

从AI设备到个人AI

智能体驱动的终端进化与产业重构



CONTENTS

目录

前言	01
第一章：产业拐点：AI超级周期下的结构性矛盾与个人终端重塑	02
1.1 繁荣与跨越：AI高渗透率下的个人终端原生之困	03
1.2 智能体破局：生态重构的推动力量	05
1.3 个人AI：以智能体为核心的产业新范式	06
第二章：供需矛盾：用户体验需求与厂商智能方案的困境	08
2.1 用户需求升维：从“功能尝鲜”到“价值刚需”的认知跃迁	09
2.2 终端厂商的智能化路径与核心挑战	11
第三章：个性化智能从愿景到现实：智能技术的系统级重构	16
3.1 Token 经济学的必然性：AI推理从“云端集中”走向“端边云分布式”	17
3.2 异构计算架构的必然性：支撑智能体自主运行、规划与跨设备协同	18
3.3 个性化智能的关键技术支柱	19
第四章：个人AI时代终端与场景创新展望	23
4.1 个人AI催生终端角色演变	24
4.2 个人终端的应用场景和交互模式展望	27
4.3 全栈能力平台加速个人AI终端落地	29
第五章：个人AI时代终端市场展望与厂商行动建议	32
5.1 个人AI终端市场发展规模预测	33
5.2 厂商行动建议	36

PREFACE

前言

2026年，全球AI产业已进入超级周期——IDC预测，全球企业本年在AI上的支出将达到9400亿美元，且有望在2029年增长至2.1万亿美元。全球市场正处于从基础设施建设转向应用爆发的阶段，企业与个人应用、智能体AI（Agentic AI）等前沿趋势正以前所未有的速度重塑智能终端的产业格局。随着终端算力、模型能力与系统能力持续演进，传统以硬件参数和单点功能驱动增长的发展模式正逐步触及边界，终端产业进入新的结构性调整阶段。

一方面，终端厂商在短期市场正面临多重压力。根据IDC预测数据，2026年，PC、手机等核心市场出货量预计将迎来两位数下滑，终端设备平均单价上涨超过10%，在传统的“参数内卷驱动销量增长”逻辑逐渐失效之时，AI终端正成为最大的增长亮点。

与此同时，AI产业的重心也正在发生变化。随着智能体AI加速发展，AI开始从以内容生成和问答响应为主，进一步演进为能够理解意图、编排任务、调用工具、执行任务并持续协同的系统能力。对于终端产业而言，这意味着设备的角色、交互方式及系统架构都将被重新定义。

在这一背景下，IDC认为，以用户为中心、以智能体为载体的个人AI，作为全新AI时代的个人终端核心系统形态，是破解当前产业结构性矛盾、重构终端产业价值的重要机遇。本白皮书旨在基于产业及技术演进趋势，系统梳理产业现状，剖析个人AI的核心价值与实现路径，为终端厂商把握产业机遇、实现战略转型提供清晰的指引。

PART 1

产业拐点

AI 超级周期下的结构性矛盾与个人终端重塑

过去数十年，个人终端市场先后经历四轮迭代浪潮，每一轮技术变革都重塑着产业形态与用户交互模式。1980年代至2000年代，PC的普及开启了桌面计算时代，独立桌面设备成为办公生产力的核心载体，依托键鼠交互支撑起数据处理、文档编辑与基础网络连接需求。2007年至2018年，智能手机爆发推动移动互联网浪潮全面到来，多点触控成为主流交互方式，APP生态的爆发式增长彻底重构了用户获取信息、社交娱乐与移动办公的模式。2019年至2025年，多形态个人设备集中涌现，逐步构建起“手机+平板+手表+耳机”的全场景设备矩阵，AI功能以单点植入的方式初步落地，行业启动跨设备任务流转与生态协同的探索。

2023年，首批集成专用AI加速硬件、具备本地AI推理功能的终端设备上市，AI终端正式进入消费市场。IDC据此提出“AI终端”¹定义——具备本地AI推理能力、集成AI加速器、并能支撑AI驱动用户体验的产品。2026年以来，芯片、模型与系统的协同突破催生市场快速迭代，产业正式迈入“AI原生+多形态终端”的全新发展阶段，新兴终端形态加速涌现，终端产品的价值评估逻辑也发生深层变化。

高速增长的同时，产业也面临着成本结构、产品逻辑与体验能力的多重深层挑战，传统以单设备为中心的发展模式已难以适配新一代AI的演进要求，终端产业的竞争重心正从单点功能升级转向系统级的架构重构与生态协同。

1.1 繁荣与跨越：AI高渗透率下的个人终端原生之困

IDC最新预测数据显示，2026年全球AI终端渗透率将迎来全面普及拐点：全球智能手机出货量将超10.88亿台，其中AI Phone出货量超7.22亿台，渗透率66.4%；PC市场出货量将超过2.55亿台，其中AI PC出货量预计接近1.27亿台，渗透率达49.9%；平板出货量将接近1.39亿台，其中AI平板出货量超1.03亿台，渗透率74.3%²。与此同时，AI技术加速向全品类终端渗透，新兴终端形态迎来爆发式增长：IDC数据显示，2026年全球智能眼镜出货量预计同比增长34.6%；轻量化AI可穿戴设备呈现多样化发展态势，AI戒指、智能吊坠等贴身设备快速涌现，凭借极致便携性与全天候感知能力，成为个人AI生态的重要补充；智能座舱、智能家居中控等场景化终端也完成了AI能力的全面升级。传统终端与新兴终端协同发展，共同构建覆盖用户全生活场景的AI终端硬件网络。

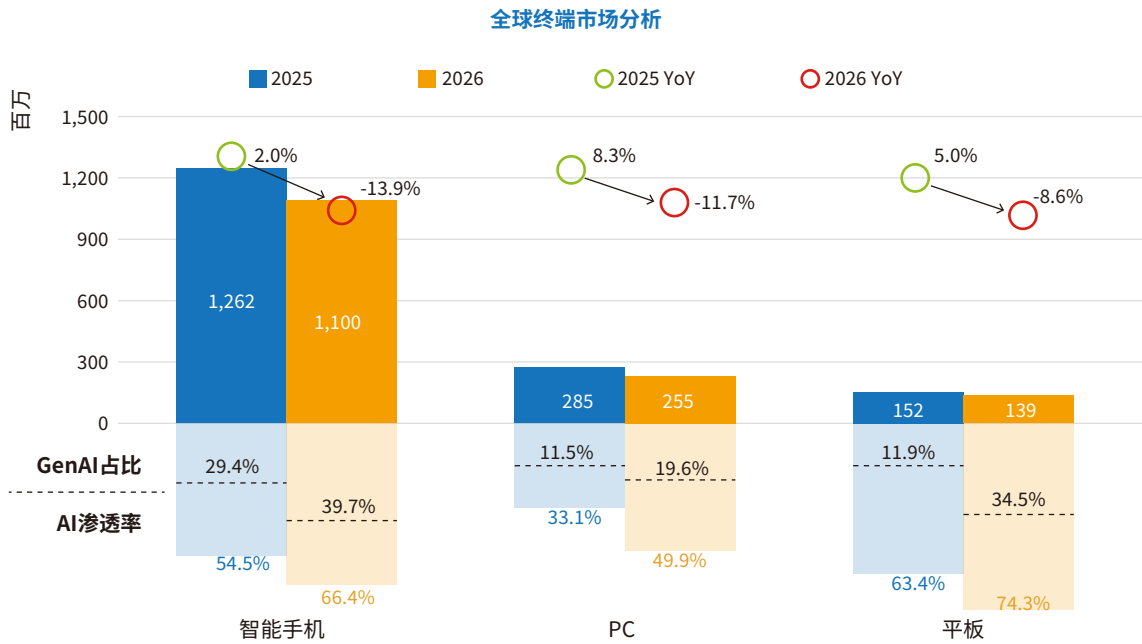
1:根据IDC定义，“AI终端”需具备本地AI推理能力，集成AI加速器（如NPU），并能支持AI驱动的用户体验，该定义适用于PC、平板与手机三类终端产品。根据搭载的算力大小，IDC将AI终端划分为Basic AI与GenAI两类，详细划分标准如下：

手机&平板：Basic AI Phone: 1-30 TOPS；Gen AI Phone >30 TOPS

PC: Basic AI 1-40 TOPS；Gen AI >40 TOPS

2:数据来源：IDC 2026 全球 AI 终端市场跟踪报告

图1 IDC 全球终端出货量预测及AI渗透率



来源：IDC，2026

高速增长的AI终端市场背后，暗藏制约行业长期发展的结构性矛盾。AI高渗透率的表象之下，传统终端厂商正面临愈发严峻的市场挑战，压力集中于两端：硬件端，供应链成本波动推高终端定价，但用户换机决策更趋理性，且对AI工具的实用价值感知不足，不愿为AI终端的硬件溢价付费；服务端，随着用户AI使用习惯逐步养成，云端推理成本随Token消耗量快速攀升，但用户付费意愿与成熟商业模式尚未形成，成本收益错配问题显著。

与此同时，智能眼镜、轻量可穿戴等新一代贴身终端加速兴起，推动交互入口多元分化，这不仅是对传统终端形态的补充，更意味着终端产业正在从“单一主设备承载全部能力”的模式，逐步转向“多终端共同承载AI体验”的发展路径。

多终端产业规模的急速扩张让变革刻不容缓。IDC全球终端保有量数据显示，目前全球市场拥有超过14亿台PC、38亿台智能手机、7.5亿台平板、13.5亿台可穿戴设备（涵盖智能眼镜、手表、手环、戒指、耳机等形态）以及约3.5亿台智能网联汽车，终端总量超77亿台³。IDC认为，当前的终端设备数量还只是起点，当以用户为中心的多终端生态体系进一步成熟，未来个人智能终端市场整体的规模与复杂性还将进一步提升。这意味着，当前围绕单一终端设备设计搭建的传统智能架构，将难以适配未来百亿级智能终端持续在线、跨设备协同的AI计算需求，产业架构变革已势在必行。

3:IDC 2026 全球 AI 终端市场跟踪报告

1.2 智能体破局：生态重构的推动力量

2026年，AI产业正加速从生成式AI(Gen AI)迈入智能体AI (Agentic AI)的时代。相较于传统AI助手主要聚焦单轮问答、内容生成与固定功能调用，智能体AI已具备更强的目标理解、任务规划、工具调用和多步骤执行能力，推动AI从“响应式能力”向“行动式能力”演进。

