CAICT 中国信通院

中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

集智 专题报告

端侧智能产业发展报告

(2026 年)

中国信息通信研究院人工智能研究所
中国人工智能产业发展联盟端侧智能组

2026 年 10 月

版权声明

本报告版权属于中国信息通信研究院和中国人工智能产业发展联盟，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院和中国人工智能产业发展联盟”。违反上述声明者，编者将追究其相关法律责任。

前 言

人工智能正加快从云端集中计算向端云协同、泛在部署和场景化服务演进，端侧智能成为人工智能能力向用户侧、现场侧和物理世界延伸的重要方向。随着大模型、多模态模型、智能体等技术持续突破，人工智能对计算、存储、感知和执行能力提出更高要求，如何在资源受限、环境多样的设备上实现高效部署和稳定运行，正在成为产业创新的重要课题。端侧智能以本地智能为能力基础、以端云协同为运行模式，涵盖模型、芯片、传感器、存储、基础软件、智能体框架、应用服务、安全机制和运营生态的全栈式能力体系。

当前，端侧智能的发展重点正在由“能否部署 AI”向“能否高效运行并形成持续服务”转变。一方面，量化、剪枝、蒸馏、稀疏化等模型轻量化技术不断降低计算、存储和功耗需求，高效模型架构、推理优化和大小模型协同持续提升有限资源条件下的智能能力；另一方面，CPU、GPU、NPU 等异构算力以及传感、存储、通信和执行部件加快与模型融合，编译器、运行时和基础软件进一步承担模型适配、资源调度和任务编排功能，推动端侧智能从单点功能嵌入走向软硬件协同和系统级智能。与此同时，模型能力与多样化硬件深度融合，正在催生新的产品形态、服务业态和商业模式。人工智能逐步从既有产品的附加功能转变为产品设计和服务组织的重要基础，个人智能、空间智能、行业现场智能和具身智能等形态加快发展，服务方式也由即时响应向持续感知、跨设备协同和智能体自主执行演进。

本报告系统梳理端侧智能产业发展脉络、政策环境、市场态势和概念特征，重点研究模型轻量化、异构硬件融合、基础软件、智能体、端云协同和安全可信等关键技术，研究个人智能、空间智能、移动智能、行业智能节点和具身智能等原生产品形态，梳理消费娱乐、智慧家庭、智慧办公、智慧出行、智慧健康及融合场景的发展特征，并进一步研判端侧智能在技术、应用、生态和安全等方面的发展趋势、主要挑战与推进路径，为技术创新、产业布局、标准研制和应用推广提供参考。由于端侧智能仍处于快速发展阶段，相关技术路线、产品形态仍在持续演进，限于研究时间和资料掌握有限，报告中难免存在不足之处，敬请批评指正。

目 录

一、端侧智能产业综述 1
（一）产业由功能嵌入迈向系统智能，智能体推动端侧能力加速跃迁 1
（二）政策持续强化人工智能向端侧和实体场景渗透 4
（三）产业价值由硬件增量向智能能力与服务增量拓展 7
（四）端侧智能是以本地为主、端云协同的融合智能形态 9
二、端侧智能技术体系 12
（一）智能硬件层：异构计算与高效存储支撑模型稳定运行 13
（二）基础软件层：编译适配与资源调度打通模型到硬件执行链路 15
（三）端侧模型层：轻量化与高效推理提升有限资源下的能力密度 18
（四）端侧应用层：智能体与原生应用推动 AI 由功能调用走向任务执行 20
（五）端侧安全信任层：构建覆盖模型、数据与智能体执行的可信边界 .. 25
三、端侧智能原生产品形态 28
（一）个人智能产品：多类硬件协同形成智能入口 28
（二）空间智能产品：感知、决策与执行融合形成环境级智能 31
（三）移动智能产品：本地实时推理支撑自主感知与移动决策 32
（四）行业智能节点：轻量模型下沉推动现场设备由采集走向决策 33
（五）具身智能产品：模型与传感、控制和执行系统融合走向物理任务闭环 35
四、端侧智能应用场景 36
（一）消费娱乐：以内容服务为核心，极致用户体验驱动商业转化 37
（二）智慧家庭：以多端协同为核心，由被动响应迈向主动服务新模式 .. 39
（三）智慧办公：以提升效率为核心，重塑知识管理与工作流体系 42
（四）智慧出行：以安全服务为核心，出行效率与交互体验持续优化 46
（五）智慧健康：以持续感知为核心，打造集监测-发现-干预为一体的健康服务体系 48
（六）融合场景：打通多端服务链路，构建全场景智能协同 50
五、总结与展望 53

（一）趋势：技术产品双向演进，端侧智能加速规模化落地 53
（二）挑战：多重瓶颈交织叠加，产业成熟仍需系统突破 55
（三）建议：强化技术生态协同，夯实端侧智能发展基础 57

图目录

图 1 端侧智能产业发展阶段图 2
图 2 端侧智能技术架构图 12

表目录

表 1 国家层面端侧智能相关政策 5
表 2 国内端侧智能典型载体市场规模及出货量参考 9
表 3 以人工智能任务需求为牵引的智能硬件技术分层 14
表 4 以人工智能任务需求为牵引的基础软件技术分层 17
表 5 以人工智能任务需求为牵引的端侧模型技术分层 20
表 6 端侧智能原生应用 24
表 7 以人工智能任务需求为牵引的安全治理技术分层 27
表 8 端侧智能原生产品典型形态与融合特征 29
表 9 不同硬件资源条件下的模型部署与协同方式 30
表 10 行业智能节点典型形态与模型硬件融合特征 34
表 11 消费娱乐主要端侧产品类型 37
表 12 智慧家庭主要端侧产品类型 39
表 13 智慧办公主要端侧产品类型 42
表 14 智慧出行主要端侧产品类型 46
表 15 智慧健康主要端侧产品类型 48
表 16 融合场景主要端侧产品方案 50

一、端侧智能产业综述

生成式人工智能产业快速发展，当前主流 AI 服务高度依赖云端集中式算力架构，在隐私安全、运行成本、服务稳定三方面的矛盾日益凸显。随着终端硬件算力提升、软件栈持续优化，原本部署于云端的 AI 推理任务开始向设备本地迁移，端侧智能成为行业重要发展方向。端侧智能是以本地为主、端云协同的融合智能形态，其发展动力既来自模型轻量化、异构计算和基础软件能力提升，也来自实时响应、隐私保护、离线可用和自主执行等需求增长。随着模型、硬件、系统和应用协同演进，端侧智能正由单点算法嵌入走向系统级智能，并进一步催生智能原生产品、持续智能服务和智能体型新业态。

（一）产业由功能嵌入迈向系统智能，智能体推动端侧能力加速跃迁

端侧智能产业大体经历功能嵌入、系统融合和智能体跃迁三个阶段，演进主线由“把 AI 功能放到设备中”逐步转向“围绕模型运行重构软硬件体系”，并进一步迈向“以智能体组织感知、推理和执行”，如图 1 所示。这一过程既体现模型从云端依赖向本地高效部署的迁移，也体现人工智能与芯片、存储、传感、系统软件和执行部件由松散适配走向深度协同。



来源：中国信息通信研究院

图 1 端侧智能产业发展阶段图

1.功能嵌入期（2011—2022 年）：轻量 AI 能力开始向设备侧下沉

AI 作为附加功能嵌入终端，主要解决“是否具备 AI 能力”的问题。端侧智能处于萌芽起步阶段，AI 主要作为终端差异化卖点存在，尚未成为系统级能力和核心竞争力。此阶段 AI 以单点独立功能为主，典型应用包括语音助手、拍照美颜、人脸解锁、场景识别等，功能之间相对割裂，缺乏全局协同和跨场景调度能力。技术上以云端推理为主、端侧轻量推理为辅，终端仅集成小规模 AI 加速单元，端侧算力较低，主要支撑简单算子推理，难以运行大参数模型。总体来看，AI 开始打破终端单纯依赖硬件参数竞争的路径，语音交互、影像优化等能力成为产品差异化亮点，但受限于云端依赖、算力不足和功能碎片化，用户体验仍停留在局部智能阶段。

这一阶段的标志性进展主要体现为语音交互规模化、本地 AI 加速能力出现以及轻量视觉算法普及。苹果 Siri、中兴星星系列和小

爱同学等推动语音交互进入高频使用，华为麒麟 970 集成独立 NPU，推动专用 AI 计算成为重要硬件能力；2018—2019 年后，AI 美颜、场景识别等本地视觉能力逐步普及。上述变化说明，模型能否以较低资源开销稳定运行，开始成为产品设计和芯片演进的重要约束。

2.系统融合期（2023—2025 年）：AI 升级为系统级能力

AI 深度融入芯片、系统和应用生态，主要解决“AI 能否成为可复用、可调度的系统能力”的问题。端侧智能进入规模化普及阶段，AI 从附加功能升级为终端操作系统和硬件平台的基础能力。此阶段 AI 能力与芯片、操作系统深度耦合，独立 NPU 逐步成为终端重要配置，操作系统开始内置原生端侧 AI 推理引擎，推动 AI 能力从单一应用调用转向全局系统调度。AI 能力由局部功能扩展为通用能力，开始支持端侧文档处理、内容生成、多模态理解、会议办公、影像创作、个人知识管理等跨应用、跨场景任务。

核心标志包括：端侧 NPU 与异构计算能力显著提升，数亿至数十亿参数级模型具备规模化本地部署条件；INT8 量化工程化成熟，INT4 等更低比特方案开始与硬件能力结合；操作系统和运行时加快集成原生 AI 服务、模型管理与统一调用接口。产业竞争由“是否具备 AI 功能”转向“模型能否高效运行、系统能否统一调度、能力能否形成持续服务”。

3.智能体跃迁期（2026—2030 年）：端侧 AI 由模型运行走向自主任务执行

端侧 Agent 成为新型交互入口，主要解决“AI 能否在本地环境中持续理解任务并完成执行”的问题。端侧智能迈入高阶发展阶段，终端将从被动响应工具加速演进为主动服务助手。此阶段端侧大模型将进一步升级为端侧智能体，具备环境感知、长期记忆、任务规划、决策执行和持续学习能力，能够围绕用户意图自主拆解任务、调用工具、联动应用并完成复杂操作。多设备端侧协同生态加快成型，各类终端将围绕用户和场景进行联合感知、联合决策和服务流转，端侧智能也将进一步向自动驾驶、人形机器人、具身智能等更复杂终端形态延伸。

产业形态也将随之变化。人工智能不再只是既有产品的附加能力，而是逐步参与产品定义、服务组织和价值创造，个人智能体、空间智能、具身智能和行业现场智能等形态加快出现。服务入口由固定应用向智能体迁移，价值实现也由一次性硬件销售向能力订阅、持续服务、任务执行和行业解决方案延伸。端侧智能的竞争焦点将更加集中于模型效率、软硬件协同、系统开放性和真实任务完成能力。

（二）政策持续强化人工智能向端侧和实体场景渗透

1.国家政策由智能硬件培育向“人工智能+”实体应用系统推进

国家层面对端侧智能的支持由早期智能硬件和终端品类培育，逐步延伸至“人工智能+”、智能制造、具身智能和新一代智能设备等方向。政策关注点也从扩大硬件供给，进一步转向模型能力下沉、

