Qualcomm

2023 年 5 月

# 混合 AI 是 AI 的未来

## 第一部分：
## 终端侧 AI 和混合 AI
## 开启生成式 AI 的未来

# 目录

# 1 摘要

混合 AI 是 AI 的未来。随着生成式 AI 正以前所未有的速度发展[1]以及计算需求的日益增长[2]，AI 处理必须分布在云端和终端进行，才能实现 AI 的规模化扩展并发挥其最大潜能——正如传统计算从大型主机和瘦客户端演变为当前云端和边缘终端相结合的模式。与仅在云端进行处理不同，混合 AI 架构在云端和边缘终端之间分配并协调 AI 工作负载。云端和边缘终端如智能手机、汽车、个人电脑和物联网终端协同工作，能够实现更强大、更高效且高度优化的 AI。

节省成本是主要推动因素。举例来说，据估计，每一次基于生成式 AI 的网络搜索查询（query），其成本是传统搜索的 10 倍[3]，而这只是众多生成式 AI 的应用之一。混合 AI 将支持生成式 AI 开发者和提供商利用边缘终端的计算能力降低成本。混合 AI 架构或终端侧 AI 能够在全球范围带来高性能、个性化、隐私和安全等优势。

混合 AI 架构可以根据模型和查询需求的复杂度等因素，选择不同方式在云端和终端侧之间分配处理负载。例如，如果模型大小、提示（prompt）和生成长度小于某个限定值，并且能够提供可接受的精确度，推理即可完全在终端侧进行。如果是更复杂的任务，模型则可以跨云端和终端运行。混合 AI 还能支持模型在终端侧和云端同时运行，也就是在终端侧运行轻量版模型时，在云端并行处理完整模型的多个标记（token），并在需要时更正终端侧的处理结果。

随着强大的生成式 AI 模型不断缩小，以及终端侧处理能力的持续提升，混合 AI 的潜力将会进一步增长。参数超过 10 亿的 AI 模型已经能够在手机上运行，且性能和精确度水平达到与云端相似的水平。不久的将来，拥有 100 亿或更高参数的模型将能够在终端上运行。

混合 AI 方式适用于几乎所有生成式 AI 应用和终端领域，包括手机、笔记本电脑、XR 头显、汽车和物联网。这一方式对推动生成式 AI 规模化扩展，满足全球企业与消费者需求至关重要。

---

[1] https://www.statista.com/chart/29174/time-to-one-million-users/
[2] https://siliconangle.com/2023/02/05/generative-ai-drives-explosion-compute-looming-need-sustainable-ai/
[3] https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/

# 2 生成式 AI 简介和当前趋势

ChatGPT 激发了人们的想象力和好奇心。自 2022 年 11 月推出后，短短两个月内其月活用户便达到 1 亿，成为有史以来增长速度最快的消费类应用和第一个杀手级的生成式 AI 应用。随着创新节奏的加快，想要紧跟生成式 AI 的发展速度，难度越来越大。大型聚合网站的数据显示，目前已有超过 3,000 个可用的生成式 AI 应用和特性[4]。AI 正迎来大爆发时期，就像此前电视、互联网和智能手机的问世，而这仅仅是一个开始。

ChatGPT 和 Stable Diffusion 等生成式 AI 模型能够基于简单的提示创作出全新的原创内容，如文本、图像、视频、音频或其他数据。这类模型正在颠覆传统的搜索、内容创作和推荐系统的方法——通过从普通产业到创意产业的跨行业用例，在实用性、生产力和娱乐性方面带来显著增强。建筑师和艺术家可以探索新思路，工程师可以更高效地编写程序。几乎所有与文字、图像、视频和自动化相关的工作领域都将受益。

网络搜索是生成式 AI 正在变革的诸多应用之一。另一个例子则是 Microsoft 365 Copilot，作为一项全新的生产力特性，它能够利用生成式 AI 帮助编写和总结文档、分析数据，或将简单的书面想法转化为演示文稿，嵌入于 Word、Excel、PowerPoint、Outlook 和 Teams 等微软应用中。

生成式 AI 的出现也标志着用户开始向探索更加多样化、个性化的数字世界迈出了第一步。由于 3D 设计师可以借助生成式 AI 工具更加快速高效地进行内容开发，3D 内容创作有望得到普及。这不仅将加速沉浸式虚拟体验的创建，而且能够降低个人创作者自主内容制作的门槛。

我们即将看到从生成式 AI 中涌现出各种各样的全新企业级和消费级用例，带来超越想象的功能。GPT-4 和 LaMDA 等通用大语言模型（LLM）作为基础模型，所具备的语言理解、生成能力和知识范畴已达到了前所未有的水平。这些模型大多数都非常庞大，参数超过 1 千亿，并通过 API 向客户提供免费或付费服务。

基础模型的使用推动大量初创公司和大型组织利用文本、图像、视频、3D、语言和音频创建应用。例如，代码生成（GitHub Copilot）、文本生成（Jasper）、面向艺术家和设计师的图像生成（Midjourney），以及对话式聊天机器人（Character.ai）。

