东吴证券 SOOCHOW SECURITIES

电子行业深度报告

如何看待 AI 手机对端侧&云端算力影响？——算力需求看点系列

2025 年 02 月 15 日

增持（维持）

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
研究助理 李雅文
执业证书：S0600125020002
liyw@dwzq.com.cn

投资要点

- **端侧 AI 在手机场景中如何落地？**苹果引领 AI 手机市场，端云混合推动 AI 落地。2024 年 6 月，苹果在 WWDC24 上正式公布了 Apple Intelligence（苹果 AI），推出了一系列 AI 功能，包括：Siri 升级、文字处理、图像生成等，上述功能大多直接通过端侧运行，云端用于弥补端侧算力的不足。我们重点梳理了 Writing Tools（写作工具）的一些功能，根据测评发现，除了 Summary、Key Points、Table、List 这四个较为复杂的功能仅支持接入网络的场景下使用，其他 Writing Tools 功能均可以在无网络的场景下直接使用。

- **如何理解端侧 AI 对云端算力的依赖？**展望 AI 手机在算力上的布局，我们认为云端 or 端侧算力都是未来中短期不可或缺的存在。一方面，在成本、功耗和隐私性优势较大的情况下，算力从云端分流到终端运行或为大势所趋。但当前无论从 SoC 算力水平还是端侧模型性能来看，仍然还有较大的提升空间。我们认为端侧算力落地的可行路径有：（1）端侧 SoC 硬件不断升级支撑 AI 需求。（2）端侧小模型针对主流功能做定向优化减轻算力负载。另一方面，端侧/云端算力不是此消彼长的关系，我们认为端侧 AI 对算力总盘子的拉动作用会长期存在。当前从 0 到 1 阶段，用户更关注端侧 AI“有没有”或“效果好不好”，而不是“推理速度快不快”。而展望从 1 到 10 的阶段，推理速度一定会是重要的优化环节，我们认为在一些特定场景中云端推理仍然是不错的选择。

- **如何测算手机 AI 算力需求？**我们将测算拆分为累计 AI 手机用户数及单日单机算力需求两方面。（1）累计 AI 手机用户数：IDC 预计全球智能手机出货量稳定在 12-14 亿部/年，根据 Canalys 预测，2024/2028 年全球 AI 手机渗透率或将达到 16%/54%。我们进一步预计到 2030 年 AI 手机有望渗透到“千元机”价格带，由此 AI 手机渗透率或将达到 80%。（2）单日单机算力需求：根据“推理算力需求=2×参数量×token 数”的公式，进一步拆分端侧/云端算力来计算。（3）结论：我们测算得到端侧算力需求在 2024-2027 年间基本维持翻倍以上的增速，2027-2030 年间增速依然在高双位数水平。云端算力需求若折算成 Blackwell GPU 卡的 FP8 算力，2025/2026 年需求量约为 12/103 万张。

- **产业链相关公司**
  - 消费电子整机：立讯精密、歌尔股份、领益智造、蓝思科技等。
  - 端侧 SoC：翱捷科技、晶晨股份、瑞芯微、全志科技、恒玄科技等。
  - 云端算力链：工业富联、沪电股份、胜宏科技、寒武纪、海光信息（与计算机联合覆盖）、龙芯中科、盛科通信（与通信联合覆盖）等。