软硬件融合、行业场景应用和智能化分级。总体看，端侧智能已成为人工智能进入消费、制造、交通、健康和公共服务等实体场景的重要落脚点。

表 1 国家层面端侧智能相关政策

年份	政策	相关内容
2010 年	《国务院关于加快培育和发展战略性新兴产业的决定》	首次在国家层面提出“智能终端”概念，提出推动新一代移动通信、下一代互联网核心设备和智能终端的研发及产业化
2016 年	《“十三五”国家信息化规划》	加快构建智能穿戴设备、高级机器人、智能汽车等新兴智能终端产业体系和政策环境
2016 年	《智能硬件产业创新发展专项行动（2016-2018 年）》	提升终端产品智能化水平，重点支持智能家居、智能车载、可穿戴设备等
2021 年	《智慧健康养老产业发展行动计划（2021–2025 年）》	明确支持适老化智能终端的研发与应用
2022 年	《虚拟现实与行业应用融合发展行动计划（2022–2026 年）》	将 VR/AR 头显等交互终端列为重点
2023 年	《电子信息制造业 2023–2024 年稳增长行动方案》	明确支持虚拟现实智能终端、车载视听、商用显示等新兴终端
2024 年	《关于加快推进视听电子产业高质量发展的指导意见》	发展适应通用智能趋势的终端，打造智能适老的医疗健康终端，突破元宇宙入口等超级终端
2024 年	《关于推进移动物联网“万物智联”发展的通知》	推动移动物联网终端从“互联”向“智联”升级
2025 年	《关于深入实施“人工智能+”行动的意见》	大力发展人工智能手机和电脑、智能网联汽车、智能机器人等新一代智能终端
2025 年	《数字中国建设 2025 年行动方案》	着力发展智能网联新能源汽车、AI 手机/电脑等新一代智能终端及智能制造装备
2025 年	《电子信息制造业 2025—2026 年稳增长行动方案》	推动手机、计算机向智能化、高端化迈进，打造新型显示、智能可穿戴等新兴产品
2026 年	《“人工智能+制造”专项行动实施意见》	加速智能终端升级，支持端侧模型突破，培育工业巡检、远程医疗等场景的新型终端（AR/VR、脑机接口）

来源：公开信息整理

2.地方政策围绕模型下沉、智能硬件和实体场景形成差异化布局

地方政策正围绕人工智能能力下沉与实体产业融合形成差异化布局。北京突出具身智能和自动驾驶制度创新，上海聚焦智能硬件、低空经济与智能机器人，广东依托消费电子和制造基础强化 AI 硬件供给，重庆、四川则围绕人工智能设备、智能网联汽车、机器人和行业应用构建特色产业体系。各地政策虽然载体不同，但共同指向模型能力与硬件、场景和产业链的深度融合。

北京出台《北京市推动“人工智能+”行动计划（2024-2025 年）》，明确提出鼓励围绕电脑、手机、家电、汽车等新硬件终端，开发嵌入式大模型组件及智能系统，重点发展具身智能机器人等新一代智能终端。在此基础上，专项文件《北京具身智能科技创新与产业培育行动计划（2025—2027 年）》进一步提出“万台落地、千亿集群”目标，力争到 2027 年培育千亿级具身智能产业集群。

上海聚焦智能终端发布专项规划《上海市智能终端产业高质量发展行动方案（2026—2027 年）》，提出到 2027 年产业总体规模突破 3000 亿元，实现三个“一千万”台规模，即人工智能计算机、人工智能手机、人工智能新终端规模各达到千万台以上。在低空经济领域，发布《上海市关于链接长三角加快建设低空经济先进制造业集群的若干措施》，定位“世界 eVTOL 之都”，对 eVTOL 及吨级无人机整机制造项目给予最高 1 亿元支持，关键技术攻关最高 5000 万元，并配套适航取证奖励。

广东将消费电子及智能终端定义为万亿级产业，2026 年发布《广东省增强消费品供需适配性进一步促进消费行动方案》，将其与绿色智能家电、汽车并列，提出到 2027 年推动空调、智能电视、AI 手机年产量均超 1 亿台，并培育民用无人机等百亿级特色消费品。

重庆市发布《重庆市加快推进人工智能终端产业创新蝶变行动计划（2026—2030 年）》顶层规划，目标到 2030 年产业规模超 3000 亿元，推出 20 款 AI 终端爆品、打造 50 个典型场景。实施四大专项行动，升级 AI PC、智能家电等优势终端，发展各类机器人、消费智能终端与算力硬件。

（三）产业价值由硬件增量向智能能力与服务增量拓展

1.硬件载体市场为端侧智能规模化部署提供基础盘

端侧智能正在成为智能终端产业新一轮增长的关键变量，推动全球市场从“存量换机”迈向“智能升级”。尽管全球智能终端行业仍处于周期调整期，但 AI 技术加速渗透正在重塑产品价值、消费预期和产业增长逻辑，AI 手机、AI PC、AI 可穿戴、智能家居、智能汽车、AI 机器人等品类由单点功能创新进入多品类规模化落地阶段。2025 年，全球智能手机、PC 和可穿戴设备出货量分别约为 12.6 亿部、2.85 亿台和 6.12 亿台，主要终端市场总体保持增长。进入 2026 年，存储器供给紧张与成本上升使传统终端总量承压，IDC 预计全球智能手机出货量同比下降 13.9%至 10.9 亿部，全球 PC 出货量同比下降 11.3%；但 AI 终端仍表现出结构性增长动能。Gartner 预计，AI PC 出货量将由 2025 年的 7780 万台增至 2026 年的 1.431 亿

台，市场占比由 31.0%升至 54.7%，率先从创新品类迈向主流形态；生成式 AI 手机也继续向中高端及更广价位段扩散。[1]

2.中国市场为端侧智能规模化落地提供丰富载体

中国市场的 AI 化进程仍在加快，但整体智能手机市场受存储及零部件成本上升、换机周期延长等因素影响，短期承压。IDC 数据显示，2026 年第二季度，中国智能手机市场出货量约 6601 万台，同比下降 4.3%；上半年累计出货量约 1.34 亿台，同比下降 4.2%。与此同时，AI 手机加快进入普及阶段，IDC 预计 2026 年中国新一代 AI 手机出货量将达到约 1.47 亿台，占整体智能手机市场的 53%。可穿戴设备保持较高景气度，2025 年中国腕戴设备市场出货量达到 7390 万台，同比增长 20.8%，其中智能手表出货量 5061 万台，同比增长 17.2%。AI 芯片、健康监测、精准传感和多模态交互等能力持续向可穿戴设备渗透，推动其由数据采集设备向持续智能服务入口演进。[2]

3.产业价值由设备规模向模型、服务与生态增量延伸

从产业价值看，端侧智能的增量不应只以设备出货和硬件规模衡量。现有手机、PC、可穿戴、家居、出行、安防和机器人等载体为模型下沉提供了规模化部署基础，更重要的增量来自模型适配、基础软件、智能体平台、数据与安全服务、行业解决方案以及持续

1 IDC：Worldwide Smartphone Market Results and Forecast，2025—2026 年；IDC：MacBook Neo: Apple's

2 IDC：《IDC：二季度中国智能手机市场延续小幅下滑，头部厂商表现亮眼》，2026 年 7 月；Omdia：《Mainland China smartphone market shipments decline 2% in 2Q26, Huawei and Apple buck the trend》，2026 年 7 月；IDC：《WAIC 智能终端产业质变已至：现场实景和 IDC 数据告诉我们什么？》，2026 年 7 月；IDC：《2025 年中国腕戴设备市场同比增长 20.8%，促销补贴对市场影响持续增强》，2026 年 3 月。

智能服务。**随着 AI 能力从功能叠加转向系统级调用，产业价值链将由“硬件销售”进一步延伸至“模型部署—系统运行—任务服务—生态运营”。**

表 2 国内端侧智能典型载体市场规模及出货量参考

设备类型	2024 年	2025 年	2030 年（测算）
AI 手机	出货 6300 万台，规模 2600 亿元	出货 1.16 亿台，规模 5100 亿元	规模 1.5 万亿元，出货 3.8 亿台，出货量 CAGR26.8%（2025-2030，下同）
AI PC	出货 580 万台，规模 280 亿元	出货 890 万台，规模 470 亿元	规模 2300 亿元，出货 4200 万台，CAGR36.2%
AI 可穿戴设备	出货 2580 万台，规模 600 亿元	出货 4670 万台，规模 1160 亿元	规模 4200 亿元，出货 1.5 亿台，CAGR26.3%
智能家居	出货 2.78 亿台，规模 7800 亿元	出货 2.81 亿台，规模 9300 亿元	规模 1.6 万亿元，出货 5.0 亿台，CAGR12.2%
AI 安防设备	出货 7000 万台，规模 1480 亿元	出货 8800 万台，规模 1743 亿元	规模 3500 亿元，出货 2.1 亿台，CAGR19%
AI 机器人	出货 850 万台，规模 1800 亿元	出货 1200 万台，规模 2600 亿元	规模 7500 亿元，出货 4610 万台，CAGR30.8%

来源：中国信通院根据公开资料整理；2030 年数据为本报告综合测算

整体来看，端侧智能正从硬件功能增量转向能力与服务增量。短期内，轻量模型、本地推理和多模态能力仍将依托成熟设备扩大覆盖范围；中长期看，模型运行时、智能体服务、空间智能、具身智能和行业现场智能将成为新的价值承载方式。产业竞争重点也将由单机参数进一步转向模型效率、硬件适配、系统调度、安全治理和任务服务能力，形成覆盖芯片、模型、系统、应用和服务的持续价值链。

（四）端侧智能是以本地为主、端云协同的融合智能形态

端侧智能是以终端设备为入口，优先使用本地软硬件资源，完

成感知、推理、决策与执行的本地化智能服务，结合云端算力、模型与知识能力，完成复杂任务的融合智能形态。其核心不是某一种具体设备，而是将模型能力与计算、存储、传感、通信、操作系统和执行部件结合，使 AI 尽可能在靠近数据和任务的位置完成处理，并通过端云协同补充复杂知识、训练更新和高强度计算。与主要依赖云端集中处理的模式相比，端侧智能更强调低时延、隐私保护、离线可用、个性化适配和现场执行。

从狭义看，端侧智能主要指“本地 AI 运行能力”，即将模型、算法和推理决策能力部署在设备侧或现场侧，依托 CPU、GPU、NPU、DSP 等计算单元完成指令识别、环境感知、任务分配、推理决策、任务执行与结果处理功能。其典型特征是“模型在本地运行、数据就近处理、结果即时产生”，重点解决云端模式在网络时延、带宽成本、隐私保护、离线可用和实时控制方面的不足。

从广义看，端侧智能并不等同于简单的“本地运行模型”，而是以本地智能为能力基础、以端云协同为运行模式，涵盖模型、芯片、传感器、存储、基础软件、智能体框架、应用服务、安全机制和运营生态的全栈式能力体系。其中，模型需要通过轻量化架构、量化、剪枝、蒸馏、稀疏化等方式适应资源约束；硬件需要通过异构计算、低精度运算、存算协同和低功耗设计提供运行基础；基础软件需要完成编译转换、算子映射、资源调度和能力调用；上层应用则进一步把模型能力转化为持续感知、任务执行和场景服务。

从概念边界看，端侧智能与边缘智能、端侧大模型和端侧智能

体既有联系也有区别。边缘智能更强调人工智能与边缘计算基础设施融合，范围可覆盖边缘服务器、边缘网关和网络侧节点；端侧智能更关注 AI 能力在数据产生侧和任务执行侧的本地运行及软硬件融合。端侧大模型是端侧智能的重要模型基础，重点解决复杂模型在有限资源下的部署效率问题；端侧智能体则进一步将模型、记忆、工具和执行能力组织起来，使 AI 从“能理解、能生成”走向“能规划、能行动”。

端侧智能的核心特征主要体现在五个方面：**一是**本地化，关键数据和高频任务尽可能在数据产生侧处理；**二是**高效化，通过模型轻量化和软硬件协同在有限资源下提升能力密度；**三是**实时性，缩短数据传输与云端往返链路，支撑即时反馈和实时控制；**四是**隐私性，通过本地处理、权限隔离和最小化外发降低敏感数据暴露；**五是**协同性，根据任务复杂度、设备能力和网络条件，在不同计算节点之间动态分工。随着智能体能力成熟，端侧智能还将进一步具备记忆、规划、工具调用和持续服务能力。

从产业边界看，端侧智能覆盖从基础硬件、模型与基础软件，到智能原生产品、应用服务和安全治理的完整链条。上游提供 AI 芯片、传感器、存储、通信与执行部件，中游围绕轻量模型、推理框架、操作系统、运行时和智能体平台完成能力组织，下游则在个人服务、空间智能、移动智能、行业现场和具身任务等领域形成新产品、新业态和新模式。总体来看，端侧智能正在推动人工智能从“集中式模型服务”走向“分布式智能能力”，并把 AI 进一步嵌入