[4] 截至 2023 年 4 月，生成式 AI 应用和特性：https://theresanaiforthat.com/

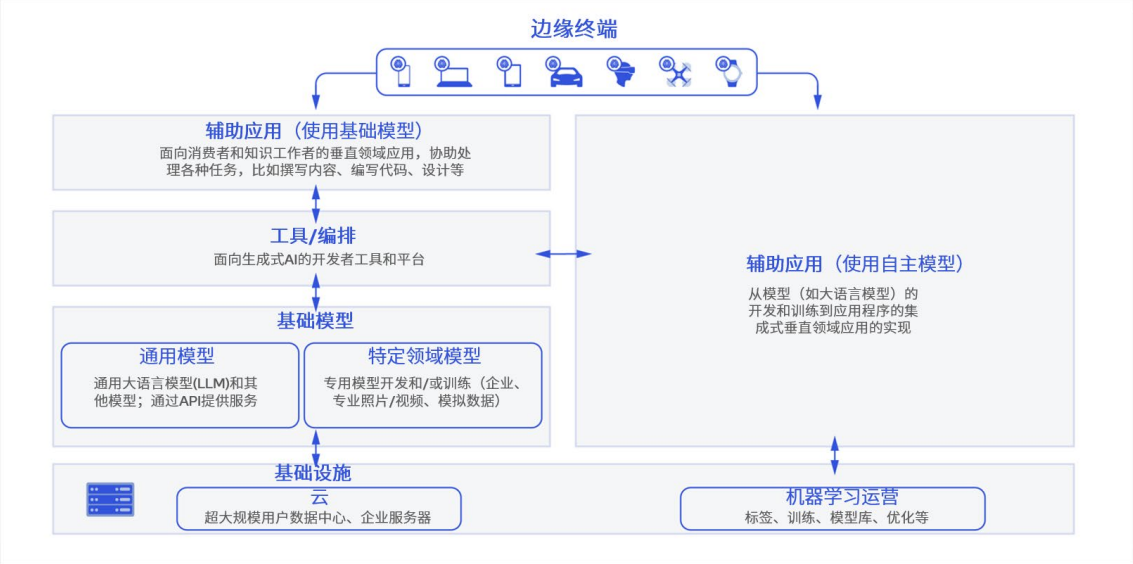


*图1：生成式 AI 生态链使应用数量激增*

据初步估计显示，生成式 AI 市场规模将达到 1 万亿美元 [5]，广泛覆盖生态链的各个参与方。为把握这一巨大机遇，并推动 AI 成为主流，计算架构需要不断演进并满足大规模生成式 AI 日益增长的处理和性能需求。

## 3 混合 AI 对生成式 AI 规模化扩展至关重要

拥有数十亿参数的众多生成式 AI 模型对计算基础设施提出了极高的需求。因此，无论是为 AI 模型优化参数的 AI 训练，还是执行该模型的 AI 推理，至今都一直受限于大型复杂模型而在云端部署。

AI 推理的规模远高于 AI 训练。尽管训练单个模型会消耗大量资源，但大型生成式 AI 模型预计每年仅需训练几次。然而，这些模型的推理成本将随着日活用户数量及其使用频率的增加而增加。在云端进行推理的成本极高，这将导致规模化扩展难以持续。

混合 AI 能够解决上述问题，正如传统计算从大型主机和瘦客户端演变为当前云端和 PC、智能手机等边缘终端相结合的模式。

[5] 瑞银，2023 年 2 月

## 3.1 什么是混合 AI?

混合 AI 指终端和云端协同工作，在适当的场景和时间下分配 AI 计算的工作负载，以提供更好的体验，并高效利用资源。在一些场景下，计算将主要以终端为中心，在必要时向云端分流任务。而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些 AI 工作负载。

## 3.2 混合 AI 的优势

混合 AI 架构（或仅在终端侧运行 AI），能够在全球范围带来成本、能耗、性能、隐私、安全和个性化优势。

### 3.2.1 成本

随着生成式 AI 模型使用量和复杂性的不断增长，仅在云端进行推理并不划算。因为数据中心基础设施成本，包括硬件、场地、能耗、运营、额外带宽和网络传输的成本将持续增加。

例如，当前面向大语言模型推理的云计算架构，将导致无论规模大小的搜索引擎企业负担更高运营成本。试想一下，未来通过生成式 AI 大语言模型增强的互联网搜索，比如 GPT，其运行参数远超 1750 亿。生成式 AI 搜索可以提供更加出色的用户体验和搜索结果，但每一次搜索查询（query）其成本是传统搜索方法的 10 倍。目前每天有超过 100 亿次的搜索查询产生，即便基于大语言模型的搜索仅占其中一小部分，每年增量成本也可能达到数十亿美元。[6]

将一些处理从云端转移到边缘终端，可以减轻云基础设施的压力并减少开支。这使混合 AI 对生成式 AI 的持续规模化扩展变得至关重要。混合 AI 能够利用现已部署的、具备 AI 能力的数十亿边缘终端，以及未来还将具备更高处理能力的数十亿终端。

节省成本也是生成式 AI 生态系统发展的重要一环，可以支持 OEM 厂商、独立软件开发商（ISV）和应用开发者更经济实惠地探索和打造应用。例如，开发者可以基于完全在终端上运行的 Stable Diffusion 创建应用程序，对于生成的每个图像承担更低的查询成本，或完全没有成本。

### 3.2.2 能耗

支持高效 AI 处理的边缘终端能够提供领先的能效，尤其是与云端相比。边缘终端能够以很低的能耗运行生成式 AI 模型，尤其是将处理和数据传输相结合时。这一能耗成本差异非常明显，同时能帮助云服务提供商降低数据中心的能耗，实现环境和可持续发展目标。

---

[6] 摩根士丹利，《How Large are the Incremental AI Costs...and 4 Factors to Watch Next》，2023 年 2 月

### 3.2.3 可靠性、性能和时延

在混合 AI 架构中，终端侧 AI 处理十分可靠，能够在云服务器和网络连接拥堵时，提供媲美云端甚至更佳的性能 [7]。当生成式 AI 查询对于云的需求达到高峰期时，会产生大量排队等待和高时延，甚至可能出现拒绝服务的情况 [8]。向边缘终端转移计算负载可防止这一现象发生。此外，混合 AI 架构中终端侧处理的可用性优势，让用户无论身处何地，甚至在无连接的情况下，依然能够正常运行生成式 AI 应用。