- **风险提示：** AI 手机出货量不及预期风险，端侧软硬件技术发展不及预期风险，AI 创新效果不及预期风险。

行业走势

相关研究

《关注 DeepSeek 推动 AI 应用带来的推理需求，利好国产设备》

2025-02-06

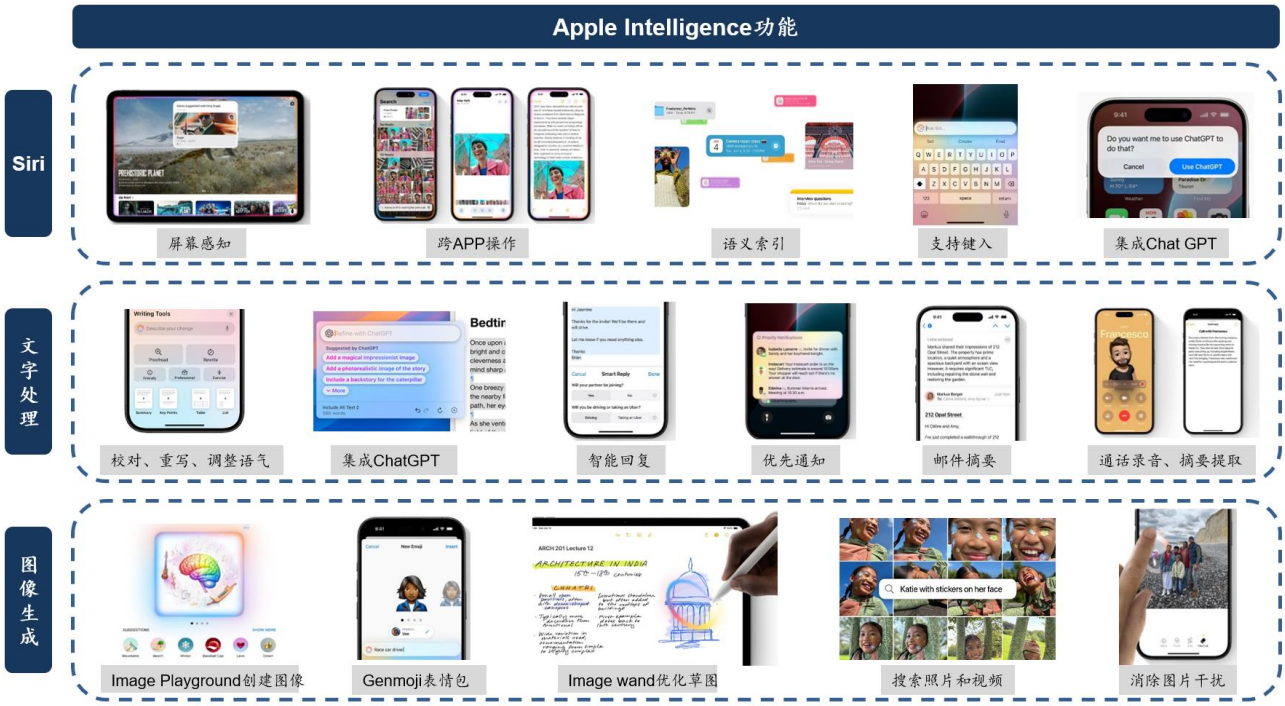
内容目录

图表目录

1. 端侧 AI 在手机场景中如何落地？

苹果引领 AI 手机市场，端云混合推动 AI 落地。2024 年 6 月，苹果在 WWDC24 上正式公布了 Apple Intelligence（苹果 AI），推出了一系列 AI 功能：**（1）Siri 升级：**Siri 实现了更丰富的语言理解能力和个人情境的感知能力，支持跨 APP 操作、语义索引、直接键入等功能。**（2）文字处理：**苹果 AI 推出了系统级写作工具，支持用户在邮件、备忘录、Pages 等第三方应用中，实现校对、重写、调整措辞语气等功能；在文字汇总方面，实现了快速提取录音、邮件的摘要；此外，还能够帮助用户自动管理消息通知。**（3）图像生成：**苹果 AI 实现了 Image Playground 文生图、Genmoji 表情、Image wand 优化草图、自定义记忆影片和消除图片干扰等功能。值得注意的是，**上述功能大多直接通过端侧运行，云端用于弥补端侧算力的不足。**苹果 AI 中使用了两种基础大语言模型：一个约 30 亿参数，用于设备端运行的 AFM-on-device；一个更大型的，基于服务器的 AFM-server。在用户发出指令后，苹果 AI 会分析用户发出的请求，如果用户需要使用的模型已经超出了端侧 AI 的能力上限，这时云端 AI 才会介入。此外，苹果在近日新上线的 IOS 18.2 Beta1 中，已经实现接入 OpenAI，将 ChatGPT 集成到 Siri 和写作工具中，用户可以切换使用 GPT-4o 以调用性能更强的云端模型。

图1：苹果 AI 功能的应用情况

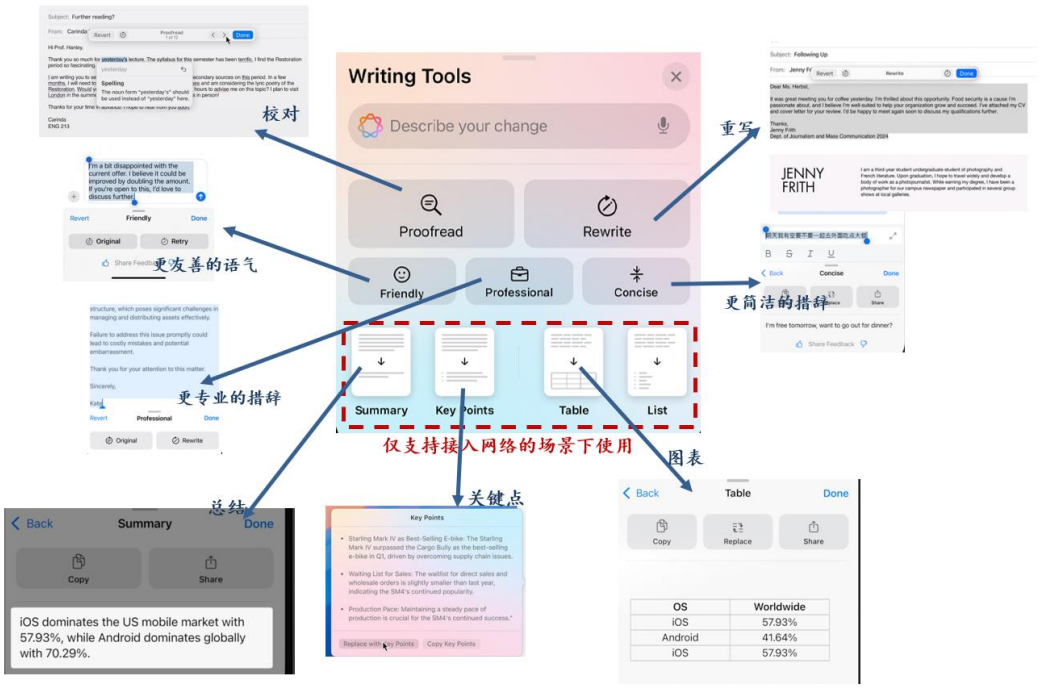


数据来源：苹果，东吴证券研究所

在这里，我们重点介绍 **Writing Tools（写作工具）**的一些功能。包括校对文本、重写不同版本，使得语气措辞恰到好处；并支持对文本进行总结、关键点提取、图表制作、分点总结的功能。根据我们的测评发现，除了 Summary、Key Points、Table、List 这四个较为复杂的功能仅支持接入网络的场景下使用，其他 Writing Tools 功能均可以在无网