真实世界的感知、决策与执行过程。

二、端侧智能技术体系

端侧智能技术架构正围绕模型轻量化与软硬件协同持续演进。面对终端设备有限的算力、存储、电量等资源条件下，如何解决“算不动、装不下、跑不久”的问题，是端侧智能技术发展演进的主要方向。不同设备在算力、存储、功耗、感知和通信条件上的差异，端侧智能需要通过高效模型架构、模型压缩、推理优化、异构计算、编译适配、资源调度和端云协同等技术，实现模型能力与多样化硬件环境的高效匹配。整体上，如图 2 所示，技术体系由智能硬件层、基础软件层、端侧模型层、端侧应用层和安全信任层构成，各层围绕人工智能任务需求协同演进，共同支撑模型从部署、运行到感知、推理和执行的完整链路。



来源：中国信息通信研究院

图 2 端侧智能技术架构图

（一）智能硬件层：异构计算与高效存储支撑模型稳定运行

智能硬件层决定端侧模型可运行的能力边界。其发展重点正从单纯追求峰值算力，转向计算、存储、感知和能效协同优化，使模型能够在不同资源条件下保持较低时延、较小内存占用和可持续运行。

1.异构计算与低精度加速提升本地推理效率

CPU、GPU、NPU **等计算单元各有优势，端侧智能逐步通过异构协同匹配不同模型算子和任务阶段。**NPU 承担矩阵与张量计算，CPU 负责控制与通用逻辑，GPU 适合并行计算和多媒体处理；INT8、INT4 等低精度计算以及结构化稀疏进一步降低计算量和能耗。硬件优化重点正在由“算力更高”转向“单位功耗下有效 AI 算力更高”，为轻量模型和多模态模型持续运行提供基础。**一是**强化 AI 专用计算与异构集成，高通 Hexagon、联发科天玑系列持续增强面向 Transformer 和生成式 AI 的专用计算能力，紫光展锐 T9100、瑞芯微 RK3588 等也通过 CPU、GPU 与 NPU 协同提升端侧模型运行能力。**二是**支持低精度与稀疏执行，Arm Ethos-U85、瑞芯微 RK3588 和天玑 9500 通过相应数据格式与计算支持，将压缩模型的资源优势转化为执行收益。**三是**由单一 AI 加速向系统级协同演进，英特尔 Lunar Lake、意法半导体 STM32N6 等通过处理器、AI 加速单元和输入处理能力集成提升端侧 AI 执行效率，紫光展锐、地平线等也在芯片、软件工具链和模型适配方面持续完善。

2.内存与存储优化缓解模型权重和运行状态压力

统一内存、高速存储与存算协同配合，支撑模型加载、长上下文和多模态处理。一是扩大权重与缓存的承载空间，此芯 P1、AMD Ryzen AI Max+ 395 和 DGX Spark 通过共享内存及容量分配支持本地模型运行。**二是**衔接运行内存与持久存储，长鑫、三星 LPDDR5X 提高数据传输能力，Lunar Lake 采用封装内存，天玑 9500 以 UFS 4.1 接口支持模型加载。**三是**减少外部访问，AMD XDNA 通过本地存储复用数据，WTM-8 和天玑 9500 进一步结合存储与计算，降低数据搬移开销。

3.感知、能效与执行协同支撑持续智能服务

端侧智能需要长期接收视觉、语音、位置和环境信息，并将推理结果转化为交互反馈或设备动作。传感器前端处理、低功耗常驻感知、动态电压频率调节和温度反馈等机制有助于降低持续运行开销；显示、音频、触觉和执行器则把模型结果连接到真实任务。硬件体系因此从“提供算力”进一步转向“支撑感知—推理—执行闭环”。

总体来看，智能硬件的演进重点是围绕模型运行效率重构资源组合，通过异构计算提升有效算力，通过内存与存储优化降低数据搬移，通过低功耗感知和执行能力延长智能服务链路。未来硬件设计将更加重视与模型结构、量化精度和任务特征的协同。

表 3 以人工智能任务需求为牵引的智能硬件技术分层

技术方向	关键技术	具体机制
1 异构计算与低精度加速技术	1.1 神经网络专用加速（NPU、AI 加速器）	矩阵、向量与张量专用计算
	1.2 CPU/GPU/NPU 异构计算与集成	多处理单元协同执行

表 3 以人工智能任务需求为牵引的智能硬件技术分层（续）

技术方向	关键技术	具体机制
1 异构计算与低精度加速技术	1.3 低精度与稀疏计算	INT8/INT4 低比特计算与结构化稀疏执行
2 内存与存储优化技术	2.1 统一内存架构	共享内存与灵活容量分配
	2.2 分层存储	模型权重、KV Cache 等分层存放与复用
	2.3 内存带宽与数据访问优化	减少数据搬移与访存开销
3 感知、能效与执行协同技术	3.1 低功耗感知与前端处理	多源传感采集、前端融合与事件触发
	3.2 动态电压频率调节与温度管理	DVFS 与温度反馈调节
	3.3 交互输出与动作执行	显示、音频、触觉及执行器驱动

（二）基础软件层：编译适配与资源调度打通模型到硬件执行链路

基础软件层承担模型与多样化硬件之间的适配和组织作用。其核心价值是把模型结构、算子和运行状态映射到具体计算资源，并根据内存、功耗、温度和任务优先级动态调度，使同一类模型能够在不同芯片和系统环境中稳定运行。

1.编译与运行时适配降低跨硬件部署成本

统一编译与后端接口，优化计算和加载过程，提升模型跨终端部署效率。一是以 LiteRT 模型转换、ONNX Runtime 执行提供程序和 ExecuTorch 子图委托，衔接模型表示、算子支持与硬件执行。**二是**通过 Core ML 注意力融合内核、MNN-LLM 数据重排和 FlagOS 公共算子、编译组件，减少重复适配与执行开销。**三是**通过模型预加载、驻留控制与执行器预热，减少模型调用时的准备开销。例如，Ollama 通过预加载和内存驻留控制，减少加载等待与重复加载；苹

果通过模型执行器预热，提前完成部分初始化工作。运行框架则负责推理过程中的会话管理、缓冲区管理和结果输出。

2.资源调度提升持续运行稳定性

端侧模型需要与前台应用、传感任务和系统后台共享有限资源。内存预算、模型驻留、并发控制、缓存复用和温度感知调度可以避免资源争抢和频繁加载；在功耗或温度达到阈值时，系统还可降低并发、调整模型规模或延后非紧急任务。由此，端侧优化从一次推理性能进一步转向长时间运行的稳定性。

结合资源预算、任务优先级与设备状态，协调常驻感知、前台推理和后台维护。一是通过 FreeToken 的 CPU-GPU 协同、专家缓存，以及 Core ML 缓存原位更新、MNN-LLM 分层存储，控制计算与数据搬移开销。**二是**以 Ollama 的并发、排队和模型驻留上限约束负载，并借助 CHRE 低功耗任务下沉、WorkManager 条件调度，减少资源争用与无效唤醒。**三是**利用 ADPF 温度反馈和性能提示，由应用调整并发与非紧急任务，将短时峰值优化延伸到持续响应。

3.智能体运行与端云协同扩展系统级能力

随着端侧应用从问答走向任务执行，基础软件开始承担上下文管理、工具调用、状态保持和任务编排。一是通过上下文管理与状态持久化，支撑多步骤任务的连续执行。如，LanguageModelSession 与 Transcript 记录交互及工具结果，配合运行框架的状态和检查点衔接暂停、恢复与结束。**二是**通过将复杂任务拆分为子任务，依据任务类型选择执行主体，按照步骤之间的依赖关系组织顺序执行或并

行处理，再汇总各子任务的结果，形成后续步骤的输入或最终输出。如，利用 LangGraph 的路由、并行和协调者—工作者机制，组织多智能体分工及结果汇总。**三是**通过权限隔离、操作确认与重复执行控制，约束智能体对应用、文件和外部服务的实际操作。例如，Windows 智能体账户、隔离工作区和运行框架的中断确认约束真实操作，并按步骤设置恢复路径。

结合任务需求、模型可用性与数据边界，衔接本地处理和云端补充。一是通过任务路由与执行策略，选择端侧或云端模型，并明确切换条件，如 Firebase AI Logic 的端侧优先、云端优先及仅端侧等模式。**二是**通过统一调用与上下文管理，衔接模型切换前后的任务信息，如苹果 Foundation Models 通过动态会话配置选择端侧或私有云计算（PCC）模型，并管理交互历史、指令与工具。**三是**通过云端环境验证与请求加密保护外发数据，如 PCC 的节点验证和加密机制；同时针对模型不可用、网络异常或配额不足配置替代模型、延后执行等处理路径。

总体来看，基础软件正从单一推理框架扩展为“编译适配—资源调度—任务编排—端云协同”的完整运行链路。其成熟度直接决定模型能否低成本适配多样化硬件，并将局部模型能力转化为稳定、连续的系统级智能服务。

表 4 以人工智能任务需求为牵引的基础软件技术分层

技术方向	关键技术	具体机制
1 模型编译与推理运行技术	1.1 模型转换与编译	模型格式转换与目标编译
	1.2 计算图优化与算子融合	计算融合与中间数据消减

表 4 以人工智能任务需求为牵引的基础软件技术分层（续）

技术方向	关键技术	具体机制
1 模型编译与推理运行技术	1.3 硬件抽象与算子映射	统一后端接口与硬件执行映射
	1.4 推理框架、引擎与运行时	推理会话与运行流程组织
2 系统资源管理与调度技术	2.1 异构计算资源调度	CPU/GPU/NPU 任务分配
	2.2 内存管理与缓存分配	KV Cache 与内存预算管理
	2.3 模型驻留与并发控制	模型预加载、驻留与排队
	2.4 温度感知任务调度	依据温度和能耗动态调整负载
3 智能体运行与端云协同技术	3.1 上下文与状态管理	保持多步骤任务状态
	3.2 工具调用与任务编排	组织工具和应用调用
	3.3 端云任务路由与上下文传递	依据任务复杂度、数据敏感性与网络条件选择执行位置
	3.4 异常回退与离线降级	模型不可用或网络异常时进行策略回退

（三）端侧模型层：轻量化与高效推理提升有限资源下的能力密度

端侧模型面向有限计算、内存、带宽和功耗环境设计或适配，关键不是简单缩小参数量，而是在任务质量、响应时延、运行内存和能耗之间获得更高的“能力密度”。**当前技术重点主要集中在高效架构、模型压缩、推理优化以及多模态与个性化适配。**

1.高效架构由“缩小模型”转向提升单位参数能力

端侧模型逐步采用深窄结构、参数共享、高效注意力和弹性子模型等设计，在较小参数规模下保持语言理解和任务处理能力。模型设计开始把目标设备的内存、带宽和执行特征纳入架构选择，通过减少冗余参数和计算路径，提高有限资源条件下的有效能力，而不是简单复制云端大模型并进行缩放。

2.量化、蒸馏与稀疏化成为模型轻量化核心手段

低比特量化通过降低权重和激活精度显著减少存储、带宽和计算开销，知识蒸馏将大模型能力迁移至小模型，剪枝与稀疏化则减少结构冗余。当前优化重点正从“压得更小”转向“在较低精度和更小规模下保持任务质量”，并根据芯片对 INT8、INT4 及稀疏计算的支持情况选择适配策略，使压缩收益真正转化为端侧运行收益。

3.推理优化与多模型协同降低持续运行开销

KV Cache 压缩、提示缓存、推测解码、动态计算和提前退出等技术可减少重复计算和运行状态开销。与此同时，大小模型、通用模型与专用模型之间的协同逐步成为重要路线：高频简单任务由小模型处理，复杂任务按需升级到更强模型或云端模型。模型路由使端侧系统能够按照质量、时延、功耗和成本动态选择最合适的执行路径。

4.多模态与个性化适配拓展端侧任务能力

视觉、语音、文本和传感数据正在进入统一模型体系，高效视觉编码和紧凑模态表示有助于降低多模态输入开销。基于 LoRA、Adapter 等参数高效方法的个性化适配，则使模型能够结合用户或场景数据形成差异化能力。端侧模型由通用问答向多模态感知、工具调用和场景适配扩展，成为智能原生产品的重要能力基础。

总体来看，端侧模型轻量化正在从“把大模型做小”转向“围绕任务、资源和硬件共同优化”。高效架构提升单位参数能力，量化、蒸馏与稀疏化降低部署门槛，推理优化和模型协同控制持续运