### 3.2.4 隐私和安全

终端侧 AI 从本质上有助于保护用户隐私，因为查询和个人信息完全保留在终端上。对于企业和工作场所等场景中使用的生成式 AI，这有助于解决保护公司保密信息的难题。例如，用于代码生成的编程助手应用可以在终端上运行，不向云端暴露保密信息，从而消除如今众多企业面临的顾虑 [9]。对于消费者使用而言，混合 AI 架构中的“隐私模式”让用户能够充分利用终端侧 AI 向聊天机器人输入敏感提示，比如健康问题或创业想法。此外，终端侧安全能力已经十分强大，并且将不断演进，确保个人数据和模型参数在边缘终端上的安全。

### 3.2.5 个性化

混合 AI 让更加个性化的体验成为可能。数字助手将能够在不牺牲隐私的情况下，根据用户的表情、喜好和个性进行定制。所形成的用户画像能够从实际行为、价值观、痛点、需求、顾虑和问题等方面来体现一个用户，并且可以随着时间推移进行学习和演进。它可以用于增强和打造定制化的生成式 AI 提示，然后在终端侧或云端进行处理。用户画像保留在终端内，因此可以通过终端侧学习不断优化和更新。

个性化不仅仅适用于消费者，企业或机构可以借助它标准化代码的编写方式，或者制作具有特殊语气和声音的公共内容。

---

[7] https://www.qualcomm.com/news/onq/2023/02/worlds-first-on-device-demonstration-of-stable-diffusion-on-android

[8] https://www.digitaltrends.com/computing/chatgpt-is-at-capacity-and-is-frustrating-new-people-everywhere/

[9] https://www.pcmag.com/news/samsung-software-engineers-busted-for-pasting-proprietary-code-into-chatgpt

## 3.3 AI 工作负载的分布式处理机制

我们期望打造能够支持不同工作负载分流方式的混合 AI 架构，可以根据模型和查询复杂度进行分布式处理，并能持续演进。例如，如果模型大小、提示和生成长度小于某个限定值，并且能够提供可接受的精确度，推理即可完全在终端侧进行。如果是更复杂的任务，模型则可以跨云端和终端运行；如果需要更多最新信息，那么也可以连接至互联网获取。

### 3.3.1 以终端为中心的混合 AI

在以终端为中心的混合 AI 架构中，终端将充当锚点，云端仅用于分流处理终端无法充分执行的任务。许多生成式 AI 模型可以在终端上充分运行（参阅图 2），也就是说终端可通过运行不太复杂的推理完成大部分处理工作。

例如，用户在笔记本电脑上运行 Microsoft 365 Copilot 或必应 Chat 时，包含高达数百亿参数的模型将在终端上运行，而更复杂的模型将根据需求在云端进行处理。对用户来说，这种体验是无缝的，因为终端侧神经网络或基于规则而运行的判决器（arbiter）将决定是否需要使用云端，无论是为了有机会使用更好的模型还是检索互联网信息。如果用户对请求处理结果的质量不满意，那么再次尝试发起请求时可能就会引入一个更好的模型。由于终端侧 AI 处理能力随着终端升级和芯片迭代不断提升，它可以分流更多云端的负载。


*图 2：在以终端为中心的混合 AI 架构中，云端仅用于分流处理终端无法充分运行的 AI 任务。*

对于各种生成式 AI 应用，比如创作图像或起草邮件，快速响应式的推理更受青睐，即使它在准确度上会稍有损失。终端侧 AI 的快速反馈（即低时延）可以让用户使用改进的提示来快速迭代推理过程，直至获得满意的输出结果。

### 3.3.2 基于终端感知的混合 AI

在基于终端感知的混合 AI 场景中，在边缘侧运行的模型将充当云端大语言模型（类似大脑）的传感器输入端（类似眼睛和耳朵）。例如，当用户对智能手机说话时，Whisper 等自动语音识别（ASR）的 AI 模型将在终端侧运行，将语音转为文字，然后将其作为请求提示发送到云端。云端将运行大语言模型，再将生成的文本回复发回终端。之后，终端将运行文本生成语音（TTS）模型，提供自然免提回答。将自动语音识别和文本生成语音模型工作负载转移至终端侧能够节省计算和连接带宽。随着大语言模型变为多模态并支持图像输入，计算机视觉处理也可以在终端上运行，以进一步分流计算任务并减少连接带宽，从而节省成本。

在更先进的版本中，隐私将得到进一步保护，终端侧 AI 能够承担更多处理，并向云端提供经过改进且更加个性化的提示。借助终端侧学习和终端上的个人数据，比如社交媒体、电子邮件、消息、日历和位置等，终端将创建用户的个人画像，与编排器（orchestrator）程序协作，基于更多情境信息提供更完善的提示。例如，如果用户让手机来安排与好友会面的时间并在喜爱的餐厅预订座位，编排器程序了解上述个性化信息并能够向云端大语言模型提供更佳提示。编排器程序可在大语言模型缺乏信息时设置护栏并帮助防止产生“AI 幻觉”。对于较简单的请求，较小的大语言模型可在终端侧运行，而无需与云端交互，这类似于以终端为中心的混合 AI。

