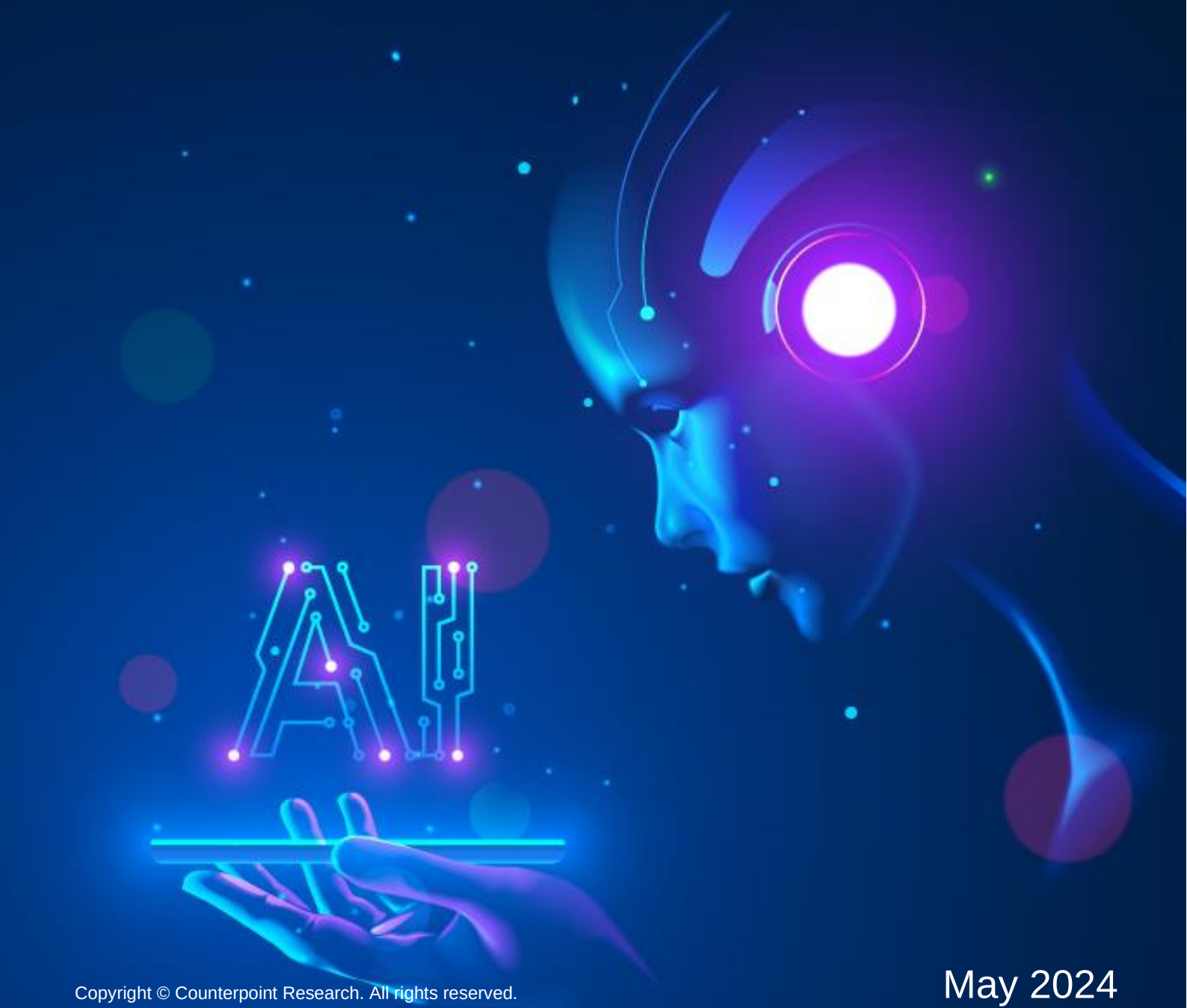
MEDIATEK | Counterpoint Technology Market Research

阿里云 | 百川智能 BAICHUAN AI | 虎牙直播 huya.com | KUGOU | 零一万物 01.AI | oppo | Soul APP | Tencent AI Lab | 腾讯混元 | vivo

生成式 AI 手机产业白皮书



May 2024

本白皮书由 Counterpoint 与联发科技共同发布，其他联合发布者包括阿里云通义千问、百川大模型、虎牙、酷狗、零一万物、OPPO、Soul、腾讯 AI Lab、腾讯混元、vivo（按公司名拼音首字母顺序排列）

目录

前言

2022 年 11 月 30 日，ChatGPT 上线，并迅速获得追捧。这场最初由 ChatGPT 引发的生成式 AI 浪潮，让全球消费者惊讶于大语言模型（LLM）所带来的全新人机交互体验的同时，也让产业界充分认识到生成式 AI 技术在消费者（C 端）市场的巨大应用潜力。而智能手机，作为当下最重要的个人智能终端，在全球范围内拥有超过 40 亿用户规模，无疑是生成式 AI 技术在 C 端应用成功与否的重中之重。

另一方面，智能手机产业在进入 5G 时代后，也需要一场真正意义上的颠覆性革新，为消费者带来更加智能、个性化，同时也更加安全的使用体验，将手机打造成全天候的私人智慧助手和移动生产力工具，从而为智能手机下个十年的发展打下坚实的基础。生成式 AI 技术与智能手机的融合刚好契合这一需求，它将全方位赋能智能手机产业，革新包括硬件、软件，以及相关的移动互联网内容生态的方方面面。

生成式 AI 与智能手机的融合既是产业各方的需要，也是 AI 普惠的必由之路，以智能手机为媒介，全球手机用户能够更便捷、更高效地享受生成式 AI 技术发展所带来的福祉。

本白皮书提出了生成式 AI 手机的概念，讨论了生成式 AI 手机生态中各个玩家，包括芯片厂商、手机厂商、大模型厂商、开发者的相关 AI 战略，以及围绕生成式 AI 手机的软硬件科技全景，最后是 Counterpoint 对生成式 AI 手机发展的预测。

第一章:智能手机开启生成式 AI 手机时代

从 2023 年底至 2024 年一季度，新一代旗舰智能手机陆续发布，越来越多的基于生成式 AI 能力的功能开始出现在这些产品中。无论是手机厂商还是其生态伙伴，在主动拥抱生成式 AI 趋势的同时，也在积极探索各种可能性，着力打造对用户有价值的高频使用场景，而这一探索将贯穿整个 2024 年。Counterpoint 认为 2024 年会是生成式 AI 手机的元年。

手机的智能化演进

大约在二十多年前，以诺基亚塞班为代表的操作系统，第一次允许用户自行下载 APP，并将其作为入口，访问服务和数字内容，这种模式一直持续到今天。也正是这种变化，使得全球的开发者可以加入到智能手机产业中来，为手机用户提供丰富多样的应用选择，促成了之后移动互联网生态的蓬勃发展，手机也逐渐发展为人们休闲娱乐、通信社交、健康和出行服务、消费购物，以及移动办公的重要载体，早已不可或缺。

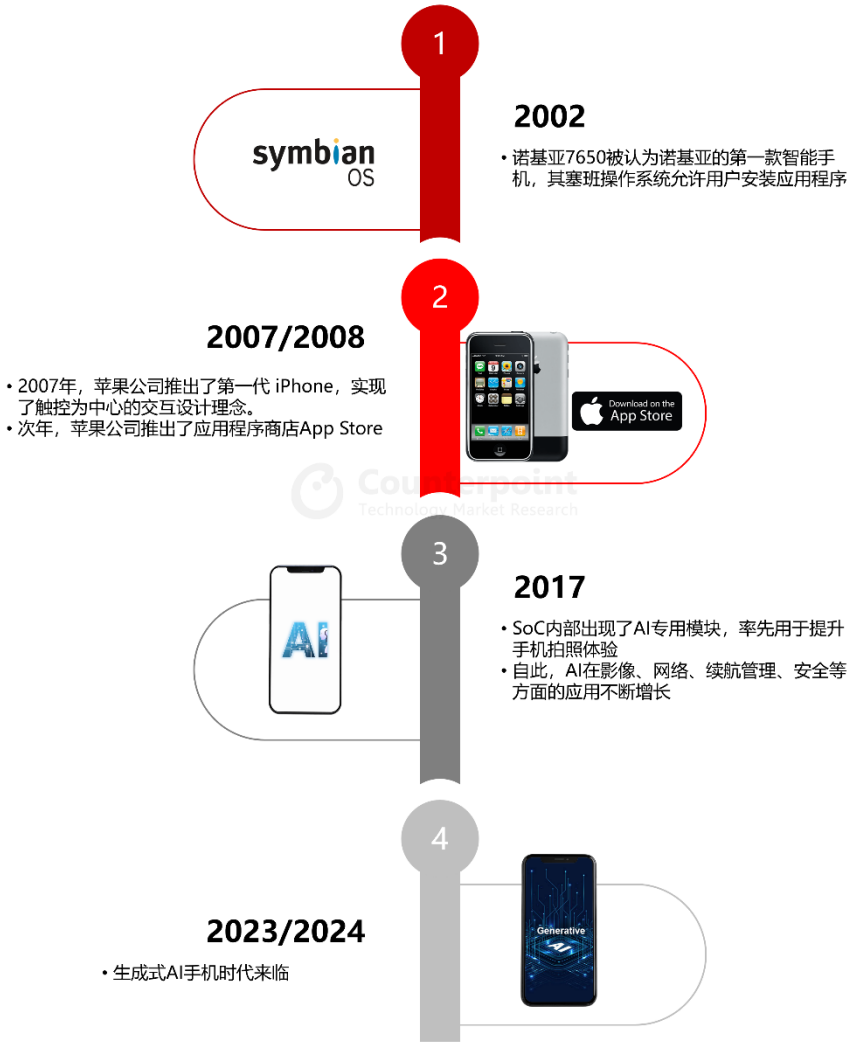
2007 年，iPhone 的问世颠覆了传统的手机设计理念，物理键盘逐渐被淘汰，触控屏幕成为人机交互的核心。然而随着时间的推移，在一些场景下，触控输入的方式变得越来越低效，常常需要多次的用户干预，才能到达最终的服务界面。在此背景下，出现了包括智能语音助手，手势、眼球追踪在内的新的交互方式，致力于打造更流畅、高效，更加用户友好的交互体验。

