AI在终端运用增长，成为手机、电脑新卖点

■中国城市报记者 孙雪霏

相较于依托数据中心、通过处理巨量数据给出洞察建议的云端AI，终端侧AI多在拍照、安全、联接等领域扮演并不显山露水的赋能角色。随着终端侧AI技术的不断进步，叠加智能手机和PC（个人电脑）市场漫长的累库周期已过，产能逐渐恢复，市场开始将AI与终端的融合视为新的创新锚点。

如今，终端侧AI正在走向台前，成为手机、电脑厂商的新卖点。虽然目前尚未出现真正的“杀手级应用”彻底改变市场格局，但端侧大模型竞赛已趋白热化，手机、电脑未来将如何被重塑？

消费电子景气度底部改善 AI成为新卖点

vivo在11月1日举行的开发者大会上，展示了其AI解决方案中的前沿技术——“蓝心小V”。这款由vivo自研大模型加持的智能手机助手，不仅能高效检索照片和文件，还可在复杂场景下实现路人消除功能，甚至助力用户提炼论文要点，以及创作社交媒体内容和生成思维导图。

vivo并未明确表态这些功能是否完全基于终端。但据透露，为“蓝心小V”提供支撑的自研蓝心大模型矩阵中，70亿参数版本已落地移动终端，而更高级的130亿参数模型也已实现终端侧跑通。11月13日，vivo即发布了搭载“蓝心小V”、端侧支持70亿大模型的X100系列手机。

与此同时，三星于11月8日展示了其Gauss大模型，OPPO则计划于11月中旬公布安第斯大模型的最新进展和特性。这些动作紧跟在10月下旬小米、荣耀公布其自研端侧大模型进展之后。不到一个月，几大主流安卓手机品牌齐聚端侧大模型赛道。另一方面，华为则早在8月就宣布其智慧助手“小艺”已经接入了自家的盘古大模型能力。

将经过精细训练的大模型引入终端设备，已是终端与芯片行业新一轮技术竞赛的核心。2022年11月，OpenAI发布聊天机器人ChatGPT，迅速引发了全球范围内关于大模型的AI热潮，互联网、云服务和AI领域企业纷纷入局。如今，这场技术革命之“火”从云端烧到终端。

为端侧大模型提供底层算力的芯片领域火药味渐浓。10月下旬，高通在骁龙峰会上发布新款旗舰处理器第三代骁龙8，成为首个在手机终端侧支持百亿参数大模型的芯片平台。该平台基于Stable Diffusion模型，其文生图功能的运行速度由2023年一季度的15秒提升至仅需0.6秒。在此次峰会期间，小米、荣耀也宣布了各自可运行于高通平台的自研大模型进展。

联发科则抢在高通发布会前夕，公布了与OPPO、vivo在端侧大模型上的合作。11月6日，联发科在其新款旗舰手机芯片“天玑9300”发布会上，展示了其在AI大模型领域的实力。这款芯片不仅支持70亿参数规模大模型落地，还成功地在端侧运行了130亿参数模型，并正在探索端侧运行330亿参数大模型的可能性。在展示用例中，联发科称，基于这款芯片的文生图速度已降至不到1秒。

PC芯片领域，竞争亦在白热化。高通在骁龙峰会上发布了其新款PC处理器骁龙X Elite，AI特性成为最大卖点。与此同时，英特尔也计划于12月发布集成NPU（神经网络处理器）的新品“酷睿Ultra”，意图直指AI PC市场。根据市场调研机构Counterpoint数据，2023年三季度全球智能手机与PC两大市场均录得环比增幅。在触底期已过的乐观预期下，市场对AI的提振作用寄予厚望，认为AI PC可能会在2024年推动新一轮出货反弹，并有望在2026年后主导PC市场。

大模型塞进终端 效果几何

塞进AI大模型的终端设备，究竟会有什么不同？华为在今年8月的新产品发布会上，基本奠定这一轮手机端侧大模型落地基础应用——智能助手。升级后的变化在于手机助手从原先仅限于语音交互，扩展为支持语音、文字、图片和文档等多种输入形式，并能从单纯的准确指令执行，进步为自然对话。

几乎与华为同步，小米也在8月对其语音助手“小爱”进行了大模型赋能，实现了从过去较局限的固定对答语库向复杂指令交互的升级，极大地增强了处理复杂指令的能力。此外，小米还宣布成功在手机端跑通了自研的13亿参数模型，部分场景效果比肩云端的60亿参数模型，并在10月进一步将端侧大模型参数升至60亿。

小米技术委员会AI实验室的负责人栾剑在接受中国城市报记者采访时表示，传统的智能助手由于需要精确指令才能给出正确反馈而被许多用户视为不实用。他提到，大模型技术可在复杂指令和多轮交互上带来全新体验，使得用户可以更自然、随意地与AI进行交流。此外，在生成能力上，AI不再仅仅是娱乐工具，而是成了一个实用的助手。

在高通骁龙峰会上，荣耀展示了其YOYO助手通过大模型技术，帮助用户以语音搜索手机中的影像，并按指令生成视频的能力。而vivo的策略有所不同，推出了独立的GPT产品蓝心小V，而不是升级原有的手机助手。vivo解释称，选择相对保守的应用方案缘于用户仍可能延续使用习惯。两款产品可能会在未来成熟后融合。

相较于终端厂商，芯片厂商携合作伙伴展示的使用场景更轻量和聚焦。例如，高通与慧鲤科技合作推出的“照片扩充”功能，可以通过AI补全已拍摄照片的周围景观，创造广角效果。联发科则展示了更贴近中国市场的应用，比如快速生成表情包的“文生趣图”。此外，高通合作伙伴元智能（RWKV）也展示了于手机端侧生成音乐的能力。

在PC市场上，AI也是新一轮角力之处。Counterpoint高级分析师William Li表示，从硬件角度看，苹果搭载M系列自研芯片的Mac系列PC已具备AI PC的特性。而随着英特尔和高通新款芯片的推出，预计到2024年上半年，我们将看到更广泛的AI PC应用。

手机端侧大模型瓶颈：性能和能耗难平衡

高通公司总裁兼CEO安蒙（Cristiano Amon）曾就如何将大模型装入手机提供了深刻见解。他表示：“尽管大模型在训练和调优上存在差异，其参数规模越大，通常意味着模型能力越强。”安蒙详细说明了量化技术在优化模型训练过程中的重要性，其不仅确保了减小模型参数时不会对准确性产生显著影响，还能实现模型计算的高效率。他进一步解释说：“我们的核心任务是实现模型的极致精简，甚至是转换浮点模型计算为更高效的低比特定点模型计算。”

受制于终端设备处理器算力、内存和存储容量及电池续航等各方面瓶颈，与云端动辄成百上千亿参数的大模型相比，端侧大模型参数量目前堪堪触及百亿，差距一望而知。因此，业界多计划采取端侧大模型搭配云端大模型的混合式AI路线。

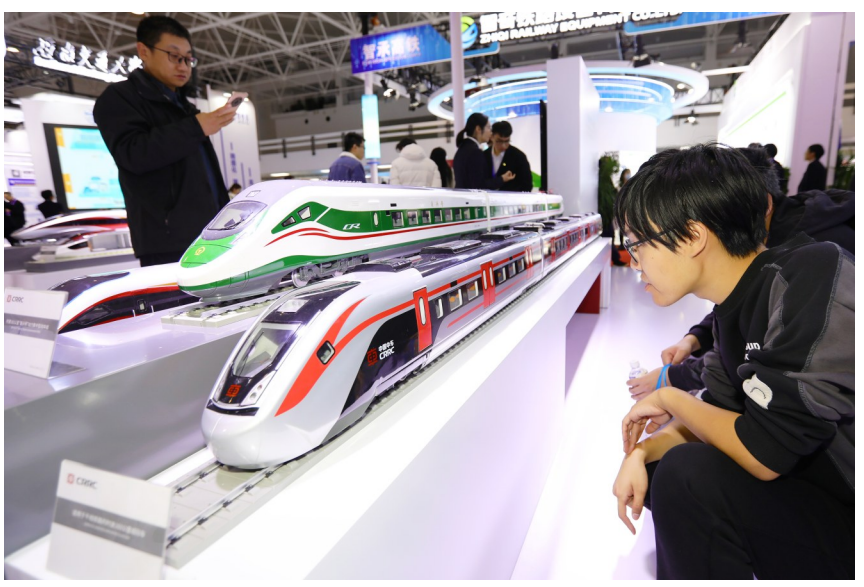
元智能联合创始人罗璇认为，未来半年到一年，手机端将能跑通140亿及以下参数规模的大模型，PC端则有望容下600亿参数的模型。在此情况下，云端与终端亦不会是割裂状态。他说：“未来可能出现的情形是，手机上运行一个140亿参数的大模型作为OS（操作系统）的‘发动机’，而云端则运行一个比GPT-4更大的模型，作为整个下一代互联网的底座。这两者将相互配合，如同当前的本地软件与互联网。

手机厂商中，vivo采取了矩阵策略，其蓝心大模型包含70亿及以下参数的端侧版本和700亿、1300亿、1750亿参数的云端版本。vivo副总裁、vivo AI全球研究院院长周围称，虽然目前70亿参数模型已能较好支持文档摘要、拆解等功能，但要具备中台级的上下文理解能力、接近类人思维能力的“智能涌现”——一个标志性的500亿参数门槛，还需要“再往上走走”。因此，130亿参数大模型成为实现智能体的更好选择。但截至目前，130亿参数大模型的内存占用已接近7GB，这几乎是高端智能手机12GB内存的一半。除去大模型占用后，就只剩下了类似中端手机的内存性能。

相较于内存占用，罗璇认为，能耗是更大的瓶颈。他指出：“内存是可以增加的，但电池容量的提升却受限于电池密度的挑战。”他认为，当前大多数大模型基于Transformer架构，其算法复杂度决定了性能和能耗难以平衡。

荣耀CEO赵明曾在与媒体交流中指出，端侧大模型必然带来更高的硬件需求。赵明透露，未来一段时间，手机厂商会宣传端侧大模型，但当前在端侧普遍使用的通常只是10亿、20亿参数规模的“小模型”。他进一步表示，任何AI应用，如果不能平衡用户隐私、算力和低功耗，就无法提供更好的消费者体验。

求解未来升级路线，栾剑认为，解决方案首先包括增加内存容量和带宽，以便终端设备能够承载更大的模型；其次是提升或优化算力，更高效地支持大模型的网络结构与算子；此外，软件层面上，持续探索模型的压缩和量化，以及推理算法的改进，从而在保持同等效果的条件下降低算力需求。“更重要的是，端侧大模型的发展将改变手机的使用方式，可能催生新的生态系统，因此软件架构需要适时调整，以适应新的需求。”他说。



现代化铁路技术装备展览会在京举行

近日，第十六届中国国际现代化铁路技术装备展览会在中国国际展览中心（朝阳馆）举行，来自14个国家和地区近400家企业参展，展品涵盖铁路全产业链的最新技术和装备产品。

中国城市报记者 全亚军摄