计、智能化场景应用以及透明化基础架构，降低 AI 应用成本，提升企业数字化转型效

率，以“智能体即服务”（Agent-as-a-Service）、“AI+人协作”，以及“决策智能”为核心范式，融合 AI 与数字化能力，加速人工智能在制造业、能源、科研、政府和跨境电商等领域的落地实践。目前，研究院已成功推动多项创新解决方案，为客户创造显著的商业价值。研究院秉持技术创新与产业赋能并重的理念，致力于成为全球人工智能产业应用的引领者，为经济高质量发展注入新动能。

研究团队：周闻钧、李宁远、黄浩权、管震、彭昭

前 言

端侧 AI，正在成为数字经济时代的智能新引擎。

随着人工智能技术从云端向终端迁移，端侧 AI 正成为推动智能设备革新的核心力量。当 AI 开始拥抱终端硬件，端侧 AI 以实体的方式切实让消费者感受到 AI 技术与终端硬件结合后带来的功能变革。端侧 AI 直接在终端设备，如智能手机、汽车、智能家居等上运行 AI 模型算法，实现本地化数据处理，具有低延迟、高隐私性、低成本等优势。

2023 年中国端侧 AI 市场规模为 1939 亿元，2025 年预计突破 2500 亿元，同比增长 35%，2030 年将达 1.2 万亿元，年复合增长率（CAGR）30.8%（中研普华数据）。在这场 AI 产业化变革中，端侧 AI 正在加速 AIoT 从 1.0 的“万物互联”迈向 2.0 的“**万物智联”**，进而推动 AI 在产业中的深度应用，实现更为彻底的智能化变革。端侧 AI 背后的模型技术与基础软硬件底座构建起了未来智能时代的蓝图，其影响不仅仅体现在提升数据处理能力，更在于推动数据驱动的闭环智能。

站在端侧 AI 产业即将爆发的前夕，如何科学解构端侧 AI 发展全景？又该如何把握新机遇，引领行业变革？由中国 AIoT 产业研究引领者“智次方研究院”联合“广东智用人工智能应用研究院”，携手倾力打造**《中国端侧 AI 全景图谱报告》**。图谱报告将深度剖析端侧 AI 发展现状，洞察发展重点与未来趋势，为业界奉上一份全面解读端侧 AI 发展的“集大成者”：

一、核心观点

1. **端侧 AI 正成为 AI 产业化的突破口**

随着 AI 从云端走向终端，端侧 AI 具备低延迟、高隐私、低带宽成本等优势，正在驱动新一轮智能终端革命。

2. **“芯模端智”一体化推动产业系统演进**

报告以“芯片-模型-终端-应用”四大模块为主干，全面解析了端侧 AI 产业链结构、协同路径和关键演进趋势。

3. **模型创新与软硬协同成为 ROI 优化关键**

以 DeepSeek 为代表的新一代大模型推动轻量化部署，结合 AI 模组与 SoC 芯片，实现端侧 AI 的商业可行性与规模落地。

4. **硬件即入口，入口即生态**

OpenAI、Apple、小米等先后布局 AI 原生硬件，端侧设备正成为 AI 应用分发与用户关系的“第一接触点”。

二、产业数据亮点

- **2025 年中国端侧 AI 市场规模将达 2500 亿元**，同比增长 35%；
- **2030 年预计突破 1.2 万亿元**，年复合增长率（CAGR）达 30.8%；
- **AI PC 出货量 2025 年将突破 1 亿台**，占全球 PC 市场 40%；
- **AI 手机 2025 年中国出货量预计达 1.18 亿部**，渗透率 40.7%；
- **AI 模组 2023–2027 年出货年复合增长率达 73%**。

三、技术趋势洞察

- **端侧 SoC 持续进化**：NPU 算力不断提升，支持轻量大模型本地推理；
- **存储芯片高带宽低功耗化**：UFS 4.0、LPDDR6、端侧 HBM 加速普及；
- **传感芯片智能化融合**：多模态感知芯片支持视觉、语音、力觉等任务；
- **AI 模组平台化演进**：模组不再只是通信组件，具备完整 AI 处理能力；
- **本地大模型部署成熟**：DeepSeek-R1、Qwen、MiniCPM 等已实现在手机、PC、机器人、模组等多终端运行。

四、典型应用爆发

终端形态	应用方向	代表厂商
智能汽车	智驾系统、智能座舱	华为、比亚迪、小米、蔚来、理想
AI PC	智能助手、文生文、AI视觉	联想、华为、苹果、戴尔
AI手机	多模态助手、生成式影像	小米、荣耀、OPPO、vivo
AI可穿戴设备	实时翻译、拍摄分析、语音交互	雷鸟、华为、万魔、疯米、OPPO、Rokid、XREAL、IN
具身机器人	家庭陪伴、工业作业	优必选、宇树、智元、星尘
智慧工业	智能制造、工控平台、设备维护	研华科技、海尔卡奥斯、格创东智、美的
智慧城市	安防监控、环境监测	海康、大华、宇视科技

五、全景图谱亮点

构建“芯-模-端-智”四层产业图谱：

- **芯**：涵盖 SoC、存储芯片、传感器、AI 模组等 70+企业；

- **模**：覆盖 DeepSeek、Qwen、MiniCPM 等端侧大模型及推理框架；
- **端**：聚焦智能汽车、AI 手机、AI PC、机器人、AI 眼镜等终端形态；
- **智**：深入剖析智慧汽车、工业、城市等重点应用场景与生态系统。

六、未来展望

- **技术趋势**：轻量化模型、异构计算、智能体框架将成为产业主线；
- **商业模式**：“模型即服务”与“智能体即入口”将重塑分发逻辑；
- **产业变革**：端侧 AI 将引领 AIoT 从“万物互联”走向“万物智联”；
- **投资机会**：SoC 芯片、AI 模组、机器人交互系统等赛道潜力巨大。

《中国端侧 AI 产业图谱报告》旨在成为业界了解端侧 AI 产业结构、技术趋势与应用落地的权威工具书。本报告将持续更新，欢迎产业各方交流合作，共绘端侧智能新图景。

智次方研究院

2025 年 7 月

目 录

目录

一、中国端侧 AI 产业年度洞察

洞察 1：2025 端侧是人工智能新战场

2025 端侧将是人工智能的新战场

随着企业开始广泛采用人工智能和数据分析技术，将数据集中到云端似乎成为一种顺理成章的选择。然而，这种方法也逐渐暴露出其局限性。

数据传输延迟、带宽成本和连接稳定性等问题，使得纯粹依赖云端的人工智能无法满足实时决策的需求，尤其是在制造业、物流业和医疗保健等对时效性要求极高的行业。不断攀升的云计算成本更是雪上加霜，加剧了这一问题。

知名研究机构 Gartner 预测，到 2026 年，由于管理公有云支出的难度不断增加，40% 的组织将放缓采用云计算的步伐。曾经被视为实现无限可扩展性的“银弹”，如今却面临着成本飙升和收益递减的窘境。

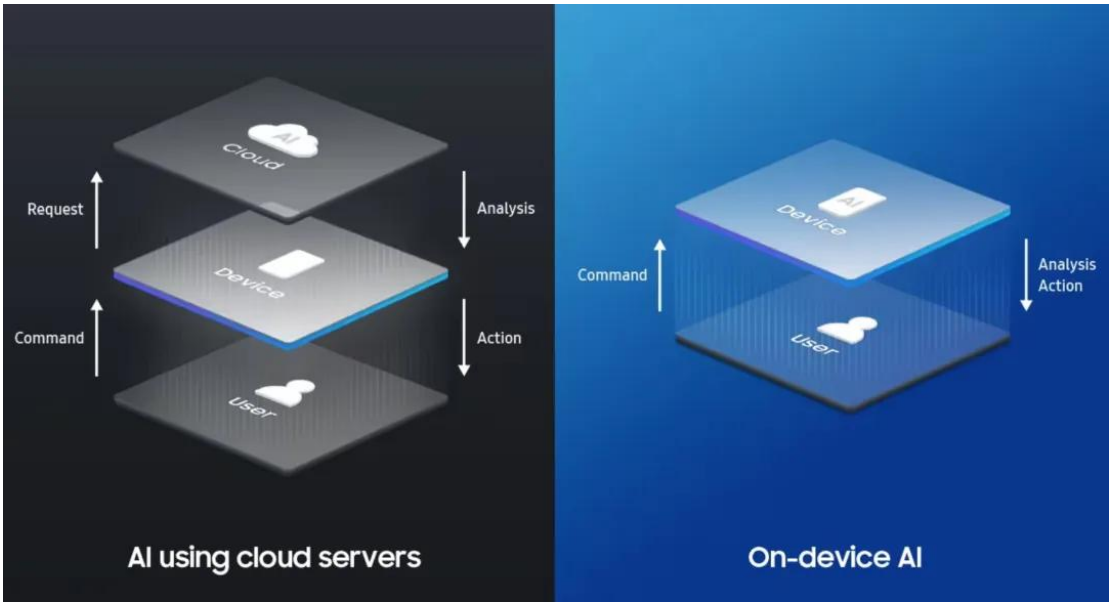

图 1：云侧端侧 AI 对比

（图源：物联网智库）

然而，AI 的未来并非“仅限于云”。它将演变为一种混合分布式模型，其中端侧和边缘在实时决策中扮演着关键角色，而云端则在训练大规模 AI 模型和存储长期数据方面仍不可或缺。这种组合巧妙地利用了两个系统的优势：边缘计算的高速和即时性，以及云计算的可扩展性和强大分析能力。

2024 年 12 月 17 日，英伟达 CEO 黄仁勋在发布的视频中，从烤箱端出最新“烹饪”的全新边缘开发套件 Jetson Orin Nano Super。这一产品的问世，不仅将会掀起一场全球范围内的 AI 开发板大战，也为中国企业在端侧和边缘 AI 领域的发展提供了难得的机遇。

“端侧 AI”意味着 AI 可以直接在移动设备上处理数据，无需连接到服务器或云端，能够在用户设备本地完成复杂的推理和决策。

端侧 AI 在对硬件提出更高要求的同时，也对其算力、能效以及软硬件协同等方面提出了新的挑战。为了满足端侧 AI 的需求，AIoT 芯片的能力日益增长。AIoT 芯片是一种集成了人工智能和物联网技术的系统级芯片，旨在实现智能化设备的连接、控制和数据处理。

回顾发展之路，端侧 AI 经历了一段辗转的旅程：从判别式 AI、增强式 AI 的领域，现在又来到了具有突破性的生成式 AI 前沿。每一步都让我们更接近未来，智能系统将无缝融入我们的日常生活，为我们带来不仅有感知，还有掌上创造的沉浸式体验。

从信息论的角度来看，这三种范式可以看作是对源熵的不同影响。判别性 AI 旨在降低熵，增强性 AI 或多或少地保持熵值不变，而生成性 AI 则会造成熵增。

随着端侧的优势逐步显现，端侧 AI 将从其当前的角色演变为 AI 驱动创新的核心推动者。

2025 年，终端和边缘设备不仅将会收集数据，还将充当智能枢纽，无缝运行先进的 AI 小模型，并推动决策更接近源头。

采用这种混合模型的组织将重新定义现代 AI 驱动的运营，为性能和适应性设定新标准。端侧 AI 将成为下一波人工智能创新塑造和实现的关键战场。

洞察 2：端侧 AI"算力 x 通信 x 存储"协同优化

从模型创新到软硬结合,端侧 AI"算力 x 通信 x 存储"协同优化决定商业价值

DeepSeek 引发的新一波生成式 AI 浪潮，本质上是端侧 AI 的红利。正是 DeepSeek 的出现，迫使我们重新评估 AI 投资回报率 ROI，并认识到端侧 AI 将成为提升 ROI 的新路径。

过去，生成式 AI 大模型一直面临着成本与价值之间的 ROI 困境。尽管 DeepSeek 等大模型的能力不断提升，但它们的训练和推理成本极高，限制了商业化落地的 ROI。目前，AI 的投资逻辑仍然围绕算力规模和模型能力，但这种模式的可持续性正受到挑战。

AI 落地的核心问题在于如何降低计算成本。

图 2：端侧 AI 价值重估矩阵

（图源：物联网智库）

传统云端 AI 计算的高昂投入，让许多企业，尤其是中小企业难以承受。如果 AI 模型不能在更低成本、更低功耗的环境下运行，那么它的商业应用将受限，投资回报率也难以提升。

而 DeepSeek 等大模型的出现，为解决这一问题提供了新的思路。从企业参与度来看，过去只有大型企业能负担端侧 AI 研发，而如今，借助 DeepSeek 低成本推理技术，中小企业也能在 AI 玩具、AI 眼镜等产品中融入强大的 AI 功能，推动端侧硬件智能化的普及。

通过端侧 AI 应用，DeepSeek 等大模型正在以更低的计算成本，在本地部署轻量化版本，提高推理效率。结合 AIoT 专用芯片，可以优化推理过程，减少云端算力消耗，提高整体 ROI。这种模式特别适用于智能制造、智能硬件、自动驾驶等应用场景，有望推动 AI 的大规模商业落地。

过去，AI 投资主要围绕提升模型能力展开，追求更大的参数规模、更复杂的神经网络架构。然而，计算成本与商业收益的平衡正在成为新时期 AI 投资的核心考量因素。

未来，AI 投资的关键，将会从“更强的 AI”，到“更高效的 AI”；从“单纯软件创新”，到“软硬结合”。AIoT 芯片、边缘设备、优化算法的发展，将重新定义大模型的商业价值。

图 3：端侧 AI 三角定律

（图源：物联网智库）

因此 AI 的商业价值将不再由单纯的模型能力决定，而是由计算成本与商业收益的平衡来定义。只有那些能够在算力、功耗、存储、通信等多个维度平衡商业价值的 AI 架构，才能真正实现可持续增长。端侧 AI 的崛起，将推动整个产业走向更加务实、可持续的发展之路。

洞察 3：AI 硬件正成为新的对决前线

硬件即入口，入口即生态：端侧 AI 硬件正成为新的对决前线

2025 年 5 月 22 日，OpenAI 宣布以近 65 亿美元（约合人民币 468 亿元）的全股权交易，正式收购由苹果前首席设计官艾维创立的硬件初创公司 IO Products。

这项交易的核心目的，不是为了扩充硬件营收渠道，而是为了解决一个更根本的问题——分发。因为模型再强，没有入口也难落地。

在生成式 AI 的早期阶段，谁能训练出参数最多、效果最惊艳的大模型，是竞争的关键。但到了今天，模型之间的差距正在逐渐缩小，AI 的“护城河”正从模型本身转向另一个维度：用户入口与应用分发能力。

OpenAI 的 CEO 奥特曼非常清楚这一点。他曾公开表示，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”。

这意味着，依赖手机 App 或网页访问 ChatGPT 并不是长久之计。相较之下，Google 拥有 Android 操作系统与 Chrome 浏览器，Meta 拥有社交网络和智能眼镜，甚至 Apple 也在布局系统级 AI 功能，这些科技巨头都掌握了原生分发渠道。

而 OpenAI，如果继续依赖这些平台，就等于将 ChatGPT 的“用户关系”交给了别人。每一次访问、每一笔订阅、每一个使用数据，都要通过别人定义的路径来完成。这不仅成本高昂，还意味着无法真正实现闭环。

因此，OpenAI 决定自建分发体系。它不是选择开发一个新的 App，而是选择直接切入

硬件底层，打造属于自己的“AI 原生入口设备”。这正是其与 IO 合作，开发所谓“AI 贴身助理”（Companion Device）的根本动因。

这种设备不是传统意义上的智能手机、平板或可穿戴设备，而是一种全新形态的 AI 交互终端：常驻、无屏、可语音交互、具备情境感知能力，贴近用户生活，成为 AI 与人之间的“第一接触点”。

如果说模型是 AI 的“大脑”，那么硬件就是它的“神经末梢”与“感知界面”。真正的 AI 应用，不可能永远停留在服务器或云端，它必须“落地到生活中”，成为用户每天都能触达的存在。

端侧 AI 硬件就是这个“落地”的物理载体。

OpenAI 对此有着清晰的认知。在其发布的官方视频中，奥特曼明确指出，“使用笔电或手机访问 ChatGPT 太繁琐，我们希望打造一种更自然、更贴近生活的设备。”这句话的背后，是对 AI 使用方式的重新定义——从“主动调用”到“被动陪伴”。

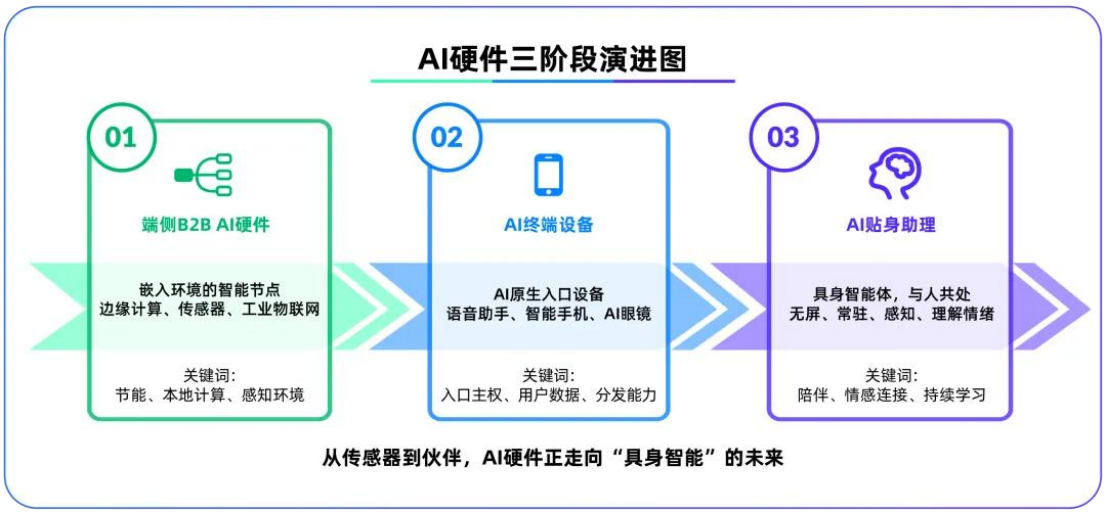

图 4：端侧 AI 硬件演进

（图源：物联网智库）

这种转变意味着，AI 硬件不再是一个“运行模型的容器”，而是“承载用户关系的入口”。它不只是执行任务的终端，更是汇聚用户数据、行为反馈、使用频次、支付习惯、生成内容的生态中枢。

可以说，硬件即入口，入口即分发，分发即生态。

从这个角度看，ChatGPT 已不再只是一个“应用程序”，而是一个正在进化为“AI 操作系统”的平台。而要实现这种平台化，就必须拥有一个专属的物理入口设备，让 AI 能够“无处不在、无感接入”。

更重要的是，这种设备一旦形成用户粘性与习惯，就会成为整个 OpenAI 生态的“控制锚点”——数据从这里产生，模型从这里学习，服务从这里启动，商业从这里闭环。

在模型能力渐趋同质化的时代，AI 公司之间的真正差异，不再是“谁的模型更强”，而是“谁更贴近用户”。端侧 AI 硬件，正成为连接算法与人、链接模型与生态的下一场决战前线。

洞察 4：AI 模组破局产业落地

端侧 AI 产业已处在爆发前夕，AI 模组正在破局 DeepSeek 在实体产业落地“最后一公里”

从生成式 AI 的云端智能到端侧 AI 落地的革命漫长的技术周期里，上下游厂商不断探索着硬件创新、端侧算法模型优化与场景落地的协同。那当 AI 走出云端落到端侧如何才能让终端设备真正“智能”？Deepseek 的横空出世给出了一份答案。

Deepseek 展现的“低成本、高性能、开源”颠覆性优势，直接点亮了终端侧 AI 的发展前景，端侧智能不再完全受限于硬件算力与能效，大模型通过蒸馏技术重构的小模型在端侧部署可行性大增。

从已发布的多个 Deepseek R1 的精简模型来看，在保持性能的前提下，能将模型参数量大幅压缩，这使得端侧模型部署难度显著减小，并突破以往端侧 AI 面临存储空间、算力消耗、推理延迟等部署障碍。知名分析师郭明錤日前也发文指出，Deepseek 爆红后，端侧 AI 趋势将加速。

端侧应用的想象空间的确在 Deepseek 的加持下不断扩大，特别是在今年端侧 AI 元年这个时间节点，AI 模组厂商纷纷布局 Deepseek，帮助下游终端客户搭建本地智能。模组与 Deepseek 的融合，这意味着产业链下游的中小型厂商能够通过模组快速集成 AI 能力推出各自的终端产品。

Deepseek 带动的资本市场热潮褪去后，落地到真正的实体产业带动终端设备升级与市场增长是下一阶段的关键。作为与终端设备关系最紧密的中游模组厂商，将 AI 模组与 Deepseek 的融合，为下游提供更精准、更高效的端侧 AI 产品与服务，为端侧实体产业落地的难题提供了解题思路。

Deepseek 能够无缝地将大模型的推理能力迁移到更小、更高效的端侧版本中，也能更方便将其融合在智能模组中。像移远通信 AI 模组 SG885G 成功实现了 在 DeepSeek-R1 蒸馏小模型端侧运行的基础上，同时完成该模型的针对性微调，提供更精准、更高效的端侧 AI 服务，生成速度超过 40Tokens/s，而且还能优化。

目前已经官宣跑通 Deepseek 的模组，在应用场景覆盖性很广，涵盖智能汽车、机器视觉、PC、机器人、智能家居、AI 玩具及可穿戴设备等多元化场景，多场景应用支持让不同行业不同终端的下游设备厂商能够全面受益于 Deepseek 带来的本地智能，加速终端智能化的发展。

Deepseek 在解决了端侧 AI 硬件碎片化、模型泛化和效能瓶颈上提供了强大助力，模组与 Deepseek 的深度结合更为端侧 AI 落地“最后一公里”难题指出了一条破局之道。这条破局之道指向的最终蓝图，是让端侧 AI 成为终端设备核心功能的定义者，让终端硬件真正智能起来。

洞察 5：端侧智能从创新功能到市场的快速转化

AI 功能以落地商业为首要目标进行迭代，力求从创新功能到市场的快速转化

不论是 C 端的消费者客户还是 B 端的行业客户，在今年对于 AI 的期待都达到了前所未有的高度。如何借助 AI 技术讲好智能时代的新故事成为供应商们的核心命题。对于 AI 的探索和创新变得尤为重要，特别是在终端侧 AI 上，厂商们想利用今年端侧 AI 落地发展周期将用户生态培养起来，以确保能够占得先机。

特别是在今年「人工智能+」行动的政策导向下，政府工作报告中明确指出了要“持续推进「人工智能+」行动，将数字技术与制造优势、市场优势更好结合起来，支持大模型广泛应用，大力发展智能网联新能源汽车、人工智能手机和电脑、智能机器人等新一代智能终端以及智能制造装备”。

提升对 AI 的应用和整合能力，在今年这样的政策驱动和市场需求下，成为厂商面临的关键问题。在这一趋势下，很多 AI 功能不再只是浅尝辄止地嵌入，而是开始深度整合到终端设备的内核中。

从目前推出的应用来看，以微软 Windows 11 AI+ PC 设计为例，这些 AI 功能能加速日常工作、生活、学习，提升效率，如 Recall 回顾、增强搜索、照片超分、实时字幕等功能。目前很多端侧 AI 产品在 AI 功能的布局上主要都聚拢在效率提升类应用上，这是目前市场上比较明确的具体需求。

另一个方向则是提供情绪价值类的端侧设备，在模型技术与配套软硬件的加持下，这些端侧设备从简单的互动设备进化到集教育、陪伴和娱乐功能于一身，除了能给予消费者情绪价值反馈，在实用性上也有了质的飞跃，一经推出就有了较高的市场接受度。

不过由于目前整个行业尚未找到“杀手级潜力的应用”，厂商在这类功能上的军备竞赛难免有些同质化，不同设备在这些功能上的核心体验差异微弱。虽然行业现阶段难免有功能趋同的现象，但这是技术演进的必然，后续随着各厂商不断将端侧 AI 与设备融合得更自然，差异化创新将逐渐显现。

此外，现阶段智能终端开始更注重落地，以落地商业为首要目标进行迭代，找场景找应用找客户需求，力求从创新功能到市场的快速转化。端侧 AI 的落地，离开了场景就永远是一个 demo，没有办法量产，没有办法商业化，只有立足于场景，才能推进端侧 AI 落地生根。

从细分应用出发，立足于解决场景差异化诉求，在细分场景细分赛道中充分挖掘端侧方案的应用价值，端侧 AI 才能最终普及。

二、芯模端智一体化

2.1、芯

随着人工智能技术的飞速发展，端侧 AI 芯片逐渐成为科技领域的焦点。端侧 AI 芯片并非单一的芯片类型，而是一个涵盖了多种芯片的体系，包括端侧 SoC、存储芯片、传感芯片以及智能模组等，它们协同工作，为各类智能终端设备赋予强大的 AI 能力。

端侧 SoC 作为各类型硬件设备的主控单元，堪称硬件的 “大脑”，承载着运算控制等核心功能。在 AI 浪潮的推动下，传统 SOC 正加速向集成人工智能和边缘计算能力的系统级芯片 ——AI SOC 转变，其算力已能达到几十甚至数百 TOPS。存储芯片在端侧 AI 芯片体系中也至关重要，它负责存储 AI 模型、数据以及运行过程中的中间结果等。随着端侧 AI 应用对模型复杂度和数据处理量要求的不断提高，对存储芯片的容量、读写速度和功耗都提出了严苛挑战。传感芯片如同智能终端设备的 “五官”，用于感知外部环境信息，如温度、压力、光线、声音、图像等。在端侧 AI 应用中，传感芯片采集到的数据是 AI 算法进行分析和决策的基础。智能模组则是将多种芯片、元器件以及相关软件集成在一起的模块化产品，它为终端设备提供了便捷的 AI 能力接入方案。

2.1.1 端侧 SoC

定义与概述

端侧 SoC（System - on - Chip，系统级芯片）是为端侧 AI 应用场景专门设计的一种高度集成的芯片。在硬件集成方面它将处理器（如 CPU、GPU 等多种计算单元）、存储器（用于存放指令和数据）、通信接口（用于设备间的数据传输，如 USB、Wi - Fi 等接口）、传感器接口（能够连接各种传感器，如摄像头传感器、麦克风传感器等）以及专门用于 AI 计算的硬件模块，如将 NPU（ Neural Processing Unit 神经网络处理器，专用于运行与神经网络/机器学习/AI 任务相关的数学函数。）等集成到一个单一的芯片上。使得该 SoC 能够在终端设备本地实现 AI 算法的运行。它具备低功耗的特性，同时还具有高能效比，能够在有限的功耗和芯片面积下提供足够的计算能力来满足端侧 AI 应用的需求。

市场规模与概况

SoC 市场规模：根据 MarketResearch 的数据，全球 SoC 市场规模预计将从 2022 年的 1548 亿美元增长至 2032 年的约 3278 亿美元，2022 年-2032 年的年复合增长率（CAGR）为 8%。这种增长主要得益于多个领域对 SoC 需求的增加。特别是在移动设

备和物联网领域，SoC 的应用极大地提升了产品的性能和效率。随着 AI、5G 连接和边缘计算时代的到来，SoC 继续演变以适应不断增长的复杂性和处理要求，为端侧设备提供了更强大的智能处理能力。

图 5：全球 SoC 市场规模

（数据来源：MarketResearch，智次方制图）

主要企业与方案

紫光展锐

紫光展锐（上海）科技股份有限公司，是全球少数全面掌握 2G/3G/4G/5G、Wi-Fi、RedCap、蓝牙、电视调频、卫星通信等全场景通信技术的企业之一。在核心的 5G 领域，紫光展锐是全球公开市场 3 家 5G 手机芯片企业之一。 紫光展锐具备大型芯片集成及套片能力，产品包括移动通信中央处理器，基带芯片，射频前端芯片，射频芯片等各类通信、计算及控制芯片等。

紫光展锐构建了一套覆盖软件、硬件、生态应用的全场景 AI 异构计算系统。在端侧大模型软硬件协同设计、深度学习算法和模型优化、AI 与传感器数据融合等方面持续演进。同时，深耕局域网、广域网、卫星通信技术，积极部署泛在联接，为云端混合 AI 奠定了坚实的通信基础。紫光展锐还持续推进芯片架构革新，将 AI 延伸到每个角落，全面提升终端设备的整体性能和 AI 计算能力。

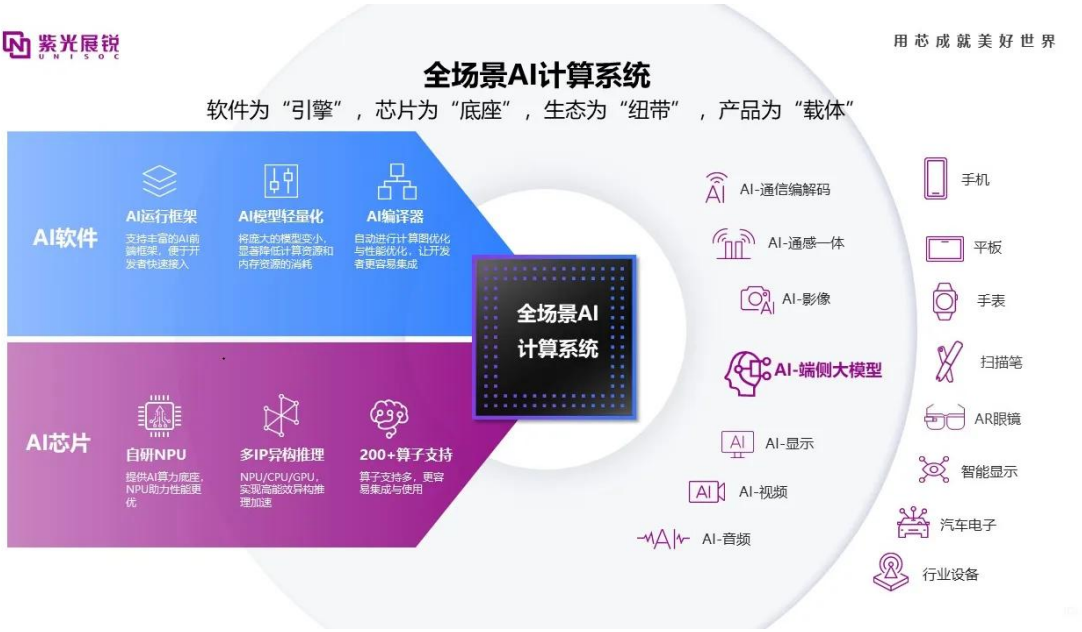

图 6：紫光展锐产品布局

（图源：紫光展锐）

瑞芯微

瑞芯微成立于 2001 年，总部位于福州，经过二十余年的发展成长为领先的物联网（IoT）及人工智能物联网（AIoT）处理器芯片企业，在处理器和数模混合芯片设计、多媒体处理、影像算法、系统软件开发上具有丰富的经验和技术储备。坚持 “IP 芯片化” 发展策略，持续更新迭代 NPU、ISP、高清视频编解码等核心 IP，成功建立起多领域技术性能的优势。作为国产头部端侧算力 SoC 企业，在 AIoT 领域具有较高的市场份额和影响力，其产品广泛应用于智能家居、智能安防、汽车电子、机器人等多个领域。

瑞芯微推出新一代 AI 视觉芯片 RV1126B 凭借 3T 强劲算力、定制化 AI-ISP 架构、动态拼接、防抖、超级编码技术及硬件级安全等性能优势，为智能安防、工业视觉、机器人、智能车载等 AIoT 领域提供高效能解决方案，推动终端设备从 "看得清" 向 "看得懂" 升级。正开启边缘智能终端的全新可能，为 "视觉即 AI 入口" 的产业趋势注入强劲动力。

瑞芯微基于自研芯片平台的系列 AI 技术，完善涵盖机器视觉、大语言模型（LLM）、多模态交互等方向的端侧 AI 技术矩阵，为工业自动化、智能制造及设备智能化提供全栈解决方案，如 RK3588 多路边缘计算方案用于智能视频分析检测、 RV1106B 的低功耗 AOV 方案、RK3588 的离线大模型 LLM 对话方案。

全志科技

全志科技成立于 2007 年，2015 年在深交所创业板上市，是一家在智能应用处理器 SoC、高性能模拟器件和无线互联芯片设计领域表现卓越的高新技术企业。公司产品广泛应用在智能硬件、汽车电子、AIoT、机器人等多个领域，已发展成为国内智能终端芯片的主流供应商，并且具备一系列显著优势。

V 系列聚焦智慧视觉与端边计算，以高性能 SoC 架构和 AI 算力融合为核心，覆盖智能安防、车载电子及工业视觉领域。其中，V821 系列采用 RISC-V 架构，可实现毫秒级别系统冷启动图像及视频抓拍，支持人形检测、人脸检测、人脸识别等端侧图像处理算法，拓展至智能穿戴和 AI 互动玩具场景。

R 系列是端侧 AI 语音交互引擎，专攻智能语音与本地化 AI 推理，已导入扫地机器人及智能家电等场景。MR 系列专为智能机器人设计，支持多核实时控制及 MPC 算法，实现动态平衡与避障功能。

恒玄科技

恒玄科技成立于 2015 年，专注于超低功耗技术、智能音视频交互技术和无线通信连接技术的研发，面向未来智能可穿戴和智能家居市场，打造低功耗无线计算 SoC 芯片。恒玄科技拥有优秀的射频/模拟/电源管理、无线通信、声学/音频、图像/视觉、NPU 技术、超低功耗 SoC，完整软件协议栈和复杂操作系统的综合研发能力。

BES2800 系列，采用 6nm FinFET 工艺，集成多核 CPU、NPU、Wi-Fi 6 及蓝牙功能，支持端侧 AI 算力与低功耗连接，单芯片支持语音交互、主动降噪等功能，该芯片凭借 CPU+ DSP + NPU 三核异构架构、24bit 无损音质和 6.4 TOPS/W 的超高能效比，为消费级与专业级音频 AI 升级提供强大驱动力。

BES2700 系列，采用 12nm FinFET 工艺，单芯片上集成了多核 ARM CPU、音频 DSP、应用于图像图形转换加速的 2.5D GPU、可穿戴低功耗显示系统控制器、神经网络加速的协处理器，在多个一线品牌的手表、手环项目中实现量产，提升了在可穿戴芯片领域的竞争力。

乐鑫科技

乐鑫科技成立于 2008 年，2019 年在上交所科创板首批挂牌上市。公司专注于为客户提供高性能、低功耗、安全可靠的物联网芯片解决方案，其产品以 RISC-V 指令集架构为核心，集成了 Wi-Fi、蓝牙等多种无线通信技术，并支持多种物联网协议，广泛应用于智能家居、工业自动化、智能穿戴设备、汽车电子等多个领域。

ESP32-P4 是乐鑫突破传统涉猎的通信 + 物联网市场，进军多媒体市场的首款非 Wi-Fi 或蓝牙的 SoC。通过增强的计算能力和高效的数据处理性能，ESP32-P4 能够支持多种复杂的 AI 应用，为用户提供更智能、更高效的技术体验。它由乐鑫自研的高性能双核 RISC-V 处理器驱动，拥有 AI 指令扩展、先进的内存子系统，并集成高速外设，充分

满足下一代嵌入式应用对 HMI 支持、边缘计算能力和 IO 连接特性等方面提出的更高需求。

在人工智能 (AI) 和自然语言处理技术 (NLP) 的推动下，乐鑫的大模型 LLM 和 HMI 解决方案为用户提供无缝、直观的交互体验。ESP32-S3 和 ESP32-P4 均支持本地化语音助手和图形化触摸屏交互，结合云端大语言模型，实现更精准的语义理解和上下文对话。

晶晨股份

晶晨股份是一家全球布局的无晶圆半导体系统设计领导企业，成立于 2003 年，专注于系统级 SoC 芯片及周边芯片的研发、设计与销售，产品有多媒体智能终端 SOC 芯片、无线连接芯片、汽车电子芯片等，为众多消费类电子领域提供 SoC 主控芯片和系统级解决方案。

晶晨 T 系列 SoC 芯片是智能显示终端的核心关键部件，最高支持 8K 视频解码，具备超高清解码、动态画面处理、MEMC 运动补偿等先进特性，在智能家居、多媒体终端应用广泛。

A 系列 SoC 芯片代表了晶晨向智能化转型的战略方向，内置神经网络处理器，支持最高 5TOPS 的 AI 算力，具备深度机器学习和高速逻辑推理能力，已广泛应用于智能音箱、智能门铃、AR 终端、健身镜等 20 多个消费电子领域。此外，A 系列芯片在边缘计算领域表现出色，支持毫秒级响应的高动态范围影像输入和超高清编码，在智能安防、工业检测等专业领域也逐渐打开市场。

炬芯科技

炬芯科技成立于 2014 年，是中国领先的低功耗 AIoT 芯片设计厂商，主营业务为中高端智能音频 SoC 芯片的研发、 设计及销售，公司在低功耗前提下提供高音质及低延迟的无线音频体验，深耕高音质音频全信号链技术和低延迟无线连接技术。其产品广泛应用于智能手表、蓝牙音箱、蓝牙耳机、无线家庭影院等多个领域，积累了完备且先进的自主知识产权，致力于成为 AIoT 芯片领域全球领先供应商。

炬芯科技打造出了下一代低功耗大算力、高能效比的端侧 AI 音频芯片平台。共三个芯片系列：第一个系列是 ATS323X，面向低延迟私有无线音频领域；第二个系列是 ATS286X，面向蓝牙 AI 音频领域；第三个系列是 ATS362X，面向 AI DSP 领域。三个系列芯片均采用了 CPU（ARM）+ DSP（HiFi5）+ NPU（MMSCIM）三核异构的设计架构，为低功耗端侧 AI 设备量身打造。

炬芯科技全新一代基于模数混合 SRAM 存内计算技术的端侧 AI 音频芯片系列。ATS362X 作为炬芯科技全新一代搭载 AI-NPU（Actions Intelligence NPU）的三核异构 SoC 芯片，采用 ARM STAR CPU + HIFI5 DSP + MMSCIM NPU 架构设计，理论算力高

达 132 GOPS，支持声纹识别、环境音分类等复杂模型端侧实时推理，为多种端侧音频产品注入 AI 动力，助力终端品牌产品迈进 AI 新时代。

中科蓝讯

中科蓝讯成立于 2016 年，是一家专注于低功耗、高性能无线音频 SoC 芯片研发、设计与销售的公司，其产品广泛应用于 TWS 蓝牙耳机、蓝牙音箱、智能可穿戴设备、无线麦克风、语音玩具、物联网设备等无线互联终端。中科蓝讯的芯片出货量已超 20 亿颗，且已进入众多知名终端品牌供应体系。

讯龙三代 BT895x 芯片采用 22nm 低功耗工艺，集成 RISC-V CPU+HiFi 4 DSP 架构，支持蓝牙 5.3 LE Audio 协议及 LC3 编解码，内置 NPU 算力达 0.8TOPS，可本地运行轻量级 AI 模型（如豆包大模型），支持 4.2mA 超低功耗模式，满足 AI 耳机长续航需求，24 位 ADC（信噪比 103dB）与 DAC（信噪比 106dB），支持 Hi-Res 认证，适配高音质场景，2025 年 1 月通过蓝牙 6.0 双模认证，成为国内首个获此认证的非手机芯片厂商，可应用于 AI 高音质开放式耳机（如 FIIL GS Links）等产品。

安凯微

广州安凯微电子股份有限公司于 2001 年成立，是一家专注于物联网智能硬件核心 SoC 芯片研发、设计、终测和销售的芯片设计公司。公司通过技术创新，为边侧端硬件进行智能化赋能，实现万物智能互联时代的“人机物”三元融合。形成了包括 SoC 技术、ISP 技术和机器学习技术在内的七大核心技术，拥有 60 多类电路设计 IP 核以及多个系统平台 IP，产品广泛应用于智能家居、智慧安防、智慧办公以及工业物联网等多个领域。

孔明 KM02 系列芯片主要应用于人工智能摄像机和门禁考勤机。该芯片内置双核处理器，集成神经网络处理器、图像处理器和 H.265/H.264 视频编码器等智能模块，能够在满足低功耗需求的同时，提供高质量图像处理和高压缩率视频编码能力，可广泛应用于目标检测/跟踪以及人脸检测/识别领域。

孔明 KM01 系列芯片则主要应用于物联网摄像机，该芯片内置 ARM926EJ-S 内核，集成神经网络处理器、图像信号处理器和 H.265/H.264 视频编码器，能够在满足低功耗需求的同时，提供高质量图像处理和高压缩率视频编码能力。

翱捷科技

翱捷科技成立于 2015 年，是一家提供无线通信、超大规模芯片的平台型芯片企业，专注于无线通信芯片研发和技术创新，拥有全制式蜂窝基带芯片及多协议非蜂窝物联网芯片设计与供货能力，还能提供超大规模高速 SoC 芯片定制及半导体 IP 授权服务。

翱捷科技是“云侧”和“端侧”同时布局的芯片设计公司。在云侧，公司为客户定制云端大型 AI 推理芯片、大型 AI 训练芯片等人工智能芯片；在端侧，公司整合已有的自研 ISP 和 AI 终端计算网络技术，启动了首款智能 IPC 芯片项目并已完成工程流片。

润欣科技

润欣科技成立于 2000 年，是国内领先的 IC 产品授权分销商和 IC 应用方案提供商，专注于无线通信 IC、射频 IC 和传感器件的分销、应用设计及技术创新，与高通、安世半导体等全球顶尖半导体供应商合作，在移动互联网、宽带接入等领域为国内主要厂商提供芯片技术服务和应用方案。在智能手机 Wi-Fi 方案、运营商无线网络热点覆盖、三网融合等市场占据较高份额。

润欣科技的智能穿戴 SoC 芯片应用于客户的 AR 眼镜和 AI 眼镜产品中。其最新 AI 算力芯片拥有 256 个算力单元，算力提升 40%，兼顾处理速度与节能设计，能满足智能眼镜运行复杂 AI 算法和高效数据处理的需求。该芯片以 CPU 为核心，集成了 GPU、DSP、ISP 等功能模块，使智能眼镜可实现音频、视频、无线上网等多种功能，同时还能有效管理功耗。

泰凌微

泰凌微电子股份有限公司成立于 2010 年，是一家专注于无线物联网系统级芯片研发、设计及销售的高新技术企业。公司产品广泛应用于智能家居、智能穿戴、智能音频、物联网等领域，是全球范围内产品种类最为齐全的无线物联网芯片供应商之一。

TL721X 国内首颗工作电流低至 1mA 量级的超低功耗多协议物联网无线 SoC 芯片，集成了 240MHz RISC - V 单核心、512KB SRAM 和 2MB flash 内存资源，兼容 Zigbee、Bluetooth LE 等多种通信协议，射频灵敏度在 Zigbee 协议下达到了 - 103dB，在 3V 供电下，BLE 传输和接收功耗低至 2.5mA 和 1.8mA，还支持 Channel Sounding 功能。适用于需要低功耗的端侧 AI 设备中直接进行 AI 运算，可满足新一代高性能智能物联网终端产品对于核心芯片的高标准要求。

星宸科技

星宸科技成立于 2016 年，主营业务为端侧和边缘侧 AI SoC 芯片的研发及销售，其芯片广泛应用于智能安防、智能物联、智能车载等领域，是全球领先的智能视频芯片设计商，在全球 IPC SoC 市场、全球 NVR SoC 市场、全球 USB 视频会议摄像头芯片市场成绩斐然。

SSC309QL 专为智能眼镜打造，采用 chiplet 技术，内置一颗 LPDDR4x，面积减少 24%，成本降低。采用第四代自研图像处理引擎 ISP4.0，具备卓越的影像处理特性，可呈现清晰细腻、色彩逼真的画面。采用软硬结合的低功耗技术架构，全天候录像功耗

仅需 30mW，算力达 1.5T，可进行本地低功耗智能应用。

机器人 AI SoC 芯片具备多核异构、高扩展性软件系统、高效自研 IPU、丰富开源模型、低光环境应对能力、图像矫正及拼接等技术特点及优势，星宸还披露了覆盖端侧轻智能、端侧大模型、边侧智能、边侧大脑的机器人主控芯片 5 年规划路线图，以家用扫地机为入口，打造覆盖各类型机器人产品的 SoC 产品线。

博通集成

博通集成电路（上海）股份有限公司成立于 2004 年，是国内物联网无线连接芯片设计领域内的知名上市企业。公司聚焦智能交通和智能家居应用领域，拥有完整的无线通讯产品平台，支持丰富的无线协议和通讯标准，为国内外众多知名品牌客户提供低功耗、高性能的无线射频收发器和集成微处理器的无线连接系统级（SoC）芯片，以及完整的无线通讯解决方案。

蓝牙音频 SoC 芯片 BK32967 集成高性能蓝牙射频收发器、基带处理器等多个模拟和数字外设以及蓝牙协议栈。内部集成 32 位 RISC MCU，支持双麦降噪算法和自适应波束成形，还集成了电源管理系统，包含电池充电器等。可提供卓越的蓝牙连接以及先进的音频处理能力，能满足 AI 耳机对于蓝牙连接稳定性、音频处理高质量的要求，其双麦降噪算法和自适应波束成形技术有助于提升 AI 耳机在语音交互时的语音采集质量，减少环境噪音干扰，更好地实现语音指令识别等 AI 功能。

BK7252N 与 BK7258 芯片与奥嘟比合作，将火山引擎豆包 AI 大模型融入玩具，推出玩具 AI 智能套件，为传统玩具注入 AI 新活力。

北京君正

北京君正集成电路股份有限公司成立于 2005 年，基于创始团队创新的 CPU 设计技术，迅速在消费电子市场实现 SoC 芯片产业化。君正在处理器技术、多媒体技术和 AI 技术等计算技术领域持续投入，其芯片在智能视频监控、AIoT、工业和消费、生物识别及教育电子领域获得了稳健和广阔的市场。

君正的 T 系列与 C 系列智能视频处理芯片凭借专业的视频处理能力，极快的启动速度，优秀的功耗表现以及极致的封装尺寸，充分赋能 AI 眼镜市场，助力合作伙伴产品快速落地。

T32 IPC SoC 面向 AI+H265 主流市场，为各类前端设备注入智能化动能，使其具备自主思考、独立决策的能力，并能灵活应对复杂作业环境，成为适配全场景的深度智能中枢。

联发科（MediaTek）

联发科技为全球第四大无晶圆半导体公司，所研发的芯片一年驱动超过 20 亿台智能终端设备。在智能电视、语音助理设备（VAD）、安卓平板电脑、功能手机、光学与蓝光 DVD 播放器的芯片技术在市场上具有领先的地位，移动通信芯片则位居世界第二。

天玑 9300+主打 AI 能力，在端侧支持双 LoRA 融合技术，可在一个大模型基础上叠加双倍功能，提升生成式 AI 效率。支持主流生成式 AI 大模型及 AI 框架 ExecuTorch，助力开发者加快端侧 AI 应用开发。

天玑 9400 是天玑第二代全大核 SoC，也是业界首款旗舰 5G 智能体 AI 芯片。搭载联发科全新第八代 AI 处理器 NPU 890，AI 性能和能效显著提升。

高通（Qualcomm）

全球领先的无线科技创新者，变革了世界连接、计算和沟通的方式，其基础科技赋能了整个移动生态系统，其产品广泛应用于智能手机、PC、物联网、汽车等领域。

高通全力推动终端侧 AI，在引领并利用从 AI 训练向大规模推理转型，以及 AI 计算处理从云端向边缘侧扩展方面具有战略优势，在开发定制 CPU、NPU、GPU 和低功耗子系统领域取得了广泛的成就。通过与模型厂商展开合作，以及面向跨不同边缘终端领域的模型部署提供工具、框架和 SDK，高通技术公司赋能开发者在边缘侧加速采用 AI 智能体和应用。

英伟达（NVIDIA）

全球最大的独立图形芯片供应商之一，凭借在 GPU 技术上的持续创新，英伟达的产品广泛应用于游戏、专业图形、数据中心、人工智能、自动驾驶等多个领域，为全球众多知名企业提供芯片及解决方案。

英伟达不仅在端侧 AI 硬件领域不断推出新品，如 RTX 50 系列显卡和 Project DIGITS 等，还在软件端发布了开源世界基础模型 Cosmos，通过全栈技术支持加速端侧 AI 商业化落地。

发展趋势与挑战

1. 算力需求提升：随着生成式 AI、多模态交互等技术的发展，端侧设备对 SoC 芯片的算力要求不断提高。AI 手机、AI PC、智能座舱等领域，需要更强大的 SoC 芯片来支持复杂的 AI 模型运行和实时数据处理，以提供更流畅、更智能的用户体验。

2. 集成化与融合化：SoC 芯片将集成更多的功能模块，如 CPU、GPU、NPU、DSP、ISP 等，实现更强大的异构计算能力，同时降低功耗和成本。此外，端侧 SoC 芯片还将与传感器、存储器等其他元件进一步融合，形成更紧凑、高效的系统级解决方

案，满足设备小型化和高性能的需求。

3. 低功耗设计：为了满足移动设备和物联网设备对长续航的要求，端侧 SoC 芯片将更加注重低功耗设计。通过采用先进的制程工艺、优化电路设计、动态电压频率调节等技术，降低芯片在运行和待机状态下的功耗，延长设备的电池使用时间。

4. 安全性增强：随着设备智能化程度的提高和数据价值的增加，安全问题将变得越来越重要。端侧 SoC 芯片需要具备更强的安全性能，如硬件加密引擎、安全启动、可信执行环境等，以保护设备和用户的数据安全、隐私以及系统的完整性，防止被恶意攻击和篡改。

5. 性能与功耗平衡：在有限的芯片面积和功耗预算下，如何实现更高的算力和性能，同时满足设备的续航要求，是端侧 SoC 芯片设计面临的一大挑战。随着 AI 模型的复杂度不断增加，如何在保证推理精度和速度的前提下，降低模型的功耗和存储需求，也是需要解决的问题。

2.1.2 存储芯片

定义与概述

端侧存储芯片为端侧设备上的人工智能应用而设计的存储芯片。端侧设备如智能手机、智能手表、智能家居设备、智能安防摄像头等，通常具有计算资源和存储资源相对受限的特点，同时要求在实时性、低功耗、成本等方面满足特定需求。端侧存储芯片就是为了在这些设备上高效地存储和处理 AI 相关数据而开发的，具有以下特点：

高带宽：能够在短时间内完成大量数据的读写操作，以支持 AI 模型的快速加载和数据处理，满足高并发的数据访问请求，解决内存墙问题，实现在受限资源下释放端侧大模型性能。

大容量：随着端侧 AI 应用的发展，如大语言模型在智能手机等设备上的应用，需要存储芯片具备更大的容量来存储模型参数、数据以及中间结果等。

低功耗：适应端侧设备通常依靠电池供电的特点，尽可能降低能耗，以延长设备的续航时间。

高可靠性：确保数据的准确存储和读取，尤其在一些对可靠性要求较高的端侧应用场景，如智能汽车、工业控制等领域至关重要。

市场规模与概况

CFM 闪存市场数据显示，2024 年全球存储市场规模达 1670 亿美元，创出历史新高。2025 年，随着端侧 AI 的加速渗透，存储器市场包括 DRAM 和 NAND 预计将实现显著

增长。例如，2025 年 NAND Flash 和 DRAM bit 容量需求较 2024 年分别增长 12% 和 15%。

国内市场，根据中商产业研究发布的《2025-2030 年中国半导体存储器市场调查及发展趋势研究报告》显示，2024 年中国半导体存储器市场规模约为 4267 亿元，2025 年中国半导体存储器市场规模将达 4580 亿元。

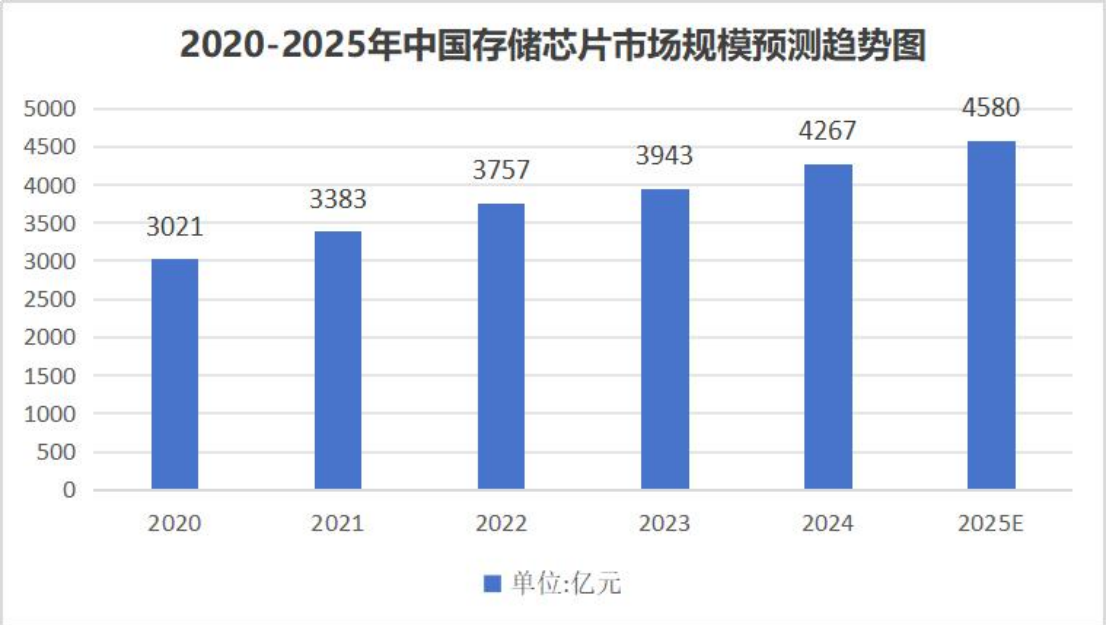

图 7：存储芯片市场规模

（数据来源：中商产业研究院，智次方制图）

AI 端侧应用的普及正在推动存储芯片需求的增加，以手机和 PC 为例，AI 手机中 16GB 的 DRAM 已成为最低配置，支持 70 亿参数大模型的智能手机至少需要 14GB 的内存，AI PC 的内存需求也从 16GB 提升至 32GB。2024 年，手机存储需求同比增长 4%，PC 市场存储需求同比增长 8%，且这一趋势预计将持续，存储大厂美光预计到 2025 年，43% 的 PC 将具备 AI 能力，到 2028 年，这一比例将上升至 64%，未来的 AI PC 将需要比当前 PC 多 80% 的内存容量。此外，智能汽车、智能家居等领域的 AI 应用也在不断拓展，对存储芯片的需求也在逐渐增加。

主要企业与方案

普冉股份

普冉半导体（上海）股份有限公司成立于 2016 年，是低功耗 SPI NOR Flash 存储器芯片和高可靠性 IIC EEPROM 存储器芯片的供应商，公司产品主要包括 NOR Flash、EEPROM、微控制器芯片及模拟产品四大类别，广泛应用于物联网、智能手机及周

边、可穿戴、服务器、光模块、工业控制、汽车电子、安防等领域。

普冉存储产品涵盖 NOR Flash 和 EEPROM 两大类别，NOR Flash 产品覆盖 512Kbit 到 512Mbit 容量，适用于低功耗蓝牙模块、TWS 蓝牙耳机、车载导航等领域；EEPROM 产品覆盖 2Kbit 到 4Mbit 容量，广泛应用于摄像头模组、汽车电子等领域。另外普冉股份是国内率先采用 SONOS 工艺设计 NOR Flash 的公司，该工艺使产品在中小容量市场具有性价比、体积、功耗和读写速度等优势，随着端侧 AI 市场的快速发展，普冉股份有望凭借其技术优势和市场布局，进一步拓展业务，提升市场份额。

兆易创新

兆易创新科技集团股份有限公司成立于 2005 年，是全球领先的 Fabless 芯片供应商，致力于各类存储器、控制器及周边产品的设计研发，是全球排名第一的无晶圆厂 Flash 供应商，在 SPI NOR Flash 领域市场占有率全球第二，也是中国品牌排名第一的 Arm 通用型 MCU 供应商。其核心产品线为存储器（Flash、利基型 DRAM）、32 位通用型 MCU、智能人机交互传感器、模拟产品及整体解决方案。

面对端侧 AI 带来的高带宽需求，兆易创新基于 3D 堆叠架构的紧凑型存储解决方案利用混合键合技术实现了 SoC 与 DRAM 之间的紧密集成。这种设计不仅能够显著降低整体尺寸，还能有效提升数据传输速率至 32-256GB/s，接近 HBM2e 标准。更重要的是，相较于传统的 HBM 技术，兆易创新的紧凑堆叠方案大幅降低了功耗，仅为前者的三分之一到四分之一，这对于依赖电池供电的移动设备尤为重要，如 VR 眼镜、智能手机及机器人等。

其在 TWS 真无线立体声耳机市场表现突出，提供的 256MB NOR 闪存获得国际客户广泛认可，积累了丰富的中大容量 NOR 产品开发经验。随着 AI 功能融入耳机和其他便携式设备，NOR 闪存容量需求预计将进一步扩大。同时兆易创新也在积极推进 NAND 产品的制程升级，以提高产品性能和可靠性更好地服务于新型端侧 AI 设备，如为 AI 眼镜提供的 EMCP 解决方案集成了 2GB DRAM 和 32GB NAND。

长江存储

长江存储科技有限责任公司成立于 2016 年，是一家专注于 3D NAND 闪存设计制造一体化的 IDM 集成电路企业，同时也提供完整的存储器解决方案。它是全球首家量产 232 层 3D NAND 闪存的厂商，通过自主研发和国际合作相结合的方式，成功设计制造了中国首款 3D NAND 闪存。截至 2025 年 3 月，长江存储在 NAND 市场的份额已超过 5%。

基于晶栈 Xtacking 架构，长存推出了 Xtacking2.0 第三代系列产品，能充分利用架构优势进一步提升闪存吞吐速率、提升系统级存储的综合性能并实现定制化闪存，为智能手机、可穿戴设备等一系列智能终端产品带来敏捷的响应速度和畅快的使用体验。

长鑫存储

长鑫存储成立于 2016 年，是一家一体化存储器制造公司，专注于动态随机存取存储芯片（DRAM）的设计、研发、生产和销售，是国内领先的 DRAM 存储芯片供应商。其产品已应用于小米、传音等国内主流手机厂商的品牌机型，同时也在积极拓展其他领域的客户合作。

长鑫存储自主研发的 DDR4 内存芯片，在数据传输速率、稳定性和能耗上表现优异，可应用于 PC、笔记本电脑、服务器、消费电子类产品等领域。LPDDR4X 内存芯片能提供高效能与低功耗的解决方案，适用于对功耗要求较高的智能终端等设备。第五代超低功耗双倍速率动态随机存储器 LPDDR5 已在国内主流手机机型上完成验证，能够为移动端电子设备带来更快的速度体验和更低的功耗消耗。

华邦电子

华邦电子成立于 1987 年 9 月，是全球知名的存储器解决方案供应商，专注于设计和制造各类存储器产品，包括 DRAM、NAND 闪存及 NOR 闪存，其产品应用于手持装置应用、消费电子及计算机周边市场，也布局于车用和工业用电子等高门槛且高质量要求的领域。

华邦电子在存储芯片领域拥有丰富的产品组合，能够满足端侧 AI 设备的多样化需求。如 CUBE 具备高带宽和低功耗特性，适合智能眼镜等穿戴式产品；LPDDR4、HYPERRAM 和 1.2V NOR Flash 等产品也能够满足智能眼镜等设备对存储的高要求。这些产品不仅具备高性能，还具备小尺寸和低功耗的特点，能够为智能设备提供理想的存储解决方案。此外，华邦电子正在持续推动 DRAM 技术向 20nm 乃至 16nm 发展，NAND Flash 和 NOR Flash 分别向 24nm 和 45nm 工艺演进。

东芯半导体

东芯半导体股份有限公司成立于 2014 年，拥有独立自主的知识产权，聚焦于中小容量 NAND/NOR/DRAM 芯片的研发、设计和销售，是国内少数可以同时提供 NAND/NOR/DRAM 设计工艺和产品方案的存储芯片研发设计公司之一。

终端设备的智能化和互联化趋势推动了存储芯片需求的增长。作为聚焦中小容量存储芯片独立研发、设计与销售的企业，东芯在 SLC NAND 方面，基于 2xnm 制程上持续开发新产品，不断扩充 SLC NAND Flash 产品线及料号，先进制程的 1xnm SLC NAND Flash 产品已进入风险量产。关于 NOR Flash，东芯一方面在 48nm 制程上持续进行更高容量的新产品开发，另一方面在持续完善 55nm 的 NOR Flash 产品线，为客户提供中高容量、高可靠性的 NOR Flash 产品。DRAM 产品方面，东芯也在不断丰富 DRAM 自研产品组合。此外，其 DDR+NAND 合封 MCP 芯片是 AI 玩具不可缺少的必要组件。

佰维存储

佰维存储 2010 年成立，专注于半导体存储器和先进封测制造领域，集存储芯片设计研发、封测制造、产品销售、品牌运营为一体，以“从芯到端，赋能万物互联，构筑万物互联时代的存储根基”为使命，旗下产品广泛应用于智能手机、PC、智能穿戴、物联网等领域。

佰维创新推出 Mini SSD 存储方案，以小型化、模块化、高性能的设计理念，针对传统 SSD 存储方案进行革新，在保持小型化设计的同时，顺序读/写速度分别高达 3700MB/s 和 3400MB/s，远超当前常规存储卡方案，并媲美主流消费级 M.2 SSD 性能。这使得 AI 模型加载、4K/8K 高清视频编辑、大型设计软件运行等高负载场景流畅无阻，充分满足 AI 时代端侧终端对于存储性能的严苛要求。

江波龙

江波龙是一家专注于半导体存储产品和应用的企业，成立于 1999 年，在存储芯片设计、封装测试、产品研发等方面拥有丰富的技术积累和经验，业务覆盖嵌入式存储、固态硬盘、移动存储和内存条四大产品线，产品广泛应用于主流消费类智能终端、工业、汽车等多个领域。

基于自研 WM7400 主控，江波龙推出 UFS 4.1，采用国际先进 Foundry 工艺，并融合了 3rd Gen. Prime LDPC 等多项高可靠特性，可同时支持 TLC 和 QLC NAND Flash，顺序读取速度高达 4350MB/s，随机读写速度高达 750K / 630K IOPS，容量最高可达 2TB。其“满血”性能领先业界同类型产品，助力端侧 AI 产品（如 AI 手机、AI 平板、智能汽车、人形机器人等）在复杂应用中实现实时决策，畅享流畅体验。

发展趋势与挑战

目前各类智能终端中应用的存储芯片主要有如下几种:

1. DRAM，动态随机存取存储器，是最常见的系统内存类型，用于存储临时数据和正在处理的数据，是最大规模的单品市场。在 AI 智能手机、AI PC 等智能终端中，DRAM 用于提供快速的数据访问和处理能力，前文已提到 AI 终端已经开始换上更大容量更高读写速度的 DRAM，随着 AI 应用的发展，对高性能 DRAM 的需求仍在不断增长。

2. NAND Flash，NAND Flash 闪存是一种非易失性存储器，用于长期数据存储。在智能终端如 AI 手机、AI PC、AI Pad 中，NAND Flash 用作内部存储，存储操作系统、应用程序和用户数据。未来端侧设备操作系统肯定会搭载大模型以及其他 AI 应用程序，端侧设备会需要更高的 NAND Flash 容量用于长期存储。

3. NOR Flash，NOR Flash 也是一种非易失性存储器，常用于存储启动代码和固件。它在需要快速启动和执行的设备中尤为重要，AI TWS 耳机就是典型的应用。市场上各类 AI 硬件对 NOR Flash 容量需求增加已经是确定性的趋势，尤其是中大容量 NOR Flash。

4. UFS 高性能的通用闪存存储方案，在高端移动设备中应用。UFS4.0 版本，其最高读取速度已经达到了 4200MB/s，写入速度也能达到 2800MB/s。端侧 AI 性能的发挥离不开高效的数据流转，UFS 4.0 存储技术正好能够为之提供支持。随着端侧 AI 软件层面的持续优化以及 UFS 技术向更多端侧设备普及，端侧 AI 设备将拥有更强大的学习能力和更快的响应速度。

5. LPDDR，一种低功耗版本的 DRAM，即低功耗双倍数据速率，是为移动设备设计以减少能耗的方案。LPDDR5 和 LPDDR5x 提供了更高的数据传输速率和更低的功耗，LPDDR6 也敲定在即，将全面适配 AI 计算需求，用更高频率更高带宽支持端侧 AI 设备。

6. HBM 本质上是一种高性能的 3D 堆叠 DRAM，拥有极高的带宽和存储密度，适用于需要处理大量数据的 AI 和高性能计算应用，主要用于服务器和数据中心等高性能计算领域。在移动端侧设备领域，头部厂商在做端侧 HBM 产品研发，初步估计 2026 年能实现商业化。

2.1.3 智能传感芯片/传感器

定义与概述

端侧智能传感芯片/传感器是应用在终端设备中，融合 AI 功能的传感芯片/传感器，具备数据采集、预处理及初步分析能力，能支持本地化智能决策，适用于各种端侧感知应用场景，使得终端设备具备自主感知、理解和决策的能力。

市场规模与概况

智能传感市场规模：根据中商产业研究院发布的《2024-2029 年全球及中国智能传感器市场调查与行业前景预测专题研究报告》显示，2023 年全球智能传感器市场规模达到约 468.9 亿美元，2019-2023 年的年均复合增长率达 10.01%，2024 年全球智能传感器市场规模预计在 520.4 亿美元。

图 8：智能传感器市场规模

（数据来源：中商产业研究院，智次方制图）

主要企业与方案

豪威集团

豪威集团是一家全球知名的 Fabless 芯片设计公司，专注于提供传感器、模拟和触控显示等解决方案，是中国最大的传感器企业、全球第三大 CMOS 图像传感器公司，其产品广泛应用于智能手机、汽车电子、安防、物联网等领域。

2 亿像素的 OVB0B、OVB0A 传感器，5000 万像素高端传感器 OV50H 等，已应用于高端 AI 手机。CIS 产品覆盖了 ADAS、驾驶室内部监控、电子后视镜、仪表盘摄像头、后视和全景影像等广泛的智能汽车应用。此外，韦尔 OG09A10 机器视觉传感器，可用于工业机器人、智能交通；OV01D1R 智能 CMOS 图像传感器，可用于 AI 眼镜、VR/AR 设备。

汇顶科技

汇顶科技是一家基于芯片设计和软件开发的整体应用解决方案提供商，主要面向智能终端、物联网及汽车电子领域提供领先的半导体软硬件解决方案。公司成立于 2002 年，围绕传感、计算、连接和安全技术领域，不断开发创新的产品和解决方案，目前已构建传感器、触控、连接、音频及安全五大产品矩阵，为客户提供具有差异化价值的创新产品和解决方案。

旗下超声波指纹传感器，2024 年推出，基于 CMOS Sensor 架构，采用晶圆级声学层加工，拥有更高信噪比，指纹模组厚度 0.17mm，在众多品牌智能手机上大规模商

用。

汇顶科技的 ToF 测距方案，如 GVW8366B、GD8573B，支持低功耗 Spot-ToF、高分辨率 Flood-ToF 和宽 FOV 的 Line-ToF 三种方案，适用于智能终端、智能家居和工业等场景。

欧菲光

欧菲光集团股份有限公司成立于 2002 年，深耕光学光电领域二十余年，拥有智能手机、智能汽车等三大业务体系，为客户提供一站式光学光电产品技术服务。其主营业务包括光学摄像头模组、光学镜头、指纹识别模组、3D ToF、智能驾驶、智能座舱、车身电子和智能门锁等相关产品的设计、研发、生产和销售。

欧菲光多功能、高精度的微型光电传感摄像模组，主要由图像传感器、镜头、音圈马达、柔性电路板、连接器等构成，广泛应用于手机、笔记本电脑、平板电脑、智能家居、智能穿戴、车载、VR/AR 等领域。

欧菲光率先实现结构光 3D Sensing 模组和 3D ToF 模组的量产，广泛应用于手机、车载、机器人、AR/VR、IoT 等领域，为手机 Face ID、车载 DMS、机器人 SLAM 避障等提供解决方案，2024 年推出双光源 ToF 专利架构方案，实现“避障+定位导航技术”，具备 10 米探测距离、低功耗、高采样率、低算力等优势。

睿创微纳

烟台睿创微纳技术股份有限公司成立于 2009 年，专注于非制冷红外热成像与 MEMS 传感技术开发，致力于专用集成电路、MEMS 传感器及红外成像产品的设计与制造。其产品广泛应用于特种装备、安防监控、工业测温、人体体温筛查、汽车辅助驾驶、户外运动、消费电子、森林防火、医疗检测设备、消防、物联网等领域。