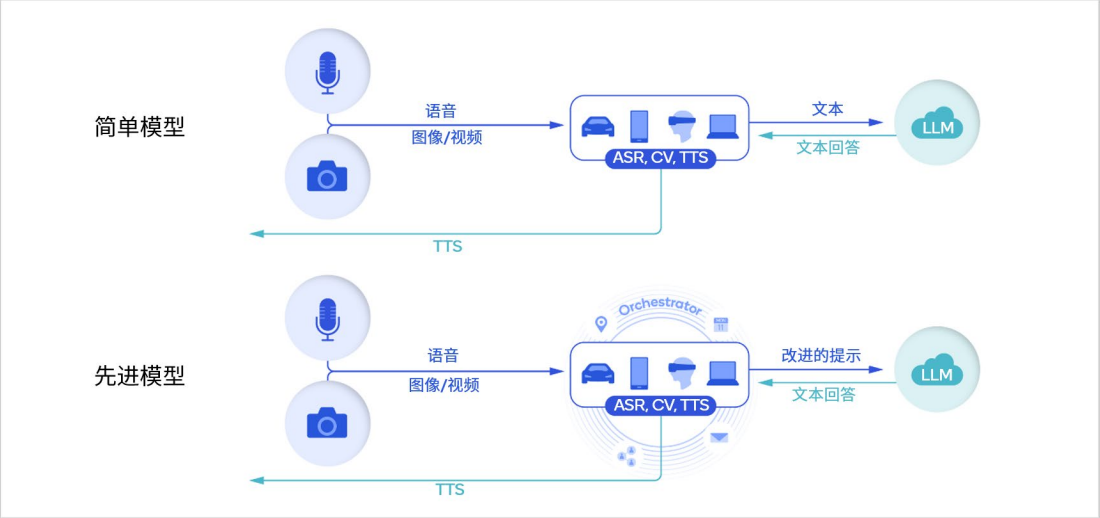


*图3：对于基于终端感知的混合 AI，自动语音识别、计算机视觉和文本转语音在终端侧进行。在更先进的版本中，终端侧编排器程序能够向云端提供经过改进且更加个性化的提示。*

### 3.3.3 终端与云端协同处理的混合 AI

终端和云端的 AI 计算也可以协同工作来处理 AI 负载，生成大语言模型的多个 token 就是一个例子。大语言模型的运行都是内存受限的，这意味着计算硬件在等待来自 DRAM 的内存数据时经常处于闲置状态。大语言模型每次推理生成一个 token，也就是基本等同于一个单词，这意味着 GPT-3 等模型必须读取全部 1750 亿参数才能生成一个单词，然后再次运行整个模型来生成下一个 token，完整的推理过程可以以此类推。鉴于内存读取是造成推理性能的瓶颈因素，更高效的做法就是同时运行多个大语言模型以生成多个 token，并且从 DRAM 一次性读取全部参数。每生成一个 token 就要读取全部参数会产生能耗和造成发热，因此使用闲置的算力通过共享参数来推测性并行运行大语言模型，可谓是在性能和能耗上实现双赢。

为了生成四个 token，一个近似的大语言模型（比原始目标大语言模型小 7 至 10 倍，因此准确性更低）要在终端上按顺序连续运行四次才可以。终端向云端发送这四个 token，云端高效运行四次目标模型来检查其准确度，而仅读取一次完整的模型参数。在云端 token 是被并行计算的，每个目标模型都有零个、一个、两个、三个或四个预测 token 作为输入。这些 token 在被云端确认或校正之前被认为是“近似的”。上述推测性解码过程将持续到完整的答案出现时为止。我们的早期实验和其他已发布结果 [10]显示，通过四个 token 的推测性解码，平均两到三个 token 是正确可被接受的，这会带来单位时间内生成 token 数的增加，并节省能耗。

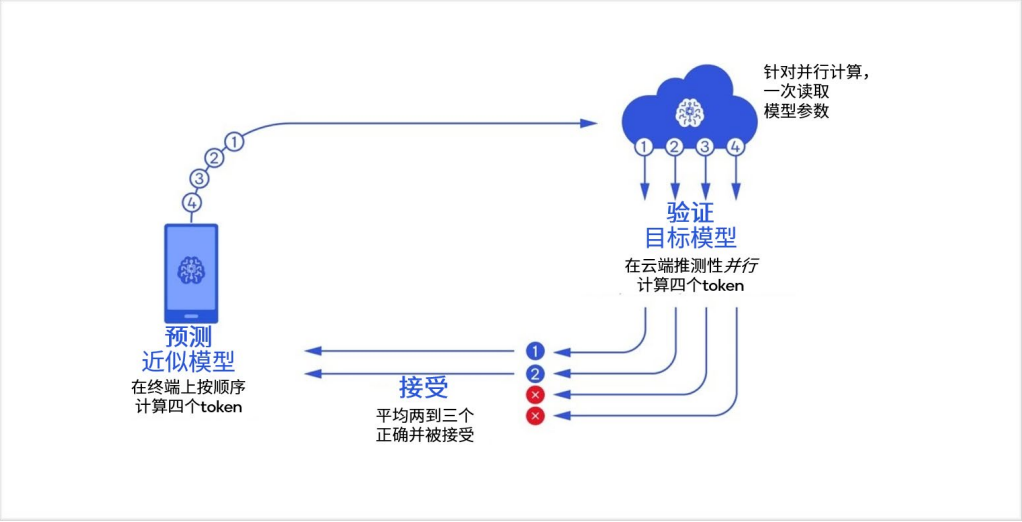


*图 4：协同处理混合 AI 的四个 token 推测性解码示例。*

[10] Leviathan, Yaniv, Matan Kalman 和 Yossi Matias。《Fast Inference from Transformers via Speculative Decoding》。 *arXiv preprint arXiv:2211.17192* (2022)

# 4 终端侧 AI 的演进与生成式 AI 的需求密切相关

终端侧 AI 能力是赋能混合 AI 并让生成式 AI 实现全球规模化扩展的关键。如何在云端和边缘终端之间分配处理任务将取决于终端能力、隐私和安全需求、性能需求以及商业模式等诸多因素（参阅第 3.3 章节）。