通过 AI 技术赋能智能手机的尝试最早可以追溯至 2017 年，彼时苹果刚刚发布了首款后置双摄手机 iPhone 7 Plus，而安卓阵营也开始在其 SoC 平台中加入独立的 AI 计算单元，用于运行和影像增强相关的深度学习模型。在这之后，AI 技术逐渐被手机厂商用于更多方面，如强化安全、优化续航、提升网络性能等，但计算摄影一直是其最主要的应用领域，直到 LLM 被装进智能手机，手机 AI 应用从中小模型时代跨越至大模型时代。

得益于 AI 大模型的赋能，智能手机将迎来新一轮的革新。首先在人机交互层面，有了 LLM 的加持，新的多模态交互将取代传统的、单一的触控屏交互，逐渐实现从图形用户界面 GUI 到语音用户界面 VUI 的跨越式转变，用户可以以更直观、更自然的方式与手机沟通。其次，多模态输入和输出能力相结合，可以极大强化智能手机的生产力工具属性：既可以基于多种形式的输入信息，生成用户需要的图表、文本、音乐、图片甚至是视频，也可以对输入的图片、视频进行编辑。

最后，随着融合的深入，生成式 AI 技术将在智能手机上孕育出一个甚至多个智能生命体（AI Agent)。智能生命体以用户为中心，不断学习用户的行为习惯，能够智能识别用户意图，适时向用户推荐个性化的内容和服务。Counterpoint 认为智能体将会成为专属于每个用户的应用入口，但预计在很长一段时间里，智能体仍将会和 APP 共存。

图表 1：手机智能化演进路线图



来源：Counterpoint Research

Counterpoint 对生成式 AI 手机的定义

Counterpoint 认为生成式 AI 与智能手机的融合无疑将引发一场深刻的变革。参照过往每一次技术革新，在初期探索阶段，新的功能和特性将首先被赋予算力资源更加充裕的旗舰和次旗舰产品，并迅速成为重要的差异化卖点。而随着时间的推移，生成式 AI 能力将加速下沉，从而能够在全球范围内，惠及更广大的消费者群体。

基于上述判断，Counterpoint 提出了生成式 AI 手机的概念，并结合现阶段生成式 AI 应用的现状，以及对生成式 A I 手机未来发展与演进的预判，给出了如下定义：

生成式 AI 手机是利用大规模、预训练的生成式 AI 模型，实现多模态内容生成、情境感知，并具备不断增强的类人能力。生成式 AI 手机开启了智能手机发展的新周期，长远看，智能手机会发展为移动智能体。Counterpoint 认为，生成式 AI 手机需要具备如下必要特征：

- 支持大模型的本地部署，或是通过云端协同的方式执行复杂的生成式 AI 任务。生成式 AI 手机本身具备强大的 AI 算力，无须完全依赖云端服务器。
- 具备多模态能力，即可以处理文本、图像、语音等多种形式的内容输入，以生成各种形式的输出，典型用例如翻译、图像生成和视频生成等。
- 确保流畅、无缝的用户体验，设备能够以自然而直观的交互方式，快速响应用户的请求。
- 拥有实现上述特征的硬件规格，包括但不限于基于领先工艺和先进架构设计的移动计算平台，拥有集成或者独立的神经网络运算单元（如 APU/NPU/TPU），大容量和高带宽的内存，以及稳定和高速的连接，硬件级和系统级的安全防御。

图表 2：Counterpoint 生成式 AI 手机定义

来源：Counterpoint Research

2024 年是生成式 AI 手机爆发的元年，在产业链的配合下，头部安卓厂商已经成功实现了 70 亿参数大模型的本地部署。在此基础上，一些基础能力被开放给开发者，其中比较有代表性是：基于 Diffusion 大模型的图片和视频生成（本地用例多为低分辨率，如 480p）；基于 LLM 的自然语言处理，包括语音转文本，文本转语音，任务型对话，实时翻译和信息问答等，以及基于 sLLM 模型（轻量型语言模型）的文字校对和文本生成、改写和总结。

端云结合将是生成式 AI 部署的主流模式

相对于手机端有限的计算和存储资源，云端无疑拥有更充足的算力，从而能够支持更大规模的 AI 模型部署和训练，当前一些复杂的生成式 AI 任务主要是通过云侧大模型来实现的。从长远看，Counterpoint 认为端云结合会是生成式 AI 在手机端侧部署的主流模式。一方面，在未来几年，本地大模型无论是规模还是效率都将保持增长，这意味着用户可以从本地获得多数基于生成式 AI 的服务，本地大模型还将为需要云端介入的任务提供数据脱敏、压缩等预处理，以保护用户隐私。另一方面，云侧 AI 大模型可以为用户带来更有价值的服务，比如提供更高品质的内容输出，如影视、动画制作等，或是专为云办公场景打造智能协同平台，可以打破物理空间的边界，允许海量人群参与到同一个项目中。

总之，要用发展的眼光看待生成式 AI 手机这一新现象，目前展示的生成式 AI 用例只是冰山一角。无论发展到哪个阶段，端云协同在满足产业各方需求的同时，也能最大化利用分布在端云两侧的算力资源。在端侧，移动计算平台的每一次迭代和升级，都意味着手机 AI 算力的大幅突破，相应的，消费者可以期待更加流畅、也更加丰富的生成式 AI 体验。同时，生成式 AI 手机的端侧多模态能力也将获得进一步的强化，Counterpoint 认为，多模态能力，包括多模态输入和输出，是生成式 AI 手机愿景得以实现的关键之一，也是实现多元化交互的基础。

多模态是实现 AI 智能体愿景的关键

上文中，我们提到多模态能力是实现生成式 AI 手机愿景的关键。在 Counterpoint 的预测里，随着生成式 AI 手机的发展，将带来如下几个维度的革新：

首先是交互方式的多元化、直觉化，一方面这要求大模型可以识别、理解不同形式的输入内容，用户可以通过文字输入，也可以是一段语音，一张表格，一张图片，一段视频。另一方面，具备多模态输出能力的大模型将以用户为中心，选择最佳的、最适合当前情境的输出方式。可以说，多模态交互是开启全新交互体验的钥匙。

长期以来，语音助手被认为缺乏实用性，但有了多模态 LLM 的加持，语音助手将变得更加智能，它能够更加准确地识别、理解人类的自然语言，不但能够快速理解和响应用户指令，还解锁了语音文本互转、多轮对话等能力，在越来越多的场景下，更自然的语音交互将会成为智能手机输入输出的首选方式。

其次，今年年初 Sora 横空出世，其展现出的文生视频能力备受行业关注，成为多模态应用创新的“新高地”。对于智能手机用户来说，短视频无疑是当下最受欢迎的移动互联网应用之一，将会是未来数字内容传播最主要的媒介，对优质、个性化、多样化短视频内容的需求也将会持续增长。因此，短视频制作是生成式 AI 最重要的应用领域之一，具备视频理解能力的生成式 AI 手机可以提供智能视频剪辑、风格转换、文案自动生成、以及生成配音等辅助功能，成为视频创作者的得力助手。受益于手机 AI 算力的不断增长，

在不久的将来，语言视觉模型（LVM）也将实现本地部署。届时，通过文字和语音提示生成、编辑短视频将成为可能，这将大幅提升视频制作效率，进一步促进短视频生态的繁荣。

最后是AI智能体的成长。多模态大模型可以同时感知不同类型的数据，包括图像、文本和语言，从而为智能体的成长提供更多维度和更加丰富的训练语料。这意味AI智慧体可以像人类一样从不同的媒介获取知识，不断提升对复杂现实世界的理解能力。在这个过程中，AI 智能体将习得“听说读写”这些类人的本领。

此外，伴随着认知能力的发展，AI 智能体将拥有更加全面的能力。联发科技认为未来的智能体还将具备自主决策能力，包括“计划”、“记忆”和“行动”的能力。

- 计划 – 具备任务分解和自我反思能力；
- 记忆 – 短期和长期记忆能力；
- 行动 – 直接执行或利用工具完成特定任务。

在联发科技的设想里，智能体可以学习、记忆手机用户的使用习惯，以及兴趣和偏好，在此基础上协助人处理日常生活 （食、衣、住、行） 以及工作(计划、执行、报告)事项，使得人可以专注在设定目标和进行决策。以旅行场景为例，智能体可以为在出发前为用户量身订制出行方案，在旅行过程中，可以根据天气、交通和景区信息，以及一些突发情况，动态调整计划，让旅行变得更轻松、更加个性化。