睿创微纳在非制冷红外探测器领域技术国内领先、国际先进。2024 年开发出世界首款 6μm 640×512 非制冷红外探测器产品，适应消费电子、车载产品小型化趋势，满足低成本需求。旗下红外热像仪整机产品，如百万像素级红外体温筛查热像仪 AT1280，可用于智慧医疗、智能安防等场景。

此外，睿创微纳子公司 Raythink 燧石技术推出的热成像 AI 智能体 “小睿”，将红外热成像仪和 AI 智能语音助手相结合，使传统的精密检测仪器变成 “会思考”“能听懂需求” 的智能助手，通过语音指令可实现复杂的参数设置和检测场景的智能诊断。

格科微

格科微电子（上海）有限公司 2003 年成立，是中国领先、国际知名的半导体和集成电路设计企业，主营业务为 CMOS 图像传感器和显示驱动芯片的研发、设计、封测和销

售。其产品主要应用于手机领域，同时广泛应用于平板电脑、笔记本电脑、可穿戴设备、移动支付、汽车电子等消费电子和工业应用领域。

GC32E1 单芯片 0.7μm 3200 万像素图像传感器，采用格科微最新 FPPI 专利技术的 GalaxyCell 0.7μm 工艺，配合 4Cell Bayer 架构可实现等效 1.4μm 像素性能，为高端智能手机前摄需求提供成熟的高像素解决方案。GC32E2 第二代单芯片 3200 万像素图像传感器，适用于智能手机等移动终端。

此外，非手机应用图像传感器为物联网摄像机、汽车电子、笔记本电脑、智慧电视、移动支付、人脸识别等非手机应用领域，提供从 VGA、HD、FHD 以及 400 万到 800 万像素不同规格的 CIS，满足各种端侧传感场景下的图像采集需求。

思特威

思特威（上海）电子科技有限公司于 2011 年成立，专注于 CMOS 图像传感器芯片的设计和研发，产品广泛应用于安防监控、机器视觉、车载电子、智能手机、移动支付、医疗影像等市场。在安防 CIS 领域，思特威出货量连续八年蝉联全球第一。

SC485SL，思特威针对安防智能化发展趋势，推动智能安防应用进一步迭代升级的 4MP 图像传感器，基于思特威 SmartClarity-3 工艺技术打造，搭载了 Lightbox IR、SmartAOV2.0 等多项先进技术，具备高感度、高动态范围、低噪声、高温成像稳定与超低功耗等优势性能，并支持全时录像(AOV)功能，助力智能安防设备实现迭代升级。

士兰微

杭州士兰微电子股份有限公司成立于 1997 年，是专业从事集成电路芯片设计以及半导体微电子相关产品生产的高新技术企业。士兰微的技术与产品涵盖消费类产品的众多领域，在多个技术领域保持国内领先地位，如绿色电源芯片技术、MEMS 传感器技术、LED 照明和屏显技术等。公司目前的产品和研发投入主要集中在功率半导体 & 半导体化合物器件、功率驱动与控制系统、MEMS 传感器、ASIC 产品、光电产品五个领域。

旗下 MEMS 传感器产品包括消费级和车载级应用的三轴加速度计、六轴 IMU 单元、骨传导加速度计、碰撞传感器、震动检测传感器、心率传感器、血氧传感器、ALS/RGB/PS 传感器、麦克风、温湿度传感器、电流传感器、MEMS 微镜传感器等。其中 SC7I22 是一款集成了 3 轴加速度计和 3 轴陀螺仪的六轴运动跟踪传感器，采用先进 MEMS 技术，可提供精确运动检测和方向定位，具备低功耗设计，适合长时间运行且对电池寿命有要求的 AR/VR 头显设备。

纳芯微

纳芯微成立于 2013 年，是一家专注于高性能、高可靠性模拟及混合信号芯片设计的公司，2013 年创立，聚焦汽车电子、泛能源和消费电子应用领域。2024 年，完成对麦歌恩的战略收购和深度整合，提升了在磁传感器领域的市场占有率。目前已量产的端侧传感产品包括磁传感器、压力传感器及调理芯片和温湿度传感器等。

旗下磁传感器基于霍尔效应/AMR/TMR 技术，具备高精度如 NSM301X 系列，可用于汽车电机控制、工业编码器、人形机器人关节控制等。

温湿度传感器通过优化集成式设计，降低功耗并提升响应速度，可以满足消费电子如智能手表、智能手环、智能家居及工业自动化对微型化、低功耗传感器的需求。

敏芯股份

苏州敏芯微电子技术股份有限公司成立于 2007 年，是国内领先的 MEMS 传感器平台型企业，掌握多品类 MEMS 芯片设计和制造工艺能力，产品广泛应用于消费电子、汽车电子、医疗、工业控制等领域。

敏芯股份是全球 MEMS 声学传感器芯片出货量头部厂商，其推出的 70dB 高信噪比 MEMS 麦克风，可积极打造 AI 声觉革命的 “核心引擎”，广泛应用于消费电子、电动汽车、智能家居等领域。

敏芯股份 MEMS 压力传感器包括多种类型，如绝压传感器、差压传感器等，适用于手机、无人机、穿戴设备、手表 / 手环、运动手表等众多终端智能产品。

速腾聚创

速腾聚创（RoboSense）是一家领先的智能激光雷达系统科技企业，成立于 2014 年，专注于激光雷达的研发、生产、销售及解决方案的提供，致力于为自动驾驶、智能驾驶辅助、机器人、车联网等行业提供高性能、高可靠的激光雷达产品及服务。

AC1 作为速腾聚创机器视觉全新品类 Active Camera 的首款产品，通过主动结构光投影 +内置 AI +端到端算法优化，构建了全新的视觉体系，为行业提供了颠覆性的视觉感知开发一站式解决方案。同时，速腾聚创正在通过 AC1 与 AI-Ready 生态系统与全球开发者共同构建智能视觉感知新范式。向“感知智能体”跨越。

镭神智能

深圳市镭神智能系统有限公司成立于 2015 年 2 月，是全球领先的全场景激光雷达与智能搬运机器人系统解决方案提供商，构建了七大激光雷达产品平台 + 三大算法 + 两大控制器 + N 个全场景解决方案的业务生态，服务覆盖自动驾驶、智慧交通、机器人、智慧物流、高端安防、港口、测绘及工业自动化等十大产业生态圈。

CH128X1 128 线车规级混合固态激光雷达采用转镜式混合固态扫描方案，减少了雷达

电机的功耗，使用寿命达 10 万小时。具有高精度高清晰的 3D 点云效果、远程探测感知性能以及小巧的外观设计等优势，已实现东风悦享汽车的应用落地，同时也成为 Intel 路侧感知和路侧边缘计算解决方案的传感器硬件。

禾赛科技

禾赛科技成立于 2014 年，是全球领先的激光雷达制造商，专注于激光雷达、气体监测及机器人应用产品的研发、生产、销售，面向全球自动驾驶、车路协同、机器人、智慧工地、智慧港口、工业气体监测等领域，提供开放、定制的激光雷达解决方案。

AT1440 车规级超高清激光雷达，搭载禾赛第四代自研芯片，内置双核 CPU、8 核 APU，测距达 300 米 @10% 反射率。采用前沿的高效感光和超高并行处理技术，激光雷达线数高达 1440 线，相比市场上同类产品提升 10 倍，点频超过 3400 万点每秒，是当前主流车载激光雷达的 45 倍以上。

发展趋势与挑战

1. 多模态融合：随着端侧 AI 渗透率提升，单一传感器已无法满足复杂场景需求，多模态融合成为必然趋势。端侧智能传感芯片将支持多模态数据（如图像、语音等）的融合处理，模拟人类的多种感知方式，从而拓展更为丰富的应用场景。

2. 集成化提高：未来端侧智能传感芯片将集成更多的功能模块，如 CPU、GPU、NPU、DSP、ISP 等，实现更强大的异构计算能力，AI 功能和信号处理功能也将更多地被引入端侧传感器，增强数据收集后的直接处理能力，分担主控信号处理负载。

3. 微型化与低功耗设计：传感器将向微型化、低功耗方向发展，满足设备小型化和高性能的要求，可应用于更广泛的场景，如可穿戴设备、物联网设备等。同时通过采用先进的制程工艺、优化电路设计、动态电压频率调节等技术，降低芯片在运行和待机状态下的功耗，延长设备的电池使用时间，推动推动端侧传感向更高性能、更低功耗方向发展

2.1.4 端侧 AI 模组

定义与概述

通信模组是物联时代信息链接的核心组件，负责将终端设备（如传感器、车载设备、智能家居等）采集数据通过无线网络传输。到了 AI 时代，端侧设备对模组提出了更高要求，需要其在满足数据传输的同时，还要能够进行本地计算，具备 AI 推理算力。端侧 AI 模组是一种集成了人工智能计算能力的硬件模块，专门设计用于在终端设备（如

智能手机、摄像头、工业设备等）上直接运行 AI 算法，实现本地化实时数据处理、智能决策和低延迟响应。其核心特点是将算力下沉至边端侧，弥补云端 AI 在隐私保护、实时性和能效方面的不足，成为推动人工智能下沉普及的关键载体。

市场规模与概况

根据 INSIGHT AND INFO 发布的《中国物联网模组行业发展深度研究与投资前景预测报告（2025-2032 年）》，人工智能模组配置用于 AI 推理的 NPU、TPU、PPU 或其他专用并行处理芯片组，随着无线智能支付、可穿戴音视频设备、车载后装设备、智慧家居等终端发展，AI 与模组的结合速度加快，AI 模组占比将由 2023 年的 2%提升至 2027 年的 9%，年复合增长率达 73%。传统模组向智能/AI 模组的产品升级预计将成为又一推动模组产业市场扩容的重要因素，除去产品本身价值量提升，新型模组对下游行业的广泛适用性或将一定程度提升产品使用的覆盖范围，进而从“量级”层面带来突破。

图 9：物联模组占比预测

（数据来源：NSIGHT AND INFO，智次方制图）

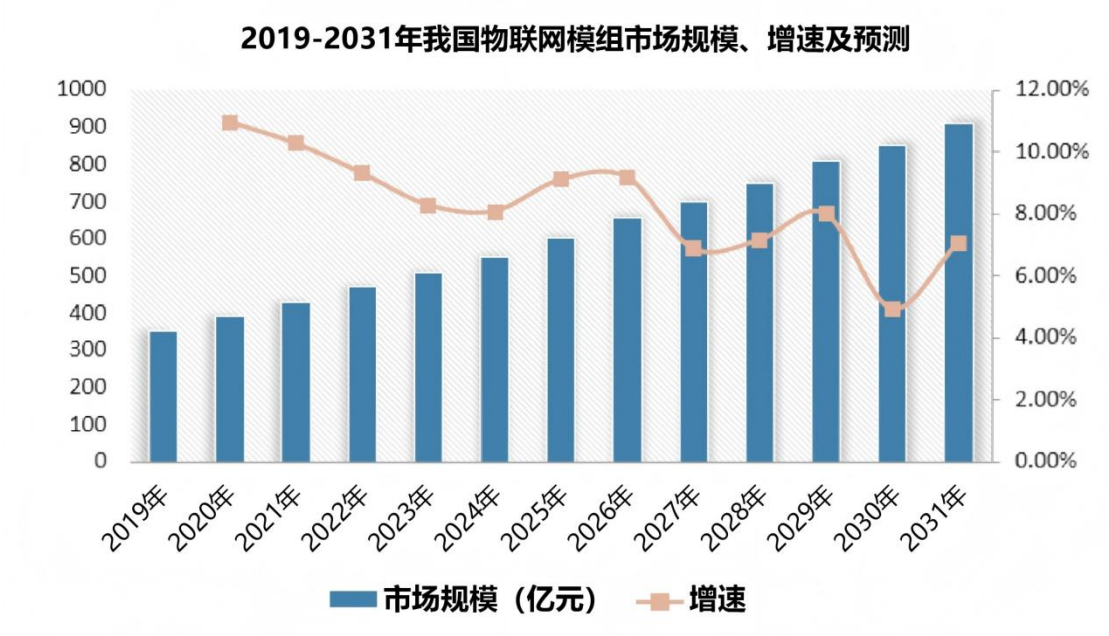

图 10：物联模组市场

（数据来源：NSIGHT AND INFO，智次方制图）

根据最新的机构预测，AI 通信模组在未来 3 年（2024-2027 年）预计 **AI 算力模组的出货量将以 76%的年复合增长率持续增长。**

主要企业与方案

比邻智联

比邻智联作为中国移动首批专精特新重点培育团队，是中国移动模组专业公司，致力于为构建智能互联的全球物联网生态作出卓越贡献。比邻智联已打造“通用+新型+行业模组”完善的产品体系，涵盖 NB-IoT、4G、5G、车载、AI、卫星模组及方案板产品，能够满足各类物联网细分场景的需求。当前，比邻智联已在能源表计、金融支付、定位追踪、共享经济等行业实现千万级销量，蜂窝模组市场份额稳居全球第二。

在 AI 模组方面，比邻智联已推出 MS351A、MS372Q、MS373Q 等系列产品，覆盖 1T～48T 多种算力配置，支持 AI 算法和模型部署，确保实时、安全、可靠的端侧 AI 推理应用，并提供深度定制化开发服务以满足多样化的应用场景需求。同时，比邻智联也针对工业质检等细分场景打造了一体化解决方案，提供“AI+网络”算网融合服务，帮助客户加速产品 AI 升级。当前，比邻智联正持续丰富高低算力搭配的端侧 AI 模组系列，扩展更多场景的应用方案，进一步推动 AI 能力与 IoT 场景的深度融合。

图 11：比邻智联 AI 模组

（图源：比邻智联）

移远通信

上海移远通信技术股份有限公司成立于 2010 年，是全球领先的物联网整体解决方案供应商，深耕物联网行业十余年，公司拥有完备的 IoT 产品和服务，涵盖模组、天线等硬件产品，以及丰富的服务和解决方案。移远通信持续将新一代通信技术、先进的计算能力、机器视觉、精准定位以及高效的天线技术融入模组与 IoT 解决方案，深度赋能千行百业，不断激发产业创新的无限潜能，推动社会迈向更加绿色、高效、智能的未来。

移远通信的智能模组产品矩阵十分丰富，可满足低、中、高不同算力需求。对于在计算能效上要求严苛的应用，移远通信能提供 12 TOPS 算力的 SG560D，满足需要兼顾算力、成本和功耗的端侧应用；主推的搭载高通 QCS8550 平台的高性能 AI 算力模组 SG885G，成功实现了 DeepSeek-R1 蒸馏模型的稳定运行，生成 Tokens 的速度超过每秒 40 个，且随着性能的不断优化，速度还在进一步提升。据悉，其 100 TOPS 左右算力的产品也即将面世。

算力硬件构筑起基础底座，在模型层面，移远通信端侧 AI 大模型解决方案以“LLM（大语言模型）+ RAG（检索增强生成）+ Agent（智能体）”技术三角为核心，通过对 AI 模型的深度优化与增强，重新定义了 AI 端侧设备的智能化逻辑。小到 PC、玩具、可穿戴设备，大到家电、智能汽车、具身智能机器人，移远通信凭借全栈 AI 研发实力，针对不同终端应用持续优化其 AI 方案，解决了多层次的软硬件 AI 场景需求，助力终端设备智能化功能落地。

图 12：移远通信 AI 模组

（图源：移远通信）

广和通

广和通（Fibocom）成立于 1999 年，是一家全球领先的物联网与移动互联网无线通信解决方案供应商。公司专注于无线通信模块及其应用行业的通信解决方案的设计、研发与销售服务，产品广泛应用于智能表计、车联网、工业物联网、智能零售、远程医疗等领域。广和通的端侧模组方案具有高性能、低功耗、高隐私保护等特点，适用于多种物联网设备和场景。其优势在于提供从高到低算力的全面模组选择，支持多种主流 AI 模型，并通过自研技术平台简化开发流程，助力客户快速部署智能解决方案。广和通的市场表现强劲，与众多知名企业合作，推动了端侧 AI 技术的广泛应用。

广和通以「AI For X」重塑千行百业，在产品设计上，通过 OpenCPU 架构将无线通信模组及解决方案升级为 “主控 + 连接 + 算力” 三合一平台，替代传统 “MCU + 模组” 的冗余设计。在软件方面，基于多操作系统和多芯片平台，模组及方案可支持 Fibocom AI Stack，实现从模型云端连接到端侧部署推理的全流程闭环。广和通积极推动 DeepSeek、ChatGPT 等优质模型在高、中、低算力 AI 模组及解决方案部署，提供不同参数模型服务，降低端侧 AI 门槛并优化成本。在端侧模组与模型的结合上，广和通 “星云”系列给出了很多详细方案，包含 1T ~ 100T 多种算力配置。

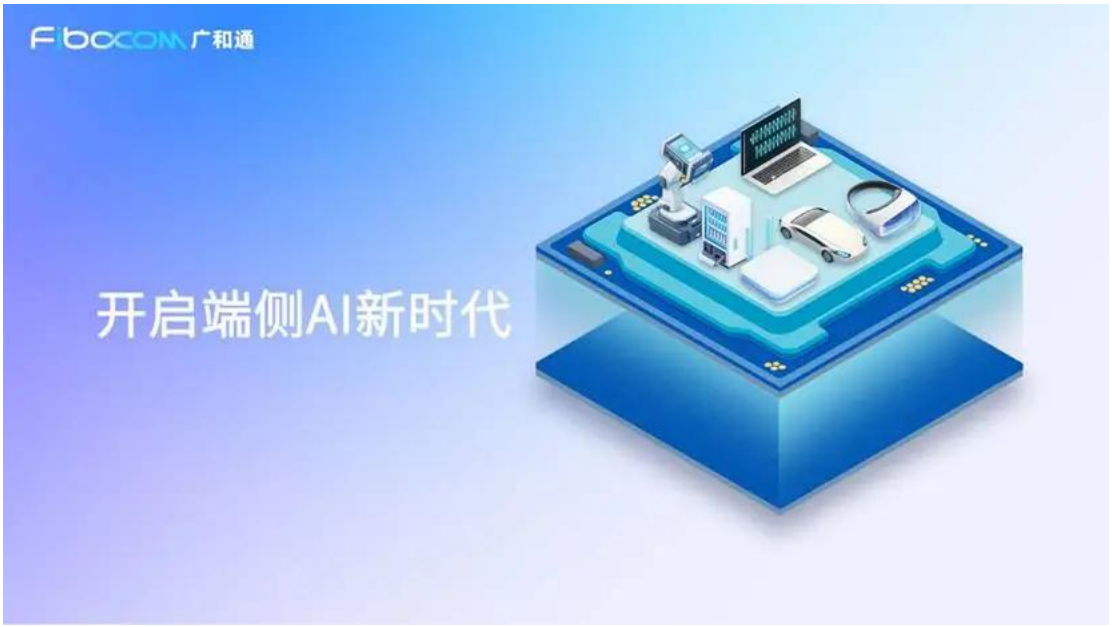

图 13：广和通端侧 AI 战略

（图源：广和通）

广和通 Fibocom AI Stack 集成了高性能模组、AI 工具链、高效推理引擎及海量 AI 模型，适用于多种智能终端的快速 AI 部署。其技术优势明显，支持 TensorFlow、PyTorch、ONNX 等主流机器学习框架的模型压缩与转换，适配不同算力等级的芯片平台，还提供完整的 AI 工具链，涵盖数据标注、模型训练和微调等功能。此外，强大的推理引擎是核心组成部分，支持多种编程语言接口，具备异构调度和硬件加速能力，能提升推理速度，适应高负载实时应用场景。

美格智能

美格智能技术股份有限公司成立于 2007 年，作为全球领先的无线通信模组及解决方案提供商，以新一代的 4G/5G 无线通信技术为基础，以万物互联的物联网行业为依托，美格智能专注于为全球客户提供以 MeiGLink 品牌为核心的标准 M2M/智能安卓无线通信模组、物联网解决方案、技术开发服务及云平台系统化解决方案，物联网行业客户已经遍及全球 100 多个国家和地区，相关产品和服务已在众多物联网核心应用领域处于领先地位。

美格智能基于骁龙 8 至尊版移动平台的高算力 AI 模组 SNM980，拥有出色的 AI 性能和多媒体能力，为广泛客户提供跨时代的超强算力，积极推动以 DeepSeek、Qwen 等优质大模型产品落地至车载、人形机器人、手持移动、AI 医疗、工业制造、无人机、AI/AR 眼镜等场景和终端产品，以创新技术持续打造领先的端侧 AI 应用生态。

2025 年初，美格智能发布 AI 智能体产品 AIMO，凭借丰富的研发经验和技术积累，美格智能快速完成了高通跃龙 QCS8550 平台在 AIMO 终端的适配，上线 AIMO Pro 版

本，并搭载 7B/14B（即 70 亿/140 亿）参数 DeepSeek，带来高速小巧的边缘计算和个人智能体体验。

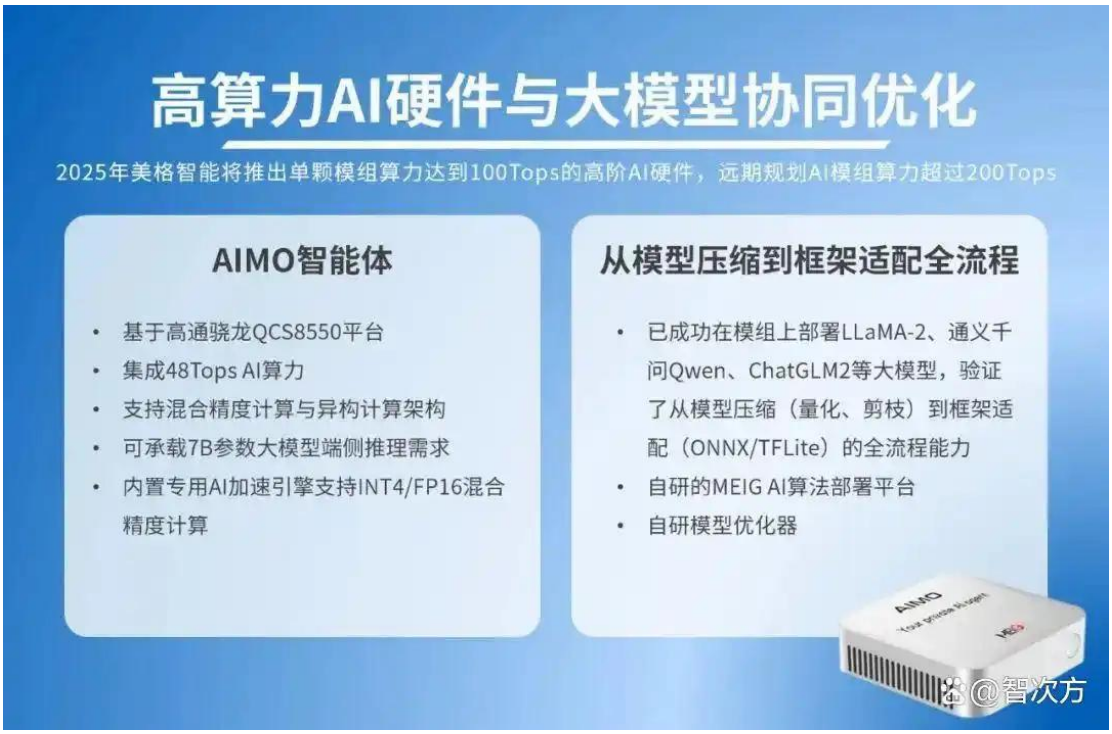

图 14：美格智能 AI 模组

（图源：美格智能）

日海智能（芯讯通）

日海智能投入数十亿资金完成多次基于人工智能物联网行业的横向产业并购，同时投入巨资进行产品及研发能力的升级，企业逐步实现物联网“云 + 端”的战略布局。围绕智慧连接为核心，日海智能构建了由无线通信模组、智能设备、通信服务、智慧物联解决方案组成的业务体系。

SIMCom AI Stack 是日海模组发布的 AI 全栈解决方案。该方案依托日海模组强大的 AIoT 产品矩阵，集成前沿算法与高效硬件架构，包含从最低的 1Tops 到最高超过 40Tops 的算力配置，可满足从智能家居的精准交互到工业质检的毫秒级响应等不同场景需求，持续赋能千行百业突破智能化转型瓶颈。

图 15：芯讯通 AI 模组

（图源：芯讯通）

有方科技

深圳市有方科技股份有限公司聚焦于物联网的“联”，专注于为物联网服务商和智能互联产品制造商等客户提供物联网接入通信产品和服务，产品涵盖接入云、管道云、2G/3G/4G/5G/NB-IoT/eMTC 等蜂窝无线通信模组和整机。正是凭借有方科技全球首创的基于云管端架构的接入通信解决方案，公司可以为物联网提供全球领先、可靠的接入通信，助力人类更环保、高效、便捷。

有方科技自主研发推出的 5G RedCap 模组 N520，采用 1T2R 双天线简化设计，能有效降低终端成本，同时以其低成本、低功耗、高可靠等特性，完美适用于智慧能源、工业控制、视频监控、车联网等场景。有方科技 5G 模组 N511、N512、N513、N521 等全面接入 AI 大模型，将持续赋能制造业数字化、智能化转型升级。

高新兴

高新兴科技集团股份有限公司成立于 1997 年，是中国物联网产业的标杆企业。公司专注于物联网、车联网、智能安防、金融物联和通信业务，为全球客户提供卓越的物联网端到端整体解决方案和服务。

高新兴科技集团正式发布“高擎警务大模型一体机”。该产品以高新兴自主研发的 AI 中台为核心，全面搭载国产 AI 芯片，面向警务垂直场景孵化出四大核心能力，标志着智

慧警务建设迈入“场景化智能体”新阶段。

映翰通

北京映翰通网络技术股份有限公司成立于 2001 年，是国内领先的工业物联网通信和整体解决方案提供商。公司专注于为客户提供工业物联网通信（M2M）产品及物联网（IoT）领域“云+端”整体解决方案，产品涵盖感知控制、网络通信、平台、应用、解决方案五个维度，应用于智能电力、智能制造、智慧零售、智慧城市等领域。

映翰通在端侧模组方面有多种方案，相关产品具备强大的数据采集和传输能力，可广泛应用于物联网设备中。如 AI 边缘计算网关采用不同的芯片硬件方案，提供多种不同的算力组合；EC5000 系列和 EC3000 系列边缘计算机成功部署 DeepSeek R1 蒸馏模型，开启了 “轻量化 + 高性能” 的边缘 AI 新范式，为工业质检、智慧交通、远程医疗等领域提供了更灵活、安全、高效的 AI 解决方案。

发展趋势与挑战

1. “通信+算力”深度融合：随着生成式 AI 的浪潮，AI 能力与通信能力不断融合，形成端侧 AI 模组作为本地 AI 推理的关键硬件支持。在通信模组的基础上，AI 模块进一步集成 NPU 和先进计算单元，为工业、城市和家庭场景提供异构计算能力，推动物联网由万物互联向万物智联演进。

2. 硬件算力与模型齐优化：端侧模组将配备更强大的算力芯片以满足复杂 AI 模型的运行需求。同时，模型优化技术如剪枝、量化、蒸馏等将被广泛应用，以降低模型的计算复杂度和存储需求，提高模型在端侧设备上的运行效率。这种硬件算力与模型优化的协同作用，将推动端侧 AI 模组在性能、功耗和成本之间取得更好的平衡，从而加速 AI 技术在智能家居、自动驾驶、机器人等众多领域的应用普及，为用户带来更高效、更智能的体验。

3. 从通用模组平台到垂直定制：端侧 AI 模组正经历从通用模组平台到垂直定制化的转变。这一趋势反映了市场对行业特定解决方案的迫切需求，以及通用模组在功能、性能和成本上难以完全满足多样化应用场景的局限性。通用模组虽具备广泛适用性，但垂直定制化模组通过针对特定行业需求的优化，提升了资源效率和应用效果。如现在正火的 AI 玩具模组方案。这种定制化不仅涉及硬件设计的优化，还包括软件的行业适配，确保模组能深度融入行业生态系统。未来，端侧 AI 模组的垂直定制化将为各行业提供更具竞争力的智能化解决方案。

2.2、模

定义与概述

什么是端侧 AI 模型？其实这是一个时效性的概念，与端侧设备的计算能力紧密相关。也许当前需要独立电源，机架式部署空间的算力设备，在一年后就能在一块单片机板卡、电池供电的场景下获得等效的算力。相应地，以前只有部署在云端、服务器端才能满足资源要求的 AI 模型，随着硬件的发展，很有可能在某个时刻流畅地运行在智能眼镜这样的穿戴式设备上。因此，比较严谨地来说，端侧 AI 模型，是指那些能运行于算力不高于主流移动设备（平板、手机、专业移动设备、可穿戴式设备、具身智能体、工业单片机、物联网设备等），基于传统机器学习和深度神经网络等算法的 AI 模型们（数值预测、视觉、语音、语言模型等），而且其运行不应显著降低设备的交互性能和电池续航能力。不太严谨的时候，还可以宽泛地把可运行于笔记本、车载算力设备和工业边缘设备的 AI 模型们也纳入端侧 AI 模型的范畴之内。

端侧 AI 模型的核心价值在于其支持数据本地处理的能力，这带来了快速响应、低延迟、高隐私性以及在无网络或网络不佳环境下的可用性等显著优势。此外，一些热门的云端模型推理服务（如 Deepseek），由于其算力稀缺、成本较高，端侧 AI 模型的这种去中心化推理模式也能大大降低模型的使用成本。

2.2.1 端侧语言模型

为了能让参数量巨大的大语言模型有可能部署于资源受限的端侧硬件和算力上，一般会采用模型压缩技术，主要的有：

1. 知识蒸馏（Knowledge Distillation）：利用一个大型的“教师模型”来指导一个小型的“学生模型”进行学习，使得学生模型能够学习到教师模型的泛化能力。

2. 剪枝（Pruning）：移除模型中冗余或不重要的权重或结构（如神经元、注意力头），分为非结构化剪枝和结构化剪枝。结构化剪枝更有利于硬件加速。

3. 量化（Quantization）：将模型参数从高精度浮点数（如 FP32）转换为低精度表示（如 FP16、INT8，甚至 INT4）。

4. 低秩分解（Low-Rank Factorization）：将一个矩阵分解为两个或多个矩阵的乘积，从而将高维数据压缩为低维表示，以减少参数量。常见的低秩分解方法包括奇异值分解（SVD）和矩阵分解（如 CP 分解、Tucker 分解等）。

除了模型压缩技术之外，一些模型架构也会专门为了端侧硬件资源而专门训练出参数量在 100 亿之内（常见的有 7B、3B、1B，甚至低至数十 M）的小语言模型（Small Language Model，SLM）。这些模型通过采用高质量、经过精心筛选的训练数据以及高效的模型架构设计，力求达到”小而精”的效果，在特定任务上逼近甚至超越参数量远大于自身的大语言模型。

在高效的模型设计方面，小模型们往往从这些方面着手：

1. Transformer 变种：大多数大语言模型基于 Transformer 架构和注意力机制，包括著名的 GPT 系列模型。而为了针对端侧资源进行优化，模型研究社区探索了多种注意力机制的变体，例如分组查询注意力（Grouped-Query Attention，GQA）、多查询注意力（Multi-Query Attention，MQA），以及由 Deepseek 所提出的多头潜在注意力（Multi-Head Latent Attention，MLA），都是以减少 KV 缓存大小和计算量为要旨。

2. 混合专家模型（Mixture of Experts，MoE）：MoE 架构通过将模型分解为多个“专家”子网络，并使用门控网络动态选择激活一部分专家来处理输入。端侧语言模型使用 MoE 架构可以有效地利用有限的计算资源。通过将多个小模型（专家）结合在一起，在保证精度的同时，减少计算量，

3. 非 Transformer 模型：Transformer 架构的算法一直为人所诟病的一点是它的复杂度是 n^2，当参数量巨大的时候，其训练所消耗的算力和存储空间增长惊人。因此模型研究社区也在研究复杂度为 n 的非 Transformer 模型，例如以循环神经网络（RNN）为核心结构的算法，也有了一些不错的候选。

评估端侧语言模型，除了常规的语言模型评价指标外，还高度关注其实际部署性能，主要包括：

1. 首字延迟（Time to First Token，TTFT）：从用户输入到模型生成第一个 token 的时间，直接影响用户体验的即时性。

2. 推理速度（Tokens per second）：模型生成后续 token 的速率，影响对话的流畅性。

3. 内存占用：模型运行时占用的内存大小，对端侧设备的硬件限制非常敏感。

4. 能耗：模型推理过程中的能量消耗，直接关系到设备的电池续航。

2.2.2 端侧视觉模型

端侧视觉模型是指在终端设备上执行图像或视频分析任务的 AI 模型，常见任务包括物体检测、图像分类、图像分割、人脸识别、光学字符识别(OCR)、姿态估计等。其典型的应用载体包括无人机、安防监控设备、视觉工业质检设备等。最新鲜出炉的应用案例就有，乌克兰军方使用博物馆中展出的图-95“熊”式和图-22M3“逆火”式战略轰炸机，训练视觉 AI 模型识别其轮廓和特征，并用搭载 AI 模型的无人机优先攻击停放在俄罗斯军事基地的战略轰炸机群。

即便计算机视觉模型远不如语言模型的参数量巨大和资源耗费高，为使这类模型更好地适配于资源受限的端侧平台，还是会在如下技术领域进行优化：

1. 轻量化 CNN 架构：模型研究社区提出一系列轻量级卷积神经网络（CNN）架构，如 MobileNet 系列（利用深度可分离卷积）、EfficientNet 系列（通过复合缩放策

略平衡深度、宽度和分辨率）、ShuffleNet 系列、SqueezeNet 以及 YOLO（You Only Look Once）系列的部分轻量化版本等。这些架构在保持较高精度的同时，显著减少了参数量和计算量。

2. 模型压缩与优化：与语言模型类似，剪枝、量化、知识蒸馏等技术广泛应用于端侧视觉模型，以进一步压缩模型大小并加速推理过程。

3. 硬件加速适配：视觉模型的许多算子（如卷积、池化）可以针对特定硬件（如 NPU、GPU、DSP）进行深度优化，以充分利用硬件的并行计算能力。这通常涉及算子融合、内存访问优化等。