这一变化不仅是AI能力本身的升级，更推动人机交互逻辑与终端角色定位发生深层转变。过去，用户主要通过操作系统、应用和设备界面完成任务；而在智能体AI时代，用户与数字系统的交互正逐步从操作设备转向意图表达，由智能体统筹协调执行。以OpenClaw为代表的智能体工具兴起，推动智能体从对话式向自主执行式演进，支撑跨设备、跨应用的统一智能体身份落地，使终端不再是孤立硬件，而是个人智能网络中的能力载体与感知端点，形成“以智能体为中心”的全新交互模式。

面对上述结构性变化，产业各参与方正从不同层面加速布局智能体相关能力。在操作系统层，谷歌以Gemini大模型为核心底座，对安卓、Chrome OS两大操作系统进行底层重构，将AI智能体能力深度融入系统架构，推动传统操作系统向主动式智能平台演进。头部终端厂商也在加速软硬件一体化布局，构建矩阵化终端体系，强化端侧AI能力与跨设备协同体验。以高通为代表的头部芯片厂商则通过统一的异构计算架构、软件栈和多重连接，跨品类为海量终端提供高性能、低功耗的智能体运行基座，支撑AI工作负载的分布式运行。

AI产业正从训练侧向推理、执行侧全面延伸。智能体从单步响应向多步骤任务执行演进，推动终端角色从传统“被动响应式工具”，向“主动服务的智能体端点”根本性跃迁。

智能体的规模化持续运行，对云端算力能耗、终端AI能效、环境感知、跨设备协同与数据隐私均提出了更高要求。IDC认为，行业下一阶段发展重点将从单一智能体能力提升，转向系统级架构与协同能力构建，以用户为中心的跨设备连续个人AI体系，正成为核心演进方向。

1.3 个人AI：以智能体为核心的产业新范式

1.3.1 智能体驱动范式跃迁

面对产业增长压力与智能体落地的技术挑战，行业已形成共识：破解矛盾的核心路径，是构建以智能体为核心、分布式智能为基础的全新AI生态架构，通过计算架构、交互范式与服务网络的三重变革，实现个人终端生态的系统性重构。

计算架构层面，打破端、边、云物理边界，形成算力可无缝流转的统一计算体系。系统根据任务时延、隐私等级与计算复杂度智能分配算力：高频轻量任务端侧处理，中等任务边缘承接，复杂任务调用云端，从根源优化全链路推理成本，实现成本、性能与能效的最优平衡。

交互范式层面，完成从“操作设备”到“表达意图”的底层转变。用户无需记忆操作路径，通过自然语言、手势、行为即可表达需求；智能体自主理解意图、规划任务、调用工具并跨设备执行，真正实现“所问即所得、所想即所成”的无感智能体验。

服务网络层面，形成以用户为中心的全场景覆盖网络。不同终端基于硬件属性承担差异化角色：手机、PC作为算力与数据中枢，智能眼镜、耳机作为自然交互入口，可穿戴设备作为全天候感知端点，座舱、家居作为场景化能力补充。服务随用户跨设备流转，实现全时、全域、连续的智能体验。

这套分布式架构并非单一设备的功能叠加，而是围绕用户构建的完整智能体系，既通过算力分流优化产业成本结构，也以跨端连续体验提升用户AI价值感知，为终端产业打开全新增长空间。

1.3.2 个人AI定义与特征

基于以智能体为核心的全新AI生态架构，IDC提出个人AI的行业标准定义：以智能体为核心运行载体、以用户为中心展开运行的个人智能系统，通过端侧智能算力与多模态交互，融合边缘与云端AI资源，在个人多终端设备生态中为用户带来全时全域服务能力。

个人AI的五大核心特征

相较于传统以单设备或单应用为边界的AI服务，个人AI以用户为中心组织数据、能力与体验，核心是让AI服务随用户流动，在不同设备与场景间保持连续一致的体验。

IDC认为，个人AI具备以下五大核心特征：



个性化服务



全时感知自然交互



连续服务



可进化



安全可靠

- **个性化服务：**以用户为中心，提供定制化、场景化的专属智能服务，基于用户行为偏好自主优化，实现千人千面的差异化体验。
- **全时感知与自然交互：**持续感知用户状态与环境变化，主动预判并触发服务；支持语音、手势等多模态自然交互，打造流畅无感的智能体验。
- **连续服务：**在终端与云端间实现数据、能力与任务的无缝流转，通过连续计算支撑跨设备的服务延续，使用户体验不受单一设备边界割裂。
- **可进化：**持续学习用户习惯与数据，自主迭代优化服务策略，越用越贴合用户需求。
- **安全可靠：**构建全链路隐私保护体系，在端、边、云间合理分配数据与计算任务，确保用户数据与交互的安全可控。

需要指出的是，个人AI的实现绝非既有AI能力的简单叠加，而是涉及产品定义、系统架构、终端协同、生态合作及商业模式的全面演进。对于产业参与者而言，推进这一转型不能脱离用户真实需求进行技术堆砌，而应先深入挖掘不同用户群体的核心诉求与使用痛点，系统梳理支撑个人AI落地的关键技术体系，为产品创新与生态建设筑牢基础。

PART 2

供需矛盾

用户体验需求与厂商智能方案的困境

2.1 用户需求升维：从“功能尝鲜”到“价值刚需”的认知跃迁

用户需求升维与体验痛点持续凸显，倒推终端产业从产品逻辑到发展路径的系统性重构；终端厂商对用户行为的深度洞察与战略转向，与需求端形成双向共鸣。供需两端的共识正在加速形成，为终端智能化落地与产业协同模式升级指明了方向。

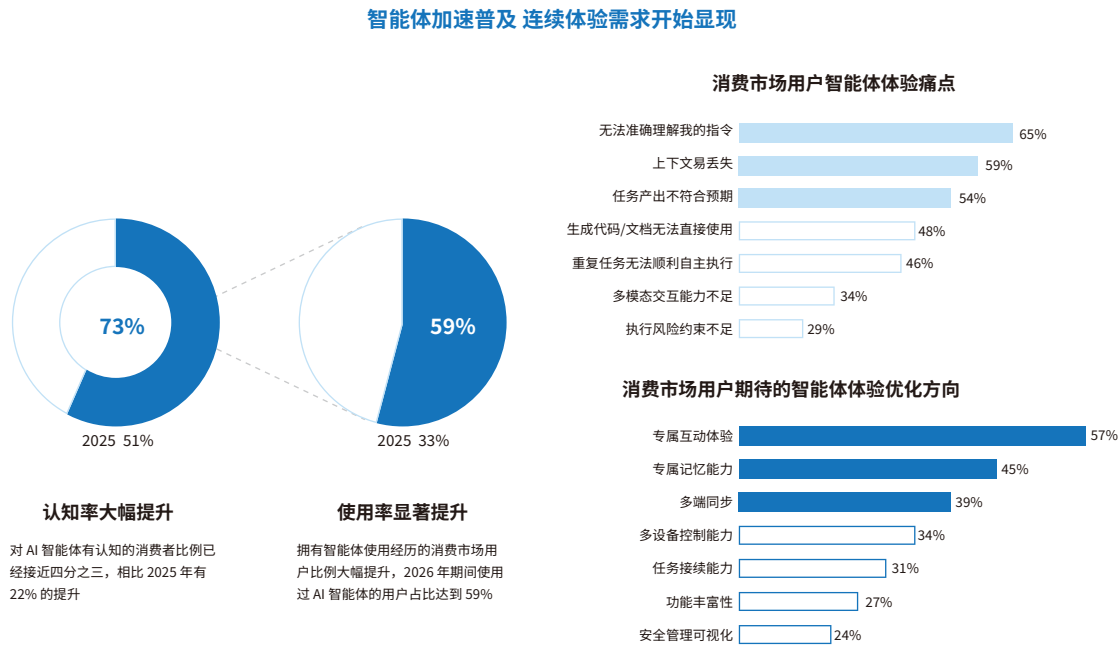
2.1.1 多设备普及催生连续体验刚需

IDC最新研究显示，2026年全球人均智能设备持有量将达7.6台，较2023年的4.5台大幅提升，用户的数字生活已全面进入多设备协同时代。与此同时，用户对AI的认知与使用深度持续增长，2026年上半年IDC用户调研显示，用户对AI智能体的认知和使用率同比均提升20%以上，AI功能已从尝鲜体验转变为设备常用服务。

使用场景的泛化与设备数量增加，推动用户对AI体验的诉求从“单设备功能增强”转向“全场景连续服务”。根据IDC针对中国市场AI先锋用户开展的调研，用户的新兴核心需求体现在两个层面：一是体验一致性需求，用户期待不同终端上的智能体保持统一的人设、记忆与能力标准，切换设备无需重新学习操作、重复告知个人偏好；二是任务连续性需求，用户希望任务进度、服务状态、上下文记忆可跨设备无缝接续，换机或切换设备时无需重复下达指令、手动同步数据。这种诉求的本质，是AI服务的核心从“绑定设备”转向“跟随用户”，与传统以单设备为中心的产品逻辑形成了根本差异。

IDC认为，这种变化正推动AI体验从“设备中心”向“用户中心”转变。未来用户关注的不再是单一设备是否具备AI功能，而是AI能否伴随用户在不同场景、不同终端间持续理解需求、延续任务状态并提供一致体验。这也意味着，终端需要具备更强的本地智能、持续感知和跨设备协同能力，才能真正支撑个人AI的发展。

图2 消费市场用户智能体认知/使用率与体验优化方向



来源：IDC消费市场AI智能体用户行为综合调研，面向中国市场在手机，平板和PC终端上有AI智能体使用经验的用户开展，N>500

* 在调研问卷中，智能体表述为“AI助手\AI智能体”

2.1.2 当前AI体验的三大核心痛点

需求的升维反衬出当前AI体验的能力短板，结合用户调研与访谈数据，IDC将用户反馈最集中的负向体验总结为三大痛点。

一是意图理解能力不足。当前多数终端AI服务仍停留在指令式响应阶段，仅能准确识别表述清晰的需求，一旦用户表述模糊、省略背景信息，系统就难以推断真实意图。根源在于当前AI缺少对用户的全天候场景感知与长期个人记忆，仅能基于单轮对话和有限上下文信息完成语义识别，难以形成对用户的完整认知，导致用户需要反复补充背景、修正需求，大幅提升了使用成本。