行开销，多模态与个性化适配进一步拓展任务边界。未来评价端侧模型的重点将更多转向任务效果、资源占用、能耗和时延的综合平衡。

表 5 以人工智能任务需求为牵引的端侧模型技术分层

技术方向	关键技术	具体机制
1 高效模型架构技术	1.1 紧凑网络与参数共享	深窄结构与权重共享
	1.2 高效注意力结构	分组查询等高效注意力
	1.3 资源约束下的弹性架构	依据内存、带宽和任务需求选择子模型或计算路径
2 模型压缩技术	2.1 低比特与混合精度量化	INT8/INT4 及 PTQ、QAT
	2.2 知识蒸馏	教师能力向小模型迁移
	2.3 模型剪枝与稀疏化	结构裁剪与稀疏执行
3 推理优化与多模型协同技术	3.1 KV Cache 压缩与复用	减少运行状态和重复计算
	3.2 提示缓存、推测解码与动态计算	候选生成集中验证并按需调整计算量
	3.3 大小模型协同与动态路由	简单任务由小模型处理，复杂任务按需升级
4 多模态与个性化适配技术	4.1 高效多模态表示与融合	紧凑视觉、语音表示与跨模态融合
	4.2 参数高效微调（LoRA、Adapter）	轻量参数更新
	4.3 用户与场景适配	结合用户或场景数据形成差异化能力

（四）端侧应用层：智能体与原生应用推动 AI 由功能调用走向任务执行

端侧原生应用将本地模型、个人数据和设备功能结合，使终端能够直接理解当前情境、处理私人信息并完成具体操作。不同算力、感知和交互能力的设备可作为智能服务入口，也提供文件、屏幕、语音、设备状态和用户授权等任务条件。围绕这些终端能力，应用形态逐步从独立问答扩展到**个人智能体**、**交易代理**、**记忆助理**、**实时交互**和**本地创作**。涉及外部信息、商业服务和交易结算的环节，则继续与云端协同。

1.个人执行类应用推动智能服务走向任务执行

Claw 类应用将任务规划、工具调用和持续记忆组织在个人设备上，使用户能够通过统一入口调用文件、应用和外部服务。一是以常驻个人智能体承接多步骤任务。例如，OpenClaw 将邮件、日历等工具接入统一交互入口，并支持通过本地模型配置连接 llama.cpp、Ollama 和 LM Studio。**二是**以本地函数调用完成明确的设备操作。例如，谷歌 AI Edge Gallery 中的 Mobile Actions 利用 FunctionGemma 解析自然语言，将指令转换为创建日历事件、开启手电筒等系统功能调用。**三是**将个人助理延伸至家庭设备控制。例如，Home Assistant 的 Ollama 集成允许本地模型通过 Assist 接口查询和控制已授权设备，将家庭状态与自然语言指令连接起来。该控制能力仍标注为实验性功能，但已提供了在家庭中枢或局域网计算设备上组织本地智能服务的具体路径。

2.交易代理类应用拓展智能服务交易闭环

Agent 支付使智能体能够在授权范围内发起交易，终端则承担交互确认、钱包调用或交易签名等关键环节。一是通过支付技能连接个人智能体与商业服务。例如，支付宝 AI 付已提供面向本地和云端智能体的接入方案，支持 OpenClaw、QClaw、QwenPaw 等应用，经身份核验和账户绑定后接入下单支付能力。**二是**通过本地钱包和硬件签名保留交易控制。既借助智能体实现交易流程的自动化发起与流程推进，又将交易控制权保留在用户终端侧，为智能体自主完成商业服务闭环探索出了可行的落地路径。

3.私人记忆类应用支撑长期记忆与任务接续

记忆类应用利用终端保存的文件和交互历史，为用户提供跨时间、跨应用的信息查找与任务衔接能力。一是将屏幕和交互记录转化为可检索的个人历史。例如，Windows Recall 在用户主动开启后，本地保存和分析屏幕快照，支持通过自然语言查找曾经浏览的内容；screenpipe 进一步整合屏幕、音频转录和应用信息，并通过本地接口向获准的智能体提供历史上下文。前者侧重找回信息，后者进一步支持基于工作记录生成会议跟进和任务摘要。**二是**将本地文档组织为可问答的私人知识库。例如，AnythingLLM 支持导入 PDF、文档、CSV 和代码等资料，并配置本地模型开展检索与问答。相较于每次重新上传文件，这类应用可持续维护项目资料和知识索引，支撑跨文档查询、资料核对与内容整理。采用本地模型和本地检索配置时，核心资料处理能够在个人设备内完成。

4.即时交互类应用提升本地实时响应能力

即时交互类应用直接处理终端采集的输入，在对话、观察和操作发生时提供反馈，减少网络等待对体验的影响。一是以端侧语音处理支撑面对面沟通。例如，三星 Interpreter 与 Live Translate 通过端侧语音识别、翻译和语音合成，支持现场交流与通话翻译。**二是**以本地视觉识别提供日常辅助。例如，谷歌 Lookout 的文字识别和物体分类采用端侧机器学习，帮助视障用户读取文字、识别周围物品。此类应用将摄像头输入转化为可听取的信息，端侧计算直接服

务于用户正在进行的观察与行动。**三是**以本地指令理解形成交互娱乐体验。

5.内容生产类应用推动端侧创作高效生成

内容生产类应用在素材所在设备上完成生成、转录或编辑，支持围绕个人资料开展连续创作。一是通过本地图像生成缩短素材创作流程。例如，Draw Things 支持在设备上离线运行图像生成模型，使用户能够围绕本地素材进行反复生成与调整。其端侧价值在于模型和创作素材可以保留在同一设备中，支持离线工作；选择云端计算选项时，则形成相应的混合处理流程。**二是**将本地转录与文本处理连接起来。例如，MacWhisper 默认在设备上完成语音转录和说话人识别，并可连接 Ollama 或 LM Studio 开展本地文本处理。录音由此可以进一步整理为访谈文字、会议摘要和创作素材，将原始音频处理与后续写作衔接在个人设备内。

综上，端侧原生应用正围绕个人信息、实时感知和设备操作能力，推动人工智能从问答助手进一步走向持续理解、即时响应和任务执行。通过个人智能体连接文件、应用和设备工具，通过私人记忆积累跨时间的个人上下文，通过实时语音与视觉处理获取当前环境信息，通过内容生产工具直接利用本地素材，并通过交易代理将任务链路延伸至外部商业服务，形成从信息获取、情境理解到操作执行的应用闭环。各类应用相互衔接，使终端由单一的模型调用入口逐步转变为能够利用个人数据、感知当前状态并调动设备能力的智能服务载体；对于外部知识、复杂计算、跨设备协同和交易结算

等环节，则继续通过端云协同完成，共同拓展端侧智能的实际应用边界。

表 6 端侧智能原生应用

应用方向	关键场景	场景介绍
个人执行	多步骤个人事务处理	将自然语言任务与文件、日历、邮件等工具连接，通过任务规划、工具调用和状态保留，连续完成资料整理、日程安排与工作跟进。如 OpenClaw、QwenPaw 等个人智能体可接入本地模型，但实际推理位置取决于模型配置。
	自然语言设备操作	在终端解析用户指令，将操作意图转换为具体的系统功能调用，完成创建日历事件、开启手电筒等任务。如 AI Edge Gallery 的 Mobile Actions，围绕明确的设备功能提供本地指令理解与执行能力。
	家庭设备查询与控制	将家庭设备状态、用户指令与已授权的控制接口连接，通过本地模型查询设备信息、调用设备功能，在家庭中枢或局域网计算设备上组织智能服务。如 Home Assistant 的 Ollama 集成，其设备控制能力仍标注为实验性功能。
交易代理	商品与数字服务购买	将下单和支付能力接入智能体任务流程，在完成身份核验、账户绑定及用户授权后，衔接商品或数字服务购买。终端承担交互确认等环节，商业服务与支付处理通过端云协同完成，如接入支付宝 AI 付的个人智能体。
	数字资产交易授权与签名	由智能体准备交易，利用本地钱包保管密钥，并通过终端或硬件设备完成交易核验与签名，再交由网络处理交易。
私人记忆	个人历史检索与任务衔接	在用户授权下，将屏幕内容、音频转录或应用活动组织为可检索的个人历史，支持找回曾经浏览的信息，并为后续任务提供上下文。如 Windows Recall 侧重历史内容查找，screenpipe 进一步支持基于工作记录生成任务摘要与会议跟进。
	本地文档知识问答	将个人文档、表格、代码等资料组织为持续维护的知识库，通过语义检索与模型问答开展跨文档查询、资料核对和内容整理。如 AnythingLLM，采用本地模型与本地检索配置时，核心资料处理可在个人设备内完成。

表 6 端侧智能原生应用（续）

应用方向	关键场景	场景介绍
即时交互	现场交流与通话翻译	在终端连续完成语音识别、语言翻译和语音合成，将对话输入及时转化为翻译结果，支持面对面沟通与通话交流，减少对网络传输的依赖。如三星 Interpreter 与 Live Translate。
	视觉识别与无障碍辅助	对摄像头采集的文字和物体进行本地识别，将视觉内容转化为用户可听取的信息，辅助视障用户读取文字、辨识周围物品。如谷歌 Lookout 的端侧文字识别与物体分类功能。
	离线互动娱乐	在设备上理解自然语言指令，并将其转换为游戏或交互应用中的函数调用，由应用依据自身规则执行操作并反馈结果。如 AI Edge Gallery 的 Tiny Garden，通过本地模型完成种植、浇水等游戏操作，也为互动教学提供参考。
内容生产	本地图像生成与调整	在素材所在设备上运行图像生成模型，围绕本地素材反复生成和调整内容，支持离线创作与连续编辑。如 Draw Things；采用云端计算选项时，则形成端云混合创作流程。
	音频转录与文本整理	将本地录音转化为文本，并结合说话人识别和本地语言模型，进一步整理为访谈文字、会议摘要与创作素材。如 MacWhisper，可连接 Ollama 或 LM Studio，将音频转录与后续文本处理衔接在个人设备内。

（五）端侧安全信任层：构建覆盖模型、数据与智能体执行的可信边界

端侧智能把模型能力进一步贴近数据和真实执行环境，也同步扩大了模型资产、敏感数据和智能体操作的风险边界。安全治理不宜停留在传统设备防护，而需要围绕身份权限、模型数据安全、智能体执行和可信运行形成贯穿任务全流程的控制机制。

1.身份权限与数据保护明确本地智能使用边界

端侧智能需要区分用户、设备和智能体身份，并按任务限定可访问的数据对象、工具和操作范围。数据最小化、本地优先处理、加密隔离和生命周期管理可减少敏感信息暴露。对于智能体通过 GUI 或应用接口执行任务，应按照最小权限原则授予必要能力，明确可访问的应用、数据对象、操作类型和授权有效范围，避免因用户已登录应用而默认继承用户全部权限。对于涉及敏感数据和高风险操作的任务，应建立与任务范围相匹配的授权机制，为后续安全执行提供明确权限边界。

2.模型与内容安全保障能力来源和输出可信

模型在压缩、微调和更新过程中可能面临投毒、后门、篡改和能力漂移等风险，需要对数据来源、模型包完整性和关键行为进行验证。输入输出风险检测、鲁棒性测试和内容防护则用于降低异常输入和不当输出影响。模型“可运行”并不等于“可信运行”，安全验证需要贯穿部署和版本演进过程。

3.智能体行为与执行安全成为新增重点

智能体具备工具调用和自主执行能力后，提示注入、工具投毒、记忆污染、越权操作和重复执行等风险会直接影响真实业务。应通过指令与外部数据分离、工具准入、最小权限、沙箱隔离等机制，把模型规划限制在授权范围内。对于支付、转账、删除文件、发送敏感信息、账号设置变更等高风险任务，应在最终执行前展示操作对象、金额或数量、接收方、影响范围等关键参数并由用户确认；关键参数发生变化时应重新确认。同时，应对批量操作、连续操作

和异常跳转设置数量、频次和范围限制，并支持任务中止、防重放和结果核验。

4.可信运行与审计追踪支撑持续风险控制

可信执行环境、安全启动、代码与模型签名等机制用于保护关键资产和运行基础；行为轨迹、调用链记录和异常检测则为任务复核、风险阻断和事后追踪提供依据。随着端侧智能从单次推理走向持续服务，安全能力也需要从静态防护转向可监测、可干预、可恢复的运行治理。

