国金证券 SINOLINK SECURITIES

2024 年 11 月 27 日

2025 AI 行业前瞻报告

行业深度研究(深度)

证券研究报告

国金证券研究所
分析师：刘道明（执业 S1130520020004） 联系人：黄晓军（执业 S1130122050092） 联系人：麦世学（执业 S1130123100111）
liudaoming@gjzq.com.cn huangxiaojun@gjzq.com.cn maishixue@gjzq.com.cn

AI 行业关键时刻：瓶颈与机遇并存

报告摘要：

2025 年，AI 将迎来模型与应用的双向奔赴：

- **模型侧**，模型将朝大小模型互补的方向演进，聚焦增强推理能力以突破当前的 Scaling Law 瓶颈。大型预训练市场逐渐收敛，由 OpenAI、Meta 的 Llama、Mistral、阿里通义等主导，更多中小厂商则专注于特定任务的微调与 Agent 业务。新兴技术路径如测试时训练、合成数据应用及感知量化训练将推动模型能力提升，而多模态融合模型在实时交互、音频与视觉生成领域展现出巨大潜力。
- **应用侧**，渗透率持续快速上升，ChatGPT 活跃度持续攀升，视频生成模型如 Runway 和可灵国际版表现稳定。我们持续看好如下应用方向：1）AI 程序员在企业中得到广泛应用，显著提升开发效率；2）数据重要性大幅提升推动 SaaS 平台如 Snowflake、Datadog 和 Databricks 等业务高速增长；3）通用 SaaS 平台如 ServiceNow 和 Salesforce 受益于大企业 AI 投入增加；4）AI 搜索有望在 2025 年诞生超级 APP；5）AI 眼镜作为综合体验最好的 AI 硬件新形态，预计将在 2025 年迎来大规模出货。
- **算力系统**，虽然英伟达最新的 Blackwell 架构算力芯片仍在云端具备绝对统治力，但是随着系统复杂性的快速提升以及核心技术及零部件供给瓶颈，硬件迭代速度可能在未来趋缓。这将给 AMD 等竞争对手以及云厂商自研芯片带来更多的发展机会。
- **电力基础设施**，随着单数据中心规模的不断扩大，局部供电压力激增。独立于传统居民/工业电网的核电站成为潜在最优解决方案。美国几大云厂亚马逊、谷歌、微软都在积极寻求核电解决方案。核电的落地速度成为制约 AI 进一步发展的重要因素。
- **端侧 AI，随着模型小型化趋势及应用场景的快速丰富，我们预计端侧 AI 在 2025 年也将迎来大发展**。在硬件、软件、生态、云等所有环节都可控并有所参与的手机厂商更容易成功，其中苹果、谷歌更为完整。苹果在硬件、软件、生态环境、云服务上具备极强竞争力。谷歌有原生安卓支持、Gemini 强大的模型能力，但在硬件上自有品牌 Pixel 渗透率低，更多需要依赖三星端侧硬件拓展用户。
- **AI PC 领域**：1）未来 X86 笔电市场竞争将会更为激烈，英特尔和 AMD 产品在性能、续航、适配性、生态方面各有千秋。2）X86 台式机领域，由于功耗的重要性大幅降低，AMD 有望依靠更出色的 CPU 性能提升市占率；3）AIPC 的渗透，重点看 ARM 芯片。苹果的优势最明显，高通 X Elite 短时间内很难与苹果竞争 ARM 架构 AI 笔电的市场。未来英伟达&联发科合作研发的处理器也会带来更多看点。ARM 架构芯片的成熟有望推动 Windows 操作系统向更适合 AI 的方向进化。

风险提示

芯片制程发展与良率不及预期

中美科技领域政策恶化

智能手机、PC 销量不及预期

内容目录

一、AI 模型趋势：大小模型互补，预训练市场快速收敛，Scaling Law 新方向增强推理需求

2024 年大模型厂商推出模型的速度仍在加快，大模型与小模型共存仍是解决模型能力上限和端侧推理的方案，各大模型厂商也会推出几 B 到 TB 级别的模型。随着大型模型训练成本的不断提升，且有更多像 Meta、Mixtral、阿里通义等公司的开源，模型预训练市场的玩家会快速缩小，针对特定任务的微调或者是 Agent 业务将会是更多中小模型厂商发展的重点。在当前算力和数据 Scaling Law 放缓的情况下，找到新的 Scaling Law 方向是明年模型发展的重点。

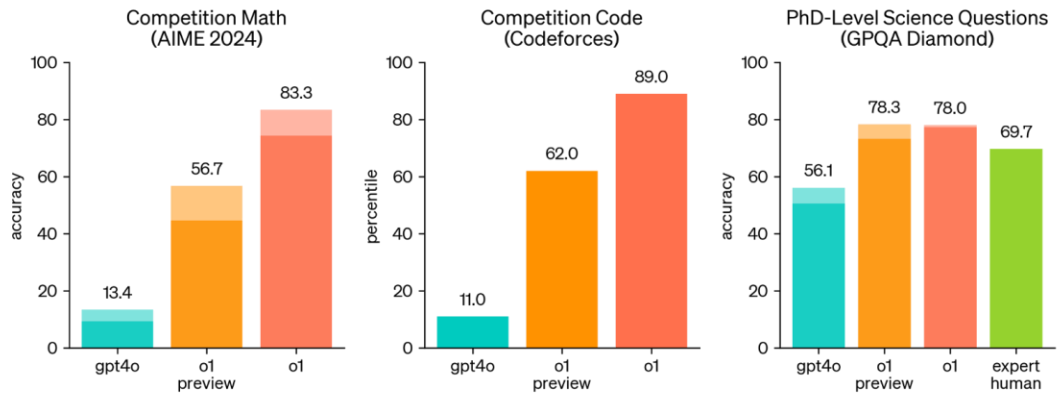
1.1 预训练和现实数据触顶，后训练时代将开启新的 Scaling Law 方向

从 24 年年初开始有论文提出模型能力提升速度随着参数规模的扩大而放缓，到 11 月份 OpenAI 前首席科学家 Ilya 在公开场合表示简单地增加数据和计算能力来扩大当前模型规模的时代已经结束。但是，预训练的 scaling law 放缓不代表大模型发展速度和算力需求的放缓，就像是芯片 gate 的实际尺寸停滞在 20nm 并不影响等效 gate 密度达到目前的 3nm，广义的摩尔定律甚至比 20 年前更快，大模型也需要找到具有更高的投入回报比的新方向。

1.2 方向一：用推理代替思考

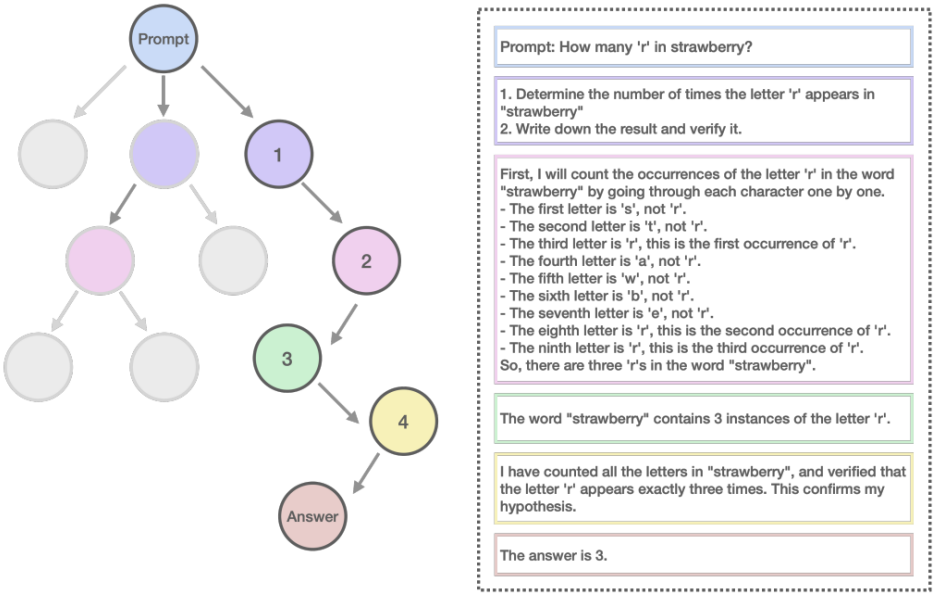
OpenAI 于 2024 年 9 月 12 日发布了新的 AI 模型系列 o1，这是 OpenAI 首个具有"逻辑推理"能力的模型系列，特别擅长处理复杂的推理任务，尤其是在科学、技术、工程和数学(STEM)领域的问题，在这些领域其评测分数都远远超过 GPT-4o。o1 模型将计算资源从大规模预训练数据集重新分配到训练和推理阶段，增强了复杂推理能力，在费用和成本上也进行了重分配，使用 o1-preview 的 API 相比于 GPT-4o 输入 tokens 价格是 GPT-4o 的 5 倍（每百万 tokens $15.00：$3.00），输出 tokens 差距 o1-preview 的价格是 GPT-4o 的 6 倍（每百万 tokens $60.00：$10.00）。

图表1：OpenAI o1 模型测评分数对比



来源：OpenAI、国金证券研究所

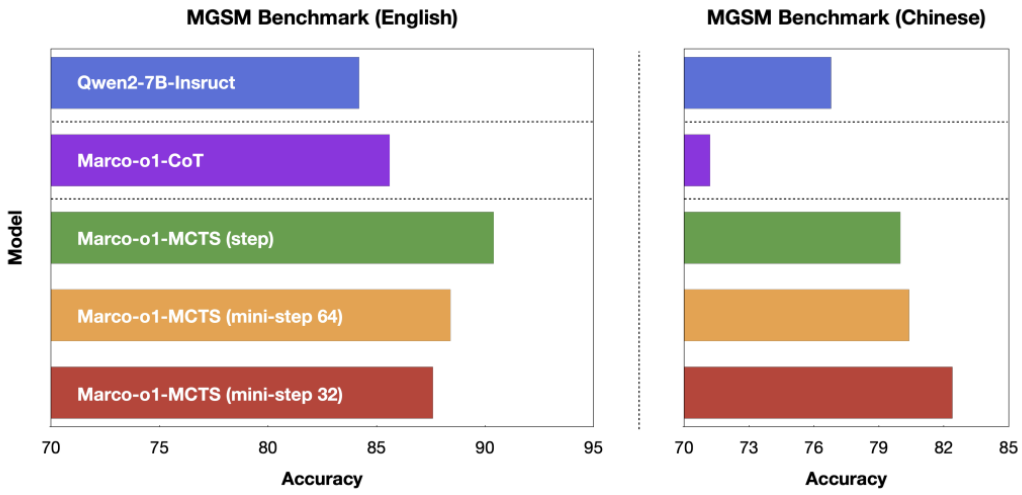
图表2：阿里通义 Macro-o1 模型的原理



来源：Macro-o1 论文、国金证券研究所

在 OpenAI 发布 o1 之后，其他大模型公司包括国内的 Deepseek 和阿里通义也发布了类似通过增强推理阶段的计算资源来提高能力的模型，并且开始有论文揭露底层技术。阿里发布的 Marco-o1 由思维链（CoT）微调、蒙特卡洛树搜索（MCTS）、自反机制和创新性推理策略驱动，专门针对复杂的现实世界问题解决任务进行了优化。同时，阿里在 Open-o1 数据集的基础上进行了筛选，并且使用合成数据方法合成了一个新的 Macro-o1 数据库，用来监督微调。最终，在应用了蒙特卡洛树微调后，模型在评测上实现了大幅超过了基底模型 Qwen2-7B 的成绩。

图表3：阿里通义 Macro-o1 模型测试成绩大幅领先基底模型



来源：Macro-o1 论文、国金证券研究所

Deepseek 也推出了一款名为 DeepSeek-R1，对标 OpenAI 的 o1 模型，同样是采用"思维链"技术，可以将复杂任务分解为多个步骤逐一解决，在 AIME 和 MATH 两项基准测试中，R1 的表现与 o1 相当或更优，但是仍未公布论文和技术详细信息。

图表4：DeepSeek-R1在复杂问题测试成绩与其他模型对比

	DeepSeek-R1-Lite-Preview	OpenAI o1-preview	GPT-4o	Claude 3.5 Sonnet	Qwen-2.5-72B-Instruct	DeepSeek V2.5
AIME (pass@1) 美国数学竞赛	**52.5**	44.6	9.3	16.0	23.3	16.7
MATH-500 (greedy) 美国数学竞赛	**91.6**	85.5	76.6	78.3	83.1	74.7
GPQA Diamond (pass@1) 理工科博士生测试	58.5	**73.3**	53.6	65.0	49.0	41.3
Codeforces (Rating) 世界级编程竞赛	**1450**	1428	759	717	732	882
LiveCodeBench (2024.8-2024.11) 世界级编程竞赛	51.6	**53.6**	33.4	36.3	31.1	29.2
Zebra Logic 自然语言解谜	56.6	**71.4**	28.2	33.4	26.6	22.1

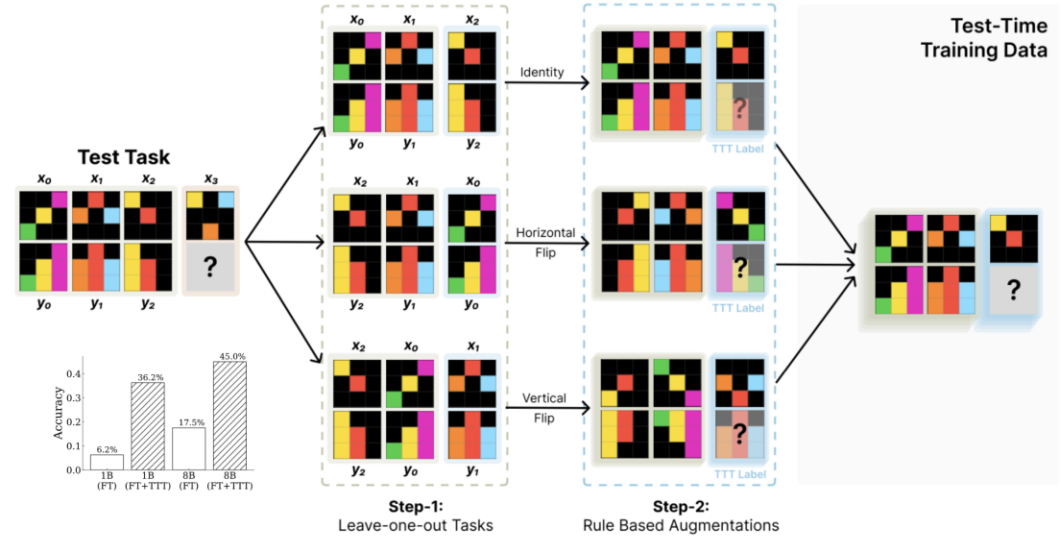
*所有测评在最大推理长度 32K 下得到，测试结果通过测试集重复测试多次求平均得到，避免温度带来的随机影响。

来源：DeepSeek 官网、国金证券研究所

1.3 方向二：测试时训练

测试时训练（Test-Time Training）是 24 年 11 月份由 MIT 提出的另一条实现大模型 Scaling Law 的路线，这是一种在推理过程中根据测试输入动态更新模型参数的技术。它不同于标准的微调，因为它在极低数据的情况下运行，通常对单个输入或一两个上下文中的标记示例使用无监督或监督目标。相当于对推理过程中的数据进行调整后合成测试时训练数据用来更新模型的参数，这种方法对抽象推理的问题效果较好，MIT 团队在 Llama3 8B 模型上使用这种方法后，相比于 1B 的基础微调模型，准确率提高了 6 倍；在 8B 参数的语言模型上应用 TTT，在 ARC 公共验证集上实现了 45% 的准确率，比 8B 基础模型提高了近 157%。但是该方法仍在初期试验阶段，对计算资源要求也很高，所以论文的评估主要在 ARC 公共验证集的一个子集上进行，并没有提交到官方排行榜。