二是产出结果落地性弱。当前AI输出的内容多为通用化、模板化结果，生成过程往往缺乏对用户个人背景、历史行为、设备状态与实时场景的深度适配。用户获取结果后，仍需大量二次编辑加工才能投入使用。这种“能用但不实用”的体验，使AI始终停留在辅助工具层面，难以成为用户可信赖的生产力工具，无法支撑用户为溢价付费。IDC认为，AI价值的核心正逐渐从“生成内容”转向“结合个人情境生成可执行结果”，这对端侧个性化数据处理和持续学习能力提出了新的要求。

三是复杂任务无法形成闭环。当前终端AI大多仅能完成单步、单应用的简单任务，面对多步骤、跨应用的复杂需求，往往无法完整落地，中间环节仍需用户手动操作。而这类复杂任务恰恰是最能体现 AI 价值、建立用户粘性的场景，相关能力的缺失直接导致用户留存率与价值感知双低。本质上，复杂任务的完成不仅依赖模型能力本身，更依赖任务规划、工具调用、跨应用协同与上下文连续性等系统能力。随着智能体AI发展，用户对于“一句话完成复杂任务”的期待快速提升，这也要求AI从单步响应逐渐演进为具备自主规划和执行能力的智能系统。

从底层逻辑看，三类痛点的根源具有较强共性：当前AI体系仍围绕单设备、单应用交互构建能力，在跨场景协同、持续理解用户需求与任务连续执行方面仍存在局限。IDC认为，未来AI能力建设重点，将逐渐从模型能力竞争转向感知、记忆、推理、规划和执行等系统能力的协同演进。

2.2 终端厂商的智能化路径与核心挑战

IDC观察认为，面对用户需求升维，终端厂商已形成行业共识，AI产品开发逻辑正从“通用功能堆砌”转向“以用户为中心”。这与个人AI强调的“个性化、情境化、可信赖”核心理念高度契合，成为行业发展的新风向标。主流厂商正沿阶梯化路径推进智能化升级，但在落地过程中仍面临多重共性挑战。

2.2.1 厂商对用户痛点的洞察与响应

IDC调研显示，终端厂商对用户AI使用行为的洞察，已从表层的功能使用频率统计，深入到场景价值、用户心理与商业模式层面，形成了四大行业共性判断，与用户端的需求升维形成了清晰的双向呼应。

第一，功能渗透率与价值留存率存在明显断层。当前信息查询、影像处理、语音交互、基础内容生成等轻量化AI功能，技术成熟度高、使用门槛低，用户使用频率与渗透率均处于较高水平。但涉及多步骤、跨应用的复杂任务，以及创意生成、深度工具类场景，用户初次尝鲜后流失率居高不下，难以形成持续使用习惯。全行业普遍判断，当前终端AI尚未跨越从“有用”到“不可或缺”的价值临界点，这是用户付费意愿不足的核心原因之一。

第二，场景化体验是终端AI的差异化核心。IDC调研发现，相较于以云端为主的通用AI服务，终端AI在实时环境感知、个性化理解，以及结合用户长期行为数据提供持续服务等方面展现出独特价值。因此，越来越多厂商开始围绕场景化体验构建差异化竞争力，强化终端侧AI的用户价值。行业正从“通用功能堆砌”转向“场景化深度定制”，依托不同终端的形态特性，打造云端AI无法替代的专属场景体验，通过场景价值构建用户粘性与差异化竞争力。随着AI进入智能体阶段，终端厂商愈发认识到，终端侧感知能力、个人记忆能力和实时情境理解能力，正成为区别于通用云端AI的重要优势来源。

第三，体验确定性与数据安全性是信任基础。调研表明，AI响应不稳定、执行结果不可靠，是导致用户流失的核心因素之一，提升执行确定性已成为全行业的共同优化方向。同时，随着AI需要调用的个人数据类型持续扩大，用户对隐私安全的关注度显著提升。当前越来越多终端厂商倾向于采用本地优先、端云协同的技术路径，在保障响应速度和服务能力的同时，强化用户数据隐私保护与数据主权管理，构建用户信任体系。

基于上述洞察，终端厂商的AI产品开发逻辑正在转向“围绕用户需求重构系统能力、深化场景定制与推进生态协同”，这与个人AI“以用户为中心”的核心理念高度契合，也为后续技术架构升级指明了方向。

2.2.2 厂商对终端智能化落地的路径探索与核心挑战

终端行业在智能化落地阶段处于共同探索阶段。尽管不同终端品类的起步节奏与能力上限存在差异，但全行业演进路径高度统一，均沿着从单点功能嵌入到智能体自主编排的方向阶梯式推进。

行业共性演进路径：从功能嵌入到系统级智能的阶梯升级

整体来看，终端智能化的落地可划分为四个递进阶段，共同构成行业升级的清晰路线图。

第一阶段为功能嵌入期，AI以单点工具形态嵌入系统与应用，仅能完成单轮问答、单功能调用等简单任务，能力分散且体验割裂。

第二阶段为助手协同期，即当前行业所处的核心阶段。AI具备基础意图理解能力，可通过系统调度完成跨应用简单操作，实现从“单功能响应”到“跨应用执行”的跨越。

第三阶段为任务代理期，预计2026-2027年规模化落地。依托端云协同架构与端侧算力升级，AI可自主完成完整单场景任务、执行稳定性大幅提升。

第四阶段为自主编排期，预计2028年及以后逐步落地。AI深度学习用户行为习惯，主动预判需求、编排多步骤复杂任务、联动多设备协同工作，向更高阶智能演进。

在落地节奏上，行业普遍采用“旗舰先行+存量升级”双线策略：一方面在旗舰机型搭载高算力硬件与完整智能体框架，迭代前沿交互体验；另一方面通过系统OTA向存量设备推送基础AI能力，快速扩大用户覆盖。

全行业共性核心挑战

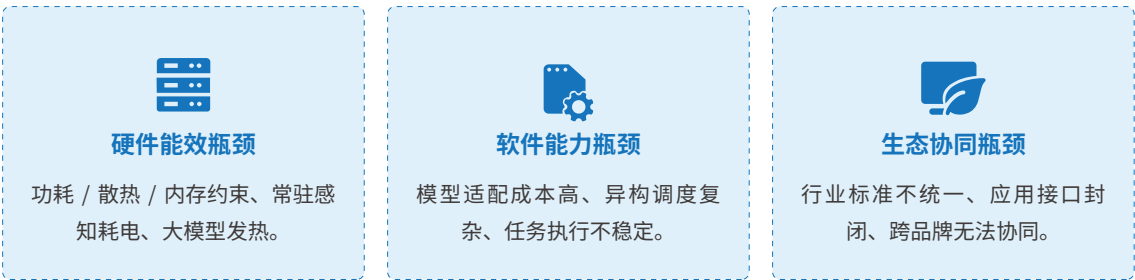
尽管演进路径清晰，但规模化落地过程中，终端厂商普遍面临三重共性瓶颈，是全行业需共同突破的核心障碍。

首要瓶颈是硬件能效，端侧AI的持续运行对每瓦特性能、功耗、散热、内存带宽都提出了更高要求，而终端设备的体积、电池容量、散热条件存在明确物理边界。多模态常驻感知带来的功耗飙升、大模型推理导致的发热与内存占用，是所有终端品类共同面临的硬约束。如何在有限的功耗、散热和体积约束下持续提供稳定的AI能力，并兼顾用户体验与设备续航，正成为终端产业共同关注的重要课题。IDC观察到，随着智能体AI逐步走向实际应用，产业关注重点正从单纯峰值算力提升，逐渐扩展到持续计算能力、资源调度能力和整体能效水平的综合优化。

其次是软件能力瓶颈：即从模型适配到任务调度的系统级难题。硬件算力只是基础，体验上限最终取决于系统级软件优化能力。IDC认为，随着大模型能力逐步普及，产业竞争重点正从模型能力本身，进一步延伸至模型部署效率、运行效率与智能体执行能力。如何实现跨计算资源的高效调度，在保障性能的同时控制时延和功耗，正成为行业持续关注的重要方向。具体体现在三方面：一是端侧模型落地部署成本高，行业缺乏可直接商用的场景化轻量化模型，厂商需投入大量资源完成模型裁剪、量化与适配；二是异构调度精度要求高，CPU、GPU、NPU的协同调度效率直接决定能效比与响应速度，微秒级的动态调度对系统优化能力提出了极高要求；三是执行确定性不足，跨应用操作缺乏可靠的容错机制，模型输出的不确定性易导致操作失误，严重影响用户信任。

第三是生态协同瓶颈。个人AI的完整体验高度依赖全生态协同，但当前行业碎片化现状严重制约了体验上限。其一，行业标准不统一，AI能力分级、跨端连接、数据流转等领域缺乏统一规范，各厂商多采用内部自定义标准，导致用户认知混乱、跨品牌适配成本高企。其二，应用生态封闭，头部应用未开放系统级接口，AI难以深入应用内部完成操作，无法形成完整任务闭环。其三，跨品牌协同不足，当前智能体能力基本局限于厂商自有品牌生态内，但用户实际使用环境往往由多个品牌、多种终端共同组成，单一品牌生态体系在满足复杂跨场景需求方面仍存在局限。

图3 终端厂商智能化落地核心瓶颈



来源：IDC，2026

产业协同模式升级：从硬件供应到协同共创

面对上述共性挑战，终端厂商与上游芯片厂商的合作模式正在逐步调整——从传统的标准化硬件供应，向需求前置、深度研发合作的协同关系演进。

传统合作模式中，芯片厂商交付标准化硬件，终端厂商再基于硬件做二次开发；需求传递链路过长、场景定制成本高，难以适配 AI 时代细分场景的差异化技术需求。而在新型共创模式下，双方在芯片定义阶段即启动深度协同：终端厂商提前输出真实的场景算力需求、模型运行特征与产品形态约束，芯片厂商将这些需求融入芯片架构设计，共同优化CPU/GPU/NPU流水线、异构调度机制与功耗控制策略，实现硬件能力与场景需求的深度匹配。