在生成式 AI 出现之前，AI 处理便持续向边缘转移，越来越多的 AI 推理工作负载在手机、笔记本电脑、XR 头显、汽车和其他边缘终端上运行。例如，手机利用终端侧 AI 支持许多日常功能，比如暗光拍摄、降噪和人脸解锁。

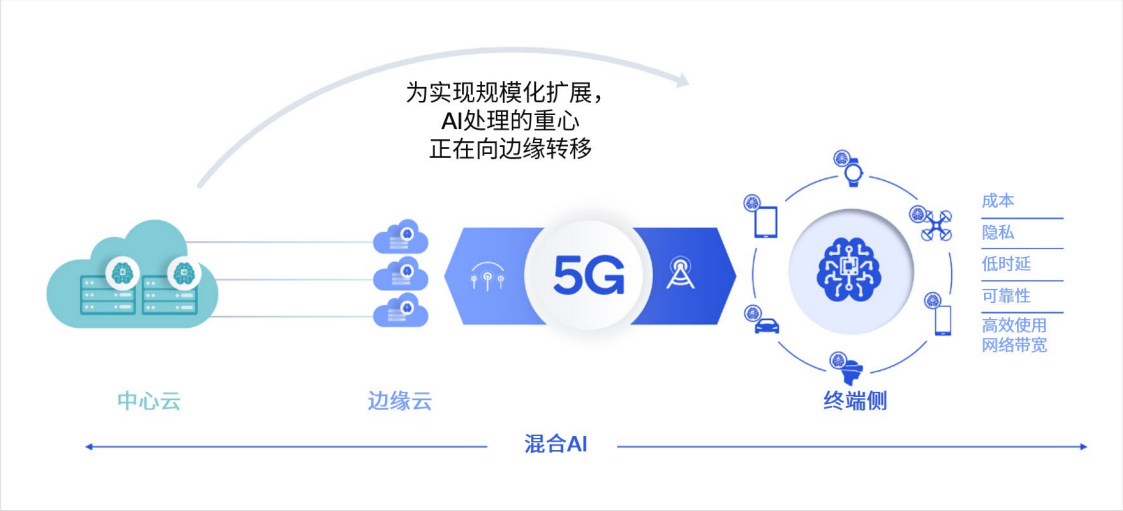


*图 5：AI 处理的重心正在向边缘转移。*

## 4.1 终端侧处理能够支持多样化的生成式 AI 模型

如今，具备 AI 功能的手机、PC 和其他品类的便携终端数量已达到数十亿台 [11]，利用大规模终端侧 AI 处理支持生成式 AI 有着广阔前景，并且将在未来几年稳步增长。

关键问题在于，哪些生成式 AI 模型能够以合适的性能和准确度在终端侧运行。好消息是，性能十分强大的生成式 AI 模型正在变小，同时终端侧处理能力正在持续提升。图 6 展示了可以在终端侧运行的丰富的生成式 AI 功能，这些功能的模型参数在 10 亿至 100 亿之间 [12]。如 Stable Diffusion 等参数超过 10 亿的模型已经能够在手机上运行，且性能和精确度达到与云端处理类似的水平。不久的将来，拥有 100 亿或更多参数的生成式 AI 模型将能够在终端上运行。

[11] https://www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai

[12] 假设使用 INT4 型的参数

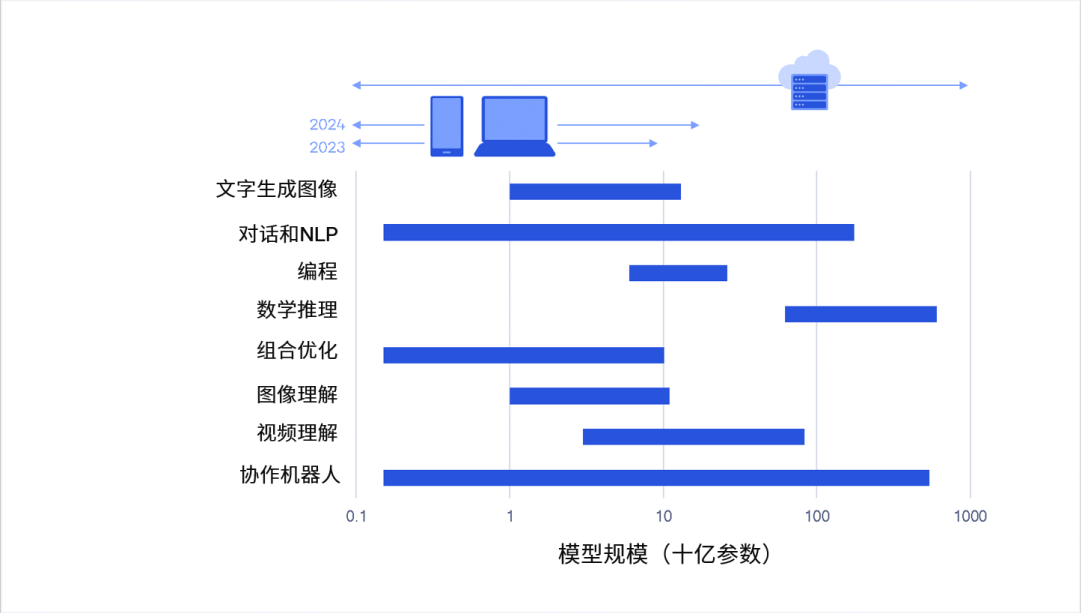


*图 6：数量可观的生成式 AI 模型可从云端分流到终端上运行。*

# 5 跨终端品类的生成式 AI 关键用例

基于基础模型的生成式 AI 迅速兴起，正在驱动新一轮内容生成、搜索和生产力相关用例的发展，覆盖包括智能手机、笔记本电脑和 PC、汽车、XR 以及物联网等终端品类。混合 AI 架构将赋能生成式 AI 在上述这些终端领域提供全新的增强用户体验。