络的场景下直接使用。

图2：Writing Tools 功能介绍



数据来源：苹果，差评 X.PIN，东吴证券研究所

2. 如何理解端侧 AI 对云端算力的依赖？

2.1. 端侧运行 AI 优势明显，AI 处理持续向端侧转移

端侧运行 AI 模型，具有成本、能耗、性能、隐私和安全、个性化五大优势。

图3：端侧运行 AI 的五大优势

优势	具体内容
成本	随着生成式AI模型使用量和复杂性的不断增长，数据中心基础设施的成本持续增加，使得云端推理的成本急剧提高。虽然生成式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）的成本是传统搜索方法的10倍。将一些处理从云端转移到端侧，充分利用数10亿级的端侧碎片化算力，可以支持开发者基于端侧算力开发应用程序，减轻云基础设施的压力并减少开支。
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。
性能	云端运行对网络要求较高，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。而端侧AI可以在本地离线运行，具有更高的可靠性，提供媲美云端甚至更佳的性能。
隐私和安全	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。此外，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的安全。
个性化	端侧AI可以在不侵犯用户隐私的情况下,根据用户独特的语音模式、表情、反应、使用模式、环境等，甚至外部数据（例如来自健身追踪器或医疗设备的数据）形成精准的用户画像，并且随着时间推移继续进行学习和演进，从而提供个性化服务

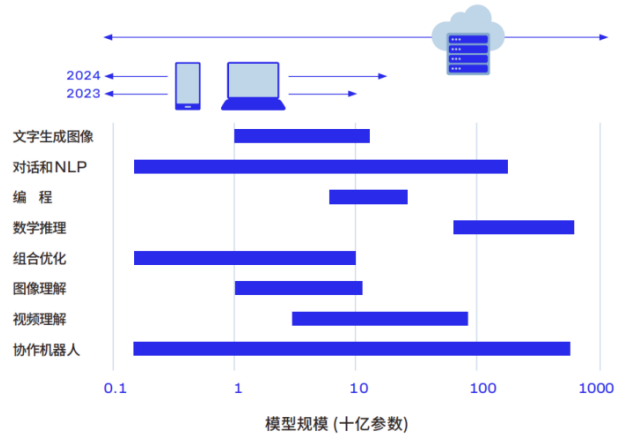
数据来源：高通，东吴证券研究所

考虑到端侧运行 AI 具有明显优势，越来越多的生成式 AI 模型从云端分流到终端运行。事实上，在生成式 AI 出现之前，AI 处理便持续向端侧转移，越来越多的 AI 推理工作在手机、笔记本电脑、XR 头显、汽车等边缘终端上运行。例如，手机利用端侧 AI 支持许多日常功能，比如暗光拍摄、降噪和人脸解锁。随着生成式 AI 模型的缩小，终端算力的提升，能在终端运行的生成式 AI 模型更加多样。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，利用大规模端侧 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。

图4：AI 处理的重心向端侧转移

数据来源：高通，东吴证券研究所

图5：生成式 AI 模型持续从云端分流到终端运行



数据来源：高通，东吴证券研究所

2.2. 受限于算力制约，端侧 AI 仍需依赖云端算力

端侧 AI 处理能力是赋能混合 AI 并让生成式 AI 实现全球规模化扩展的关键。随着大量的工作负载正从云端转向边缘终端，对于端侧处理的高性能和出色能效需求愈发增长。如何在云端和端侧之间分配处理任务将取决于端侧能力、隐私和安全需求、性能需求以及商业模式等诸多因素。下文将从端侧 AI 的算力供给与算力需求两个维度进行梳理，分析当前端侧/云端 AI 所适合的应用场景。

2.2.1. 高通、联发科发布旗舰 SoC，端侧 AI 算力不断增强

高通、联发科作为 SoC 芯片制造领域的领军企业，陆续发布全新旗舰 SoC。2024 年 10 月 9 日，联发科发布旗舰 5G 智能体 AI SoC 芯片——**天玑 9400。（1）工艺方面：**采用台积电第二代 3nm 工艺，较上代同性能功耗降低 40%。**（2）架构方面：**采用第二代全大核 CPU 架构，包含 1 个主频 3.62GHz 的 Cortex-X925 超大核，以及 3 个 Cortex-X4 超大核和 4 个 Cortex-A720 大核。值得注意的是，Cortex-X925 与新一代旗舰 12 核 GPU Immortalis-G925 相得益彰，能够提供卓越的图形性能和效率。**（3）NPU 方面：**搭载了第八代 AI 处理器 NPU 890，首发支持端侧 LoRA 训练和端侧视频生成，支持端侧运行 Meta Llama3.2 的 1B 和 3B 模型，多模态 AI 运算处理速度至高达 50 tokens/s，至高 32K tokens 文本长度；相较于上一代产品，天玑 9400 的 LLM 提示词处理能力提升了 80%。

2024 年 10 月 22 日，高通正式发布新一代骁龙旗舰 SoC——**骁龙 8 Elite。（1）工艺方面：**采用台积电的第二代 3nm 工艺制程。**（2）架构方面：**搭载第二代定制的高通 Oryon CPU，采用 2+6 的架构方案，2 颗 4.32GHz 的超大核，6 颗 3.53GHz 大核，相比上代骁龙 8 Gen3 的 3.3GHz 提升极为明显。**（3）NPU 方面：**骁龙 8 至尊版搭载了最新的 Hexagon NPU，该 NPU 中搭载了 6 核向量处理器和 8 核标量处理器，支持端侧多模态，其出词速度达到了 70 tokens/s 以上，支持 4k 上下文窗口。