端侧部署AI大模型的优势

如前所述，Counterpoint 认为端云结合、优势互补会是生成式 AI 技术与智能手机融合的主流模式。同时，Counterpoint 认可端侧部署AI大模型具有如下优势：

- 低延时：在许多生成式AI任务场景下，网络传输时延是用户无法获得流畅体验的主因，而本地大模型可以更快地响应用户需求，将时延控制在秒级甚至是毫秒级别。
- 安全和隐私：AI 大模型本地部署可以确保用户个人数据不离开手机，结合芯片公司提供的基于底层硬件的防护机制，可以最大程度保护用户的数据和隐私安全。
- 减少对网络的依赖：AI 大模型本地部署可以极大降低对网络的依赖，即便是在弱信号，甚至没有网络的情况下，手机仍然可以提供必要的生成式AI能力，为用户提供不间断的服务。
- 个性化：具有自学习能力的本地大模型可以成长为每个用户专属的智能体，从而有能力为用户提供个性化的服务和推荐。

- 减轻基础设施负载：考虑到全球范围内超过 40 亿的智能手机用户，随着生成式 AI 应用在手机上的普及，对 AI 算力的需求将呈现爆发式增长。通过部署本地大模型，可以在本地完成尽量多的生成式 AI 任务，一方面降低了对网络带宽占用，另一方面将极大减少手机用户对云计算资源的访问和占用。

第二章：生成式 AI 手机生态系统

LLM 现状以及预测

全球范围内已经掀起了生成式 AI 创新的浪潮，Google 和 Meta 在海外是 LLM 的重要创新者和参与者，其各自 LLM 的战略覆盖了从云侧到边缘，再到端侧的全域场景。

2017 年，Google 首次提出了 Transformer 架构，并将其应用于自然语言处理，使其成为最早投入 LLM 模型开发的头部互联网企业之一。2022 年 4 月，Google 推出了 LLM PaLM（Pathways Language Model），参数规模达到了 5400 亿。一年后，Google 发布升级版 PaLM2，通过采用 Compute-Optimal Large Model 技术，优化了模型的规模，并强化了多语言、推理和编程能力，从而使 PaLM2 在真实世界中的表现更加优异。2023 年底，随着全新多模态模型 Gemini 的发布，Google 加快了 LLM 的商用化步伐。Counterpoint 判断 Google 会优先将 Gemini 用于强化自身业务，如搜索、Chrome 浏览器、YouTube 等，并尝试赋能安卓生态。Gemini 目前有三个重要版本，分别是面向端侧应用场景的 Gemini Nano （包括 1.8B 的 Nano-1 和 3.25B 的 Nano-2），以及面向云侧应用场景的 Gemini Pro 和 Gemini Ultra。

2023 年 2 月，Meta 发布自研的基础 LLM LLaMA，包含 70 亿、130 亿、330 亿和 650 亿等四个不同参数规模的版本。同年 7 月，Meta 宣布开源其最新版本 LLaMA2，Meta 还和微软达成协议，后者将在 Azure 和 Windows 上支持 LLaMA2。这两项举措帮助 LLaMA2 赢得了众多产业伙伴的支持，开发者可以选择以 LLaMA2 为基础，围绕自身业务场景定制开发大模型，极大推动 LLaMA2 的产品化进程。在手机领域，部分手机厂商已经将 Meta 和 Google 作为海外部署生成式 AI 的重要合作伙伴。在 LLaMA2 取得显著成就之后，Meta 于 2024 年推出了最新版本 LLaMA3。LLaMA3 在先前版本的基础上进行了多项优化和改进，包括模型参数(7B 到 8B)和上下文长度的扩展(4K 到 8K)、算法效率的提升以及更加精细的模型训练。LLaMA3 进一步提升了自然语言处理的准确性和生成能力，将帮助 Meta 巩固其在生成式 AI 模型领域的领导地位，为开发者和产业伙伴提供了更加强大和灵活的工具，以支持他们在各自的业务场景中创造更多的可能性。

在中国，近两年也涌现出一大批的国产 LLM，其中不乏一些佼佼者，综合性能持续对标国际先进水平，中文能力上的表现更是做到了行业领先。

阿里云通义大模型

通义千问是阿里云通义实验室自主研发的 LLM，2023 年 4 月上线并邀请用户测试体验，9 月首批通过备案并正式向公众开放。2023 年 10 月，通义大模型发布 2.0 版本，模型参数达到千亿级，模型在复杂指令理解、文学创作、通用数学、知识记忆、幻觉抵御等方面的能力都有明显跃升，位居行业前列。通义大模型先后推出了 PC 和 APP 应用，能为用户提供生活、工作、学习、娱乐等多种支持。通义还是业界最具影响力的开源大模型之一，先后开源了 0.5B、1.8B、4B、7B、14B、32B 和 72B 等 7 个不同尺寸的大语言模型，包括 Base 和 Chat 版本，以及视觉理解大模型 Qwen-VL 和音频理解大模型 Qwen-Audio，在业界率先实现全尺寸、全模态开源，通义千问开源系列模型下载量已经突破 300 万。

百川大模型

百川大模型最早发布于 2023 年 6 月，初代产品为 Baichuan-7B，拥有 70 亿参数，主要由中文拼音、汉字和词语，以及英文单词构成，用于中英文的自然语言处理、机器翻译和问答。仅仅三个月后，百川智能宣布开源第二代基础模型 Baichuan 2，包括 70 亿的 Baichuan2-7B，以及 130 亿的 Baichuan2-13B，两者的训练语料均达到了 2.6TB，最高支持 4K 的上下文输入，在文档生成、多轮对话等场景下有着出色的表现。而随着 Baichuan 3 的发布，百川大模型的参数规模已经超过千亿，基础通用能力全面提升，逻辑推理和语义理解能力显著增强。

零一万物 Yi 模型

零一万物是一家致力打造 AI 2.0 时代的前沿大模型技术及软件应用的全球化公司。平台业务核心首重建构行业领先的大模型，如开源的 Yi-34B、Yi-6B 和多模态模型 Yi-VL，以及微调模型 Yi-34B-Chat (0205)、长文本的 Yi-34B-Chat-200K 和 Yi-VL-Plus，其 Yi 系列模型支持中英文等多语种对话和图像识别对话能力，千亿模型 Yi-Plus 将于近期正式亮相。零一万物在模型小型化技术方面有着深厚的积累，剪枝、蒸馏、量化等模型小型化技术一直是零一万物研发的重点之一，此外，零一万物将逐步发布完善的开放平台中间件和开发者工具，助力基于 Yi 系列的“端测模型 + 远端大模型”组合开发消费级和商务级应用。消费级应用业务着重在研发新型态个人效率工作软件“万知”及社交方向；商务级应用层面，零一万物也积极与企业客户合作探索商务级 To B 应用层面的落地场景。

腾讯混元

腾讯混元大模型技术架构已升级为混合专家模型 (Mixture of Expert, MoE) 架构，参数规模达万亿，更擅长处理复杂场景和多任务场景，中文整体表现上处于业界领先水平，尤其在数学、代码、逻辑推理和多轮对话中性能表现卓越。同时，腾讯混元还提供不同尺寸的模型，适应更多的需要低成本和高推理性能的应用场景。作为腾讯全链路自研的大模型，腾讯混元通过持续的迭代和实践，积累了行业领先的技术能力，受到多方认可，中国电子学会 2023 科学技术奖评选，腾讯《面向大规模数据的 Angel 机器学习平台关键技术及应用》获科技进步一等奖。