## 5.1 智能手机：搜索和数字助手

面对每日超过 100 亿次的搜索量且移动端搜索占比超过 60%的情况[13]，生成式 AI 的应用将推动所需算力的实质性增长，尤其是来自智能手机端的搜索请求。由于基于生成式 AI 的查询能够提供更令人满意的答案，用户的搜索方式已经开始发生转变。

对话式搜索的普及也将增加总体查询量。随着对话功能不断改进，变得更加强大，智能手机将成为真正的数字助手。精准的终端侧用户画像与能够理解文字、语音、图像、视频和任何其他输入模态的大语言模型相结合，让用户可以自然地沟通，获取准确、贴切的回答。进行自然语言处理、图像理解、视频理解、文本生成文本等任务的模型将面临高需求。

[13] https://www.statista.com/statistics/297137/mobile-share-of-us-organic-search-engine-visits/

## 5.2 笔记本电脑和 PC：生产力

生成式 AI 基于简单提示就能快速生成优质内容，它也正在凭借这项能力变革生产力。以笔记本电脑和 PC 上的 Microsoft Office 365 为例，全球有超过 4 亿 Microsoft Office 365 商业付费席位和个人订阅者，如果将生成式 AI 集成至用户日常工作流将带来重大影响 [14]。此前需要数小时或数天的任务，现在仅需几分钟就能完成。Microsoft 365 Copilot 同时利用大语言模型的功能和 Microsoft Graph 与 Microsoft 365 应用中的用户数据，能够将提示转化为强大的生产力工具 [15]。

Office 工作者可通过后台运行大语言模型，在 Outlook 中阅读或撰写电子邮件，在 Word 中编写文档，在 PowerPoint 中创建演示文稿，在 Excel 中分析数据，或在 Teams 会议中协作。生成式 AI 模型（比如自然语言处理、文本生成文本、图像生成、视频生成和编程）需要经过海量处理，才能支持这些被重度使用的生产力任务。在以终端为中心的混合 AI 架构中，大部分处理能够在 PC 上进行。

## 5.3 汽车：数字助手和自动驾驶

得益于车内和车辆周围环境相关数据所提供的信息，如今 AI 驱动的座舱能够提供高度个性化的体验。类似于智能手机和 PC，车载数字助手将能够让驾乘人员通过免提的友好用户界面保持无缝互联，同时为生态系统创造全新的创收机会。

数字助手可以访问用户个人数据，比如应用、服务和支付信息；以及来自车辆的传感器数据，包括摄像头、雷达、激光雷达和蜂窝车联网（C-V2X）等。企业 API 也支持第三方服务提供商集成他们的解决方案，将客户关系延伸到车上。例如，主动式驾驶辅助将大幅改善导航体验，比如会影响驾驶员常用出行路线的交通和天气信息更新，汽车充电或购买停车券提醒，此外，用户可以通过简单地请求即可用已绑定的信用卡预订自己喜欢的美食。如果汽车能够识别每位驾乘人员并提供定制化的音乐和播客等体验和内容，座舱的媒体娱乐体验也将会变革。随着车载 AR 应用变得更加普遍，数字助手可以按照驾乘人员的偏好提供定制化的显示。

汽车维修保养和服务也将变得更加自主和无缝。通过分析传感器输入、维修保养历史和驾驶行为等数据，数字助手可以预测何时需要进行保养。利用生成式 AI，数字助手可针对汽车如何维修提供信息，或为用户提供咨询，找到合适的服务提供商，提高车辆可靠性，同时减少时间和成本。

---

[14] 微软财报

[15] https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

感知软件栈从未遇到过的罕见或陌生物体，经常会对高级驾驶辅助系统和自动驾驶（ADAS/AD）解决方案产生干扰。这种情况通常由光线不佳或恶劣天气条件造成，会导致驾驶策略软件栈产生难以预测、有时甚至很危险的结果。为了在未来预防类似情况，必须妥善采集和标记这些极端场景的数据并重新训练模型。这个循环可能耗时费力，而生成式 AI 可以模拟极端场景，预测不同道路行为主体的轨迹和行为，比如车辆、行人、自行车骑行者和摩托车骑行者。规划者可以利用这些场景确定车辆驾驶策略。