图6：主流手机 SoC 性能梳理

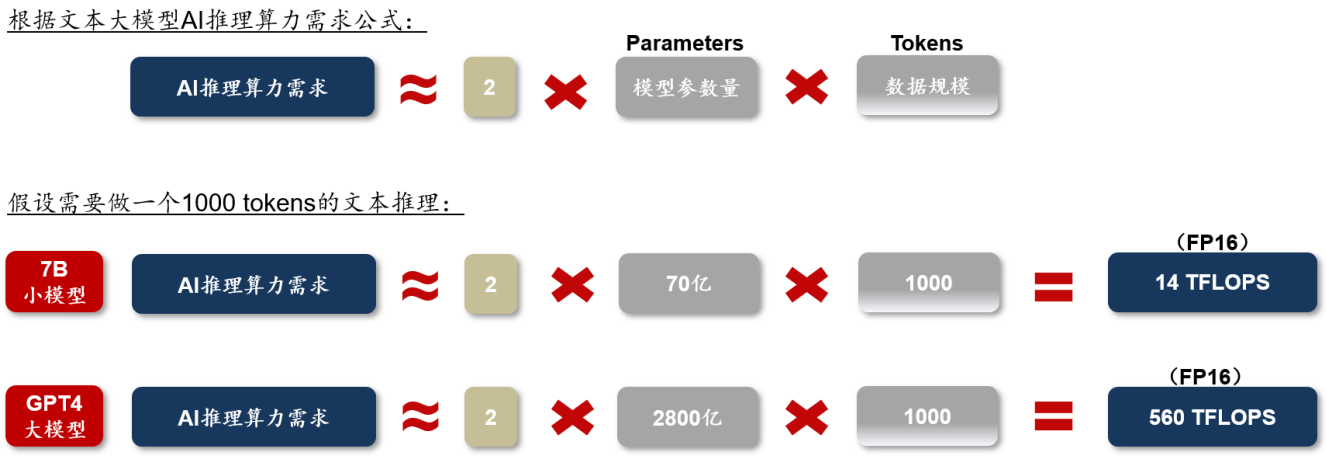
芯片型号	骁龙8Gen3	骁龙8至尊版	天玑9300	天玑9400	A17 Pro	A18 Pro
芯片厂商	高通	高通	联发科	联发科	苹果	苹果
工艺制程	4nm	3nm	4nm	3nm	3nm	3nm
CPU架构	单核X4+五核A720+双核A520	两个超级内核+六个性能内核	四核X4+四核A720	单核X925+三核X4+四核A720	六核(2+4)	六核(2+4)
CPU核心频率	3.3+3.2+2.3GHz	4.32+3.53GHz	3.25+2.0GHz	3.62+3.3+2.4GHz	3.78+2.11GHz	4.04+2.2GHz
GPU	Adreno 750	Adreno	Immortalis-G720 MC12	Immortalis-G925 MC12	自研六核心	自研六核心
GPU核心频率	903MHz		1300MHz		1398MHz	1398MHz
NPU	34 TOPS		33 TOPS		35 TOPS	35 TOPS
存储	LPDDR5X-4800	LPDDR5X-5300	LPDDR5T-9600	LPDDR5X-10667	LPDDR5-6400	LPDDR5X-7500
存储容量	24GB		24GB		8GB	
基带	骁龙X75 5G 10Gbps/3.5Gbps	骁龙X80 5G 10Gbps/3.5Gbps	5G 7Gbps	5G 7Gbps	骁龙X70	骁龙X75
出货时间	2023Q4	2024Q4	2023Q4	2024Q3	2023-09	2024-09
代表机型	小米14/14 Pro/14 Ultra/MIX Flip/MIX Fold 4、iQOO 12/12 Pro/Neo9S Pro+/Neo10、荣耀Magic6/Magic 6 Pro/Magic 6至臻版/荣耀Magic 6 RSR保时捷设计/荣耀Magic V3/荣耀GT/荣耀300 Pro/荣耀300 Ultra、OPPO Find X7 Ultra、一加12/Ace 3 Pro/Ace 5、realme GT5 Pro/GT6、Redmi K70 Pro/K80、魅族21/21 Pro、ROG游戏手机8/8 Pro、努比亚Z60 Ultra/Z60 Ultra领先版、三星Galaxy S24/S24+/S24 Ultra/Galaxy Z Flip6/Galaxy Z Fold6/W25/W25 Flip、索尼Xperia 1 VI、vivo X100 Ultra/X Fold3 Pro、Polestar Phone、华硕Zenfone 11 Ultra、蔚来NIO Phone、红魔9S Pro(领先版)、LynkCo Phone Pro	小米15/小米15 Pro/小米15 Ultra、REDMI K80 Pro、ROG游戏手机9/9 Pro、红魔10 Pro、一加13/一加Ace 5 Pro、努比亚Z70 Ultra、realme GT7 Pro、iQOO 13、荣耀Magic7/Magic7 Pro/Magic7 RSR保时捷设计	vivo X100、vivo X100 Pro、OPPO Find X7、iQOO Neo9 Pro	vivo X200、vivo X200 Pro、vivo X200 Pro mini、OPPO Find X8、OPPO Find X8 Pro、iQOO Neo10 Pro	iPhone 15 Pro、iPhone 15 Pro Max	iPhone 16 Pro、iPhone 16 Pro Max