总体来看，端侧安全信任的关键是将身份、数据、模型和智能体执行纳入同一任务链路，在本地处理和端云协同之间保持清晰的数据与权限边界，并通过可信运行和行为审计控制实际操作风险。安全能力将成为智能原生产品规模化落地的重要基础。

表 7 以人工智能任务需求为牵引的安全治理技术分层

技术方向	关键技术	具体机制
1 身份权限与数据保护技术	1.1 用户、设备与智能体身份认证	主体绑定与委托关系核验
	1.2 最小权限与动态授权	按任务限定数据、工具和操作权限
	1.3 数据最小化、加密隔离与生命周期管理	优先本地处理并控制数据留存与外发
	1.4 高风险操作确认	删除、支付、外发等操作绑定具体参数确认
2 模型与内容安全技术	2.1 模型完整性校验	对模型包和版本进行签名与完整性核验
	2.2 训练数据投毒与后门防护	识别异常样本与触发行为
	2.3 对抗输入与输入输出风险检测	通过鲁棒性测试和风险检测控制异常输入输出
	2.4 内容过滤与拒答	按场景过滤、脱敏或拒答

表 7 以人工智能任务需求为牵引的安全治理技术分层（续）

技术方向	关键技术	具体机制
3 智能体行为与执行安全技术	3.1 提示注入与工具投毒防护	分离用户指令与外部数据
	3.2 记忆写入审核	核验长期记忆来源与适用范围
	3.3 越权防护与沙箱执行	限定实际调用范围并隔离执行
	3.4 支付与交易操作安全	对交易对象、金额和状态实施授权绑定与防重放
4 可信运行与审计追踪技术	4.1 可信执行、安全启动与签名校验	保护关键资产和运行基础
	4.2 行为轨迹与调用链记录	关联身份、授权、模型版本、工具调用与结果
	4.3 异常检测与风险阻断	识别异常行为并暂停任务或撤销凭据
	4.4 操作回滚与恢复	对可逆操作进行恢复或业务补偿

三、端侧智能原生产品形态

模型轻量化与多样化硬件融合，正在催生一批区别于传统“硬件+功能”的智能原生产品。其共同特征是：**产品从设计阶段就围绕感知、模型、记忆、推理和执行能力组织软硬件资源，模型不再只是附加应用，而是参与交互入口、系统调度和任务闭环。**根据使用环境、实时性和执行方式差异，可形成个人智能产品、空间智能产品、移动智能产品、行业智能节点和具身智能产品等典型形态。不同形态的差异本质上来自模型能力与硬件资源、传感方式和执行部件的不同组合。

（一）个人智能产品：多类硬件协同形成智能入口

1.模型能力与交互硬件融合形成多形态个人智能产品

个人智能产品主要围绕用户的高频交互、知识处理、持续感知和即时服务形成。高算力设备适合承载多模态理解、内容生成、个人知识检索和系统级任务执行，低功耗可穿戴设备更适合语音唤醒、健康感知、环境识别和轻量提醒，第一视角和空间交互设备则进一步拓展视觉感知与空间信息叠加。不同硬件形态主要体现算力、续航、传感和交互条件的差异，其智能能力取决于模型运行效率以及与传感器、系统接口和用户数据的协同程度。

表 8 端侧智能原生产品典型形态与融合特征

产品形态	代表载体	模型与硬件融合重点	主要演进方向
个人智能产品	高算力交互设备、低功耗可穿戴设备、空间交互设备	在不同算力与续航条件下分配多模态理解、个人知识、持续感知和任务执行	由单机功能向个人智能入口演进
空间智能产品	家庭中枢、智慧屏、传感与环境控制设备	融合多源感知、本地模型、空间状态与多设备执行	由单品联动向环境级主动服务演进
移动智能产品	车辆、无人配送设备、无人机等	模型与视觉/雷达/定位、实时计算和控制系统协同	由信息交互向自主感知与移动决策演进
行业智能节点	摄像头、边缘网关、巡检与检测设备	轻量模型适配低功耗/复杂环境硬件，形成现场识别与即时响应	由数据采集向现场智能闭环演进

2.系统级模型、个人数据与设备能力协同推动产品 AI 原生化

个人智能产品的演进重点正在从“增加 AI 功能”转向“围绕 AI 重新组织产品能力”。一方面，本地模型与个人知识、文件、屏

幕、传感器和应用能力结合，形成可持续使用的上下文；另一方面，系统级智能体通过工具调用把自然语言理解进一步连接到真实操作。多设备之间也不再只是同步内容，而是根据算力、交互方式和使用状态动态分工，使服务在不同硬件之间连续流转。

表 9 不同硬件资源条件下的模型部署与协同方式

硬件资源类型	典型环境	模型部署重点	协同方式
低功耗感知型	穿戴、传感、门锁等	小模型、事件触发、低比特推理、常驻感知	本地完成唤醒/检测，复杂任务升级至高算力节点
高算力交互型	个人计算与交互设备	多模态模型、个人知识、内容生成、系统级工具调用	本地承担高频与隐私任务，云端补充复杂推理
实时控制型	车辆、无人设备等	低时延模型、确定性执行、安全冗余	模型推理与控制系统闭环，云端用于训练和能力增强
现场计算型	网关、工业与安防设备	轻量模型、多路数据处理、长期稳定运行	端侧实时识别，边缘/云端进行复杂校验与运维

3.资源差异、系统接口与数据边界决定个人智能产品体验上限

个人智能产品规模化发展仍面临资源差异、系统开放和数据治理等共性问题。不同硬件在算力、内存、续航和散热条件上差异明显，同一模型能力难以直接复制；系统级任务执行又依赖应用接口、权限体系和工具开放程度；个人记忆、屏幕内容、语音图像、健康与位置等数据高度敏感。未来需要通过模型分级、动态路由、统一能力接口和最小权限机制，把差异化硬件组织成稳定、连续、可信

的个人智能服务体系。

（二）空间智能产品：感知、决策与执行融合形成环境级智能

1.多源感知与本地模型融合推动设备升级为空间智能节点

空间智能产品面向家庭、办公、酒店、园区等相对稳定空间，将摄像头、毫米波、声音、温湿度、位置和设备状态等多源信息与本地模型结合，使环境具备“感知—理解—决策—执行”能力。智慧屏、音箱、门锁、家电、环境控制设备和本地中枢等不再只是独立产品，而是承担不同的感知、交互或执行角色。空间智能的核心不是连接设备数量，而是能否通过轻量模型和统一编排把分散硬件组织成面向空间状态和服务目标的整体系统。

2.统一感知、空间记忆与任务编排推动环境级主动服务

空间智能正从固定规则联动转向基于情境理解的主动服务。本地中枢可汇聚设备状态、短期上下文和长期习惯信息，端侧节点负责实时感知与即时执行，较复杂的知识和推理再由云端补充。通过空间记忆、任务编排和多设备协同，同一服务可以跨照明、环境、安防、健康、会议或能源等设备完成，形成由“设备控制”向“环境服务”的产品形态升级。

3.协议互通、感知可靠与存量设备适配影响空间智能规模化

空间智能规模化仍受协议互通、感知可靠性、隐私安全和存量设备适配影响。不同厂商能力描述和控制接口不一致，跨品牌任务编排容易中断；人员身份、行为和环境状态识别不准会放大错误联动风险；持续采集的空间数据也要求更严格的本地处理和权限隔离。未来空间智能需要在开放接口、模型鲁棒性、隐私边界和低成本改造之间形成平衡。

（三）移动智能产品：本地实时推理支撑自主感知与移动决策

1.高实时与高安全要求推动模型、传感和控制形成端侧闭环

移动智能产品面向持续变化的环境和高实时任务，将视觉、雷达、定位、设备状态等传感输入与本地模型、实时计算平台和控制系统连接。车辆、无人配送设备和无人机等形态虽然硬件不同，但共同要求模型在严格时延、功耗和安全边界下持续运行，并能够把感知结果直接转化为导航、规划和控制动作。其产品竞争力因此来自模型能力、传感融合、计算平台和控制闭环的协同，而非单一功能配置。

2.端到端模型与协同感知推动移动智能由辅助判断走向任务决策

移动智能正在从分立感知与规则控制，向端到端模型、统一计算和协同感知方向演进。模型开始连接感知、决策和控制链路，外部节点与云端提供更大范围的环境和交通信息，端侧则保留实时判

断与安全执行。随着智能体能力进入移动空间，导航、通信、任务接续和设备控制将进一步被统一组织，形成从“辅助功能”向“场景任务”演进的新型产品。

3.可靠性验证、制度规则与硬件成本决定移动智能规模化节奏

移动智能产品直接作用于真实环境，规模化前提是模型可靠性、系统安全、制度规则和硬件成本同步成熟。复杂天气、突发场景和网络波动可能放大模型失效风险，高算力芯片、传感器和冗余系统增加成本，高阶自动驾驶、无人配送和无人机还涉及准入、责任与运营规范。随着 AI 进一步参与真实控制，需要把模型测试、系统监测、故障降级和责任边界纳入产品全生命周期。

（四）行业智能节点：轻量模型下沉推动现场设备由采集走向决策

1.轻量模型与场景化硬件融合推动数据采集节点升级为现场智能节点

行业智能节点面向安防、电力、煤炭、石化、交通、制造等现场环境，将轻量模型部署在摄像头、工业相机、边缘网关、巡检设备和各类传感控制节点上，使数据在产生位置完成初步识别、判断、告警甚至联动。相比通用计算环境，行业现场更强调低功耗、长时间运行、复杂环境适应和协议兼容，因此模型结构、精度位宽、硬件加速和运行时适配需要围绕具体任务共同设计。

表 10 行业智能节点典型形态与模型硬件融合特征

节点类型	代表硬件	模型部署重点	演进方向
视觉感知节点	智能摄像头、工业相机	轻量视觉模型、图像增强、事件触发与本地告警	由视频采集向场景理解演进
边缘计算节点	边缘网关、边缘盒子	多路数据处理、大小模型协同、协议转换	由数据汇聚向现场推理演进
身份安全节点	门禁、门锁、身份核验设备	活体检测、身份模型、离线授权与风险联动	由单点认证向本地风险判断演进
现场巡检节点	巡检与移动检测设备	多模态采集、异常识别、任务反馈	由人工辅助向智能巡检演进
开放平台节点	可部署第三方应用的行业设备	模型更新、应用隔离、能力 API 与场景扩展	由封闭设备向可编程智能节点演进

2.模型分级、大小模型协同与开放运行环境提升行业适应能力

行业智能节点正在从通用硬件叠加算法，转向面向任务的软硬件共同优化。端侧小模型负责高频检测、状态识别和即时告警，边缘或云端较大模型承担复杂事件分析和低置信度校验；CPU 与 NPU 异构计算、事件驱动和低功耗设计保障长期运行；标准化接口、应用隔离和模型灰度更新则提高跨行业扩展能力。轻量模型、场景化硬件和开放运行环境共同决定现场智能能否规模复制。

3.复杂环境、模型可靠与平台碎片化制约现场智能规模复制

行业智能节点从项目部署走向规模复制仍面临真实环境、模型可靠性、平台碎片化和数据合规等问题。逆光、遮挡、噪声和极端

温度会改变输入分布，不同芯片与协议增加模型适配和运维成本，第三方应用与模型更新又扩大安全边界。未来需要通过统一模型格式、硬件抽象接口、真实场景评测和持续运维机制，提高轻量模型在复杂现场中的可迁移性和可靠性。

（五）具身智能产品：模型与传感、控制和执行系统融合走向物理任务闭环

1.多模态模型与执行硬件融合推动具身产品进入真实任务验证

具身智能产品以机器人等具有物理执行能力的设备为代表，是模型能力与传感器、执行器、控制系统融合最深的产品形态之一。其不仅需要识别环境和理解指令，还要把模型输出转化为稳定、安全、连续的物理动作。人形、四足、服务和陪伴等不同形态对应不同本体结构和资源边界，但都需要在感知、规划、控制和执行之间形成低时延闭环。产业正在从展示单项动作，转向验证真实环境中的连续任务成功率和泛化能力。