4. 多模态融合：某些应用场景中，端侧视觉模型可能需要与其他模态信息（如文本描述、音频信号、传感器读数）进行融合，以实现更精准的场景理解和决策。

2.2.3 端侧语音模型

端侧语音模型是指在终端设备上实现语音到文本（Speech to Text，STT）转换、声纹识别、关键词唤醒、语音合成（Text to Speech，TTS），以及部分自然语言理解（NLU）功能的 AI 模型。端侧语音模型的典型应用场景包括增强现实（Augmented Reality，AR）眼镜等可穿戴式设备，智能家居助手以及汽车的智能座舱等。

语音模型为端侧环境进行技术优化的方向有：

1. 模型端到端（End-to-End）：现代语音识别系统趋向于采用端到端架构，直接将原始音频波形映射到文本序列，简化了传统声学模型、发音模型和语言模型分离的复杂流程。

2. 模型小型化与优化：针对主流的基于循环神经网络 RNN（如 LSTM，GRU）、卷积神经网络 CNN 和 Transformer 架构的声学模型进行压缩和优化，以适应端侧部署需求。例如，通过参数共享、低秩分解等技术减少模型参数。

3. 噪声鲁棒性：提升模型在真实嘈杂环境下的识别准确率是端侧语音识别的关键挑战。技术手段包括数据增强、噪声抑制算法的集成，以及模型对噪声特征的学习。

4. 低功耗唤醒：对于语音助手等需要持续监听唤醒词的应用，低功耗唤醒技术至关重要，通常采用一个极小型的声学模型持续运行，检测到疑似唤醒词后再激活主识别模型。

5. 自适应学习与特征工程：通过动态调整模型参数以适应特定用户的口音或环境噪声（自适应学习），以及优化梅尔频率倒谱系数（MFCC）等声学特征的提取过程（特征工程），可以提升识别性能。

2.2.4 其他端侧模型

除了上述端侧 AI 模型类别，还存在大量针对特定应用场景在端侧部署的专用 AI 模型，例如用于个性化推荐的端侧推荐系统模型、用于健康监测的生理信号分析模型、用于工业设备预测性维护的物联网设备数据分析模型、基于电磁波回波信号处理的场景感知模型等。

这些模型的具体技术特点因应用场景而异。例如，端侧推荐模型可能需要高效处理用户行为序列，并结合上下文信息进行实时推荐；工业预测模型可能涉及时间序列分析、异常检测算法，并对模型的可靠性和实时性有极高要求。而对于这些 AI 模型的共性需求依然是轻量化、高效性和对特定硬件的适配。除此以外，数据敏感型的应用场景可能还会要求端侧 AI 模型额外采用一些技术：

1. 数据高效学习（Data-Efficient Learning）：端侧设备往往难以获取大规模标注数据。因此，小样本学习（Few-Shot Learning）、迁移学习（Transfer Learning）、自监督学习（Self-Supervised Learning）等技术对于提升端侧模型性能至关重要。

2. 隐私保护技术：由于数据在本地处理，端侧 AI 本身具有较好的隐私性。联邦学习（Federated Learning）允许在不共享原始数据的情况下协同训练模型；差分隐私技术则可以在数据收集或模型输出时添加噪声，提供更强的隐私保障。

市场规模与概况

全球端侧 AI（Edge AI）市场正经历显著增长。根据不同市场研究机构的报告：

1. Fortune Business Insights 报告称，全球 Edge AI 市场在 2023 年估值为 203.9 亿美元，预计到 2032 年将增长至 1864.4 亿美元。

2. market.us 的一份报告指出，2023 年市场规模为 190 亿美元，预计到 2033 年达到 1630 亿美元，复合年增长率（CAGR）为 24.1%。

3. Grand View Research 估计 2024 年全球 Edge AI 市场规模为 207.8 亿美元，预计到 2030 年该数字能达到 664.7 亿美元，从 2025 年到 2030 年的复合年增长率为 21.7%。

4. GMInsights 的报告显示，2024 年 Edge AI 市场价值为 125 亿美元，预计 2025 年至 2034 年的复合年增长率为 24.8%。

5. Precedence Research 的数据则显示，2024 年全球 Edge AI 市场规模约为 211.9 亿美元，预计到 2034 年将达到 1430.6 亿美元，复合年增长率为 21.04%。

尽管具体数字略有差异，但各机构均预测未来 5-10 年内，全球 Edge AI 市场将保持 20%以上的高速年均复合增长率。北美市场是 Edge AI 的主要市场之一， GMInsights 指出，2024 年北美市场占据全球份额的 30%以上，其中美国市场预计到 2034 年将达到约 200 亿美元。Precedence Research 也提到，2023 年北美 Edge AI 市场规模已达到 70 亿

美元。

中国端侧 AI 市场同样展现出强劲的增长势头和巨大的发展潜力。

在市场规模方面，根据东方财富网援引的数据，2023 年中国端侧 AI 市场规模已达到 1939 亿元人民币。预计到 2028 年，市场规模将增长至 19071 亿元人民币，2023-2028 年的复合年增长率高达 58%。回顾 2018 年至 2023 年，其复合年增长率更是达到了惊人的 116.3%。华经产业研究院也给出了相同的数据，2023 年市场规模约为 1939 亿元，2024 年该数字就猛增到近 5000 亿元。

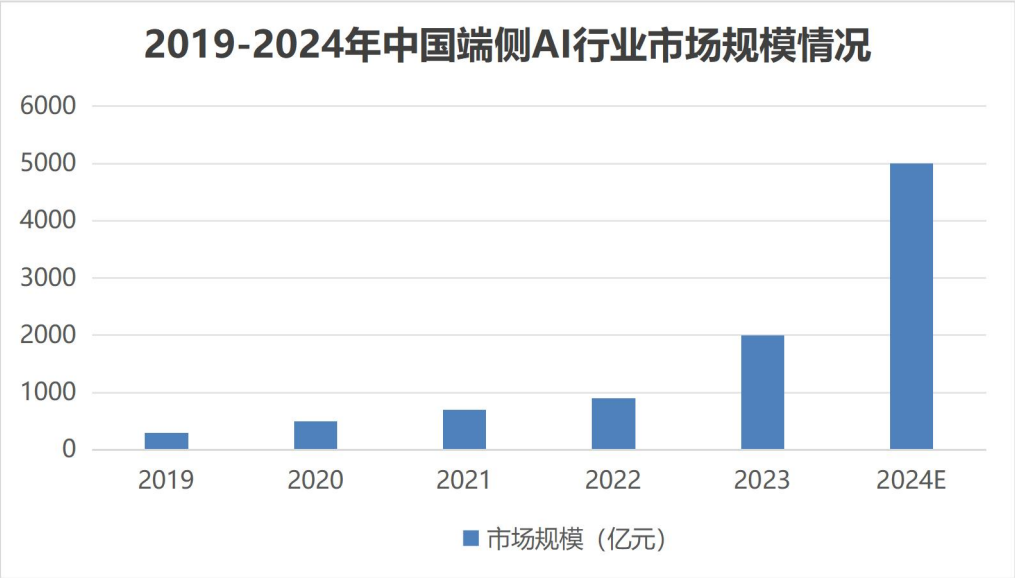

图 16：国内端侧 AI 市场

（数据来源：公开资料，智次方制图）

全球端侧 AI 模型市场形成了由中美两强主导，众多创新企业参与的多元化竞争格局。

美国，凭借其在底层操作系统、芯片设计和 AI 算法研究方面的长期优势，占据了技术制高点。谷歌和苹果通过其强大的移动生态系统，深度整合端侧 AI 能力；高通则通过其骁龙平台，为海量安卓设备提供硬件和模型优化支持；Meta 等公司也在积极布局 AR/VR 等新兴终端的端侧模型。

中国则依托庞大的终端市场、丰富的应用场景和快速迭代的创新能力，在端侧 AI 应用层面展现出巨大活力。华为、小米、OPPO、vivo 等手机厂商，以及阿里巴巴、腾讯、字节跳动和百度等互联网巨头，都在积极自研或合作开发端侧模型，以打造差异化的智能体验。同时，涌现出一批专注于 AI 模型小型化和特定场景应用的初创公司。

中国主流科技企业，特别是智能手机制造商（如华为、小米、OPPO、vivo）和 AI 技术公司，正在端侧 AI 领域加大投入，并在移动端操作系统层面（如开源鸿蒙、小米等）推动 AI 的深度集成与创新。端侧 AI 大模型被视为中游厂商实现应用落地的关键因素，

许多厂商选择自研端侧大模型。

论及端侧模型的重点应用领域，在 2023 年之前，智能安防和智能车载设备是中国端侧 AI 技术的重要落地领域。而随着软硬件技术的进步，AI Phone 和 AIPC（AI 赋能的手机和个人电脑）正迅速兴起，预计这两大领域庞大的市场需求将成为未来支撑中国端侧 AI 行业高速发展的重要驱动力。2025 年央视春晚上人形机器人的集体二人转表演，更是在中国这一制造业超级大国把具身智能再次推入了大众的想象空间。

主要企业与方案

中国移动物联网 AIoT 平台及行业大模型

中移物联深度融合卫星遥感、物联网、AI 等技术打造物联水利大模型、视联网监管大模型、生态环境 AI 大模型、万象耕耘农业大模型。此外，中国移动充分发挥通信网络与算力资源优势，在物联网领域深度融合大模型、多模态等 AI 技术，倾力打造“网联+物联+视联+智联+数联”一体化 AIoT 平台。网联构建稳定高效连接，物联支撑设备快速接入与统一管理，视联打造高清低时延视频服务，智联赋能端到端全流程 AI 智能升级，数联驱动数据价值深度挖掘。

物联水利大模型升级“水利推理模型、水利遥感智能解译、水利智能体”三大能力，打造智慧水利新范式，赋能水文水质监测、资源调度、水利工程维护、智能决策河湖治理、农业灌区等智慧水利场景。

图 17：物联水利大模型

（图源：中移物联）

生态环境 AI 大模型整合智能办公、业务智能体、舆情监测等核心能力，深度融合 AI 技术与环保专业知识，通过打造大气环境、水生态环境、环境统计、地下水环境四大智能体，实现环境数据的高效智能处理与分析，为环境质量监测、污染源溯源等领域提供精准解决方案，驱动环保工作向智能化升级。

图 18：生态大模型

（图源：中移物联）

视联网监管大模型具备强大高效的场景识别、万物检测、语义交互等核心能力，面向业务应用提供监管文件解读、零样本推理、L2 场景模型训练、大小模型协同、摄像头场景打标等应用服务，已上架 30+种 L2 算法，通过大模型能力已支撑安责险事故预防监控解决方案和自动化摄像头巡检方案打造。

图 19：视联网监管大模型

（图源：中移物联）

万象耕耘农业大模型构建“云端大脑+田间神经”的数字化体系，面向农业生产全场景提供农技智能问答、灾害预警与辅助决策、智能灌溉、长势监测与估产等核心功能。目前该平台已以河南为中心，服务全国近 300 万亩高标准农田，显著提升资源利用效率与粮食产能，为乡村振兴注入数字新动能。

图 20：万象耕耘农业大模型

（图源：中移物联）

此外，中国移动物联网 AIoT 平台将大模型、多模态交互等前沿 AI 能力，与超 10 亿级规模的连接能力深度融合，打造汇聚物联、网联、视联、智联、数联五大核心能力的平台底座，以"网联筑基、物联拓界、视联赋形、智联铸魂、数联聚势"为核心引擎，构建起端、边、云、智一体化协同的数字化中枢，支撑智慧城市、智能制造等万亿级场景。

图 21：中国移动物联网 AIoT 平台

（图源：中移物联）

北京中关村科金技术有限公司

中关村科金成立于 2014 年，是全球领先的大模型技术与应用公司，人工智能领域垂类大模型独角兽与领导者，2024 胡润中国人工智能企业 50 强。作为国家高新技术企业、北京市专精特新中小企业，中关村科金凭借在大模型垂直领域的深耕与突破，以“平台+应用+服务”三级引擎战略为核心，基于自研的得助大模型平台，打造覆盖多行业的垂类大模型解决方案，已服务 2000+央国企、政务、金融、制造等众多头部企业，引领企业从通用智能化向垂直专业化跃迁，已成为企业智能化转型的首选伙伴。

中关村科金自主研发的得助大模型平台作为业内领先的大模型平台，覆盖算力、数据、模型、智能体等全链路大模型开发和应用能力。支持国内外主流算力平台、自由搭配主流基础模型，是国内首批接入 MCP 协议的平台之一。平台基于客户实践沉淀出丰富的场景化能力，内置 200 多个 AI 能力组件和 100 多个开箱即用的行业智能体，深度结合行业 Know-How，支持四类编排模式： AutoAgent、Workflow、Knowledgeflow、Dataflow，帮助企业快速构建专属智能体。同时，平台拥有五重安全防护体系：内容合规、数据隔离、权限管理、审计溯源、生成式标识。支持国产信创适配及国际化合规部署，是垂类大模型落地的坚实安全基座。该平台可将项目交付周期缩短 40 天，GPU 利用率提升 20%，推理速度加快 4.5 倍以上，同时降低 30%的显存需求，大幅提升了企业 AI 应用的性价比和部署效率。

图 22：垂类大模型全景

（图源：中关村科金）

中关村科金已构建起从底层技术到业务价值的垂直穿透力。依托得助大模型平台，深度结合行业 know-how，中关村科金已联合客户与伙伴推出全国首个船舶工业大模型「百舸」、全国首个交通基建垂类大模型「灵筑智工」、医保垂类大模型、政务垂类大模型及金融垂类大模型等覆盖多行业的解决方案。这些大模型通过深度适配行业场景，为企业带来降本增效、精准决策、体验升级、风险管控等核心价值。

阿里巴巴

阿里巴巴作为中国业务版图最为全面的互联网巨头之一，很早就已经把不少端侧 AI 能力运用在电商、推荐、支付安全等移动应用场景中。在基础大模型领域持续不断地投入，使得阿里旗下的模型在国内的“百模竞速”这场马拉松之后，成为领先梯队中的佼佼者。

旗下通义千问开源模型当前最新的开源版本（Qwen3）包括混合专家架构（MoE）的 235B 和 30B 两款模型，以及 32B、14B、8B、4B、1.7B 和 0.6B 六款密集模型。8B 面向 PC 级别的端侧设备，4B 及以下模型针对的是手机及资源更受限的端侧应用场景。千问 3 原生支持 MCP 协议，并具备强大的工具调用（function calling）能力，为即将到来的智能体 Agent 和大模型应用爆发提供了很好的支持。

MNN (Mobile Neural Network)作为阿里巴巴开源的轻量级深度学习端侧推理引擎，为移动设备、PC、嵌入式设备等多种硬件提供高效的模型部署能力。MNN 支持

TensorFlow、Caffe、ONNX 等主流模型格式，兼容 CNN、RNN、GAN 等多种网络结构。MNN 具备轻量性、通用性、高性能和易用性特点，能在不依赖特定硬件 NPU 的情况下运行大型模型，支持模型量化和内存优化技术，能适应不同设备的算力和内存限制。MNN 提供模型转换、压缩工具和丰富的 API，让开发者能轻松地将深度学习模型部署到各种平台上。MNN 早已广泛应用于淘宝、支付宝等众多国民级应用中。

此外天猫精灵也在智能音箱中部署端侧 AI 模型，用于语音唤醒、声纹识别和部分本地指令处理。

阿里巴巴凭借其开源大模型性能在国内的领导地位，以及阿里云多年运营带来的端+云协同能力，在 AI 应用开发者中享有广泛的支持度。

百度

百度在大模型业务上，选择了与 OpenAI 类似的闭源且会员订阅的商业模式。但在 2025 年 2 月，百度宣布将于 6 月份把新一代大模型开源。百度如何平衡技术影响力和商业变现之间的平衡，还需要进一步的观察。

旗下文心大模型当前最新版本是 4.5，另有深度思考模型 X1。由于是在线服务，尚无法部署于端侧。

飞桨 Paddle Lite 是飞桨深度学习平台的端侧推理引擎，针对移动端和嵌入式场景进行了深度优化，支持广泛的硬件后端。

小度助手则在智能音箱、智能屏等自有硬件中大量应用端侧语音识别、自然语言理解模型，以降低对云端的依赖，提升响应速度。

此外 Apollo 自动驾驶平台是一个开放的、完整的平台，旨在帮助汽车行业及自动驾驶领域的合作伙伴结合车辆和硬件系统，快速搭建一套属于自己的自动驾驶系统。它有自己的参考硬件、开发工具和仿真平台。当前版本为 Apollo 10.0。

Deepseek

Deepseek 作为一家年轻的中国初创企业，以探索通用人工智能（AGI）的底层技术为己任。Deepseek 在大模型训练的工程化实践方面达到了世界级的创新水准，不仅开源的模型达到了当时 OpenAI 的旗舰级模型的性能水准，更是为全球的大模型技术社区开辟了低成本训练的崭新思路。为 AI 基础技术中美两强相争的格局奠定了基础。

旗下 Deepseek V3 采用混合专家（MoE）架构，总参数达到 6710 亿，每 token 激活 370 亿参数。对标 OpenAI 的 GPT-4，属于 L1 级别的聊天机器人。适用于大规模自然语言处理任务，如对话式 AI、多语言翻译和内容生成等。它能够为企业提供高效的 AI 解决方案，满足多领域的应用需求。

旗下 Deepseek R1 基于强化学习优化的架构。满血版 R1 参数为 6710 亿，属于 L2 级别

的推理优化模型，在复杂推理任务（如数学、编程、逻辑推理）中表现出色，上下文理解能力强，适合处理长文本分析和高精度需求的任务。

此外，蒸馏版 R1 如 DeepSeek-R1-Distill-Qwen-1.5B 等，这些版本基于满血版 R1 通过蒸馏优化技术得到。蒸馏版 R1 在推理速度、计算成本以及部署灵活性方面具有明显的优势。它们能在不同的计算资源和应用场景下，提供高性价比的体验。此外，蒸馏后的小模型在推理能力上显著超越了原始的 Qwen2.5 和 Llama 模型。蒸馏出的小参数量模型，也适合基于端侧部署。

华为

华为致力于构建万物互联的智能世界，其端侧 AI 战略是全场景智慧生活战略的重要组成部分，通过自研芯片、操作系统、AI 框架和模型，打造端云协同的全栈 AI 能力。

旗下盘古基础大模型目前提供 NLP、多模态、CV、预测、科学计算五大盘古基础模型。盘古大模型概念上为“5+N+X”三层架构，三层分别指 L0 层的 5 个基础模型、L1 层的 N 个行业通用大模型，以及 L2 层可以让用户自主训练的更多细化场景模型。

还有昇思 MindSpore Lite & MindSpore 端侧 AI 套件，提供覆盖云、边、端的全场景 AI 框架，MindSpore Lite 专为端侧推理设计，具有轻量、高效的特点。

鸿蒙操作系统（HarmonyOS）同样将端侧 AI 能力作为其分布式能力的核心，通过 AI 异构计算、意图框架等，实现跨设备的智能协同。在系统层面集成 AI 能力，提供统一的 AI 调度和开发接口。

还有 HiAI Foundation 这一端侧 AI 计算库和工具链，支持模型量化、算子优化，支撑高效端侧 AI 推理。

华为在 AI 能力上面是全栈的，这在中国 IT 厂商中是独一份。从底层芯片(昇腾、麒麟) 、AI 框架(MindSpore) 、AI 模型(盘古大模型)到操作系统(鸿蒙)和应用的全链路自研和协同优化。这也给了华为充足的底气来构建鸿蒙生态和昇腾计算生态，吸引开发者和合作伙伴，提供开放的 AI 能力和开发工具，赋能千行百业。

元始智能

元始智能是国内一家年轻的基础大模型初创企业，致力于旗下新一代大模型架构 RWKV 及其衍生 AI 应用的研发。RWKV（Receptance Weighted Key Value）是国产开源的首个结合了循环神经网络（RNN）和 Transformer 模型优点的新架构大语言模型，具有线性计算复杂度且支持并行化训练。

旗下 RWKV 模型当前最新版本是 RWKV-7 G1 2.9B、1.5B、0.4B 和 0.1B 四个版本，非常适合端侧部署和运行。RWKV-7 在文本生成、机器翻译、情感分析、对话系统、多语言处理、代码生成应用上都有超越同等参数量开源模型的表现。

谷歌

作为全球领先的科技巨头，谷歌在 AI 领域拥有深厚积累。其端侧 AI 战略旨在通过 Android 平台、自研的张量处理芯片（TPU）以及领先的 AI 模型，将强大的 AI 能力普及到数十亿设备，提升用户体验并赋能开发者生态。

Gemini 是谷歌的旗舰多模态大模型系列，并未开源，当前最强版本是 Gemini 2.5 Pro。而家族中 Gemini Nano 是专为端侧 Android 设备设计的高效版本。谷歌训练了两个版本的 Nano，参数分别为 Nano-1 的 1.8B 和 Nano-2 的 3.25B，分别针对低内存和高内存设备，采用 4 位量化进行部署，并提供一流的性能。

开源模型 Gemma 3 家族，是基于 Gemini Nano 架构的端侧多模态 AI 模型，支持文本、图像、短视频和音频输入，分为 1B、4B、7B 和 12B 四种参数规格，都适用于单块 GPU/TPU 的端侧运行场景。

TensorFlow Lite： 一个轻量级的、跨平台的机器学习推理框架，专为移动和嵌入式设备优化。

Google AI Edge SDK：为开发者提供在 Android 设备上部署和运行端侧 AI 模型的工具和服务。

谷歌的在端侧 AI 的商业模式与生态构建策略还是非常清晰的：消费者市场（ToC）方面，主要通过提升 Android 和 Pixel 设备的吸引力来驱动硬件销售；通过 Google AI Studio 和相关 SDK 赋能开发者， 构建繁荣的 Android AI 应用生态。而在企业市场（ToB）方面，Gemini 模型家族与 OpenAI 的旗舰模型相抗衡，争夺商用闭源大模型的企业客户，同时开源了轻量化的 Gemma 3 模型家族，保持在开源社区的技术领先优势。

微软

微软作为全球排名领先的 IT 企业服务巨头，是 AI 领域的长期领导者之一。微软又是巨头中第一个慧眼识珠投资了 OpenAI，并敢于把 OpenAI 的产品深度集成在自己的产品和销售业务中。微软的战略是通过 Azure 智能云服务和 Windows 平台，将 AI 能力赋能给企业和个人用户。在端侧，微软着力发展高效的小型语言模型和多模态模型，推动 AIPC 的发展。

Phi 系列小语言模型是微软自主研发并且开源的轻量级语言模型系列，其设计的核心理念就是在性能和规模之间取得极致平衡，使其非常适合在手机、笔记本电脑等端侧设备上直接运行，实现复杂的本地 AI 任务。目前的最新主力模型包括 Phi-4（14B 参数）、Phi-4-mini（3.8B 参数）、Phi-4-multimodal（5.6B 参数，多模态）。另外，另外混合专家架构的模型还是上一代的 Phi-3.5-MoE（42B 参数，活跃参数为 6.6B）。

ONNX & ONNX Runtime，开放神经网络交换格式（ONNX），微软是这一开放标准的主要推动者。ONNX Runtime 则是一个高性能的跨平台推理引擎。这套工具链使得开

发者可以轻松地将各种框架训练出的模型，优化并部署到包括 Windows 在内的各种终端设备上，是其端侧 AI 生态的重要基石。

Windows & Copilot+ PC: 这是微软端侧 AI 战略的核心。通过在 Windows 操作系统层面深度集成 AI 能力，并推出“Copilot+ PC”这一全新的 PC 营销品类，要求设备必须具备强大的神经网络处理单元（NPU）以在本地高效运行 AI 模型。目前，Copilot+ PC 通常搭载高通骁龙 X Elite 处理器。不过，Intel 和 AMD 也在计划推出符合 Copilot+ PC 标准的处理器。Copilot+ PC 与 Windows 11 24H2 结合的特色 AI 功能包括：Windows Recall、Windows Studio Effects、实时字幕、自动超级分辨率、语音清晰、“画图”应用的 Cocreator 以及“照片”应用的 Restyle 图像等功能。

微软是端侧（主要在 Windows PC 领域）AI 体验的主要推动者。最初通过 Windows Copilot 和集成的 Phi 模型，在 PC 上提供智能助手、内容创作、系统优化等功能，让 AI PC 的概念初步进入消费者的认知。当非 Windows 的 PC 和平板也开始为自己打上 AI PC 的概念后，微软进一步差异化，提出了 Copilot+ PC 的定义，为 AI 在端侧的体验和硬件标准设定了门槛。

微软还通过广受好评的开发者工具，如 Visual Studio Code 和 Github Copilot，帮助开发者更高效地产生和测试代码；Azure AI Studio 和 Copilot Studio，也能为企业提供定制化的 Copilot 解决方案。

Meta（原 Facebook）

Meta 曾一度是大模型领域的活雷锋，率先把用海量算力训练出的、可以对标 OpenAI 当时主力模型 GPT-3 的 Llama 模型及其训练权重开源了，一时风头无两。但当 Deepseek 以及 Grok 也纷纷开源了更强大的预训练模型之后，Meta 的 AI 技术栈有了一段时间的沉寂。

Llama 系列模型，曾经的开源模型社区的扛把子。2025 年 4 月发布的 Llama 4 多模态模型家族包括 Llama 4 Scout（109B，16 个专家的 MoE 架构，活跃参数为 17B）、Llama 4 Maverick（400B，128 个专家的 MoE 架构，活跃参数为 17B）、Llama 4 Behemoth（参数达到了惊人的 2T，16 专家的 MoE 架构，活跃参数为 288B）。虽然当前 Llama 系列的模型本身较大，但因为其在开源社区的良好基础，推动了大量针对其进行压缩和微调以适应端侧部署的研究和实践。

PyTorch Mobile:，主导开发的 PyTorch 框架的移动版本，是另一个主流的端侧推理框架，在开源社区拥护者众多

Ray-Ban Stories 和 Quest 3，在微软的 Hololens 和苹果的 Vision Pro 这类 AR 头显设备逐渐在市场上淡出之后，巨头中仍孜孜不倦推出可穿戴式 AR 设备的就只有 Meta 了。Meta 在最新的智能眼镜和 VR 头显中部署端侧 AI 模型，用于语音控制、情景感知和手势追踪等。

Meta 的生态策略多少有点令人迷惑。一方面，为了在开源模型社区保持其影响力，Llama 模型家族的参数量越来越庞大，越来越难以直接应用于端侧；另一方面，Meta 继续坚守可穿戴式设备阵营，发布自己第一方的头显和眼镜，但要说看好这一领域，却没有为可穿戴式设备的生态贡献多少 AI 赋能的基石。左手和右手看起来渐行渐远。

苹果

苹果以其软硬件高度整合的封闭生态系统著称。其端侧 AI 战略核心是"Apple Intelligence"， 强调在提供强大 AI 功能的同时，将用户隐私保护置于首位，大部分 AI 处理在设备本地完成。

Apple Intelligence 架构，在不久前举行的苹果 WWDC 2025 上，苹果对其发布仅一年的 Apple Intelligence 架构做了更新。包括一个约 3B 参数的端侧基础模型（Foundation Models），采用 2 位量化优化，专门为苹果的芯片而设计，同时开放了第三方开发者直接访问模型的权限。还有私有云计算框架，可以在保护隐私的前提下，将一部分 AI 请求放到苹果的云端进行计算。这是典型的端+云混合 AI 架构。

OpenELM，是苹果开源的一系列小型语言模型，参数规模从 0.27B 到 3B 不等，采用层级缩放策略优化参数分配效率。

Core ML 框架与 Neural Engine，Core ML 为开发者提供了在苹果设备上集成机器学习模型的工具，开发者可以轻松地将训练好的模型集成到 iOS, iPadOS 和 macOS 应用中。Neural Engine 是苹果 A 系列和 M 系列芯片中专门用于加速 AI 计算的硬件单元。

苹果其实研发了很多小而专的端侧 AI 模型，用它们提升苹果硬件产品在各种应用场景下的智能化水平和用户体验，从而来驱动设备的销售。但苹果并不热衷这些模型的开源，仅仅为开发者提供 Xcode 26、Core ML 等工具和框架，鼓励其在 App 中集成苹果限定的 AI 功能，丰富苹果生态。OpenELM 的开源，基本上可以看作是苹果在小语言模型方面的跟随之举。

发展趋势与挑战

1. 技术演进加速：端侧 AI 模型以在资源受限设备上实现高效率、低延迟、低功耗的 AI 推理为核心目标，可以根据应用任务可分为语言、视觉、语音等多种模型，充分发挥其优化的技术特点。目前有从专用小模型向端侧多模态、更智能的小型化大模型演进，并强调与硬件的深度协同的发展趋势。另一方面，神经架构搜索（NAS）和自动化模型压缩（AMC）等技术被广泛应用，能够根据特定硬件的算力、内存和功耗约束，自动设计出最优的端侧模型结构，实现了"千机千面"的深度优化。

2. 端云协同/混合式 AI 成为主流：业界普遍认识到，端侧 AI 与云端 AI 并非简单的替代关系，而是一种互补和协同的关系。混合式 AI(Hybrid AI) 架构允许根据任务的复杂

度、数据敏感性、实时性要求以及设备当前的网络和计算状态，动态地在端侧和云端分配 AI 计算任务。简单、高频、低延迟的任务在端侧处理，而复杂、需要海量算力的任务则交由云端，这种模式能够最大化地发挥两者的优势。苹果公司的 Apple Intelligence 就是这一模式的典型代表。

3. 硬件推动持续增强：AI 芯片，特别是专为 AI 计算设计的神经网络处理单元（NPU），其算力的快速提升和功耗的不断优化，是端侧模型得以发展和普及的基石。高通、联发科、苹果、英伟达、英特尔等芯片厂商持续推出性能更强的端侧 AI 芯片和平台，为更复杂模型的端侧运行提供了硬件保障。

4. 生态构建全面提速：端侧 AI 的成功依赖于一个繁荣的生态系统。芯片制造商、终端设备厂商、操作系统提供商、模型开发者以及应用开发者正以前所未有的紧密度合作，共同推动从硬件适配、模型优化、工具链完善到应用创新的全链条发展。而开源生态也持续繁荣。开源框架（如 TensorFlow Lite, PyTorch Mobile, ONNX Runtime）和开源模型（如 MobileNet, EfficientNet 系列）的蓬勃发展，也为端侧 AI 模型的普及奠定了坚实基础，吸引了大量开发者和中小企业参与创新。

5. 模型即服务（MaaS）向端侧延伸：领先的 AI 公司不仅提供云端 API，也开始提供高度优化的预训练端侧模型库，开发者可以根据需求直接调用或进行微调，极大地降低了开发门槛。

6. 端侧智能体协同崭露头角：包括鸿蒙、智用开物等生态将智能体和端侧模型的能力做了充分解耦，将“感知、推理、执行、迭代、调度”拆解在传感器、通讯链路、大模型、控制模块、执行模块和专有模型上，创造了“万物互联”的智能化雏形。

7. 端侧 AI 和端侧模型市场竞争中，有一些共性的挑战：模型优化与性能权衡、硬件碎片化与适配、数据孤岛与个性化、商业模式探索、安全风险：端侧模型也可能面临对抗性攻击等安全威胁，需要新的防护机制、地缘政治因素。

2.3、端

在端侧 AI 技术的强力驱动下，智能终端设备正经历一场从“功能执行者”到“主动服务者”的范式革命。通过计算架构、模型轻量化、多模态交互等技术的深度融合，AI 终端不仅重构了人机关系，更在消费电子、工业物联、健康医疗等领域催生了全新的智能化场景。

端侧 AI 正推动终端以数据本地化、响应实时化、服务个性化为核心的变革，重构人机关系本质。随着众多新形态终端的涌现，一个由“本地算力+智能体+多模态传感与交互”定义的终端智能时代已然开启，终端设备正在以更高效、更可信的方式融入人类生活的每一寸空间。

2.3.1 智能汽车终端

定义与概述

智能汽车终端是端侧 AI 技术在汽车领域的重要应用载体。它以端侧 SoC 为核心，集成了存储芯片、传感芯片以及智能模组等关键部件。基于端侧 AI，智能汽车终端可在本地完成复杂任务，如实时路径规划、危险预判、自动泊车，以及根据用户习惯调节车内环境、推荐娱乐内容等，在保障行车安全与效率的同时，带来个性化、智能化的驾乘体验，成为移动场景下端侧 AI 应用的重要载体 。

端侧 AI 正在助力汽车功能革新：

1. 智能驾驶升级：端侧 AI 推动汽车智能驾驶功能不断优化，芯片算力提升为智驾落地提供了动力。此外，汽车传感硬件向端侧智能发展，整合多项车内传感安全功能，采用运行终端侧 AI 算法的单芯片，提高驾驶期间的计算准确性和缩短处理时间。

2. 智能座舱升级：端侧 AI 让智能座舱支持更大规模的端侧大模型，实现端侧强智能，推动智能座舱向汽车智能体转变。2024 年 1 - 11 月，中国市场乘用车前装标配智能座舱搭载率提升至 72.36%，预计 2026 年中国智能座舱市场规模将达到 2127 亿元，年复合增长率约 17%，渗透率有望从 59%提升至 82%。端侧大模型应用以及多模态交互融合成为智能座舱的发展方向。

市场规模与概况

智能汽车作为汽车产业与人工智能、通信技术深度融合的产物，正推动交通出行向智能化、网联化方向变革。近年来我国积极推动智能汽车产业发展，中国智能汽车市场规模快速增长。中商产业研究院发布的《2025-2030 年中国智能汽车行业市场深度分析及投资前景研究预测报告》显示，2024 年中国智能汽车市场规模约 2152 亿元，近五年年均复合增长率为 29%。中商产业研究院分析师预测，2025 年中国智能汽车市场规模将达到 2822 亿元。

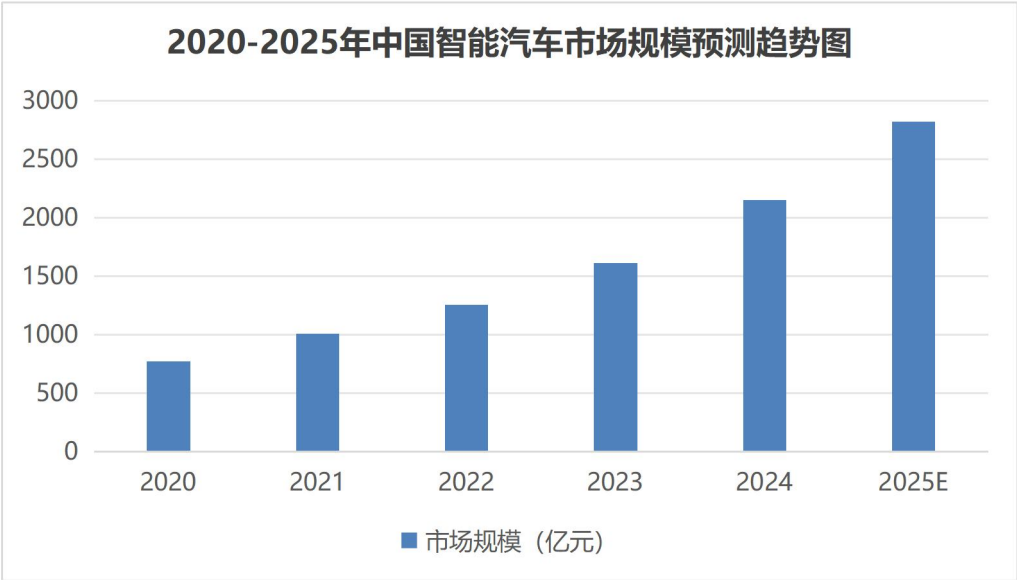

图 23：智能汽车市场

（数据来源：中商产业研究院，智次方制图）

智能座舱市场，根据毕马威的数据，2023 年智能座舱市场规模达到 1300 亿元；预计 2026 年中国智能座舱市场规模将达人民币 2127 亿元，2022-2026 年复合增长率超过 17%。

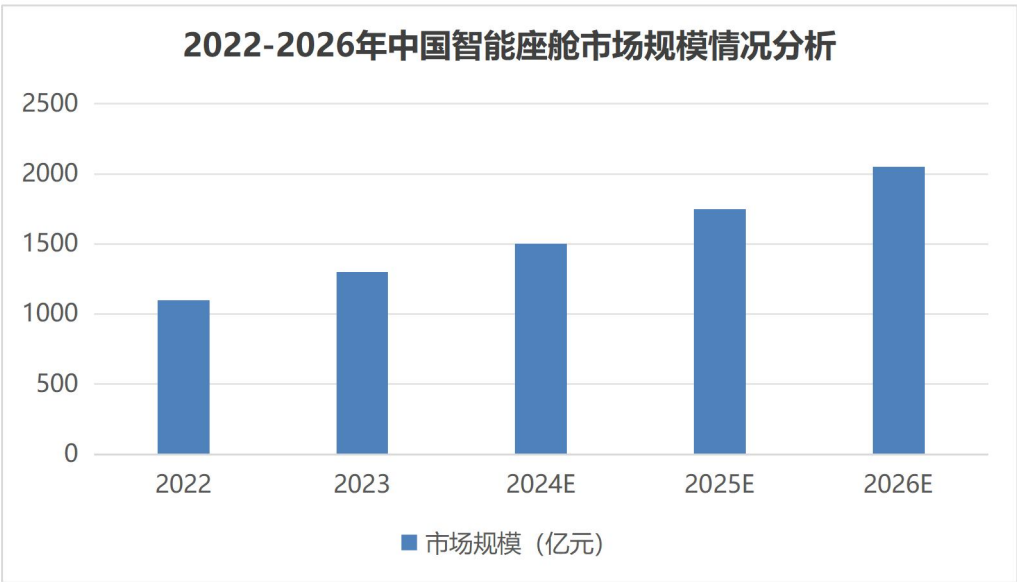

图 24：智能座舱市场

（数据来源：数据来源：毕马威，智次方制图）

主要企业与方案

华为

华为智能汽车解决方案业务聚焦智能网联电动汽车产业的增量部件，助力汽车产业的智能化、电动化升级，提供智能驾驶、智能座舱等一系列产品和解决方案。2024 年，华为智能汽车解决方案发布了以智能驾驶为核心的乾崑品牌，推出乾崑智驾 ADS3.0。2025 年，乾昆 ADS 4 高速 L3 商用解决方案正式发布。ADS 4 采用全新的 WEWA 架构，云端有世界引擎 World Engine，利用扩散生成模型技术生成案例场景；车端有 World Action Model 世界行为模型，用传感器等数据训练，为司机提供信息并控制车辆。相比上一代架构，ADS 4 端到端时延降低 50%，变道更丝滑，无效变道次数减少，通行效率提升 20%，重刹率降低 30%。

图 25：华为智驾方案

（图源：华为）

华为智能座舱方案采用 HarmonyOS 操作系统，具有面向全场景的分布式特性，能实现多设备融合与数据共享。如通过分布式软总线技术，可将手机、平板等智能设备与车机无缝连接，让用户在车内便捷地使用其他设备的资源和功能。硬件上采用麒麟芯片方案，如麒麟 990A 车规级座舱芯片，算力可达 3.5TOPS，为智能座舱的各种复杂功能提供强大的计算能力；麒麟 9610 车机模组具备标准化接口，能实现多屏协同，让不同屏幕之间的交互更加流畅和高效。

比亚迪

比亚迪股份有限公司主要经营包括以新能源汽车为主的汽车业务，手机部件及组装业务，二次充电电池及光伏业务，并积极利用自身技术优势拓展城市轨道交通及其他业务。智驾方面，比亚迪提出全民智驾战略，在整车智能战略下，构建起天神之眼技术

矩阵，全系车型将搭载高阶智驾技术，让高阶智驾人人可享。天神之眼技术矩阵包含三套技术方案，分别是天神之眼 A - 高阶智驾三激光版（DiPilot 600）、天神之眼 B - 高阶智驾激光版（DiPilot 300）、天神之眼 C - 高阶智驾三目版（DiPilot 100）。其中，天神之眼 C 配备前视三目 5R12V 感知硬件及端到端控制算法，实现架构、传感器、算法、数据四大领先。

智能座舱方案上，比亚迪 DiLink 智能座舱平台技术架构基于安卓系统深度定制，拥有开放的软件生态，支持海量第三方应用。DiLink 150 采用 4nm 制程的高性能车载芯片，拥有强大的 CPU、GPU 和 NPU 算力，为座舱系统的流畅运行和复杂功能的实现提供了坚实的基础。配备先进的语音识别和自然语言处理技术，支持多轮对话、语音控制、语义理解等功能，用户可以通过语音控制车内的大部分功能，如导航、音乐、空调、车窗等。多音区识别技术的应用，使得车内每个乘客都能享受便捷的语音交互体验

长安

长安汽车作为中国汽车品牌的重要代表，近年来积极布局智能化领域，尤其在智能驾驶和智能座舱方面取得了显著的进展。其智驾方案 —— 天枢智驾拥有覆盖激光雷达和主视觉融合感知方案，主视觉方案可实现高速、快速、高架等路况的 NOA 领航辅助驾驶能力，能满足 95% 用户的需求。同时，天枢智驾首发支持极黑环境下的避撞功能，夜视能力领先。

其智能座舱方案——天域座舱基于 SDA 天枢架构，推进智慧升级，实现 AI 渗透率 100%。融合天枢大模型，满足用户更智能的互动需求。支持多模态情感交互系统，通过语音、语义、环境、手势、眼动等多维度感知用户意图，实现无感交互，并在持续补充 AI 出行服务、AI 补能服务等功能。

蔚来

蔚来是注重核心科技的独立正向研发，在智能电动汽车的 “三电” 系统（电机、电控和电池包）和 “三智” 系统（智能网关、智能座舱和自动辅助驾驶系统）方面全部拥有自主知识产权。蔚来智驾方案超感系统配备 33 个高性能感知硬件，包括 1 个超远距高精度激光雷达、7 颗 800 万像素高清摄像头、4 颗 300 万像素高感光环视专用摄像头、1 个增强主驾感知、5 个毫米波雷达、12 个超声波传感器、2 个高精度定位单元和 V2X 车路协同，每秒产生 8GB 图像数据；超算平台搭载 4 颗 NVIDIA Drive Orin 芯片，算力高达 1016TOPS。

智能座舱方案以 NOMI 为核心，具有强大的学习能力和语音指令识别功能，能够表达情感，提供有温度、有感情的人机交互体验。拥有 4 组麦克风阵列和专用 NPU 核心，具备声纹识别、精准听音辨位、免打扰的独立音区交互能力，通过眼神注视可轻松唤醒，实现自由对话，还接入百科问答，能回答复杂问题。

上汽

上汽集团作为中国汽车行业的龙头企业，凭借强大的研发实力和广泛的产业布局，在智能驾驶和智能座舱领域取得了显著进展。通过与众多科技巨头的跨界合作，上汽不断推动技术边界，致力于为用户提供更智能、更便捷、更安全的出行体验。上汽集团通过多种方式推进智能驾驶技术发展，一方面，旗下零束科技自研智能驾驶平台，采用“高实时软件计算平台、业内领先的 BEV 算法、全域闭环数据工场”等关键技术，另一方面，上汽积极与外部企业合作，如投资国内领先的汽车边缘计算芯片厂家地平线，双方基于征程全家族计算方案达成全方位量产合作，地平线征程 6E 智驾方案将于 2025 年起搭载于上汽集团旗下荣威、名爵等多品牌车型上。

智能座舱上，上汽自研的银河全栈座舱 3.0 能够全面对接鸿蒙、安卓、IOS 三大手机系统，并与深度战略伙伴的深度融合，拓展了包括手机、手表、耳机、眼镜等海量生态场景。上汽还将 DeepSeek-R1 推理 AI 模型深度集成到其智能座舱中，建立了“集成边缘云 AI 模型中心”，支持多个高级 AI 模型的集成，实现了双 AI 模型的协作进化。

零跑

零跑汽车自成立以来，始终致力于核心技术自主研发，已成功自研智能动力、智能网联、智能驾驶三大核心技术，其智驾方案零跑 B10 的智驾系统基于高通骁龙 8650 智驾芯片开发，集成单激光雷达、毫米波雷达、超声波雷达以及摄像头等传感器。其中前向视觉识别模块为 800 万前视双目摄像头，激光雷达采用禾赛科技的 ATX 超远距激光雷达，探测范围最远 300 米，拥有 140°视场角。

其智能座舱方案基于 5nm 工艺的高通骁龙 8295 芯片打造，AI 算力达到 30TOPS，域控中集成 16GB 内存和 128GB 硬盘。支持多模态大模型（如 DeepSeek、通义千问），能够提供更自然、更智能的交互体验。

吉利汽车

吉利汽车是中国知名的汽车品牌，是全球具有广泛影响力的汽车制造商，业务涵盖汽车整车制造、动力总成研发、汽车零部件生产等多个领域。其智驾方案 —— 千里浩瀚以全域 AI 技术为核心，综合算力达 23.5EFLOPS，算法上星睿大模型与 DeepSeek 深度融合，并联合开源两款阶跃 Step 系列多模态大模型。

智能座舱方案银河 Flyme Auto 搭载吉利自研的国内首颗 7nm 车规级座舱芯片“龙鹰一号”，内置 8 核 CPU、14 核 GPU，AI 算力是高通骁龙 8155 的 2 倍，支持 2.5K 高清视频播放，具备 AI 应用持续拓展实力。

小鹏

小鹏汽车是广州橙行智动汽车科技有限公司旗下的互联网电动汽车品牌，小鹏图灵 AI 智驾搭载同级唯一双 Orin - X 芯片，总算力升至 508TOPS，是同级主流算力的 4 - 6 倍。感知系统由 27 个高精度感知元件构建，还搭载了同级中数量最多、性能卓越的超高清车感 SR，能精准识别超 50 种道路参照物。软件及功能方面拥有更强车端大模型赋能，实现了感知、决策到控制一体化。

智能座舱方案 AI 天玑系统小鹏不断提升硬件性能以支持更强大的智能座舱功能，在小鹏自研 XGPT 大语言大模型赋能下，座舱语音操控更全面，场景功能覆盖增加 30%，全车超 90% 的功能可用语音控制。

理想

理想汽车是一家中国领先的新能源智能汽车制造商，于 2015 年成立，在 2024 年发布“双能战略”，“智能” 与 “电能” 全面发力。其端到端 + VLM 的智驾方案采用“全场景端到端” 采用一体化的端到端技术架构，全部由一个模型来实现，今年还推送了首创的 “AI 推理可视化” 功能，将智能驾驶模型的思考推理过程以视觉形式展现，让驾驶员提前理解 AI 的思考和执行过程。

理想汽车的智能座舱系统不断优化，结合先进的神经网络算法，实现了更高效的面容识别、语音交互和多模态感知能力，逐步进化为生活助手 Agent。

小米

小米汽车是小米公司旗下的新能源汽车品牌，于 2024 年推出首款车型小米 SU7，Xiaomi HAD 端到端全场景智驾采用端到端大模型架构，将感知、预测和规划集成，直接将感知输入的原始数据通过大模型运算输出最终轨迹，减少信息传输误差。同时接入 VLM 视觉语言大模型，让智驾系统能识别更复杂的通行场景，并通过语音和文字与驾驶者交流。

小米澎湃智能座舱基于小米澎湃 OS 打造，交互体验与手机平板一致。桌面的 3D 模型可精准还原车辆状态，点击车模能进行开合尾翼、调节悬架高低等操作。采用五音区语音交互，小爱同学可精准应答不同位置乘客的指令，且支持分区权限管理。

发展趋势与挑战

1. AI 能力深化与多技术融合：端侧算力芯片不断进化，算力将更强劲，能效比持续优化，助力智能汽车终端实时处理海量传感器数据，实现更精准的环境感知、更复杂的决策控制以及更流畅的智能交互。同时，AI 模型架构创新，如 Transformer 等，将在端侧更好地落地，提升模型性能和泛化能力。此外，5G 通信技术的普及以及通信与计算能力的融合，使智能汽车终端能够与外界进行高速、低延迟的数据传输，实现

V2X 通信，即车辆与车辆、基础设施、行人等的实时信息交互，为自动驾驶协同和智能交通管理提供支撑。在云计算与边缘计算的协同下，将根据不同的任务需求，合理分配计算资源，进一步提升智能汽车终端的响应速度和处理效率。

2. 智慧驾驶功能拓展与优化：首先是自动驾驶升级，从辅助驾驶到更高级别的自动驾驶功能演进，智能汽车终端将具备更强大的环境感知、路径规划和决策控制能力。车辆能够自主应对复杂路况，如城市道路的拥堵场景、特殊天气条件下的行驶等，提高行车安全性和舒适性，未来有望实现全场景的无人驾驶；其次智能座舱体验提升也是必然趋势，智能座舱将成为车主的生活和工作空间的延伸，具备多模态交互功能，如语音、手势、眼神控制等，为用户提供更加自然、便捷的操作体验。同时，大尺寸、高清分辨率的车载显示屏，以及多屏联动技术，将提供更加丰富的娱乐、信息和导航内容，打造沉浸式的座舱体验。

3. 个性化程度提升：借助端侧 AI 的数据分析和用户画像构建能力，智能汽车终端能够为用户提供更加个性化、定制化的服务。车辆可以根据用户的驾驶习惯、偏好和需求，自动调整驾驶模式、座舱环境、信息娱乐内容等，满足不同用户的独特需求。此外，通过 OTA 技术，用户还可以根据自己的意愿，灵活选择和升级车辆的功能软件包，实现车辆的持续进化和个性化配置。

4. 数据安全与隐私保护：智能汽车终端在行驶过程中会产生大量的用户数据和车辆数据，这些数据涉及用户的个人隐私和商业机密，一旦泄露将给用户带来严重的损失。因此，如何确保数据的安全存储、传输和使用，建立健全的数据安全管理体系和隐私保护机制，是智能汽车终端面临的重要挑战之一。

2.3.2 具身智能机器人终端

定义与概述

机器人是典型的智能硬件产品，随着感知技术、运控技术、计算能力的逐步提升，以及 AI 模型快速迭代，具身智能机器人正在成为端侧的智能硬件集大成者。具身智能即通过智能体与环境的交互来获取信息、理解问题、做出决策并执行行动，从而展现出智能行为和适应性。传统的人工智能通常依赖于抽象的符号计算，而具身智能更强调通过物理身体的感知、运动以及与外部环境的交互来实现认知，并基于这些认知实现高阶自主智能。

具身智能机器人终端是集成端侧 AI 技术与机器人技术的具身智能的具体物理载体。具身智能机器人终端自主性与适应性极强，可自主学习优化行为模式，适应不同环境任务，实现真正自主智能。其应用广泛，覆盖医疗康养、智能客服、接待导览、零售导购等领域，为各行业智能化升级提供强大助力，未来发展前景广阔。

市场规模与概况

根据中国人形机器人与具身智能产业大会发布的《2025 人形机器人与具身智能产业研究报告》显示，2025 年，中国具身智能市场规模预计达 52.95 亿元，占全球约 27%；人形机器人市场规模预计达 82.39 亿元，占全球约 50%。

同时根据麦哲产业研究院数据，国内机器人市场规模 2023 年已突破 1572 亿元，其中工业机器人占据半壁江山，服务与特种机器人紧随其后。

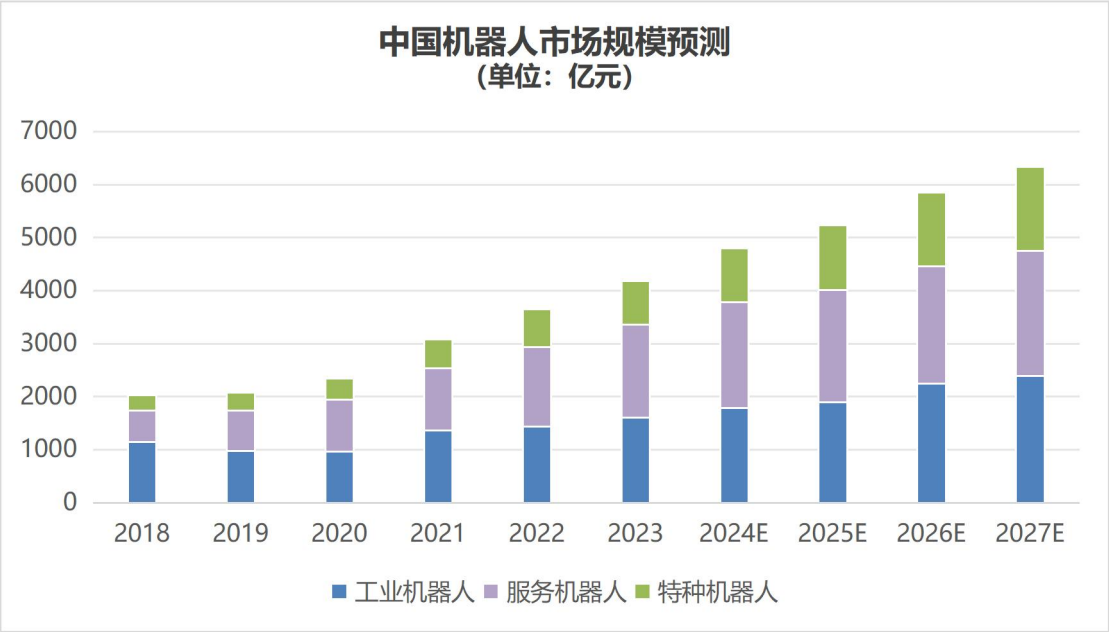

图 26：机器人市场规模与分类

（数据来源：麦哲产业研究院，智次方制图）

主要企业与方案

优必选

优必选科技成立于 2012 年 3 月，是人形机器人的领导者和智能服务机器人的领航企业。布局了人形机器人全栈式技术，开展智能服务机器人解决方案的研发、设计、智能生产和商业化应用，业务涵盖多个行业的企业级和消费级广泛应用场景，是全球极少数具备人形机器人全栈式技术能力的公司，拥有行业领先的人形机器人硬件与控制技术、人工智能技术、机器人与人工智能融合技术。

商用版人形机器人 Walker C 搭载了优必选自主研发的具身智能交互大模型，支持多语言交互应用，具备强感知、强适应性、强通用性等泛具身智能特点，可为多个商用场

景提供智能化服务。不久前，在 2025 年日本大阪世博会中国馆，Walker C 作为具身智能 “导览大使”，为参观者提供智能化的导览接待服务和全新的人机交互体验。

宇树科技

宇树科技 2016 年成立，是一家在民用机器人领域具有显著影响力的公司，专注于消费类和行业级高性能通用腿和人形机器人、六轴机械手等的研发、生产和销售。自 2016 年成立以来，宇树科技围绕仿生机器人、智能运动控制和 AI 感知交互展开技术创新，逐步形成了从四足机器人到人形机器人的产品演进路径，并在商业化应用上不断推进。

其具身智能机器人方案以 Unitree G1 人形智能体为例，硬件设计上配备深度相机和 3D 激光雷达等感知传感器，可精确感知周围环境，搭载 UnifoLM（Unitree 机器人统一大模型），提供机器人世界模型和共创平台，推动机器人从感知到认知决策的智能化发展，使机器人能够理解任务和环境，并做出合理的决策。

智元机器人

智元机器人（AgiBot）成立于 2023 年 2 月，创始人是稚晖君，致力于通过 AI 与机器人的融合创新，打造世界级的具身智能机器人产品。

远征系列，智元通用型具身智能机器人，其“具身智脑”包括云端超脑、大脑、小脑、脑干等部分，分别负责任务级、技能级、指令级、伺服级的任务。WorkGPT 是智元自研的任务级具身多模态大模型，赋予了机器人理解用户意图、感知环境、编排任务的能力。灵犀 X2，2025 年发布，集运动智能、交互智能、作业智能于一体，完美诠释了具身智能“AI+本体”的产品技术特点。

傅利叶智能

傅利叶智能成立于 2015 年，总部位于上海张江机器人谷，是国内领先的具身智能机器人研发企业。公司以康复机器人起家，后来转向通用人形机器人领域，推出了多款具身智能机器人产品。

2025 年傅利叶发布了 GR-2 通用人形机器人，进一步提升了运动控制能力和具身智能交互体验。同时，傅利叶与上海国际医学中心宣布双方将围绕具身智能机器人在康复医疗场景的应用标准建设、康复方案共创、科研攻关等展开全面合作，携手打造国内首个具身智能康复示范基地。

达闼机器人

达阀机器人成立于 2015 年，是一家专注于机器人研发的高科技企业，其产品广泛应用

于公共卫生、智慧城市、商业零售等多个领域。

达闼提出 “云脑 + 安全网 + 机器人”架构并实现商业化。其具身智能机器人方案以 RobotGPT 多模态具身大模型为核心，RobotGPT 以多模态 Transformer 为基础，具备多模态（文本、语音、图片、视觉、运动、点云等）融合感知、认知、决策和行为生成能力，能够基于人工反馈的强化学习快速智能进化。通过与机器人的具身智能相结合，RobotGPT 使机器人能够理解人类语言，自动分解、规划和执行任务，进行实时交互，完成复杂的场景应用，推动具身智能的自主进化。

普渡科技

普渡科技是全球服务机器人领域的领导品牌，专注于服务机器人的研发、设计、生产和销售，具备 “移动、操作、AI” 三大核心技术，在业内率先实现了专用、类人形和人形机器人产品的完整布局。

普渡科技具身智能机器人方案，以 PUDU D7 为例，采用 “大脑大模型” 与 “小脑大模型” 分离策略，通过多层次模型联动实现 AI 智能交互和具身智能学习能力，能执行多场景复杂任务。

云深处科技

杭州云深处科技有限公司（DEEPRobotics）成立于 2017 年，是一家专注于四足机器人及具身智能技术研发的高新技术企业，其产品在电力巡检、应急救援、工业检测等多个领域实现了落地应用。

其“绝影”系列机器人在电站、工厂、管廊巡检、应急救援等多种应用环境中落地应用。方案特点是通过多模态感知与智能决策技术，实现复杂环境下的自主巡检和应急响应。

追觅科技

追觅科技成立于 2017 年，是一家全球化科技公司，在高速数字马达、智能算法、流体力学及机器人控制等方面拥有一系列授权专利并处于世界领先地位。

追觅科技具身智能机器人方案主要体现在扫地机器人产品上，以仿生多关节机械手和具身智能大模型为核心。具身大模型包含感知模型和决策模型。感知模型基于视觉和大模型，用于环境感知；决策模型是语言模型，能把扫地机器人的技能和状态空间用网络做特征表达，直接输出技能序列。机器人搭载端侧模型，不联网也可使用，联网后能力更强。在抓取任务前，会调度多模态视觉大模型识别地面物品，决定是否抓取及如何处理，能将物品分为垃圾类和有用物品类，并基于常识规划放置位置。通过具身智能大模型与仿生多关节机械手的融合，追觅扫地机器人具备了家庭 “物理交互 +

空间智能” 的双引擎驱动，从 2D 清洁走向 3D 清洁，实现了从清扫到服务维度的进化，开启了服务机器人 3.0 时代。

星尘智能

星尘智能（深圳）有限公司成立于 2022 年，是一家专注于研发“新一代最强 AI 机器人助理”的高科技企业。公司创始团队核心成员来自腾讯、谷歌、华为、大疆等知名企业及国内外顶尖高校和人工智能研究院。星尘智能以其独特的“Design for AI”理念，致力于开发具备人类级操作能力的 AI 机器人助理，推动通用具身智能技术的广泛应用。

星尘智能提出了行业首发的面向 AI 的软硬件一体化系统架构，将“AI 智能”与“最强操作”强耦合，让机器人高度仿人，能像人一样学习、思考和劳动，与人流畅智能地交互，使用人的工具和设备，帮人完成枯燥、困难或危险的任务。星尘智能的自研端到端大模型正在积极寻求具身模型及迁移能力的突破，能在具身智能数据获取上取得关键性突破，能够低成本、高效率地利用已有的真实世界视频数据及人体动作数据，并通过第一人称高效收集多维度高质量数据，实现视觉、触觉、力觉等人类多模态数据交互。

帕西尼感知科技

帕西尼感知科技（PaXini Tech）是一家专注于多维触觉技术与人形机器人研发的创新型企业，致力于构建更自然的人机交互体系，推动触觉智能在机器人领域的规模化商用。

旗下 TORA - ONE 通过 VTLA - Model（视觉 - 触觉 - 语言 - 动作多模态感知模型），将视觉、触觉等多模态数据融合，使机器人能更全面地理解环境和任务。如 DexH13 仿生灵巧手集成多维触觉传感单元与高清手眼相机，实现 “多维触觉 + AI 视觉” 双模态感知，结合运控算法，可达到毫米级操作精度。

星动纪元

北京星动纪元科技有限公司于 2023 年 成立，由清华大学交叉信息研究院孵化，是一家专注于具身智能以及人形通用机器人技术和产品研发的企业，也是唯一一家清华大学占股的人形机器人企业 。其致力于成为引领行业的原生通用具身智能体定义者，通过打造 “端到端原生机器人大模型 ERA - 42” 和 “为 AI 定义的全新硬件平台”，聚焦研发高性能通用具身智能体。在短时间内，星动纪元凭借其创新技术和产品，在具身智能领域崭露头角。

其端到端原生机器人大模型 ERA-42，是首个端到端原生机器人大模型，直接从原始数据中学习，并根据环境和任务进行自主调整，极大地提高了学习效率和泛化能力。结合星动纪元为 AI 打造的全新硬件平台，可快速实现具身智能体软硬件协同进化和商业

化落地。

乐聚机器人

乐聚（深圳）机器人技术有限公司是一家专注机器人关键共性技术研究、智能机器人产品研发和生产的高科技企业，成立于 2016 年，产品主要包括通用人形机器人、中小尺寸双足人形机器人、编程教育系列机器人、医院物流机器人、重载运输机器人等几大类。

乐聚机器人与华为合作，推出了搭载盘古大模型的人形机器人“夸父”，展示了其智能化和泛化能力。方案特点是通过盘古具身智能大模型，实现软硬件层面的协同优化，能够进行物品识别、问答互动、击掌、递水等操作。同时，乐聚机器人与地瓜机器人达成战略合作，推出 Aelos Embodied 具身智能人才培养平台。该平台以地瓜机器人 RDK X5 开发者套件为核心算力载体，使机器人在步态规划、传感器融合、多模态交互等复杂任务中实现毫秒级响应，支持二次深度开发与算法快速迭代。

自变量机器人

自变量机器人是一家专注于具身智能通用大模型研发的公司，致力于通过该技术实现通用机器人，其软件算法团队兼具 Robotics Learning（机器人学习）和大模型的双重背景，硬件团队汇集了来自头部硬件公司的核心技术骨干及高管，拥有成熟的工程能力和量产经验。

自变量机器人实现了全球目前最大参数规模的具身智能通用操作大模型，Great Wall 系列（GW）的 WALL-A 模型，采用的技术路线为“统一具身智能大模型”，方案特点是实现了所有步骤“端到端”的完全纵向统一和不同任务的横向统一，用一个模型解决所有操作任务。

发展趋势与挑战

1. 感知升级：从发展趋势来看，感知能力的进化首当其冲，感知是机器人进行决策、执行的前提。如 3D 视觉技术将让具身智能机器人的 “眼睛” 更锐利，能够精准识别环境中的细微特征。与此同时，触觉感知会成为提升操作精细度的关键。借助电子皮肤、力矩传感器等，机器人抓取物品时能像人类一样感知力度与质地，避免出现用力过猛夹坏物品或拿不稳掉落的情况，这对于需要进行工业类、服务类机器人而言至关重要。

2. 大模型与本地模型相结合：二者是推动具身智能机器人前行的 “双引擎”，大模型凭借强大的学习与数据处理能力，基于多模态数据助力机器人提升感知、决策及自主学习能力，尤其在人形机器人领域，能让机器人在复杂环境中做出更合理的行为决

策。而轻量化模型则专注于在低算力条件下实现高效运行，适配不同硬件平台，为机器人在执行具体任务提供灵活决策支撑。

3. 载体形态多元化：具身智能机器人的载体形态不一定非要是人形，协作机器人、移动机器人、商用服务机器人等纷纷融入人工智能技术，将在各自擅长的领域发挥作用，如协作机器人助力工业生产更高效协同，移动机器人承担物流配送任务。人形机器人作为高阶形态，因其与人类相似的外形和功能，在对物理环境通用适应方面优势显著，未来有望在更多领域大显身手，如在家庭服务中辅助老人生活，在商业场景中提供个性化服务等。

4. 商业化难题待突破：首先当前机器人泛化与适应能力不足，如何使机器人在面对新任务、新环境时表现出真正的理解和适应能力，而非仅仅依赖于训练数据的模式匹配，是通往通用具身智能的核心难题。这仍旧需要较长一段时间来实现突破，例如发展更强的因果推理和常识理解能力。

2.3.3 AI PC

定义与概述

AI PC 是指将人工智能技术与传统 PC 深度融合，具备自然语言交互、本地大模型部署、混合算力支持以及强大隐私保护能力的个人计算机。在硬件构成上，AI PC 通常配备专用的神经处理单元（NPU）或集成 NPU 的芯片组 / 模块，用于专门处理 AI 工作负载，加速 AI 任务的执行，如图像识别、语音处理、自然语言理解等。同时预装了 AI 操作系统和 AI 生态系统，能够实现智能语音助手、智能写作助手、AI 智能剪辑等功能，还可以让用户方便地访问和使用各种 AI 应用程序及服务。AI PC 可以本地运行大语言模型，并在设备上高效处理数据，无需依赖云端，从而减少延迟，更好地保护用户隐私，同时能够快速响应用户的语音指令、文字输入等操作，实现智能化的人机交互。

市场规模与概况

据 Omdia 旗下 Canalys 最新报告，2024 年第四季度，AI PC 出货量显著增长至 1540 万台，占该季度 PC 总出货量的 23%。AI PC，即配备专门处理 AI 任务（如 NPU）的芯片或模块的台式机和笔记本，正逐渐成为市场主流。随着供应加速，AI PC 出货量环比增长 18%，全年占比达到 17%。该机构预估 2025 年全球 AI PC 出货量超过 1 亿台，占 PC 出货总量的 40%；到 2028 年，全球 AI PC 出货量 2.05 亿台，2024 年至 2028 年期间的复合年增长率将达到 44%。

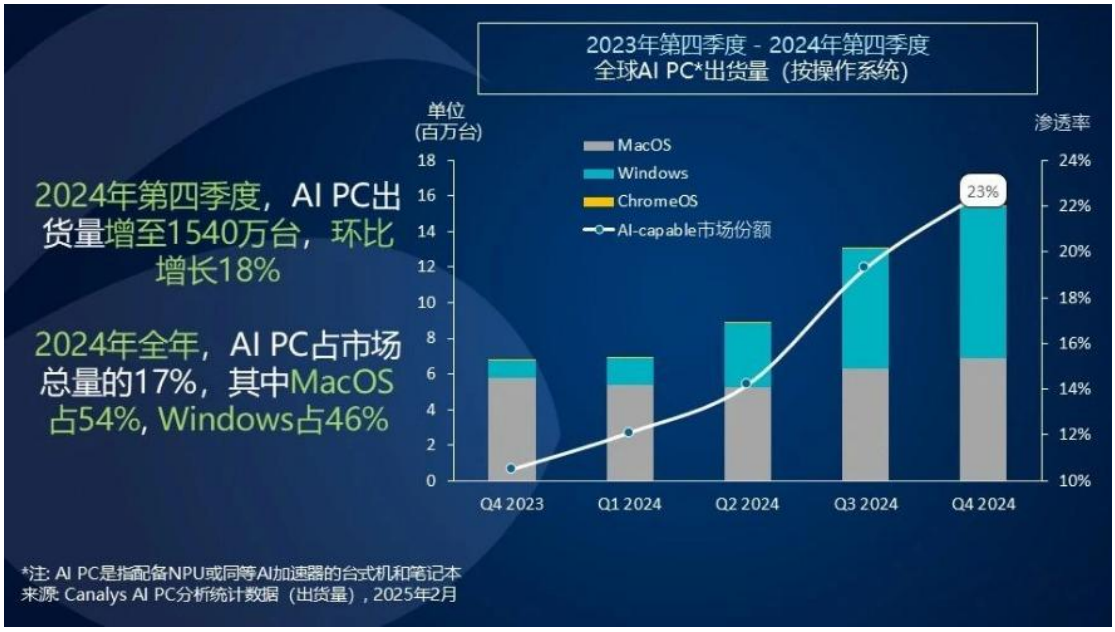

图 27：AI PC 市场情况

（数据来源：Canalys）

主要企业与方案

联想

联想集团成立于 1984 年，是全球领先的智能设备与解决方案提供商。作为全球 PC 市场的领导者，联想连续多年稳居出货量第一，2025 年第一季度以 24.1% 的份额持续领跑全球市场。其业务覆盖个人电脑、智能手机、服务器、数据中心解决方案及 AI 终端等领域。

联想在 2025 年推出多款突破性 AI PC，覆盖高端商务、创意设计、游戏电竞等场景。YOGA Book 9i AI 元启版是全球首款双屏卷轴屏 AI PC，采用英特尔酷睿 Ultra 处理器，支持隔空手势交互与天禧个人智能体系统；ThinkPad X1 Yoga 2025 AI 超能本搭载 Ultra 7 处理器与 32GB 高速内存，配备 2.8K 120Hz 触控屏，支持 HPD 人体感应与 5G 连接；拯救者 Y9000P 2025 AI 元启，电竞旗舰机型，采用酷睿 Ultra 9 处理器与双显卡方案，支持本地运行 70 亿参数大模型，满足游戏与 AI 创作需求。此外，天禧 个人智能体系统将发布基于个人云的 DeepSeek 大模型，实现端侧、个人云、公有云的协同，解决大模型 “高性能、低成本、安全” 的矛盾。

华为

2024 年华为在鸿蒙生态春季沟通会上推出了首款 AI PC 产品 MateBook X Pro 2024，MateBook X Pro 2024 首次应用华为盘古大模型，还集成了 WPS AI、文心一言、讯飞星火、智谱清言等三方合作大模型，首发 AI 空间功能，聚合一站式 AI 能力，支持 100

+ 个智能体。该机采用柔性 OLED 屏幕，重量仅 980g，是全球最轻薄高性能 OLED 笔记本，搭载酷睿 Ultra 9 处理器，配备 AI 调度大师，支持多任务毫秒级响应。

2025 年华为计划全面推出鸿蒙 PC，搭载鲲鹏芯片与纯血鸿蒙 PC 系统，叠加 AI 智能体与多终端设备协同功能，进一步完善其在 PC 领域的产品布局，为用户提供更丰富的产品选择。MateBook Pro 2025 超轻薄旗舰机型，搭载华为自研 NPU，支持 AI 慧眼、AI 音效及盘古大模型驱动的 AI 概要功能。

苹果

作为 AI PC 领域的领导者，苹果凭借自研芯片与端侧 AI 技术，在 2024 年第四季度以 54% 的份额占据全球 AI PC 市场主导地位。其核心战略聚焦 “软硬协同创新”，通过 M 系列芯片的神经引擎、端侧 AI 算法优化及 Apple Intelligence 功能整合，推动 MacBook 系列从生产力工具向智能终端演进。

2025 年春季发布的 MacBook Pro 搭载新一代 M4 芯片，其集成的 16 核神经引擎算力达 120 TOPS，较前代提升 200%，支持本地运行 70 亿参数大模型，实现 4K 视频实时降噪与智能剪辑，处理效率提升 40%。

惠普

作为 AI PC 领域的先行者，惠普通过 “硬件 + 软件 + 服务” 全栈式布局，在 2024 年第四季度以 12% 的份额位居全球 AI PC 市场第三。其核心战略聚焦 “本地化 AI”，通过惠小微智能助手、行业定制化解决方案及生态合作，推动 AI PC 从功能集成向场景赋能升级。

惠普 2025 年推出的 EliteBook Ultra G1i 与星 Book Ultra 系列，搭载英特尔酷睿 Ultra 7 处理器（NPU 算力 48 TOPS）与 AMD 锐龙 AI 300 系列芯片，支持本地运行 70 亿参数大模型。

戴尔

戴尔其核心战略聚焦 “开放生态 + 场景化创新”，通过与英特尔、AMD、英伟达等芯片厂商深度合作，构建 “CPU+GPU+NPU” 异构算力架构，推动 AI PC 从功能集成向场景赋能升级。

2025 年，戴尔推出全新品牌体系，将 PC 产品线整合为戴尔（Dell）、戴尔 Pro（Dell Pro）、 戴尔 Pro Max（Dell Pro Max）三大系列，覆盖日常办公、专业应用与极致性能场景。戴尔 Pro Max 系列为 AI 开发者与创意工作者打造，配备 NVIDIA Blackwell 架构 GPU（如 RTX 5000 Ada），本地算力达 1000 TOPS，支持 2000 亿参数大模型推理训练，采用模块化设计（可拆卸 GPU 扩展坞），可满足文生视频、数字人开发等高负

载需求。

华硕

华硕是全球领先的 3C 设计制造集团，其主要产品包括主板、图形卡、个人电脑、平板电脑、智能手机、显示器、服务器、存储设备以及网络产品等，在 2025 年推出了多款 AI PC 产品。

华硕无畏 14 AI 版 2025 年 2 月发布，搭载高通骁龙 X 处理器，基于 ARM 架构 4nm 先进制程，8 核 CPU，44%性能提升，NPU AI 算力高达 45TOPS，内置“小硕知道”AI 助手和 StoryCube 一站式智能媒体中心。灵耀 14 Air 骁龙版 2025 年 2 月发布，采用高科技陶瓷铝材质，整机重量低于 1kg，搭载 2.8K 120Hz OLED 华硕好屏、32GB 高频内存，配备骁龙 X 平台，NPU AI 算力高达 45TOPS。

荣耀

荣耀作为从华为独立的科技品牌，以 “创新、品质、服务” 为核心战略，业务涵盖智能手机、笔记本电脑、平板及 AIoT 设备，在 AI PC 领域打造 “端云协同 + 场景创新” 的差异化优势。

2025 年 2 月，荣耀发布 AI PC 2.0 战略并推出全新旗舰 ——荣耀 MagicBook Pro 14，搭载英特尔酷睿 Ultra 9 285H 处理器，首次实现离电插电同性能，彻底解决传统笔记本 “离开电源即降频” 的痛点。其全域调校技术 HONOR Turbo X 通过六大 AI 引擎（感知、学习、决策、芯片调优、智能调度、场景识别）实现软硬件深度协同，使笔记本在性能、续航、静音等六大维度全面进化。

宏碁

宏基业务涵盖个人电脑、服务器、显示器及 AIoT 设备。作为全球头部 PC 厂商，宏基核心战略聚焦 “AI 全场景赋能”，通过与英特尔、AMD 等芯片厂商深度合作，构建 “CPU+GPU+NPU” 异构算力架构，推动 AI PC 从功能集成向场景化创新升级。

2025 年，宏基以 “打破 AI 的藩篱” 为主题，在 Computex 展会上推出覆盖全线产品的 AI PC 生态体系，涵盖轻薄办公、专业创作与高阶电竞场景。如掠夺者•刀锋 14 AI，定位旗舰级 AI 电竞创作本，搭载英特尔酷睿 Ultra 9 处理器与 NVIDIA RTX 5070 独显，在 1.6kg 机身中实现 120W 性能释放，支持 120Hz OLED 触控屏与主动式触控笔，满足高精度创作需求。AI 视觉传感器可自动检测用户状态，执行锁屏、唤醒与视线感应操作，提升安全性与交互体验。

小米

小米公司作为专注于智能硬件和电子产品研发的全球化企业，在 2025 年推出了 AI PC 新品。

REDMI Book Pro 2025 系列首批搭载英特尔酷睿 Ultra 处理器（第二代），首次搭载「小米 AIPC 引擎」，内置端云融合大模型，打通 AI 调度、AI 感知、AI 办公、AI 创作、AI 隐私和跨端智联六大核心能力。其自研端侧模型可深入底层硬件识别场景，智能区分插电和离电状态，按需调度算力资源，实现性能与续航的双提升。

雷神科技

雷神科技成立于 2010 年，是海尔集团旗下专注于电竞硬件与智能设备研发的科技企业。产品覆盖电竞笔记本、台式机、显示器及 AIoT 设备。

2025 年 3 月发布的雷神 ZERO 18 Pro 是雷神首款 18 英寸旗舰 AI 游戏本，搭载英特尔酷睿 Ultra 9 275HX 处理器（24 核心 24 线程，150W 性能释放）与 NVIDIA RTX 5090 笔记本 GPU，整机功耗释放达 270W，支持本地运行 200 亿参数大语言模型。此外雷神 AI 智能体作为端云协同的智能中枢，全系 AI PC 预装雷神 AI 智能体，接入 DeepSeek 大模型，支持智能对话、文本生成、代码补全等功能，并拓展心光同步、Aippt（AI PPT 生成）、小 IN（智能会议助手）等生态应用。

发展趋势与挑战

1. 算力与能效的双重突破： AI PC 搭载的 NPU（神经网络处理器）、GPU 的高性能芯片持续优化，将实现本地 AI 算力的进一步跃升。同时，芯片制程工艺的进步与异构计算技术的普及，将进一步平衡算力与功耗，延长笔记本电脑的续航时间。

2. 大模型本地化部署加速：随着轻量化 AI 模型技术（如蒸馏、量化）的成熟，AI PC 将从依赖云端转向本地运行大语言模型（LLM）、多模态模型。例如，华为 MateBook 系列已支持本地 AI 大模型，可实现离线文档生成、代码补全；未来，更多 PC 将搭载更高参数级模型，用户无需联网即可完成复杂的 AI 创作任务，降低对网络的依赖，提升数据隐私保护能力。

3. 智能交互与场景化应用升级：AI PC 将突破传统键鼠交互，融合语音识别、手势控制等多模态交互方式。例如，用户可通过语音指令快速生成 PPT、自动整理会议纪要；在创意设计领域，AI 辅助工具可加速图像渲染、视频剪辑效率。

4. 软件生态碎片化与兼容性问题：不同厂商的 AI PC 在硬件架构、软件接口上存在差异，导致 AI 应用适配成本高，开发者难以构建统一的生态标准。同时，大模型的本地化部署依赖底层硬件支持，低配置设备难以运行复杂模型，加剧了用户体验的两极分化。

2.3.4 AI 手机

定义与概述

AI 手机是集成了先进人工智能技术的智能手机，搭载高性能芯片，如专用的神经网络处理器（NPU），能够快速执行 AI 算法；配备多种传感器，可敏锐感知用户与环境的复杂信息，如语音、图像、手势等；能通过机器学习算法，手机能够根据用户习惯自我优化，如智能调整拍照参数、预测用户需求等。此外，AI 手机通过端侧部署 AI 大模型，实现多模态人机交互，展现为非单一应用智能化的手机终端，以用户为中心提供个性化智能体验。

市场规模与概况

手机品牌争相布局 AI，对手机厂商而言，智能体可能成为一个强大的变现工具——当手机的 AIagent 收到指令时，它可以推荐与互联网公司服务相当的初创企业产品。IDC 则预计，AI 手机成为行业亮点，2025 年中国新一代 AI 手机市场出货量预计达 1.18 亿部，同比增长 59.8%，占整体市场的 40.7%。其他调研机构也相当看好 AI 手机的未来，根据 Counterpoint Research 预估，2024 年 AI 手机占比将达整体手机市场约 11%; 至 2027 年，占比更将提升到 43%，预估出货量成长 4 倍，年销量可望冲破 5.5 亿支大关。

市场调研机构 Canalys 近期也发布最新报告，预计 2025 年 AI 手机渗透率将达到 34%，端侧模型的精简以及芯片算力的升级将进一步助推 AI 手机向中端价位段渗透。2025 年芯片厂商发布的新款次旗舰 SoC 已经具备了流畅运行端侧大模型的能力，DeepSeek 的出现也在很大程度上降低了大模型对于芯片算力的开销，在这两大因素的共同作用下，2025-2026 年 AI 手机仍预计会保持高速渗透的趋势。

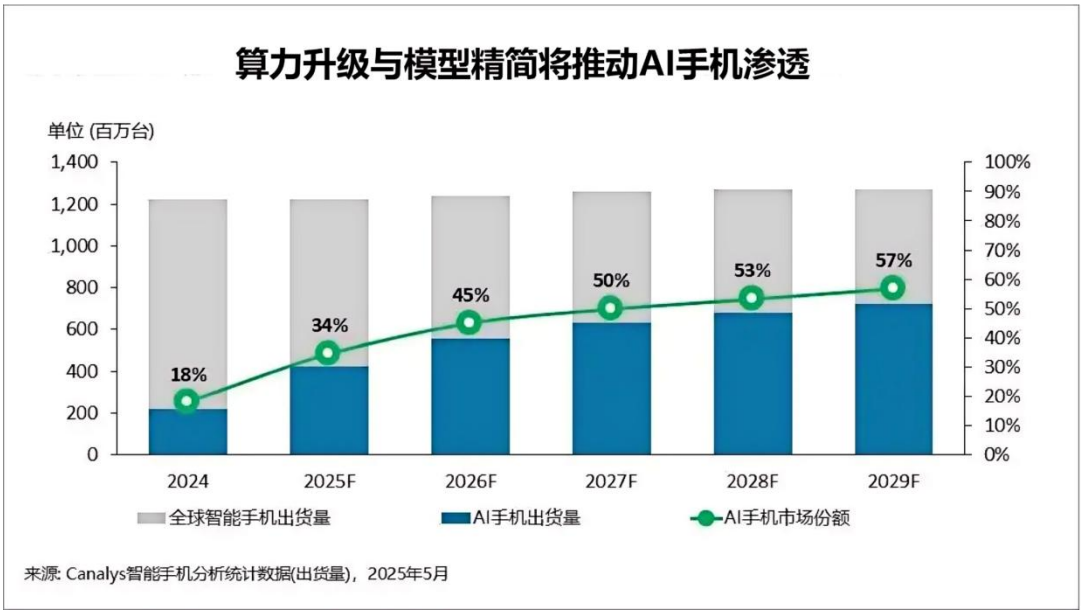

图 28：AI 手机发展情况

（数据来源：Canalys）

主要企业与方案

华为

华为正以 “端侧大模型 + 鸿蒙生态” 为核心，构建从硬件、软件到服务的全栈式 AI 手机解决方案，推动个人计算从 “工具属性” 向 “智能伙伴” 的代际跨越。

2025 年 6 月公布的 Pura80 系列（含 Pro 与 Pro + 两款）是华为 AI 手机的标杆之作。该系列搭载麒麟 9020 升级版芯片（采用 3D 封装工艺集成 SoC 与运存），预装原生鸿蒙 5 操作系统，首次实现盘古大模型与 DeepSeek 双引擎协同。nova 14 系列定位年轻消费群体，首次全面搭载鸿蒙 NEXT 操作系统，强化 AI 影像与通信能力。作为华为高端旗舰，Mate 70 系列则在持续深度整合鸿蒙 NEXT 的原生智能特性。

联想

联想在 2025 年发布的 moto razr 60 系列（含 Ultra AI 元启版、Pro AI 元启版及标准版）是联想 AI 手机的战略级产品，在硬件架构上 Ultra 版首发高通骁龙 8 至尊版芯片，集成 Hexagon NPU（算力 45 TOPS），支持同时运行语言、视觉、语音多模态大模型，AI 翻译 2.0 实现 112 种语言实时字幕叠加。在端侧 AI 能力上全系接入 DeepSeek-R1 满血版大模型（671B 参数），支持本地运行 70 亿参数模型，实现自然对话、文档智能处理（总结 / 翻译 / 润色）及 AI 绘画生成。此外，联想自主研发的天禧 AS 是 “一体多端” 战略的核心，2025 年 Q2 起在手机、平板等设备逐步部署。

荣耀

荣耀 Magic6 系列（含 Pro、至臻版及 RSR 保时捷设计）是荣耀 AI 手机的标杆之作。该系列搭载第三代骁龙 8 移动平台，支持 70 亿参数端侧大模型，实现端云协同的多模态交互。

荣耀 400 系列（含 Pro 与标准版）定位中端市场，首次全面搭载 MagicOS 9.0，将 AI 功能下放，标准版支持 AI 超清长焦与桌面无字模式，Pro 版引入 AI 重构编译引擎，应用启动速度提升 30%，后台保活应用数量增加至 18 个。

OPPO

OPPO 正以「端侧大模型 + ColorOS 生态」为核心，构建从硬件、软件到服务的全栈式 AI 手机解决方案。最早在 Find X7 上搭载天玑 9300 旗舰平台，搭载端侧 70 亿参数大模型，具备生成式 AI 能力，提升了影像、语音等多方面表现。Find X8 系列进一步升级，Ultra 版首发骁龙 8 Gen4 芯片，集成 Hexagon NPU（算力 50 TOPS），支持同时运行语言、视觉、语音多模态大模型。

小米

小米 15 系列将 AI 融入影像系统的多个层面，其影像模组接入了大模型，升级为首个 AI 大模型计算摄影平台 Xiaomi AISP。通过全面整合 CPU、GPU、NPU 和 ISP 算力，提供“超级抓拍”和“超级底片”功能。其次，其搭载最新的澎湃 OS 2.0 系统，基于 AI 大模型重构，拥有 AI 写作、AI 识音、AI 字幕、AI 妙画等功能，HyperAI 将端云大模型矩阵、多设备端侧感知、跨端执行能力全面整合，还加强了安全和隐私保护，可加密保存敏感数据。

Redmi Turbo 4 Pro 定位 AI 性能旗舰搭载联发科天玑 8350 Extreme 芯片，集成小米自研 NPU 加速单元，支持端侧大模型本地运行。

中兴通讯

中兴通讯成立于 1985 年，是一家全球领先的综合性通信设备制造商和解决方案提供商，在通信行业拥有深厚的技术积累和广泛的市场覆盖，其产品和解决方案涵盖无线通信、有线通信、光通信、数据通信等多个领域。2025 年中兴通讯以「AI for All」为核心战略，构建全场景智慧生态 3.0，整合商务出行、运动健康、智能驾驶等五大场景，推动 AI 技术普惠化。

努比亚 Z70S Ultra 摄影师版是中兴 AI 手机的代表，搭载骁龙 8 至尊版芯片，集成 Hexagon NPU，支持星云大模型、豆包、DeepSeek 多模型协同。此外，中兴通讯还构建了星云大模型家族，包括基础模型、通信大模型及行业大模型。

魅族

魅族科技是一家以智能手机为核心，覆盖 AIoT、智能汽车等领域的科技公司。通过「AI 平权」战略加速技术普惠，其 AI 手机产品在影像、交互、生态协同等领域展现差异化竞争力。

7 月即将发布的魅族 22 系列是里程碑之作，定位「旗舰 AI Device」，支持端侧部署轻量化大模型任务，并首发 Flyme AIOS 2.0，深度整合阿里云 Qwen2.5-Omni 全模态大模型与 DeepSeek-R1 视觉模型，支持动态调用通义千问、文心一言等第三方模型，根据场景智能分配任务。

vivo

vivo 成立于 2009 年，是全球领先的智能手机及智能生态解决方案提供商。vivo 以「蓝心智能」战略为核心，通过自研蓝心大模型、OriginOS 5 操作系统及跨设备协同技术，持续推动 AI 技术在手机领域的深度落地。

AI 手机领域，vivo X200 Ultra 其 AI 功能覆盖多模态交互，例如通过蓝心大模型实现 AI 实时 HDR 融合、动态范围扩展及多帧降噪。此外，X200 Ultra 支持与 iOS 生态的无缝互联，通过「双机流转」功能实现短信、通知跨设备同步，并搭载骁龙 8 至尊版芯片与 6000mAh 蓝海电池，兼顾性能与续航。

发展趋势与挑战

1. AI 技术深化：AI 大模型将加速在手机端的部署和应用，如苹果将 3B 参数量模型集成至其操作系统中，实现 Siri 功能增强和 APP 重构。手机 AI 芯片性能不断提升，AI 技术的应用将与手机的拍照、通信、办公等基础功能深度融合，如 AI 修图大师等功能，实现从功能升级到功能创新的转变，为用户提供更智能、便捷的体验。

2. 个性化与智能化服务增加：在端侧 AI 的助力下，AI 手机将作为智能体为用户提供专属的定制服务也将是重要趋势。借助 AI 技术，手机能够深度学习用户的使用习惯、兴趣爱好、工作生活需求等，为用户量身定制个性化的服务。比如，根据用户日常的出行轨迹和时间，智能推荐合适的通勤路线，并提前推送交通路况信息；依据用户的阅读偏好，精准推送感兴趣的文章、书籍、资讯等。

3. 算力与能耗亟待解决：高算力芯片在运行时会产生大量热量，若散热和能耗管理不佳，将严重影响手机的续航能力和稳定性，甚至导致设备过热降频，降低使用体验。如何平衡手机算力与能耗问题，是亟待解决的难题。

2.3.5 AI 眼镜

定义与概述

AI 眼镜是一种集成人工智能技术的智能穿戴设备，通过摄像头、传感器、微型处理器和显示屏等组件，结合内置或云端的 AI 算法实现多模态交互与智能化服务。其核心功能包括语音助手、图像识别、增强现实（AR）体验等，例如实时翻译、智能导航、健康监测及工业场景的远程协作。今年各大展会上，AI 眼镜都是焦点产品，其中中国厂商以压倒性优势占据舞台中心，雷鸟、XREAL、Rokid 等品牌推出的新一代智能眼镜，不仅实现了硬件性能的突破，更通过场景化生态合作，重新定义了人机交互的边界，2025 年也因此被业内视为 “AI 眼镜元年”。

市场规模与概况

根据 IDC 报告显示，2024 年全球 AI 眼镜出货量预计突破 900 万台，中国市场以 210 万台的出货量占据全球 23%份额，同比增速达 135%。Counterpoint Research 预测 2025 年市场规模将突破 120 亿美元，年复合增长率达 62%。

根据 QYResearch 数据显示，2023 年全球 AI 眼镜市场规模大约为 1.27 亿美元，预计 2030 年将达到 17.2 亿美元，2023—2030 年期间 CAGR 为 45.1%。2025 年 AI 眼镜行业将迎来新品密集发布期，自研芯片成为差异化发展的重要策略。当前，国内大型互联网公司均有相关产品正在设计或开发中，国内手机制造商也在进行类似的规划。

主要企业与方案

华为

在 AI 眼镜领域，华为发布了智能眼镜 2 钛空圆框光学镜，配备小艺翻译、头部控制等功能，支持面对面翻译、同声传译、全天候智慧播报。其 AI 功能基于鸿蒙 OS 5 系统与盘古多模态大模型 5.0，支持智能翻译与交互、健康监测与提醒、全场景生态协同三大核心能力，致力于构建「个人智能助手」新形态。

歌尔股份

歌尔股份是全球领先的智能声学整体解决方案提供商，业务涵盖智能硬件、声学器件、光学模组及微电子等领域。作为全球 VR 设备代工市占率超 70% 的头部企业，歌尔为 Meta Quest 系列、苹果 Vision Pro 等高端产品提供核心制造服务，并深度绑定华为、小米、索尼等品牌。

在 AI 眼镜领域，歌尔于 2025 年 CES 展会上推出两款轻量化 AI + 显示智能眼镜参考设

计——Mulan2（36 克）和 Wood2（58 克），标志着其在消费级 AI 眼镜技术上的突破。Mulan2 采用自研超薄碳纤维框架和全息波导镜片，集成音乐播放、AI 问答等功能，支持 Micro-LED 光机实现高清晰度显示；Wood2 则通过定制 SiP 模组和透明天线技术优化通信性能，搭载 VHG 体全息光栅波导和 1200 万像素摄像头，支持 4K 拍摄及多模态交互（语音 + 手势 + 智能指环），端到端延迟低至 2 秒，在嘈杂环境中仍能保持高拾音准确率。

雷鸟创新

雷鸟创新成立于 2021 年，是一家专注于 AI+AR 眼镜整机研发及软件生态构建的科技公司，由 TCL 创新实验室孵化而来。作为全球领先的消费级 AR 品牌，雷鸟创新在近眼显示光学设计、自研 AI 算法模型、多模态人机交互等领域拥有深厚积累，是业内唯一具备核心光学方案全链路自研及量产能力的企业，通过并线布局 MicroLED + 光波导、MicroOLED+BirdBath 等技术实现技术跨越。

在 AI 眼镜领域，雷鸟创新于 2025 年发布多款新品。旗舰产品雷鸟 X3 Pro 是全球首款量产的全彩 MicroLED 光波导 AR 眼镜，搭载自研萤火光引擎和 RayNeo 波导技术，该产品接入阿里云通义大模型定制化开发的 “云 + 端” 模型，支持可视化 Live AI 交互、实时语音翻译、物体识别等多模态功能，并开创性引入安卓虚拟机，可运行社交、文档等手机应用，构建 “虚实融合” 的新生态。雷鸟 V3 AI 拍摄眼镜则聚焦轻量化与 AI 交互，搭载高通骁龙 AR1 芯片和索尼 IMX681 传感器，支持 4K 照片拍摄及 1.3 秒极速 AI 响应，独家定制的连续视觉大模型可通过摄像头实现环境理解，同时具备 音乐点播、全场景录音总结等功能。

Rokid

Rokid 成立于 2014 年，是一家专注于人工智能与增强现实技术融合的科技公司，核心业务涵盖智能硬件研发、空间计算技术及多模态交互解决方案。作为国内首批布局 AI+AR 领域的企业，Rokid 在语音识别、单摄像头空间定位（误差精度达厘米级）、YodaOS-Master 空间操作系统等核心技术上拥有深厚积累。

Rokid 在 2025 年 CES 上展示了 Rokid Glasses 和 Rokid AR Lite 等产品，其中 Rokid Glasses 支持实时翻译、AI 问答、提词器等功能，采用衍射光波导技术，提供轻薄舒适的佩戴体验。此外，Rokid 还推出了全球首款消费级空间计算产品 Rokid AR Spatial，具备智能屈光度与瞳距调节系统，支持多窗口操作、3D 内容观看和多设备连接。

李未可

李未可成立于 2021 年，是一家专注于 AI 与 AR 技术融合的科技公司。作为国内领先的 AR 科技潮牌，李未可聚焦轻量化智能眼镜研发，核心技术包括自研 WAKE-AI 大模

型、多模态交互算法及超轻机身设计，产品覆盖商旅、运动、翻译等场景。公司通过开放 WAKE-AI 2.0 架构，联合阿里云、微软云成立「智能眼镜生态联盟」，推动 AI 能力在硬件端的规模化落地。

今年李未可推出多款 AI 智能眼镜及“零级智能体 ZeroAgent”。AI 音频眼镜与 AI 拍摄眼镜，均以轻量化为核心，分别定位不同用户群体需求。李未可发布的 AI 眼镜专属大模型“WAKE-AI 任务式交流系统”，其 AI 大模型支持 180 多种语言翻译，可实现跨语言交流、信息记录与整理、文旅场景应用等功能。未来，李未可将致力于群体智能研究，计划推出多个智能体协作系统。

XREAL

XREAL 成立于 2017 年，是全球 AR 眼镜领域的领军企业，作为 Android XR 平台的首批战略合作伙伴，XREAL 在光学显示、空间计算、自研芯片等核心领域拥有深厚积累。

今年，XREAL 创始人兼 CEO 在全球范围内首次对 AI 眼镜进行明确分级，XREAL 正瞄准 L4 级高阶 AI 眼镜，并计划在 2027 年推出。L4 级 AI 眼镜将具备高清 AR 显示、脑电和情绪识别能力，搭载端云协同的 AI 大模型，可实现主动预判并与用户建立记忆功能。同时，在 Google I/O 全球开发者大会上，XREAL 与谷歌联合发布了全球首款搭载 Android XR 系统的 AR 眼镜 ——Project Aura。此外，XREAL 与海信视像完成签约，双方将在 AI/AR 眼镜领域开展技术协同和生态共建，并于下半年推出 AI/AR 新产品。

影目 INMO

影目 INMO 成立于 2020 年，INMO 以 “定义未来眼镜” 为使命，产品以极致轻量化、时尚设计及 AI 融合为核心竞争力，自主研发的 IMAR 光学引擎、衍射光波导技术及多模态交互系统，实现了专业级显示效果与全天候佩戴舒适性。

影目科技今年发布了多款 AI 眼镜新品，其中 INMO AIR3 是全球首款量产的 1080P 无线 AR 智能眼镜，搭载自研 IMAR 光学显示引擎和新一代骁龙空间计算协作处理器，融合高清全彩显示、空间计算和 AI 语义交互功能。其还接入腾讯应用宝平台，成为腾讯应用宝内容生态在 XR 终端的首个落地合作，为用户提供全面的内容支持。此外，影目与中国移动合作，推出搭载九天大模型的 AI 智能眼镜，支持端云协同计算与多模态交互。另一款产品 INMO GO2 则是面向翻译场景的智能同传翻译 AI 眼镜，支持多语种在线和离线翻译，具备会议纪要 AI 整理和演讲提词功能，外观设计日常化，满足商务和生活需求。

闪极科技

闪极科技是一家先锋智能硬件品牌，专注于打造消费级 AI 眼镜等智能硬件产品。公司

以创新设计和强大功能为特色，致力于通过前沿技术提升用户体验，满足用户在生活、工作和娱乐等多场景下的需求。其产品不仅注重硬件性能，还在软件和 AI 技术方面进行了深度研发，为用户提供高效、便捷、智能的使用体验。

其首款 AI 眼镜闪极 AI 拍拍镜 A1 搭载索尼 1600 万像素摄像头、紫光展锐 W517 芯片及自研 Loomo OS 系统，支持全天候佩戴、4K 拍摄、Hi-Fi 级音质及多模态 AI 交互。公司通过与 LOHO 眼镜、科大讯飞、云天励飞等企业深度合作，构建起覆盖硬件设计、AI 算法、渠道服务的完整生态，并计划在北美及欧洲设立分公司，加速全球化布局。

创维数字

创维数字是知名的智能硬件研发与生产公司，公司积极创新，不断拓展业务边界，旗下创维 XR 推出了首款全场景 AI 智能眼镜。此外，创维数字还与华为、腾讯、微软等知名企业展开深度合作，共同推进 AI 与 AR 技术的发展。

创维数字旗下创维 XR 推出首款全场景 AI 智能眼镜，以 34.7 克超轻量化设计、24 小时全天候续航及强大的 AI 交互能力为核心突破点，通过超集成化硬件架构与开放式 AI 生态的深度融合，不仅解决了传统智能眼镜在实用性、舒适性及续航能力上的痛点，更以“科技隐形化”理念推动智能穿戴设备从极客专属向大众消费品转型。

发展趋势与挑战

1. 多技术集成与性能提升：未来 AI 眼镜将集成更强大的端侧 AI 芯片，结合低功耗 NPU 与高速 GPU，实现本地实时处理复杂任务。例如，在翻译场景中，AI 眼镜可瞬间识别多国语言文字并完成语音转换；在导航时，能通过环境识别自动规划最优路线。同时，显示技术将持续突破，如 Micro LED 与光波导技术的普及，会让画面更清晰、视野更广阔，提升 AR 显示效果。

2. 功能多元化与场景拓展：AI 眼镜的功能将从基础的语音交互、拍照翻译，向健康监测、工业辅助、娱乐交互等领域延伸。在健康领域，通过集成生物传感器，AI 眼镜可实时监测心率、血压、用眼疲劳等数据，甚至提前预警健康风险；在工业场景中，工人佩戴 AI 眼镜能获取远程专家指导，实现复杂设备的智能检修；在娱乐方面，AI 眼镜将打造沉浸式游戏、观影体验，成为移动娱乐新载体。

3. 轻量化与舒适性优化：随着材料科学与微型化技术的进步，AI 眼镜将变得更轻薄、更贴合人体工学。柔性电路板、微型电池、超薄镜片等技术的应用，能大幅减轻眼镜重量，降低长时间佩戴的不适感，推动 AI 眼镜从 “尝鲜产品” 向 “日常穿戴设备” 转变。

4. 续航短板影响使用体验：当前 AI 眼镜在续航、散热、显示效果等方面仍存在明显

短板。高功耗的 AI 芯片与小型化电池的矛盾，导致设备续航时间短；长时间运行易产生过热问题，影响性能稳定性；同时，AR 显示的清晰度、视场角以及画面延迟等问题，也制约着用户体验的提升。

2.3.6 其他智能终端设备

定义与概述

除 AI 眼镜外，其他智能穿戴设备也是端侧 AI 落地重要硬件。其他智能穿戴设备是指集成了先进传感器、无线通信技术和计算能力的便携式电子设备，通常以贴近人体的形式(如手表、手环、眼镜、耳机或衣物)佩戴，用于实时监测、记录和分析用户的生理数据、运动状态或环境信息。这类设备通过与智能手机或其他终端互联，实现健康管理、运动追踪、信息交互、娱乐控制等功能。智能穿戴设备作为物联网与消费电子融合的产物，近年来在全球范围内呈现爆发式增长。随着传感器技术、云计算及人工智能的成熟，智能手表、健康监测器等设备已从单一功能的"科技玩具"演变为健康管理、运动监测、生活助手的综合平台。

还有 AI 一体机，这是一种软硬件深度协同优化的一体化智能设备，集成了计算硬件（如专用算力模块、CPU、存储模块）、预装 AI 模型（含预训练模型与行业定制模型）、推理框架、开发工具及场景化应用程序，旨在为用户提供 “开箱即用” 的 AI 能力部署方案。AI 一体机是 AI 技术 “轻量化、场景化” 落地的载体，而端侧 AI 则定义了其 “本地化、低延迟” 的核心场景需求。两者的结合，本质是通过软硬件一体化设计，解决端侧算力有限、数据敏感、网络不稳定等痛点，让 AI 能力从 “云端集中式” 走向 “端侧分布式”。

市场规模与概况

除 AI 眼镜外，智能手表、智能手环、智能耳机是目前规模较大的智能可穿戴设备种类。

根据 TechInsights2024 年智能手表单位销量将同比增长 5%，TechInsights 可穿戴设备研究服务指出，预计 2024 年，全球智能手表的销量将达到 9100 万台，同比增长 5%；2025 年增长率将进一步上升至近 8%，到 2026 年将保持在 7%以上。2023-2029 年的全球智能手表市场复合年增长率为 5%。此外，根据观研天下数据，2024 年中国智能手表市场出货量达到 4317 万台，同比增长 18.8%。2024 年智能腕戴设备全渠道市场销量达 5704 万台，同比增长 14.2%，其中智能手表占比较大。

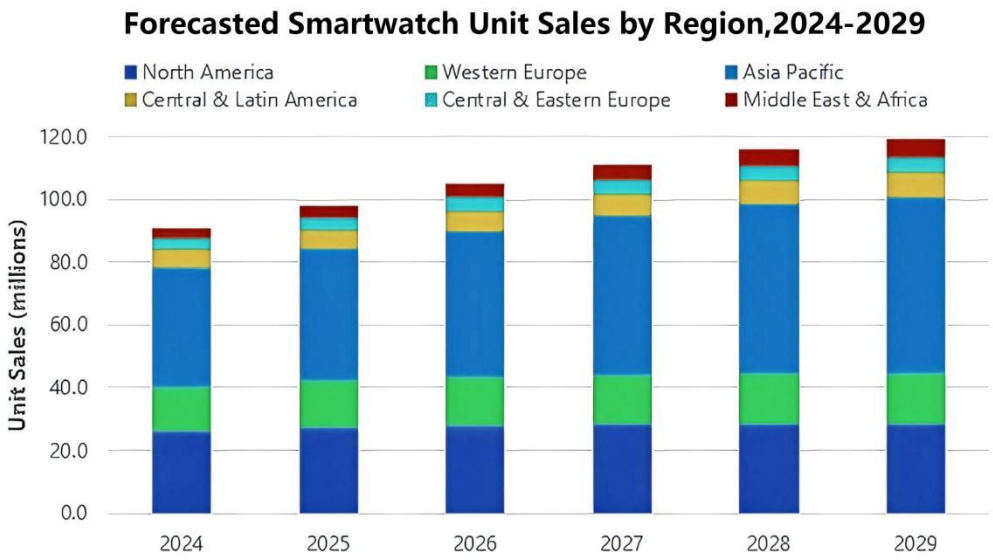

图 29：智能手表发展情况

（数据来源：TechInsights）

在手环领域，全球智能手环市场正处于爆发期，据中研普华产业研究院数据显示，2020 年全球市场规模为 78 亿美元，2024 年已突破 150 亿美元，年复合增长率达 18.2%。中国市场的表现尤为亮眼，2024 年出货量达 1.2 亿台，占全球总量的 40%，成为全球最大消费市场。从细分市场看，健康监测型手环占比达 65%，运动辅助型占 25%，时尚配饰型占 10%。值得注意的是，老年群体成为新兴增长极。2024 年老年健康手环市场规模同比增长 47%，血糖、跌倒监测等功能的渗透率显著提升。

智能耳机在 AI 技术驱动下也迎来爆发式增长，据智研资讯数据，2024 年全球智能耳机出货量约 5.38 亿台，其中，全球 TWS 耳机出货量约 3.32 亿台。近年来，中国耳机市场在生成式 AI 的洪流席卷之下，应用场景更加丰富，市场发展迅猛。智能耳机是中国耳机市场是主要推动力之一。2018 年我国智能耳机出货量为 0.36 亿台，2021 年增长至 0.96 亿台，2024 年我国智能耳机出货量约为 1.5 亿台。随着性能升级，智能耳机除了具有播放、采集声音信息的功能，未来将实现语音控制、语义识别、主动降噪、运动健康监测、虚拟现实声学和其他智能设备互联等功能升级，满足消费者工作和生活中的多种复杂应用需求，提升潜在多样化智能需求空间，智能化耳机市场具有广阔的市场前景和发展机遇。

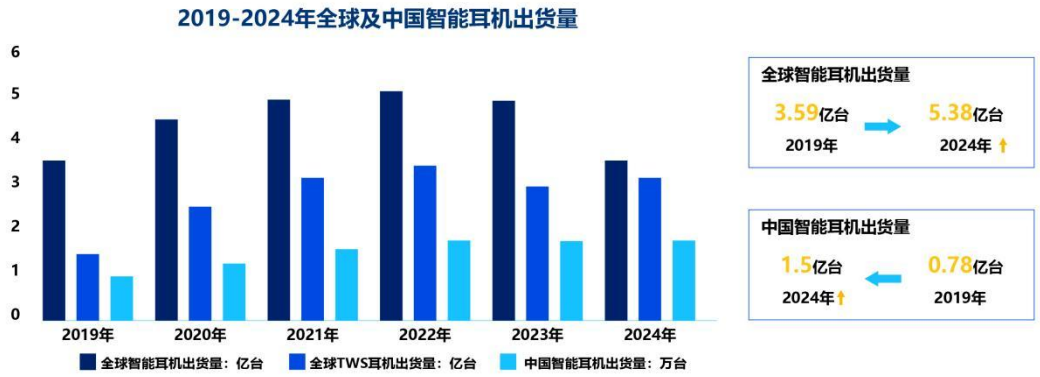

图 30：智能耳机市场情况

（数据来源：智研资讯）

主要企业与方案

中移物联行业一体机

中移物联依托中国移动九天人工智能平台，推出行业 AI 一体机系列产品，通过“硬件 + 模型 + 场景”深度融合，为水利、农业等领域提供本地化、低延迟、高可靠的 AI 解决方案。支持从芯片、算力到模型的全链路国产化部署，保障数据安全与供应链稳定。

中移物联万象耕耘一体机搭载万象耕耘农业大模型，集成卫星遥感、物联网传感器、无人机影像等多源数据，采用低功耗 NPU 芯片，支持本地化处理土壤墒情、气象数据等，提供 LoRa、蓝牙、RS485 等物联网协议接口，直接对接农田传感器、智能灌溉设备、无人机等终端。

中移物联水利一体机提供双引擎模型支持，预装 DeepSeek-R1-Distill-Qwen-32B 与九天水利大模型，支持自然语言理解、多任务推理及水利专业知识生成。集成气象、水文、地质等数据接口，构建“天空地水工”五位一体感知体系，洪涝灾害预警精度达秒级。作为首个通过“双备案”的央企大模型一体机，支持全链路国产化部署，已在全国 31 省水利厅规模化应用。

华为

在智能穿戴设备领域，华为不断推出创新产品，智能手表领域华为 WATCH 5 在 6 月 11 日正式亮相，其采用全新双引擎智慧架构，搭载 NPU 神经网络单元，同时手表接入 DeepSeek，并融合盘古大模型与运动健康专业模型，用户可通过腕上智能体智能分析 20 多个运动健康领域近 200 项指标，进一步提升了手表在健康监测和运动分析方面的智能化水平。

在智能手环方面，目前华为有多款手环产品在售，如华为手环 10 等。华为手环系列通常以其出色的健康监测功能和便捷的日常使用体验受到消费者喜爱，后续会在现有功能基础上，进一步优化健康数据监测的准确性和智能提醒功能。

智能耳机方面，华为 FreeBuds 6 搭载 HarmonyOS NEXT，在通话、音质等方面带来重要升级。还加入了星闪核心技术支持多设备无感切换和同时连接，耳机盒具备星闪查找功能。搭载麒麟 A2 芯片，整机支持 IP54 级防尘防水，耳机还具备空间音频功能，支持头部追踪，并增加了智能翻译和听力保护等 AI 功能。

荣耀

智能手表方面，荣耀手表 5Ultra 以 15 天超长续航、高端运动健康监测功能及 AI 技术引发关注，内置的 AI 健康监测功能，能实时提供用户健康信息，并支持多种运动模式，助力健康管理。

荣耀 EarbudsOpen 是荣耀音频耳机产品线的重要突破，设计方面，该耳机采用符合人体工程学的挂耳式设计，独特之处在于其避免了耳道，减轻了长时间佩戴带来的不适感。该耳机使用抗过敏液态硅胶，适合多种皮肤类型，并获得 SGS 亲肤认证，确保用户在长时间使用时也能享受到舒适体验。荣耀 EarbudsOpen 不仅在外观和舒适度上有所创新，其核心功能之一是 AI 实时翻译功能，可以将对方说的内容实时翻译为预设语言。

智能手环方面，荣耀手环 10 全面屏实现了媲美智能手机的丝滑交互体验，无论是日常翻页浏览菜单、快速切换运动模式，还是滑动查看健康数据，画面都过渡自然无卡顿，能结合 AI 算法精准分析用户健康趋势。

科大讯飞

科大讯飞作为知名的智能语音和人工智能上市企业，是中国最大的智能语音技术提供商，在中文语音合成、语音识别、口语评测等多项技术上拥有国际领先成果。

智能手表方面，科大讯飞正在开发一款嵌入其自主研发的星火大模型的儿童智能手表。星火大模型集成了自然语言处理、知识图谱、机器学习等多种技术，若嵌入儿童百科手表中，孩子可通过语音交互获取知识、解答疑问，还能记录孩子的运动量、睡眠质量等数据，同时可提供故事播放、音乐欣赏等娱乐功能。

在智能耳机方面，讯飞旗下未来智能发布了讯飞 AI 会议耳机，讯飞 AI 会议耳机 Pro 3 外观简约，内置 viaim AI 会议助理，支持多语言实时翻译、自动生成会议摘要等功能，还可进行内容提问与智能反馈，围绕会议全流程提供解决方案。

小米

智能手表方面，小米以其最新推出的纪念版智能手表——Xiaomi Watch S4/S4eSIM 15 周年纪念版，充分展现了其在 AI 创新和技术领先优势方面的深厚实力，搭载的小米自主研发芯片玄戒 T1，采用了先进的深度学习算法和神经网络优化技术，有效降低了功

耗，显著提升了续航能力，充分展现了 AI 芯片在能耗管理和智能优化方面的巨大潜力。

小米手环 9 Pro 在 2.133 固件版本中，在交互界面、操作智能化等方面实现了多项突破，此次固件升级的核心技术亮点集中在全新的 HyperOS 2.0 操作系统，其基于深度学习算法的优化，使得设备交互更为流畅自然。控制中心的可编辑功能，依托于深度神经网络的训练模型，能够精准捕捉用户偏好，实现智能化推荐与操作，体现了人工智能在穿戴设备中的深度应用。

OPPO

智能手表方面，OPPO Watch X2 系列更新了一系列 AI 相关的功能。通过 AI 智能算法，能够提供更加精准的健康监测服务，如内置 ECG 心电电极、8 通道心率监测、16 通道血氧检测和腕温传感器，能在 60 秒内完成覆盖大场景的 14 项指标体检。

智能耳机方面，OPPO Enco Free4 无线耳机拥有 AI 智能翻译功能，能为用户提供语音翻译服务。还具备智能自适应模式，可根据环境噪音自动切换降噪和通透模式，同时配备三麦 AI 通话降噪功能，通过三个麦克风配合 AI 技术，精准拾取用户声音并消除背景噪音。

VIVO

vivo WATCH 5 于 2025 年 4 月 10 日公布多项核心配置和功能，主打健康监测与轻量化设计，搭载多通道健康传感器，支持 30 秒快速血压风险评估及全天候监测，还新增心脏健康研究功能，可分析心律不齐等异常状况，结合 AI 血压分级算法与升级版 PPG 光学传感器，实现无感化监测，周期性生成血压趋势报告，并在异常时主动提醒，整机重量仅 32 克，支持长达 22 天蓝牙续航，搭载蓝河操作系统，内置“蓝心小 V”AI 助理，运动场景中还具备专业 AI 跑步指导功能，覆盖新手与进阶跑者，助力科学训练。

疯米科技

疯米科技（Funcl）成立于 2018 年，是一家专注于无线耳机研发、设计、生产与销售的互联网耳机品牌，其产品以高性价比和创新技术为特点。

疯米科技推出的疯米 AI 无线智慧耳机，搭载高通 QCC3026 芯片，支持 aptX 和 AAC 音频编码，具备 CD 级高保真立体声和 60 毫秒超低延迟。其重量仅 5 克，续航时间长达 24 小时，支持语音助手、智能语音问答和实时翻译等功能。

万魔声学

万魔声学是一家专注于声学耳机、研发、设计智能软硬件为主的创新型互联网公司，

成立于 2013 年，致力于呈现灵动、逼真的音乐，为用户提供高品质的音频产品和卓越的使用体验。

万魔声学推出了多款具有 AI 功能的耳机产品，如 1MORE 万魔开放式蓝牙耳机 S70 AI 版，具备同声传译、会议记录、大模型 AI 接入以及本地播放功能，可实现全时间的双向实时传译，打破语言壁垒。1MORE 万魔听力保护学习耳机 SonoFlow Mini HQ20 专为学习与听力保护设计，拥有三档智能音量控制、独家高频能量压制技术等听力保护功能，以及三档实时返听等学习辅助功能，搭载 AI ENC 降噪技术。

发展趋势与挑战

1. 功能细化与专业化：未来智能手表、智能手环将集成更多专业功能并逐步细化，如更精准的健康监测、移动支付、智能家居控制等，同时在运动、医疗等专业领域不断深化应用。例如华为的智能手表已具备专业运动模式和医疗级健康监测功能，未来还可能拓展至血糖监测等更多医疗应用。

2. 终端生态系统拓展：智能手表、手环将与手机、平板、电脑等设备更紧密地互联互通，形成更加完善的智能生态系统，实现信息共享、功能协同等，为用户提供更加便捷的智能生活体验。

3. 市场同质化与竞争加剧：随着市场规模扩大，智能可穿戴设备同质化问题凸显。多数产品在功能、外观上缺乏创新，依赖价格战抢占市场。此外，头部品牌占据主导地位，中小厂商面临生存压力，行业亟须差异化竞争策略。

2.4、智

端侧 AI 推动下的智慧应用，是指依托终端设备本地的人工智能计算能力，实现数据实时处理、智能决策与自主执行的新型应用模式。与传统依赖云端计算的模式不同，端侧 AI 将计算资源下沉至终端设备，通过高性能端侧芯片、边缘计算等技术，使设备具备独立完成复杂 AI 任务的能力，显著降低延迟、提升响应速度，并增强数据隐私保护。在这一变革浪潮下，智慧汽车、智慧工业、智慧城市等领域迎来了跨越式进化。

2.4.1 智慧汽车应用

定义与概述

从 “出行工具” 升级为 “移动智能终端”，端侧 AI 在汽车领域的应用正推动行业从电动化

向智能化深度跃迁，其核心价值体现在将 AI 算力与算法部署于车辆本地，实现实时决策、隐私保护与场景化服务的融合创新。在智能驾驶领域，端侧 AI 通过车载芯片与传感器协同，构建全场景感知能力。智能座舱是端侧 AI 的另一主战场。端侧大模型的引入重构了人机交互体验，端侧模型的部署，平衡了性能与成本，推动全民智慧驾驶进程。

端侧 AI 的意义不仅在于技术革新，更在于重构产业价值链条。其核心优势体现在：一是隐私安全范式革命，车内敏感数据如语音、图像无需上传云端，避免泄露；二是实时响应与可靠性，在隧道、郊野等弱网环境下仍能稳定运行，解决云端方案延迟痛点；三是成本优化，减少对云端算力的依赖，整车智能化成本下降。此外，端侧 AI 加速了汽车向“具身智能”进化，通过多模态模型打通物理世界与数字世界，未来或成为整合交通系统的智能实体。端侧 AI 正成为汽车智能化的核心引擎，引领出行方式的全面升级。

市场规模与概况

智能汽车作为汽车产业与人工智能、通信技术深度融合的产物，正推动交通出行向智能化、网联化方向变革。近年来我国积极推动智能汽车产业发展，中国智能汽车市场规模快速增长。中商产业研究院发布的《2025-2030 年中国智能汽车行业市场深度分析及投资前景研究预测报告》显示，2024 年中国智能汽车市场规模约 2152 亿元，近五年年均复合增长率为 29%。中商产业研究院分析师预测，2025 年中国智能汽车市场规模将达到 2822 亿元。

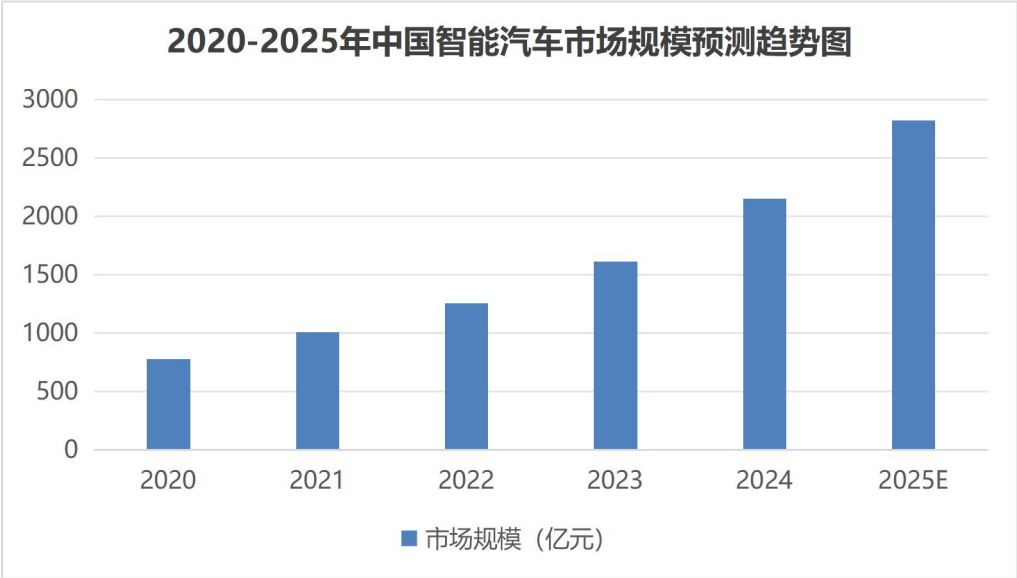

图 31：智能汽车市场预测

（数据来源：中商产业研究院，智次方制图）

主要企业与方案

中科创达

中科创达成立于 2008 年，是一家全球领先的智能操作系统及端侧智能技术和产品提供商，自 2013 年进入汽车领域以来，在智能汽车领域深耕不辍，构建了从智能座舱、智能驾驶、自驾域控平台、整车操作系统和中央计算平台等智能汽车全生命周期的产品和解决方案。

中科创达于 2025 年发布的全新的滴水 OS 1.0 Evo 操作系统，该系统以端边云协同 AI 原生架构为核心，灵活利用主流芯片平台算力，实现从车内外多模态感知到车辆底层能力，再到生成式 HMI 的座舱系统的全面重构，将 AI 融入系统的方方面面，为用户带来 “超真实、超智能” 的驾乘体验，并为汽车产业出海带来 “超融合生态” 的完整应用体系。同时依托端边云协同 AI 框架，深化边缘侧大模型 AI 能力，融合多模态交互与智能体技术，精准覆盖登车、用车至停车全场景，为用户提供直观、智能响应的全流程服务体验，覆盖 50 + 座舱意图域，实现毫秒级响应。此外，中科创达还携手面壁智能双方通过优势互补、资源整合，共同推动智能汽车大模型应用的优化升级。

百度

百度是中国领先的科技公司，在人工智能领域拥有深厚积累，尤其在自动驾驶和智能汽车领域处于行业前沿。百度推出了国内唯一、世界唯二的纯视觉高阶智能驾驶产品，具备城市、高速以及智能泊车全场景的点到点领航辅助驾驶功能，并基于百度 Apollo 自动驾驶大模型 ApolloADFM 全面升级。此外，小度车载智慧助手是基于文心大模型专为智能汽车客户打造的专属座舱智能体。

百度还推出了智能汽车云 3.0，从车端、云端、开发端、路端等多维度出发，端到端适配自动驾驶量产落地，其百舸 4.0 平台可实现主流异构多芯片训练，通过多种技术手段确保大模型的有效训练时长。此外，百度的 Paddle Lite 推理引擎是端侧高性能轻量化 AI 应用部署的利器，为端侧 AI 在智能汽车中的应用提供了重要支持。

岚图

岚图是东风汽车集团旗下的高端新能源汽车品牌，成立于 2018 年，聚焦智能电动 SUV 与 MPV 市场，以 “技术创新” 与 “用户体验” 为核心竞争力，2025 年战略目标冲刺年销 20 万辆并推出 4 款以上新车型。在端侧 AI 与智能汽车融合领域，岚图 2025 年上半年发布自研 AI 语音对话系统，车控响应速度小于 1 秒，唤醒识别率超 98%。该系统基于大模型的 AI Agent 架构，结合 Deepseek 的 CoT（思维链）训练方案，支持多步骤复杂指令的精准执行，并研发端侧大模型（On-Device LLMs）离线架构，优化弱网环境下的交互稳定性。同时，岚图自研的鲲鹏智驾系统基于天元架构，通过 31 个高性能感知硬件与端侧 AI 芯片协同，实现 L2.9 级智驾辅助。

面壁智能

面壁智能成立于 2022 年，是一家专注于大模型技术创新与应用转化的人工智能公司。2025 年面壁智能推出了首个纯端侧智能助手 “小钢炮超级助手 cpmGO”，该助手由面壁小钢炮 MiniCPM 模型驱动，是智能座舱领域的首个纯端侧方案，具备跨越舱外至舱内的全链条感知、决策与执行能力，实现了端到端的智能化应用。它拥有行业首个纯端侧 GUI Agent 屏幕助手，支持语音、手势、屏幕多模态交互，真正做到所见即可说、所指即所达。

面壁智能还与多家企业展开合作，共同推动端侧 AI 在智能汽车领域的发展，与英特尔建立战略合作关系，双方将共同推动端侧原生智能座舱的发展，定义下一代车载 AI；与长安、大众等顶尖车企达成深入合作，并与长城汽车签订战略合作，共研端侧大模型应用，深度布局汽车智能化的未来。

斑马智行

斑马智行是国内最早探索智能网联汽车的企业之一，专注于智能网联汽车操作系统的研发与应用，2025 年斑马智行发布了融合端到端框架和交互智能体两大 AI 智舱黑科技。其融合端到端框架将 pipeline 流水线架构与端到端架构深度整合，兼具二者优势，可将人机交互速度从平均 2 秒提升至 0.3-0.4 秒，交互效能提升 5 倍，还具备高准确率、有情感、可管理、可协调、可运营等特性，采用该架构的方案最早于 2025 年二季度首版上车，三季度量产。

同时斑马智行发布的元神 AI 智舱“一箭十星”交互智能体，包含 1 个 System Agent 和 10 个平台级应用 Agents，如用车控车 Agent、导航出行 Agent、亲子生活 Agent 等，通过自然语言即可调起对应场景化服务，解决 AI Agent 重构车服务生态的关键问题。

东软睿驰

东软睿驰成立于 2015 年 10 月，是汽车行业领先的软件及系统解决方案供应商，专注于基础软件、SOA 中间件、自动驾驶和跨域融合车云一体技术的研发。

东软睿驰推出的面向 AIDV 时代的整车智能操作系统 NeuSAR OS，该系统作为整车的神经中枢，实现了跨平台、跨域高效开发，大幅缩短应用开发周期，助力车企加速智能化迭代。其 NeuSAR AI Framework 中间件为 AI Agent 提供便捷开发框架和基础服务，提升开发效率。此外，东软睿驰还推出了 AI + 车云协同平台，通过场景服务引擎与车云算力底座，结合云端 AIGC 生成式场景组件和车云弹性算力扩展平台，解决算力挑战，支撑智能化场景高效落地。

Momenta

Momenta 成立于 2016 年，致力于通过突破性的 AI 科技，创造更美好的生活，通过数据驱动的“飞轮”实现技术持续进化，并结合量产自动驾驶（Mpilot）与完全无人驾驶（MSD）两种方案，推动智能驾驶技术发展。Momenta 率先实现“基于一段式端到端大模型”规模化量产，将感知与规划整合进一个大模型，形成端到端、深度学习的自动驾驶解决方案，降低了系统响应延迟。其智驾大模型采用“短期记忆”和“长期记忆”两条支路，提高训练效率，降低成本。

此外，Momenta 还计划在 2025 年下半年推出基于强化学习的 Momenta R6 飞轮大模型，进一步提升驾驶的安全性和可靠性。未来，Momenta 将继续秉持开放合作的态度，携手更多伙伴，为市场提供更优质的智能驾驶解决方案，推动智能驾驶行业迈向新高度。

发展趋势与挑战

1. 自动驾驶技术加速落地：随着 AI、大数据、云计算等技术的不断进步，L3 级及以上高级别自动驾驶技术将加速从研发测试向商业化应用过渡。特斯拉的 FSD 系统、华为的 ADS 3.0 等高端智能驾驶系统的推出，标志着自动驾驶技术已经达到一个新的高度，未来端到端自动驾驶技术将逐渐成为主流。

2. 车路协同与单车智能协同发展：通过车路协同、V2X 通讯、单车本地决策等关键技术，实现车辆与道路基础设施之间的信息交互，提高整体交通效率和安全性。未来车路协同与单车智能将相互补充、协同发展，共同推动智慧交通的进步。

3. 高质量数据匮乏：智能汽车应用需要依靠海量且精准的数据进行算法训练，但由于数据采集的范围与场景存在局限性，数据的准确性与一致性难以保障，缺乏有效的标准与体系，以及数据安全与隐私保护等问题，目前我国仍需要该方面的积累。

2.4.2 智慧工业应用

定义与概述

端侧 AI 在智慧工业的应用正推动着工业领域的深刻变革。通过在工业设备和传感器上部署端侧 AI 模型，能够实现实时数据处理和智能决策，大大提高生产效率和设备可靠性，降低维护成本和停机时间。在设备故障诊断中，端侧 AI 可以对设备运行数据进行实时分析，快速识别异常模式并预测潜在故障，使维护人员能够提前采取措施，避免生产中断。生产流程优化方面，端侧 AI 能够根据实时生产数据调整生产参数，实现生产流程的自动化和智能化，提高生产效率和产品质量。此外，端侧 AI 还可以应用于工业机器人，实现机器人的自主导航、路径规划和任务分配，提升生产灵活性和效率。

端侧 AI 在智慧工业中的意义不仅在于提高生产效率和降低成本，还在于增强工业系统

的智能化水平和自主性。与云计算相比，端侧 AI 具有低延迟、高带宽和数据隐私保护等优势，可以满足工业场景中边端侧对计算实时性和安全性的严格要求。同时，端侧 AI 有助于推动工业物联网的发展，实现设备之间的互联互通和协同工作，构建更加智能化的工业生态系统。

随着技术的不断进步和应用场景的拓展，端侧 AI 在智慧工业的发展前景广阔。未来，端侧 AI 将与云计算、边缘计算等技术深度融合，形成更加完善的工业智能架构。同时，端侧 AI 模型的性能将不断提升，能够处理更复杂的工业数据和任务，实现更加精准的预测和控制，推动工业 4.0 时代的加速到来，为工业生产带来更高的效率、质量和可靠性，助力工业领域实现数字化转型和可持续发展。

市场规模与概况

根据 fortunebusinessinsights 数据，2024 年全球智能制造市场规模为 3494.8 亿美元，预计该市场将从 2025 年的 3943.5 亿美元增长到 2032 年的 9989.9 亿美元，2025 年至 2032 年期间 CAGR 为 15.21%。

根据前瞻产研院数据，2022 年中国智能制造行业市场规模约为 4 万亿元，其中智能制造装备市场规模约 3.2 万亿元，智能制造系统解决方案市场规模约 0.8 万亿元。2023 年中国智能制造行业市场规模达到 4.3 万亿元，同比增长 7.5%。预计到 2027 年，中国智能制造行业市场规模将达到 6.6 万亿元，其中智能制造装备市场规模约 5.4 万亿元，智能制造系统解决方案市场规模约 1.2 万亿元。

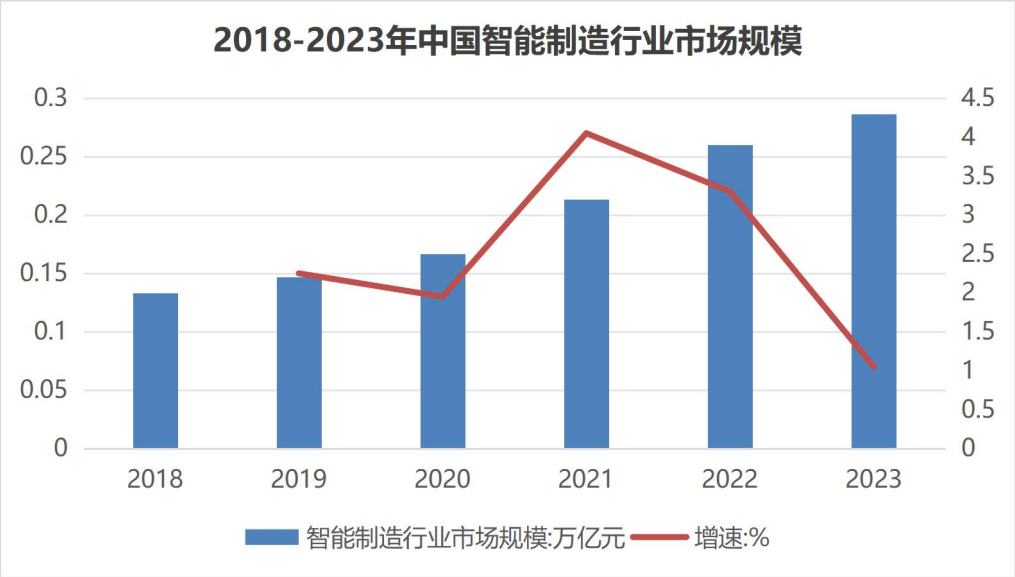

图 32：智能制造市场预测

（数据来源：前瞻产研院，智次方制图）

、

主要企业与方案

中移物联 OneOS

中移物联 OneOS 是中国移动物联网有限公司自主研发的高实时、高安全、高可靠国产化实时操作系统，内核自主度 100%，系统能力经院士专家组评审达到国际先进水平。全面适配 ARM、MIPS、RISC-V、LoongArch 等六种指令集架构，构建互联互通、自主可控的产业生态。物联网公司已完成基于宏内核精简架构 OneOS Lite、宏内核标准架构 OneOS Multi 以及 OneOS HYP 三个版本研发，OneOS 已通过 IEC61508 SIL3、ISO 26262 ASIL D、CCRC EAL4+等国际权威认证。满足能源电力、石油化工、智能制造、工业控制、轨道交通、家居、穿戴等多场景需求；并推出智能卡操作系统、工业操作系统等行业版本。为企业提供打通异构设备及数据、复杂流程与业务应用的智能中枢，助力企业构建自主可控的数字化基座。

OneOS 立足工业操作系统，打造“操作系统+系统组件+核心板+场景化方案”的产品体系；基于工业控制产品的工业场景化方案能将工业操作系统价值最大化，赋能工业各个领域。OneOS 解决方案，满足客户复杂场景下的定制化需求，整合云、网、算力能力及生态合作资源，聚焦 AI，通过兼容主流芯片架构、融合端侧 AI 算法，联合硬件厂商与生态能力平台，共同构建从终端到云端的完整行业解决方案，加速物联网行业解决方案应用落地。

图 33：中移物联 OneOS

（图源：中移物联）

研华科技

研华科技成立于 1983 年，总部位于中国台湾，是全球领先的工业物联网、边缘计算解决方案提供商，专注于边缘计算、嵌入式系统与智能硬件的研发与整合，业务覆盖智能制造、智慧交通、能源管理等领域。2025 年研华科技开始战略转型，从“工业电脑领军企业”向“Edge AI 引领者”转变，即从提供行业硬件平台与软件工具，向边缘运算硬件与智能软件彻底融合的 AI Agent on Edge 方向发展，将 AI 硬件与软件深度绑定协同解决工业 AI 落地中的诸多技术瓶颈，在 OT 与 IT 的深度耦合基础上合力推动产业应用的全面智能化。

图 34：研华科技边缘解决方案

（图源：研华科技）

研华科技正在围绕工业智能体的核心技术进行布局，如开发 Edge AI 加速模块、Edge AI 产业应用系统、Edge AI 大型语言模型训练系统及 Edge AI 服务器等产品，并提供整合式 AI 软件平台工具 Edge AI SDK，协助产业客户评估验证 AI 平台效能及应用开发，同时与主流芯片厂商共同开发高效能边缘 AI 计算平台。研华科技正在通过 WISE-Edge 链接边缘端的软硬整合策略，打造工业智能体生态系统。

格创东智

格创东智成立于 2018 年，由 TCL 集团战略孵化，是中国领先的工业互联网与智能制造解决方案提供商，格创东智以软硬一体、场景化赋能为核心策略，构建起覆盖边缘计算、AI 视觉检测、工业大模型的完整技术体系。

格创东智自主研发东智边缘计算产品体系，包括 EdgeBox 工业网关，采用容器化架构

实现边缘应用快速部署。其推出天枢 AI 视觉检测系统，集成零代码模型训练平台与智能标注工具，可将检测模型开发周期从 3 个月缩短至 2 周。2025 年发布的章鱼 Agentic AI 平台，整合 TCL 集团星智 X-Intelligence 大模型，支持自主 Agent 开发，实现了智能知识推荐、知识库系统,报表智能生成、告警内容深度分析、智能派工等核心功能。

美的

美的集团成立于 1968 年，业务覆盖智能家居、楼宇科技、工业技术、机器人与自动化、数字化创新五大板块。目前美的正在构建 AI 智能体中台，负责调度与协调生产线上各类智能体的工作，美的的 AI 智能体已在工厂落地，如品质控制智能体能够自动匹配关键工艺参数，确保产品质量符合标准。目前，美的集团已在旗下工厂部署了 10 余个 AI 智能体，并计划将智能体功能拓展至更广泛的业务场景

针对工业制造应用场景接近 100%识别成功率的需求，美的研发了基于视觉大模型的通用强泛化高精度识别定位的技术，应用于焊接，装配等 5 类典型制造场景，完成 100 余条产线规模化复制。

海尔卡奥斯

海尔卡奥斯是海尔集团基于近 40 年制造经验，于 2017 年 4 月首创的以大规模定制为核心、引入用户全流程参与体验的工业互联网平台，专注于为行业提供数字化转型服务。卡奥斯自主研发的工业大模型 COSMO-GPT，在通用模型基础上融入丰富的工业场景数据，具备工业知识问答、工业代码生成和工业理解计算等专业能力，致力于打造成为工业 AI 的 “最强大脑”。

分阶段发展： 卡奥斯将工业互联网的发展划分为三个阶段：“工业 OS”“工业大脑” 和 “智能交互引擎”。目前，其正在逐步推进各阶段目标的实现，通过构建工业大脑，依托 “大数据 + 大模型”，实现生产决策的智能化，并进一步探索智能交互引擎，让设备、产线、车间、工厂、企业、园区具备自主学习和协同能力。

宝信软件

宝信软件成立于 1978 年，是中国宝武集团旗下核心工业软件与自动化解决方案提供商，深耕钢铁行业智慧制造领域，依托自主工业互联网平台 xIn³Plat，宝信软件构建 “云边端” 协同架构，形成覆盖工业大数据、AI 模型、智能装备的全栈技术体系，赋能全球钢铁企业及离散制造行业数字化转型。

自主研发全栈国产化 PLC（天行系列），实现从操作系统到编程环境的自主可控，支持毫秒级工业控制与数据交互。联合宝武集团发布 “宝联登钢铁行业大模型”，首创 “通专融合” 架构，覆盖高炉、转炉等关键工序，在硅钢研发中实现 “数据预测 + 实验验

证” 模式，研发周期缩短 30%；在热轧排程中通过 AI 优化，轧硬卷周转周期缩短 12%，年均增效超千万元。生产安全视觉 AI 平台集成 120 + 智能识别算法，隐患处置响应时间从小时级降至分钟级。

中控技术

中控技术成立于 1999 年，总部位于中国杭州，是中国流程工业自动化与数字化领域的领军企业，专注于为石化、化工、电力等行业提供工业软件、自动化控制系统及智能制造解决方案，连续多年位居国内集散控制系统（DCS）市场占有率首位。

中控技术 2024 年正式启动 “ALL IN AI” 战略，致力于全方位将人工智能技术融入工业生产流程，同年发布了流程工业首款 AI 时序大模型 TPT，该模型已在氯碱、热电、石化、乙烯等装置上取得突破性应用。2025 年中控技术将 TPT 与 DeepSeek 进行深度融合，推出流程工业首个 “时序智能 + 认知智能” 双引擎架构的 TPT 大模型升级版。此外，中控技术还积极布局机器人业务，2024 年作为第一大股东投资入股浙江人形机器人创新中心，并发布了首款全域自研人形机器人整机 “领航者 1 号” 和 “领航者 2 号 NAVIAI”，以及流程工业机器人解决方案 “Plantbot”。

赛意信息

赛意信息成立于 2005 年，深耕制造业数字化转型领域 20 年，为电子、家电、汽车、新能源等 23 个行业提供工业软件、智能制造解决方案及全栈信创服务。赛意信息以 “AI + 数智化综合解决方案” 为核心竞争力，旗下谷神工业互联网平台入选国家级 “双跨” 平台。

赛意信息以全栈信创能力驱动行业大模型落地为核心策略，构建 “国产算力 + 行业模型 + 工具链” 技术体系，推动制造场景从经验驱动向 AI 驱动跃迁，自主研发善谋 GPT 平台，构建 “通专融合” 行业大模型体系。针对 PCB 行业，赛意信息推出垂直大模型，针对光伏领域，构建电池丝网印刷工艺优化模型。

天工人工智能工业平台

天工人工智能工业平台”是在上海市经济和信息化委员会指导下，由上海市经信委、中国电信上海公司、上海市人工智能行业协会、上海库帕思科技有限公司以及人工智能产业链上下游企业共同发布的平台。该平台旨在通过集成先进的人工智能技术，降低工业企业使用人工智能的门槛，提高 AI 应用开发验证效率，助力企业实现智能化转型。

天工人工智能工业平台”自发布以来，在端侧 AI 结合智慧工业应用方面取得了显著进展。平台发布了工业语料库 1.0 版，支持对工业大模型进行针对性训练，为企业提供更精准的决策支持和解决方案。这一语料库涵盖了制造业各个环节的专业术语、工艺

流程、案例数据等海量内容，经过精心整理和标注，具有极高的准确性和实用性。平台还启动了 “模塑申城” 智能制造行业应用基地，聚焦人工智能创新技术、构建协同创新生态、加快应用场景落地。

羚数智能

上海羚数智能科技有限公司成立于 2021，是一家专注于工业人工智能大模型智能体应用的高科技创业公司，运用云原生、大数据、低代码等前沿技术，结合多年工业沉淀，自主研发了 Lead Agents AI 大模型智能体和 Lead Series 新一代自主工业软件系列等产品。

羚数智能旗下上海羚一人工智能科技的“百工大模型”是上海首个通过国家备案认证的工业垂类大模型。公司基于“书生·浦语”大模型，开发了多款 AI Agent 产品，如面向高端装备制造业核心场景的产销研一体化 AI Agent 等，已在数家国央企头部装备制造企业中落地应用。此外，羚数智能还构建了 Multi-Agent 系统，可通过自然语言交互协调设计、采购、生产等环节，提升项目调整效率。

发展趋势与挑战

1. 工业智能体驱动的智能化转型加速：工业智能体作为融合了人工智能、物联网、大数据和边缘计算等多种先进技术的智能系统，正逐渐成为智慧工业的核心 “大脑”。它能够理解自然语言指令，自主决策并控制物理设备完成复杂的工业任务，实现从传统自动化系统向新一代认知智能系统的跨越。从发展阶段来看，工业智能体目前正处于感知增强与工程突破阶段，不仅提升了关联分析能力，还在解决小数据、实时性、可解释性等难题方面取得了显著进展。未来，工业智能体将迈向认知提升阶段，构建全局性工业知识图谱，实现推理能力的协同，进一步提升其在复杂工业环境中的决策能力和适应能力。

2. AI 与工业的深度融合拓展应用边界：在生产制造环节，人工智能驱动的机器视觉系统能够实现高精度的产品检测和质量控制，有效提高产品合格率。在供应链管理方面，人工智能可以通过对市场需求、物流信息、库存数据等多源数据的分析，实现精准的需求预测和智能的库存管理，提高供应链的响应速度和灵活性。此外，生成式人工智能技术的发展也为工业设计带来了新的机遇。设计师可以利用生成式 AI 快速生成多种设计方案，并通过与工业智能体的交互，对设计方案进行优化和评估，大大缩短产品设计周期，提高创新能力。

3. 数据质量参差不齐工业模型迭代缓慢：智慧工业、工业智能体的价值实现高度依赖数据闭环，对数据的数量、质量、可得性和流动性要求极高。然而，在实际工业场景中，数据往往存在多源异构、噪声干扰、数据缺失等问题，影响决策准确性。此外工业环境复杂多变，存在大量的不确定性因素，这对工业智能体所依赖的模型精度和

适应性提出了很高的要求。传统的建模方式难以应对复杂的工业场景和突发工况变化，需要结合人工智能技术实现自适应模型更新。但目前在模型的实时更新、多尺度建模以及模型在复杂环境下的可靠性等方面仍存在技术难题。

2.4.3 智慧城市应用

定义与概述

端侧 AI 的应用有力推动了智慧城市中细分场景的智能化升级，涵盖智慧安防、交通管理、环境监测、能源管理等诸多领域。在智慧安防方面，端侧 AI 赋能摄像头等设备，可实时识别异常行为、可疑人员，实现快速预警和响应，提升城市公共安全水平。于交通管理而言，借助端侧 AI，交通摄像头能精准识别交通违法行为，还能实时监测路况，自动调控信号灯，缓解拥堵。在环境监测领域，各类传感器借助端侧 AI 实现对空气污染、水质等数据的实时采集与分析，及时发现环境问题并采取应对措施。在能源管理上，端侧 AI 可实时监测能源使用情况，优化能源分配，提高能源利用效率。

端侧 AI 与智慧城市的结合，提升城市运行效率，通过对各类数据的实时处理和智能分析，实现资源的合理配置与服务的高效供给；同时增强城市管理的精细化程度，能够及时发现问题并做出精准决策。随着技术的不断进步，端侧 AI 芯片性能将不断提升，相关算法也将更加优化，为端侧 AI 在智慧城市中的深入应用提供有力支撑。另一方面，端云协同的混合 AI 架构逐渐成为主流，端侧 AI 将与云端 AI 优势互补，共同推动智慧城市的发展。同时，5G、物联网等技术的普及，为端侧 AI 设备的广泛部署和互联互通创造了有利条件，将进一步拓展其在智慧城市中的应用空间。

市场规模与概况

根据《2020-2025 年全球及中国智慧城市行业市场现状调研及发展前景分析报告》数据，按投资价值计，中国智慧城市市场规模由 2020 年的 15 万亿元增长至 2024 年的 36.8 万亿元，复合年增长率为 25.2%。预计 2025 年中国智慧城市市场规模将达到 45.3 万亿元。

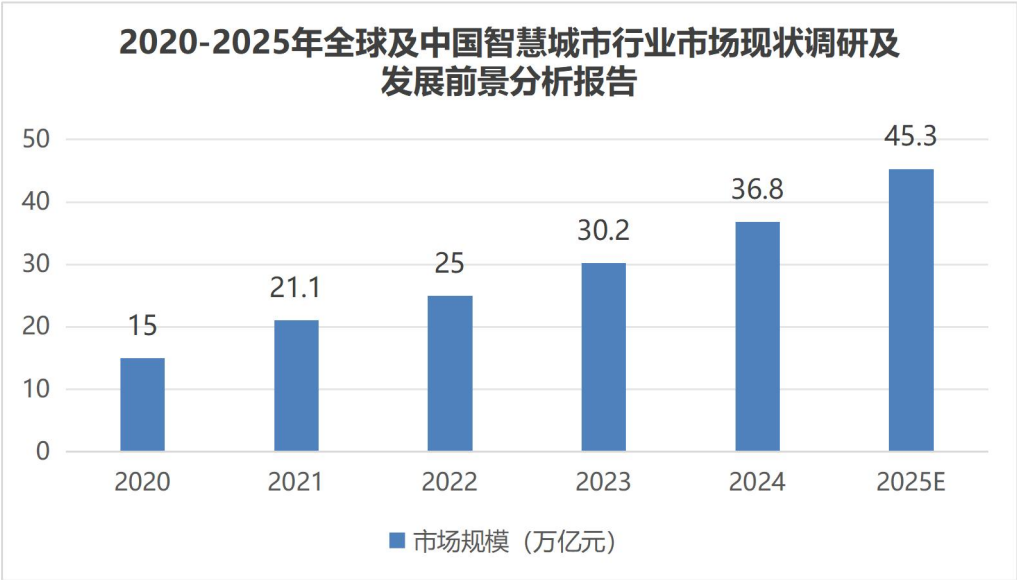

图 35：智慧城市市场规模

（数据来源：中商产业研究院，智次方制图）

在全球城市化进程持续加速与信息技术迅猛发展的双重推动下，智慧城市建设已成为全球城市发展的核心趋势，蕴含着巨大的市场潜力与发展机遇。

主要企业与方案

中移物联 OnePark

中移物联 OnePark 是中国移动面向全场景园区管理推出的智能化综合解决方案，以智慧园区数字化底座为核心，融合 5G 专网、物联网、AI 大模型等技术，构建覆盖园区通行、安防、运营、招商、能源、服务的全场景管理体系。聚焦商办园区、白酒烟草生产园区、仓储物流园区、化工园区、政企培训园区、社区六大领域，通过标准化产品与定制化服务适配园区多类客户需求，实现园区管理智能化、运营高效化、服务精准化，助力客户降本增效与数字化转型。2024 年 OnePark 荣获“光华杯”全国总决赛一等奖、第七届“绽放杯”最佳国际化应用奖等数十项荣誉，并连续三年被权威机构评为智慧园区行业 TOP3。

OnePark 智慧园区聚焦人工智能和双碳新能源发展趋势，重点发展公租房、共享充电、社区治理等主力产品，并积极推动园区智服、信息发布等潜力产品。同时，将开放标品平台接口，提升一级平台能力，并构建园区 AI 三件套“智模、智用、智体”，推动智慧园区产业升级。

针对智慧园区标准产品、小场景和高价值 ICT 项目，协同推进三大业务板块，实现差异化规模发展。标准产品灵活组合，依托商客等渠道批量销售；小场景方案以模块化能力服务中小园区，轻量化版本实现快速交付；高价值 ICT 项目整合自有能力与生态

应用，聚焦工业、化工等重点园区，复制标杆案例。

图 36：OnePark 产品体系

（图源：中移物联 OnePark）

OneNET 城市物联网平台

中移物联以探索数字化城市治理为推进方向，以城市底座为基础，构建“城市底座+行业 aPaaS”模式，在城市物联网平台的底座基础上，升级构建“1 + 3”AI 感知网产品体系。“1”是指城市感知智能通，作为核心引擎与统一入口，构建动态城市知识图谱，精准解答复杂感知问题，“3”是指提供事件洞察智能体、健康度诊断智能体、运维排障智能体等三大核心能力，覆盖城市治理、设备管理、健康评估及运维排障全链条，打造“感知－分析－决策－行动”一体化智能中枢。以全新产品体系建设平台运营内外循环，内循环推动资产沉淀，外循环推进价值变现，提供成熟的售前、产研、交付、运维一体化支撑体系，综合推动城市治理效率提升，实现智能化和精准化决策，获评 IDC 中国城市物联感知平台综合能力、市场份额排名双第一。

图 37：OneNET 智慧城市

（图源：中移物联）

海康

海康威视成立于 2001 年，构建了以物联感知、人工智能、大数据为核心的智能物联技术体系，为千行百业提供安防和场景数字化产品与服务。其产品和解决方案应用于公安、交通、司法、文教卫、金融、能源等行业，助力客户实现智能化升级与数字化转型。

在端侧 AI 结合智慧城市应用方面，海康威视依托海康观澜大模型技术体系，将大模型能力直接部署至端侧，推出一系列视觉大模型摄像机。这些摄像机不仅成像画质更优，还突破了场景普适性弱、复杂目标识别难等瓶颈，进一步提升目标检出率，大幅降低误报。海康威视还构建了完善的大模型部署技术体系，支持低成本硬件承载大模型能力。例如，在郑州经开区等地部署大量视频监控前端，通过大数据和人工智能技术，实现对城市的全方位实时监控，自动识别如违规摆摊、垃圾堆积等问题，提升管理效率。

大华

浙江大华技术股份有限公司成立于 2001 年，是一家全球领先的以视频为核心的智慧物联解决方案提供商和运营服务商。大华聚焦于城市和企业两大业务战略，围绕以视频为核心的智慧物联解决方案，通过持续的技术创新和产品迭代，为全球用户提供更加精准、高效的解决方案和产品。

2025 年，大华推出大华星汉大模型，其端侧设备通过小模型可检测到“人员闯入”等行为，并将分析结果上报云端，云端大模型对上报结果进行再次判断，确认事件真实性后生成精准的事件摘要，并通知管理人员，降低 90%以上的事件处理成本。此外，大华还构建了完善的大模型部署技术体系，通过模型轻量化技术，使边缘端算力需求下降，支持低成本硬件承载大模型能力，推动端侧 AI 在智慧城市中的广泛应用。

宇视科技

宇视科技是全球 AIoT 产品、解决方案与全栈式能力提供商，以“ABCI”（AI 人工智能、BigData 大数据、Cloud 云计算、IoT 物联网）技术为核心，自 2011 年成立以来发展迅速。

宇视科技行业大模型“梧桐”，经过迭代，实现了大模型软件硬件化、硬件装备化、装备序列化，其能力不仅在图像质量上实现突破，推出了第二代 AI-ISP 图像处理引擎猎光 2.0，还完成了长尾算法开发效率的指数级提升。基于“梧桐”大模型，宇视科技在智慧城市等多领域实现了规模部署，如在锡林浩特市上线“AI 智慧城管”。今年宇视创新首发的 Agent Link 智能体算力链技术，突破性实现云边端算力动态协同，配合“关山湖”系列边端智能体装备，构筑起工程化落地的技术闭环。

萤石

萤石网络以“2+5+N”战略为框架，以 AI 和物联网云平台为核心驱动力，构建了覆盖智能家居摄像机、智能入户、智能服务机器人、智能穿戴及智能控制五大 AI 交互类核心产品线，并通过生态控制器接入多种子系统生态，形成完整的 AIoT 生态体系。

公司自主研发的“萤石蓝海大模型”采用 MoE 架构，支持端云协同、具身智能和跨模态交互，显著提升了智能摄像机的交互体验和机器人自主决策水平。其推出的视频通话摄像机 S10 和可移动宠物看护摄像机 TAMO，通过 AI 算法实现人脸追踪、异常行为预警等创新功能。此外，萤石网络在智能家居摄像机领域持续创新，推动产品向视觉化、场景化、交互化、智能化升级，发布了多款创新产品，如双摄、黑光、4G 相机等。

商米科技

上海商米科技集团股份有限公司成立于 2013 年，是一家致力于为商用领域提供智能 IoT 硬件及软硬结合的数字化解决方案的物联网科技公司。通过搭建设备互联互通基础架构与智能设备管控体系，不断增强设备感知与连接能力，开发出系列 IoT 产品，并形成了商米 IoT 云服务平台，以开放方式践行技术普惠理念，携手全球各行业开发者及合作伙伴，形成面向商业全场景的数字化解决方案。

商米科技通过其智能商用硬件设备搭载 AI 算法，如支持金融级别的人脸支付产品等，实现支付、点单、收银等场景的智能化。其 BIOT 战略中的端侧设备，如智能 POS 机、收银机等，可实时收集商业数据并进行初步处理，再通过商米云 OS 将数据传输至云端进行深度分析，帮助商家实现精准营销和运营管理，推动线下商业的数字化转型，助力智慧城市建设中的商业智能化发展。

卓视智通

卓视智通专注于计算机视觉及数字孪生技术的原创研发，致力于为智慧交通和安全垂直行业提供解决方案。公司以 5G+AI 技术为基础，自主研发了车脸识别、交通视频融合感知及车路协同等产品，并在智慧公路、城市交管、车路云一体化等领域实现广泛应用。

卓视智通推出“智通卓识”多模态大模型，通过端云协同实现道路、人、车、环境的精准融合感知识别。其边缘端设备基于国产化芯片研发，具备视频图像的智能识别、二次分析及设备智能运维能力，支持图搜万物功能，可自动生成违法数据报告，为执法取证提供智能化支持。此外，卓视智通还构建了交管智能化全场景解决方案，推动智慧交管领域的应用创新，助力智慧城市建设。

海纳云

海纳云是海尔集团旗下的数字城市物联科技平台，依托星海数字平台，海纳云打造了城市大脑、数字市政、数字应急与城市生命线、数字孪生 BIM/CIM、数字城市治理、数字社区/园区和数字安全等多个应用场景，赋能城市生活、经济和治理的数字化转型，已成长为数字城市、新城建、“AIoT”领域的优秀企业。

海纳云以 AIoT、大数据、AI 算法和数字孪生技术为核心，构建了数字城市底座能力。其打造的青岛市安全风险综合监测预警平台接入了 1.6 万余台智能感知设备，汇聚了 9 个部门 176 项数据类目及 40 余种应急安全算法，覆盖燃气、供水、排水等 9 大专题场景，助力青岛构建“全域感知、实时监测、分级响应、高效处置、动态发布”的智慧应急体系。此外，海纳云还推出了智慧桥梁管理平台等产品，通过智能监测和预警提升桥梁养护效率和安全性，相关项目已在多个城市落地并形成样板。

力维智联

深圳力维智联技术有限公司成立于 2005 年，前身为中兴通讯子公司 “中兴力维”，是国家高新技术企业与全球领先的 AIoT 产品及解决方案提供商。公司深耕智能运维与数智转型领域，以 “泛在连接、数据智能” 为核心技术，构建了 Atlantis 机器视觉平台与 Ganges 泛在连接与智能协同平台，为城市治理、工业智能、公共安全等领域提供软硬一体的行业解决方案。

力维智联于 2025 年申请基于大语言模型的机房空调节能专利，通过端侧 AI 实现空调控制策略的动态优化，在深圳某数据中心试点中降低能耗 15%。生态协同上，公司与华为昇腾、百度智能云等合作，推动端云协同架构落地，如在上海 “一网统管” 项目中，边缘节点与云端训练形成数据闭环，模型迭代效率提升 60%。政策驱动下，力维智联深度参与 “东数西算” 工程，在西部枢纽节点部署边缘算力中心，支撑智慧城市中交通流量预测、能耗管理等实时应用。

凌华科技

凌华科技成立于 1995 年，总部位于中国台湾，是全球领先的边缘计算解决方案提供商，在端侧 AI 结合智慧城市应用上，凌华科技展现出显著的技术布局与场景落地能力。其推出的 DLAP-211-Orin 和 DLAP-411-Orin 边缘 AI 平台搭载 NVIDIA Jetson Orin 模块，AI 性能较前代提升 8 倍，达 275 TOPS，支持智能停车、交通管理等场景。

此外，凌华科技与致星科技合作推出边缘联邦学习一体机，应用于金融、医疗等注重隐私的领域，对比传统 CPU 架构性能提升 7 倍，功耗降低 40%。其自主研发的 EVA SDK 边缘视觉分析软件，可帮助用户在两周内完成 PoC 项目部署。

发展趋势与挑战

1. 数字化与智能化转型加速：城市对数字技术的重视程度不断提升，数据资源成为核心要素，将服务于城市治理与经济发展。如从城市治理、民生服务到政务服务等领域，智慧城市应用不断拓展。例如，智慧交通、智慧医疗、智慧物流等细分领域应用场景日益丰富，端侧 AI、5G、物联网、大数据等技术与这些领域的深度融合，催生了大量新的应用场景和解决方案。

2. 智能平台与算力中心建设增强支撑：城市大脑、城市智能平台、城市运管中心等智能平台建设加快，各地着力打造城市数据汇聚中心、交换中心，逐步将智能应用从城市管理、服务向经济发展等领域拓展。同时，构建完善的算力调度机制，助力构建柔性、智能城市算力中心。

3. 智慧城市运营成为重点：随着智慧城市数字基础设施的逐步完善以及多元应用场景需求的升级，针对城市物理空间与数字空间的规划、设计、管理、运维、运营、安全等全流程全领域的服务成为重点，结合端侧 AI 技术对数据资源化、资产化、场景化等内容的智慧城市运营价值凸显。

三、总结与展望

总结与展望 1：端侧 AI 到具身智能体

端侧 AI 硬件的未来，不是更强的功能，而是“具身智能体”

AI 的未来，不是更强的问答，而是更自然的陪伴。为了实现这样的“陪伴式智能”，技术能力只是基础，真正的突破在于人机关系的重新设计。

这种转变在设计语言上尤为关键：

- 从“科技感”转向“生活感”与“温度感”；
- 从“功能性”转向“存在感”与“情感亲和力”；
- 从“用户操作”转向“设备主动理解与协作”。

我们可以将这种新型设备称为一种“具身智能体”（Embodied Intelligence Agent）——它具备以下三大核心特征：

1. 情感接口：通过语音、语调、节奏、语言内容与语义理解，构建“有温度”的交流；
2. 感知能力：设备能理解用户所处的环境、情绪、需求与历史行为；
3. 行为自主：不仅被动响应，更能主动协助、提醒、规划，成为“智能体”而非“工具箱”。

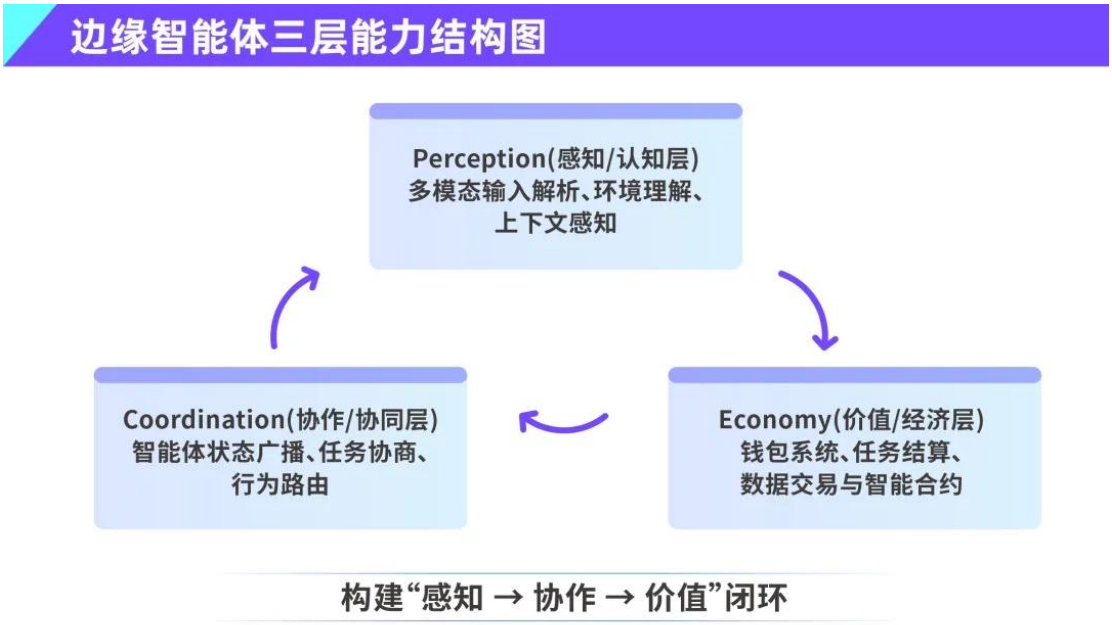

图 38：智能体能力结构

（图源：物联网智库）

这标志着 AI 硬件从“被动工具”进化为“主动伙伴”，从“智能终端”演化为“日常智能体”。在这一进程中，硬件不再是模型的载体，而是人格化的延伸。AI 不再只回应你的问题，而是与你共处、共感、共生。

当 AI 模型拥有了“身体”，AI 将不再只是“技术”，而是你日常生活中的“存在”。

总结与展望 2：揭示垂类模型进化路径

大模型的下半场是垂类模型的主场，揭示垂类模型从工具到平台的跃迁的 4 阶段进化路径

边缘智能的飞速发展正在倒逼 AI 模型“下沉”到端和边，而垂类模型更适合在资源受限的边缘环境中高效运行。垂类大模型的引入，提升了端侧智能的“智能等级”，让端侧设备不仅能“感知世界”，还能“理解场景”，推动端侧 AI 从“感知”向“认知”跃迁。

垂直 AI 模型的演进，本质上是一个从“解决单一任务”到“辅助整个流程”，再到“承载行

业生态平台"的能力跃迁过程。

通过对产业先行者的实践路径进行梳理，我们可以将其归纳为四个典型的发展阶段：

图 39：垂类模型演进

（图源：物联网智库）

阶段一：垂直切入，解决刚需痛点

在初始阶段，企业需要聚焦于一个高价值、高频次、数据结构化程度适中的垂类场景，率先突破 AI 的实用性边界。这些场景通常具有明确的痛点需求和可衡量的价值回

报，如制造业中的缺陷检测和良率预测等。

在选择切入场景时，除了考虑商业潜力和 AI 可行性外，还要重点评估数据的可用性和质量。一个理想的垂类模型场景，应该具备相对完整、标注充分的数据积累。这往往需要企业在特定领域有深厚的行业积淀和数字化基础。因此，与行业头部企业合作，或选择已经实现良好数字化的细分场景切入，往往是明智的选择。

这一阶段的关键词是：痛点明确、商业价值可衡量、数据可获得、流程可闭环。只有同时满足这四个条件，垂类模型项目才有可能在起步阶段取得突破。

阶段二：构建能力飞轮，形成垂类护城河

当垂类 AI 模型在某个具体任务上实现稳定的、可靠的表现后，它就具备了形成“能力飞轮”的基础。这个飞轮的运转逻辑是：首先，随着模型的不断优化，其准确性、效率等核心性能指标不断提升，用户使用体验越来越好；好的体验吸引更多客户接受和使用该 AI 系统，从而在实际业务中产生更多真实数据；更多的数据反过来又可以用于持续训练和优化模型，从而进一步提升模型性能和用户体验，形成正向循环。

要驱动这一飞轮，关键是构建“模型能力”与“产品体验”的双引擎。

在“模型能力”方面，垂类模型企业基于不断积累的真实业务数据，对模型进行持续的微调和优化，针对行业特定任务、知识、语言和规则定制算法，最终形成高度可靠、性能卓越的行业专属模型（Domain-Specific Foundation Model）。

在“产品体验”方面，仅有强大的模型是不够的，还需要从产品层面，围绕行业用户的真实需求和使用习惯，精心设计人机交互、任务流程和系统功能。很多时候用户的需求有待进一步澄清，Agent 的优势在于它可以跟用户进行多轮对话交互，理解用户的真实意图，将用户的高层指令转化为可执行的具体任务，完成端到端成果交付。

阶段三：流程重构，迈向业务成果即服务（BOaaS）

当 Agent 掌握了理解用户需求、调度算法模型、协同多方资源、开展端到端任务交付的能力后，垂类 AI 企业就站上了从“工具”到“平台”跨越的台阶。

这一阶段的核心特征，是从“提供模型”转向“交付服务”：企业不再把 AI 视为单点的功能工具，而是以之为杠杆，撬动行业流程的全面重构，最终实现业务成果即服务（BOaaS）的阶跃。

在 BOaaS 模式下，企业交付给客户的，不再是一套解决方案，而是一个“端到端的服务承诺”：客户只需输入目标和约束，智能系统就可以自动调度算法、数据、知识等数字资源，完成整个业务流程，交付客户所期望的结果。

对客户而言，他们所购买的不再是一个“死”的软件系统，而是一种“活”的智能服务，一种随需应变、持续优化、快速响应的业务能力。这就是 BOaaS 的本质：业务流程的全栈智能化，价值交付的服务化与柔性化。

当机器可以自动执行 80%的流程性任务时，人的角色就从“流程的执行者”转变为“流程的设计优化者”。

当然，BOaaS 绝非一蹴而就，它对产业智能化的深度和广度提出了极高要求。单点技术、单点场景的突破还远远不够，必须通过持续的技术创新和场景扩展，打造一张覆盖业务全流程、全要素的“智能化地图”。

阶段四：平台化演化，占据行业的流程控制点

当越来越多的客户开始习惯于通过 AI Agent 完成各项任务，当垂类模型开始掌控行业中最关键的业务流程时，垂类模型就迎来了从“应用”到“平台”的最后一次跃迁。

当然“平台化”并非所有垂类模型企业的必由之路，许多企业可能会选择专注于某个细分领域，成为该领域的“小而美”的服务提供商。能否最终完成平台化转型，取决于企业的战略定力、技术实力、行业理解以及生态运作能力。

这条路径的本质，并不是做一个更强的模型，而是通过 Agent 能力，逐步重构行业流程，并最终赢得稳固的生态位。

总结与展望 3：端侧模型带动智能硬件爆发

国产端侧模型带动智能硬件爆发，助力端侧 AI 生态重构与产业链协同

回望 2022 年，整个智能硬件行业颓势尽显，相关厂商哀鸿遍野，而后，生成式 AI 和大模型的爆火为智能硬件发展注入了新的生机。AI 耳机的热销可以看作整个智能硬件行业回暖的标志性风向之一，随着终端侧软硬件条件的成熟，不只耳机品类，包括智能眼镜、智能水杯、智能手环等更多硬件类产品近半年来都在海外取得了不错的销量。

随着 DeepSeek 问世，优质端侧模型的面世驱动着端侧 AI 的爆发，为整个智能硬件行业带来了更大的想象空间。对终端企业而言，接入优质端侧模型是用终端 AI 为用户切实带来价值的重要策略，从而在硬件这条同质化竞争严重的赛道中塑造差异化的竞争优势。

未来，预期还会有更多硬件企业加速接入或开发优质端侧模型、垂类模型。如何通过 AI 直击消费者的需求痛点，同时实现技术与商业的平衡，是制胜市场的关键。

此外，端侧模型的出现推动了端侧 AI 生态重构与产业链协同。过去几年，智能硬件之所以卖得不好，是因为没有回答好一个核心问题，即“凭什么让消费者为你买单”？AI 的确是有用的，但是如果有用的“代价”是高昂的溢价，那么依然难以取得商业上的成功。

对硬件厂商而言，技术与商业平衡的艺术在于用更低的成本为消费者带来更好的智能体验，从这个维度来说，优质端侧模型通过技术创新和生态优化，降低了硬件厂商在端侧部署高性能 AI 的门槛。

首先，优质端侧模型训练与推理成本大幅压缩，在保持性能的同时降低终端硬件资源消耗和算力需求。

第二，硬件兼容性与国产化适配是国产优质端侧模型助力硬件厂商降低成本的另一大利器。

第三，产业链协同与规模化效应未来会进一步摊薄端侧 AI 的边际成本。

从芯片到模组厂商的适配，从算力到云服务再到应用，整个产业链的协同工作正在推进，某种意义上，优质端侧模型的出现就如同提供了一个交点，让通向端侧智能未来的各个环节都能够更好地互联互通。而因协同带来的规模化效益，未来将进一步摊薄端侧 AI 的边际成本。

总结与展望 4：端侧 AI 商业价值

端侧 AI 商业价值新的评估维度和指标正在形成

对于端侧 AI，传统的 AI 商业价值评估模型已不再适用，新的评估维度和指标正在形成。

以下是一个端侧 AI 的商业价值重估矩阵，对比了传统模型和新一代端侧 AI 在不同维度上的差异：

图 40：端侧 AI 商业价值重估矩阵

（图源：物联网智库）

- 核心指标：从“绝对性能”到“效能密度”

传统 AI 模型追求绝对性能，如准确率，但端侧 AI 更关注效能密度，即在有限功耗下实现更高的性能。每瓦准确率成为衡量端侧 AI 商业价值的新指标。

- 价值锚点：从“规模崇拜”到“效率革命”

过去，AI 模型的价值锚定在参数量上，认为更大的模型必然带来更强的性能。但端侧 AI 时代，推理能效比成为新的价值锚点，即在有限算力下实现更高效的推理。

- 竞争壁垒：从“数据垄断”到“架构创新”

传统 AI 竞争的壁垒在于数据规模，掌握更多数据的企业往往占据优势。但在端侧 AI 领域，架构创新成为新的竞争壁垒，更高效、更智能的计算架构将决定企业的市场地位。

- 商业模式：从“软件服务”到“硬软融合”

传统 AI 商业模式以云端 API 调用为主，用户按使用量付费。但端侧 AI 时代，硬件和服务的融合成为新的商业模式。企业通过销售智能硬件，并提供订阅服务，实现持续的收入。

端侧 AI 的 ROI 优化路径包括：利用低成本边缘计算芯片，减少对昂贵数据中心资源的依赖；优化推理效率，提高能效比，降低 AI 计算在终端设备上的成本；结合智能存储、通信技术，提升数据利用率，减少冗余计算。

过去，AI 投资主要围绕提升模型能力展开，追求更大的参数规模、更复杂的神经网络架构。然而，计算成本与商业收益的平衡正在成为新时期 AI 投资的核心考量因素。

未来，AI 投资的关键，将会从“更强的 AI”，到“更高效的 AI”；从“单纯软件创新”，到“软硬结合”。AIoT 芯片、边缘设备、优化算法的发展，将重新定义大模型的商业价值。

因此 AI 的商业价值将不再由单纯的模型能力决定，而是由计算成本与商业收益的平衡

来定义。只有那些能够在算力、功耗、存储、通信等多个维度平衡商业价值的 AI 架构，才能真正实现可持续增长。端侧 AI 的崛起，将推动整个产业走向更加务实、可持续的发展之路。

结语

端侧 AI 的崛起，正深刻地重塑着人工智能技术的应用版图和发展轨迹。通过将智能直接赋予终端设备，端侧 AI 模型赋能下的终端设备、场景应用在实时性、隐私保护、低网络依赖性以及个性化体验方面展现出无可比拟的优势。这些核心价值使其成为推动“芯模端智一体化”——即端侧芯片、端侧模型、智能终端与智慧应用的深度融合——这一新范式的关键引擎。