数据来源：快科技，CPU Monkey，芯参数网，东吴证券研究所

2.2.2. 云端、端侧算力仍存在差距，复杂场景下云端大模型更为适宜

对于云端大参数模型而言，AI 手机端侧算力较难支持。根据 OpenAI《Scaling Laws for Neural Language Models》论文中“算力需求=2×参数量×token 数”的公式，我们很容易能够给出一个简单的测算来感受端侧和云端算力需求的显著差距。如下图所示，假设需要对一个 1000 tokens（不同的大模型均有各自的分词器设计，以 OpenAI 为例，1000 个 token 通常代表 750 个英文单词或 500 个汉字）的文本进行推理：（1）目前主流的 7B 小模型，算力需求大约为 14TFLOPS；（2）GPT4 大模型训练参数量 1.8 万亿，推理时激活参数 280B，算力需求大约为 560TFLOPS。根据 IDC 定义，AI 手机是 NPU 算力大于 30 TOPS（INT8）、搭载支持生成式 AI 的 SoC 并支持端侧大模型的手机。然而，截至 2024 年 2 月，符合 IDC 要求的 SoC 只有苹果 A17 Pro、联发科天玑 9300 和高通骁龙 8Gen3。此外，最新发布的天玑 9400 和骁龙 8 至尊版也符合 IDC 定义的 AI 手机芯片，但目前端侧 AI 能实现的算力仍远远不足以支持云端大模型的推理。

图7：端侧、云端算力需求的差距



数据来源：OpenAI《Scaling Laws for Neural Language Models》，华尔街见闻，思瀚产业研究院，OpenAI Platform，ittbank，大脑助理，腾讯科技，东吴证券研究所

虽然轻量化的小模型可以满足在端侧运行，但其效果较云端大模型仍有较大差距。根据 SuperCLUE 2024 年 12 月模型榜单，云端大模型的得分明显高于端侧小模型，传统大模型使用长文本时，会把整个上下文都放进模型的输入中，而大型的计算开销会因为输入的提升而极速上升，尤其在端侧算力有限的场景下，会对性能产生制约。

图8：大模型性能总排行榜（2024 年 12 月）

排名	模型名称	机构	总分	Hard	理科	文科	使用方式	发布日期
-	o1	OpenAI	80.4	76.7	87.3	77.1	网页	2025年1月8日
-	o1-preview	OpenAI	74.2	63.6	80.6	78.5	API	2025年1月8日
-	ChatGPT-4o-latest	OpenAI	70.2	57.8	72.1	80.7	API	2025年1月8日
🥇	DeepSeek-V3	深度求索	68.3	54.8	72	78.2	API	2025年1月8日
🥇	SenseChat 5.5-latest	商汤	68.3	51.5	71.6	81.8	API	2025年1月8日
-	Gemini-2.0-Flash-Exp	Google	68.2	55.5	72.6	76.6	API	2025年1月8日
-	Claude 3.5 Sonnet(20241022)	Anthropic	67.7	54.6	71.4	77.2	API	2025年1月8日
🥇	360zhinao2-o1	360	67.4	51.4	72.1	78.7	API	2025年1月8日
🥈	Doubao-pro-32k-241215	字节跳动	66.5	50.6	72.3	76.6	API	2025年1月8日
🥈	NebulaCoder-V5	中兴通讯	66.4	48.6	69.5	80.9	API	2025年1月8日
🥈	Qwen-max-latest	阿里巴巴	66.2	51.3	67.4	80	API	2025年1月8日
-	Qwen2.5-72B-Instruct	阿里巴巴	65.4	49.7	66.2	80.3	API	2025年1月8日
🥉	Step-2-16k	阶跃星辰	65.2	50	65.1	80.3	API	2025年1月8日
🥉	GLM-4-Plus	智谱AI	65.1	48.5	68.1	78.8	API	2025年1月8日
-	Grok-2-1212	X.AI	63.9	49.2	66.8	75.5	API	2025年1月8日

数据来源：Super CLUE，东吴证券研究所

图9：10B 小模型性能排行榜（2024 年 12 月）

排名	模型名称	机构	参数量	总分	理科	文科	Hard	参数量.1	使用方式	发布日期
🥇	Qwen2.5-7B-Instruct	阿里巴巴	70亿	55.5	54.4	76.4	35.7	70亿	API	2025年1月8日
🥈	GLM-4-9B-Chat	智谱AI	90亿	52.4	50.6	75.1	31.6	90亿	模型	2025年1月8日
-	Gemma-2-9b-it	Google	90亿	48.6	49.5	73.7	22.7	90亿	模型	2025年1月8日
🥉	360Zhinao2-7B-Chat-4K	360	70亿	47.8	50.7	75.2	17.5	70亿	模型	2025年1月8日
4	Qwen2.5-3B-Instruct	阿里巴巴	30亿	46.1	44.2	75.5	18.6	30亿	API	2025年1月8日
5	Yi-1.5-9B-Chat-16K	零一万物	90亿	44.3	41.3	71.3	20.3	90亿	模型	2025年1月8日
5	MiniCPM3-4B	面壁智能	40亿	44.2	45.9	73	13.7	40亿	模型	2025年1月8日
-	Llama-3.1-8B-Instruct	Meta	80亿	43.9	42.8	68.1	20.9	80亿	API	2025年1月8日
-	Phi-3.5-Mini-Instruct	微软	38亿	42.4	42.4	70.7	14	38亿	模型	2025年1月8日
-	Gemma-2-2b-it	Google	20亿	39.2	36.4	69.4	11.8	20亿	模型	2025年1月8日
-	Mistral-7B-Instruct-v0.3	Mistral AI	70亿	33.2	31.2	56.9	11.4	70亿	模型	2025年1月8日