这种合作模式的转变，是AI时代终端产业价值链重构的缩影。芯片厂商从单一芯片方案供应商，逐步转变为跨多终端软硬件统一方案的提供者，并在个人AI底层基础设施建设与体验定义中发挥更重要的作用；终端厂商则从被动的方案使用者，转变为提前输入场景需求、与芯片厂商共同打磨产品的深度参与者；软件厂商从单设备功能开发者，升级为面向全场景的智能体体验优化服务商，聚焦用户侧体验打磨与垂类场景能力拓展，成为连接底层算力与终端体验的关键纽带；模型

厂商则从独立提供模型能力，转变为打造跨端统一智能体框架的能力中枢，输出模块化、可调度的智能服务，支撑全场景任务的自主执行。IDC认为，随着终端AI向更复杂的智能体能力演进，产业链协同的重要性正持续提升。通过更紧密的软硬件协同与产业合作，有望进一步提升创新效率，并推动相关技术与应用场景的规模化发展。

图4 产业协同模式对比

	传统产业模式：垂直分立	个人AI协同模式：生态共生
 芯片厂商	 以算力供给为核心 • 竞争重心集中在制程、主频等硬件参数 • 对上层模型与应用场景的系统级适配相对有限	 全栈计算底座 • 面向智能体做系统级优化，打通“端-边-云”算力调度 • 提供异构计算与模型推理加速
 终端厂商	 自有生态为主 • 品牌内软硬件一体化布局，系统与核心应用深度绑定 • 跨品牌、跨品类的协同与数据流转存在明显壁垒	 场景化体验枢纽 • 构建多设备智能矩阵，作为感知、交互与执行物理载体 • 实现跨端无缝流转与融合
 软件厂商	 聚焦单设备功能 • 多基于单一操作系统开发，能力受单设备硬件条件制约 • 较难打通全局数据与算力实现体系化智能进化	 智能体体验优化服务商 • 深耕垂类场景的智能体服务能力，聚焦意图识别、任务执行与个性化体验优化，打磨用户侧全链路智能体验
 模型厂商	 痛点 • 各环节侧重内部技术迭代 • 用户体验受设备与应用生态边界制约明显 • 全链路价值传递存在断层	 优势 • 打破壁垒实现横向协同 • 围绕用户构建连续一致的智能体验 • 最大化释放生态价值

来源：IDC，2026

综合来看，无论是用户体验需求的持续升级，还是终端厂商在硬件、软件及生态层面面临的现实挑战，都表明AI正在从单点功能向系统能力演进，从设备智能向用户智能发展。IDC认为，未来产业发展的重点或将逐步转向感知、记忆、规划、执行和协同等能力的系统化建设。这也为下一阶段个人AI相关技术架构的发展提供了重要方向。

PART 3

个性化智能从愿景到现实

智能技术的系统级重构

3.1 Token经济学的必然性：AI推理从“云端集中”走向“端边云分布式”

智能体AI正在重塑AI计算的经济模型。智能体可自主运行、持续保留上下文，并高效调用应用、系统与服务执行多步骤任务。由此，Token不再仅是单次问答的计算消耗，更成为智能体规划、推理、工具调用和跨系统协同中的持续性资源消耗。随着个人AI从单次问答走向全时在线、持续感知和主动服务，叠加高昂的推理成本、难以避免的网络时延与用户日益增长的隐私安全需求，纯云端计算已无法支撑个人AI的规模化落地，AI计算必须从云端集中走向端边云协同的分布式架构。

个人AI全时持续、跨端协同的系统特性，推动了AI计算范式的底层变革，其核心逻辑源自Token经济学的客观规律。智能体时代，多轮推理、多智能体协同、持续上下文保留等核心特性，推动单用户Token消耗量进入指数级增长通道。IDC测算显示，单用户智能体的日均Token消耗可达传统单轮交互的生成式AI应用的百倍至千倍，若全部依赖集中式云端推理，单用户年均服务成本将突破现有商业模式的承载上限，同时还将面临算力供给、网络时延、能效、隐私风险等多重技术瓶颈。这一矛盾在终端厂商侧尤为突出：当前多数终端厂商需承担大部分云端推理成本，却难以转化为用户可感知的体验差异与持续付费意愿，导致“成本高企但价值倒挂”的商业模式困局。无论是经济模型还是技术可行性，单一依赖云端集中式计算模式，已难以支撑个人AI的规模化落地。

AI产业的破局方向并非无限扩展云端算力，而是重构算力分配模式。核心思路是构建端边云一体化的计算连续体——打破端侧、边缘、云端各自独立的物理节点边界，形成算力可无缝流转、负载可动态调度的统一计算体系——在这一体系中实现算力的智能分配，在每一个计算层级达成每瓦特性能与单位Token成本的最优解，从而整体实现成本、性能与能效的最优平衡。

通过统一的AI计算架构与动态任务调度机制，可构建跨计算连续体的协同能力：系统可根据任务的时延要求、隐私等级与计算复杂度，自动将计算负载分配到最优节点运行，高频轻量任务在端侧处理，中等复杂度任务由边缘节点承接，高复杂度推理任务按需调用云端算力。例如，在健康场景中，用户通过智能手表发起健康状态查询时，实时心率、步数等轻量数据直接在手表端处理，运动轨迹分析与短期健康报告由手机端完成，而长期健康趋势预测与个性化运动方案生成则按需调用云端算力。

这种分布式计算模式，本质是让算力供给围绕用户的场景需求动态流动，而非集中于单一节点。

相比传统以云为中心的模式，既从根源上降低了全链路推理成本，也在低时延与数据隐私之间取得更优平衡，为个人 AI 的规模化商用奠定了经济与技术双重可行的基础。

3.2 异构计算架构的必然性：支撑智能体自主运行、规划与跨设备协同

随着AI从被动响应走向主动执行，终端侧的计算需求正发生根本性变化。智能体不再只是完成单次推理任务，而是需要持续理解用户意图、拆解复杂任务流程、调用多种工具并驱动多设备协同执行，这对算力提出了“持续运行 + 多任务并行 + 实时响应”的全新要求。这一变化要求计算架构需要从单一通用走向专用异构，才能在算力、时延、功耗和不同处理任务需求之间实现动态平衡。CPU的核心角色从通用计算单元演进为智能体运行的调度中枢，承担系统调度、多模块协同、端云算力分配与跨端任务调度四大核心职责，支撑智能体跨设备稳定运行。

底层能力的持续演进，让智能体的“行动式能力”具备了落地基础。智能体AI需要主动理解用户意图，拆解复杂任务、调用多种工具并协调多个设备协同完成任务，同时还要持续学习用户的行为偏好和使用习惯，因此低功耗感知能力、多模态处理、任务规划与资源调度成为系统的核心能力，这要求终端计算架构从设计之初就围绕智能体打造，通过CPU、GPU、NPU及各类不同功能模块的深度协同，构建贯通“感知-理解-规划-执行-记忆”的完整技术链路。

单一通用计算架构无法同时兼顾多类型任务的算力需求、时延控制与功耗平衡，难以支撑智能体的持续稳定运行。在专用异构架构体系中，CPU负责整个系统的调度管理与通用计算；高效能、高计算密度的NPU和GPU则专注于高效AI推理、生成与并行计算；还需要专用低功耗感知单元、ISP、多媒体处理模块与内存子系统的深度协同。在智能体运行流程中，CPU的全局调度在任务执行中发挥着至关重要的作用，通过统一调度规划多模态 workflow、协调管理各类异构计算模块、动态分配本地与云端算力和调用各领域相关API，保障复杂任务的高效、流畅运行。

专用异构计算架构将重塑个人 AI 的技术底座：一是实现能效的全面提升，破解智能体全时运行与终端续航的核心矛盾；二是支撑多负载并行处理，满足AI推理、感知、多媒体编解码等多元任务需求；三是统一全层级计算范式，为跨端边云的算力协同奠定硬件基础。

IDC认为，这种高性能、高效能的专用异构计算架构，是实现个人AI系统级能力的核心算力基础。个人AI时代的计算架构不仅需要单设备内部的算力配比优化，还需要在此基础上进一步延伸，打造跨端、边、云的统一异构计算架构：以智能体为调度中枢，围绕用户的任务目标，在多设备之间实现动态分配算力资源并完成任务编排调度。

3.3 个性化智能的关键技术支柱

要实现千人千面、服务随时开启的个性化智能体 AI 体验，需构建“感知 - 理解 - 规划 - 执行 - 记忆”的完整能力闭环。其背后依赖四类核心技术的协同支撑：全时感知与自然交互、端侧算力与个人记忆、连续计算与跨端协同、安全可信与隐私保护。

以用户为中心的个人AI生态，核心要求是智能体能够跨终端无缝流转，随时随地为用户提供主动、深度个性化的服务。物理距离上与用户贴身、全天候伴随用户的各类移动终端，是提供个人AI体验的核心载体与感知端点。终端侧的原生AI能力直接决定了智能体服务的个性化程度、响应速度、体验连贯性与隐私安全性。因此，边缘终端领域的底层计算架构和端侧 AI 技术储备与演进方向，对个性化智能的规模化落地具有决定性意义。根据IDC的持续观察，当前移动计算芯片领域的头部厂商已构建起覆盖异构硬件加速、跨平台AI软件栈、跨终端分布式协同、连接能力和隐私安全保护的完整技术体系，为四大关键技术支柱的成熟落地提供了核心底层支撑。这些技术共同构成了个人 AI 的技术底座，支撑着从感知用户意图到执行复杂任务的完整闭环，也从底层保障了“以用户为中心”的系统本质。

全时感知与自然交互

该技术是支撑个人 AI 实现全天候主动服务的关键。对智能体而言，传感器数据的价值不只是支持新的交互方式，更在于形成持续情境信息，使智能体能够理解用户所处环境、当前状态和任务意图，并在合适时机主动提供服务。缺乏情境信息，智能体很难从被动响应走向真正实用的主动服务。

其核心包含两大层级：一是低功耗感知硬件层，依托主芯片集成的超低功耗传感器中枢，配合外设智能传感器阵列，实现对用户行为、生理状态、环境信息的全天候持续采集与初步处理。传感器中枢可在主芯片低功耗或休眠状态下独立工作，并行处理加速度、心率、环境光等十余类传感数据，待机功耗仅毫瓦级，不会影响设备续航。这一设计让手机、智能手表、智能眼镜等各类随身消费终端，能够在不影响日常使用体验的前提下，为智能体提供持续的感知输入。同时，不同终端的感知能力形成互补：手机覆盖全面的环境、位置与行为感知，智能手表更适合专注生理体征监测，智能眼镜提供第一视角的视觉、听觉和空间感知，共同构建覆盖用户全生活场景的感知网络。硬件层面的全天候低功耗采集，是实现“长效无感感知”的核心前提：在用户授权和系统策略允许的前提下，智能体无需用户手动唤醒，就能持续感知用户状态，主动适应用户的行为节奏，在合适时机推送精准匹配的服务。

二是情境理解算法层，通过多模态感知融合技术，整合视觉、听觉、运动姿态、位置、生理体征等多维度数据，结合端侧多模态大模型的实时推理能力，精准判断用户所处场景与当前状态。这是支持个人AI意图理解能力的关键：即使用户描述模糊、需求不完整，系统也能结合历史上下文、情境信息和用户偏好推断真实诉求，自动生成可执行任务。在此基础上，个人AI支持语音、手势、眼动、微表情、头动追踪等全模态自然交互方式，无需用户适应系统的操作逻辑，以最自然的方式与智能体完成交互。

端侧算力与个人记忆

个人AI时代，用户对于智能体交互的实时性、连续性和个性化要求达到前所未有的高度。要支持智能体对用户意图的快速响应和准确理解，强大的端侧算力与本地记忆技术不可或缺。端侧算力并非简单的硬件性能堆砌，而是针对个人AI“轻量高频、后台常驻”的运行特点进行深度定制：构建专属的软硬件协同和算力调度机制，利用CPU快速处理意图理解、偏好匹配、任务编排等与用户个性化直接相关的任务，通过GPU、NPU加速AI推理任务，同时实现用户前台操作与智能体后台运行的资源隔离，互不干扰。通过混合精度等模型压缩、量化技术与内存优化，在手机、PC等终端有限的功耗和内存约束下，实现多模态大模型的本地流畅运行，让绝大多数日常个性化需求都能在终端本地快速响应。

如果说端侧算力是个性化服务的动力引擎，那么个人记忆就是实现专属体验的灵魂。端侧构建的分层结构化个人记忆体系分为两层：短期记忆层在系统内缓存近期用户交互上下文与临时任务状态，保障多轮对话的连贯性；长期记忆层则通过在芯片层构建用户个人知识图谱，持续记录和学习用户偏好、行为模式、知识关联、社交关系与日程安排等多维度数据。搭配高效本地向量检索技术，智能体可在毫秒级时间内从海量历史数据中精准调取与当前任务相关的记忆片段，实现“越用越懂你”的持续进化。这种端侧原生的技术架构，从根源上保障了所有个人上下文记忆、行为偏好数据的所有权完全归属用户本身，这正是“以用户为中心”的核心权益体现——不仅将用户需求放在首位，同时实现了更安全、更隐私且响应更快速的体验。

连续计算与跨端协同

AI在推动终端品类和规模快速扩展的同时，也让可跨终端延续的AI服务成为刚需。因此，连续计算与跨端协同技术成为个人AI不可或缺的技术支柱之一。传统模式下，任务的载体是设备，换机、切换设备意味着任务中断、数据迁移、体验割裂；而在个人AI体系中，服务的核心是用户，智能体的状态、上下文、任务进度始终跟随用户身份同步，不受设备边界限制。

这需要构建覆盖全层级的统一底层计算架构：形成从毫瓦级贴身可穿戴终端、消费级手机、PC，到边缘计算节点、高性能数据中心的完整芯片谱系，实现硬件指令集、计算范式与软件开发框架的底层统一，为计算和任务的跨端无缝迁移提供统一的软硬件解决方案基础。该架构确保了智能体的核心能力可在不同算力层级的终端上一致运行，将加速个人AI智能体应用和服务跨云端和广泛边缘终端的规模化普及。

高速可靠的连接技术是跨端协同的关键支撑。当前5G-Advanced网络已具备超高速上行、低时延的稳定连接能力，可充分支撑现阶段个人AI跨端交互、情境数据同步等核心需求。面向未来更长周期的技术演进，6G作为AI原生的新一代无线通信技术，有望在现有能力基础上，进一步拓展个人AI的跨端情境感知协同和沉浸式体验边界。

安全可信与隐私保护

从底层确立用户的数据主权、授权机制和安全边界，是打造“以用户为中心”的个人AI体验的关键保障。由于个人AI需要持续理解用户行为、偏好、上下文和长期记忆，其所处理的数据比传统应用更加敏感，也更接近用户个人生活和工作场景。因此，安全可信与隐私保护必须成为个人AI系统设计的内生能力，而不是后置补丁。

硬件层面，终端可通过专用安全处理单元，构建独立于主系统的可信执行环境（TEE），实现关键用户信息处理的硬件级隔离。软件层面依靠联邦学习、差分隐私两类隐私增强技术保障数据安全：多台终端可联合优化通用AI模型，但所有原始用户数据全程保存在本地；差分隐私技术会在数据中添加可控噪声，避免逆向推导还原用户隐私信息的风险。

数据流转层面，建立可审计、可追溯的细粒度数据授权机制，让用户完整掌握自身数据的使用权限与流转记录。产业层面在通过提前对齐技术路标、统一安全标准等方式，推动AI服务隐私保护技术的规模化落地，让用户在享受个性化服务的同时，无需担心数据泄露风险。

综上，个人AI所需的核心技术是跨端、边、云的基础设施、计算架构和软件技术栈的系统级重构。以Token经济学规律为底层驱动，以跨终端的统一异构计算架构为硬件基础，以四大关键技术为核心支撑，整套体系的最终指向始终是同一个目标：让智能适配人。这正是“以用户为中心”的个人AI区别于传统AI的核心本质，也是其持续创造用户价值、推动产业新一轮增长的核心动力。

PART 4

个人AI时代终端与场景创新展望

4.1 个人AI催生终端角色演变

个人AI的落地，正从底层重塑终端设备的产业定位与价值逻辑。传统终端时代，每类设备都是独立的功能载体，价值由自身硬件参数与内置功能决定；而在个人AI体系下，所有终端都成为统一智能体的分布式能力端点，核心价值不再局限于单机功能，而是取决于其在全场景生态中贡献的感知、计算与交互能力，服务始终围绕用户跨设备流转。

全品类终端角色的系统性重构，需要各终端侧高效AI处理能力的集体跃升，以及统一的底层计算基础设施支撑。IDC观察到，随着个人AI的发展，产业对跨终端异构计算、低功耗AI处理与统一软件生态的需求正在快速提升，具备跨终端产品布局、统一计算架构和持续生态投入的平台型企业，正逐步成为个人AI时代的核心基础设施提供者。在这一趋势下，以高通为代表的行业实践者，其覆盖手机、PC、智能眼镜、可穿戴设备、数字座舱到数据中心的统一异构计算架构与AI软件栈，使不同形态终端可在同一套智能体系下分工协同，为个人AI生态的落地提供了坚实的硬件支撑。

在2026年的短期时间窗口内，终端演进仍将在现有品类框架内展开，核心是各终端基于自身硬件属性完成角色重构，通过差异化分工共同构建覆盖用户全生活场景的智能服务网络。而高通覆盖全计算连续体的系统级计算解决方案能力，将为长期个人AI生态建设奠定基础。

4.1.1 智能体AI手机：从应用容器到个人AI体验头号入口

在IDC的AI终端分级体系中，GenAI Phone⁴将构成端侧生成式AI能力的通用硬件底座；而智能体AI手机正是在此算力基础上，叠加系统级智能体调度能力的高阶演进形态，也是智能手机从“应用运行容器”向“个人AI体验入口”升级的核心方向。作为使用频率最高、生态最完整的核心终端，手机正成为当前最成熟、最易承载智能体体验的个人AI落地起点，其新角色体现在三个维度：一是算力调度枢纽，承接各类日常AI任务的端侧处理，同时可为智能眼镜、可穿戴等轻量化终端提供算力中继，实现端边云的动态算力分配；二是数据与身份中枢，承载用户个人记忆、身份认证与核心数据存储，是全生态设备的数据同步底座；三是生态连接中枢，统筹跨设备的任务流转与智能家居、车载等场景设备的联动调度。

4:根据 IDC 分类标准，GenAI Phone指搭载算力>30TOPS 的智能手机

IDC认为，当前手机的升级方向正从单纯提升AI算力，逐步转向构建持续感知、个人记忆、任务编排和跨终端协同能力。因此，能够同时满足性能、能效、连接和隐私需求的平台，将更好地支撑智能体AI的持续运行体验。以第五代骁龙8至尊版为例，该平台围绕性能、能效与端侧AI能力进行了系统级重构，为可持续运行的个性化智能体AI体验奠定了基础。平台搭载第三代Qualcomm Oryon CPU，在单核、多核性能及响应速度和能效方面均实现显著提升，可为即时性AI任务和复杂多任务场景的规划调度提供算力支持。Hexagon NPU在架构层面引入更多标量与向量加速器、更高性能的张量加速器，并支持64位内存架构，使终端侧能够高效部署更大参数规模的大模型，同时提升每瓦特性能，为终端侧AI的长期运行提供关键保障。

与此同时，高通传感器中枢通过超低功耗设计持续感知用户情境，在本地构建个人知识图谱与个人记录，使AI可在合适的任务场景下在本地完成低时延、隐私敏感的处理；也可与其他计算资源协同处理，理解用户偏好、主动采取行动，并始终确保数据隐私与安全可控。Adreno GPU与AI增强的5G及无线连接方案，通过异构计算与高效连接，能够加速AI推理与多模型并发运行，提升跨端协同体验。整体来看，第五代骁龙8至尊版并非单一性能升级，而是一个以高效能计算为基础的系统设计，兼顾性能、能效、隐私与生态协同的移动平台，让个人AI走向可落地、可规模化的实际体验。

4.1.2 智能眼镜：从辅助配件到轻量化交互入口

智能眼镜正加速从手机的附属配件，向独立的轻量化交互与感知终端演进。智能眼镜对个人AI的重要性并不局限于显示，更在于其天然具备第一视角、多模态感知和全天候伴随能力，这使其成为个人AI获取现实世界上下文信息的重要来源。依托第一视角的传感器布局，其核心价值在于两大不可替代的能力：一是情境感知入口，为智能体提供最贴近用户视角的环境视觉、空间姿态与听觉信息，是实现主动情境服务的重要感知载体；二是沉浸式交互，以丰富直观的语音、视线、手势等交互模式，承载高频轻量化的信息获取与指令下达，打造“所见即所得”的无感交互体验。

针对不同定位的XR产品，头部芯片厂商提供分层算力支撑：面向高端沉浸式空间计算设备，骁龙Reality Elite平台可提供48TOPS终端侧AI算力，支持大语言模型与大视觉模型本地运行，同时优化视觉透视效果与手势追踪能力，打造更自然的混合现实体验；针对大众轻量化智能眼镜，高通的骁龙AR平台能够在更加严格的尺寸、功耗要求下提供眼镜侧AI、影像、音频和空间计算能力，满足更加轻薄、舒适的佩戴体验。

4.1.3 AI可穿戴设备：从数据采集工具到全天候感知核心

AI可穿戴设备正在从单一的生理数据采集工具，升级为个人AI体系的全天候贴身感知端点。凭借低功耗、随身佩戴的天然优势，其核心角色转变包括三个方面：一是全时感知载体，7×24小时持续采集用户生理体征、运动状态与行为习惯数据，为智能体的主动服务提供持续的情境输入；二是服务延续触点，在手机不在身边的场景下，承接快捷提醒、语音交互等轻量服务，保障AI体验的连续性；三是隐私安全屏障，依托本地算力闭环实现敏感数据不出设备，从源头保障用户生理、行为数据的安全性。

骁龙可穿戴平台至尊版作为业内首个集成专用NPU的可穿戴平台，为这一角色升级提供了关键技术底座。其可在毫瓦级功耗约束下通过Hexagon NPU运行十亿参数级模型，结合多传感器融合与低功耗连接技术，在保障隐私与低时延的同时，实现对用户情境的持续感知与实时决策；覆盖手表、吊坠、胸针等多种形态，让可穿戴设备真正成为个人AI的贴身感知入口。依托性能能效设计、多模连接架构和多操作系统支持，该平台可支持个人AI始终在线、跨场景稳定协同运行。

IDC认为，可穿戴设备未来最大的价值不在于显示更多信息，而在于成为用户持续状态感知与长期个人记忆的重要组成部分。

4.1.4 协同与扩展设备：从独立单品到场景化能力补充

IDC认为，个人AI的发展并不会由单一终端完成，而是依赖多个终端围绕用户形成协同网络。在这一体系中，PC、座舱和智能家居等设备更多承担特定场景下的能力扩展角色，为个人AI补充更强算力、更丰富场景以及更长时间跨度的服务能力。其中，搭载骁龙X2 Elite系列平台的AI PC，凭借最高80TOPS的NPU算力与领先的CPU性能和能效，承接复杂内容创作、应用开发等高算力任务，是个人AI的“生产力算力端点”；基于骁龙数字底盘解决方案的智能座舱，依托车载多模态感知与交互能力，打造出行场景的连续智能体验，实现服务在通勤场景的无缝接续；智能家居中控则作为家庭场景的服务枢纽，联动全屋设备提供主动式场景服务。

这类设备共同补全了个人AI的场景边界，让智能服务能够覆盖用户办公、出行、居家全场景；而统一的算力架构与AI软件栈，也为跨场景的算力流转与体验一致性提供了底层保障。

图5 个人AI时代终端角色变化总结表

终端品类	核心定位	核心算力单元	核心感知能力	代表性骁龙平台	主要承载任务
 智能体AI手机	全域算力&身份调度中枢	CPU+GPU+NPU+传感器中枢	屏幕、位置、视觉、语音、运动、多模态环境感知	第五代骁龙8至尊版	任务统筹规划、日常AI任务处理、轻量化算力中继、个人记忆存储等
 智能眼镜	第一视角轻量化交互入口	专用轻量化NPU+ISP视觉单元	视觉、动作、空间、语音	骁龙Reality Elite平台/骁龙AR平台	实景感知、AR实时交互、轻量化指令输入等
 AI可穿戴 (手表/戒指/吊坠等)	全天候贴身感知节点	专用低功耗NPU	心率、运动、生理体征、视觉、语音	骁龙可穿戴平台至尊版	7×24生理采集、轻量提醒、隐私本地处理等
 AI PC	生产力高算力节点	CPU+GPU+NPU	近距离办公环境感知	骁龙X/X2系列平台	多模态创作、应用开发等复杂工作流和云端前置运算等

来源：IDC，2026

4.2 个人终端的应用场景和交互模式展望

从中长期视角看，个人AI对终端的重塑将深入到交互范式与场景形态层面，推动终端产业从设备竞争转向体系化竞争。IDC认为，这一进程分两个阶段递进：中期阶段终端角色发生根本性转变，意图驱动交互全面普及；远期阶段个人终端市场迎来全面重构，形成以用户为中心的全新产业生态。

4.2.1 中期演进：角色重构与意图驱动交互普及（2027-2030 年）

这一阶段的核心变化集中在三个维度，推动个人AI从辅助工具升级为用户数字生活的核心载体。

- **一是手机角色深层跃迁。**随着智能眼镜、可穿戴设备承接更多一线交互任务，手机的核心价值将从直接交互入口，转向边缘侧算力载体与智能体调度中枢。硬件设计重心从屏幕、影像等交互参数，向算力、散热、异构调度效率倾斜，闲置状态下的手机可动态作为家庭边缘算力节点，为轻量化终端提供后台推理支撑，实现算力资源复用。
- **二是终端边界逐步融合。**功能、形态与生态的边界持续消解：特定功能不再绑定单一设备，智能体可根据场景自动调度最优终端承载任务；模块化设计兴起，终端可通过模块组合适配不同场景；跨品牌智能体协同逐步落地，用户体验不再受设备品牌与品类的边界限制。
- **三是意图驱动交互全面成熟。**用户无需记忆操作指令与功能入口，通过语音、肢体动作甚至行为模式即可自然表达需求；智能体结合情境信息与个人记忆主动推断真实意图，跨设备任务接力成为常态，主动式服务模式逐步普及，真正实现无感智能、以用户为中心的生态系统体验。

IDC认为，进入远期阶段，终端产业的底层设计逻辑有望逐步转变——从“为人类操作设计工具”，更多地转向“为人类与智能体共同操作构建能力载体”，终端市场将从形态、能力到产业生态等多维度迎来系统性重构。

- **第一，终端形态趋向分布式演进。**IDC观察到，以人机交互为核心的传统终端形态将逐步弱化，分布式部署的感知、算力、执行功能单元的重要性将持续提升。各类单元通过统一智能体体系协同工作，形成贴合用户的全域智能网络，用户感知到的更多是无处不在的 AI 服务，而非孤立的硬件设备。
- **第二，核心能力趋向解耦与分布式部署。**IDC判断，算力、感知、执行三大核心能力将逐步解耦，并在更适配的载体上分布式布局：算力形成“贴身终端 - 边缘设备 - 云端集群”的多层级网络，实现全局动态调度；感知形成“终端侧 + 网络侧 + 环境侧”融合的全域感知体系，覆盖用户与环境全量信息；执行能力从数字世界延伸至物理世界，逐步贯通“感知 - 理解 - 决策 - 行动 - 记忆”的完整闭环。
- **第三，价值模式加速演进。**IDC研判，以意图理解为核心的个人AI服务网络将逐步成形，用户直接表达需求，智能体自动调度全生态资源完成任务；产业价值链有望从传统的硬件销售驱动，向基于“硬件+AI服务价值”的利益共享模式演进：芯片厂商作为算力与连接底座的核心供给方，是个人AI基础设施的关键建设者；终端厂商聚焦差异化能力节点，打造硬件产品，应用服务商提供可复用的服务能力，全产业链各环节围绕用户价值协同重构分工体系。

IDC认为，实现上述愿景仍依赖于底层计算架构、连接技术和软件生态的持续协同演进。当前产业界已开始围绕端边云分布式计算、空间计算、智能体运行框架以及下一代连接技术展开前瞻布局。以高通为代表的产业参与者，正依托自身在终端高性能低功耗计算、先进连接与端云协同领域的技术积累，打通从终端到云端的完整技术链路，为个人AI生态的稳步发展提供底层技术支撑。

4.3 全栈能力平台加速个人 AI 终端落地

IDC观察到，随着个人AI从概念走向规模化部署，产业竞争正在从单点技术突破转向平台能力竞争。行业领先企业正通过硬件平台、软件工具链、产业协同和生态建设等多维能力，帮助终端厂商和开发者降低创新门槛，加速个人AI体验落地。

作为终端智能计算的核心技术提供商，以高通为代表的头部芯片厂商不仅通过芯片产品构建算力底座，更打造了覆盖硬件参考设计、软件工具链、深度合作研发到生态共建的全栈赋能体系：一方面帮助终端厂商降低研发门槛、加速产品上市节奏、提升体验竞争力；另一方面帮助开发者打造能够覆盖云端、边缘和终端的一体化AI体验，推动整个个人AI产业从单点创新走向生态繁荣。

4.3.1 全场景硬件矩阵与标准化方案，降低研发准入门槛

以混合AI（Hybrid AI）为顶层技术路线，搭建覆盖端-边-云的统一计算连续体，构建了覆盖全品类终端的算力产品矩阵，从毫瓦级的可穿戴设备、消费级手机与智能眼镜，到PC、车载、物联网场景和数据中心超大规模推理，形成完整算力梯度。所有产品共享统一的专用异构计算架构，确保不同终端之间的算力调度与模型部署具备底层一致性，为跨端协同体验筑牢硬件基础。

针对不同成熟度的终端品类，提供分层硬件支撑方案：对于手机、PC 等成熟品类，通过旗舰平台的持续迭代不断刷新端侧AI的每瓦特性能上限，支撑更复杂的智能体能力落地和智能体应用的规模化部署；对于智能眼镜、可穿戴等快速成长的新兴终端品类，通过骁龙Reality Elite、骁龙可穿戴平台至尊版等平台，针对不同终端的形态与场景需求，提供高能效的端侧AI推理能力，拓展个人AI的场景边界；通过骁龙START等个人AI设备生态系统支持计划，将芯片解决方案封装为标准化硬件模块，搭配兼容主流AI框架的完整软件栈，让品牌厂商无需从零搭建硬件体系，大幅缩短研发周期。该计划还整合了制造合作伙伴网络，为厂商提供从研发到量产的全链路支撑，帮助更多品牌快速推出个人AI终端产品。