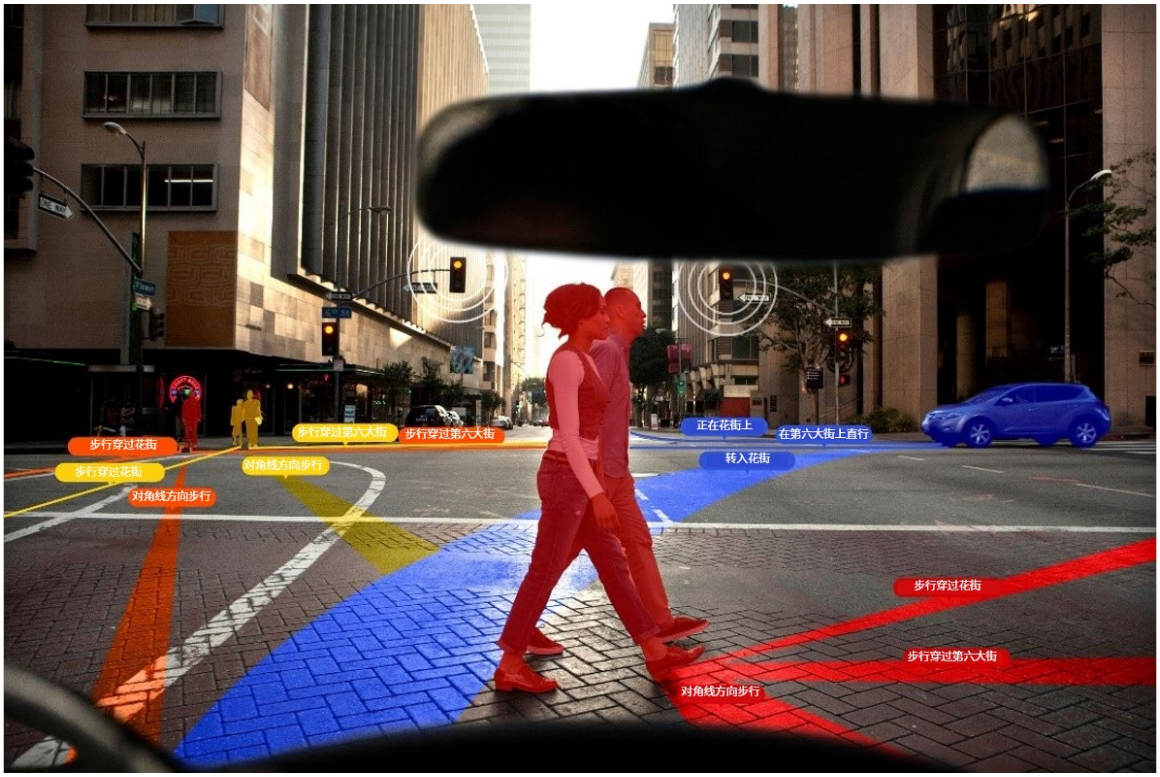


*图7：生成式AI 可用于先进驾驶辅助系统/自动驾驶（ADAS/AD），通过预测不同行为主体的轨迹和行为，帮助改进驾驶策略。*

驾驶策略软件栈以及感知软件栈始终在汽车的 AI 算力可支持的情况下本地运行。严苛的时延要求决定了云端无法针对这些 AI 工作负载在决策过程中发挥任何作用。随着 ADAS/AD 解决方案采用支持适当后处理的生成式 AI 模型，汽车必然需要具备显著高能效的 AI 计算能力。

## 5.4 XR：3D 内容创作和沉浸式体验

生成式 AI 能为 XR 带来巨大前景。它有潜力普及 3D 内容创作，并真正实现虚拟化身。下一代 AI 渲染工具将赋能内容创作者使用如文本、语音、图像或视频等各种类型的提示，生成 3D 物体和场景，并最终创造出完整的虚拟世界。此外，内容创作者将能够利用文本生成文本的大语言模型，

为能够发出声音并表达情绪的虚拟化身生成类人对话。总而言之，这些进步将变革用户在 XR 设备上创造和体验沉浸式内容的方式。

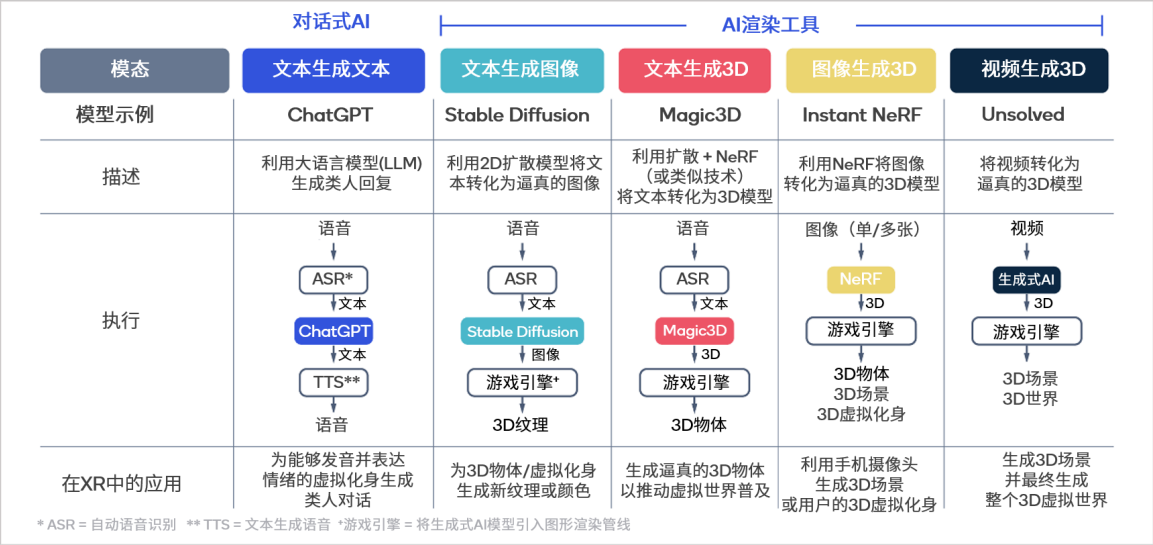


*图 8：生成式 AI 模型将面向 XR 赋能对话式 AI 和全新渲染工具。*

生成式 AI 为 XR 提供的前景无疑令人兴奋，但很难预测这些技术何时才能被广泛采用。不过，根据近几个月快速的创新步伐，可以肯定地说，我们可以期待在未来几年内取得重要进展。

对于沉浸式世界，Stable Diffusion 等文本生成图像类的模型很快将赋能内容创作者在 3D 物体上生成逼真的纹理。我们预计，一年内这些功能将在智能手机上实现，并延伸到 XR 终端。XR 中的部署需要“分布式处理”，即头显运行感知和渲染软件栈，与之配对的智能手机或云端运行生成式 AI 模型。未来几年，首批文本生成 3D 和图像生成 3D 类的模型将可能实现边缘侧部署，生成高质量的 3D 物体点云。几年后，这些模型将通过提升，达到能够从零开始生成高质量 3D 纹理物体的水平。在大约十年内，模型将更进一步，支持由文本或图像生成的高保真完整 3D 空间和场景。未来，文本生成 3D 和视频生成 3D 类的模型最终或能让用户踏入从零开始生成的 3D 虚拟世界，例如自动构建满足用户任何想象的 3D 虚拟环境。


*图9：生成式AI将有助于基于简单提示创造沉浸式3D虚拟世界的过程，比如"超现实世界、水母四处游动、美丽的瀑布、神秘的湖泊、巍峨的高山"*

虚拟化身将遵循类似的发展过程。文本生成文本的模型，比如有130亿参数的LLaMA，将运行在边缘终端，为虚拟化身生成自然直观的对话。此外，文本生成图像的模型将为这些虚拟化身生成全新的纹理和服装。未来几年内，图像生成3D和编/解码器模型将能够为人类生成全身虚拟化身，支持远程通信。最终，人们将能够利用语音提示、图像或视频生成逼真、全动画、智能、可量产的类人虚拟化身。

## 5.5 物联网：运营效率和客户支持

目前，AI已广泛应用于各种物联网垂直领域，包括零售、安全、能源和公共设施、供应链和资产管理。AI依靠近乎实时的数据采集和分析改进决策质量，优化运营效率，并赋能创新以打造差异化竞争优势。通过生成式AI，物联网细分领域将进一步从AI的应用中受益。

以零售业为例，生成式AI可以改善顾客和员工体验。在售货亭或智能购物车旁的导购员可以基于每周特价商品、预算限制和家庭偏好帮助顾客定制带有菜谱的菜单。商店经理可以根据即将发生的事件预测非周期性的促销机会并进行相应准备。如果一个运动队来到其所在的城市，那么商店经理可以利用生成式AI查询粉丝喜爱的商品品牌，并相应地增加库存。另一个用途是参考来自相似社区的商店的优秀案例和成功经验，重新进行店面规划。生成式AI可以利用简单提示帮助商店

经理重新排列货架商品，为利润高的产品腾出空间，或者利用附近连锁店的数据，尽可能降低产品缺货情况的发生。

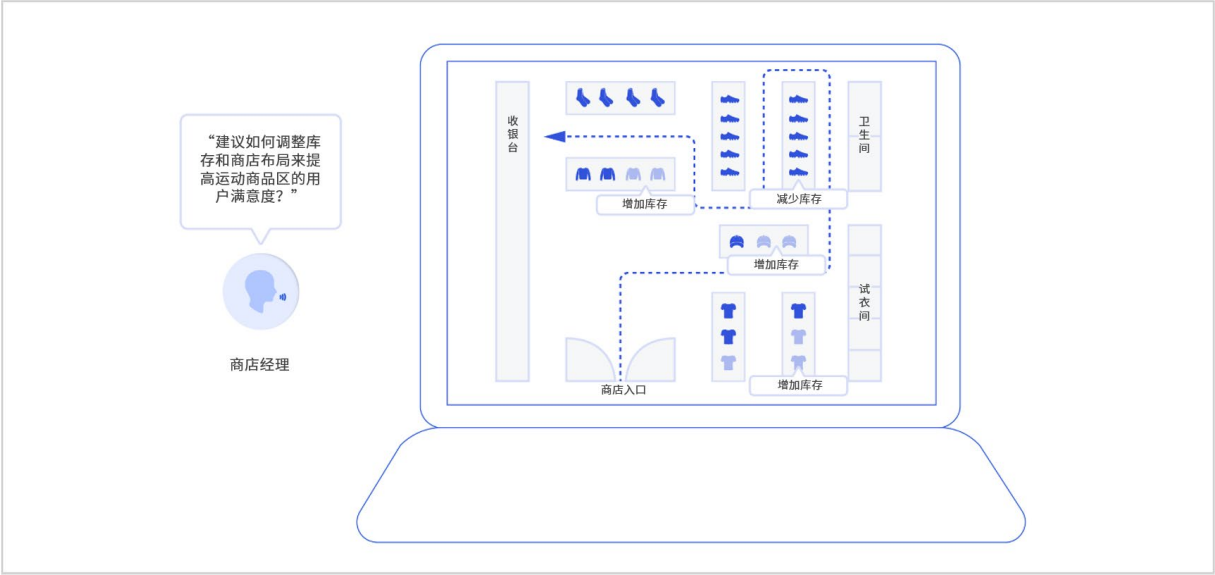


*图10：以零售业为例，生成式 AI 有助于提升顾客和员工体验，比如提供库存和商店布局推荐。*

能源和公共设施领域也将受益于生成式 AI。运营团队可以创建极端负荷场景并预测电力需求，以及特殊情况下潜在的电网故障，比如农村地区在炎热的夏季出现强风和局部火灾的情况，从而更好地管理资源、避免电力中断。生成式 AI 也可以用于提供更好的客户服务，比如解答断电或账单计费问题。

## 6 总结

混合 AI 势不可挡。生成式 AI 用例将持续演进并成为主流体验，云端和其基础设施需求将不断增加。凭借终端侧 AI 的先进能力，混合 AI 架构将规模化扩展，以满足企业和消费者的需求，带来成本、能耗、性能、隐私、安全和个性化的优势。云端和终端将协同工作，依托强大、高效且高度优化的 AI 能力打造下一代用户体验。

Qualcomm

请关注我们：

欲了解更多信息，请访问

qualcomm.com