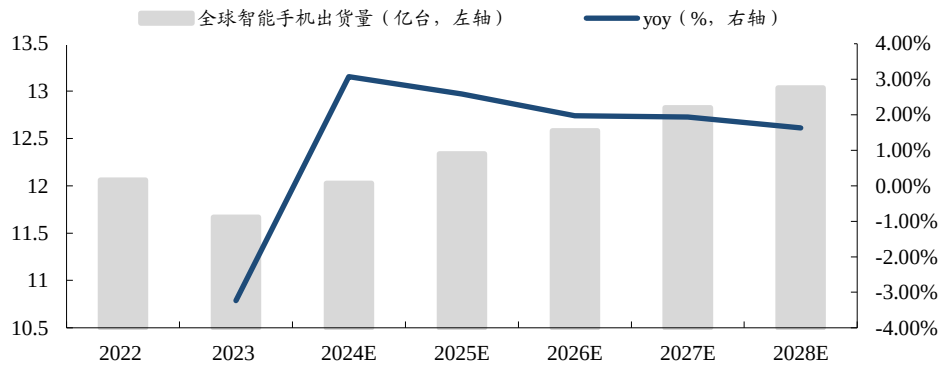
数据来源：Super CLUE，东吴证券研究所

3. 如何测算手机 AI 算力需求？

3.1. 累计 AI 手机用户数如何定义？

全球 AI 手机年出货量：我们采用全球智能手机出货量×AI 手机渗透率的方式进行测算，其中（1）全球智能手机出货量：根据下图数据，IDC 预计全球智能手机出货量稳定在 12-14 亿部/年。（2）AI 手机渗透率：根据 Canalys 对具有生成式 AI 能力智能手机市场的预测，2024 年 AI 手机出货量预计占全球智能手机出货量的 16%，到 2028 年，这一比例将激增至 54%。从 2023 年到 2028 年，AI 手机市场年均复合增长率（CAGR）将达到 63%。往更长期看，我们认为到 2030 年 AI 手机有望渗透到“千元机”价格带，根据 Bloomberg 转引 IDC 数据可知，截至 24Q2，$150 以下的智能手机约占全球智能手机市场的 25%-30%，我们预计随着 AI 手机的不断渗透，AI 手机价格带有望穿越“千元机”水平进一步下探，由此我们预计到 2030 年 AI 手机渗透率或将达到 80%。

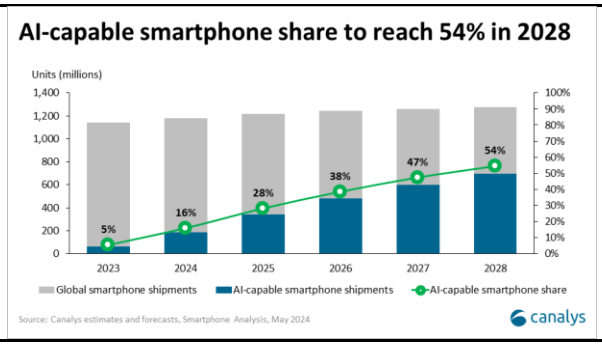
图10：全球智能手机出货量及预测



数据来源：IDC，钛媒体 APP，CNMO 手机中国，东吴证券研究所

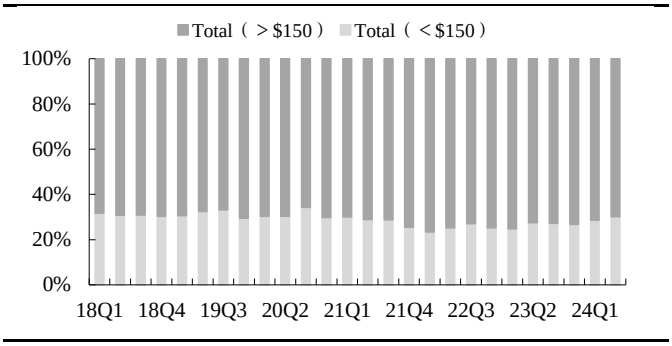
累计 AI 手机用户数：根据 TechInsights，全球智能手机的换机率为 23.8%，周期为 51 个月。我们假设由于 AI 手机对智能手机硬件要求会随着 AI 功能的升级换代而不断提升，AI 手机换机周期会在此基础上有所缩短，假设换机周期为 4 年。由此，我们测算得到 AI 手机存量用户数有望在 2024-2026 年期间翻倍式增长，在 2027-2030 年间或将逐年放缓。

图11：AI 手机渗透率情况



数据来源：Canalys，东吴证券研究所

图12：全球智能手机价格带情况



数据来源：IDC，Bloomberg，东吴证券研究所

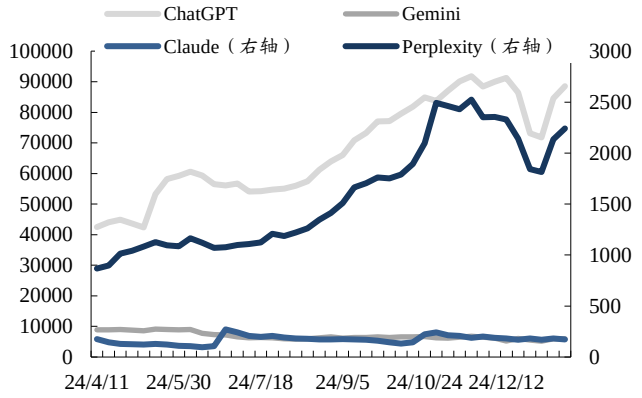
3.2. AI 手机文本推理需求与端侧模型、端侧硬件的关系

对于使用场景的划分：根据本文第一章对苹果 Apple Intelligence 功能的归纳总结（尤其是对 Writing Tools 的理解），我们发现目前 AI 手机对于文本的处理功能比较丰富，由此我们进一步将当前的推理任务划分为简单功能和复杂功能。由于当前的复杂功能主要为 Summary、Key Points、List、Table 等，我们倾向于认为以上复杂功能正常情况下会用于长文本的处理（经测评发现，以上场景需接入网络，由此我们判断需接入云端大模型来实现以上功能）；而简单功能更适用于短文本的处理（经测评发现，以上场景不需要接入网络）。我们认为苹果对于全球手机行业都具备一定的示范效应，且以上功能基本也可以覆盖当前手机文本推理的大多数使用场景，由此我们近似认为全球 AI 手机均有望具备以上类似功能。

我们判断上述**简单功能**有望在 **7B 端侧小模型**上实现。尽管苹果为设备端储备了 3B 小模型，但据腾讯研究院，当前主流的端侧小模型的参数体量为 7B。由此我们近似假设端侧简单功能主要是在 7B 小模型上实现。未来随着 AI 大模型行业发展，端侧小模型有望从 7B 向 13B 逐渐渗透。

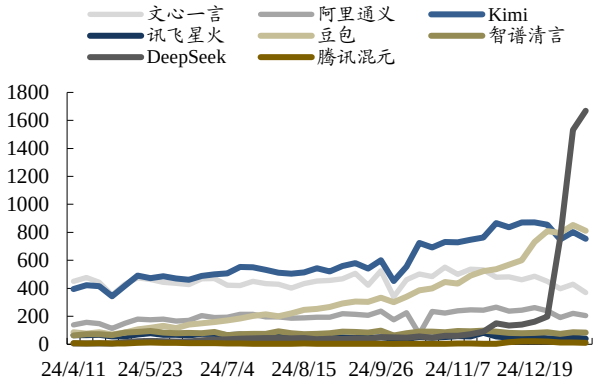
我们判断上述**复杂功能**有望通过**外接云端大模型**实现。根据我们对各 AI 大模型用户访问量的统计，发现当前主流的 AI 推理需求主要来自对 ChatGPT 的访问，因此我们的测算以 ChatGPT 为例，当前 GPT-4 的参数量为 1.8 万亿。**但站在长期视角，我们判断上述复杂功能有望逐渐被端侧算力消化。**随着端侧硬件的不断升级，以及端侧小模型对于细分功能的不断强化，我们假设在端侧硬件上运行复杂推理功能或将于 2027 年开始逐步承接，在 2027-2030 年间加深渗透程度。

图13：文本大模型网站访问量周度数据（单位：万次）



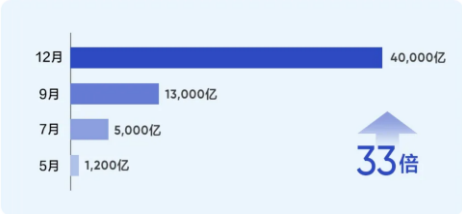
数据来源：Similarweb，东吴证券研究所

图14：文本大模型网站访问量周度数据（单位：万次）



数据来源：Similarweb，东吴证券研究所

图15：豆包大模型 2024 年日均 tokens 使用量



数据来源：火山引擎，东吴证券研究所

图16：2024/10-12 月豆包大模型各应用场景调用量

数据来源：火山引擎，东吴证券研究所
注：截至 2024/12/18

3.3. 单日单机算力需求（文本）测算

根据 OpenAI《Scaling Laws for Neural Language Models》论文，我们得到“推理算力需求=2×参数量×token 数”的公式，即每参数每 token 推理所需计算量为 2 Flops。（1）参数量：相关假设详见第 3.2 节。（2）token 数：我们将 token 数进一步拆分为单次推理需求对应字数×单日调用次数×单个字对应的 token 数（为方便理解，我们将测算放在中文场景下假设）。若按一屏所示文字数（500 个汉字）作为基准，我们假设一次简单功能对应的文本推理需求约为 300 个汉字，一次复杂功能对应的文本推理需求约为 1000 个汉字。单日调用次数随着 AI 功能的不断创新升级而逐年增加。

最终，基于以上假设，我们实际测算过程如下表所示。我们测算得到**（1）端侧算力需求：**在 2024-2027 年间基本维持翻倍以上的增速，2027-2030 年间增速依然在高双位数水平。**（2）云端算力需求：**假设将 2025-2026 年 AI 手机产生的云端算力均折算成 Blackwell GPU 卡的 FP8 算力；假设算力利用率（Flops 利用率）约为 50%；假设在算力峰值情况下，50%的算力需求集中在 5 个小时（波峰时段）内释放，以此作为算力部署需求测算。由此得到，2025/2026 年端侧算力对 Blackwell GPU 卡的需求量约为 12/103 万张。

图17：手机 AI 算力需求（文本）测算

	单位	2023	2024E	2025E	2026E	2027E	2028E	2029E	2030E
单日全球AI手机算力需求 | 文本									
端侧算力需求	**亿TFlops**		**457**	**1760**	**5866**	**15541**	**26934**	**45236**	**64855**
yoy				**285%**	**233%**	**165%**	**73%**	**68%**	**43%**
云端算力需求	**亿TFlops**			**214657**	**1856530**	**5244942**	**9202494**	**12880665**	**17991043**
yoy					**765%**	**183%**	**75%**	**40%**	**40%**
Blackwell单卡FP8算力	TFlops			10000	10000				
算力利用率				50%	50%				
Blackwell需求量	万张			12	103				
AI手机	**亿部**								
累计AI手机用户数		**0.6**	**2.5**	**6.0**	**10.7**	**16.2**	**21.3**	**26.6**	**32.7**
全球AI手机出货量		0.6	1.9	3.5	4.8	6.0	7.0	8.8	10.8
全球智能手机出货量		12	12	12	13	13	13	13	14
AI手机渗透率		5%	16%	28%	38%	47%	54%	66%	80%
算力需求 | 文本									
简单功能									
单日调用次数	次		20	30	50	70	80	90	100
调用大模型参数									
7B			90%	80%	65%	50%	35%	20%	10%
13B			10%	20%	35%	50%	65%	80%	90%
单次调用算力	TFlops								
7B			8	8	8	8	8	8	8
13B			16	16	16	16	16	16	16
复杂功能									
单日调用次数	次					3	5	9	10
调用大模型参数									
7B			90%	80%	65%	50%	35%	20%	10%
13B			10%	20%	35%	50%	65%	80%	90%
单次调用算力	TFlops								
7B						28	28	28	28
13B						52	52	52	52
外接大模型（以ChatGPT为例）									
单日调用次数	次			5	8	6	5	4	3.5
单次调用算力	**TFlops**			**7200**	**21600**	**54000**	**86400**	**120960**	**157248**
参数量	亿			18000	54000	135000	216000	302400	393120

数据来源：IDC，钛媒体 APP，CNMO 手机中国，Canalys，Bloomberg，TechInsights，苹果，英伟达，OpenAI《Scaling Laws for Neural Language Models》，腾讯研究院，腾讯科技，大脑助理，东吴证券研究所

注 1：图表中的算力单位统一按 TFlops 表示，仅用于直观对比算力体量，实际应用中端侧算力以 TOPS 为主要单位

注 2：假设简单功能推理需求对于文本量约为 300 字，复杂功能约为 1000 字

3.4. 未来展望

展望 AI 手机在算力上的布局，我们认为云端 or 端侧算力都是未来中短期不可或缺的存在。从长期趋势上看，我们主要探讨以下两个问题：

1、在成本、功耗和隐私性优势较大的情况下，算力从云端分流到终端运行几乎成为大势所趋。但当前软硬件基础设施能力有限，无论从 SoC 算力水平还是端侧模型性能来看，仍然还有较大的提升空间。展望端侧算力落地的可行路径，我们认为有两种：

路径#1 端侧 SoC 硬件不断升级，通过足够强大的本地运算能力支撑 AI 需求。根据前文梳理，我们发现当前只有少数旗舰 SoC 能够满足端侧 AI 算力要求，且通过我们的测算，发现在复杂功能或长文本处理方面，仅仅依靠端侧 SoC 还是远远不够的。而根

据物理世界的客观规律，我们认为硬件的升级速度也是存在瓶颈的。因此，我们认为端侧 SoC 硬件的升级会是电子行业未来很长一段时间的主题任务。

路径#2 端侧小模型针对主流功能做定向优化，从参数量的维度减轻算力负载。事实上我们已经能够看到有不错的端侧小参数模型不断涌现，表现出亮眼的 AI 性能。而从手机 AI 功能的角度，主流功能其实是有限的，如果短期内通过硬件升级无法完全满足用户需求，那么我们认为端侧模型的定向优化或许是可以同步进行的思路。

2、端侧算力和云端算力不是此消彼长的关系，我们认为端侧 AI 对算力总盘子的拉动作用会长期存在。结构上看，**我们认为云端算力发展的机会点在于：用户对 AI 推理速度的容忍度。**当前端侧 AI 推理还处于从 0 到 1 阶段，用户对端侧 AI 更多的关注点在于“有没有”以及“效果好不好”，而不是“推理速度快不快”。也即，很多在硬件基础不足的情况下，也可以先通过拉长时间来实现 AI 推理。展望端侧 AI 推理从 1 到 10 的阶段，推理速度一定会是一个重要的优化环节，我们认为在一些网络、用电设施部署相对完备的室内场景中，云端推理仍然是一个不错的选择。

4. 风险提示

AI 手机出货量不及预期风险。智能手机市场竞争激烈，当前 AI 手机正处于市场渗透的早期阶段，并且消费者对 AI 手机的需求和认知参差不齐。若 AI 功能的研发进展不及预期、或消费者对 AI 功能支付溢价的意愿不及预期，则 AI 手机出货量也可能不及预期。

端侧软硬件技术发展不及预期风险。AI 应用对芯片算力、存储容量、电池续航等硬件要求极高，目前手机硬件在进行 AI 复杂运算时，常出现性能不足、发热严重、续航短等问题。若硬件升级滞后，将限制 AI 手机功能发挥。另外，要实现 AI 手机软件与硬件的深度适配，还需要进行算法优化，若 AI 算法在不同手机硬件上出现适配兼容性差、运行不稳定等情况，或将导致用户体验不佳。

AI 创新效果不及预期风险。部分 AI 手机功能看似新颖，但实际应用中对用户帮助不大，比如一些 AI 语音助手的理解和执行指令能力弱，无法真正解决用户痛点。展望未来，若 AI 手机创新缺乏 AI 底层技术的深度挖掘和创新应用，用户对 AI 创新的效果体验可能会不及预期。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴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。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn