

# 新一代智能终端蓝皮书

开启万物智联、协同智慧新纪元

(2024 年)

中国信息通信研究院

2024年12月

---

## 版权声明

---

本蓝皮书版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本蓝皮书文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

---

## 更名声明

---

原“集智”白皮书更名为“集智”蓝皮书。“集智”蓝皮书将继续秉承原有的编撰理念和高质量标准，致力于提供有价值的信息和洞见。



## 前 言

当前，全球新一轮科技革命和产业变革正深入发展，信息通信技术创新已从过去单点突破转变为多技术协同推进及群体性演变。人工智能（AI, Artificial Intelligence）、物联网、卫星通信和大数据等关键技术不断突破，加速终端行业创新进程，新一代智能终端逐渐走深向实，推动人类社会进入智慧化时代。

新一代智能终端指基于信息通信技术，以强感知、强计算、强交互、强体验提升为目标，能够执行多元化复杂任务，为用户提供强智能服务的新型智能终端。新一代智能终端扩展了智能终端的功能和边界，支持人类对物理世界更深刻的感知与理解，更加强调系统整体的智能互联和智能化管理，旨在通过更全面、综合的智能化应用，促进人、机、物、境的深度互联融合，并成为推动数字化转型和智慧生活的新质生产力工具。

本报告旨在通过深入剖析新一代智能终端的发展现状、技术特点、产业生态、市场前景及对未来的憧憬展望，为相关领域的决策者、研究者及从业者提供全面而深入的洞察与参考。报告编制过程中得到了业内相关企业的大力支持，在此特别感谢华为终端有限公司、荣耀终端有限公司、维沃移动通信有限公司、OPPO 广东移动通信有限公司、联通华盛通信有限公司、腾讯科技（深圳）有限公司、北京邮电大学。由于智能终端相关技术仍处于快速变化发展阶段，我们对新一代智能终端的认识还有待持续深化，报告中如存在不足之处，敬请谅解。

# 目 录

|                             |    |
|-----------------------------|----|
| 一、智能终端整体发展脉络.....           | 1  |
| 二、新一代智能终端总体发展情况.....        | 5  |
| （一）呈现“四新”趋势特征，人机物境深度融合..... | 5  |
| （二）产品应用不断涌现，重塑终端智能边界.....   | 14 |
| （三）扩展市场增长空间，提振行业发展信心.....   | 24 |
| 三、新一代智能终端技术创新重点.....        | 27 |
| （一）终端硬件创新升级，助推端侧性能加速提升..... | 27 |
| （二）平台软件加速重构，端云协同成为主要路径..... | 34 |
| （三）多重技术综合应用，催生人机交互全新范式..... | 39 |
| （四）安全防护策略演进，筑牢用户安全隐私屏障..... | 42 |
| 四、新一代智能终端产业应用生态.....        | 44 |
| （一）产业结构优化调整，形成多方协同生态.....   | 44 |
| （二）应用创新持续深化，开发逻辑逐步迁移.....   | 45 |
| （三）多元流量入口并存，构建全新市场格局.....   | 47 |
| （四）智能分级纵深推进，引领产业高质发展.....   | 48 |
| 五、新一代智能终端发展建议.....          | 50 |
| （一）创新软硬核心技术，加速终端智慧升级.....   | 50 |
| （二）健全智能终端标准，推动标准落地实施.....   | 50 |
| （三）壮大终端产业主体，培育繁荣共生生态.....   | 51 |

图 目 录

图 1 智能终端发展历程..... 1

图 2 AI 智能体架构图..... 12

图 3 全球 AI PC 市场规模及预测..... 27

图 4 新一代智能终端产业链各方角色..... 45

表 目 录

表 1 全球手机代表产品发布情况汇总..... 15

表 2 全球 AI PC 代表产品发布情况汇总..... 17

表 3 AI 赋能的可穿戴设备发布情况汇总..... 19

表 4 全球人形机器人代表产品情况汇总..... 21

表 5 全球 AI 手机市场规模及预测..... 26

表 6 部分旗舰级芯片性能表现..... 28

表 7 头部手机厂商操作系统优化情况..... 36

表 8 端云协同应用场景划分..... 39

## 一、智能终端整体发展脉络

智能终端作为驱动数字生产方式质效变革的核心引擎、数字化转型的重要工具，深度融合了硬件制造、操作系统研发、中间件设计、应用软件创新及内容创作等多个核心环节，通过高度集成前沿技术与多元化功能，能够为用户提供更加高效、便捷、智能的使用体验，深刻改变了人们的生活方式和模式，成为推动全球迈向数字化、智慧化时代的关键力量。随着技术的持续革新及市场需求的不断变化，智能终端设备将会在未来发挥更加重要的作用，为人类和社会发展带来更多的便利和价值。

根据技术发展路线、应用服务智能化水平等要素，智能终端演进历程可概括为以下四个关键阶段。



来源：中国信息通信研究院

图 1 智能终端发展历程

一是计算智能起步期（1995 年—2007 年），终端初步具备基础的数据计算和传输能力，硬件和软件技术尚在发展初期，应用生态处于早期萌芽阶段。此阶段，智能终端的智能化水平尚处于初级阶

段，终端的功能和应用较为基础，主要以计算能力的提升为主要发展方向，市场上的终端产品种类较少、应用类型较单一，应用生态处于逐步探索阶段。1996 年，Palm 公司发布初代 PalmPilot 1000，开创了掌上电脑（PDA）的新纪元，PDA 设备逐渐演化为后续的智能手机形态，推动了移动智能终端的市场拓展。1997 年，IBM 公司推出的“深蓝”（Deep Blue）超级电脑，成为历史上第一个在标准国际象棋比赛中成功击败卫冕世界冠军的计算机系统，其所采用的并行计算、混合算法等技术对后续的人工智能和计算机技术的发展产生了深远的影响。2000 年全球第一款 Symbian 操作系统手机发布，融合了 PDA 与手机的功能，并允许用户安装应用程序，奠定了后续智能手机的发展基础。

二是应用智能导入期（2007 年—2011 年），iOS、Android 及 Windows Phone 等移动智能操作系统正式问世，应用程序被广泛引入终端设备以提供多类型服务，智能应用生态初步构建。2007 年，苹果公司发布第一代 iPhone，搭载 iOS 1.0 系统，以其独特的多点触控技术和简洁直观的用户界面交互方式，彻底改变了人们对智能终端的认知，引领了手机设计的方向。2008 年 10 月，首款搭载 Android 系统的设备 HTC G1 上市，标志着 Android 系统正式进入市场。iOS 和 Android 两大移动智能操作系统成为主导智能终端市场发展的核心力量，推动智能终端市场竞争新格局的形成。2010 年，微软发布 Windows Phone 操作系统，期望与桌面操作系统形成互补，

完善计算设备领域的生态布局。此阶段，3G 网络开始普及，移动网络数据连接为智能终端的智能化应用发展提供了更为广阔的空间，各种应用程序如雨后春笋般涌现，应用商店模式逐步兴起，智能终端应用生态初步构建。随着应用程序的丰富和硬件性能的增强，智能终端的用户体验得到了显著提升，用户可以通过智能终端完成更多类型的任务和服务，如社交、购物、娱乐、出行等。

**三是单点智能普及期（2011 年—2020 年），新一代信息技术驱动终端软硬件升级变革，应用生态进入多元化发展阶段，万物互联迎来新机遇。**随着移动互联网技术的不断突破、云计算基础设施的逐步完善、物联网应用场景的持续拓宽，智能终端产业快速发展，掀起了移动互联网的创新浪潮，缔造了有史以来最具规模的终端产业生态。2011 年，苹果发布 iPhone 4S，首次引入语音助手 Siri，改变了用户与终端产品的传统交互方式。2017 年起，苹果、海思、高通、联发科等主要芯片厂商相继发布支持 AI 加速功能的新一代芯片，支持终端处理更复杂的任务。同时，4G 网络在全球范围内得到普及，为应用落地提供了高速、稳定的网络连接，推动了移动互联网应用的爆发式增长。此阶段，以智能手表、智能眼镜、智能手环等为代表的可穿戴设备受到消费者的广泛关注，智能门锁、智能摄像头、智能温控系统等智能家居和物联网终端日趋普及。随着智能终端不断拓展其应用边界并逐步与各类物联网设备进行深度融合，实现了设备间的无缝连接与高效协同，为用户提供了更加便捷、智

能的生活体验，推动应用生态向更加多元、丰富的方向发展。

四是协同智慧发展期（2020 年以来），人工智能、物联网等技术取得关键性突破，终端交互方式及服务方式发生深度革新，推动智能终端进入万物智联、协同智慧新纪元。万物互联注重设备之间的连接和数据传输，但智能化程度相对较低，无法满足人们对智能化、个性化服务的新需求。万物智联、协同智慧是在万物互联的基础上，通过融合人工智能等先进技术，实现设备与设备之间更高层次的智能连接和数据交互，更加强调设备与设备之间的智能协同和智能决策。人工智能、物联网、5G 通信、大数据等信息通信技术的快速发展与融合应用，为智能终端产业提供了强大的技术支撑和创新动力，促使智能终端设备在功能丰富性、性能卓越性以及用户体验友好性等方面实现显著提升，能够更好地满足人类对智慧化生活的追求。2020 年，华为发布面向家庭和企业场景的“全场景智能连接解决方案”。2022 年 11 月，OpenAI 发布基于 Transformer 神经网络架构的聊天机器人 ChatGPT。ChatGPT 展现出强大的语言理解及生成能力，再次引发全球对人工智能的关注，全面推动生成式 AI 技术的寒武纪式爆发。2023 到 2024 年，新一代智能终端创新产品及应用相继涌现，华为、OPPO、vivo、荣耀、小米、苹果、微软等国内外终端厂商加速升级 AI 战略，发布搭载大模型能力的 AI 手机、AI PC 等，AI Pin、Rabbit R1、智能戒指等新形态产品不断推出，具有感知和交互能力的人形机器人快速发展，商业化进程大幅提速。

随着深度学习技术的突破和算力资源的指数级增长，AI 大模型技术取得了革命性进展，以前所未有的规模和性能重新定义人工智能的边界，为智能终端提供了更强大的智能处理能力，终端交互方式、服务方式等都发生了深度变革，使其能够更好地适应各种复杂场景和应用需求，新一代智能终端新特征逐渐凸显。同时，随着物联网技术的发展，各种智能终端设备能够相互连接、相互协同，实现了全场景、全区域的智慧化连接。智能终端迈向“万物智联、协同智慧”全新发展阶段。

## 二、新一代智能终端总体发展情况

### （一）呈现“四新”趋势特征，人机物境深度融合

智能终端指能够接入通信网，搭载能够提供应用程序开发接口的开放操作系统，并能够安装和运行第三方应用程序的终端。新一代智能终端指基于信息通信技术，以强感知、强计算、强交互、强体验提升为目标，能够执行多元化复杂任务，为用户提供强智能服务的新型智能终端。传统智能终端侧重于单一设备或系统内部的功能增强与智能化，主要依赖于预设的规则和算法进行自动决策和执行，用户与系统的交互层次相对有限。新一代智能终端扩展了智能终端的功能和边界，支持人类对物理世界更深刻的感知与理解，更加强调系统整体的智能互联和智能化管理，旨在通过更全面、综合的智能化应用，促进人、机、物、境的深度互联融合。新一代智能终端与传统智能终端在交互方式、服务方式、应用范式及业务场景

等方面具有显著区别。

随着信息技术的持续创新与突破,应用场景的不断延伸与丰富,用户个性化、多样化需求的日益变化与增加,新一代智能终端的范畴将呈现不断扩展的趋势,并囊括更多融合新一代数字技术和创新科技、满足更广泛消费及行业需求的设备与服务。产品应用创新方面,以强智能为主要发展方向,以 AI+终端为主要实现方式,代表性产品包括 AI 手机、AI PC、AI 可穿戴设备、AI 车载终端、北斗终端、智能工业终端等;产品形态创新方面,以场景应用为导向,以多样化、融合化为特征,涌现出裸眼 3D 终端、三折叠手机、AI PIN、Rabbit R1、人形机器人、eVTOL(电动垂直起降飞行器)、智能戒指、扩展现实终端等新型产品。当下,人工智能技术是智能终端产业发展的关键驱动力,AI+终端正显著提高智能终端产品功能、不断扩展应用场景、持续提升消费者及行业用户的体验,促进消费类终端向强智能升级,推动新一代智能终端技术创新和产业变革,助力人工智能赋能新型工业化落地实施。

以大模型技术为代表的人工智能正驱动新一轮的行业转型深化,并引领新一代智能终端走深向实。全球大模型井喷式发展,语言大模型多维度能力持续进化,视觉大模型和多模态模型加速迭代。随着缩放定律下大模型技术的不断成熟,其在各行业的深入应用带来显著的效率提升和模式创新,深刻改变制造业的生产模式和创新路径。在大模型的驱动下,人工智能全方位、深层次、多维度重塑终

端开发逻辑、服务模式、交互方式，智能终端产业开启新一轮智能革命周期，进入全产业链条变革阶段。人工智能、大数据、卫星通信等新一代信息技术的深度融合赋能，推动终端设备智能化水平持续提升，助力终端应用场景及功能范围不断拓展，从终端形态、硬件性能、操作系统、中间件到上层应用软件，都在持续优化升级，新一代智能终端的特性更为凸显。

新一代智能终端呈现出“四新”趋势特征，即新的交互方式、新的服务方式、新的应用范式以及新的业务场景。

1.新的交互方式：从“单一模态交互”变为“多模态智能交互”

人机交互历史是不断演进的过程，主要分为 5 种形式。一是命令行交互。在计算机发展的早期阶段，人机交互主要依赖于命令行界面，用户需要通过键盘输入特定的命令来控制或访问系统资源，如程序、文件等。二是图形界面交互（GUI）。图形界面交互增加了形象的可视化元素，如按钮、图标、菜单、窗口等，使用户能够更加直观地与终端系统进行互动。三是多点触控交互。随着触摸屏精度和响应速度的不断提高，多点触控交互通过模拟人类自然的触摸和手势动作，使用户能够通过简单的触摸、滑动、缩放、旋转等手势来控制屏幕上的内容，降低了使用门槛。四是自然语言交互。用户通过自然语言与终端系统进行交互，系统能够理解用户的语言意图，并给出相应的回应，无需手动操作，极大地提高了交互的便捷

性。**五是多模态智能交互。**模态指信息的来源或形式。模态基于人的感官可分为视觉、听觉、嗅觉、味觉、触觉。基于信息的承载方式或传递的媒介可分为图像、视频、语音、文本、触觉信息等，此外，情景环境中也会涉及通过雷达、红外、GPS 等多类型传感器所获取的各种模态数据。多模态智能交互指用户能够通过肢体语言、情感表达、信息载体（文字、图片、音频、视频）、事件环境等多个通道与终端设备进行交流，终端融合处理多维信息并提取核心特征内容进行意图理解判断，根据输入的信息和用户的意图，生成多模态的输出。

新一代智能终端开启人机交互新体验，形成以多模态智能交互为主的多元化并存范式，交互方式从“人适应机器”逐渐演变为“机器适应人”。人类生活在一个由多种模态构成的世界，新一代智能终端更加强调人与终端在特定任务中相互配合的效率。基于多模态大模型强大的泛化能力，终端可以同时处理大量文本、图像、视频、音频等不同模态的信息，并能够结合智能感知的其他数据信息，通过统一的表示空间完成跨模态的信息交互和知识融合，实现更加自然、流畅的人机交互，此种方式更加符合人类感知和处理信息的方式，使得交互方式逐渐从“人适应机器”演变为“机器适应人”。随着技术的不断更迭，终端计算智能、感知智能将逐步走向认知智能，进而促进更多创新交互模式的诞生。新兴的交互方式能够满足用户更复杂的交互需求，催生更多的应用场景。

## 2.新的服务模式：从“被动服务”转向“主动服务”

新一代智能终端构建个性化服务新模式，服务模式由“以用户指令为驱动”转变为“以用户意图为导向”，由“被动服务”转变为“主动服务”。在终端传统的服务模式中，系统通过接收用户的指令或输入执行相应的操作，缺乏对用户深层次意图的理解，用户需要主动寻找、打开特定的应用来满足自己的需求，服务模式主要以反应式为主，缺乏主动性和预见性。新一代智能终端基于深度学习、语义匹配检索、上下文理解、端云协同等关键能力，不仅能够理解和处理用户明确表达的需求，还能够通过分析用户的行为、偏好、情景、上下文等信息，理解用户的深层次意图，甚至提供用户未曾明确表达但实际需要的服务，为用户提供主动性的、预见性的、个性化的服务，满足用户更加复杂多元的需求。新一代智能终端服务模式逐步由“以用户指令为驱动”转变为“以用户意图为导向”，交互方式更加智慧、人性，实现真正“懂你”的用户体验。例如，智能家居设备能够根据用户的生活习惯自动调整室内温度、湿度等参数；智能客服则能够根据用户的咨询内容主动推荐相关的产品或服务；智能医疗助手通过连接用户的医疗记录和健康数据，为用户提供个性化的健康管理建议，如用药提醒、健康饮食建议等。

## 3.新的应用范式：从“AI 赋能”迈向“AI 原生”

以大模型为代表的 AI 技术改变了智能终端应用软件的设计理念及开发模式，激发了基于大小模型的智能应用开发需求，衍生出

**应用+AI、AI 原生应用两类代表性范式。应用+AI 即 AI 赋能传统应用**，主要在原应用场景基础上叠加新的 AI 模块或在已有服务流程上叠加基于 AI 技术的决策调度，以增强软件的功能和用户体验，提升原有软件的智能化水平，这种集成使得传统软件能够处理更复杂的任务。如在内容消费类应用中集成 AI 推荐算法，根据用户的浏览历史、购买记录等数据，为用户提供个性化的内容或商品推荐；照片编辑应用中叠加 AI 修图功能；线上会议应用中增加自动会议纪要、语音转文字功能；英语学习应用叠加 AI 仿真对话等。**AI 原生应用**是以 AI 技术和理念为出发点，依托于大模型技术构建的全新应用程序，从基础架构和设计初始便立足将 AI 能力作为核心，更加强调利用 AI 能力进行原生创造。AI 原生应用能够实现 AI 技术与业务场景的深度融合，具备高度的自适应能力、可扩展性和 AI 含量。

**应用+AI 与 AI 原生应用短期共存，随着 AI 技术的不断成熟和应用场景的持续拓展，AI 原生应用有望成为主流趋势。**应用+AI 仅是在现有产品或服务中引入 AI 技术，未从应用底层重构开发逻辑，在智能化程度、数据利用与整合、用户体验、可扩展性与维护性以及安全性与隐私保护等方面均存在一定的局限性。AI 原生应用能够进行智能化决策、自动化处理，此范式下的应用软件通常具有更高的智能化程度和更好的用户体验。随着 AI 原生应用开发思维、开发方法、商业模式的逐渐成熟，会涌现更多的创新性 AI 原生应用，并有望成为主流趋势，为人们的生活和工作带来更多的便利和效率提升。

AI 智能体作为典型 AI 原生应用，正逐步成为新一代智能终端的核心构成要素与创新的交互入口，通过整合资源内容为用户提供智能化服务。随着技术的不断演进，先后出现基于规则的智能体、基于强化学习的智能体、基于大模型的 AI 智能体（AI Agent）。AI 智能体<sup>1</sup>指能够感知外部信息，自主进行任务规划、决策与执行的软件或系统，旨在响应、理解、分析人类输入，像人类一样与环境互动并完成特定任务目标，通常具备主动性、交互性、反应性和适应性等特性。智能体主要包括系统级智能体及应用级智能体。系统级智能体通常与操作系统集成，旨在管理、控制和优化整个系统或环境的运行，可能涉及多个子系统的协同工作，以及复杂的数据处理和分析任务，能够为用户提供体系化、可扩展的智能服务。系统级智能体更注重整体性能的优化和系统资源的调度。应用级智能体主要聚焦于特定应用领域及细分应用场景，以通过智能技术解决特定的任务。

智能体引领终端进入智能感知、多模态交互、全场景服务的一体化新时代。智能体基于大模型等技术底座，具有信息感知、学习记忆、推理规划、工具调用及反思改进五大核心能力，并赋能人类学习、办公等多种应用场景。通过感知系统状态、环境信息并结合用户提供的输入信息进行意图理解判断，运用内置的智能算法和模型，进行数据分析和逻辑推理，做出任务分解、任务编排及任务优化等规划决策。能够基于一定的学习记忆能力，根据过往经验和反

<sup>1</sup> 本报告中如无特别说明，智能体均指 AI 智能体。

馈不断优化决策策略，以提高任务完成的准确性和即时性。结合多模态大模型能力，可以对语音、图像、手势、触摸等多种形式信息的理解及处理，使得用户能用更加自然、直观的方式与智能终端进行交互，提高了交互的便捷性和效率。智能体所具备的自主性和主动性使终端能够更好地理解用户复杂需求，并通过智能体的集成和协同工作，形成一个紧密相连、互相支撑的智能生态系统，进而为用户提供更加个性化、智能化的服务。



来源：中国信息通信研究院

图 2 AI 智能体架构图

#### 4.新的业务场景：从“万物互联”进入“万物智联”

技术创新从过去的单点突破进入多种技术协同推进、群体性演变新阶段，人工智能、大数据、移动互联网等前沿技术的广泛应用与融合渗透推动业务场景从“万物互联”进入“万物智联”。随着技术的不断进步和应用场景的日益丰富，“万物互联”逐渐显现出其局

限，单纯的连接无法满足人们对智能化、自动化的更高追求，用户所关注的不仅是单设备的体验，更加关注场景化的体验，比如出行是否智能、家居是否智能。万物智联、协同智慧时代更加强调场景化、泛在化、智慧化、生态化、协同化、安全化。“万物智联”不仅要求设备之间互联互通，更强调通过人工智能、大数据、云计算等先进技术，实现设备间的智能识别、智能交互和智能决策。在这个框架下，每一个智能终端都不再是孤立的存在，而是能够感知环境、理解用户需求，并主动提供服务的智能系统，可以根据用户的习惯、偏好和场景，自动调整工作模式，提供个性化的服务，真正实现“懂你所想，予你所需”。在万物智联时代，各种智能终端和设备将形成一个开放、协同、共赢的生态系统，通过数据共享和协同工作，共同为用户提供优质的服务和智能化、便捷性的体验。

新一代智能终端是万物智联的基础和载体，在数据采集与感知、智能分析与决策、互联互通与协同工作等方面发挥着重要作用。万物智联的核心在于多种设备的接入与识别、各种数据的融合与处理、终端的智能决策与控制，通过对海量感知及输入数据的分析，终端需要做出实时反应，并能基于用户的信息及情景环境，提供智能化的服务。新一代智能终端作为万物智联的感知层，能够实时采集和感知周围环境的数据，如温度、湿度、光照、声音等，通过集成 AI 算法和大数据等技术，对采集到的数据进行自主分析和决策，并与其他智能终端设备进行无缝连接和数据共享，做出更加智能的响应

和行动。同时，5G 网络支持更多的设备连接，为设备间的智能交互提供了广阔的空间和可能性，灵活地分配网络资源，满足了万物智联场景下的多样化需求。

## （二）产品应用不断涌现，重塑终端智能边界

新一代信息通信技术加速变革终端能力和应用生态。随着人工智能工程化的加速推进，新产品新模式层出不穷，催生了 AI 手机、AI PC、AI 可穿戴设备等系列人工智能终端，并成为新一代智能终端的代表性应用产品，旨在通过深度集成人工智能技术赋能终端完成更复杂的任务，实现终端更强的性能。AI Pin、Rabbit R1、智能戒指等新形态终端，具有独特的设计理念和功能特点，能够通过学习用户的偏好和行为模式，简化用户与应用程序的交互过程，提升用户使用的便捷性。人形机器人作为具身智能的理想载体、未来产业的新赛道、经济增长的新引擎，将深刻变革人类的生产生活方式，有望迎来更加广阔的发展前景。AI 车载终端集成导航、娱乐、通信等基础功能，并具备车辆监控、智能驾驶辅助、安全预警等高级功能，智能化水平不断提升，并呈现出快速增长的发展态势。本报告中主要聚焦 AI 手机、AI PC、AI 可穿戴设备、人形机器人及 AI 车载终端进行重点介绍。

### 1.AI 手机：从生成式 AI 向全域 AI 升级优化

AI 手机作为人工智能与移动通讯设备融合的新范式，在传统智

能手机的基础上，通过软硬件全方位创新提升移动用户交互体验，推动手机智能化程度迈向新高度。硬件层面，SoC、存储性能不断提升，电池、散热、面板等技术持续进步，助力手机硬件能力不断增强。软件应用层面，从生成式 AI 逐步向全域 AI 升级优化，更加关注 AI 在系统级、场景级、跨应用级中的应用及赋能。手机智能体重塑软件架构和应用模式，整合并联动各种资源工具，为用户提供更加人性化、智能化的服务。全球手机厂商加快升级 AI 战略，相继推出搭载大模型的 AI 手机，抢抓重要发展机遇。2024 年，国内外手机厂商陆续发布新款机型，升级终端硬件能力及服务能力，重点强化 AI 相关功能应用，如图像处理、智能推荐、智能日程管理、原子化服务，并通过系统级操作、跨应用调度、复杂场景分析、智能决策执行等方式提升手机智能化服务水平。在消费电子产品同质化严重的当下，个性化服务已成为品牌差异化、提升用户黏性的关键因素。随着用户需求的不断变化和升级，终端厂商将持续探索新的服务模式和应用场景。

表 1 全球手机代表产品发布情况汇总

| 公司名称 | 代表产品             | 发布时间    | 模型名称        | 部署方式 | 芯片               | 代表性 AI 应用             |
|------|------------------|---------|-------------|------|------------------|-----------------------|
| 三星   | Galaxy S24 Ultra | 2024.01 | 高斯大模型       | 端云协同 | 高通骁龙 8Gen3       | 即圈即搜、通话实时翻译、笔记助手、图片助手 |
| 谷歌   | Pixel 9          | 2024.08 | Gemini Nano | 端云协同 | Google Tensor G4 | 照片修复、自动构图、实时翻译、魔法橡皮擦  |

| 公司名称 | 代表产品          | 发布时间    | 模型名称               | 部署方式 | 芯片         | 代表性 AI 应用                           |
|------|---------------|---------|--------------------|------|------------|-------------------------------------|
| 苹果   | iPhone 16     | 2024.09 | Apple Intelligence | 端云协同 | 苹果 A18     | 通知摘要、通话摘要、写作工具、照片整理                 |
| OPPO | Find X8 Pro   | 2024.10 | 安第斯大模型             | 端云协同 | 联发科天玑 9400 | AI 图像助手、AI 一键问屏、AI 识万物、AI 办公        |
| vivo | X200 Pro      | 2024.10 | 蓝心大模型              | 端云协同 | 联发科天玑 9400 | 原子岛、小 V 圈搜、小 V 帮记、小 V 写作、AI 输入法     |
| 荣耀   | Magic7 Pro    | 2024.10 | 魔法大模型              | 端云协同 | 高通骁龙 8 至尊版 | 任意门、AI 换脸检测、AI 消除、AI 扩图、AI 翻译、AI 办公 |
| 小米   | xiaomi 15 Pro | 2024.10 | MiLM               | 端云协同 | 高通骁龙 8 至尊版 | AI 绘画、AI 写作、AI 抠图                   |
| 华为   | Mate 70 Pro+  | 2024.11 | 盘古大模型              | 端云协同 | 麒麟 9020    | AI 动态照片、AI 互动主题、AI 隔空传送、AI 降噪通话     |

来源：公开信息整理

## 2.AI PC：新型数字化新质生产力工具

AI PC 基于人工智能推理硬件算力和内存的升级，通过大模型为用户提供个性化、高效的人工智能应用服务，驱动传统个人电脑向新型数字化新质生产力工具发展。AI PC 核心特征包括智能化交互、个性化服务、高效能计算、自主学习与进化等，利用人工智能模型和算法，使电脑能够更智能地理解用户需求，自动完成或辅助完成办公、生活、娱乐等应用场景任务。硬件层面，AI PC 芯片架构、算力、工艺、存储升级以支持端侧大模型。CPU+GPU+NPU 集成协同为当前端侧大模型推理架构的主要模式。软件层面，大模型

驱动应用生态向 AI 终端智能体生态演进，对软件的应用架构和模式进行全新重构。应用场景方面，大模型赋能个人电脑提升办公效率，AI PC 正逐渐成为新型数字化生产力工具。全球主要企业加快布局 AI PC，谋求在新一轮产业变革中抢占竞争制高点。国际巨头谷歌、微软、Meta、苹果、高通、英伟达等在端侧模型、芯片算力、赋能应用等多方面发力，加快布局大模型端侧推理部署。如微软与高通实现了 Copilot 在 PC 端推理部署，将其 150 余项任务能力整合到 Windows11，支持本地化智能办公。苹果发布 Apple Intelligence 战略，将 3B 端侧模型集成于 iOS 18、iPadOS 18 和 macOS Sequoia 操作系统中。2024 年起我国联想、荣耀等发布多款 AI PC 产品，以办公、搜索等为主要应用场景。

表 2 全球 AI PC 代表产品发布情况汇总

| 公司名称 | 代表产品                | 发布时间    | 搭载处理器         | 模型名称                       | 部署方式 |
|------|---------------------|---------|---------------|----------------------------|------|
| 微软   | Surface Pro 10      | 2024.03 | 英特尔酷睿 Ultra 7 | Microsoft Copilot（GPT-4 等） | 云侧部署 |
|      | Surface Laptop 7    | 2024.05 | 高通骁龙 X Elite  |                            |      |
| 苹果   | MacBook Pro         | 2024.09 | Apple M4 系列   | Apple Intelligence         | 端云部署 |
| 联想   | Yoga Book 9i        | 2024.07 | 英特尔酷睿 Ultra 7 | 天禧大模型                      | 端云部署 |
|      | ThinkPad T14s Gen 6 | 2024.05 | 高通骁龙 X Elite  |                            |      |
|      | ThinkPad T14p       | 2024.04 | 英特尔酷睿 Ultra 9 |                            |      |
|      | Yoga Pro 9i         | 2024.01 | 英特尔酷睿 Ultra 9 |                            |      |

|    |                  |         |               |       |      |
|----|------------------|---------|---------------|-------|------|
| 荣耀 | MagicBook Pro 16 | 2024.02 | 英特尔酷睿 Ultra 7 | 魔法大模型 | 云侧部署 |
| 华为 | MateBook X Pro   | 2024.04 | 英特尔酷睿 Ultra 9 | 盘古大模型 | 云侧部署 |

来源：公开信息整理

### 3.AI 可穿戴设备：智能化创新应用持续涌现

AI 算法的进步、智能传感器精度的提升及低功耗硬件的发展，使得可穿戴设备能够处理大量数据，并可进行深度学习和智能分析。AI 可穿戴设备是一类能够通过内置传感器实时收集用户数据、环境信息，并结合 AI 算法分析处理，最终为用户提供个性化服务的设备。这类设备具备感知、计算、通信和交互能力，能够自动学习用户的行为和习惯，并通常通过与智能手机等其他设备互联为用户提供更加便捷的智能化服务。当前，AI 可穿戴设备的类型已从智能手表、智能手环扩展到眼镜、服饰、耳机、鞋子等形式，以满足不同场景中的用户需求。

智能手表和智能手环凭借集成健康监测、消息提醒、导航、无接触支付等功能，在健康预警、运动指导和个性化服务方面取得显著进展，并主导可穿戴市场。Apple Watch 和 Fitbit 等设备通过检测异常心率（如心房颤动）帮助用户及早发现潜在健康风险。同时，这些设备能够通过长期数据积累，提供个性化的健康建议，如根据睡眠模式优化作息或根据运动数据调整健身计划。智能眼镜的体积和重量也逐渐轻量化，如雷朋与 Meta 合作推出的智能眼镜，外形

与普通近视眼镜相差无几，但配备了摄像头、麦克风和扬声器等组件，基于内置的 Meta AI 助手，可通过语音进行实时交互。AI 驱动的智能耳戴设备，如智能耳机，不仅提供音频播放和通话功能，还具备主动降噪、声音增强和实时语言翻译等功能，进一步提升用户的生活便利。随着人工智能在可穿戴设备领域的持续发力，越来越多令人瞩目的创新应用将不断涌现。

表 3 AI 赋能的可穿戴设备发布情况汇总

| 产品名称             | 产品类别  | 发布时间    | 公司名称         | 代表性 AI 应用                                       |
|------------------|-------|---------|--------------|-------------------------------------------------|
| Ray-Ban Meta     | 智能眼镜  | 2023.09 | Meta 与雷朋合作推出 | 设置提醒、AI 视频功能、实时翻译                               |
| Ikko Active Buds | AI 耳机 | 2023.09 | Ikko         | AI 语音聊天、40+种语言同声翻译等功能                           |
| AI PIN           | 可穿戴设备 | 2023.11 | Humane       | 连接到基于 GPT - 4 开发的专有大模型，并通过投影方式完成交互              |
| Cleer Arc3 音弧    | AI 耳机 | 2024.04 | Cleer        | 内置 AI 运动算法，升级 AI 语音控制，实现 AI 降噪效果                |
| Galaxy Ring      | 智能戒指  | 2024.07 | 三星           | 健康监测与分析、个性化健康建议                                 |
| QIDI Vida 智能眼镜   | 智能眼镜  | 2024.08 | 中国联通、奇点临近    | 采用双目全彩 AR 抬头显示与 AI 大模型技术，实现运动监测指导、AI 导览和语音助手等功能 |
| Oura Ring 4 智能戒指 | 智能戒指  | 2024.10 | Oura Health  | 推出个人 AI 健康顾问，为用户提供短期和长期健康建议                     |
| Olafriend        | AI 耳机 | 2024.10 | 字节跳动         | 接入豆包大模型，通过语音唤起豆包进行对话                            |

| 产品名称     | 产品类别 | 发布时间    | 公司名称 | 代表性 AI 应用                                        |
|----------|------|---------|------|--------------------------------------------------|
| 小度 AI 眼镜 | 智能眼镜 | 2024.11 | 百度   | 搭载中文大模型的原生 AI 眼镜，具备第一视角拍摄、边走边问、识物百科、视听翻译、智能备忘等功能 |

来源：公开信息整理

#### 4.人形机器人：通用人工智能的理想载体

人形机器人作为人工智能与机器人技术深度融合的新载体，与传统机器人相比，具备智能感知、行为规划、自主学习和决策执行的能力，在技术革新和场景应用方面展现出巨大潜力。当前，人形机器人以大模型驱动的大脑技术为核心，结合类脑计算、多模态感知与融合、自主学习、自主决策和精准控制等前沿技术，形成了从“大脑”到“小脑”再到系统集成的协同创新体系。软硬件的深度优化使得人形机器人不仅能适应复杂多变的环境，还能高效执行任务并实时学习提升。其脑力模拟与运动协作的能力已初步成熟，未来将在工业、教育、医疗等多个领域释放更大的社会和经济价值，逐步成为智能化时代的关键赋能终端，加速推动人类社会在生产方式、服务模式和生活质量上的全面跃升，为智能化变革提供重要支撑。

全球科技企业正加速推动人形机器人技术与应用的创新突破。谷歌推出的 RT-1、RT-2 和 RT-X 大模型大幅提升了机器人认知与执行能力，为多场景智能化应用奠定了基础；英伟达开发的 GR00T 基础模型，覆盖医疗、物流、家务等多个应用领域，展现出卓越的适应性与可扩展性。在产品层面，特斯拉 Optimus 结合自研 AI 模型

与高度集成化设计，专注于制造和物流场景的任务执行。其基础模型通过强化学习和实时决策能力，有效支持机器人完成多任务流水线作业，同时实现低成本大规模量产。Figure 02 则基于多模态感知与自主学习模型，聚焦家庭、零售等多样化场景，其核心算法将视觉、语音与运动控制相结合，赋予机器人在复杂环境中的高效互动与场景切换能力。国内企业亦积极布局，优必选推出接入 AI 大模型的多模态交互系统，智元机器人实现了基于大语言模型的任务规划与行为执行；华为云与乐聚机器人联合发布的夸父机器人，融合了华为盘古大模型和开源鸿蒙系统，在智能化与系统稳定性方面表现突出，为行业注入更多创新动能。

表 4 全球人形机器人代表产品情况汇总

| 公司名称          | 产品名称         | 技术参数                           | 产品介绍                                                                                                                   | 应用场景                 |
|---------------|--------------|--------------------------------|------------------------------------------------------------------------------------------------------------------------|----------------------|
| 特斯拉           | Optimus Gen2 | 高：172cm<br>重：57kg<br>自由度：52DoF | 采用特斯拉自研执行器和传感器，全新设计的双足结构结合自动驾驶系统 FSD（Full Self-Driving）、Autopilot 等神经网络技术和动态平衡算法，实现更稳定的行走和灵活的动态调整。为工厂等复杂场景的任务执行提供强大支持。 | 工业制造<br>物流仓储<br>服务领域 |
| 波士顿动力         | Atlas 电动版    | --                             | 采用电机驱动，具有全向关节设计和 360° 视觉能力。除模拟和模型预测控制外，还配备了新的人工智能和机器学习工具。                                                              | 工业制造<br>特种应用         |
| Agility Robot | Digit        | 高：175cm<br>重：65kg<br>自由度：20DoF | 拥有高度灵活的运动能力，可以在复杂地形中平稳行走，且具备抓取和搬运物体的能力，单次搬运重量可达 16 公斤。                                                                 | 物流仓储<br>快递配送<br>服务行业 |

| 公司名称      | 产品名称      | 技术参数                              | 产品介绍                                                                                                         | 应用场景                 |
|-----------|-----------|-----------------------------------|--------------------------------------------------------------------------------------------------------------|----------------------|
| Figure AI | Figure 02 | 高：170cm<br>重：70kg<br>自由度：手部 16DoF | 通过与 OpenAI 合作训练的定制 AI 模型，实现了与人类的自然对话能力，支持自主行走、物体操作和复杂环境下的任务执行。                                               | 工业制造<br>物流仓储<br>服务领域 |
| 优必选       | Walker S1 | 高：172cm<br>重：76kg<br>自由度：未提及      | 配备 360° 多模态感知系统，包括集成化头部设计、双耳鱼眼相机和 3D 立体视觉技术，以及仿人灵巧手和 6 个阵列式触觉压力传感器，使其能够精准感知周围环境，实现高效的交互和作业。已进入比亚迪工厂执行搬运任务实训。 | 工业制造<br>物流仓储         |
| 傅利叶智能     | GR-2      | 高：175cm<br>重：63kg<br>自由度：53DoF    | 采用了优化的仿生设计和内走线设计，提升了整体美观度和场景适配性。灵巧手单手自由度增至 12 个，并搭载了触觉传感器和运动控制算法，能够实现对不同物体的精准力控抓取。                           | 工业制造<br>医疗康养<br>服务领域 |
| 智元机器人     | 远征 A2-Max | 高：175cm<br>重：85kg<br>自由度：53DoF    | 专为重载特种场景设计，具备力大和灵巧的优势。能轻松搬动 40 公斤的航空箱，展示了其在重载作业方面的能力。                                                        | 物流仓储                 |
| 宇树科技      | G1        | 高：135cm<br>重：35kg<br>自由度：43DoF    | 可进行高难度的动态动作，如动态站起、坐下折叠、舞棍等。基于深度强化学习和仿真训练，借助 AI 的加速发展，在不断进行升级和演进。                                             | 服务领域                 |
| 乐聚机器人     | 夸父        | 高：150cm<br>重：45kg<br>自由度：26DoF    | 国内首款面向家庭场景的华为开源鸿蒙人形机器人，能够实现任务作业、规划导航等功能，步速最高可达 4.6 公里/小时，快速连续跳跃高度超过 20 厘米，能够在各种复杂环境中自由行动。                    | 工业制造<br>服务领域         |

来源：公开信息整理

## 5.AI 车载终端：开启舱驾融合一体化、智能化新时代

AI 车载终端通过集成智能座舱和智能驾驶系统，为用户提供更

**加智能化的服务体验。**人工智能已经全面渗透至汽车产业的各个环节，并推动汽车行业的竞争从电动化、智能化逐渐转向 AI 化。传统汽车制造商、科技公司、互联网巨头等纷纷加大在 AI 领域的投入，通过合作与竞争共同推动汽车产业的快速变革。**数据驱动型端到端智能驾驶正在颠覆过去人工规则驱动的传统智能驾驶开发范式，有望在更广泛的驾驶场景中实现更高级别的自动驾驶功能。**端到端智能驾驶通过构建统一的神经网络模型，从原始传感器数据中直接提取信息，能够学习到各种驾驶场景下的潜在规律和特征，从而具备更强的适应性和泛化能力，实现感知到控制的无缝衔接。多家车企和智能驾驶技术公司正在研发和推广数据驱动型端到端智能驾驶技术，如特斯拉的 FSD 系统、小鹏的图灵 AI 智驾、长城汽车的 Coffee Pilot Ultra 等都已经实现了量产或试量产。**智能座舱由功能座舱逐渐升级为以大模型为基础的感知智能座舱，拥有更高、更全面的智能化交互能力与权限。**功能座舱是智能座舱的早期形态，主要侧重于提供基本的车辆信息和娱乐功能，用户可以通过触摸屏或语音控制来操作车载系统，实现音乐播放、导航查询等功能。这一阶段的智能座舱虽然具备了一定的智能化特征，但交互方式相对单一，且智能化程度有限。大模型技术赋予了感知智能座舱更准确、更流畅的语音识别和自然语言处理能力，以及更丰富的知识储备和语义理解能力。在感知智能座舱中，用户可以通过更自然的多模态方式与车辆进行交互，如通过语音指令控制车辆功能、通过手势操作调整车

载系统等。海外车企特斯拉、奔驰等，以及国内车企广汽、比亚迪、极氪、岚图、理想、蔚来、小鹏等均已在座舱中接入大模型。未来，智能座舱与智能驾驶系统将实现更深度的融合，通过共享算力、数据等资源，提高整体系统的性能和效率。

### （三）扩展市场增长空间，提振行业发展信心

受益于国家政策扶持、新技术融合应用及市场需求扩增等多重因素，智能终端产业进入新一轮增长周期。新一代信息技术通过升级、重组和延伸不断实现自身的迭代进化，从集成电路、互联网到云计算、5G、卫星通信、人工智能等，形成了庞大且交叉组合的技术集群来赋能新一代智能终端发展。此外，人们已经不再满足于普通的电子设备，而是渴望拥有更加智能、便捷、个性化的产品。随着国家相关政策的效果显现、人工智能等创新技术的落地应用及消费者市场需求的推动，全球智能手机、PC 和智能穿戴等代表性产品出货量呈不同幅度增长，推动市场的进一步扩大和升级。

国家部委及地方政府发布系列政策支持新一代智能终端产业发展，为行业创新提供强有力支持。2024 年 1 月，工业和信息化部等七部门印发《关于推动未来产业创新发展的实施意见》（工信部联科〔2024〕12 号），提出“突破下一代智能终端。发展适应通用智能趋势的工业终端产品，支撑工业生产提质增效，赋能新型工业化。发展量大面广、智能便捷、沉浸体验的消费级终端，满足数字生活、数字文化、公共服务等新需求。打造智能适老的医疗健康终端，提

升人民群众生命健康质量。突破高级别智能网联汽车、元宇宙入口等具有爆发潜能的超级终端，构筑产业竞争新优势”。2024 年 5 月，国家发展和改革委员会等五部门印发《关于打造消费新场景培育消费新增长点的措施》（发改就业〔2024〕840 号），提出“打造电子产品消费新场景。加大柔性屏、超级摄影、超级快充、人工智能助手、端侧大模型、跨屏跨端互联等软硬件功能开发，增强人机交互便利性。支持智能穿戴设备在通信娱乐、运动健身、健康监测、移动支付等领域应用，开拓柔性可穿戴、环境自适应智能纺织品应用领域。拓展智能机器人在清洁、娱乐休闲、养老助残护理、教育培训等方面功能，探索开发基于人工智能大模型的人形机器人。鼓励探索反向定制、个性化设计和柔性化生产等新模式，创新电子产品应用场景。鼓励以市场化方式举办电子产品展销会，提高智能产品知晓度和渗透率。”各地方聚焦于推动智能终端产业的创新发展、技术升级和应用拓展，以提升区域智能终端产业的竞争力和影响力。

如《深圳市推动智能终端产业高质量发展若干措施》提出“提升链主企业产业链带动能力”“增强关键核心技术攻关能力”“推进优质产品提质增量”等推动地区智能终端产业集群发展。《重庆市推动成渝地区双城经济圈建设行动方案（2023—2027 年）》提出“培育形成智能家居终端、智能安防终端等百亿级新型智能终端产业”。

IDC 数据显示，我国 2024 年智能终端市场出货量将同比增长

4%，预计 AI 终端<sup>2</sup>占比将达到 55%，搭载 AI 功能的终端设备超过 70%。作为智能终端领域的核心产品，智能手机的市场销量始终受到业界的高度关注。根据 Canalys 发布的报告，2024 年第三季度全球智能手机出货量同比增长 5%，已连续四个季度实现增长。2024 年，具备生成式 AI 能力的智能手机出货量预计将占全球智能手机总出货量的 16%，并在 2028 年激增至 54%。从 2023 年到 2028 年，AI 手机市场的年均复合增长率（CAGR）将达到 63%。

表 5 全球 AI 手机市场规模及预测

| 产品类型             | 2023 年 | 2024 年 | 2025 年 | 2026 年 | 2027 年 | 2028 年 |
|------------------|--------|--------|--------|--------|--------|--------|
| 全球智能手机出货量（百万台）   | 1100   | 1150   | 1180   | 1200   | 1220   | 1250   |
| 全球 AI 手机出货量（百万台） | 55     | 184    | 330.4  | 456    | 573.4  | 675    |
| AI 手机占比          | 5%     | 16%    | 28%    | 38%    | 47%    | 54%    |

来源：Canalys

Gartner 数据显示，2024 年 AI PC 的出货量将达到 4300 万台，AI PC 出货量占比 PC 终端总出货量达 17%以上；预计到 2025 年，AI PC 的全球出货量将达到 1.14 亿台，占比达 43%左右。AI PC 渗透率加速提升，未来 2~3 年占比有望超过 50%，成为市场主力，走向成熟和规模化。

<sup>2</sup> 本统计中，AI 终端是指处理器集成 AI 引擎的终端设备。



来源：Gartner

图 3 全球 AI PC 市场规模及预测

MarketsandMarkets 报告数据显示，全球可穿戴人工智能市场规模预计从 2024 年的 627 亿美元增长至 2029 年的 1385 亿美元，复合年增长率预计 17.2%。这一增长主要受到消费者对健康认识的提升、可穿戴设备技术进步等因素驱动。在预测期内，端侧 AI 细分市场将占据较高的市场份额。智能手表仍有期望占据较大市场份额，AR/VR 头戴设备将引领较高的增长率。Statista 报告数据显示，2024 年全球 AI 机器人市场规模将超过 190 亿美元，较 2023 年市场规模增长近 30%，到 2030 年，预计这一规模将超过 350 亿美元（约合人民币 2497 亿元）。

### 三、新一代智能终端技术创新重点

#### （一）终端硬件创新升级，赋能端侧性能加速提升

##### 1. 全栈智能计算技术夯实新一代终端智慧化基础

###### （1）算力大幅提升，支持端侧更强 AI 性能

端侧部署大模型对 AI 芯片算力提出了更高要求。端侧 AI 芯片不仅需要满足高算力需求，还要在有限的功耗和散热条件下提供稳

定的计算性能。为了实现这些目标，AI 芯片的发展逐渐从以往的通用处理向专用处理过渡，特别是面向神经网络的深度学习任务优化。早在 2017 年，芯片设计公司已经开始在其商用的 SoC 中加入独立神经网络运算单元。根据 Counterpoint 数据，现有手机 AI 算力已经增长了 20 倍（以 TOPS 为单位），且芯片峰值 AI 算力水平还将继续增长。端侧芯片发展主要带来两方面变化。一方面，AI 芯片的架构不断优化，用以深度适配主流 Transformer 架构的 APU、NPU 被推出；异构计算架构（如 CPU+GPU+NPU）逐步应用，在提高算力和性能的同时还可以降低功耗和成本，同时还具备多类型任务的处理能力，可以更好地应对多种计算任务。另一方面，芯片制造工艺不断升级，制程从 4nm 向 3nm 迈进，并有望突破 2nm，芯片的计算密度不断提升，功耗显著降低，更符合端侧设备对能效比的要求。

表 6 部分旗舰级芯片性能表现

| 公司 | 芯片        | 制程          | AI 处理器         | 应用终端 | 性能表现                                                                                                                      |
|----|-----------|-------------|----------------|------|---------------------------------------------------------------------------------------------------------------------------|
| 高通 | 骁龙 8 Gen3 | 台积电 N4P 5nm | Qualcomm AI 引擎 | 手机   | CPU 最高频率达 3.3GHz，性能提升 30%，能效提升 20%，GPU 性能提升 25%，能效提升 25%，AI 性能提升 98%，支持终端侧运行 100 亿参数的模型，面向 70 亿参数大语言模型每秒生成高达 20 个 Tokens。 |
| 高通 | 骁龙 8 至尊版  | 第二代 N3E 3nm | Hexagon NPU    | 手机   | 二代 Oryon CPU 速度提高 45%，电源能效提高 45%，Hexagon NPU 融合了标量、向量、张量等 AI 加速器架构，支持 INT4、INT8、INT16 和 FP16 精度，AI 性能提升 45%。              |

| 公司  | 芯片           | 制程          | AI 处理器      | 应用终端  | 性能表现                                                                                                                               |
|-----|--------------|-------------|-------------|-------|------------------------------------------------------------------------------------------------------------------------------------|
| 高通  | 骁龙 X Elite   | 4nm         | Hexagon NPU | PC    | GPU 算力可达 4.6TFLOPS，异构 AI 引擎性能达 75TOPS，支持设备端运行参数量超过 130 亿的大模型。采用 Hexagon NPU 最高可提供 45 TOPS 算力；CPU 采用 Oryon，未来将进入智能手机、汽车、XR 领域。      |
| 高通  | 骁龙 XR2 Gen2  | --          | Hexagon NPU | VR、XR | GPU 相比前代性能提升 2.5 倍，在终端侧 AI 能力上相比第一代骁龙 XR2 平台提升高达 8 倍。支持 INT8，并优化了不同 AI 模块之间的连接，同时扩展了终端侧的内存。                                        |
| 联发科 | 天玑 9300      | 第三代 4nm     | NPU 890     | 手机    | 内置硬件级生成式 AI 引擎，支持最高可达 330 亿参数的 AI 大语言模型，可基于 1 个基础大模型支持 N 种技能 AI 多模态处理能力达到 50Tokens/秒，长文本处理能力提升 8 倍，支持高达 32K 的长文本内容。                |
| 联发科 | 天玑 9400      | 台积电 N3E 3nm | NPU 890     | 手机    | AI 算力相比前代提升了 40%，从而在 AI 计算方面更具优势。对比天玑 9300，NPU890 在大语言模型（LLM）的提示词处理性能提升了 80%，功耗降低 35%。                                             |
| 联发科 | Kompanio 838 | 6nm         | APU 650     | PC    | 相较于 Kompanio 500 系列显著升级。GPU 性能提升达 76%，CPU 基准测试性能提升达 66%，网页基准测试性能提升达 60%。                                                           |
| 苹果  | A18 Pro      | 3nm         | 16 核 NPU    | 手机    | 集成了 16 核 NPU，运算速度高达 35TOPS，具备更快和更高的效率，内存带宽相比上代增加 17%，与 Apple Intelligence 协同设计，相比 A17 Pro 在运行 Apple Intelligence 的各项功能的速度最高提升 15%。 |

| 公司  | 芯片                 | 制程     | AI 处理器         | 应用终端 | 性能表现                                                                                                                                   |
|-----|--------------------|--------|----------------|------|----------------------------------------------------------------------------------------------------------------------------------------|
| 苹果  | M4                 | 3nm    | 16 核 NPU       | PC   | M4 芯片配备了最多 10 核 CPU, 包括 4 颗性能核心和最多 6 颗能效核心。CPU 性能相比 M1 提升最多可达 1.8 倍。10 核 GPU 提供卓越图形处理性能, 相比 M1 芯片提升最多可达 2 倍。16 核 NPU, 运算速度最高可达 38TOPS。 |
| 英特尔 | 酷睿 Ultra 9185H     | intel4 | Intel AI Boost | PC   | 酷睿 Ultra 9 185H 处理器采用了 6 个性能核, 8 个能效核再加上 2 个超低功耗核构成, 总计 16 核 22 线程。最高睿频可以达到 5.1GHz, 拥有 24MB 三级缓存。                                      |
| AMD | 锐龙 8040 PRO 系列     | 4nm    | XDNA 架构 NPU    | PC   | Zen 4 CPU 架构、RDNA3 GPU 架构、XDNA NPU 架构的强大组合, 并加入了 AMD PRO 技术组合, 适用于商用笔记本电脑, 可以从协作、能效、生产力、隐私安全等各方面全方位提升运营效率。                             |
| AMD | 锐龙 AI 9 HX Pro 375 | 4nm    | XDNA 2 NPU     | PC   | NPU 算力为 55TOPS, 支持 Copilot+ 的实时字幕和实时翻译、Cocreator 等功能。                                                                                  |
| 三星  | Exynos 2400        | 4nm    | NPU            | --   | 与 Exynos 2200 相比, Exynos 2400 的 CPU 性能提升了 1.7 倍, AI 处理能力增加 14.7 倍。                                                                     |
| 三星  | S2MU A02 电控芯片      | --     | --             | 智能戒指 | --                                                                                                                                     |
| 谷歌  | Tensor G3          | 4nm    | edge TPU       |      | 能运行更复杂的 ML 模型                                                                                                                          |
| 英伟达 | DRIVE Orin         | 7nm    | --             | 汽车   | 可提供 254TOPS 的算力, 是适用于智能汽车的中央计算平台, 能够为自动驾驶功能、置信度视图、数字仪表盘以及 AI 座舱提供强力支持。                                                                 |
| 英伟达 | Jetson Orin        | --     | --             | 机器人  | 包含 7 个具有相同架构的模组, 运算速度高达 275TOPS, 其性能是上一代多模态 AI 推理的 8 倍,                                                                                |

| 公司 | 芯片 | 制程 | AI 处理器 | 应用终端 | 性能表现                                                                            |
|----|----|----|--------|------|---------------------------------------------------------------------------------|
|    |    |    |        |      | 并支持高速接口。这款强大的软件堆栈包含预训练的 AI 模型、参考 AI 工作流和垂直应用框架，可加速生成性 AI 的端到端开发，以及边缘 AI 和机器人应用。 |

来源：公开信息整理

## （2）内存容量、带宽、功耗性能成为核心关注点

**终端芯片在大容量、高带宽和低功耗方面的性能提升成为关键要素。**端侧大模型通常包含数十亿甚至更多的参数量，每次处理复杂 AI 任务时，都需要在内存中完成对数据的高频调用和处理，模型参数量越大，对内存带宽需求越高。例如，在 int4 量化级别下，3B（30 亿参数）、7B（70 亿参数）模型压缩后所占用内存空间分别为 1.4G 和 3.5G 左右，因此需要足够的内存带宽来支持这种大规模的数据传输。同时，大模型在推理过程中需要进行频繁的数据交换和快速的存储访问，智能终端不仅需要具备大容量的存储空间，还需要在高数据传输速率和低功耗条件下确保存储效率。尽管处理器算力逐年提升，端侧设备在大模型推理中的计算速度显著加快，但存储带宽相对滞后，使得数据传输效率成为限制计算性能的主要瓶颈。

LPDDR5X、LPDDR 6、HBM 等存储技术被提出，可以有效应对数据传输效率和功耗的问题。LPDDR5X（Low Power Double Data Rate 5X）是一种新型的低功耗双通道内存，传输速率最高可达

8533Mbps，是 LPDDR4X 的两倍，LPDDR5 的 1.33 倍，可以在更短的时间内完成更多的数据传输，从而提高设备的运行速度和响应能力。通过动态电压和频率调节（DVFS）等节能技术，LPDDR5X 的功耗在待机和运行状态下相比 LPDDR5 分别降低了约 20% 和 10%，使得搭载 LPDDR5X 的设备在提供高性能的同时，能够拥有更长的续航时间。LPDDR5X 作为 LPDDR5 的升级版，在传输速率、功耗、带宽等方面实现了显著提升，并在智能手机、平板电脑、自动驾驶汽车、边缘计算设备以及人工智能领域等多个领域展现出广泛的应用前景。业内正积极筹备新一代低功耗内存 LPDDR6 的研发，以适应智能手机、轻薄笔记本等设备的性能需求。超微半导体和 SK 海力士提出 HBM 存储，旨在解决传统内存架构的带宽和功耗瓶颈问题，是一种基于 3D 堆栈工艺的高性能 DRAM，适用于高存储器带宽需求的应用场合。如英伟达的 H100 GPU 和谷歌 TPU 等都使用 HBM 技术，其高数据吞吐能力可以支持更大规模参数量的模型训练与实时推理。HBM 技术主要应用于高性能计算、数据中心、人工智能和自动驾驶等领域，由于生产成本和智能终端设备电池续航等原因，尚未实现在终端上的规模化应用。对于终端设备的存储芯片，成本仍被视为重要考量因素，未来发展也将更注重提高存储密度、数据传输效率等方面。

**“存算一体”是一个重要演进方向，有望降低“存储墙”、“功耗墙”等影响。**传统处理器在执行人工智能算法时，由于存储和计

算分离，在存储器与运算器之间存在大量的数据搬运，造成巨大的功耗和延时开销，导致数据搬运的功耗远远高于计算功耗，成为冯诺依曼架构加速器的发展瓶颈。存算一体技术的核心思想是将存储器与运算器融合在一起，实现在存储器内部执行矩阵向量乘法计算，在完成高并行数据访问和计算的同时有效避免了大量权重数据的搬运，从而达到提高运算速度与能量效率的目的，适合用来加速人工智能算法中的矩阵和向量运算。用于存算一体架构的存储技术路线主要包括近存计算、存内处理和存内计算三大类。每种技术路线都有其独特的特点和实现方式，适用于不同的应用场景和需求。随着半导体工艺的不断进步和新型存储介质的研发，存算一体技术有望在未来实现更广泛的应用和突破。

## 2. 电池、散热、显示等技术演进助力端侧 AI 落地

电池续航、散热等性能也将直接影响到智能应用的用户体验。若电池续航能力弱，则难以支持智能应用长时间运行；散热不佳将会导致设备过热降频，影响性能发挥。随着技术的不断进步，这些问题正逐步得到解决。在续航方面，电池容量呈上升趋势，2024 年智能手机旗舰机型电池容量增长至 5500mAh 以上。端侧 AI 的算力激增导致功耗上升，要求更大容量电池和更快速充电，技术策略主要有负极材料改进和电解质材料改进两种。快充技术是保障 AI 手机续航能力的重要手段，2024 年新推出的安卓旗舰级顶配机型均配备了功率过百瓦的私有快充技术并大部分支持 50W 无线快充，大功

率快充技术的渗透率有望不断提升。端侧 AI 运行时处理器能耗高，产生大量热量，要求终端具备更强的散热性能，液冷技术、VC 均热板、石墨烯散热膜成为当前新的散热解决方案。显示技术方面，DC 调光、Real RGB 等技术支撑长时间使用下护眼体验，窄边框、折叠折痕优化支撑 AI 终端更大的交互界面。在 XR 近眼显示设备中，光波导方案清晰度高、可视角度较大、同时将显示屏和成像系统远离眼镜移到额头顶面或者侧面，极大降低光学系统对外界视线的阻挡，并且使得重量分布更符合人体工程学，从而改善设备佩戴体验。波导的传输方式为近眼显示系统的结构设计提供了更多的选择性，是未来 XR 设备光学方案的主流方向。声学方面，语音输入是人机交互的关键接口，预计后续终端将持续升级麦克风收音质量，增强语音交互体验。

## （二）平台软件加速重构，端云协同成为主要路径

### 1. 原生 OS 重构开发逻辑，提升开发效率

AI 技术在终端产品的融入度逐步深化，当前处于“系统导入 AI”阶段。AI 技术在终端产品落地的过程，可分为“应用加入 AI”、“系统导入 AI”、“AI 重构原生 OS”三个阶段。第一阶段“应用加入 AI”的 AI 功能较为简单，强调单点特性的智能增强，如翻译，图片处理，文字识别等领域引入 AI 能力，实现方式为将各类 AI 能力集成至应用软件中，主要的计算处理任务在云端运行；第二阶段“系统导入 AI”是指 AI 能力下沉至 OS 侧，并对系统高频应用和

服务进行智能化改造，AI 功能融入终端预装应用，如图片一键消除、图库检索、日程自动生成等；第三阶段“AI 重构原生 OS”是以 AI 为中心重构系统及应用，包括交互方式、意图理解、执行方式、系统资源调度等，以系统级入口和全局可用的超级智能体，实现更自然的多模态交互。

当前多家终端厂商旗舰机操作系统中加入 AI 模型和 AI Agent，尝试以 AI 重构原生 OS。传统终端 OS 一般由硬件层、内核层+应用框架层、应用层组成，新一代终端的内核层逐步加入 AI 模型和 AI Agent，实现模型访问、模型管理、意图理解等功能，应用层也将从传统应用演化出原生化服务组件和智能体应用。重构后的系统将有五大特征，一是将生成式 AI 能力内化于操作系统中，进一步落地到终端应用场景，提供智能问答、语义搜索、内容创作、文字处理与总结等功能；二是多模态聚合输入输出，将文本、图像、声音、视频等不同类型数据通过大模型聚合分析，提供统一的输出响应；三是进一步整合 UI 界面，以大模型驱动个性化功能，通过智能体链接各个应用；四是面向开发者提供包含基础大模型能力 API，即操作系统提供的 API 将被扩展，使开发者更便捷地创建智能应用；五是通过用户对用户习惯的自主学习以优化资源管理，如对高频应用进行前台化、重新分配带宽和存储等方式来提升系统性能，为用户提供量身定制的计算体验。

表 7 头部手机厂商操作系统优化情况

| 公司   | 操作系统             | 优化情况                                                                                          |
|------|------------------|-----------------------------------------------------------------------------------------------|
| 华为   | HarmonyOS NEXT   | 智慧助手小艺能力提升，具备更强的感知、推理能力，可实现 23 类 Top 场景记忆感知，知识量突破万亿级 Tokens，300+项服务通过意图框架接入小艺智能体，支持全场景多模态分发。  |
| OPPO | ColorOS 14       | 配备 AIGC 消除功能，支持路人、物体等多类型元素识别与消除，并生成自然的填充画面。AI 通话摘要可以智能识别通话内容，生成重点信息摘要，提升办公效率。                 |
| vivo | OriginOS 5       | 通过提供圈搜、全局唤醒和全局拖拽等全新交互方式，借助端云结合的蓝心大模型，将 AI 技术融入系统的各项能力，如照片去路人、电话智能接听、会议转录摘要、外语沟通翻译和听觉视觉无障碍等场景。 |
| 小米   | 澎湃 OS            | 跨越多端的智能思考中枢，通过原子感知能力，学习人的习惯，让周边设备基于习惯来运作和协同。使用的设备越多，学到的习惯越多，常用的设备无需操作就能按照习惯运行。                |
| 荣耀   | MagicOS 9.0      | 搭载智能体的个人化全场景 AI 操作系统，从传统内核基于设备的资源计算，演进为 AI 内核基于人的关联因子计算，如个人习惯、意图理解、环境感知，AI 算力按需调度。            |
| 魅族   | Flyme 10.5       | Flyme AI 大模型的能力，图库升级了 AI 搜图、图片扩展、魔法消除、AI 写真等功能。                                               |
| 三星   | Samsung One UI 6 | 离线场景 13 种语言互译，并具备笔记助手、浏览助手、智能修图建议和生成式编辑的功能。                                                   |
| 苹果   | iOS 18           | Apple Intelligence 能够深入理解设备中的各类数据，包括照片、邮件、信息等，并据此为用户提供个性化的智能服务，带来了功能更加强大的 Siri。               |

来源：公开信息整理

## 2. 大模型轻量化成为热点方向，端云协同成主流部署方式

模型轻量化部署能够显著降低模型的复杂度，降低对终端计算资源和存储资源的需求。模型规模定律表明，基于海量参数和训练

数据的大规模预训练模型能够有效提高人机交互和推理能力，从而增强可完成任务的多样性和丰富性。随着模型参数数量和训练数据量的增加，模型的性能通常会得到提升，其处理复杂任务的能力也随之跃升。然而，随着模型规模的增大，存储、传输以及计算资源的消耗也随之增加，对轻量化技术提出迫切需求。模型轻量化通过优化模型结构、压缩数据大小以及提高计算效率等手段，能够显著降低模型的复杂度、资源消耗和部署成本，从而使其能够在有限的算力和存储空间内实现高效推理。

端侧轻量化模型正逐渐成为 AI 领域的重要发展方向，各大科技公司纷纷加大在端侧模型层面的投入和研发力度。2024 年，微软、谷歌、Meta 等在端侧模型领域取得了显著进展，并发布了多项重要成果。微软发布 Phi-3 Mini，可测量参数仅为 38 亿，现已在 Azure、Hugging Face 上可使用。Phi-3 Mini 的性能比前一版本更好，与 GPT-3.5 等 LLM 不相上下，微软后续还计划发布 Phi-3 Small（7B 参数）和 Phi-3 Medium（14B 参数）两个版本。Google 发布 Gemma 2B 和 7B，这两款模型更适合简单的聊天机器人和语言相关的工作。Meta 发布全新的量化版 Llama 3.2 模型 1B 和 3B 版本，适用于移动设备和边缘计算。国内科技企业陆续发布轻量化模型，如智谱清言 GLM-4 9B 版本、阿里 Qwen2 0.5B/1.5B/7B 版本、面壁智能 MiniCPM 2B 版本。同时，国内终端厂商持续发布自研端侧模型，如 vivo 发布蓝心大模型 3B 版本，荣耀发布魔法大模型 3B 版本，OPPO 发布

AndesGPT 7B 版本，小米发布 MiLM 大模型 4B 版本。

为了更好地适配端侧运行需求，厂商通过参数共享、模块化架构、紧凑表示等方式创新模型架构设计，通过量化、剪枝、知识蒸馏和低秩分解进行模型压缩。面向特定任务，常在自注意力层或前馈网络层采用 LoRA 等技术微调，避免全参数微调带来的高昂计算成本和巨大内存消耗。通过低秩矩阵分解，高效调整模型的权重，而无需修改或重新训练所有参数，快速适配新任务，如摘要总结、内容生成、文本润色等特定任务。未来端侧模型仍面临平衡性能与资源限制、确保模型鲁棒性、持续学习和个性化等挑战，预计端侧模型参数不会爆发式增长，而是向专业化任务小模型、多功能与多模态能力集成的方向拓展。

端云协同部署模式实现计算资源动态分配，增强系统灵活性与可扩展性，成为主要发展路径。端侧模型优势点在于低延迟、强隐私保护、免受网络强弱限制，满足自动驾驶、智能机器人等实时性要求高的场景，处理敏感信息（如医疗、金融数据）更具安全性。但受终端计算、存储和能耗限制，无法支持较大规模的 AI 模型，所以当前端侧大模型常用于处理算力要求较低的任务，如翻译、图片编辑、本地搜索等。云侧大模型优势点在于算力强、知识库丰富、可扩展性与灵活性好，能够支持更大规模的模型部署和训练，所以云侧负责算力要求高的 AI 工作，如提供更高品质的内容生成及更复杂的推理任务。综合终端的成本、性能、功耗等要求，以及目前

大模型的能力，当前端云协同是主流解决方案。端侧模型兼顾本地化运行的算力限制，负责执行实时数据的采集和处理、敏感数据本地处理、用户意图理解、贴合用户的个性智能化、端侧计算资源调度等功能，做本地化快速响应；云侧大模型做深度用户意图理解、复杂指令拆分与执行、跨设备协同、远程设备管理等。未来端云协同仍为主流部署模式，并将不断优化云侧、端侧资源调度和使用，平衡能力、效率、功耗等关键性能，有望在智能家居、自动驾驶、工业制造等行业场景提供更多服务。

表 8 端云协同应用场景划分

| 类别       | 云端                                                   | 端侧                                                               |
|----------|------------------------------------------------------|------------------------------------------------------------------|
| 算力       | 算力强大、并行可扩展、适合训练和推理阶段计算                               | 算力有限、可扩展性差、更适合推理阶段前向计算                                           |
| 时延       | 网络时延+计算开销                                            | 本地计算，无网络开销或很低，实时响应                                               |
| 网络依赖     | 强依赖                                                  | 弱依赖；隐私保护                                                         |
| 系统架构     | 灵活可扩展                                                | 封闭、比较分散                                                          |
| 端云协同场景划分 | 提供通用泛化能力，负责全局规划及复杂任务处理。如：深度用户意图理解、内容创作、跨设备协同、远程设备管理等 | 提供即时处理能力，负责低延迟、高敏感及离线应用场景。如简单用户意图理解、翻译、通话摘要、文本润色与总结、本地搜索、敏感信息处理等 |

来源：中国信息通信研究院

(三) 多重技术综合应用，催生人机交互全新范式

1.多元感知丰富用户体验和服务

终端感知技术发展经历四个阶段。一是基本功能阶段，以机械

和简单电子设备为主，限于最基本的功能，如信号强度检测、电池电量监测等，感知能力有限；二是多媒体互动体验阶段，电子技术的发展促进了感知设备的精度和种类的增加，加速度计、光线传感器等开始被集成到手机中，传感器逐渐小型化、集成化；三是智能感知阶段，感知设备能够采集、处理和传输大量数据，传感器网络开始兴起，数量大幅增加，种类也更加丰富，包括人脸识别、指纹识别、心率监测、环境噪声检测等；四是全场景融合感知阶段，广泛收集环境数据、人体数据、位置数据等多维度信息，并通过人工智能算法完成数据处理分析，识别推测用户意图，从而做出最合理的交互决策和推荐，为用户提供“主动服务设计”。

**当前感知能力正从智能感知阶段逐步向用户全场景融合感知阶段过渡。**随着终端产品日益丰富，感知源不断扩充，逐渐实现多场景多维度的感知能力，包括：基于 LBP 特征、HOG 直方图等图像感知，基于定位、加速度等的位置感知，基于体温、心率等的人体生理指标感知，基于行为分析的意图感知等。然而，目前大部分感知呈碎片化，各终端和各软件间相互独立，通常不分享感知数据和感知结果。未来随着操作系统重构，将实现感知数据汇总、感知结果共享、系统调度优化，支撑新一代智能终端对物理世界全方位的感知需求。

## 2. 意图框架促进数据共享和协同

意图框架作为用户与多样化生态服务的桥梁，促进了终端与应

用软件的数据共享与服务协同，将为用户提供更智能的交互体验及更便捷的服务操作。意图框架是一种系统级的意图标准体系，通过多维系统感知、大模型等能力构建全局意图范式，为终端与应用软件之间提供标准化的接口和协议，使得不同设备和应用软件能够遵循统一的规范进行通信和交互。这种标准化降低了设备和应用软件之间的耦合度，提高了系统的可扩展性和兼容性。意图框架还促进了终端与应用软件之间的数据共享和协同。通过意图框架，不同设备和应用软件可以共享用户数据、意图数据等资源，减少了用户人工干预的流程，从而实现更加精准的服务推荐和个性化的用户体验。例如，在移动设备上，用户可以通过语音指令要求系统打开某个应用、发送消息或进行其他操作，而无需手动触摸屏幕。

### 3. 屏幕理解实现任务自动化操作

**屏幕理解实现感知与交互自动化，开创人机交互新范式。**屏幕理解技术指利用计算机视觉、图像处理、自然语言处理等多种技术手段，对屏幕上的图像、文本、图标等元素进行识别和分析。通过这一技术，系统能够深入理解屏幕内容，并根据用户意图自动执行相应的交互操作，实现更加智能化、便捷化的人机交互体验。此技术还为视障用户等特殊群体提供了无障碍辅助服务，通过解析屏幕内容，系统可以将屏幕上的信息转化为语音输出给用户听，从而帮助他们更好地使用电子设备。屏幕理解技术在一定程度上打破了终端与应用的连接壁垒，助力终端与应用之间的数据和服务共享，实

现终端感知和交互的自动化流程。

#### （四）安全防护策略演进，筑牢用户安全隐私屏障

##### 1.提升安全风险检测能力

传统安全风险检测主要依赖人工巡查、经验判断和物理测试等手段，存在检测效率低、准确性不足、难以发现潜在风险等问题。利用人工智能的多模态分析处理能力，新一代智能终端可全面高效地发现端侧网络和数据安全中的规律和异常，如病毒、恶意代码植入或非法数据访问等，对安全风险进行预测和评估，从而实现对安全风险的精准识别和预警处理。

针对合成语音、伪造换脸、屏幕共享等内容安全风险问题，新一代智能终端利用人工智能精准识别伪造内容和异常行为，及时采取处置措施，维护公共安全和社会秩序。如部分终端内置大模型可以对文本转语音、语音克隆、变声器等方式的合成语音进行端侧实时检测，若检测出合成音频，将向用户弹窗提示风险，预防深度合成语音诈骗。

##### 2.打造智能安全防护新范式

新一代智能终端设备将具备先进的分析能力，可以综合用户的语音、屏幕内容、传感设备等信息来洞察用户意图，进而自主预测用户行为，编排触发多个应用调用完成复杂任务，呈现出围绕用户意图和场景实施自组织、多应用协同的新特点，传统的基于静态规

则访问控制权限和数据安全方法已难以满足当前的新挑战。

AI 大模型凭借高效的思维推理与研判能力，通过端云协同方式自动化识别应用场景，匹配场景安全策略，动态调整本地多应用/服务的访问权限，确保移动应用交互安全；通过多智能体代理协同工作，对高敏感数据执行加密、隔离操作，确保用户数据在应用间动态自适应安全流动，打造智能安全防护新范式。

### 3.构建端云协同的大模型安全防护体系

端云协同逐渐成为终端部署 AI 大模型的方式，端侧和云侧模型分别执行轻量和复杂的推理任务。然而，AI 大模型因其不可解释性、训练数据繁多、安全对齐困难等问题，存在数据投毒、遭受提示词攻击、推理幻觉和不安全内容输出等全新安全风险。

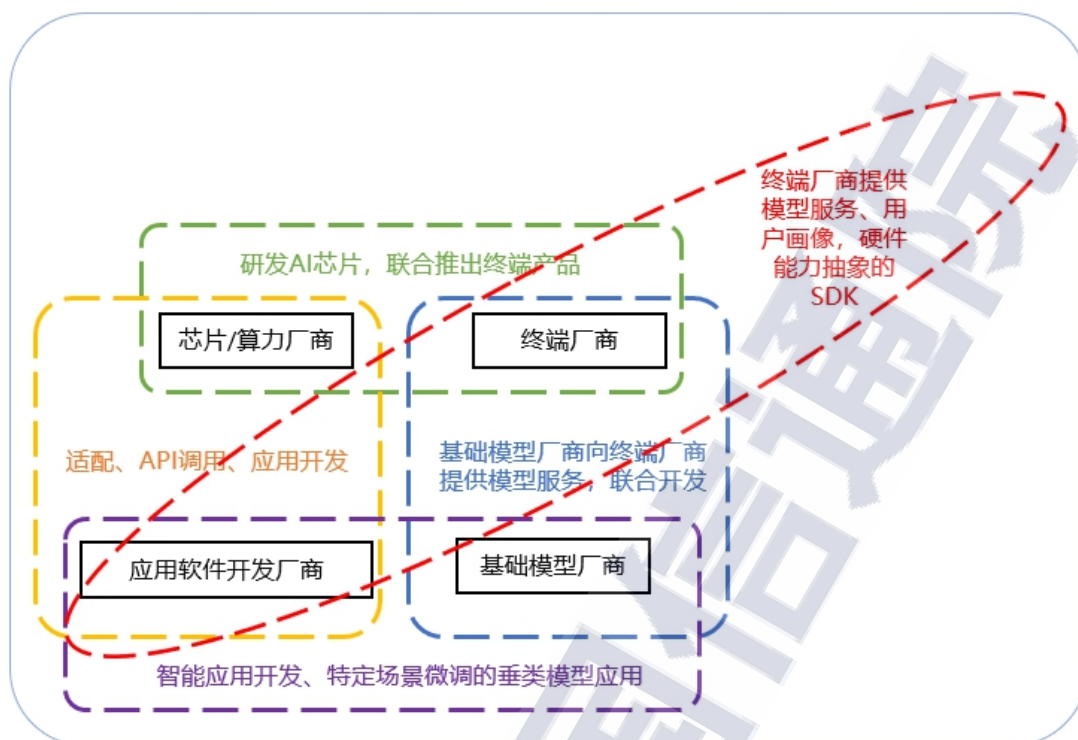
新一代智能终端为确保 AI 在端侧的安全性，围绕大模型全生命周期对端侧和云侧大模型实施安全防护措施。在大模型训练阶段，通过数据安全合规清洗、大模型安全微调等数据增强和模型加固技术，云侧确保大模型使用安全的公开数据和用户私人数据进行训练；在大模型部署阶段，通过红队测试和大模型自查等技术模拟攻击和发现大模型漏洞，以确保大模型的安全性；在大模型应用与推理阶段，实施大模型输入和输出安全的过滤，按照数据敏感性实施端云协同推理，仅在端侧进行高敏感数据业务推理，保障用户隐私安全。

## 四、新一代智能终端产业应用生态

### （一）产业结构优化调整，形成多方协同生态

AI 重塑产业结构，初步形成软硬协同、多方共建的产业生态。

新一代智能终端市场的主要参与者包括终端厂商、芯片/算力厂商、基础模型厂商以及应用软件开发厂商，深度参与到智能终端生态的各个环节，形成五类协同开发关系。一是芯片/算力厂商研发 AI 芯片，联合终端厂商推出终端产品；二是芯片/算力厂商联合应用软件开发厂商完成适配、API 调用、应用开发；三是基础模型厂商向终端厂商提供模型服务，联合开发应用，将模型能力集成到终端中；四是应用软件开发厂商联合基础模型厂商完成智能应用开发、特定场景微调的垂类模型应用等；五是终端厂商向应用软件开发厂商提供模型服务、用户画像和硬件能力抽象的 SDK，助力应用开发厂商实现语义解析、模型微调等功能。在当前阶段，各类型厂商都具备自己的优势能力。以 AI PC 为例，芯片厂商以提供算力支撑为主、模型厂商主要负责提供价值支撑、PC 厂商主要负责集成并落地产品，但都很难完全覆盖 AI PC 所需的全部价值点，市场整体以合作适配为主。未来，随着终端产品进一步智能化升级，市场竞争将更趋向于产品体验与用户黏性的竞争，各厂商将继续深化竞争壁垒，而数据储备强、软硬件兼备且能持续投入能力的企业可能会更加具备竞争优势。



来源：中国信息通信研究院

图 3 新一代智能终端产业链各方角色

## （二）应用创新持续深化，开发逻辑逐步迁移

以大模型为代表的 AI 技术正引领着智能终端应用的创新浪潮，并衍生出两类具有代表性的应用：AI 原生应用、应用+AI（即 AI 赋能传统应用）。这两种应用类型不仅推动了技术的创新，也为用户带来了全新的体验。AI 原生应用是指依托于大模型技术构建的全新应用程序，这类应用从基础架构和设计开始就将 AI 能力作为核心，能够充分利用大模型强大的推理能力完成复杂的操作。如智谱清言推出 AutoGLM，可以基于 GUI 模拟人类在手机上的操作，自动完成点外卖、编辑评论、淘宝购物、朋友圈点赞、总结文章生成摘要

等日常活动。**应用+AI**是指将 AI 集成到现有的软件中，以提升原有软件的智能化水平，这种集成使得传统软件能够处理更复杂的任务。如 Adobe 在 Photoshop 引入了多项 AI 功能，包括生成式形状填充、快速应用插图、识别编辑字体等，提高创作效率。**终端厂商系统级智能体作为 AI 原生应用的一种主要形式**，通常与终端硬件、操作系统深度集成，能够提供更加个性化和优化的用户体验。如荣耀发布 YOYO 智能体可完成 600 项需求意图理解、950 项个人习惯记忆、270 项复杂任务规划，可实现一句话点咖啡等 900 项任务自动执行。联想的“小天”个人智能体，能实现意图理解和主动感知、自我能力认知、短期长期记忆、任务规划和分解等功能。“小天”使用户与多类 AI 终端设备自然交互，打造无缝连接的服务体验，是用户在工作、学习和生活等不同场景下的“个人 AI 助理”。这两种应用的共同点在于它们都利用了 AI 技术的强大能力，但各自的侧重点有所不同。AI 赋能传统应用更侧重于提升现有软件的智能化水平，AI 原生应用以 AI 为核心驱动，拓展技术的应用边界，终端厂商系统级智能体更强调与硬件、系统的深度集成和个性化服务。

随着两类应用不断探索落地，应用开发逻辑也正在改变。传统应用软件开发流程包括需求定义、开发平台选择、前端开发与 UI 设计、服务器及数据库配置、测试发布。AI 应用软件开发流程则更强调大模型选择、数据准备、模型训练，以及模型集成至应用软件。

### （三）多元流量入口并存，构建全新市场格局

平台级超级应用、应用级智能体、终端系统级智能体正构建新一轮流量入口，为新一代智能终端带来更多商业可能。移动互联网时代，流量入口呈现出多样化和碎片化的特点，超级应用、应用商店、浏览器、小程序和社交媒体成为主要流量入口，用户获取内容的链路变短。现阶段，智能终端通过整合内容和联动资源应用，解耦数据、软件、硬件之间的壁垒，产生了新的交互入口，驱动流量入口生态进一步演进，商业模式也随之发生变化。

超级应用凭借其庞大的用户基数及生态优势，通过融合大模型能力进一步强化其入口地位。超级应用通过融合大模型能力，并探索跨应用交互，进一步提升其用户体验和服务质量，吸引更多的用户流量，增强用户的黏性和忠诚度。同时，超级应用正在利用其庞大的用户基数和生态优势，构建更加完善的生态系统和服务体系，以在新一轮流量入口变革初期进一步巩固和强化其入口地位。如支付宝推出智能助理，覆盖吃、喝、行、娱乐等多种生活场景，进一步强化其日常生活服务领域的入口地位。钉钉上线 AI 助理市场，进一步扩大其在办公领域的优势。微信与腾讯元宝深度绑定，在 AI 搜索、AI 创作方面为用户提供更深度、专业的信息。

应用级智能体是指以大模型为核心技术而构建设计的智能体应用，有望成为用户获取信息和服务的新入口。应用级智能体将改变用户的使用路径，即智能体理解用户意图，自主完成问题拆解和决

策，直接调用应用工具实现既定目标并自主执行任务，一定程度上掌握应用生态的分发权。在智能体的基础上，模型厂商推出自家应用商城，进一步构建以大模型为中心的产业生态。如 OpenAI 推出的 GPT Store 汇集开发者和用户来创建、分享和销售基于 GPT 模型的定制化 AI 应用，用户可以构建个性化智能体，包括写作、研究、编程、教育和生活方式等领域，有望改变传统的 AI 应用开发和分发模式并形成新的入口及生态圈。

**终端厂商升级语音助手至系统级智能体，有望成为人机交互新入口。**与语音助手的严格遵循用户明确指令相比，智能体更能感知用户意图主动提供服务，对非明确指令主动规划并通过多轮交互完成用户需求，善于调用工具达成用户目标。一方面，终端厂商具备底层硬件和操作系统天然优势，可以将自己的智能体深度集成到操作系统中，提供更加自然无缝的用户体验。另一方面，终端厂商围绕自己的智能体构建生态，吸引更多开发者，依托自有应用商店，为开发者提供应用分发渠道，进一步丰富生态。

#### **（四）智能分级纵深推进，引领产业高质量发展**

**终端智能化**是指基于先进的信息技术，对各类终端设备进行深度赋能，使其具备数据处理能力、意图分析能力、决策执行能力和学习进化能力等智能功能，为人类提供更加便捷、高效、安全的服务。**终端智能化分级**为新一代智能终端的技术创新提供明确的方向和目标。随着分级标准的不断提升，智能终端厂商需要不断投入研

发，提升设备的智能化水平，以满足市场需求，有助于推动整个智能终端行业的技术创新和产业升级。技术的创新不仅推动了智能终端产业的发展，也带动了相关产业链上下游的协同发展。

中国信息通信研究院联合业内二十余家重点企业，开展了终端智能化分级研究，并发布《终端智能化分级研究报告》。依据终端对用户情景和意图的理解程度、对用户信息的学习记忆程度、任务规划程度、任务完成度四大核心要素，将终端智能化水平分为 L1-L5 五个等级。L1 级（智能响应级），具备极少的智能化，能够根据用户确定性的操作指令完成相关任务。L2 级（智能辅助级），具备一定的智能化，能够准确感知识别用户简单意图，并基于用户偏好完成单类型任务。L3 级（智能助理级），具备较高的智能化，能够准确感知识别用户复杂意图，并基于用户信息进行自主规划完成多类型任务。L4 级（智能协同级），具备更高的智能化，能够准确感知识别用户潜在意图，并基于用户情景进行自主规划完成多类型任务。L5 级（自主智能级），具备全面的智能化，能够准确主动预测识别用户意图，并基于全场景进行自主规划完成全类型任务。终端智能化程度是评估行业现状、驱动产业升级的关键因素，成为推动数字化进程、构建智慧社会、实现高质量发展的重要力量。现有智能化分级研究内容是基于当前技术及产业发展特点所形成的行业共识，后续产业界需要根据技术演进情况、市场及用户需求的变化情况等持续滚动研究，协同合作、凝心聚力，共同推动终端智能化水平提升。

## 五、新一代智能终端发展建议

### （一）创新软硬核心技术，加速终端智慧升级

模型算法、操作系统以及芯片仍是下一阶段的技术创新重点，在后续的发展过程中，可通过针对性、组合性、协同性的有效措施支持核心技术的突破与发展。一是聚焦模型轻量化、可扩展性与适配性的研究与提升，大力增强国内厂商在端侧模型方面的自主研发水平，持续挖掘模型压缩的潜力，力求达到压缩极限，从而为端侧模型规模化部署应用打好基础，并积极向专业化小模型优化及多功能、多模态能力集成等方向拓展，以丰富模型的应用场景与功能表现；二是重点突破芯片、多模态传感等关键硬件技术，解决端侧部署面临的内存带宽不足、计算能力有限以及功耗受限等问题，逐步缩小我国与国外在零部件、制造工艺、架构设计上的差距，降低技术封锁和供应链受阻等可能出现的国际局势影响；三是关注模型底层架构搭建和研发，培育具有全球影响力的通用大模型企业，支持开源社区建设，构建具有国际化竞争力的产业生态。

### （二）健全智能终端标准，推动标准落地实施

2024 年 6 月，工业和信息化部等四部门联合发布《国家人工智能产业综合标准化体系建设指南（2024 版）》，梳理了人工智能标准体系框架，包含基础共性、基础支撑、关键技术、智能产品与服务、赋能新型工业化、行业应用、安全/治理等七个部分。未来应进一步

落实人工智能标准顶层规划，聚焦端侧模型能力、交互能力、软硬件能力以及智能化水平评估等方面，丰富智能产品与服务中有关新一代智能终端的相关标准，加强标准制定与产业发展的协同联动，确保标准制定与产业实际需求紧密结合，推动标准在产业中的广泛应用和落地实施。**一是**推动智能终端生态系统的标准化建设，包括硬件、软件、服务等方面的标准化接口和协议，促进不同品牌、不同厂商之间的生态系统互联互通，制定统一的智能终端平台标准，包括操作系统、应用开发框架、云服务接口等，降低开发者门槛，促进应用生态的繁荣。**二是**推进终端智能化分级测试评估，针对 AI 手机、AI PC、XR、可穿戴等典型产品开展智能化分级测试评估工作。**三是**研究具有前瞻性与针对性的新一代智能终端个人信息保护标准体系，构建生成式人工智能用户权益保护系列标准，规范风险分类分级、用户数据处理、个人权利响应及服务保障能力，强化标准对产业、社会发展的牵引作用。**四是**推动细分场景下的标准化工作，针对细分场景下的智能终端产品，制定详细的产品标准，包括功能要求、性能指标、安全规范等，确保产品质量。**五是**与国际标准化组织对接，参与国际标准的制定和修订工作，提升国际话语权和影响力。

### （三）壮大终端产业主体，培育繁荣共生生态

在智能终端产业培育方面，我国拥有庞大的消费市场和多样化的应用环境，这些为行业发展提供了坚实的内需支撑和广阔的增长

空间，也为智能终端的快速迭代与应用提供了丰富的试验田。未来可以借鉴欧盟“领先市场计划”，以大市场培育大产业，通过税收优惠、购买补贴等举措有效激发终端用户的参与热情与消费活力，进而实现终端产业主体的持续壮大。

当前端侧 AI 模型百花齐放，终端厂商、模型厂商及应用厂商纷纷聚焦于应用入口的激烈争夺，商业模式还未成熟，培育繁荣共生生态是当前产业发展的重点。未来应积极发挥政府部门以及行业组织的引领与协调作用，加强硬件提供商、大模型服务商、终端厂商、运营商、AI 应用开发商多方协同，深入挖掘智能终端应用，尤其是探索教育、工业、金融、医疗等垂类场景，综合多样化技术路线，力争构建繁荣共生的生态体系。

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62300393

传真：010-62300586

网址：[www.caict.ac.cn](http://www.caict.ac.cn)