4.3.2 全栈软件工具链，打通模型到应用的落地路径

硬件算力是基础，软件工具链则是终端厂商快速落地AI能力的关键。高通通过统一AI软件栈，为开发者提供跨终端一致的开发体验，支持一次开发、多端部署，大幅降低了厂商在不同品类终端上的适配成本。

软件栈配套GENIE大模型推理优化库与全新智能体开发SDK：前者针对端侧大模型推理做深度性能优化，大幅提升低比特量化模型的运行效率；后者封装了智能体任务编排、上下文管理、跨端调度等通用能力，简化Agent类应用的开发流程。

AI软件栈全面兼容主流AI框架与运行时环境，集成成熟的模型量化、裁剪与优化工具，帮助厂商快速将大模型部署到不同算力层级的终端上。面向智能体开发需求，还提供从模型到应用的全流程工具支撑：高通AI规划器作为高通AI软件栈的组成部分，可整合用户个性化偏好、本地情境信息与终端应用能力，帮助厂商构建智能化的AI体验，实现端侧个性化知识图谱、情境感知服务等核心能力，大幅降低模型部署和智能体能力开发难度。

高通还通过一系列产业合作，加速AI模型在搭载其产品的各类终端上的部署。来自Hugging Face生态的AI模型，现可通过智能体接入高通平台，由智能体负责配置、优化与部署，全程无需人工干预。这使开发者能够通过从边缘到云端的统一工作流，更轻松地将先进的AI能力引入搭载高通产品的智能手机、PC、可穿戴设备、智能座舱之中。同时，高通与Hugging Face计划共同支持一个分布式AI框架，使智能体能够在端侧、云端系统间协同运行，并根据性能、成本、隐私及时延需求，对模型与工作流进行动态规划。

对AI软件基础设施企业Modular的收购，将进一步补齐跨端边云的统一推理架构软件底座，为广大终端厂商、开发者提供统一、开放的AI原生软件栈，支持AI应用跨各类硬件架构高效运行。

4.3.3 推动产业标准协同，共建开放生态体系

个人AI的完整体验离不开全生态协同，而行业标准的统一是生态开放的基础。高通凭借其产业影响力，联合终端厂商、模型厂商与应用伙伴，共同推动跨终端智能体连接、数据传输、隐私安全等领域的行业标准建设，逐步打破不同品牌之间的智能孤岛，为跨设备的智能体协同与流转创造条件。

IDC认为，个人AI的发展需要跨产业链协同推进，覆盖连接技术、数据治理、隐私保护与跨终端协同机制等多个领域。当前产业各方已开始围绕相关方向展开合作探索。以高通为代表的厂商通过“AI加速计划”等产业项目，携手运营商、终端厂商、模型厂商与ISV伙伴，共同探索边缘领域的智能能力和应用场景，推动一系列智能化体验在终端领域落地。通过生态联动，高通不仅为终端厂商提供底层技术，更助力其对接更广泛的模型与应用资源，丰富终端的AI服务能力，提升产品的用户价值与市场竞争力，共同推动个人AI产业的规模化发展。

图6 高通全栈式赋能终端厂商加速智能化方案落地

赋能层级	核心内容	具体构成	产业价值
第一层： 全场景标准化硬件矩阵	提供覆盖全终端品类的AI算力标准化芯片平台	第五代骁龙8至尊版（手机）、骁龙X2 Elite（PC）、骁龙Reality Elite（XR）、骁龙Wear Elite（可穿戴）等	降低硬件研发门槛，厂商无需自研芯片即可快速落地顶尖AI能力
第二层： 统一全栈软件工具链	提供统一AI开发工具与模型库，简化AI应用开发	AI Hub（预优化模型库）、GENIE（大模型推理优化库）、智能体开发SDK	加速AI应用落地，开发者可直接调用工具与模型，大幅缩短开发周期
第三层： 行业标准共建与开放生态建设	推动跨品牌、跨设备的生态合作与行业相关技术标准建立	跨端协同协议、隐私安全框架、开放API接口	打破生态孤岛，实现不同品牌设备无缝协同，保障用户连贯智能体验

来源：IDC，2026

PART 5

个人AI时代终端市场展望与 厂商行动建议

个人AI正成为终端产业新一轮增长的核心驱动力。本章将基于IDC产业数据与趋势研判，结合市场规模走势，为产业链各环节参与者提供战略方向与行动参考。

5.1 个人 AI 终端市场发展规模预测

短期来看，终端市场仍面临需求波动与成本压力，但AI已成为拉动行业增长的核心变量。伴随端侧算力、模型能力与生态体系的持续成熟，个人AI终端的渗透节奏将逐步加快，推动终端产业价值重心从硬件本身向体验与生态迁移，整体市场规模将在此周期内迎来结构性扩容。

5.1.1 细分品类市场规模与渗透率预测

智能手机：个人AI的核心承载中枢

智能手机是个人AI能力落地的核心载体，也是端侧智能体技术规模化渗透的核心阵地。基于IDC对AI智能手机的算力分级体系，终端市场正沿着“基础AI能力普及 — 生成式AI能力下沉 — 智能体能力升级”的路径持续演进。

据IDC 2026全球AI终端市场跟踪报告数据，2026年全球智能手机出货量预计超过10.88亿台，其中AI Phone整体出货量超7.22亿台，渗透率达66.4%，其中GenAI Phone占比已接近40%；端侧AI功能已逐步成为主流机型的标配能力。

从市场演进节奏来看，GenAI Phone是当前AI手机市场扩容的核心基本盘，整体规模仍处于快速提升周期；而智能体AI手机作为GenAI Phone中具备系统级智能体能力的高阶细分形态，将成为中高端市场增长的核心驱动力。预计到2029年，全球智能手机年出货量将超11.78亿台，GenAI Phone整体市场渗透率将超过60%，其中搭载完整系统级智能体能力的智能体AI手机机型将成为中高端市场的主流配置；到2030年，GenAI Phone市场渗透率将进一步接近64%，智能体AI能力将同步向中端机型渗透，跨端协同、本地记忆、主动服务等系统级个人AI能力将成为中高端机型的核心竞争维度。

智能眼镜：新兴交互与感知增长极

智能眼镜正处于从小众尝鲜向大众普及的过渡阶段，是个人AI终端赛道中增速最快的品类之一。IDC数据显示，行业当前已进入高速增长周期，2026年全球智能眼镜出货量预计达1989万台，同比增长34.6%。2026-2030年，市场复合年均增长率（CAGR）约为25.5%，预计2030年全球年出货量接近5000万台。

IDC认为，智能眼镜在个人AI终端体系中的生态战略地位将持续提升——它将成为与智能手机协同搭配的最重要生态产品之一，与作为个人AI核心承载中枢的智能手机形成能力互补。凭借第一视角感知、解放双手交互的独特优势，智能眼镜将成为个人AI生态中不可或缺的视觉感知入口与轻量化交互终端。

可穿戴设备：全天候贴身感知入口

IDC数据显示，当前全球可穿戴设备保有量已超13.5亿台，覆盖手表、手环、耳机、戒指等多元形态。预计2030年全球可穿戴设备年出货量将超过6.6亿台，其中智能戒指、耳机等产品的年复合增长率更高。具备本地AI感知、轻量推理能力的AI可穿戴设备占比将持续提升，功能从单一数据采集升级为个人AI体系的全天候贴身感知节点。

PC：生产力场景的核心算力端点

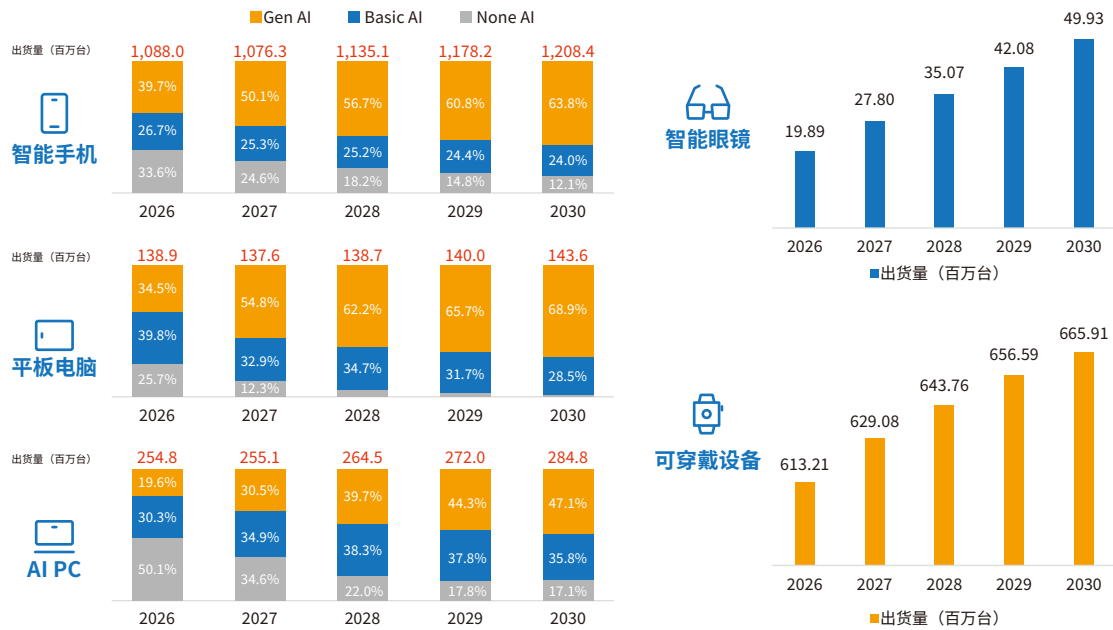
AI PC是近年个人终端市场增长最快的品类之一，也是个人AI在生产力场景的核心算力支撑。2026年全球PC市场出货量预计将超过2.54亿台，其中AI PC出货量预计约1.27亿台，渗透率达到49.9%，正处于快速渗透的上升期。IDC预测，到2030年全球PC年出货量约2.84亿台，AI PC渗透率将超过82.9%，基本实现全面普及。端侧AI算力与智能体调度能力将成为PC的标准配置，推动产品从传统生产力工具，升级为个人AI分布式算力网络中的核心生产力端点。

平板电脑：跨场景协同的补充节点

平板电脑是AI渗透速度最快的终端品类之一，2026年全球平板出货量预计约1.39亿台，其中AI平板出货量预计超1.03亿台，渗透率达74.3%，基础AI功能已实现广泛覆盖。预计到2030年，全球

平板年出货量将超1.43亿台，其中GenAI平板占比将接近69%。作为手机与PC之间的核心场景补充节点，平板将依托统一的智能体调度体系深度融入个人AI生态，实现多设备间的任务无缝流转与协同接续。

图7 IDC 全球主要智能终端出货量预测 2026-2030



来源：IDC，2026

5.1.2 核心驱动因素与市场趋势

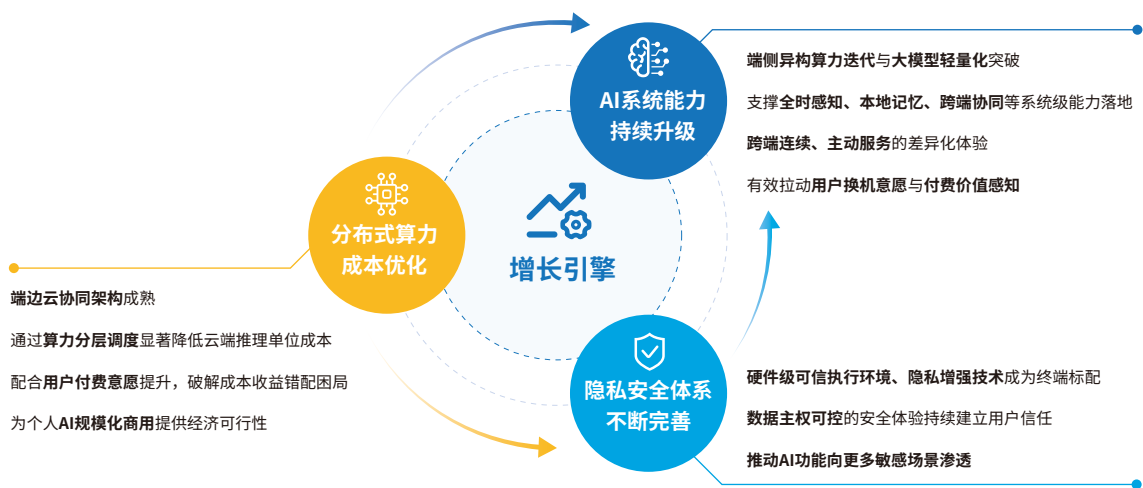
个人AI终端市场扩容，将带来终端产业竞争逻辑的深层迭代——行业竞争将从单一硬件参数内卷，转向AI系统能力、分布式算力成本控制、用户隐私安全三位一体的综合较量。这一转变是市场增长的核心驱动力，将塑造终端市场的长期结构特征。

三大维度共同推动产业向上：一是AI系统能力持续升级，端侧异构算力迭代与大模型轻量化技术取得突破，支撑全时感知、本地记忆、跨端协同等系统级能力落地；跨端连续、主动服务、深度个性化的差异化体验，将有效拉动用户换机意愿与付费价值感知；二是分布式算力成本逐步优化。基于Token经济学规律的端边云协同架构走向成熟，通过算力分层调度显著降低集中式云端推理的单位成本，从供给侧改善产业成本结构；配合AI体验价值提升带动的用户付费意愿增强，将逐步破解行业成本收益错配的困局，为个人AI的规模化商用提供经济可行性；三是隐私安全体

系不断完善。硬件级可信执行环境、隐私增强技术逐步成为终端标配，数据主权可控的安全体验持续建立用户信任，推动 AI 功能向更多敏感场景渗透。

伴随竞争逻辑转变，终端市场结构呈现三大核心趋势：一是价值重心向服务端上移，产业价值分配从一次性硬件销售向“硬件+AI服务”、内容订阅延伸，智能体服务能力与体验品质取代纯硬件参数，成为产品核心定价依据；二是终端形态基于个人AI生态角色分工而分化，不同品类终端围绕感知、算力、交互的差异化定位相互协作，传统“万能终端”的产品思路，逐步被多终端分布式协同体系替代；三是个人AI核心终端的行业门槛系统性提升，系统级个人AI能力成为手机等核心终端市场的主流产品标配，仅叠加单点 AI 功能的终端将逐步丧失竞争力。

图8 驱动个人AI市场向上的三重力量



来源：IDC，2026

5.2 厂商行动建议

个人 AI 时代，行业竞争逻辑已从单机硬件参数比拼，转向全域智能体系的综合竞争。产业链各环节厂商需匹配自身定位落地针对性动作，协同突破算力能效、生态互通、体验确定性三重瓶颈，共同分享终端产业结构性升级的增长红利。

5.2.1 供应链厂商（芯片、底层方案商）

核心目标：搭建统一软硬件技术底座，降低全行业AI终端商业化门槛，打通端-边-云分布式算力链路。

- **迭代分层异构算力硬件。**针对手机、AI PC、智能眼镜、可穿戴设备的体积与功耗约束差异，分梯度优化处理器和AI加速单元运算流水线、超低功耗传感器中枢的功耗调度逻辑；同步完善硬件级可信执行环境（TEE）布局，统一硬件加密、身份隔离技术标准。
- **芯片设计面向智能体任务进行系统级优化。**在硬件架构层面原生支持感知、记忆等智能体核心能力，围绕多步骤智能体任务优化CPU编排调度与内存子系统设计，强化上下文缓存、低时延数据通路与本地记忆存储，为端侧智能体的持续感知、自主规划与稳定执行提供底层支撑。
- **打造软硬件协同的跨终端统一解决方案。**落地标准化AI开发套件，包含统一推理引擎、智能体开发框架与预优化模型库，实现模型一次适配、多品类终端部署；优化INT4/INT8低比特混合量化工具链，简化十亿参数级模型在各类终端的本地部署流程，降低厂商适配成本。
- **推行芯片-终端深度合作模式。**前置参与终端产品定义，提前收集厂商的场景算力需求、续航约束，反向优化芯片架构；派驻软硬件工程师联合调优向量检索、多模态感知等高频个人AI能力，加速技术产品化落地。
- **牵头共建行业统一规范。**联合运营商、终端品牌、模型企业制定跨端智能体数据传输、数据授权、感知交互三类通用标准；开放底层硬件接口，消除不同品牌设备之间的算力调度、数据互通壁垒。

5.2.2 终端厂商（手机、可穿戴、PC等厂商）

核心目标：跳出单机功能内卷，围绕用户完整生活场景搭建多终端协同智能体系，落地个人AI四大核心技术能力。

- **重构产品设计逻辑，以智能体验定义硬件规格。**不再单纯比拼屏幕、影像等传统参数，优先匹配每瓦特AI性能、NPU算力、低功耗常驻传感器配置、本地内存容量；区分主设备（手机/PC）、感知设备（穿戴）、交互设备（眼镜）的差异化硬件配置，适配各自在个人AI体系中的端点定位。

- **布局全品类终端矩阵，补齐场景覆盖缺口。**主力机型预装完整的本地记忆、跨端协同能力；同步推进轻量化智能眼镜、智能手表、智能吊坠等贴身感知终端研发，覆盖居家、办公、出行全场景的感知与交互入口。
- **打通系统与应用两层体验闭环。**系统侧落地全时感知、本地向量检索、跨端任务无缝接续三大核心能力；推动主流应用API调用接口开放合作机制，支撑智能体完成多应用连贯任务。
- **落地分层算力调度机制。**接入端边云动态计算调度框架：心率、手势等轻量感知任务在可穿戴设备本地处理，用户提问等复杂任务交由手机/PC完成，超大模型复杂预测任务按需调用边缘、云端算力。

5.2.3 模型厂商与应用开发商

核心目标：打造可被全域智能体调用的标准化服务组件，提升AI任务执行稳定性，融入多终端分布式生态。

- **模型厂商：**构建跨端统一智能体模型框架，输出分级轻量化方案适配不同算力终端，保障上下文与任务状态跨设备同步，筑牢全场景智能体通用能力底座。
- **软件及应用开发商：**拆解垂类场景能力为可复用功能单元，开放标准化接口供智能体统一调度；深耕多场景用户体验打磨，优化意图理解精度与任务执行确定性，持续拓展智能体服务边界。

产业链上下游围绕核心能力定义、调用规范与算力协同框架协同演进，逐步统一基础技术标准，才能共同推动个人AI从概念走向规模化商用，释放终端产业新的增长空间。

图9 厂商建议汇总表

厂商类型	核心战略目标	硬件层面动作	软件&生态动作
 芯片/供应链厂商	搭建统一算力底座	分层异构算力迭代， 芯片设计从底层适配智能体 工作流，覆盖穿戴到PC	统一AI软件栈， 共建行业基础标准与 互通协议
 终端厂商	打造差异化AI系统体验	布局全品类终端矩阵， 优化设备AI能效比	围绕核心技术重构系统能 力，落地跨端任务流转与分 层算力调度
 模型/应用厂商	构建场景化 可编排服务体系	适配多终端轻量化模型与 算力分级约束	开放标准化应用接口， 适配跨设备智能体统一调度

来源：IDC，2026

IDC认为，个人AI的发展绝非单一维度的技术升级，更是驱动终端产业从“设备中心”向“用户中心”底层跃迁的核心变革。未来产业竞争将跳出单一终端的性能比拼，转向以感知、记忆、推理、协同、服务五大能力为支柱的全生态体系较量。率先搭建起端、边、云一体化的个人AI能力架构，在生态协同效率与用户体验价值上构建持续壁垒的市场参与者，将在下一轮产业重构中掌握发展主动权。

关于 IDC

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC帮助IT专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC在全球拥有超过1100名分析师，他们针对110多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在IDC超过50年的发展历史中，众多企业客户借助IDC的战略分析实现了其关键业务目标。IDC是IDG旗下子公司，IDG是全球领先的媒体出版、会展服务及研究咨询公司。

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

blogs.idc.com

www.idc.com

版权声明

凡是在广告、新闻发布稿或促销材料中使用 IDC信息或提及IDC都需要预先获得IDC的书面许可。如需获取许可，请致信 gms@idc.com。翻译或本地化本文档需要IDC额外的许可。

获取更多信息请访问www.idc.com，获取更多有关IDC GMS信息，

请访问<https://www.idc.com/prodserv/custom-solutions>。

版权所有 2026 IDC。未经许可，不得复制。保留所有权利。