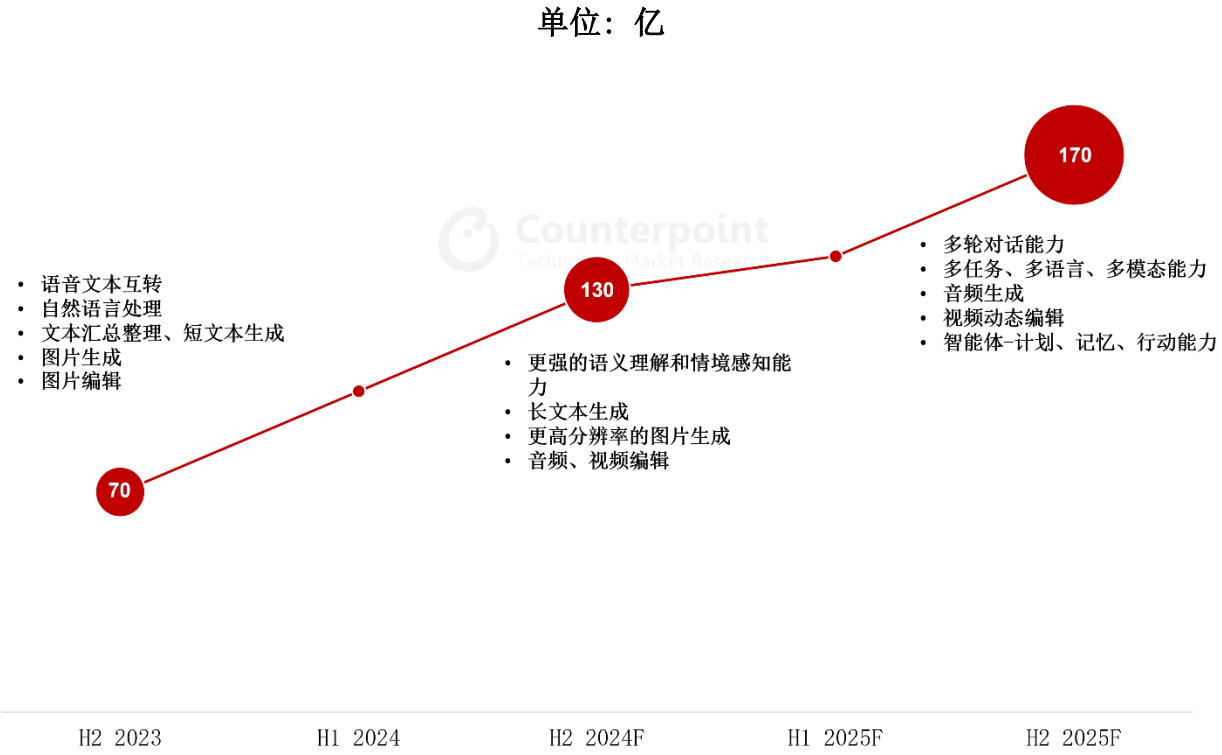
基于扎实的基础能力积累，腾讯混元大模型积极推进相关应用落地，让大模型创造更多价值。目前腾讯内部超过 400 个业务及场景已接入测试，企业微信、腾讯会议及腾讯文档部署了生成式 AI 功能，腾讯广告基于混元大模型推出 AI 广告创意平台妙思，有效提高广告主生产及投放效率。不仅如此，腾讯还联合生态伙伴，将大模型技术与 20 多个行业结合，提供超 50 个行业大模型解决方案，也将与手机行业伙伴一起探索应用生成式 AI 技术为消费者带来全新服务与体验的各种可能性。

受益于众多的开源基础大模型，开发者可以结合自身业务需要定制专属的应用模型，从而加速了 LLM 在端侧、云侧的应用部署。同时，开源更有利于芯片设计企业、手机 OEM 厂商对上述端侧 LLM 进行深度适配和优化。以联发科技为例，在天玑 9300 和 8300 平台，已经完成了针对 Google Gemini Nano，Meta LLaMA2、LLaMA3，百川大模型和通义千问等模型的底层调优。

未来两年端侧大模型参数规模将继续增长

当前，包括 vivo X100 系列，OPPO Find X7 系列，以及荣耀 Magic 6 系列在内的一众安卓旗舰产品已经成功实现了 70 亿 LLM 的本地部署，预计 AI 算力会是未来两代旗舰 SoC 升级的重中之重，从而使端侧部署更大规模的 LLM 成为可能。Counterpoint 预测，本地大模型参数的上限将在 2024 年增长至 130 亿和在 2025 年增长至 170 亿。

图表 3：本地大模型参数预计逐年增长，2023 H2-2025(F)



来源：Counterpoint Research

大模型参数规模的增长将进一步拓展生成式 AI 手机的能力，比如更强大的音频、视频的处理能力，以及对多语言的支持，一些原本面向 AI PC 开发的办公和创作辅助功能也将被迁移至手机端。而另一方面，更大的模型通常意味着更高的硬件成本，同时还会影响到手机的续航能力。考虑到算力、内存、发热的限制，端侧需要采取与云端不同的发展策略，并非一味追求更大的参数规模。

有鉴于此，芯片设计企业如联发科技，正在探索新的技术手段，解决单一模型的限制，如采用 MoE 架构，将多个专家模型融合在一起，既可以针对不同的任务需求动态选择不同的专家模型，又可以对每个独立的专家模型进行微调，从而显著提升整体性能。有了 MoE 的加持，就有可能突破单一模型的天花板，在端侧实现数百亿甚至千亿模型方能达到的准确度。

而云端大模型的发展将继续遵循规模理论（Scaling Law），模型参数越大，模型的通用性和精度越高。未来两年，将有越来越多的云端大模型达到万亿参数规模。

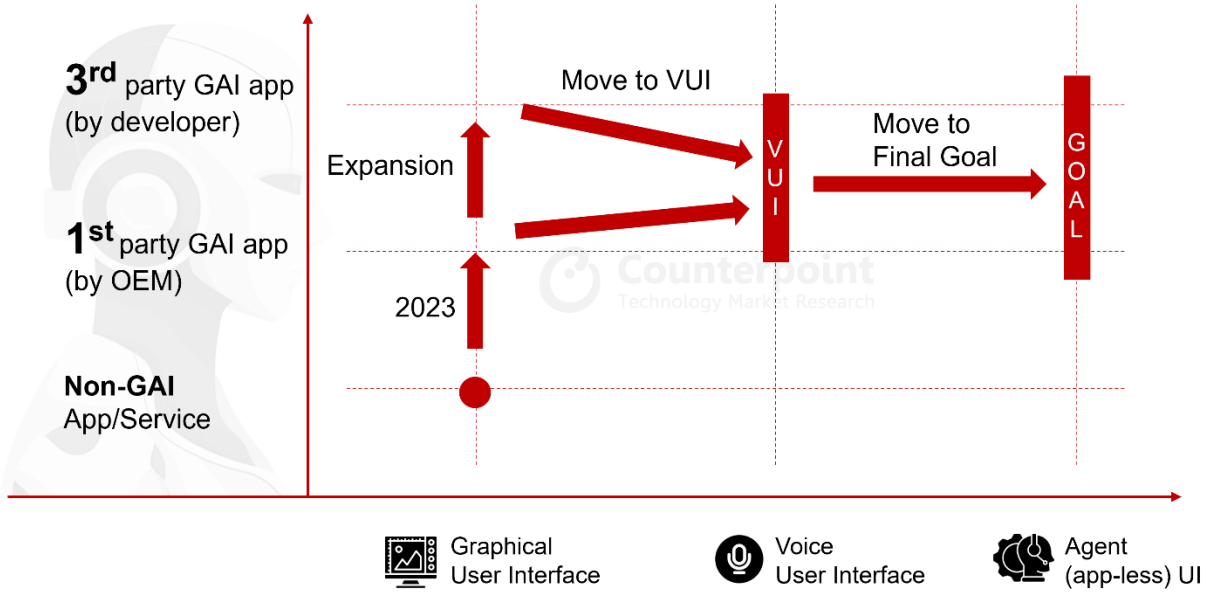
APP 为基础的用户界面与 AI 智能体将会在未来几年内共存

在生成式 AI 手机时代，仍然需要产业生态中的各方群策群力，在各自擅长的领域推动产品和服务的革新。预计移动互联网厂商和广大应用开发者将继续主导移动内容生态的发展，在此背景下，APP 仍将会是构成手机交互的界面的重要元素。

随着越来越多的手机厂商开始贯彻以用户为中心的操作系统设计理念，智能手机可以根据用户的使用场景自动生成个性化的界面设计，包括自动调整图标布局、颜色主题和字体大小等。

另一方面，基于 AI 智能体多元的、自然的交互体验将受到手机用户的青睐。随着生成式 AI 手机的进化，以及生成式 AI 应用生态的繁荣，越来越多的功能和服务将被接入 AI 智能体。在此基础上， AI 智能体将革新智能手机的交互体验，即从传统的 GUI 发展为 VUI，最终升级为全新的 Agent (app-less) UI，这意味着更多的交互将发生在 AI 智能体和用户之间，从而弱化了 APP 的存在感。

图表 4：AI 智能体将发展为未来生成式 AI 手机交互的中心



来源：联发科技；Counterpoint Research

AI 智能体会逐渐成为连接数字生态的入口，用户只需输入想要获得的服务（通过语音、文字等形式），AI 智能体会直接跳转到服务页面，同时 AI 智能体基于对用户习惯的了解以及当前使用情景，以更加安全

和个性化的方式为用户提供服务，或是由 AI 智能体直接完成用户所需要的服务，这是传统 APP 访问模式所不能做到的。然而上述演进不会一蹴而就，在很长一段时间里，Counterpoint 判断 AI 智能体与 APP 会同时活跃在生成式 AI 手机里。此外，头部互联网厂商有可能围绕自己的业务生态打造专用的 AI 智能体，这将使得未来的生成式 AI 手机中可能有多个 AI 智能体共存，并能够相互沟通，携手为用户提供智能体验。

芯片设计公司的生成式 AI 战略

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。时至今日，手机系统中的 AI 算力增长了近 20 倍，而本地部署多模态大模型对 SoC AI 算力提出了更高的要求。在未来几年，受益于半导体代工厂量产更先进的工艺（2nm 和 1nm），以及芯片设计公司和 IP 企业开发出更适合生成式 AI 大模型运行的新型计算架构，SoC 的 AI 算力还将保持两位数的年增长率。

头部芯片设计公司，如联发科技、高通已经率先投入到生成式 AI 手机的浪潮中，发布了多款支持多模态大模型端侧部署的移动计算平台，其中包括 2023 年四季度发布的天玑 9300、天玑 8300、高通骁龙 8 Gen3，三星猎户座 Exynos2400，以及 2024 年第二季度发布的骁龙 8s Gen3 和天玑 9300+，为 2024 年生成式 AI 手机的大规模商用铺平了道路。

此外，联发科技、高通两家公司基于多年来在 AI 手机应用方面积累的丰富经验，为 OEM 和开发者提供了一整套开发平台，分别是联发科技的天玑 AI 开发套件 NeuroPilot 和高通 AI Stack、AI Hub。在这里，开发者可以获得大量训练好的 AI 模型，并有配套工具可以帮助快速将模型嵌入到应用中。

联发科技的生成式 AI 战略主要体现在硬件、软件和生态三个方面。硬件上，全新的 APU 引擎深度适配主流的 Transformer 架构，并支持包括 INT4、INT8、INT16 和 FP16 在内的各种整数和浮点格式，具有行业领先的高能效。软件方面，为在天玑平台部署大模型和开发 AI 应用提供了完整的开发工具链以及集成开发环境。而在生态合作方面，联发科技不断拓展与手机厂商、大模型厂商以及应用开发者的合作，携手共建生成式 AI 手机应用生态。

综上，头部芯片设计公司将在生成式 AI 手机时代发挥更重要的作用，将帮助构建从芯片设计到大模型，再到工具链的全局解决方案。

手机 OEM 厂商的生成式 AI 战略

手机 OEM 厂商也是生成式 AI 手机浪潮的重要推动者，一方面，OEM 希望通过生成式 AI 技术来全面提升现有产品，打造差异化的硬件产品，继续推动智能手机功能的革新。另一方面，借助智能手机向智能生命体演化的机遇，OEM 可以进一步提升品牌科技内涵和用户粘性，与用户建立起长期而紧密的联系。

截至目前，多数头部 OEM 已经发布自研 AI 大模型，如 vivo 的蓝心大模型 BlueLM，包括从最小的 10 亿参数到最高 1750 亿参数云端大模型，目前落地了 10 亿、70 亿端侧大模型，端侧跑通 130 亿参数模型；

OPPO 的安第斯大模型 AndesGPT，Tiny 版本为目前端侧主流采用的 70 亿参数规模，Titan 版本包含 1800 亿参数；小米发布了面向端侧部署的 AI 大模型 Xiaomi MiLM，拥有 130 亿参数，三星和荣耀也分别发布 100 亿和 70 亿参数的多模态大模型。

OPPO 生成式 AI 战略

OPPO 从很早以前就开始布局 AI 赛道，截至 2023 年底，其在 AI 领域的全球专利申请超过 3160 件，主要分布在计算机视觉、语音技术、自然语音处理、机器学习等方面。在此基础上，OPPO 已经成功落地了 100 多种 AI 功能，包括视觉模型支撑的 AI 消除功能（日均使用次数达到了 15 次），以及 LLM 支撑的全新的小布助手。为了更好地迎接生成式 AI 手机时代的到来，OPPO 在深圳大湾区建设了滨海湾数据中心，部署了千卡规模的训练算力，并在 2024 年 1 月成立了 AI 中心，预示着 OPPO 将继续加码 AI 布局。2024 年 3 月，OPPO Find X7 系列顺利通过泰尔实验室 AI 手机测试，使得 OPPO 成为首批获得上述认证的手机厂商。

在 LLM 自研方面，OPPO 早在 2020 年就发布了基于 BERT 架构的预训练语言模型，并于 2023 年发布了自主训练的安第斯大模型（AndesGPT），在 SuperCLUE 给出的十大基础能力排行中，AndesGPT 获得了“知识与百科”能力国内大模型第一。OPPO 同时开发了面向端侧和云侧的 AI 大模型，打造了端云协同框架。OPPO 的端侧模型目前聚焦在 130 亿以下规模，云侧训练了包括 130 亿、340 亿和 700 亿及以上多个规模的版本，可以灵活适配不同业务场景。

在端侧，OPPO 不断在小布助手中加入新技能，包括演讲稿生成、头脑风暴、文档助手、AI 作画等，用于提升生产效率。同时，OPPO 还将重点打造面向影像创作、通话摘要、实时翻译等高频场景的 AI 功能。在云侧，OPPO 希望通过与第三方平台的合作，实现全局知识搜索和广域知识问答，并帮助用户缩短知识和服务获取的路径，OPPO 已经发布包括小布英语老师、小布面试官、小布问答在内的一系列相关功能。

OPPO 秉持用户价值驱动，开放协作共赢的理念，通过采用 MoE 架构，OPPO 与芯片厂商、国内外头部大模型技术方案及生态伙伴密切合作，其自研的 70 亿 LLM 在联发科技的天玑 9300 平台上实现了极快的推理速度，10 亿视觉模型可以做到秒级文生图（512x512 分辨率）。

vivo 生成式 AI 战略

vivo 是较早设立人工智能团队的手机厂商之一，于 2017 年组建 AI 团队，2018 年成立 AI 全球研究院（一直保持千人规模的专家团队），并在 2023 年 11 月率先发布自研 AI 蓝心大模型，并成为行业首批在手机端落地大模型并实现端云协同的厂商。

vivo AI 蓝心大模型的优势：

- 海量数据：自 2018 年成立图谱团队，vivo 已积累了 18,000T 的多种模态数据以及 6,750T 高质量中文文本数据，其中训练数据超过 30T，相当于 5 个国家图书馆的藏书量;

- 全面的模型矩阵：针对不同的应用场景、计算能力以及成本需求，vivo 提供矩阵式的解决方案，覆盖十亿、百亿、千亿的参数级别，通过提供从小到大的选择来保证更好的产品体验。在 2024 年，vivo 图像大模型将迎来重大升级，vivo 还将发布音频大模型和多模态大模型；

- 强大的算力：截止目前，vivo 已拥有充沛的训练算力规模，以强大的算力和工程能力支撑自研大模型；

- 高效的算法：vivo 已在 AI 顶级学术会议（AAAI、ICLR、ECCV、CVPR、InterSpeech 等）发表 70 多篇论文，申请了 800 多项发明专利；

- 丰富的场景：vivo 在 AI+影像，办公、视频、音乐、游戏、无障碍、生活、健康、出行、安全等多领域，拥有丰富的应用场景。

- 凭借算法、工程、数据、系统、应用层面的持续积累，vivo 蓝心大模型在中文整体表现上处于业界领先水平。在中国信通院组织的可信 AI 大模型标准符合性验证测评中，蓝心大模型在大语言模型能力评级中，获得最高等级的 4 星+认证。

基于蓝心大模型的人工智能助手蓝心小 V/蓝心千询，能够提供高效、智能、贴心的语言交互服务，可以在学习、工作、生活上为用户提供便捷高效的贴心服务，包括但不限于任意问题答疑解惑、文本创作、美图生成、文档处理、情感聊天等，从而为用户带来更加便捷和愉悦的智能生活体验。

在与生态伙伴的合作方面，vivo 已经自主研发了成熟的大型模型端侧业务能力，将来会向生态伙伴开放使用。同时，vivo 已经开源了 7B 模型，并将开放基于 7B 模型的 1+N 架构，以支持互联网生态伙伴基于大型模型的业务开发。

为了更好地服务全球消费者，在大模型的采用，以及生成式 AI 应用生态发展上，OEM 厂商秉持开放、合作、共赢的理念，正积极与第三方大模型平台以及云服务商展开合作，并为开发者从事生成式 AI 应用研发提供必要的技术和资金支持。

Counterpoint 判断领先的 OEM 厂商会通过自建云计算中心，或是租用第三方云计算资源，不断优化自有 AI 大模型。此外，能否与产业生态伙伴构建深层次的、面向 AI 大模型和生成式 AI 应用的战略伙伴关系，也将影响到各 OEM 厂商在生成式 AI 时代的表现。

开发者生成式 AI 战略

毫无疑问，广大开发者将是生成式 AI 应用生态的重要贡献方。智能手机已经渗透到人们生活和工作的方方面面，庞大的 APP 生态满足了消费者社交、娱乐、出行、移动支付、线上购物以及移动办公的需要。

生成式 AI 技术将会赋能开发者社区，帮助 APP 开发者革新现有应用，并创造出惊艳的基于 LLM 的新应用，从而围绕生成式 AI 手机形成新的应用生态。届时，生成式 AI 手机将成为连接消费者和数据世界的纽带。

虎牙

虎牙直播是以游戏直播为主的互动直播平台，正在积极探索生成式 AI 技术在直播领域的应用，包括 AI 美容、AI 主播助手、AI 观众助手、千人千面的虚拟主播以及 AI 专区等，为主播和观看直播的用户打造虎牙特有的个性化使用体验。

- AI 美容：基于图生图能力，虎牙开发了 AI 美容功能，可以在手机端实现低延时的 AI 美颜，提高了才艺主播的开播数量。
- AI 主播助手：基于多模态大模型，AI 助手可以理解游戏主播的说话内容和直播画面，在此基础上和弹幕进行互动，解说游戏事件、发起聊天话题，提升直播过程中的互动体验。
- AI 观众助手：AI 观众助手可以根据观看历史分析观众喜好，通过文本、语音交互等方式为观众推荐有趣的或是有用的直播内容。新的交互推荐可以弥补传统算法理解观众意图能力弱，不能互动式进行推荐的弱点。
- 千人千面的虚拟主播：基于多模态大模型，观众可以指定主播的个性、样貌、直播内容等，并可实时地与生成的虚拟主播进行互动。不仅是虚拟主播，本地部署了 LLM 的生成式 AI 手机还可以根据观众的喜好修改主播的容貌、声音、画面风格，做到一个主播开播，不同观众可以得到个性化定制的直播观看体验。
- AI 专区：为了丰富观众在直播前和观看后的体验，虎牙创建了 AI 专区，其中包括了基于 AI 文生图的 AI 相机功能，基于 LLM 的文字挑战游戏，角色扮演，主播分身等丰富的功能。
- 此外，虎牙还十分关注 AI 智能体，以及文生视频技术的发展，这些高阶能力将可能彻底改变直播行业的生产模式和直播中的交互体验，并带来更加丰富的应用场景。

酷狗

在音乐领域，AI 大模型也将带来新的价值。作为中国领先的数字音乐交互服务平台，酷狗音乐正积极地探索如何通过生成式 AI 技术赋能音乐内容的创作，并为听众带来个性化的音乐体验。借助 AI 大模型的多模态能力，酷狗音乐将赋能音乐人创作出更丰富优质的音乐内容，包括个性化定制创作等。音乐也可以反向作为输入，AI 可以输出更全面的描述，为用户生成个性化歌单，帮助用户更好地感受音乐的魅力。通过端云协同部署 AI 大模型，可以赋能音乐的创作、分类、推荐以及风格转换等，这将产生出全新的业务形态和与之匹配商业模式，实现音乐产业的升级。

Soul

Soul 是上线于 2016 年的新型社交平台，是少有以虚拟人设提供即时交流互动体验的应用和 AI Native 的社交网络。在 AI 社交时代，通过 AI 实现关系推荐、对话辅助、表达门槛降低、社交体验提升，是 AI Native 社交网络的关键，也是 Soul 的重要机遇。2020 年，Soul 启动对 AIGC 的技术研发工作，并在智能对话、图像生成、语音技术等方面拥有技术积累。2023 年，Soul 正式上线自研语言大模型 SoulX。围绕年轻一代的核心社交需求，Soul 陆续上线、内测了智能对话机器人“AI 苟蛋”、AI 辅助聊天、虚拟陪伴等诸多工具和创新功能，进一步丰富平台用户的社交体验。此外，Soul 也在尝试将更多 AI 能力引入到产品体系之中，包括开发陪伴型 AI 来提升用户的游戏化社交体验，文生图、文生视频等工具降低表达门槛，以及虚拟人能力提升交互沉浸感等。同时，Soul 也一直在关注和跟进 AI 技术在端侧推理设备上的结合与落地，从而更好提升用户使用体验和提供隐私保护。

腾讯 AI Lab

腾讯 AI Lab 自 2016 年起投身游戏 AI 研究，迄今已发布绝艺、绝悟、开悟等多项世界领先的科研成果，技术应用覆盖亿级玩家。在 2024 全球游戏开发者大会（GDC）上，腾讯 AI Lab 发布了自研游戏 AI 引擎 GiiNEX，基于生成式 AI 和决策 AI 技术，为游戏全生命周期提供丰富的 AI 解决方案。在生成式 AI 方面，GiiNEX 面向 AI NPC、场景制作、内容生成等场景，提供了包括 2D 图像、动画、3D 城市、剧情、对话、关卡以及音乐等多样化的生成式 AI 能力，帮助开发者提升高质量内容生成的效率。

GiiNEX 包含了前沿算法模型、高效训练平台、在线推理引擎三大核心，可支持游戏从研发到运营的全生命周期需求。其中，基于强化学习、自然语言处理等 AI 基础研究能力构建的统一算法底层模型，能够支持 MOBA、FPS、派对游戏等十余种游戏类型；专为智能体和大模型定制的高性能训练平台，可支持万卡规模资源调度；兼容主流设备的在线推理引擎，实现了移动端与云端的混合部署，保障多端协同。

GiiNEX 致力于助力开发者打造生机涌现的游戏世界。落地手机游戏场景，GiiNEX 已为玩家打造了智能教学功能，让 AI 成为玩家的私人教练。此外，UGC 关卡生成解决方案为玩家提供了一系列游戏内可用的 AI 工具，包括文生灵感图、建筑 PCG 生成、3D 模型组合生成、配色方案生成、NPC 动作生成等，助力玩家提高创作效率，丰富个性化内容。

当前，AI 大模型发展迅速，数量呈现快速上升的趋势，而大模型在手机上部署和应用仍处在早期探索阶段，包括腾讯、虎牙、酷狗、Soul 在内的应用开发者正在与手机、芯片厂商紧密配合，一方面基于手机上预装的 LLM 开发上层应用，让用户真正感受到生成式 AI 的魅力。另一方面，开发者可以深度参与到模型的研发中，开发面向特定业务场景的专属大模型，一起为生成式 AI 手机的生态共建努力。

第三章：生成式 AI 手机的软硬件科技全景

SoC/AP 是生成式 AI 手机中最重要的 AI 运算单元，在很大程度上决定本地部署的 AI 大模型的参数规模，从而决定了一款智能手机所能支持的生成式 AI 能力。除了来自硬件方面的支持，完整的工具链支持也不可或缺，这里包括软件开发工具包（SDK）、预训练的大模型等，以帮助开发人员快速创建应用场景。本章节会分别阐述生成式 AI 对智能手机硬件配置的要求，以及配套软件生态的重要性。

端侧部署 AI 大模型的硬件要求

AI 大模型的能力和性能，很大程度上是由模型包含的参数数量和质量决定的。一般说来，大模型拥有的参数越多，其精确度越高，能实现的功能越强大，生成式 AI 手机也更加智能。

智能手机的运行环境是比较苛刻的，需要考虑散热、功耗和续航、以及 PCB 布板面积等一系列制约因素。同时，为了让消费者对生成式 AI 赋能有更深切的感受，端侧运行的 LLM 参数规模也不能过低，需要支持多模态、多语言等能力。此外，消费者普遍希望端侧大模型的响应速度可以达到几乎“实时”的效果。以联发科技的天玑 9300 为例，面向 70 亿参数 LLM 的端侧推理可以做到每秒 20 tokens，良好匹配人类的平均阅读速度。

端侧 AI 算力已经成为芯片设计公司手机厂商和消费者越来越为重视的性能指标。手机 SoC 中普遍开始集成诸如 APU 或 NPU 等独立的 AI 计算单元，专门负责处理重载的 AI 任务。与此同时，芯片设计厂商正在针对 Transformer 架构重新设计这些 AI 运算单元，使其可以在 1-2 瓦的功耗预算下实现更高的 TOPS（每秒万亿次操作）/Watt，从而完成更加复杂的生成式 AI 任务，这对本地部署和运行大模型至关重要。

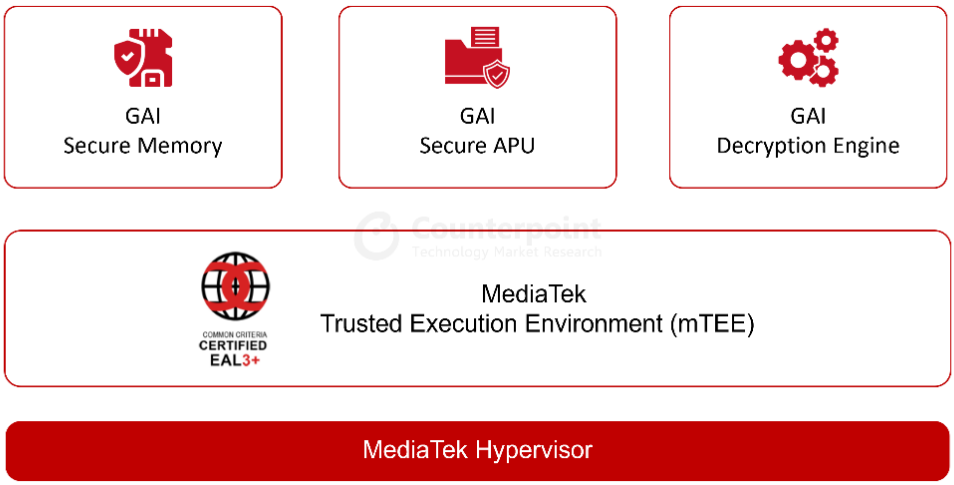
当前，手机 SoC 中集成了多个不同的处理器单元，包括 CPU、GPU、DSP 以及独立的 AI 运算单元。Counterpoint 判断，生成式 AI 手机在未来需要同时运行多个 AI 模型，包括参数量大的多模态模型，也有体量较小的影像计算和图像生成模型，系统会根据不同的任务需求调用不同的处理单元，从而实现更低的功耗和延时。基于上述判断，专用 AI 计算单元和异构计算架构将共存于手机 SoC 中，不同运算单元之间通信带宽的重要性也将日益凸显。