图表5：测试时训练（TTT）合成数据的原理

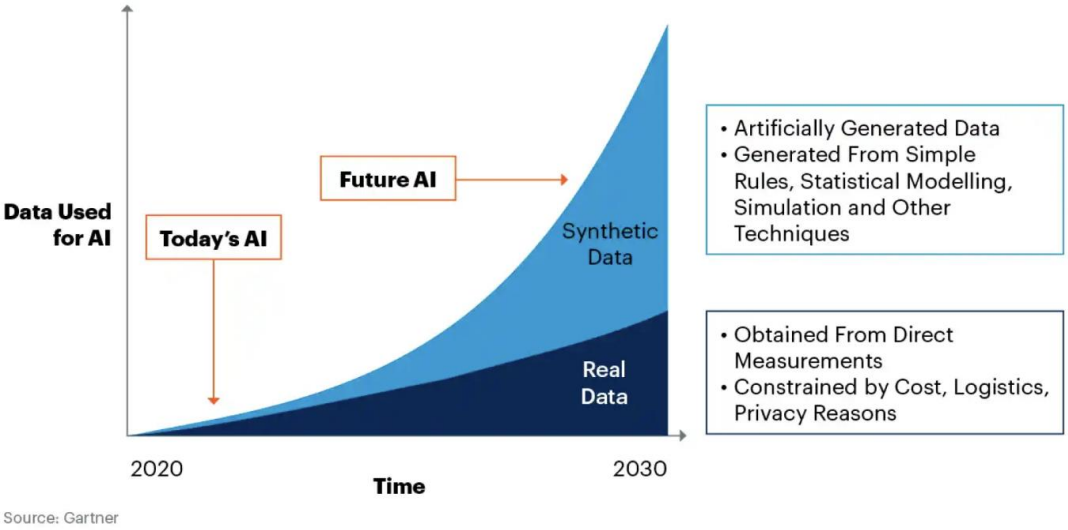


来源：Test-Time Training 论文、国金证券研究所

1.4 方向三：合成数据

合成数据在 LLM 开发中的应用正在迅速扩大，从预训练到微调阶段都发挥着重要作用。它不仅解决了数据获取和隐私问题，还能有针对性地增强模型在特定任务上的表现。OpenAI 的模型训练和 Alignment 项目大量使用合成数据；Anthropic 公司在 Claude 系列模型中采用了 Constitutional AI（CAI）方法，通过合成数据显著提升了模型的稳健性，使得 Claude 模型能够更准确地识别和拒绝回答不确定的问题；阿里通义的 Qwen 系列则采取了一种独特的方法，利用早期版本的 Qwen 模型来生成合成数据，用于增强预训练数据集的质量，同时在训练过程中创新性地使用合成数据生成多个候选响应，再通过奖励模型筛选出最优答案；Apple 的 AFM 模型也在这一领域做出了重要尝试，特别是在预训练阶段使用合成数据来延长上下文长度，并且特别关注数学和代码任务相关的高质量合成数据生成。

图表6：大模型训练中合成数据占比不断提升



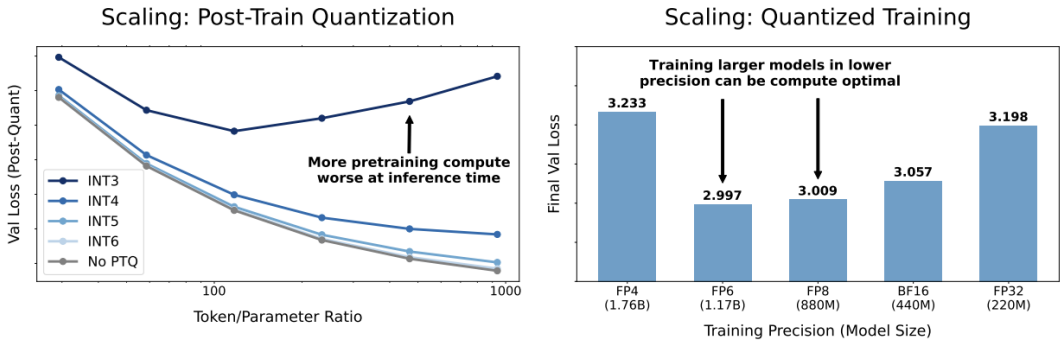
来源：Gartner、国金证券研究所

据 Gartner 预测，到 2030 年，合成数据将在 AI 模型中完全超过真实数据的使用，而合成数据的生成过程需要消耗大量计算资源。以使用 OpenAI 的模型为例，使用 GPT-4 生成十万个 JSON 合成数据元素预计成本高达 506 美元，随着现实世界数据被逐渐发掘用尽，合成数据消耗的推理资源会快速上升。

1.5 方向四：模型量化逐渐失效

量化是把模型里的数字用更少的位数表示，比如用整数代替小数，这样计算更快，占用的空间也更小。在模型推理时使用量化后的模型是主流的节约推理成本的方法，但是在 24 年 11 月，哈佛和斯坦福大学等顶尖学府学者发布的《Scaling Laws for Precision》引起了大模型行业科学家的广泛讨论，研究发现在预训练阶段使用更低精度的参数会会降低模型的有效参数数量，而推理量化后的模型的性能下降会随着模型训练数据量的增加而增加，意味着数据太多反而对推理低精度模型有负面影响。论文还提出了感知量化训练技术，是一种有效的模型量化技术，模型仍然使用高精度（例如 FP32 或 BF16）进行训练，但在每次前向和反向传播过程中，都会模拟低精度量化的操作，感知到降低哪些部分的参数精度对模型效果的影响较小，可以在保持较高推理性能的同时降低模型的计算和存储成本。

图表7：训练后量化和训练时量化效果对比



来源：Scaling Laws for Precision、国金证券研究所

1.6 方向五：多模态融合模型发展空间大

尽管各大厂商如 Meta 和阿里巴巴积极布局多模态大模型领域，分别推出了 Llama 3.2 系列（包括其首个大型多模态模型）以及通义 Qwen-VL 升级版（Qwen-VL-Plus 和 Qwen-VL-Max），在图像推理等能力上取得了显著进展，但在整体架构设计、性能效果以及支持的模态数量等方面，相较于 OpenAI 推出的 GPT-4o 仍存在明显差距，比如 Llama 3.2 仍然是将音频模型叠加到大语言模型上获得的多模态能力，而 GPT-4o 具备的以下能力仍然是多模态模型的标杆：

1. 多模态理解与生成：支持文本、图像、音频、视频理解，文本、图像、音频生成
2. 统一模型：使用单一神经网络处理所有模态，而非多个独立模型的管道
3. 端到端训练：跨文本、视觉和音频进行端到端的联合训练
4. 实时交互：音频输入响应时间平均为 320 毫秒，接近人类对话反应速度，支持近实时的语音对话和翻译

图表8：GPT-4o 多模态能力展示，实时逐步指导用户解答习题

来源：OpenAI、国金证券研究所

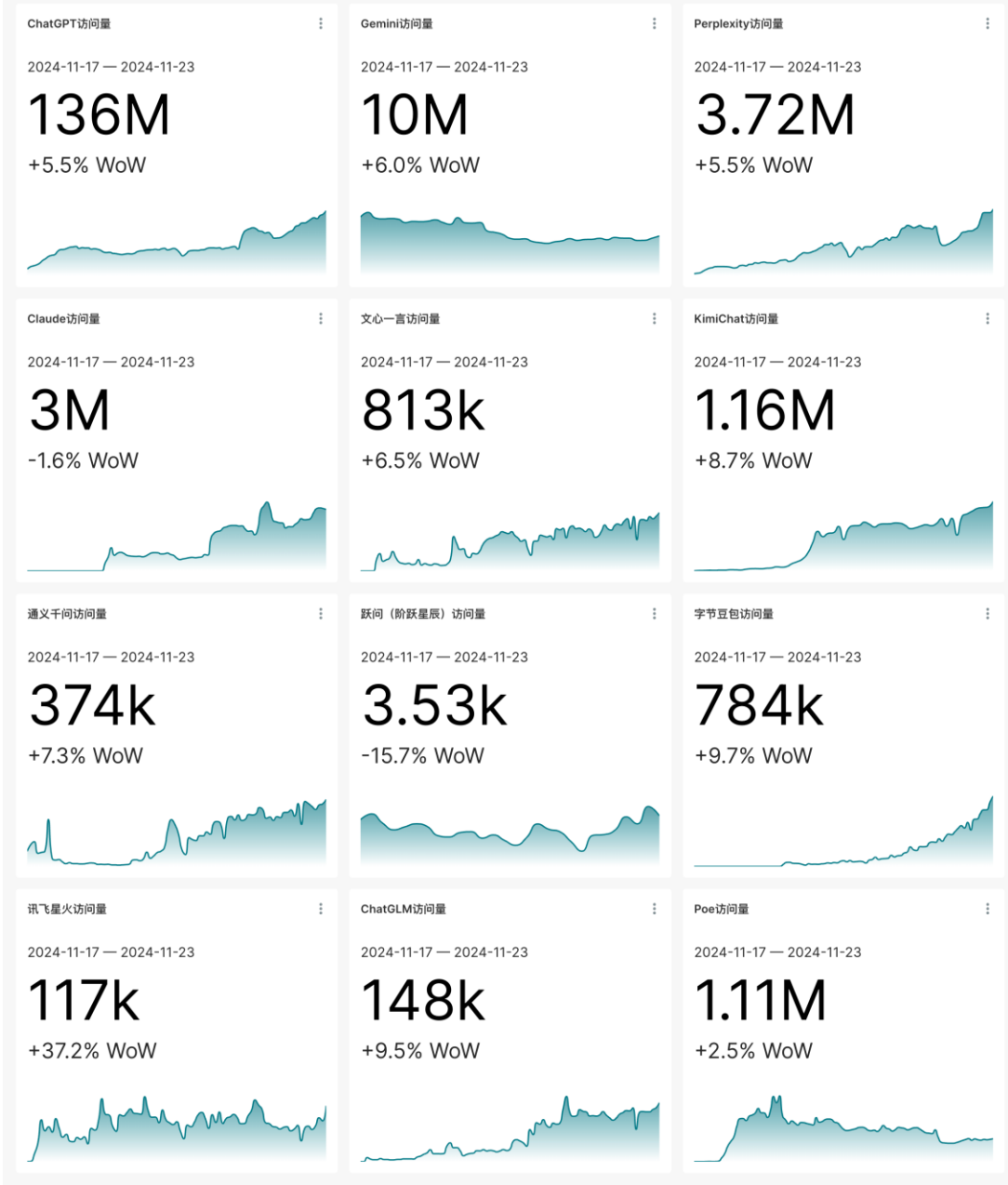
多模态模型的发展可以给予 AI 应用和 AI 硬件的落地更大的想象空间，比如可以根据上下文来理解和生成不同语气语调的音频；使用语音直接进行图片编辑；在 AI 硬件上直接进行实时对话，将看到或者听到的内容实时翻译成另一种语言；实时逐步的对眼前的题目进行解答等。

二、AI 应用渗透率持续增长，落地场景多点开花

2.1 AI 应用活跃度持续增长，应用场景得到认可，进入快速获客期

从 AI 应用的日活跃度数据看，ChatGPT 活跃度持续增长，其他 AI 聊天助手应用也保持增长态势，AI 应用渗透率不断提升。从国内市场看，头部应用如 Kimi、文心一言、通义千问、豆包等的活跃度也在不断提高，AI 聊天助手应用场景得到用户认可，进入快速获客期。

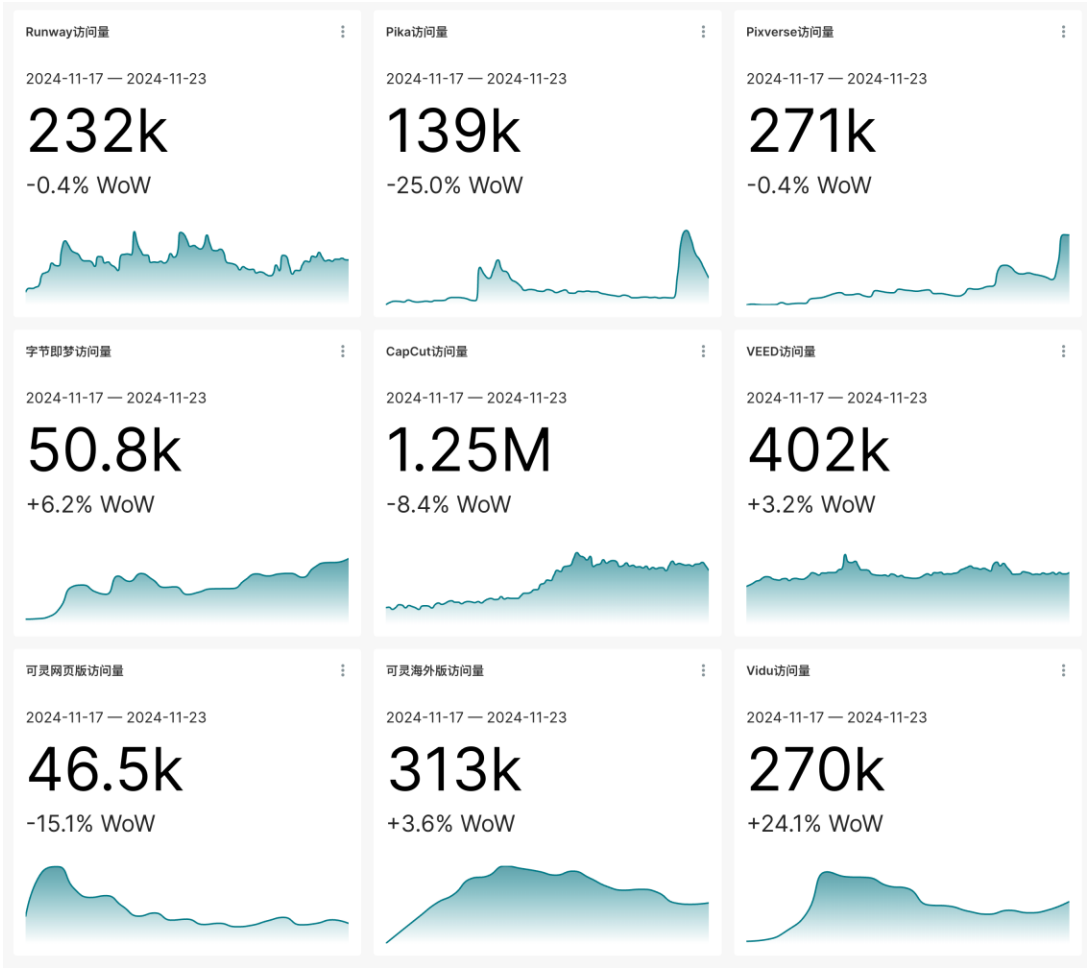
图表9：聊天助手类应用周均日活变化



来源：SimilarWeb、国金证券研究所

视频模型在快速发展阶段，闭源模型如 Runway 和可灵的活跃度较为稳定，新发模型对应用活跃度仍然有较大的提升。快手的可灵国际版实现了 AI 模型出海，属于现在可用模型中在海外的评价较高的视频生成模型。开源的视频模型也在出现，包括 Meta 的 Movie Gen 和 Mochi 1。视频模型对算力需求的提升符合我们的预期，比如未量化版本的 Mochi 需要 4 个 H100 才能进行推理。

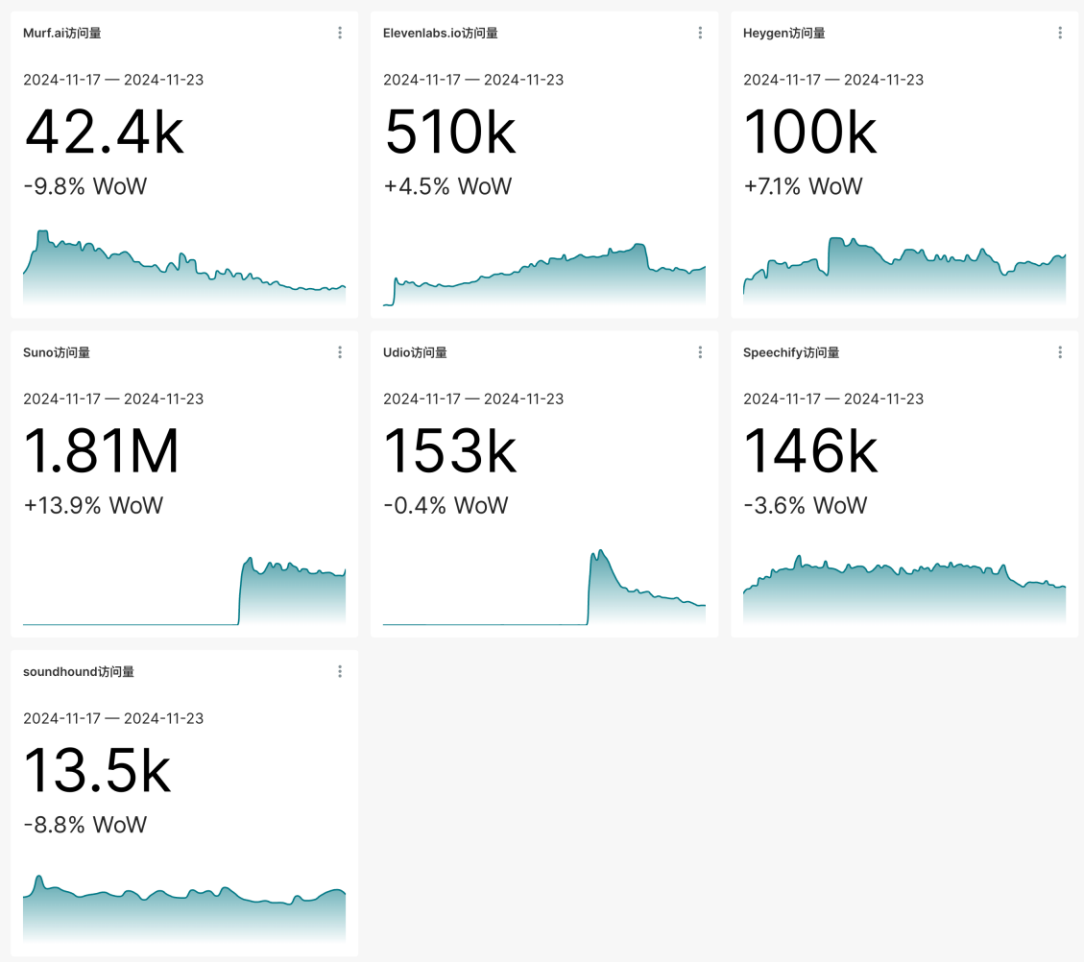
图表10：视频生成类应用周均日活变化



来源：SimilarWeb、国金证券研究所

音乐和音频模型应用的市场空间仍有局限，部分新应用昙花一现，在爆发增长后用户没有留存，活跃度持续下滑如语音合成应用 Murf 和音乐生成应用 Udio。但是部分应用如音乐生成应用 Suno 和语音视频融合应用 Heygen 的活跃度较为稳定，用户留存率较高。随着多模态模型的发展，音乐和音频应用的市场空间会被进一步压缩，创意和易用性是这类应用发展和生存的关键。

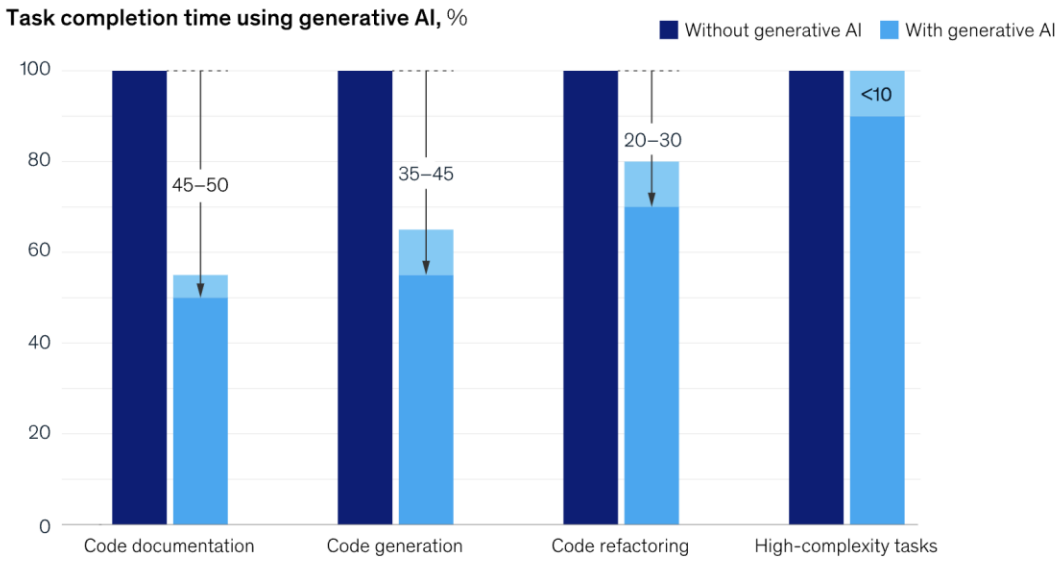
图表11：音乐和音频模型应用周均日活变化



来源：SimilarWeb、国金证券研究所

2.2 AI 程序员是确定性的强需求

图表12：AI 代码生成对不同任务效率的提升



来源：麦肯锡、国金证券研究所

AI 代码生成已经从概念验证阶段进入企业实际应用阶段，并在提升研发效率方面展现出明显价值。根据麦肯锡的调查，使用生成式 AI 进行代码文档编写时，可以节省约 45%到 50%的时间；在代码生成任务中，节省时间在 35%到 45%之间；而代码重构的时间节省幅度较小，为 20%到 30%。对于高复杂性任务，生成式 AI 的效果最弱，时间节省不足 10%。整体来看，生成式 AI 在较简单的任务上表现出显著的效率提升，而在处理复杂任务时，优势相对较小。

从海内外科技公司来看，AI 程序员的渗透率也在不断提升，Google 在财报会上公布，目前超过 25%的新代码是由 AI 辅助生成的，使用 AI 工具的开发者在软件开发任务上的效率提升了 21%。Meta 内部广泛部署的 CodeCompose 工具为数万名开发者提供代码建议和片段，其建议的接受率达到 22%，约 8%的代码来自于这些建议的采纳。在中国市场，阿里巴巴的通义灵码(Tongyi Lingma)工具获得了 20%的采用率，显著提升了开发效率，特别在测试代码实施方面减少了超过 70%的工作量。百度的智能代码助手 Comate（基于文心一言大模型）更是贡献了该公司 27%的日常新增代码。

2.3 AI 搜索是 25 年最有希望诞生超级 APP 的赛道

在大模型上加入搜索功能，可以丰富模型的知识库，缓解模型无法获取新知识和幻觉问题的出现，是最有希望诞生超级 APP 的赛道。Perplexity 作为主打 AI 搜索的应用，活跃度数据再不断提升，同时 ChatGPT 推出的 Search 功能助力其活跃度再创新高，说明 AI 搜索市场仍在快速发展期。Google 作为传统搜索引擎厂商，也在搜索中加入了 AI Overview，对搜索结果进行总结，同时在 AI Studio 中也提供了 AI 搜索功能，其日活跃度目前还未受到负面影响，但 AI 搜索工具都在积极替代浏览器默认搜索引擎，我们认为随着 AI 搜索渗透率提高，传统搜索引擎厂商会受到更严峻的挑战。

2.4 AI 为通用型和数据类 SaaS 平台打开增长空间

在大模型时代，数据的重要性在快速提高，数据不仅是 AI 训练的基础，更是创新、性能提升和商业成功的关键。数据的管理与安全 SaaS 平台业务迎来高速增长期。例如，Snowflake 产品收入达到 9.003 亿美元，同比增长 29%，产品收入超过 100 万美元的客户相比上一季度的 510 个增加到 542 个，同样保持着高增长的还有 Datadog 和还未上市的 Databricks。除了数据类 SaaS 平台，通用类 Horizontal SaaS 平台如 ServiceNow、Salesforce 也积极在业务中加入 AI 功能，比如 ServiceNow 引入了生成式 AI 功能，如 Now Assist 和 Generative AI Controller，这些工具帮助企业提高工作效率，简化项目部署，并提供智能化的用户体验，Salesforce 也发布了 Einstein AI 平台，集成了多种人工智能技术。

图表13：不同类型的 SaaS 公司列表

来源：OpenView、国金证券研究所

我们认为，AI 为 SaaS 公司带来了新的功能和机会，使其能够开发出以前无法实现的解决方案，这种创新能力帮助企业在竞争激烈的市场中保持领先地位，并通过提供更具吸引力的产品来扩大市场份额。对于细分领域定制化的 Vertical SaaS，我们认为机会会远小于通用型 SaaS，垂类 SaaS 应用本身市场空间有限，并且随着 AI Agent 的成熟，其业务会受到更大的冲击。

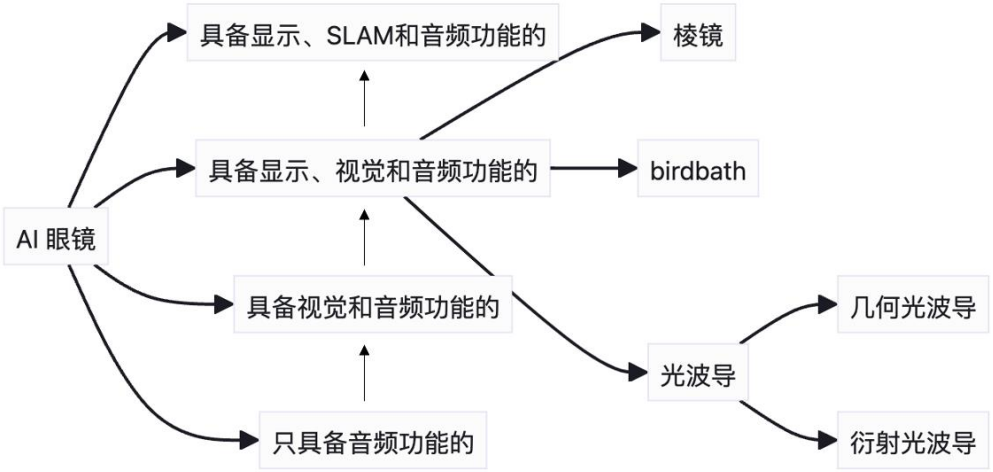
2.5 AI 眼镜是 AI 应用落地的最佳硬件，25 年将迎来发布潮和出货量大增

AI 落地需要硬件载体，目前主要包括 1）传统的消费电子产品，如手机、PC、智能音箱等；2）创新型消费电子产品，如 AI Pin、Rabbit R1 等；3）眼镜形态的 AI 硬件。AI 赋能传统消费电子，基于现有的成熟硬件，推动传统硬件 AI 化，继承传统硬件原有的生态，有助于 AI 应用落地。对于创新型产品，可以探索新的硬件形态，想象力丰富，但需要市场和消费者的验证， 无论是基于传统的消费品嵌入电子硬件，还是针对 AI 应用构建 AI 专用硬件，对于用户的使用习惯、接受程度都是一个很大的挑战。

图表14：探索过程中的 AI 设备类型

来源：Friend、Limitless、Rabbit、AI Pin 官网、国金证券研究所

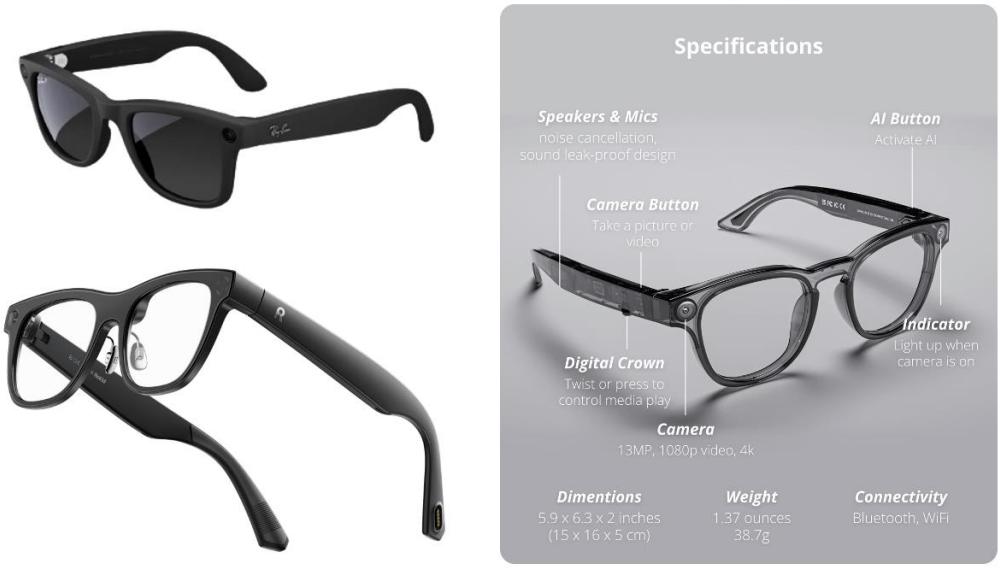
图表15：AI 眼镜技术发展路线



来源：国金证券研究所

从输入输出方式上看，眼镜是最靠近人体三大重要感官的穿戴设备：嘴巴、耳朵和眼睛。嘴巴是语言输出器官、耳朵是语言接受的器官、眼睛则是人类最重要的信息摄入器官，人类 80%的信息来源于视觉。眼镜是人类穿戴设备和电子设备中最靠近这三大感官的群体，是 AI 最好的硬件载体，可以非常直接和自然的实现声音、语言、视觉的输入和输出。目前具备显示功能的眼镜重量仍然会远远超过日常佩戴的眼镜，但是只具备视觉和音频的眼镜已经可以做到接近日常佩戴眼镜的重量。并且目前大模型发展的方向也是多模态和实时性，作为聊天助手返回的内容主要还是文本，但是可以理解图片、视频、音频，只具备视觉和音频的眼镜提供的交互方式契合大模型的使用方式，为目前 AI 应用最好的载体。

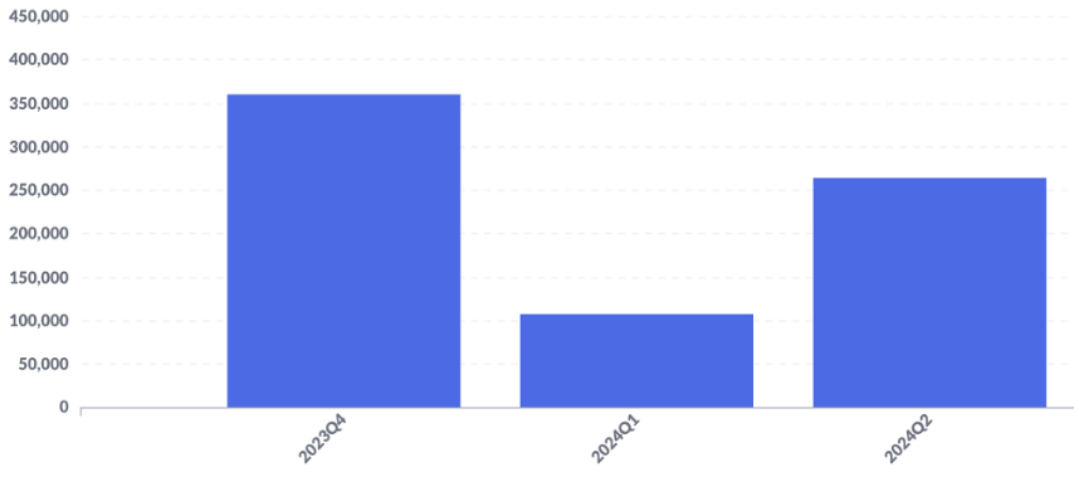
图表16：已经发布的AI眼镜外观和形态



来源：Meta、Rokid、LookTech 官网、国金证券研究所

从具体产品看，Meta 与 Rayban 联名推出的眼镜在 2024 年 4 月开放 Meta AI 功能有已经有放量的趋势，到 2024 年 Q2 有约 80 万的出货量。国内厂商也在积极布局类似形态的 AI 眼镜，2025 年将进入 AI 眼镜大量出货元年，并且随着光波导技术的成熟和模型多模态和实时性能力的进步，AI 眼镜会有更好的体验。我们预计明年率先大量出货的仍是不具备显示功能的类 Meta Rayban 形态眼镜，随着光波导中光机和波导片成本的下降和体积的缩小，后年具备显示功能的 AI 眼镜有大量出货的机会。

图表17：Meta Rayban 季度出货量（台）



来源：IDC、国金数字未来实验室、国金证券研究所

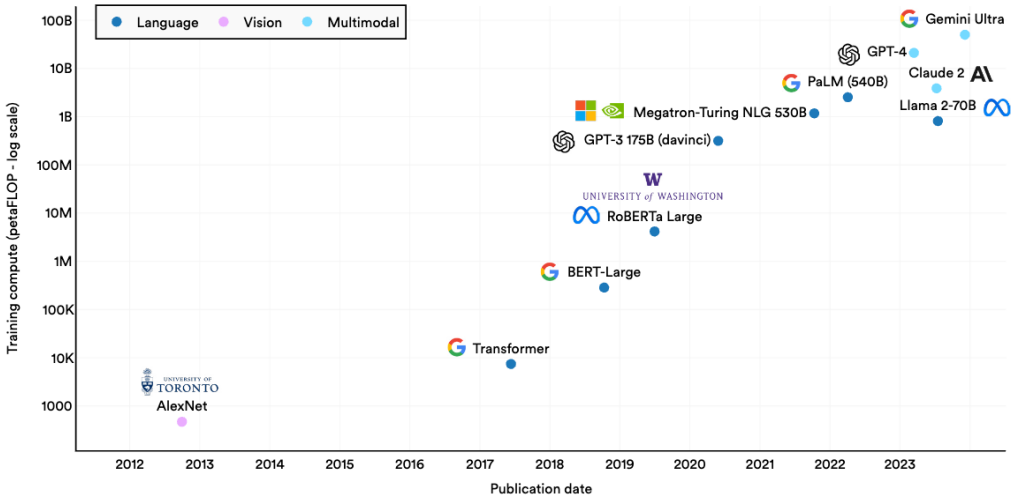
三、算力系统面临“木桶效应”挑战，供给端瓶颈或成主要矛盾

3.1 人工智能算力系统面临诸多挑战

随着人工智能的发展，模型的规模和复杂性呈现出指数级增长，自 2012 年 AlexNet 问世以来，算力需求迅速攀升。AlexNet 作为深度学习在计算机视觉领域的开创性成果，训练时依赖于两块 NVIDIA GTX 580 GPU，耗费约 470 petaFLOP，标志着深度学习时代的来临。而此后模型的扩展速度令人瞩目：2020 年推出的 GPT-3 模型拥有 1750 亿参数，训练消耗约 $3.14x10^{8}$ petaFLOP，GPT-4 进一步升级至 1.8 万亿参数，依赖 25000 个 A100 GPU，计算需求达 $2.1x10^{10}$ petaFLOP，耗时 90 至 100 天，硬件与能源的需求达到新高度。

在最新的超大规模模型——Gemini Ultra 上，算力要求再度跃升至 $5x10^{10}$ petaFLOP。谷歌为此部署了大量 TPUv4 和 TPUv5e 加速器，以应对计算需求和硬件挑战。Gemini Ultra 的训练使用了多个数据中心中跨集群的 TPUv4 加速器，配置在 4096 个芯片组成的 SuperPod 中。每个 SuperPod 通过高速互联进行数据通信，并利用专用光开关在大约 10 秒内动态重配置为 3D 环面拓扑。

图表18：人工智能模型训练所消耗算力需求快速提升



来源：Epoch、国金证券研究所

随着超大规模模型对硬件资源的需求不断增加，系统故障率也相应上升，平均故障间隔时间成比例下降。谷歌通过减少抢占和重新规划的比率尽量减少硬件故障的影响，但在如此规模的硬件部署中，故障不可避免。Gemini Ultra 的计算复杂性推动了多模态 AI 架构和大规模硬件集群的极限，尽管当前的硬件性能接近瓶颈，但要满足这种庞大模型的训练需求仍需数月的时间和大量的能源投入。

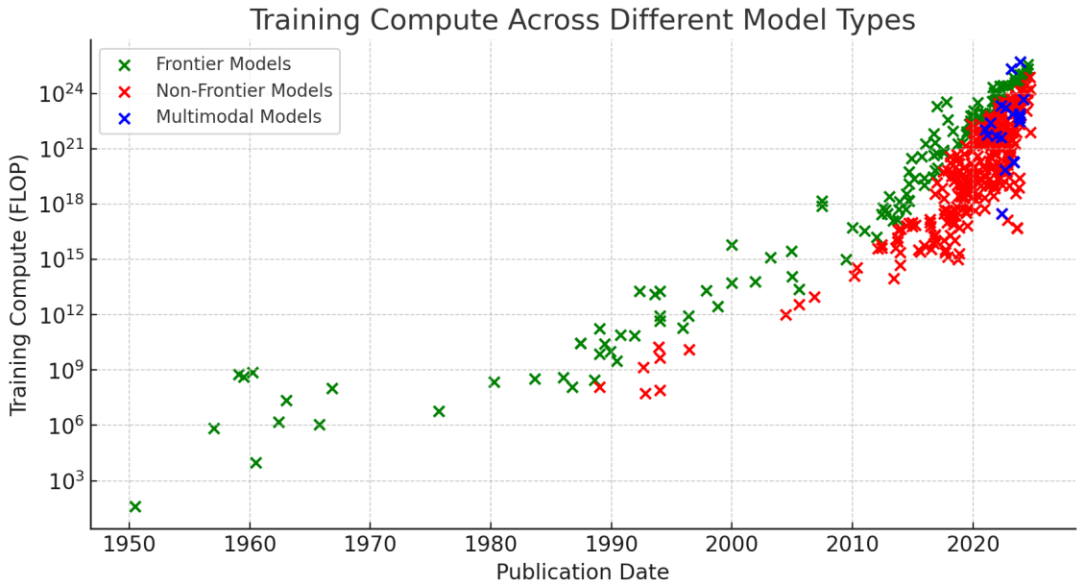
然而，单卡算力、互联性能和能源供应的发展速度已逐渐趋缓。即便硬件性能逐年提升，模型规模的增长速度却更为迅猛，带来了计算瓶颈和能耗压力。因此，AI 模型的未来发展将面临这些硬件和能源限制的制约，解决这些关键短板将成为 AI 系统持续迭代和优化的核心挑战。

3.2 单卡算力升级速率落后于模型迭代速率，Blackwell 延后预示系统摩尔进一步降速

在我们之前的报告中，我们已经指出，当前基于传统摩尔定律的发展速度（即单芯片晶体管数量的扩展速度）已经远远落后于人工智能模型对于算力的需求。最新的 Epoch AI 数据表明，当前一些知名模型的训练所需算力仅需大约六个月就能实现翻倍。我们对这些数据进行了更深入的分析后发现，如今更为主流的多模态大模型（例如 Gemini Ultra 和 GPT-4）在算力需求方面的增长趋势更为陡峭。这些模型的算力需求翻倍时间已经缩短到不到六个月。

与此形成鲜明对比的是，传统摩尔定律所定义的晶体管数量翻倍周期是 18 个月，这意味着上述多模态大模型的算力需求增长速度已经显著超越了传统芯片晶体管数量增长速度。换句话说，大模型算力需求的增长速度已经明显领先于传统摩尔定律所能支撑的硬件性能提升速度。

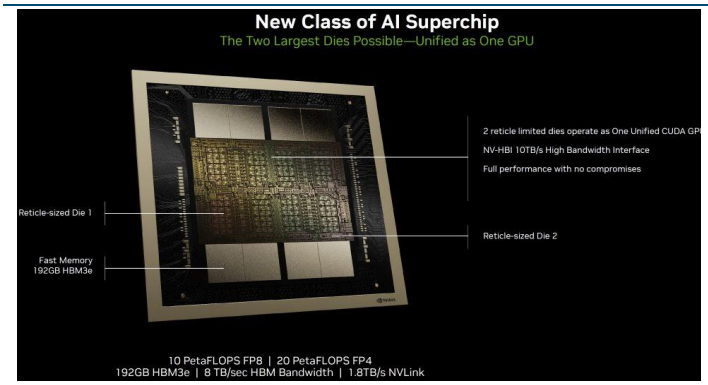
图表19：多模态大模型训练所需算力增长速率显著快于普通大模型



来源：Epoch、国金证券研究所

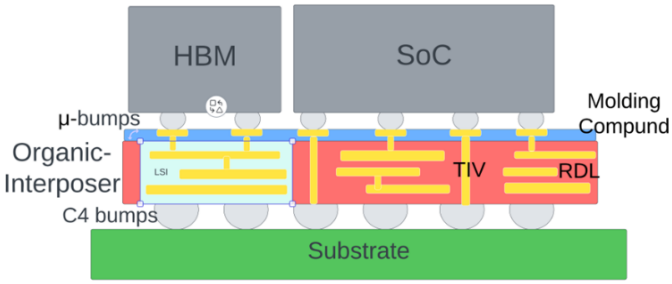
系统摩尔是业界为应对摩尔定律放缓的解决方案。英伟达最新的 Blackwell 架构的核心特性之一是其多芯片模块（MCM）设计，B200 芯片将两个接近光罩极限面积的芯片通过 NV-HBI 技术连接在一起，该技术基于 NVLink5.0 协议，提供高达 10TB/s 的带宽。

图表20：Blackwell 芯片接近两倍光罩极限面积



来源：igorslab、国金证券研究所

图表21：Blackwell 出于成本考虑采用 CoWoS-L 封装



来源：anyscilicon、国金证券研究所

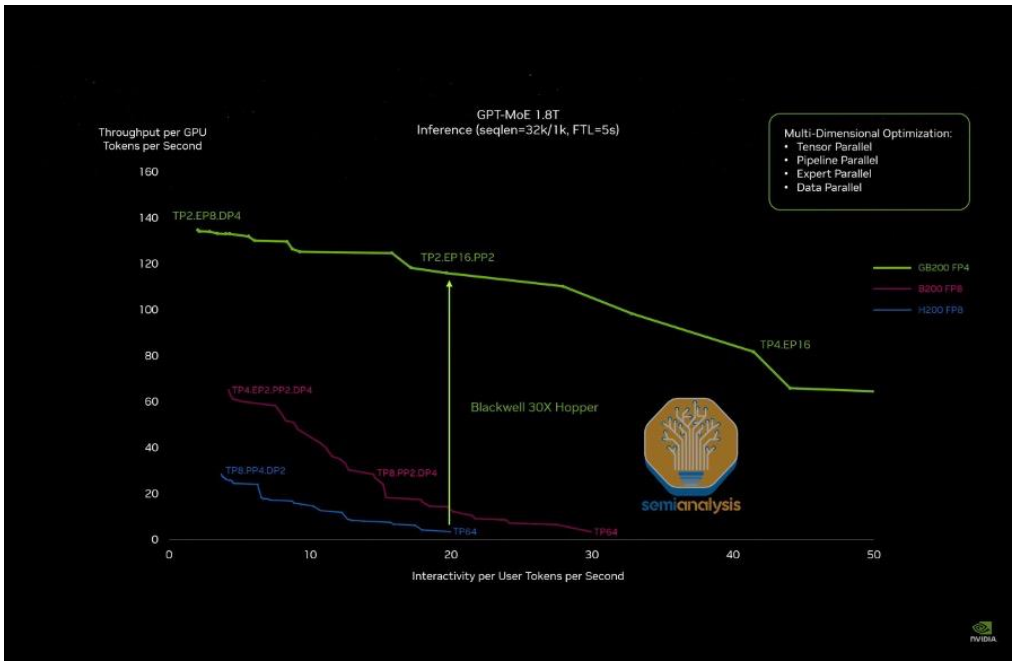
从单卡性能来看，以芯片面积增益进行归一化计算后，空气冷却的 B200 在 FP16 FLOPS 性能上每单位芯片面积仅提升了 14%，这与人们对全新架构的期望相去甚远。这是因为大部分性能提升主要依赖于更大的芯片面积和量化优化。

由于计算芯片（die）的面积不断扩大，封装所需的中介层面积也相应增加，导致整体成本上升。与采用完整硅中介层的 CoWoS-S 技术相比，CoWoS-L 技术通过在有机基板中局部嵌入硅桥的方式，减少了硅的使用量，从而有效降低了成本。这也是 Blackwell 选择采用 CoWoS-L 封装技术的主要原因。但与此同时带来的，是工艺上的新难题，Cerebras 联合创始人指出，此次 Blackwell 延后的核心原因是，GPU 之间以及 HBM 和 GPU 之间的局部硅桥的位置校准出现了偏差，尤其是在 Blackwell 所采用的接近两倍光罩极限面积的中介层上，其工艺难度进一步增加，另外，计算 die、CoWoS-L 中局部硅桥、以及 CoWoS-L 中介层中的 RDL 部分三者的热膨胀系数之间的差异也会导致封装结构出现弯曲，影响系统性能。

发布会上英伟达表示 GB200 相较于 H200 在 1.8T 参数的 GPT-MoE 模型上的推理性能将提升 30 倍，然而，这一数据是基于一个非常特定的最佳场景得出的。需要明确的是，这一场景在理论上确实可以实现，但并不能完全代表市场中的普遍应用场景。解释 30 倍性能提升的一个关键因素是将 GB200 NVL 在 FP4 下的性能与 H200 和 B200 在 FP8 量化下的性能进行对比，而且比较基准选取的是最不适合 H200 的 64GPU 张量并行，根据

Semianalysis 模拟分析，这一情形下实际性能提升仅有 18 倍，如果在更贴近现实的情况下，性能提升幅度将更低。

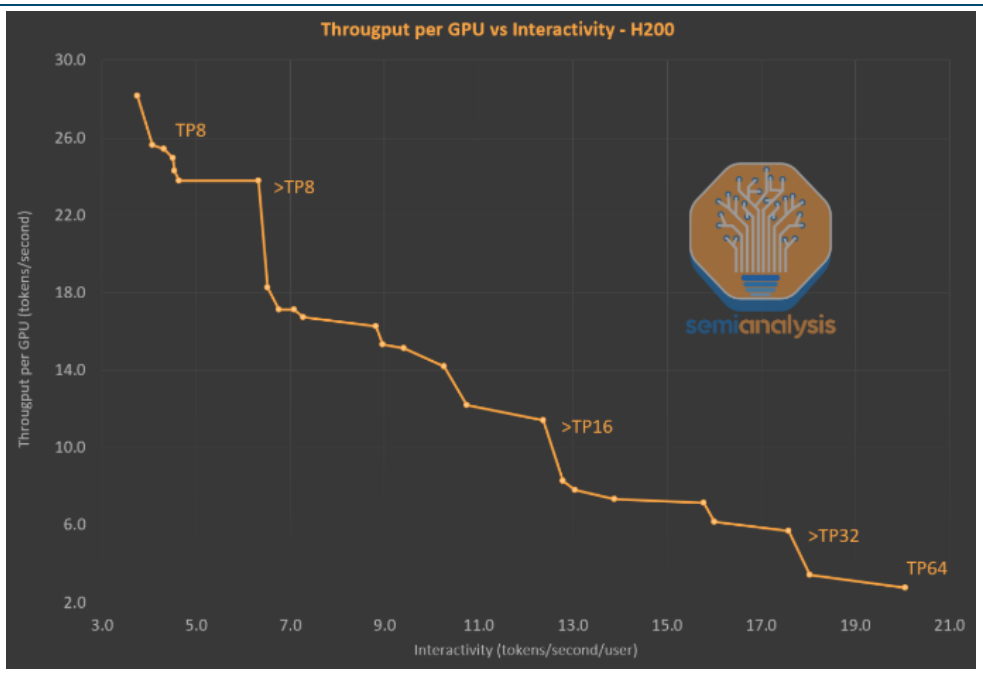
图表22：发布会上英伟达表示 GB200 的 GPT-MoE 推理性能能够达到 H200 的 30 倍



来源：英伟达、Semianalysis、国金证券研究所

我们认为 Blackwell 因设计问题延迟出货已经反映出了数据中心高性能计算芯片在制造段继续迭代的瓶颈，尽管英伟达可以通过节点内和节点外互联提升总体系统性能，但我们认为单卡算力(计算性能/功耗)的提升仍旧是必要的，节点内 GPU 间通信(NVLink)慢于片上通信，节点间通信(Infiniband/Ethernet)又显著慢于节点内通信，导致并行化带来的算力提升是边际递减的，单卡 PPA 的提升仍是后续系统性能继续提升的关键。

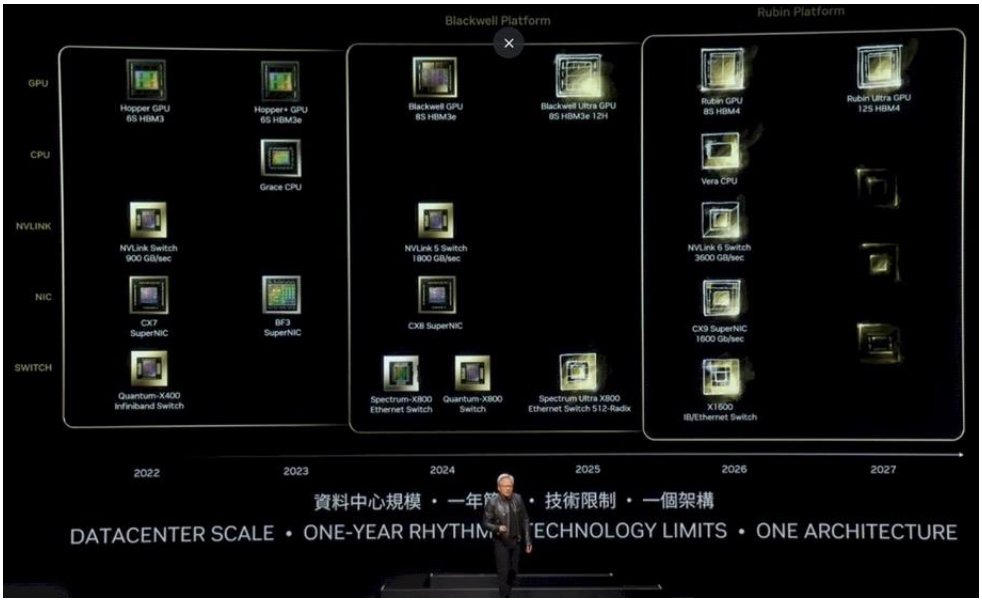
图表23：H200 张量并行系统中，节点间互联比例越高，整体性能越低



来源：Semianalysis、国金证券研究所

当前市场对英伟达的预期相当充分，根据彭博一致预期，市场预期英伟达 FY2025Q4 至 FY2026Q3 毛利率分别为 73.5%、72.2%、72.9%、74.2%，说明市场对未来三个季度 Blackwell 研发部署对毛利率的压制是有所认知的，但认为 FY2026Q3 对毛利率的压力将有所缓解。

图表24：英伟达预计将在2026年发布并出货Rubin GPU



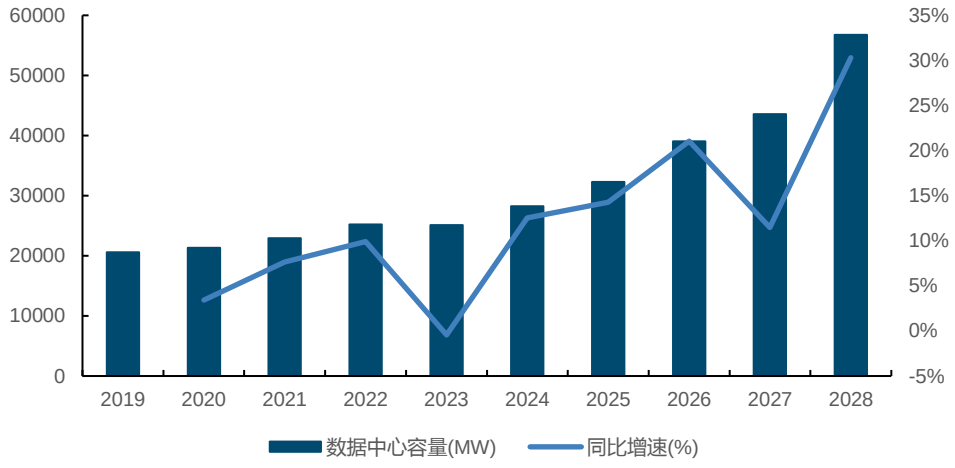
来源：nextplatform、国金证券研究所

从时间线上来看，FY2026Q3英伟达或将开始出货Blackwell Ultra，Blackwell Ultra即为Blackwell的HBM升级版本，技术上难度相对Blackwell并没有显著提高，市场预期FY2026Q3毛利率有所回升是合理的。我们不同于市场的观点是，应当警惕下一代产品即Rubin不能如期发布的风险，对英伟达的下一代产品来说，从芯片制造的角度，我们认为无论是从单位面积晶体管缩放还是先进封装角度，实现大幅度性能提升的难度都不容小觑。

3.3 数据中心电力消耗呈指数级增长，核电或成最优解决方案

根据IDC数据，24年云服务厂商数据中心容量达到28240兆瓦（MW），2028年将达到56756兆瓦（MW），CAGR为19%。24年云服务厂商数据中心预计消耗电力约达到563亿千瓦时，按全球23年发电量29.92万亿千瓦时来算，云厂数据中心耗电量占比将达到0.2%，而如果按全部数据中心耗电量4170亿千瓦时来计算，则这一比例达到1.4%。按2028年8568亿千瓦时用电量来计算的话则占比达到2.9%。数据中心耗电量的快速上升将会影响到正常生活中的用电。且全球主要数据中心集中在中国、美国、欧洲等地区，这些国家发电量仅为全球的一半左右，但数据中心用电量基本没有减少，数据中心耗电量的比例在这些国家中的还会继续上升。如果再进一步集中到这些国家中数据中心密集的地区，则地区的用电压力还会进一步提升。

图表25：全球云服务厂商数据中心容量（MW）



来源：IDC、国金数字未来实验室、国金证券研究所

为了应对越来越高的能源需求，主要的云服务厂商都打算将能源供应的责任放在核电站上。

独立于居民、工业用电的核电具备许多优势。1）尽管核电站的建设成本历来较高，但其运营成本相对较低，单个反应堆的发电容量通常超过 800 MW。此外，核电站发电过程中不直接排放二氧化碳，对于那些投资高能耗数据中心且试图实现减排目标的科技公司来说，核能具有重要吸引力。2）与住宅或许多其他行业的用电需求不同，数据中心的用电需求在一天中的各个时间段相对稳定。这种持续的用电需求非常契合核电站的运营特点，后者通常无法快速调整发电功率以应对需求波动。核电站持续稳定的发电能力能够确保数据中心在全天候都能获得足够的电力，同时还为其提供了零碳排放的大规模能源来源。3）当数据中心与发电源直接连接时，数据中心可以直接从发电厂获取电力，而无需经过更大的输电网络。尽管购电协议的存在并不要求发电厂和数据中心必须在同一地点，甚至不需要在同一时间发电和用电，但这种安排可以通过直接将需求增长与发电来源匹配，降低整体电网成本。

24 年 3 月，亚马逊斥资 6.5 亿美元，从电力运营商 Talen 能源手里买下一座占地 1200 英亩的“核电数据中心园区”——数据中心就坐落在两个核反应堆边上。除此之外，亚马逊也在积极和 Constellation Energy 寻求更多核电站合作。24 年 9 月，Constellation Energy 宣布了一项为期 20 年的购电协议（PPA），将为微软位于美国中大西洋地区的数据中心提供电力。这些电力将来自宾夕法尼亚州三哩岛核电站的 1 号反应堆。谷歌 24 年 10 月宣布，与 Kairos Power 公司签署协议，将利用小型核反应堆来生成支持其人工智能（AI）数据中心所需的巨大能源。根据协议内容，谷歌计划在本十年内开始使用首个核反应堆，并在 2035 年前引入更多的核能设施。

图表26：云厂数据中心用核电供电计划

云厂	时间	核电站情况	核电站	容量
谷歌	2024 年 10 月	达成协议，待建	地点未知	6-7 个小型 SMR 核反应堆，累计 500MW，2035 年交付
微软	2024 年 9 月	核电站重启	宾州三哩岛核电站	835MW
亚马逊	2024 年 3 月	在建，进度过半	宾州萨斯奎哈纳核电站	960MW
	2024 年 10 月	签署协议，待建	弗吉尼亚州北安娜核电站	5 亿美元合作开发 SMR 小型核反应堆项目，累计 300MW
	2024 年 10 月	签署协议，待建	华盛顿州	由亚马逊投资的 X-energy 开发 4 个 SMR 小型核反应堆项目，累计 960MW

来源：eia、bbc、财联社、亚马逊官网、国金证券研究所

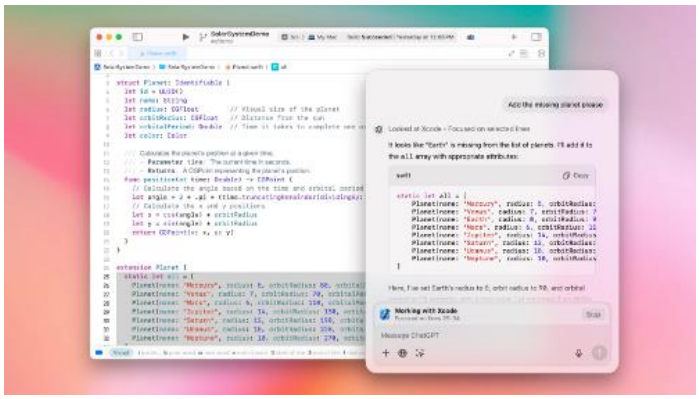
但核能并非毫无风险。除了核反应堆安全问题之外，数据中心&核电站供电方式也存在新的问题。虽然新建核电站可以单独供给数据中心，但已建成的核电站通常与电网相连，数据中心过度供电依旧会威胁电网的可靠性，产生的额外费用目前也没有规定该由谁支付。这也是 24 年 11 月，美国联邦能源管理委员会否决亚马逊增加数据中心供电提案的主要原因。

考虑政治&安全因素，一些大国如中国、美国国内核电站新建政策可能会收紧，但东南亚一些国家正在积极扩张核电。除了越南、缅甸、马来西亚等已经建设或考虑建设核电站的国家外，泰国于 11 月 15 日签署核电站项目合作备忘录，首次启动核能发展，以推动清洁、低成本能源建设。该项目以 SMR 小型模块化反应堆技术为核心，旨在降低电价并减少碳排放。该技术具有高安全性、空间需求小等优势，并计划将核电纳入 2037 年国家清洁能源发展目标。

四、大模型推理服务大规模部署，如何影响硬件市场？

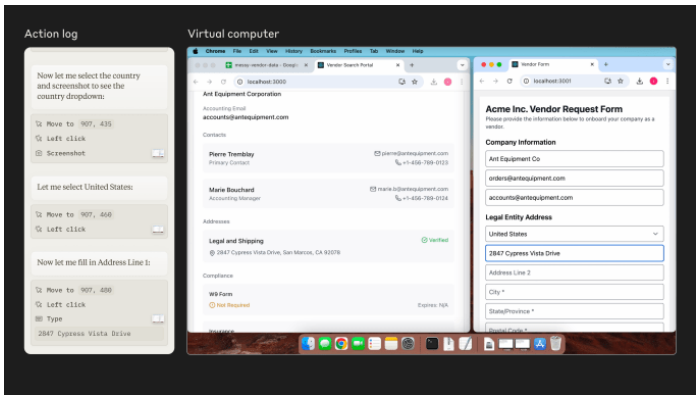
4.1 大模型性能提升，推动推理算力需求加速增长

图表27：OpenAI 发布 Work with Apps 功能



来源：OpenAI、国金证券研究所

图表28：Anthropic 发布 Computer USE API

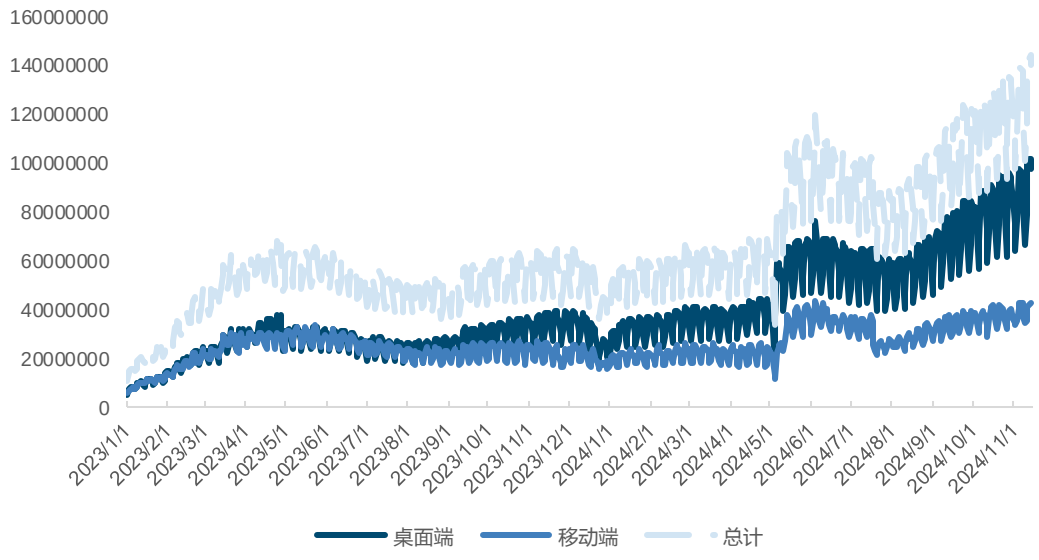


来源：Anthropic、国金证券研究所

大模型服务已从聊天机器人进化为严肃生产力，十一月中，MacOS ChatGPT 客户端已经开始支持读取用户屏幕上的代码并给出编程建议，这是 OpenAI “Work with Apps”功能在编程工具上的体现，从名字上可以看出，该功能可能不仅面向编程工具，未来可能支持更多工具。Anthropic 也已在十月中发布了其 Claude 3.5 Sonnet 更新版本，通过其“Computer USE”API，Claude 被训练具备屏幕视觉理解能力，能够“观察”屏幕上发生的事情，并通过分析屏幕截图理解用户界面（UI）的布局和内容。当开发者将特定任务交付给 Claude 并授予其必要的权限时，它可以通过解析截图计算光标需要移动的具体像素距离（包括垂直和水平方向），以便精准定位到目标区域进行操作。

尽管作为第三方模型，Work with Apps 和 Computer USE 并没有接入系统底层，大模型在系统层面的集成已经初见雏形，从推理算力的结构上来看，系统集成大模型提供类似于 AI Agent 功能，输入和输出 Token 的数量将大大增加，单位 Prompt 所需的推理算力将显著增长。

图表29：ChatGPT 访问量加速增长



来源：Similarweb、国金证券研究所

根据我们追踪的 ChatGPT 访问量数据，我们认为大模型正在被加速应用，其生产力属性已经在消费级市场获得了验证，GPT-4o 和 Claude 3.5 的发布代表着大模型能力进入了一个新的阶段，将驱动推理算力需求的大幅提升。

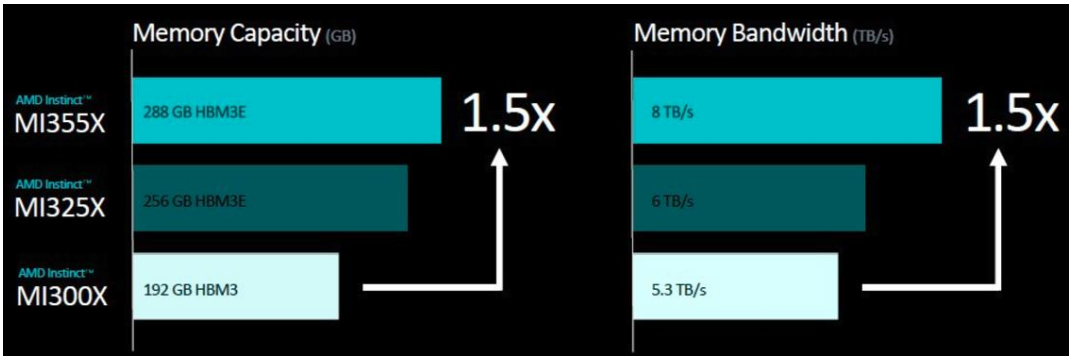
4.2 服务器推理：内存墙难破，HBM 容量仍为竞争要点

GPT 类模型通过给定前文来预测下一个 token(即单词或符号)进行训练。在生成文本时，需要首先输入提示词（prompt），然后模型预测下一个 token，并将其添加到提示词中，随后再预测下一个 token，重复这一过程直到完成生成。这一生成机制每次生成下一个 token 时，所有模型参数必须从内存传输到处理器，而这些庞大的参数需要尽可能靠近计算单元存储，以降低数据传输的延迟，必须确保这些参数能够在需要时精准加载到芯片上。这种推理模式对硬件的内存带宽、容量以及数据传输效率提出了严苛要求，也成为当前生成式 AI 技术突破的重要瓶颈之一。

Instinct GPU 相对于英伟达 GPU 一直提供更高的存储容量和存储带宽，MI325X 和 Hopper 架构中存储容量最大的 H200 相比，Instinct 在显存容量上具有 1.8 倍的优势，这意味着加载特定模型参数时所需的 GPU 数量减少了 1.8 倍，同时，AMD 在带宽上也具备 1.25 倍的优势，这表明在将模型参数传输至 GPU 的过程中所需时间更短。

在 AMD 的路线图中，未来的 Instinct GPU 的存储容量和带宽将持续增长，MI355X 将采用 HBM3E，存储容量将达到 288GB，带宽将达到和 B200 相同的 8TB/s，而存储容量将显著高于 B200 的 192GB。

图表30：MI355X HBM3E 容量将达到 288GB



来源：nextplatform、国金证券研究所

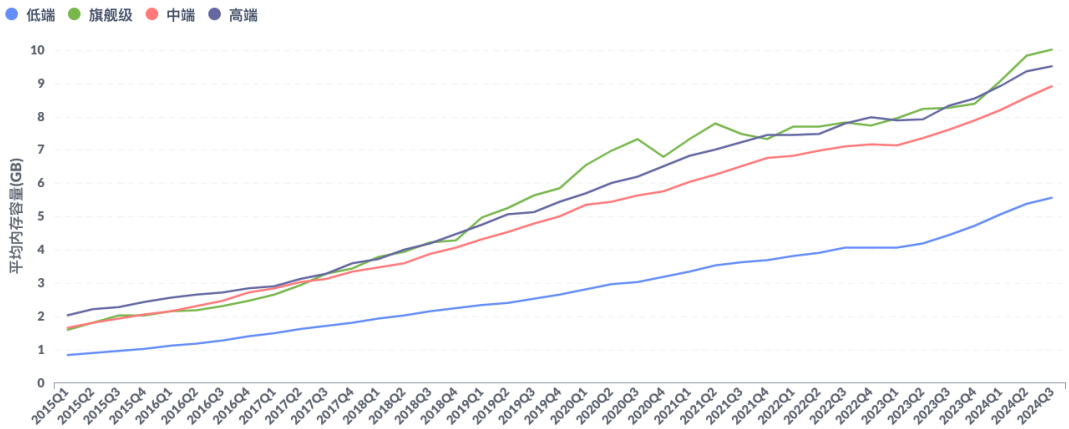
Instinct GPU 除了在存储容量方面具有显著优势，其相对英伟达 GPU 较低的定价更为低廉，三星于去年采购过一批 MI300X GPU，单价约为一万美金，相较于当时 H100 三万至四万美金，有显著的成本优势。根据我们的产业链调研，出于降低成本和寻找第二供应商的考虑，海外云场正在积极尝试使用 AMD GPU 集群。尽管 AMD 受制于软件生态和互联性能在训练领域尚难以于英伟达竞争，我们认为随着推理算力需求大幅提升，AMD 在该领域将持续收益。

4.3 端侧推理：单用户推理导致内存端高成本，端云结合将是未来趋势

AI 手机和 AIPC 提供的端侧 AI 允许用户将数据留在本地，但端侧 AI 的单用户场景意味着 Batch Size 为 1，这意味着每次从内存加载模型参数到芯片上时，其成本只能分摊到单个 token 上，无法通过其他用户的并发计算来缓解这一瓶颈，服务器端的推理我们先前已经讨论过，内存墙仍然存在，但多个用户的推理请求使内存加载参数的成本分摊到多个 token 上，大幅降低单个 token 生成的开销，对生成式模型的推理效率提升有显著作用。

从模型参数占据的存储空间来看，当前 AI 手机的内存容量仍旧是严重不足的。以 Llama 7B 模型为例，在 FP16 格式下，每个参数占据两个字节，对应 14GB 的内存容量，除此之外，手机 RAM 中还需要存储应用程序和操作系统相关数据，在 AI 手机本地存储并运行这一规模的端侧模型还是颇有难度的。根据 IDC 数据，端侧 AI 需求尚未推动智能手机单机内存容量显著增长，我们认为这并非手机厂商没有意识到端侧 AI 的重要性，而是在端侧实现高性能模型所需的存储容量远高于目前技术所能提供的，即便手机厂商将存储容量从 16GB 提升至 32GB 能够显著增大可容纳模型参数规模，但和超大云端模型当前所能提供的性能相比，我们认为仍旧是不具有可比性的。

图表31：各价格段智能手机平均内存容量



来源：国金数字未来实验室、国金证券研究所

从消费者的角度来看，端侧并非严肃生产力场景，用户并不需要频繁处理复杂任务。即使是在类似 AIPC 这样的端侧场景中，复杂任务往往可以通过网页或客户端接入云服务来完成，而非依赖本地化运行复杂内容。因此，单纯为了本地化复杂任务而额外增加内存开销并不具备充分的合理性。

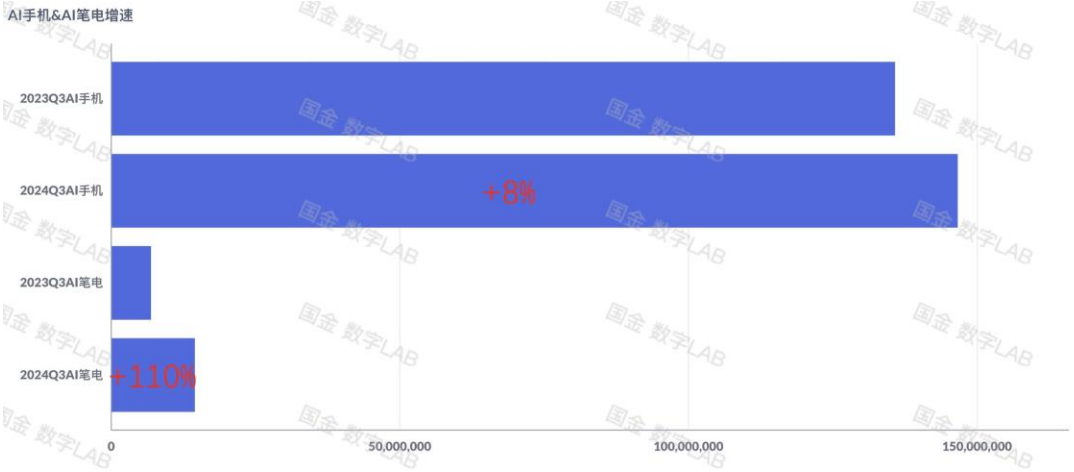
我们认为，端侧 AI 用户的核心需求并不在于直接在本地处理复杂任务，而是通过 AI 实现系统层面的非标准化操作。例如，自动将个人住址信息填写到电商应用中。相比单纯依靠提升 AI 能力来满足这些需求，我们认为更优的解决方案是将系统底层的数据接口和指令接口与 AI 模型深度集成。具体而言，AI 模型可将用户指令拆解为具体操作指令，并通过脚本直接与操作系统交互，从而以更高效、更经济的方式实现个性化功能。

在本地模型性能显著提升需要大量额外内存容量，而端侧 AI 用户的核心需求能够通过数据接口和脚本操作来满足的背景下，我们认为端侧 AI 硬件厂商大幅增加内存容量并非明智之举。当前市场数据也验证了这一观点，在内存技术尚未实现单位体积容量大幅提升或单位容量成本显著下降的前提下，端侧 AI 硬件厂商对内存容量配置的谨慎态度可能将持续。

五、AI 设备销量正在提升

24 年三季度，AI 手机销量达到 1.47 亿台，同比增长约 8%，AI 笔电销量达约 1450 万台，同比增长 110%。

图表32：全球 AI 手机&AI 笔电销量（台）及增速



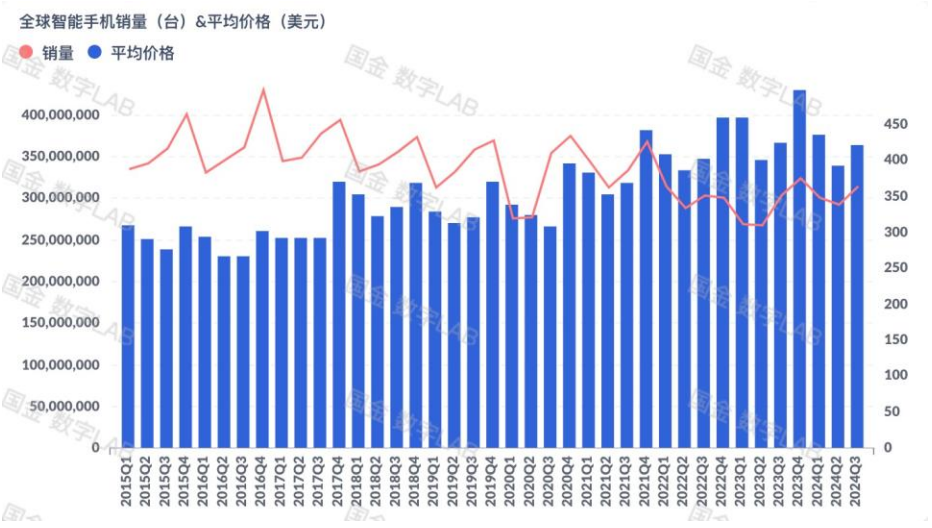
来源：IDC、国金数字未来实验室、国金证券研究所

5.1 AI 手机焦点在于旗舰机

相较于传统智能手机，AI 手机在硬件上重点突出了 NPU 和内存的提升。AI 手机主要由 NPU 来负责端侧的推理，部分解放了 GPU 资源，使得手机可以更好的运转。同时内存容量的提升也是在进行端侧推理却不降低手机流畅性的必要条件。作为新加入者的 AI 功能，用户对其的要求是不能影响以前的视屏、游戏、办公等功能。因此，高配置、高价格的旗舰机在硬件需要提升带来成本增长的情况下更能满足消费者的需求，涨价也更容易被消费者接受。从价格上看，全球智能手机价格在 21 年后缓慢回升，且 700 美元以上的智能手机销量在不断提升，24 年三季度销量达到约 7400 万台，同比增长 2%。

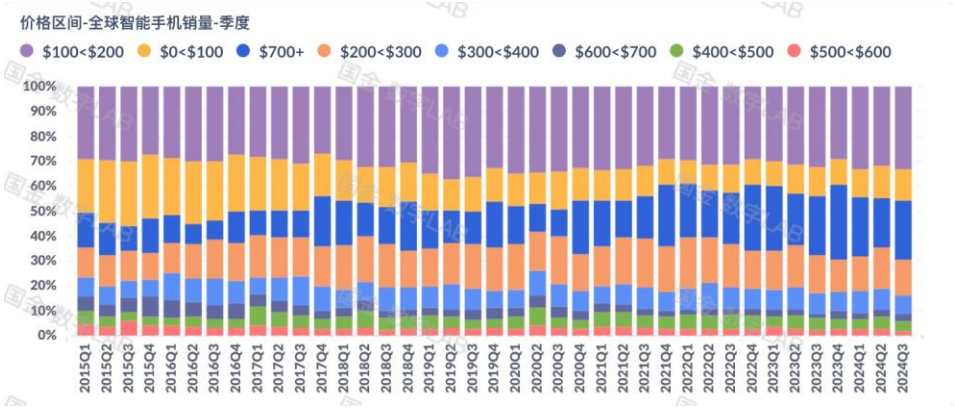
10 月安卓系厂商公布的数据表明了消费者对于新一代旗舰机的热情。根据 vivo 官方公告，X200 系列手机全渠道销售金额已经突破了 20 亿元，这一数据打破了 vivo 历史上所有新机销售记录，显示了消费者对这一系列新品的热烈欢迎。截至 2024 年 10 月 19 日，X200 系列的销量估计在 29.4 万到 46.5 万台之间（按最低价与最高价估计）。11 月 9 日，在小米直播中，小米集团总裁卢伟冰透露，小米 15 系列的销量已破 100 万台，破百万的速度要快于前代小米 14 系列。

图表33：全球智能手机销量（台）&平均价格（美元）



来源：IDC、国金数字未来实验室、国金证券研究所

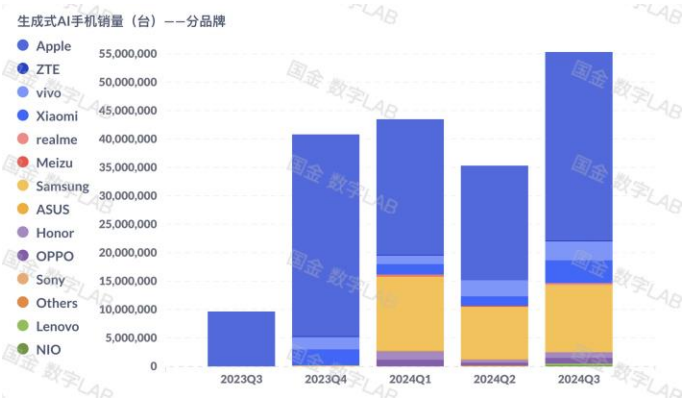
图表34：全球智能手机平均价格区间（美元）



来源：IDC、国金数字未来实验室、国金证券研究所

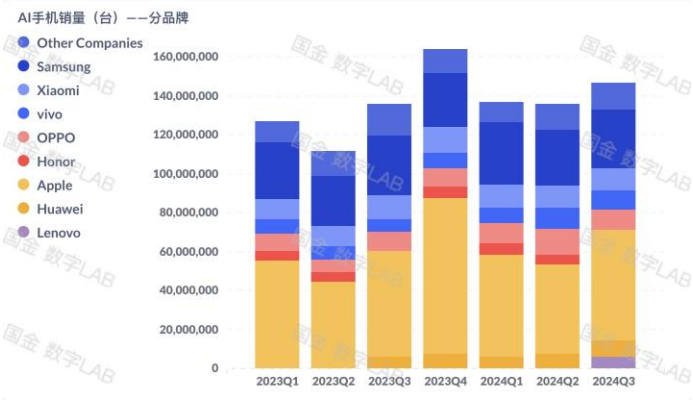
24 年三季度，在具备基础 AI 功能的手机中，苹果占约 42%的份额，安卓系厂商占到约 58%的份额。而在可以完成本地推理的 AI 手机中，苹果 24 年三季度占据约 60%的份额，安卓系厂商约占到 40%。我们认为在 2022 年 ChatGPT 爆火后的两年中，安卓系厂商和苹果在推进将 AI 功能融入进自己的生态中。即使现在爆款的 AI 应用还没有出现，但各家大厂已经在硬件、软件、生态上布局，目标是当爆款 AI 应用真正出现时，自家的旗舰手机能够支持这些应用。

图表35：生成式 AI 手机销量（台）



来源：IDC、国金数字未来实验室、国金证券研究所

图表36：AI 手机销量（台）



来源：IDC、国金数字未来实验室、国金证券研究所

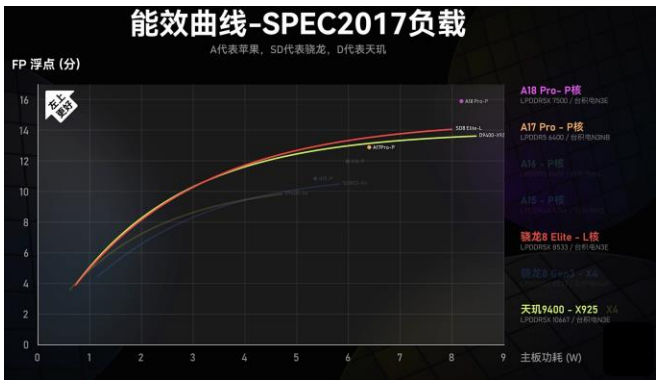
硬件的提升是手机更新换代中至关重要的一环，但与以往不同的是，这次 AI 革命中，硬件、软件、生态缺一不可。

SoC 端，23 年以来，联发科天玑系列芯片的崛起，为安卓系厂商旗舰机新增了一个选择。以往，高端手机除了苹果外，高通旗舰芯片几乎是唯一选择。但在 2023 年联发科使用天玑 9300 的 ARM 公版 4 超大核+4 大核的策略以来，联发科凭借其优异的性能使得今年 vivo、OPPO 等厂商选择天玑 9400 作为旗舰机的 SoC 选择。在目前的竞争中，苹果 A18 Pro 搭载了和 A17 Bionic 相同的 16 核 NPU，支持每秒高达 35TOPS 的计算能力。高通骁龙 8 Elite 采用增强的 Hexagon NPU 技术，具备 80TOPS 算力，性能提升了 45%，能效提升了 45%，支持更长的 token 输入、多模态 AI 助手的本地部署，综合 AI 性能增强达到 45%。天玑 9400 凭借全新第八代 NPU 890，不仅 AI 跑分再夺得苏黎世理工学院的 AI Benchmark 测试第一，同时还首发带来了天玑 AI 智能体化引擎，端侧视频生成及端侧 LoRA 训练，全面提升了端侧 AI 的体验。我们认为，安卓系手机厂商在有了更多的旗舰 SoC 选择后对于高通的依赖性将会有所降低，更有效的策略会促使厂商研发更能满足消费者的产品。

图表37：手机旗舰 SoC CPU 能耗曲线（整数）

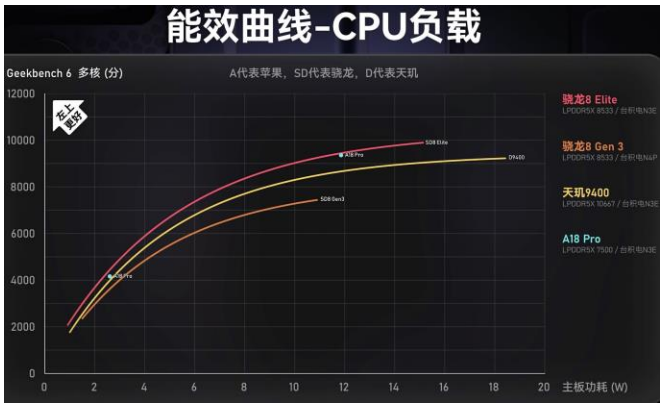
来源：极客湾，国金证券研究所

图表38：手机旗舰 SoC CPU 能耗曲线（浮点）



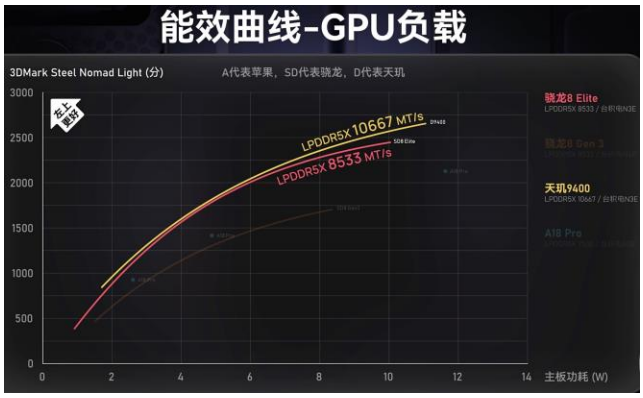
来源：极客湾，国金证券研究所

图表39：手机旗舰 SoC CPU 负载能耗曲线



来源：极客湾，国金证券研究所

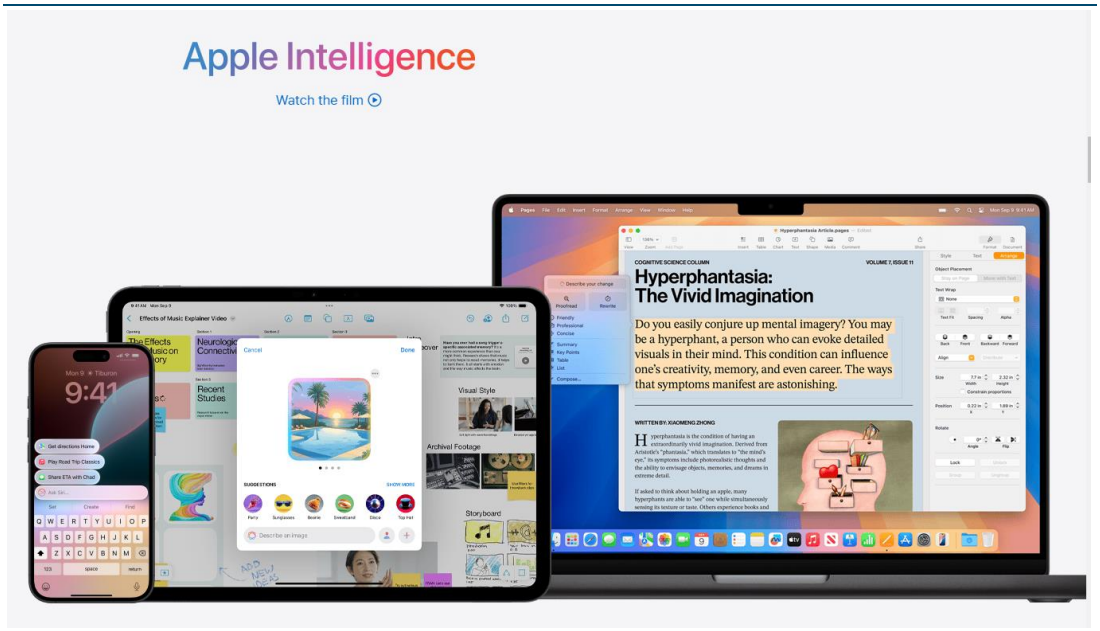
图表40：手机旗舰 SoC GPU 负载能效曲线



来源：极客湾，国金证券研究所

正如前文提到的那样，我们认为在未来，端云结合是更适合 AI 手机的方案。操作系统的提升是端侧 AI 推理对使用体验提升最直观的场景，也是目前手机厂商关注的重点。在硬件支持端侧推理后，苹果着重强化了 Siri 的能力，规划中 Siri 将扮演智能语音助手的职责，帮助用户用语音或简短的文字操作手机。新 AI 技术加持后，Siri 的交互更加自然灵活。用户可随时与 Siri 交流,还可在文本与语音之间随心切换,即便用户表述不够连贯，Siri 也能顺畅理解，并能为连续的请求沿用语境场景。实际操作过程中，系统的反馈速度、对用户要求的准确性都是影响体验的关键因素，因此端侧硬件的性能以及模型的能力都至关重要。小米也在不断更新小爱同学，让其完成 AI 手机助手的职责。在娱乐场景可以通过端侧的推理简单、高效的强化用户体验，一些小型的的生产力场景如会议纪要/文件摘要在端侧完成也可以大幅提高效率。而在一些大型生产力场景如生成视频或者多路并行任务，云端推理可以解放本地资源，使得手机运转更为流畅。

图表41：Apple Intelligence 相关功能



来源：苹果官网、国金证券研究所

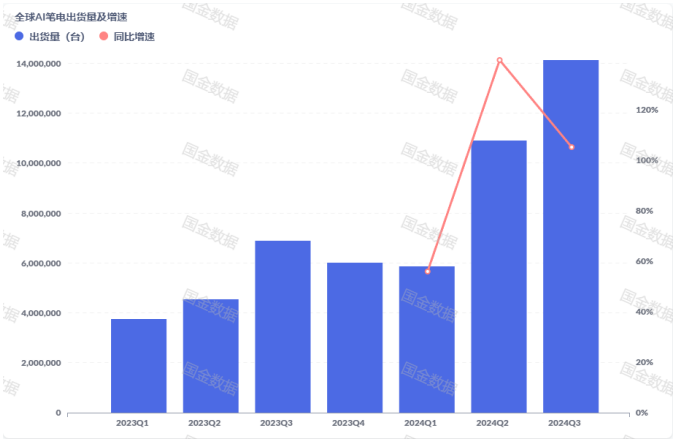
通常来说，端云结合的模式下，需要硬件（性能）、软件（适配操作系统）、生态环境（AI 应用的数量及质量）、云（模型质量及云服务质量）等环节参与。我们认为全部环节都可控并参与的手机厂商更容易成功。苹果、谷歌在某项上有些缺陷，但整体来看链路更为完整。苹果在硬件、软件、生态环境、云服务上能力都很强，但是在模型领域需要暂时和 OpenAI 合作。谷歌有原生安卓支持、Gemini 强大的模型能力，但在硬件上自己的 Pixel 手机渗透率低，需要仰仗三星端侧硬件拓展用户。两者相比的话，苹果的 AI 服务现金化率我们认为将会更高。1）苹果本身用户群体更容易接受付费。iPhone 的价格在所有大厂的智能手机中属于偏高的档次，苹果的用户付费能力也会更强一些。从苹果 FY23-24 财报来看，服务收入为 961.69 亿美元，同比增长约 13%，收入占总收入的 24.59%，是除 iPhone 外收入最高的产品。此次 Apple Intelligence 的订阅费为 20 美元/月，与 ChatGPT 订阅费一致，用户平替的阻碍较小。谷歌的 Gemini 近期从免费变为收取 20 美元/月的订阅费，相对来说，用户付费阻力更大一点。根据数据，24 年一季度虽然谷歌 Play 商店应用总下载量（255 亿次）远超过苹果 App Store 的 84 亿次，但苹果在 2024 年一季度的收入为 246 亿美元，反而是谷歌同期 112 亿美元收入的“近两倍”。

2）苹果硬件软件生态更为完整，同时与 OpenAI 的合作摩擦更少。作为谷歌 AI 的载体，三星即将收取 AI 服务费，这降低了用户对 Gemini 的付费意愿。三星作为全球安卓手机市场的龙头，对谷歌的态度处于合作&竞争并存的状态，这种合作形式不利于谷歌全额赚取 AI 服务的费用。

5.2 AI PC 的竞争将会越发激烈

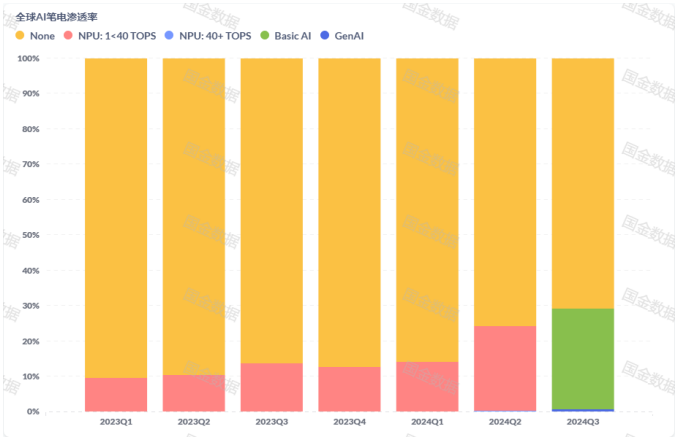
24 年三季度，全球 AI 笔电销量达到约 1400 万台，同比增长 105%，AI 笔电渗透率达到了约 30%，相比二季度提升了约 5 个百分点。其中英特尔核 AI 笔电销量约为 720 万台，占比约 50%，苹果、AMD、高通核的 AI 笔电销量分别为 540 万、134 万、55 万台，占比分别为 37%、9%、4%。

图表42：全球 AI 笔电销量（台）及增速



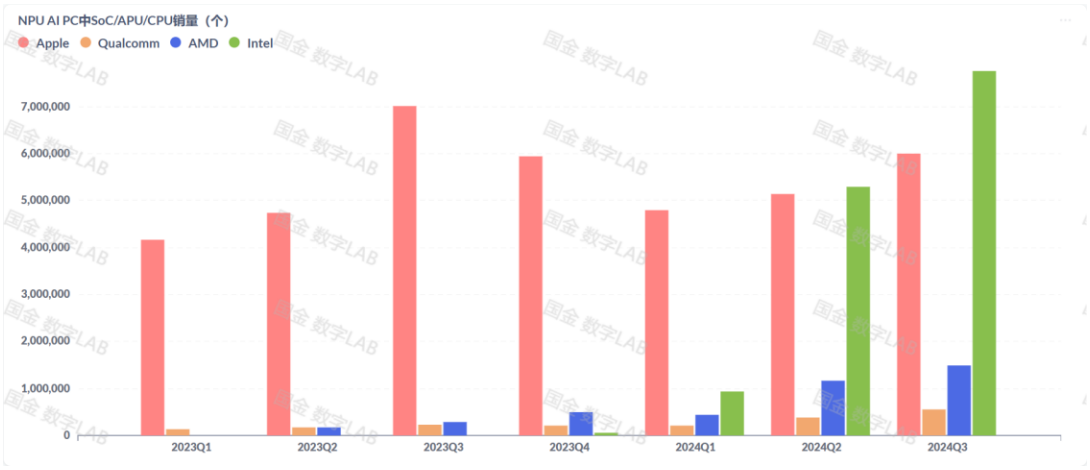
来源：IDC、国金数字未来实验室、国金证券研究所

图表43：全球 AI 笔电渗透率



来源：IDC、国金数字未来实验室、国金证券研究所

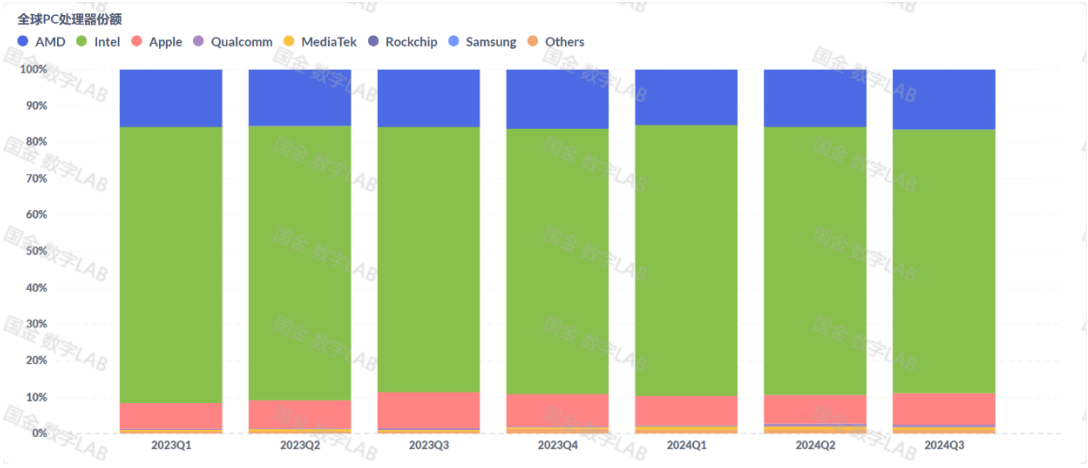
图表44：AI 笔电 SoC/CPU/APU 销量（个）



来源：IDC、国金数字未来实验室、国金证券研究所

作为传统 CPU 厂商，英特尔在互联网时代凭借的设计牢牢把控着服务器、PC 设备 CPU 的市场，同时和微软的“Wintel”联盟以及芯片先进制成使得英特尔的护城河足够宽。但在近些年移动互联网（手机等移动设备）的发展过程中，英特尔错失了手机 CPU 市场。再加上三星、台积电工艺迅速提升而英特尔晶圆技术发展陷入停滞，除了传统服务器&PC CPU 业务外，英特尔近年来业务范围和规模不断缩小，导致收入与利润水平不断下滑。因此，在“最后的大本营”CPU 处理器上，英特尔实际已经不能再失败。但与二十年前一家独大不同的是，在 X86 架构处理器上 AMD 已经摆脱需要英特尔这个对手帮助存活的阶段，开始抢占英特尔的份额。而在 ARM 架构处理器上，苹果依靠着 M 系列芯片出色的设计、台积电先进工艺的加持以及强大的生态环境等优势，稳定的占据 ARM 架构 PC 的份额。虽然，在全球所有 PC 设备处理器市场中，英特尔依旧保持着领先地位，72%左右的市占率虽然相比之前有所下滑，但依旧遥遥领先份额缓慢上升的 AMD（约 16%份额）。但在 AI 笔电领域，份额下滑到约 50%的英特尔将会遇到更多竞争。

图表45：全球 PC 处理器份额



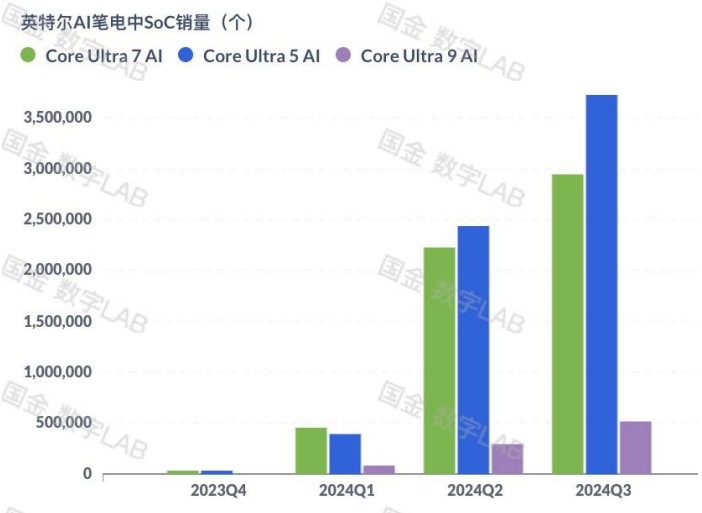
来源：IDC、国金数字未来实验室、国金证券研究所

在 X86 与 ARM 的竞争中，虽然 ARM 功耗低带来续航的优势，但 X86 架构下好用的各种 APP 为 X86 构建起了软件护城河。即使苹果在努力扩大自己的软件生态，同时提升 ARM 转译 X86 能力，短时间内仍不足以在 MacOS 跑通全部应用，X86 架构将继续享有软件生态上的优势。在 X86 生态内，AMD 是英特尔最大的对手。凭借着 AI 笔电中 SoC 优异的性能，以及反超的发布规划，AMD 正走在抢占英特尔 AI 笔电份额的路上。

近几个月，英特尔、AMD 分别推出 Ultra Lake 和 Ryzen AI 系列芯片，两家 SoC 厂商在苹果以外的 X86 架构 AI 笔电市场展开了竞争。AMD 在芯片推出时间上占据了优势，但随着英特尔采用台积电 3nm 先进工艺，英特尔 Ultra 200s 系列芯片在功耗&性能上有所改善，抢占了大批 OEM 市场。高通 X Elite 系列芯片近期表现不佳，X Elite 核笔电的价格偏高的同时，性能却不如同为 ARM 架构的苹果。在适配性上，高通 X Elite 芯片在转译 32 位 X86 软件时性能下滑严重，同样也让想要良好软件生态的用户有所顾虑。

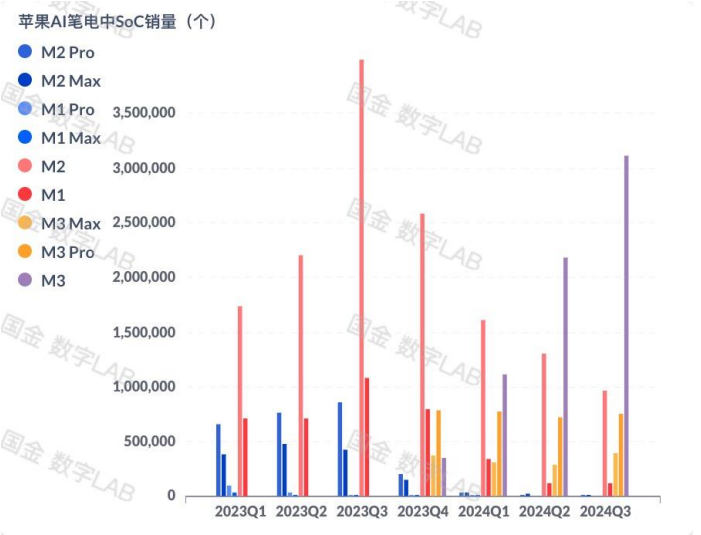
英特尔 Lunar Lake 系列芯片在 24Q2&24Q3 都有不错的销量，四季度英特尔发布了新一代 Ultra 架构芯片，在功耗上有明显提升，我们认为英特尔核 AI 笔电四季度的销量将会继续增长。苹果凭借 M3 芯片的强大性能维持销量的稳定，我们认为四季度 M4 核 AI PC 推出后，苹果销量将继续增长。AMD Ryzen 300 AI 系列在 8 月正式发布，在芯片性能表现优异的情况下我们认为 4 季度 Ryzen 300 AI 芯片的销量占比将会提高。高通由于 X Elite 在性能&价格等方面的劣势暂时销量较低。

图表46：英特尔核 AI 笔电中 SoC 销量（个）



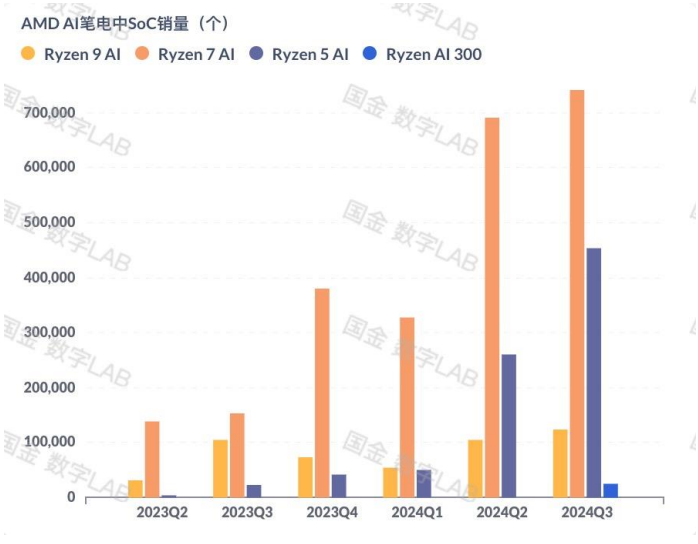
来源：IDC、国金数字未来实验室、国金证券研究所

图表47：苹果核 AI 笔电中 SoC 销量（个）



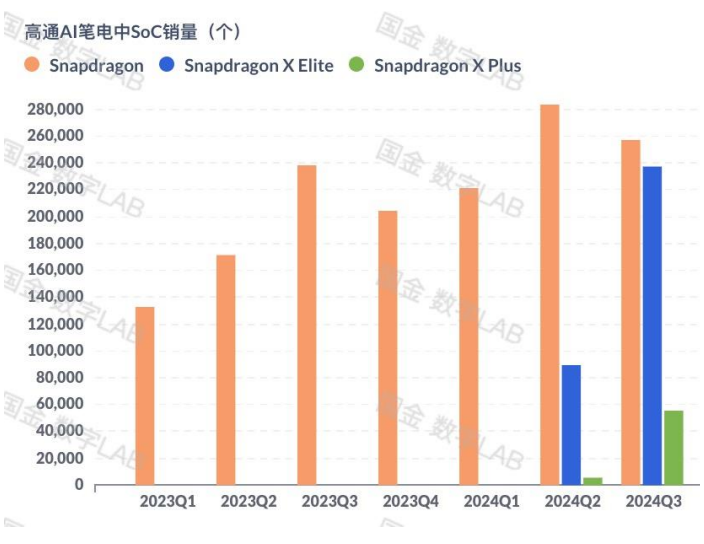
来源：IDC、国金数字未来实验室、国金证券研究所

图表48：AMD 核 AI 笔电中 SoC 销量（个）



来源：IDC、国金数字未来实验室、国金证券研究所

图表49：高通核 AI 笔电中 SoC 销量（个）



来源：IDC、国金数字未来实验室、国金证券研究所

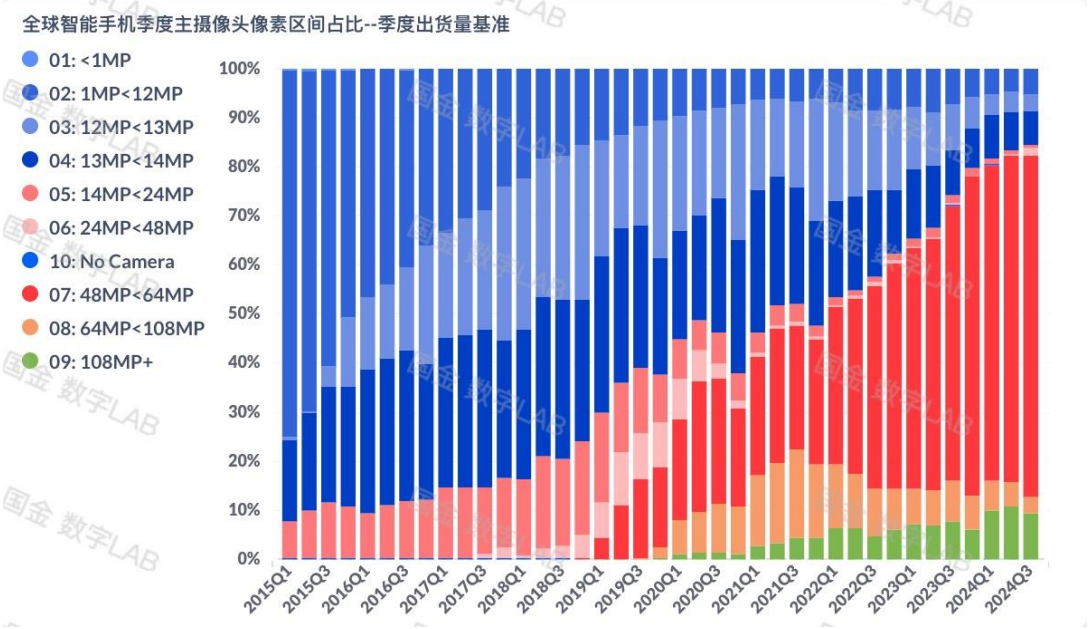
我们认为未来 X86 笔电市场竞争将会更为激烈，英特尔和 AMD 产品在性能、续航、适配性、生态方面各分秋色。而在 X86 台式机领域，由于功耗的重要性大幅降低，AMD 的 CPU 性能更为出色使得用户更偏向于采用 AMD CPU 的个人台式机或者工作站电脑。在 ARM 领域，苹果的优势更为明显。高通 X Elite 目前性能仅与苹果 M1、M2 芯片类似，同时在生态上远远落后于苹果及 X86 架构对手。短时间内高通很难与苹果竞争 ARM 架构 AI 笔电的市场。ARM 市场除了苹果与高通外，联发科&英伟达合作开发的 AI PC 也将于 2025 年进入市场。据报道，这款 3nm AI CPU 将结合联发科的 CPU 技术与 NVIDIA 的 GPU 技术，旨在实现高效的图形处理与 AI 计算能力。该芯片的流片阶段预计已在 24 年年 10 月正式展开，英伟达的 AI PC 芯片预计将在 2025 年下半年首次亮相市场，并在 2026 年进入商业生产阶段。短期内，英伟达与联发科合作的 AI PC 处理器受制于生态以及产品经验不足，很难与苹果竞争，但长期来看，英伟达、联发科合作的新片性能值得期待。

5.3 AI 设备产业链随着 AI 加入将迎来更新换代

除了处理器之外，在 AI 设备产业链中，我们认为散热、光学、OLED 和内存都是确定性较强的产业机会。即使现在最先进处理器的制程达到了 3nm，但在本地推理的计算压力和高频内存的读取压力下，芯片散热依旧是需要解决的大问题。苹果在 iPhone 16 上首次增加了均热板来解决问题。虽然第一次添加的均热板面积不够大，也没有像安卓系旗舰那样通过多种导热材料散热，但从零到一表现出了苹果对散热问题的重视，同时也为供应链新增了增量。

在光学领域，我们认为随着大模型多模态性能不断提升，作为输入口的摄像头及 CMOS 模组有更新升级的机会。24 年三季度 4800 万像素以上的手机占比已经超过了 80%，我们认为手机摄像头还会继续迭代向 6400 万以上像素。除此之外，LiDar、ToF 等激光传感器或光学镜头也会随着 AI 眼镜、AR/VR 设备渗透率提升而放量。

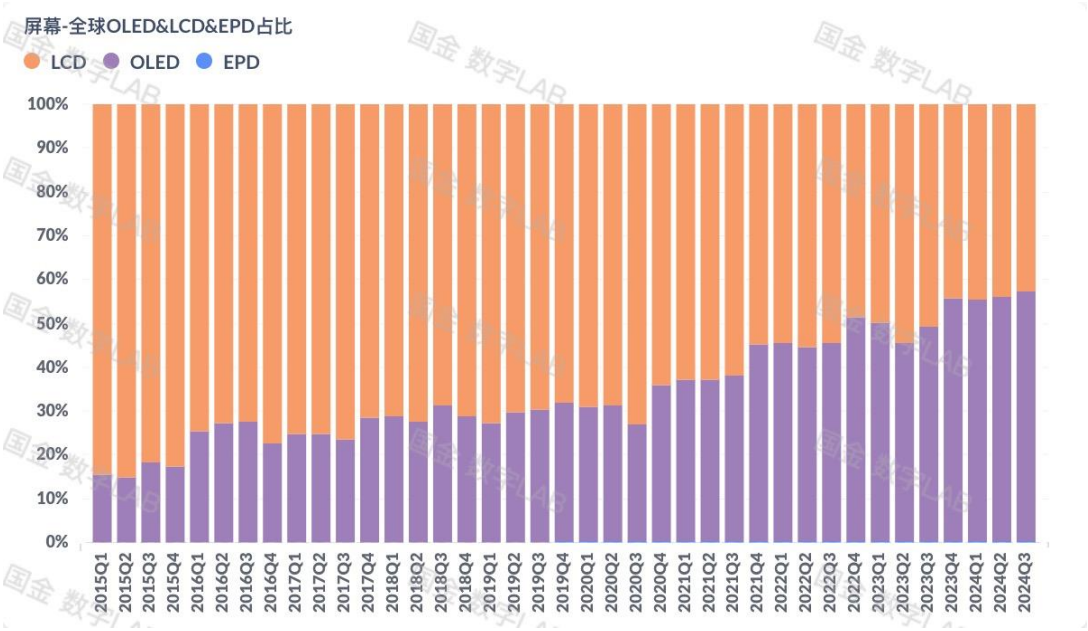
图表50：全球智能手机季度主摄像头像素区间占比



来源：IDC、国金数字未来实验室、国金证券研究所

OLED 屏幕凭借其更好的显示效果、响应速度快、功耗低等优势击败 LCD 成为主流旗舰机的标配。而随着人们对旗舰机的接受程度更高，OLED 渗透率也在不断提升。目前，在一些更追求显示效果的场景中，比如苹果 Vision Pro，即使 Micro OLED 目前成本极高，但其凭借其出色的显示效果成为了显示屏的选择。等到 Micro OLED 技术更为成熟，我们认为 OLED 屏幕将迎来新一轮产品结构优化。

图表51：全球手机 OLED&LCD&EPD 屏幕占比



来源：IDC、国金数字未来实验室、国金证券研究所

DRAM 市场经过短期整固并且低端产品逐渐淘汰后，会继续强势。目前，由于手机端侧模型的部署的推迟，手机内存大小虽然在旗舰机中普遍有所上调，但涨幅并不算大。我们认为在 AI 手机端云结合的理想模式下，内存的带宽的要求带来 DDR4 向 DDR5 以及 DDR6 提升，同时内存的频率及大小也会提升。由于体积&功耗的限制，手机很难将内存提升到类似 Macbook 128G 内存的水平，但为满足大模型性能的提升，16G 至 32G 的内存在未来是必须的。

六、智能驾驶&机器人行业正在摸索技术路径

6.1 智能驾驶：模块化方案与端到端方案之争

在智能驾驶领域，技术路径的选择不仅关乎技术发展的方向，还深刻影响商业化落地的进程。目前，行业内主要存在两种技术路线：模块化方案和端到端方案。模块化方案以 Mobileye 和 Waymo 等企业为代表，而端到端方案则更多由特斯拉和一些前沿研究团队推动。

模块化方案将自动驾驶任务分解为多个独立模块，包括：1）感知：通过摄像头、激光雷达等传感器识别环境中的物体。2）定位：使用高精度地图和 GNSS 定位技术确定车辆的准确位置。3）决策规划：根据感知和定位的信息，制定行车路径。4）控制：将规划转化为车辆的具体动作。模块化方案通过这四步循序渐进完成自动驾驶。这种方案的优点是可解释、可验证、易调试，但缺点也很明显：传递过程中信息损耗、任务多且散导致低效、存在复合误差。这种方案中，Mobileye（摄像头+高精地图）和 Waymo（摄像头+激光雷达+毫米波雷达等多种传感器）走的最远，也和 Uber 和 Lyft 等大型出租车平台展开了 RoboTaxi 相关合作。24 年 11 月，Lyft 一次性宣布了三项 Robotaxi 相关合作，分别是与自动驾驶方案商 Mobileye、自动驾驶接驳车公司 May Mobility，以及基于 AI 的行车记录仪制造商 Nexar。根据计划，Lyft 将从 2025 年开始逐步引入上述公司的技术，以提供自动驾驶运送乘客的服务。

端到端方案基于统一的神经网络从原始传感器数据输入直接到控制指令输出的连续学习与决策过程,过程中不涉及任何显式的中间表示或人为设计的模块。特斯拉发布的 FSD V12 则是端到端方案。

我们认为目前两种方案在复杂场景和简单场景中各具优势，未来的方向可能是在简单环境下使用端到端方案，而在较为复杂的城市中心场景使用模块化方案。

从技术角度看，无论使用什么方案，车厂仍需一段时间才能达到 L4+级别的自动驾驶，而国内新能源汽车渗透率较高将一定程度上降低未来消费者对新能源汽车的需求。

图表52：2024 年周度汽车上险量（台）及渗透率

来源：国金数字未来实验室、国金证券研究所

6.2 具身智能想要放量需要更实用的场景及更低的价格

具身智能可以通过机器人或其他具身设备与真实世界进行交互，实现感知、决策和行动的闭环。这一领域结合了人工智能、机械设计和感知技术，赋予机器自主学习和适应复杂环境的能力，使其在无人配送、灾害救援、仓储物流等场景中大放异彩。目前包括波士顿动力、特斯拉、小米、宇树等厂商都积极投入。

我们认为国内龙头公司如宇树更容易放量。宇树在四足机器人领域实现了高集成度的小型化和模块化设计，使其机器人具有强大的灵活性和高性价比。相比于其他品牌，如波士顿动力的 Spot，宇树的产品不仅在成本上更具优势，而且在可拓展性上也非常突出。同时，宇树通过技术优化，将四足机器人价格拉低至几万元的级别，大幅降低了企业和个人的购买门槛，使得这一技术能够从实验室走向实际场景。低成本的实现并未牺牲性能，其产品依然能够应对复杂的地形环境和高强度任务，这为其在物流、安防和教育等市场开辟了广阔的应用空间。

风险提醒

1. 芯片制程发展与良率不及预期：半导体工艺的发展面临诸多挑战，主要包括技术瓶颈、良率提升难度、研发成本高企以及供应链不确定性等问题。随着工艺节点微缩变得愈发复杂，先进制程的实现难度和成本不断攀升，可能导致量产延迟，甚至影响产品性能和成本控制。此外，地缘政治风险和出口管制可能扰乱供应链，进一步拖累产能扩张。

2. 中美科技领域政策恶化：中美在 AI 领域竞争激烈，美国限制先进芯片和半导体对中国的出口，随着竞争的加剧，未来可能会推出更严格的限制政策，限制国内 AI 模型的发展。

3. 智能手机、PC 销量不及预期：智能手机销量与产品本身质量关系紧密，若产品本身有缺陷则智能手机销量可能收到影响。同时宏观经济变化也有可能导致消费者消费意愿发生变化从而影响智能手机销量。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究