2.VLA 与端侧实时控制推动模型能力从感知理解走向自主行动

视觉—语言—动作（VLA）、多模态大模型和智能体框架正在把视觉理解、语言指令与动作生成连接起来，使具身产品能够围绕任务目标形成操作序列。端侧承担实时感知、安全响应和低时延动作控制，云端可补充训练、知识和复杂推理。随着模型轻量化、动作策略和本体控制进一步协同，具身智能将从“识别环境、执行固

定动作”迈向“理解目标、规划步骤、反馈修正并完成任务”。

3.本体成本、模型泛化与控制系统协同决定具身产品商业化进程

具身智能产品实现规模化仍取决于本体成本、核心部件、模型泛化和系统协同。灵巧手、传感器、执行器等需要在精度、可靠性和成本之间平衡，具身模型仍依赖高质量场景数据并面临开放环境泛化不足，模型规划、实时控制和安全机制之间也需要更紧密衔接。未来突破重点不是单纯增加模型参数或提升单项动作，而是把模型能力稳定映射到不同本体和执行系统上，形成可复制的任务完成能力。

四、端侧智能应用场景

端侧智能正加快向消费娱乐、智慧家庭、智慧办公、智慧出行、智慧健康等多类场景渗透，推动人工智能能力从单点功能应用走向深度嵌入用户日常生活和生产活动。不同场景对端侧智能能力提出差异化需求：消费娱乐场景更加关注内容生成、实时交互和个性化体验；智慧家庭强调多设备联动、环境感知和主动服务；智慧办公聚焦知识处理、任务协同和工作流程重塑；智慧出行突出低时延感知、连续交互和安全辅助；智慧健康则依赖持续感知、隐私保护和长期健康管理。随着端侧算力、轻量化模型、多模态感知和智能体能力持续提升，端侧智能应用正在从“单设备、单功能、单场景”向**“多设备、多任务、跨场景”**演进，服务方式由被动响应逐步转向**主动感知、主动推荐和主动执行**。跨设备协同能力的发展，使不同计算、感知、交互和执行节点可以围绕同一任务形成连续服务链

路，推动不同应用场景之间探索更加连贯的服务协同。

（一）消费娱乐：以内容服务为核心，极致用户体验驱动商业转化

表 11 消费娱乐主要端侧产品类型

序号	主要品类	代表产品
1	智能影音终端	智能电视、投影仪、智能音箱、家庭影院
2	游戏娱乐设备	AI 游戏主机、掌上游戏机、云游戏终端、体感游戏设备
3	智能影像创作	AI 相机、AI 运动相机、智能云台、AI 视频创作终端
4	音乐娱乐	智能耳机、智能 K 歌设备、AI 音乐播放器、智能乐器
5	数字阅读娱乐	AI 阅读器、彩色电子书、智能学习平板、内容消费平板
6	陪伴娱乐	AI 陪伴机器人、AI 玩具、桌面机器人、虚拟形象交互终端

1. 消费娱乐以强用户感知、高频使用为主要特点

消费娱乐场景集中体现内容生成、影像处理、游戏增强和个性化推荐等端侧能力。AI 手机通过端侧大模型实现智能摄影、实时翻译、语音助手和内容推荐，AI PC 支持本地图像生成、文档处理和会议翻译，智能游戏设备利用端侧 AI 进行画面增强、NPC 行为生成和反作弊检测，AI 相机、AI 耳机、AI 玩具等新形态也在持续拓展消费娱乐边界。总体来看，消费娱乐逐渐成为端侧 AI 从技术能力走向用户高频体验的重要入口。

2. 消费娱乐由单点功能向多模态和个性化体验拓展

消费娱乐场景的端侧智能正由单点功能向多模态、个性化和软硬件协同方向扩展。首先，端侧大模型能力持续增强，旗舰终端可

支撑更复杂的自然语言理解和多模态交互。其次，多模态融合交互成为主流，文本、语音、图像、手势等输入方式共同提升交互自然度。再次，个性化与情境感知能力显著提升，端侧 AI 能够基于历史行为、环境上下文和情绪状态动态调整内容推荐与服务策略。最后，端云协同混合架构成为标准，高实时任务在端侧完成，复杂生成任务通过云端补充，同时 AI 眼镜、AI 耳机、AI 玩具等新型硬件开辟新的消费入口。

从典型案例看，消费娱乐场景已经形成覆盖摄影、游戏、创作和可穿戴助手的多元化落地路径。端侧 AI 摄影可在本地完成场景分析、参数优化和智能消除，在保障隐私的同时实现秒级响应；AI 游戏增强通过端侧超分辨率重建和自适应难度调整提升画质与个性化体验；端侧 AI 内容创作可自动识别视频高光时刻、生成配乐和转场，降低专业创作门槛；AI 智能体耳机则通过语音唤醒和大模型问答，让用户无需打开手机即可完成信息查询和日程管理。

3. 场景普及需进一步平衡体验、内容安全、能耗与付费价值

消费娱乐场景的进一步普及仍面临体验稳定性、内容治理、能耗成本和付费价值等问题。端侧生成和多模态交互对算力、存储与续航提出更高要求，不同档位终端之间体验差异明显；内容生成可能引发版权、虚假信息和未成年人保护等治理风险，个性化推荐也需避免过度采集和算法沉迷；同时，部分功能仍停留在尝鲜阶段，用户持续使用意愿和付费意愿不足。未来需要在高频需求、低功耗

运行、可信内容和可持续商业价值之间形成平衡。

（二）智慧家庭：以多端协同为核心，由被动响应迈向主动服务新模式

表 12 智慧家庭主要端侧产品类型

序号	主要品类	代表产品
1	家庭智能中枢	智能音箱、控制屏
2	智能环境控制	照明、新风、空调、净化器
3	家庭安防	智能门锁、智能摄像头、烟雾警报器
4	家务服务	清洁机器人、智能厨电
5	健康与陪伴	陪伴机器人、智慧屏、智家老人房/儿童房
6	智慧家电	电视、洗衣机、冰箱、烤箱
7	能源管理	家庭能源系统方案

1. 智慧家庭的价值正在由设备互联走向主动服务

智慧家庭是端侧智能技术落地最系统化、场景最复杂的应用领域，正在推动家庭空间从设备联网升级为主动服务系统。具体来看，智慧家庭覆盖环境管理、安全守护、健康管理、烹饪洗护、娱乐交互和全屋能耗管理等多个维度，依托家庭大脑、多模态感知、认知推理、智能决策和家电智能体，实现家庭空间的自然交互与主动服务。空间交互、语义交互和情感交互能力的增强，使系统能够更好理解家庭成员状态、空间环境变化和生活服务需求。总体来看，智慧家庭的产业价值不在于连接更多设备，而在于让设备围绕家庭成员生活状态、家庭空间变化和服务任务目标协同工作。

2. 云边端协同成为智慧家庭规模化落地的关键路径

智慧家庭正在从设备联网走向空间智能、从被动控制走向主动

服务。首先，家庭大脑成为核心架构，集成通信连接、空间感知、用户理解、认知推理和服务编排能力，支撑家庭空间自主理解与任务调度。其次，长短期记忆与共享记忆机制逐步成为主动服务的重要基础，系统既能够基于短期上下文理解用户当前意图，也能够结合家庭成员习惯、设备运行状态和历史交互记录，形成连续、个性化、可进化的服务能力。再次，云边端协同成为智慧家庭规模化落地的关键路径，端侧设备负责实时感知与低时延响应，边缘家庭中枢负责本地决策与隐私保护，云端大模型负责知识增强、复杂推理和跨场景优化。最后，多智能体协同正在推动家电从单机智能走向群体智能，冰箱、空调、洗衣机、厨电、门锁、照明、机器人等设备可围绕“烹饪、洗护、安防、康养、节能”等任务自主分工、协同执行。总体来看，智慧家庭未来竞争的核心，是能否形成统一感知、统一记忆、统一决策和统一执行的空间智能体系。

从典型案例看，智慧家庭的落地价值集中体现在空间交互、家电智能体和主动服务闭环上。海尔智慧家庭以家庭大脑为核心，融合智家大脑屏、智家 APP、语音入口、智能门锁、AI 视觉、毫米波雷达、家庭机器人和多品类智能家电，构建面向全屋空间的主动服务能力。其中，3D APP 交互能力将家庭空间、设备位置、运行状态和服务场景进行可视化呈现，使用户能够以空间视角理解和管理全屋设备，并支持设备控制、场景编排、能耗查看、环境调节等功能在同一空间界面中协同完成。相比传统列表式设备控制，3D 空间交互更符合家庭真实使用场景，有助于降低用户操作复杂度，提升全

屋智能的可理解性和可用性。

在具体场景中，海尔智慧家庭正在从“用户发出指令、设备被动响应”向“系统理解状态、智能体主动服务”演进。在厨房场景中，冰箱可结合食材识别、存储状态和用户偏好，提供食材管理、临期提醒和菜谱推荐，并联动烟机、灶具、蒸烤箱等设备完成烹饪过程协同；在客厅与卧室场景中，系统可基于用户归家、休息、观影、睡眠等状态，联动灯光、空调、新风、窗帘等设备，提供舒适环境调节；在康养场景中，毫米波雷达、视觉 AI 和陪伴类终端可支撑跌倒检测、睡眠健康管理、紧急呼叫和用药提醒；在全屋能耗管理场景中，系统可基于用水、用电、用气等运行状态开展可视化管理、异常提醒和节能建议。

3. 可靠感知、隐私保护和改造成本是影响智慧家庭规模化发展关键要素

智慧家庭规模化落地仍受到生态割裂、感知误判、隐私安全和改造成本等因素制约。不同品牌、协议和平台之间互联互通不足，跨设备任务编排往往受生态边界限制；家庭场景高度动态，人员身份、行为状态和空间环境识别不准可能引发错误联动和过度打扰；摄像头、门锁、健康与行为数据长期留存在家庭空间，对本地处理、权限隔离和数据最小化提出更高要求；存量住宅设备改造、系统运维和售后服务成本也影响用户接受度。

（三）智慧办公：以提升效率为核心，重塑知识管理与工作流体系

表 13 智慧办公主要端侧产品类型

序号	主要类别	代表产品
1	智能个人办公	AI PC、AI 平板、AI 办公手机、双屏办公终端
2	智能会议	AI 会议一体机、智能会议屏、会议摄像机、AI 会议音箱
3	智能音视频办公	AI 耳机、降噪会议耳机、智能麦克风、会议录音设备
4	文档与信息处理	AI 扫描仪、智能打印机、智能电子墨水本、AI 录音笔
5	协同交互	智能白板、协同大屏、交互式会议桌、无线投屏终端
6	移动办公	AI 眼镜、便携式显示器、移动办公平板、智能键盘终端
7	专业创作办公	AI 工作站、图形创作终端、AI 设计笔记本、专业影像工作站

1. 智慧办公围绕高频工作流探索端侧智能应用

智慧办公是端侧智能率先形成规模化需求牵引和商业化落地的重要方向，推动办公终端从信息处理工具升级为智能生产力入口。随着大模型、多模态感知、端侧推理芯片、智能操作系统和企业知识工程能力持续演进，办公终端正在具备理解、生成、检索、协同和任务执行能力。总体来看，智慧办公的产业确定性来自高频刚需、明确效率收益和企业付费基础。

从通用办公场景看，智慧办公正在围绕文档创作、知识检索、会议协同、邮件处理、数据分析和演示生成等高频工作流重构生产效率。具体来看，AI 可根据主题、提纲和参考资料生成文档初稿并辅助润色改写；可对本地文档、网盘资料、企业知识库、会议记录

和邮件进行统一语义检索；可完成会议转写、纪要生成、待办抽取和会后跟进；可把自然语言需求转化为表格公式、数据图表和分析结论；也可围绕主题生成汇报结构和页面草稿。总体来看，通用办公的核心价值，是把“资料处理”转化为“任务推进”。

端侧智能在通用办公中的价值，不仅体现在内容生成效率提升，更体现在对用户工作上下文的持续理解。具体来看，本地文件语义检索、屏幕内容理解、个人知识摘要、离线会议转写和敏感文档预处理等高隐私、高个性化任务更适合在端侧完成；复杂推理、跨域综合分析和大规模知识检索则可由云端大模型和企业知识系统协同支撑。总体来看，智慧办公正在形成“端侧即时响应、云端深度推理、企业知识增强、应用工具执行”的端云协同模式。