此外，大模型在运行时，需要驻留在内存中，每次处理生成式 AI 任务，都可能涉及到海量的数据搬运。相应的，生成式 AI 对手机的内存容量和带宽有着更高的要求，以目前主流的 70 亿参数模型为例，模型运行需要占用约 4GB 的内存空间，建议采用至少 8GB 的 LPDDR5x（推荐 60GB/s 以上的 I/O 带宽）。因此高性能的处理平台和高规格的内存，对生成式 AI 手机而言缺一不可，只有合理配置，才能提供给用户理想的体验。

其次，相较于云端部署的 AI 大模型，端侧部署的 AI 大模型由于用户隐私数据完全在手机端侧存储、计算，具有更高的安全性。但是芯片设计厂商也应提供端侧的安全解决方案，确保用户在享受生成式 AI 带来的福祉的同时，保障模型数据与用户个人的隐私安全。

针对生成式 AI，联发科技将既有的手机硬件级别的安全保护机制，升级成生成式 AI 级别的安全解决方案，提供从模型参数到用户隐私，再到运行数据的全方位保护。考虑到生成式 AI 对巨量参数存储需求，联发科技的 AI security 架构针对 Secure Memory、Secure APU、Secure Decryption Engine 进行了升级，提升了存储容量的上限，APU 运行的效率，以及解密速度，新的 AI Security 2.0 更是在现有基础上进行了大幅升级。

图表 5：联发科技生成式 AI 安全解决方案



Target (AI Security version)		AI Security 1.0	AI Security 2.0
mTEE cc EAL 3+Certificated	Secure memory allocation	V4.0 Max. 3.8 GB (0.38s for 1GB allocation)	V5.0 Max. 30 GB (0.05s for 1GB allocation)
Decryption	Gigabytes Data	**0.3 GB/s** (Big core CPU)	**3 GB/s** (CPU crypto ext.)
Secure APU	AI Performance	**15% perf.** **Drop up to 1 VM**	**5% perf.** **Drop up to 1 VM**

来源：联发科技；Counterpoint Research

最后，调用云端 AI 大模型需要智能手机拥有快速、稳定的网络连接。5G 的高速数据传输能力使得手机可以更快地上传和下载数据，这对依赖云侧大模型的 AI 应用尤为重要。用户可以更快地调用云端的 AI 大模型，运算结果也可以快速返回给用户，保证了合理的时延。

软件生态的需求

端侧部署 AI 大模型仅有硬件上的支持是远远不够的，需要软件生态的配合，帮助大模型更好适应当下智能手机硬件的实际情况，并向开发者和手机厂商提供完备的开发环境，帮助快速实现生成式 AI 应用开发和部署，促进产业健康良性发展。

首先，智能手机存储空间有限，尤其是内存空间相对紧张，需要采取包括量化（Quantization）、剪枝（Pruning）等技术来优化预训练模型的规模，使其占用更小的内存空间。量化是一种模型压缩技术，旨在减少模型的大小并加速模型的推理速度，同时尽量保持模型输出的准确性。通常 AI 大模型采用的是 32 位浮点数（FP32）类型，当前主流的解决方案是将其转化为 IN4 或 INT8 整数类型。以 130 亿参数大模型为例，借助量化技术，可以将几十 GB 大小的模型压缩至 13GB 以内，但在考虑使用 INT4 量化算法压缩模型时，需要确保目标部署平台支持此类低精度运算。同时，芯片厂商还通过采用快速译码技术，如推测解码（speculative decoding）、美杜莎（Medusa）等，进一步提升既有硬件上的模型运行速度。

除此之外，对于应用开发者而言，如果芯片厂商可以提供与底层计算平台深度适配的、完成训练的生成式 AI 大模型，应用开发周期将会被大大缩短，同时其更快的 AI 推理速度，也会为消费者带来更好的使用体验。

目前，生成式 AI 大模型，特别是 LLM 如雨后春笋般涌现，诸多国内外企业，包括谷歌、Meta、阿里巴巴、百川等，都发布了面向智能手机本地部署的生成式 AI 大模型，并且联合芯片厂商对模型进行了优化，帮助生态伙伴实现快速部署。

隐私也是消费者目前日益关心的重点，这就需要开发者也将隐私保护纳入到整体产品框架之中。如果芯片厂商可以为开发者提供完备系统级的安全解决方案，方便开发者调用，也可以帮助开发者缩短应用开发者周期，降低消费者对隐私安全的忧虑，实现开发者和芯片厂商的共赢。

考虑到生成式人工智能技术的高度复杂性，需要来自各方面的协作，共同助力产业稳健快速的发展。特别是在生成式 AI 领域，训练大型模型的高门槛不仅为大模型开发者创造了机会，同时也增加了芯片厂商、手机厂商、大模型供货商以及应用开发者之间合作的复杂性。为了更好地协调多方合作，联发科技发起了“天玑 AI 先锋计划”，邀请国内领先的手机厂商、大模型开发商和创新应用开发者加入，旨在加快生成式 AI 技术在移动端的应用落地，从而为消费者提供更加丰富功能和更好的使用体验。

目前可支持端侧 AI 大模型手机的 SoC 平台

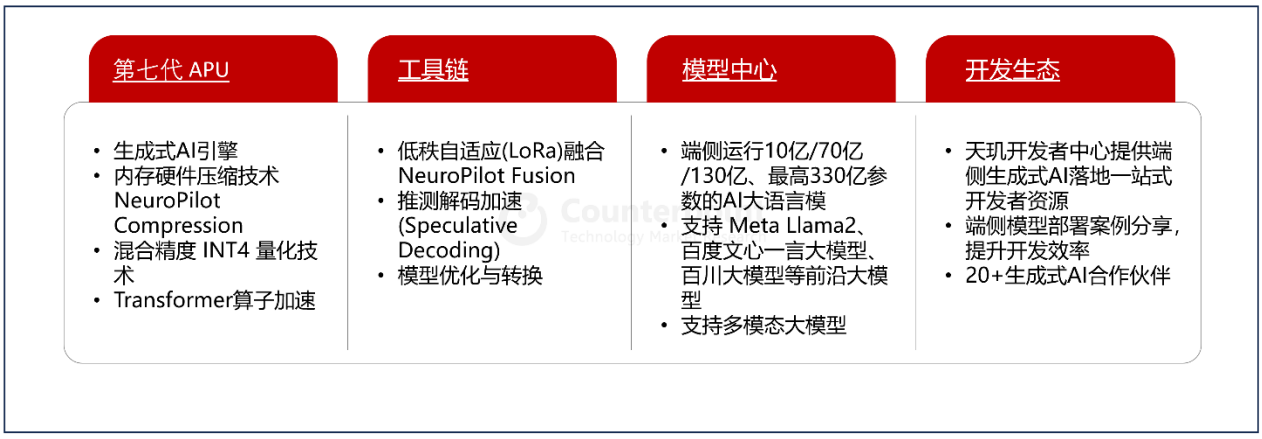
智能手机行业对生成式 AI 技术的应用的探索还处在起步阶段，并且端侧部署的生成式 AI 模型对手机规格配置提出了新的要求，这决定了端侧 AI 大模型与智能手机的融合，目前主要出现在高端手机上。作为头部芯片设计厂商，联发科技和高通分别为多模态大模型端侧部署提供了强大的移动计算平台，并完成了针对现有端侧大模型的优化工作。

联发科技于 2023 年 11 月发布的天玑 9300 旗舰手机平台集成了联发科技第七代 AI 处理器 APU 790，根据联发科技公布的数据显示，整数运算和浮点运算性能均是上一代的 2 倍，同时功耗降低 45%。此外，APU 790 内置了硬件级的生成式 AI 引擎，可实现更加高速且安全的边缘 AI 计算，相比上代，它专门针对目前 LLM 常用的 Transformer 架构进行算子加速，大模型的处理速度是上一代的 8 倍。

天玑 9300 还率先支持 LPDDR5T 技术，高达 9600Mbps 的传输速率使得数据能够快速地在内存和处理器之间传输，提高了端侧生成式 AI 模型的响应速度，也能让设备更有效地运行复杂的 AI 算法和模型。尽管传输速率提高，但 LPDDR5T 技术仍然保留了低功耗的特性。

除了硬件方面，联发科技还致力为开发者构建良好的开发环境。联发科技面向开发者推出了天玑 AI 开发套件，它是联发科技构建生成式 AI 生态的中心，现阶段涵盖了联发科技最新的第七代 APU，以及配套的工具链、不断发展的模型中心（GAI Model Hub）和开发生态。

图表 6：联发科技天玑 AI 开发套件 NeuroPilot



来源：联发科技；Counterpoint Research

其中，Neuropilot Compression 内存硬件压缩技术，可以在模型量化的基础上，进一步压缩模型体量，进一步缓解端侧部署 AI 大模型对内存带宽的压力。天玑 AI 开发套件还利用 LoRA （Low-Rank Adaptation）技术，帮助端侧大模型实现技能扩充，而不必重新部署多个模型，极大降低了开发难度和对手机存储空间的需求。而为开发者们搭建的 GAI Model Hub，里面包含了针对特定领域的、丰富的 AI 模型，开发者可以

方便、快捷地使用这些模型。依托天玑 AI 开发套件提供的一整套核心组件，可以快速、高效部署端侧生成式 AI 创新应用。

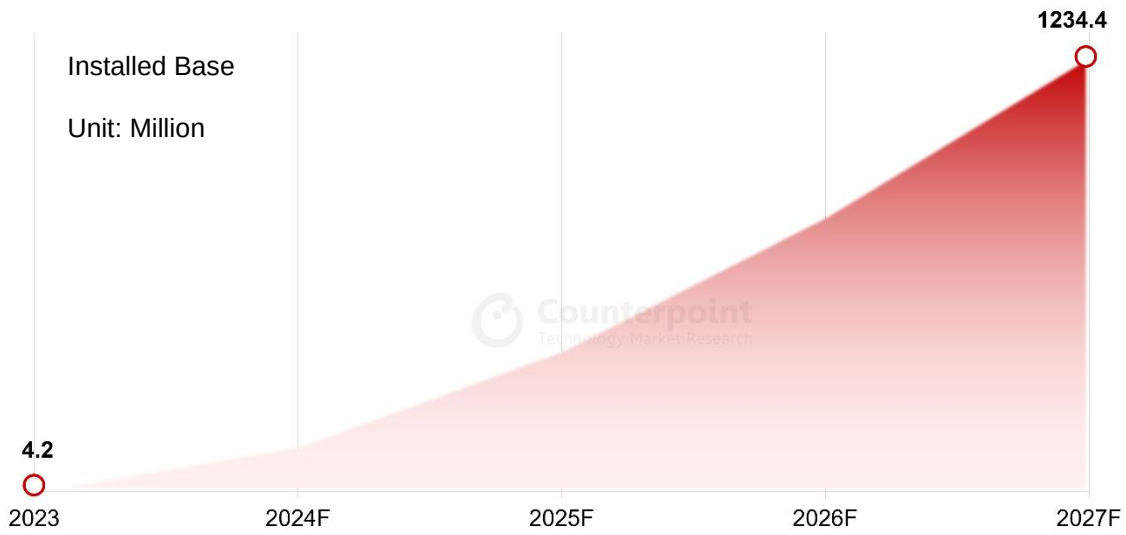
高通于同年发布的骁龙 8 Gen3 同样也针对生成式 AI 做了相应的优化。根据高通公布的数据显示，其 NPU 性能较上一代提升 98% ，在持续的 AI 推理上场景下，每瓦性能较上一代提高了 40%。骁龙 8 Gen3 支持 LPDDR5x，数据传输速率最高可达到 8500Mbps，较上一代更为适应端侧部署生成式 AI 大模型。在开发者生态方面，高通推出了高通 AI Hub，目前包含 75 个经过预训练的 AI 模型，可以快速部署在高通支持的设备上。

第四章：生成式 AI 手机预测

生成式 AI 与智能手机的结合首先从旗舰手机产品线开始。根据 Counterpoint 的数据显示，在 2023 全年出货的 11.7 亿手机中，只有不足 1%的手机满足了 Counterpoint 对生成式 AI 手机定义。但是来到 2024 年，受益于智能手机产业链上下游都在积极拥抱变革，各大手机厂商也将生成式 AI 能力作为中高端产品升级的重点，这将加速生成式 AI 手机的普及，预计在 2027 年将会达到 43%左右的渗透率。与此同时，Counterpoint 认为生成式 AI 手机存量规模将会从 2023 年的只有百万级别增长至 2027 年的 12.3 亿部。

图表 7：生成式 AI 手机总规模预测

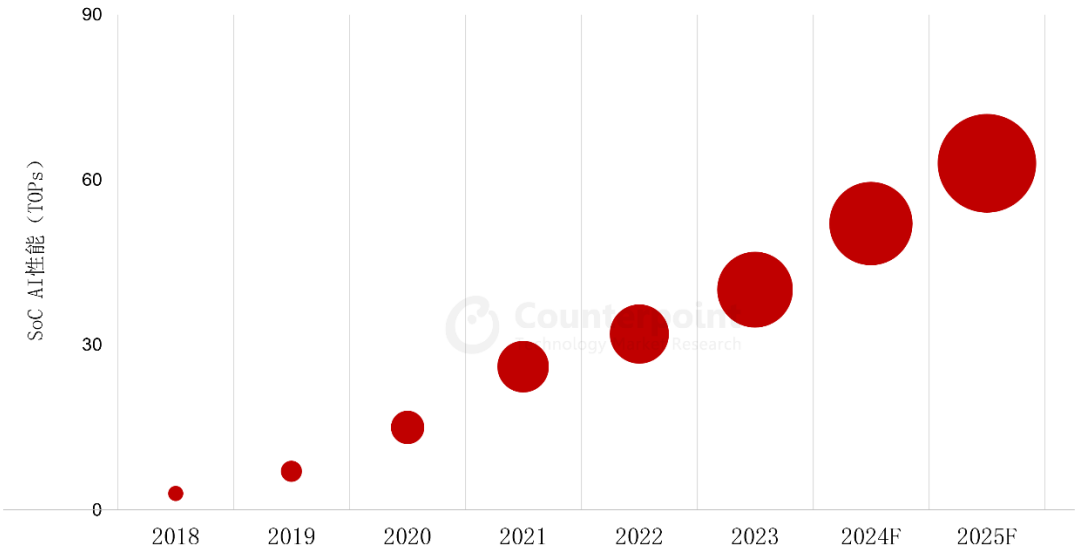
单位：百万台



来源：Counterpoint AI 360 Service

SoC 的 TOPS 性能与生成式 AI 手机的 AI 能力紧密相关。旗舰智能手机以 TOPS 为单位的 AI 算力已经增长了 20 倍，智能手机 AI 能力正变得越来越强大，而手机芯片厂商在这一转变中扮演了重要角色。Counterpoint 预计，旗舰智能手机的芯片峰值 AI 算力水平还将继续增长，在 2025 年将会达到 60TOPS 以上。

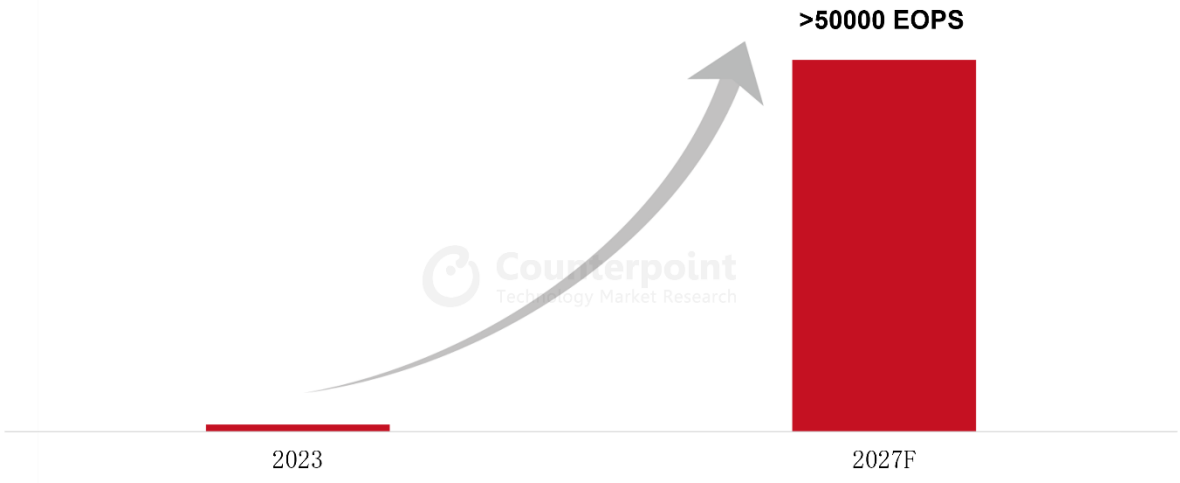
图表 8：智能手机 AI 峰值算力（TOPS）



来源：Counterpoint AI 360 Service

如上所述，生成式 AI 手机存量规模不断增长，同时每台手机所拥有的 AI 算力也在不断增加，两相作用下，全球智能手机的总体 AI 算力将会呈现爆发性增长态势。根据 Counterpoint 的估算，2027 年生成式 AI 手机端侧整体 AI 算力将会达到 50000EOPS 以上。在未来，生成式 AI 手机将会成为不可忽视的 AI 计算资源池，这也印证了端云结合部署模式的可行性、经济性和必要性。

图表 9：生成式 AI 手机端侧总 AI 计算资源



来源：Counterpoint AI 360 Service

结论

生成式 AI 手机开启了手机智能化演进的新周期，在芯片厂商、手机厂商、大模型厂商以及广大开发者的共同努力下，生成式 AI 手机将为用户提供全新的交互体验、多模态内容生成能力、个性化的服务能力，以及革新的应用生态。受益于端侧 AI 算力和大模型能力的持续提升，生成式 AI 技术与智能手机的融合将不断深入，生成式 AI 手机将发展为移动智能体，带来新的想象空间，助力智能手机产业和移动应用生态的持续繁荣，更好的迎接 6G 时代的到来。

作者、版权、用户协议和其他一般信息

Tarun Pathak
Research Director
tarun@counterpointresearch.com

Ethan Qi
Associate Director
ethan@counterpointresearch.com

Archie Zhang
Research Analyst
archie.zhang@counterpointresearch.com

COUNTERPOINT TECHNOLOGY MARKET RESEARCH
Hong Kong | USA | South Korea | India | UK | Argentina | China
info@counterpointresearch.com