从产业生态看，办公软件正在从单点 AI 功能走向原生办公智能体。具体来看，微软 Copilot 依托 Office、Teams、Outlook、OneDrive 等办公生态，将邮件、文件、会议、聊天和业务数据纳入上下文体系；OpenAI 通过连接器形态接入 Google Drive、SharePoint、Dropbox、Box 等外部工具和数据源，支持在对话中检索组织内容并生成综合分析；国内金山办公从 WPS AI 到 WPS 灵犀，围绕内容创作、智慧助理和知识洞察持续迭代。总体来看，办公软件的竞争焦点正从“提供工具”转向“组织知识和执行流程”。

垂直专业办公正在成为智慧办公深化应用的重要方向，其核心价值在于把 AI 能力嵌入高知识密度、高流程规范和高合规要求的专业工作。与通用办公相比，垂直办公对专业语料、行业知识、流程

规范、权限管控和结果可追溯性要求更高，AI 辅助范围更聚焦，但对业务效率和专业质量的提升也更显著。总体来看，垂直专业办公将成为智慧办公从效率工具走向行业生产力系统的重要路径。

法律办公是垂直智慧办公的典型方向，其核心需求是让 AI 在权威知识、证据链和风险边界约束下辅助专业判断。具体来看，法律检索、合同审查、尽职调查、诉讼材料分析、条款比对和法律意见书起草等任务，需要 AI 建立在法规、判例、合同模板和机构知识库基础上，并提供可追溯、可复核的辅助结果。总体来看，法律 AI 的价值不只是提高写作效率，而是推动法律服务从通用问答走向专业工作流智能体。

金融投研办公的核心价值，是从海量信息中快速形成可验证的研究线索、投资观点和风险提示。具体来看，金融机构需要处理公告、财报、研报、新闻、宏观数据、市场行情和内部研究资料，彭博 ASKB 等对话式 AI 通过接入数据、新闻、研究、文档和分析工具，支持多智能体并行访问数据和分析能力。总体来看，金融投研 AI 的关键不在于生成结论，而在于提高信息发现、证据整合和观点验证效率。

财税审计办公具备较强落地潜力，其核心原因是流程标准化程度高、资料审阅量大、合规要求明确。具体来看，税务、审计和会计工作涉及客户资料审阅、凭证检查、底稿生成、准则核对、税务申报和合规判断，汤姆森路透 CoCounsel 和国内犀友财税 AI 智能体已在税务申报、审计底稿、政策核对和报告起草等流程中探索应用。

总体来看，财税审计 AI 的价值在于把重复性审阅和流程性工作自动化，让专业人员更多聚焦判断和复核。

2. 智慧办公由“功能辅助”迈向“流程自主协同”

端侧智慧办公将从“功能辅助”向“流程协同”演进，办公终端将成为连接用户、知识、模型与应用的关键枢纽。首先，交互方式将由菜单驱动向自然语言和多模态意图驱动演进。其次，办公软件将由单点功能智能化升级为全流程智能体化，AI 不仅生成内容，还参与资料收集、任务拆解、应用调用、结果校验和后续跟进。再次，端侧能力在智慧办公中的比重将持续提升，承担本地感知、个人知识索引、隐私保护、低时延反馈和系统级操作入口。最后，垂直专业场景将从辅助写作走向专业流程协同。总体来看，智慧办公未来竞争的关键，是能否把 AI 从文档工具嵌入到企业工作流和知识体系中。

3. 结果可信、权限可控、系统可接入影响智慧办公应用成效

智慧办公由功能试用走向规模化应用，仍需解决准确性、知识可信、权限边界和系统集成等问题。大模型在专业任务中可能出现幻觉和引用失真，结果需要可追溯、可校验；企业文件、邮件、会议和业务系统涉及多级权限，智能体跨应用调用容易扩大数据泄露和误操作风险；不同办公软件、知识库和业务系统接口不统一，也增加了流程编排与部署维护成本。此外，企业对效率收益、使用频率和投入产出比的评估尚不成熟，影响采购和持续付费决策。

（四）智慧出行：以安全服务为核心，出行效率与交互体验持续优化

表 14 智慧出行主要端侧产品类型

序号	主要类别	代表产品
1	智能座舱	AI 座舱、智能中控屏、车载语音助手、车载大模型终端
2	智能驾驶辅助	ADAS 系统、智能驾驶域控制器、驾驶员监测系统、智能行车记录仪
3	智能导航显示	AR-HUD、智能仪表、车载导航终端、智能后视镜
4	骑行出行	智能自行车、智能电单车、智能骑行头盔、骑行码表
5	充换电服务终端	智能充电桩、家庭充电终端、移动充电机器人、智能换电设备
6	公共交通交互	智能票务终端、智慧公交屏、自助值机终端、车站服务机器人

1. 出行工具正在由交通载体向移动智能空间演进

智慧出行的价值不只是车辆智能化，而是推动人、车、路、云形成协同交通体系。智慧出行是端侧智能在交通领域集成度最高、可靠性要求最强的应用场景，其核心目标是在移动空间中提升安全、效率和服务连续性。具体来看，智慧出行覆盖智能座舱、车路云协同、智能导航、车载 AI 服务和能源管理等维度，智能座舱从信息娱乐系统升级为综合智能空间，车路云一体化通过 C-V2X 构建协同感知网络，导航、车载助手和能源管理则进一步提升驾驶安全、用户体验和续航效率。

2. 舱驾融合与车路云协同成为重要探索方向

智慧出行正在从单车智能走向舱驾融合、车路云协同和跨场景

生态。首先，端侧大模型加速上车，重塑座舱交互和车载助手能力，车载大模型向多智能体协同演进。其次，多模态融合交互成为座舱标配，系统可基于驾驶员状态、语音、视觉和环境信息提供主动服务。再次，舱驾融合架构从理念走向量产，座舱域与智驾域逐步融合为中央计算架构。最后，车路云一体化从试点走向规模化部署，“人车家”跨场景生态深化，数据安全与隐私保护成为产业可持续发展的基础。总体来看，智慧出行未来竞争的关键，是在安全可靠前提下实现车端智能、路侧感知和云端调度协同。

从典型案例看，智慧出行正在形成车路云协同、智能座舱和端到端智驾并行推进的落地路径。百度 Apollo 通过路口智能化改造验证车路云一体化对通行效率和安全性的提升；华为乾崑智能座舱与 ADS 通过端到端大模型和鸿蒙车机系统，强化智驾辅助、多屏流转和自然语音交互；特斯拉 FSD 以纯视觉端到端方案和真实场景数据支撑端侧实时推理；蔚来 NOMI GPT 则将大模型能力融入车载助手，提升多轮对话和情感交互体验。总体来看，智慧出行案例的共同趋势，是汽车正在从交通工具升级为移动智能终端。

3. 安全合规与高成本制约智慧出行的规模化落地

智慧出行规模化应用仍面临安全可靠、制度规则、数据合规和成本基础设施等多重约束。车辆和路侧系统需要在复杂天气、突发交通和网络波动条件下保持稳定运行，算法误判可能直接影响人身安全；高阶自动驾驶的准入、责任认定、保险和数据跨境规则仍需完善；车端持续采集驾驶行为、位置和舱内信息，对权限管理与隐

私保护提出更高要求；高算力芯片、传感器和道路基础设施建设成本较高，也制约方案从示范试点向大范围复制。

（五）智慧健康：以持续感知为核心，打造集监测-发现-干预为一体的健康服务体系

表 15 智慧健康主要端侧产品类型

序号	主要类别	代表产品
1	健康监测穿戴	智能手表、智能手环、智能戒指、健康监测耳机
2	家庭健康监测	智能血压计、智能血糖仪、智能体脂秤、睡眠监测设备
3	健身运动	AI 健身镜、智能健身器械、智能跳绳、体感运动终端
4	康复辅助	康复机器人、智能助行器、智能护理床、康复训练设备
5	老人陪护	陪护机器人、跌倒监测设备、智能呼叫器、居家健康屏
6	视听辅助	智能助听器、AI 眼镜、智能视觉辅助设备、语音辅助终端
7	健康管理终端	智能药盒、家庭健康一体机、便携健康检测仪、健康管理机器人

1. 智慧健康突出持续监测与长期服务需求

智慧健康把端侧智能嵌入民生服务，逐渐推动健康管理从低频诊疗转向连续监测和预防前置。端侧智能终端、轻量化模型和端云医协同架构正在构建覆盖个人、家庭、老年、慢病、运动和心理等全生命周期健康服务体系，形成“端侧监测—本地分析—异常预警—个性化指导—医疗机构联动”的闭环服务。智能手表、手环、便携监测仪等设备可实时采集心率、血氧、血糖等数据，并由本地 AI 模型完成初步解析，在便捷性、精准性和隐私保护之间取得平衡。总体来看，智慧健康是端侧智能最具公共价值和长期粘性的垂直场

景之一。

2. 持续性、个性化、全闭环是智慧健康发展的主要特点

从演进趋势看，智慧健康正迈向医疗级、个性化、全闭环和普惠化发展新阶段。首先，监测能力从基础体征向血糖、尿酸、动态心电等医疗级检测延伸，无创监测设备加速小型化和民用普及。其次，端侧 AI 分析持续专业化，融合临床数据和权威诊疗指南，强化慢病管控、健康风险预判和个性化康复方案。再次，服务模式从单点监测走向全周期闭环，打通家庭、社区和医疗机构，实现预防、监测、干预和康复衔接。最后，端边云一体化和数据安全合规将成为行业发展基础，助力优质医疗资源下沉。总体来看，智慧健康的核心竞争力不在于单个指标监测，而在于能否形成可信、连续、闭环的健康服务体系。

个人健康监测的作用，是通过端侧设备把健康管理从低频就医转化为日常连续服务。深圳市多个社康中心已在家庭病床和基层医疗服务中引入智能监护设备，实时监测心率、血压、血氧饱和度等生命体征，并把异常信息反馈给医护人员和患者家庭。总体来看，个人健康监测场景说明，端侧设备能够延伸基层医疗的感知触点，但相关结果仍需由专业医护人员结合临床情况判断。

老年健康照护的作用，是通过低门槛智能终端提升居家和社区养老的健康管理与安全响应能力。中兴通讯推出 AI 健康屏综合方案，通过带屏终端连接血压计、血糖仪等设备，支持健康数据记录、健康档案生成、远程咨询和社区服务协同。总体来看，适老化端侧智

能的关键不是叠加复杂功能，而是降低交互门槛、持续记录健康状态并缩短服务响应链路。

慢病管理的作用，是通过连续监测和智能分析提升疾病管理的及时性和个性化水平。京东健康与鱼跃、三诺等企业围绕连续血糖监测（CGM）产品开展合作，构建覆盖“监测—分析—干预—跟踪”的智能血糖管理体系，并探索以 AI 辅助用户理解指标变化和获得持续健康管理服务。总体来看，慢病管理场景说明，端侧智能能够把一次性诊疗延伸为长期、连续、可干预的健康服务。

3. 医疗准确性、数据互通与责任边界仍待突破

智慧健康从消费级监测走向医疗级服务，仍需跨越准确性验证、数据互通、责任边界和隐私合规等门槛。可穿戴设备在不同人群和复杂使用条件下的测量误差，可能影响风险判断和医疗决策；设备数据、医院系统和基层平台之间标准不统一，难以形成连续健康档案；算法建议涉及诊疗辅助和健康干预后，厂商、医疗机构与用户之间的责任边界仍需明确；健康数据高度敏感，长期采集、共享和模型训练必须强化知情同意、最小化使用和安全审计。

（六）融合场景：打通多端服务链路，构建全场景智能协同

表 16 融合场景主要端侧产品方案

序号	主要类别	代表方案
1	家庭空间	通过家庭中控、环境感知和设备联动，将家电、影音、安防、健康等设备统一接入，实现家庭场景下的协同服务。

表 16 融合场景主要端侧产品方案（续）

序号	主要类别	代表方案
2	办公空间	通过账号、会议、文档和办公终端协同，实现办公设备与办公应用之间的任务连续接续。
3	移动空间	通过手机、可穿戴设备与车载终端协同，实现出行过程中导航、通信、娱乐和个人服务的连续切换。
4	公共空间	通过身份识别、位置感知和公共智能终端，实现用户与商场、机场、车站等公共服务设施之间的智能协同。
5	跨空间协同	通过统一身份、状态同步和智能体调度，实现家庭、办公、移动及公共空间之间任务和服务的无缝接续。

1. 建设以人为核心的多设备协同体系已成为主流趋势

全场景智能交互体验的核心趋势，是终端体验正在从“用户操作单台设备”转向“多设备围绕用户和场景协同完成任务”。它不是简单增加设备连接数量，也不是在每台设备上叠加 AI 功能，而是让手机、PC、平板、手表、耳机、车机、智慧屏、家居设备等终端共同形成一个连续的服务系统：设备能够自动发现彼此，场景能够被系统理解，服务能够跨端流转，交互方式能够随场景切换，最终让用户在不同空间、不同设备、不同任务之间获得自然、连续、低打扰的体验。

2. 端侧互联互通呈现六类协同演进方向

一是从设备独立运行向场景协同扩展。不同设备可根据自身的感知、计算、显示和执行能力参与同一任务，系统根据场景状态和任务目标进行能力选择与组合。相较于单设备承载完整服务，这种方式有利于发挥不同节点的资源特征，但也对设备发现、能力描述和任务编排提出更高要求。

二是从跨端迁移向多端协同扩展。跨端迁移主要关注内容、状态和进度能否在不同设备之间接续，多端协同则进一步关注多个节点能否同时参与同一任务，例如由高算力节点承担复杂处理、低功耗节点持续感知、显示节点负责呈现、音频节点负责采集或反馈。两类方式均依赖统一身份、状态同步和能力接口。

三是从单点感知向融合感知扩展。位置、生理、环境、应用状态和设备状态等信息可以来自不同传感与计算节点，融合处理有助于形成更完整的场景上下文。但多源信息并不天然意味着判断更准确，仍需解决传感误差、数据时序、身份匹配和隐私授权等问题。

四是从固定界面向多模态协同扩展。不同场景可根据屏幕、语音、触觉、空间显示等交互条件选择更适配的输入输出方式，减少用户在多个界面之间重复操作。具体交互方式需要综合考虑场景安全、信息复杂度、用户偏好和设备能力，避免因自动切换带来额外干扰。

五是从固定应用入口向情境化服务分发扩展。在多设备环境下，系统可结合时间、位置、状态、任务上下文和可用设备，为用户推荐或衔接相应服务。其价值在于减少重复切换和信息中断，但主动服务需要以明确授权、可关闭和低打扰为前提，避免过度推送或错误触发。

六是从功能联动向任务闭环扩展。互联互通不仅关注投屏、文件传输、通知同步等单项功能，也开始关注任务能否在多个应用和设备之间连续推进。系统级智能体可承担任务拆解、流程编排和工

具调用，但任务完成质量仍取决于接口开放程度、权限控制、执行可靠性和结果校验。

3. 设备互通度不高、多端协同可靠性不足正在成为制约开放生态建设的关键要点

融合场景从单品牌体验走向开放生态，仍面临协议标准、身份权限、数据协同和运行可靠性等系统性问题。不同厂商设备发现、能力描述和任务接口缺乏统一规范，跨生态服务难以稳定接续；多设备联合感知和智能体调用扩大了数据流转范围，用户身份、设备权限和第三方服务授权需要精细管理；多个设备和智能体同时参与任务时，还可能出现策略冲突、重复执行和责任不清。未来需要建立统一的协同协议、可撤销授权、冲突仲裁和故障降级机制，确保服务连续而不过度打扰。

五、总结与展望

端侧智能正从“模型能否部署”进入“模型、硬件、系统与服务能否协同”的新阶段。轻量化模型持续提高单位资源下的能力密度，异构硬件和基础软件把压缩收益转化为真实运行效率，智能原生产品则进一步把本地感知、推理和执行嵌入真实任务。与此同时，持续智能服务、跨设备协同和智能体型业态加快形成，产业价值从硬件功能增量向能力、服务和生态增量拓展。

（一）趋势：技术产品双向演进，端侧智能加速规模化落地

技术层面，模型轻量化与软硬件协同将成为端侧智能持续演进

的主线。端侧模型的发展将从单纯缩减参数规模，进一步转向模型能力、资源效率和硬件适配的协同优化。低比特量化、蒸馏、稀疏化、高效注意力、动态计算等技术将持续降低模型运行开销，高能效 NPU、异构计算、统一内存和存算协同等能力则进一步提升模型执行效率。模型设计将更加关注目标硬件的算子支持、精度能力、内存带宽和功耗约束，编译器、运行时和基础软件也将根据任务复杂度、设备状态和网络条件动态组织端侧与云端能力，推动端侧智能由“能够部署”向“高效、稳定、持续运行”演进。

应用层面，人工智能将由功能增强加速走向智能原生和任务闭环。随着模型、多模态感知和智能体能力成熟，AI 不再局限于内容生成、语音识别、图像处理等单点功能，而是逐步参与感知环境、理解情境、规划任务、调用工具和执行操作。个人智能、空间智能、移动智能、行业现场智能和具身智能等新型产品形态将加快发展，服务方式也将由即时响应向持续服务、跨设备协同和智能体自主执行延伸。用户与智能系统的交互逻辑将逐步从“操作功能”转向“表达目标”，人工智能在真实任务中的价值将成为应用创新的重要牵引。

生态层面，产业竞争将从单点产品能力转向模型、硬件、系统和服务的全栈协同。芯片企业将更加关注模型友好型架构和高能效计算，模型企业将持续提升轻量化、多模态和任务能力密度，基础软件企业将强化跨硬件适配、智能体运行和能力接口，产品与服务企业则围绕场景数据、任务流程和用户关系构建持续服务能力。随

着智能原生产品和智能体服务发展，模型部署、开发工具、能力平台、智能体市场和运营服务等新环节将进一步扩展，产业价值由一次性产品销售向持续服务和生态运营延伸。

安全层面，治理重点将由数据保护向模型、智能体和实际执行全过程延伸。模型在多样化设备中部署后，模型资产保护、版本完整性和本地数据安全重要性进一步提升；持续感知和长期记忆使个人数据、环境数据和业务数据的治理边界不断扩大；智能体具备工具调用和自主执行能力后，安全风险也由内容输出进一步延伸至真实操作。未来安全能力将与模型加载、数据访问、身份授权、工具调用和任务执行深度融合，可信执行、最小权限、行为审计和风险阻断等能力将成为端侧智能规模化应用的重要基础。

（二）挑战：多重瓶颈交织叠加，产业成熟仍需系统突破

技术层面，有限资源与持续增长的模型能力需求之间仍存在突出矛盾。多模态、长上下文和复杂推理持续增加计算、存储、带宽和能耗压力，而不同设备在芯片架构、算力水平、精度支持、散热条件和能源供给等方面差异显著，同一模型难以直接跨平台部署。量化、剪枝、蒸馏等技术虽然能够降低资源消耗，但可能带来模型能力损失；不同芯片的算子支持、编译工具和运行时接口也存在差异，进一步提高模型适配成本。如何在性能、时延、功耗、存储和任务质量之间实现动态平衡，仍是端侧智能发展的基础技术难题。

应用层面，智能原生产品和服务仍处于价值验证向规模化落地过渡阶段。当前部分应用仍以功能增强和演示性体验为主，高频刚

需和不可替代价值尚未充分形成。持续智能服务可能产生过度提醒和错误判断，跨应用、跨设备任务容易受到权限、接口和生态边界制约，智能体在复杂长链任务中仍存在意图理解偏差、执行不稳定和结果不可验证等问题。不同行业的数据基础、数字化程度和业务流程差异较大，也使通用能力难以直接复制，端侧智能仍需通过真实任务持续验证产品价值和商业可行性。

生态层面，产业链环节众多、技术路线多样，碎片化问题仍较突出。端侧智能横跨芯片、模型、操作系统、推理框架、终端设备、应用服务和安全治理等多个环节，不同厂商在模型格式、算子体系、硬件接口、开发框架、智能体协议和能力描述等方面尚未形成充分统一。模型跨芯片部署、应用跨系统迁移和服务跨设备协同成本较高，产业链上下游之间也缺乏统一的能力评价和测试方法。开放接口、标准规范和开发者生态仍需进一步完善，以降低创新与规模复制成本。

安全层面，模型下沉和智能体执行正在带来新的风险边界。模型部署在本地设备后，可能面临模型提取、参数篡改、后门植入等风险；持续感知和个性化学习扩大了数据采集和使用范围，增加隐私泄露风险；智能体通过工具和系统接口执行真实操作后，提示注入、工具投毒、记忆污染、越权访问和错误执行等风险可能由数字内容进一步传导至物理和业务过程。目前行业针对端侧模型、智能体和跨设备协同的统一安全标准、测试方法和认证机制仍不完善，治理体系仍需加快建设。

（三）建议：强化技术生态协同，夯实端侧智能发展基础

技术层面，应强化模型轻量化与软硬件协同创新，提升端侧智能基础能力。围绕高效模型架构、低比特量化、模型压缩、多模态轻量化、异构计算、存算协同和低功耗运行等方向开展系统性攻关，推动模型设计阶段充分考虑目标芯片、内存、带宽和能耗条件。加强编译器、推理运行时、算子库和硬件抽象层建设，提高模型跨芯片、跨系统部署效率；完善模型质量、时延、内存、功耗和稳定性等综合评测方法，推动模型优化由单项指标提升向整体任务效率优化转变。

应用层面，应坚持真实任务和场景价值牵引，培育智能原生新产品、新业态和新模式。围绕个人智能、空间智能、行业现场智能、移动智能和具身智能等重点方向，加快形成能够解决实际问题的产品与服务，推动 AI 由单点功能辅助向持续感知、任务协同和自主执行演进。支持智能体与办公、家庭、工业、医疗、交通等业务流程深度结合，形成一批可复制、可推广的标杆应用；围绕高频需求和可量化价值探索能力订阅、智能体服务、按任务付费和行业解决方案等模式，提升产业可持续发展能力。

生态层面，应加强产业链上下游协同，构建开放兼容的发展环境。充分发挥产业联盟、龙头企业、科研机构、开源社区和开发者生态作用，推动芯片、模型、基础软件、产品和应用企业建立更加紧密的协同机制。加快模型格式、硬件抽象、算子接口、智能体工具调用、设备能力描述和互联互通等标准建设，降低跨平台适配成

本；加强开发工具、公共测试平台和开源基础设施建设，为中小企业和开发者提供低门槛创新环境，促进技术成果快速向产品和服务转化。

安全层面，应统筹发展与安全，建立覆盖模型、数据、系统和智能体的全流程治理体系。应坚持安全与发展并重原则，加快构建覆盖设备、模型、智能体和数据的全生命周期安全体系。推动建立端侧智能分层分类安全标准和评测认证体系，形成统一安全基线。加强可信执行环境、硬件根信任、模型签名、数据保护和安全隔离等基础能力建设，完善智能体工具调用与实际执行的安全控制机制，并通过持续监测、评估和更新应对新型风险，为端侧智能新产品、新业态和新模式规模化发展提供可信保障。

编制说明

本研究报告自 2026 年 4 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段，面向端侧智能产业链上下游的技术供应方、终端制造方、服务集成方和应用需求方开展问卷调查、深度访谈和专家征询等工作，以期为产业研究、技术创新、标准研制和应用推广提供参考。

本报告由中国信息通信研究院人工智能研究所撰写，撰写过程得到了中兴通讯股份有限公司、华为技术有限公司、维沃移动通信有限公司、上海阶跃星辰智能科技股份有限公司、浙江大华技术股份有限公司、小米科技有限责任公司、OPPO 广东移动通信有限公司、联想集团有限公司、北京面壁智能科技股份有限公司、联通（广东）产业互联网有限公司、灵伴科技（杭州）股份有限公司、青岛海尔科技有限公司、第四范式（北京）技术有限公司、广州万协通信息技术有限公司、深圳市机密计算科技有限公司、北京元支点信息安全技术有限公司、中国信通院西部分院等单位的大力支持。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn