国元证券 GUOYUAN SECURITIES

技术硬件与设备行业研究报告

2024 年 06 月 17 日

Apple Intelligence 将带动 iPhone 换机需求

推荐|首次

报告要点：

苹果将强化 AI 功能在其生态应用的落地，也印证了市场关于苹果本次 WWDC 的几个预期：1）苹果正式推出 Apple Intelligence，不同于其他竞争对手的地方在于苹果强调的 AI 不是人工智能，而是个人智能（personal intelligence），在运算方式上，采用端侧算力+私有云端算力相结合的方式。2）全新的 Siri 升级。Siri 能够开始理解用户的自然语言，并具备情境感，还可以“读懂”用户 iPhone 的操作界面，更好的处理用户需求。当苹果自带模型无法解决用户需求时，可在用户同意下连接 Chatgpt-4o。3）苹果对 IOS18、ipadOS、VisionOS、watchOS 和 MacOS 等系统进行全面 AI 升级，并基本实现了照片 Cleanup 等类似在 Android 手机上的相关 AI 功能。

从 Apple intelligence 的技术路线看苹果 AI 发展思路。苹果的 AI 发展思路可以归纳为两点：1）端侧 2）隐私。预计苹果的主体模型主要由端侧 on-device model 推动，数据无需上传到服务器，更好保护用户隐私。我们认为苹果在移动端侧通过 CORE ML 架构和 Apple Silicon 相结合，已形成自己独特的护城河。苹果自 2017 年开始引入 Core ML，到目前为止苹果基于自己的系统闭环，实现端侧模型性能和效率的提升，多模态、软硬件无缝集成等方面均已具备很强的能力。

预计 AI 功能带动换机潮，端侧 AI 应用促使 BOM 提升。若使用 iOS 18 的 AI 功能则要求芯片达到 A17 pro 以及 M1 以上规格，Apple intelligence 的功能体验需求有望推动和加快部分换机需求。我们预计 2024Q2 iPhone 出货量在 3900 万部左右，达到相对谷底，2024 年全年销量达到 2.23 亿部，25 年有望达到 2.39 亿部。从 BOM 成本变化上看，预计今年的 iPhone 16 全系列会搭载 A18 芯片，Pro 系列搭载 A18 Pro 芯片，NPU 算力更强大；其他硬件方面，Baseband 部分可能升级到高通的 X75；射频方面升级到 WIFI 7；Pro Max 机型上的长焦镜头或将升级到 48M；屏幕尺寸也将从 6.7 英寸升级至 6.9 英寸；对电源管理和散热要求的提升，以及电池容量需求增加带动钢壳锂电池需求和电源管理部分的半导体价值量提升，也将带动对快充的需求。

投资建议

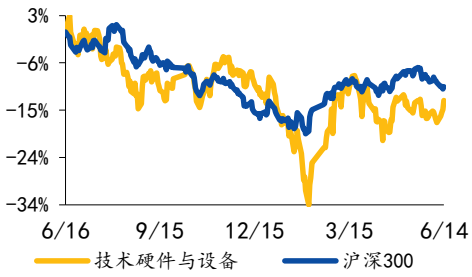
随着苹果在 WWDC 展示其在端侧加速对 AI 方面的布局，将进一步带动行业发展，随着产品逐步落地，硬件产品有望从中受益，我们持续看好苹果产业链和相关公司，建议关注：

组装：立讯精密、歌尔股份；**模拟芯片：**圣邦股份；**微型传动：**兆威机电；**快充：**安克创新；**电池：**珠海冠宇；**散热：**中石科技；**显示面板：**京东方；**光学零件：**水晶光电

风险提示

上行风险：消费电子景气度加速提升，AI 设备渗透率加速，AI 应用加速落地。**下行风险：**智能手机需求不及预期，AI 应用不及预期

过去一年市场行情



资料来源：Wind

相关研究报告

报告作者

分析师 彭琦

执业证书编号 S0020523120001

电话 (021)5109 7188

邮箱 pengqi@gyzq.com.cn

附表：重点公司盈利预测

公司代码	公司名称	投资评级	昨收盘（元）	总市值（百万元）	EPS			PE		
					2023A	2024E	2025E	2023A	2024E	2025E
002475	立讯精密	买入	34.35	246635.22	1.54	1.91	2.45	22.37	17.98	14.02
003021	兆威机电	买入	52.37	12537.98	1.05	1.36	1.94	89.51	38.52	26.96
300661	圣邦股份	增持	80	37658.17	0.60	0.96	1.37	133.81	83.53	58.31

资料来源：Wind，国元证券研究所

目 录

图表目录

1. 本次 WWDC 的几个预期

我们观察本次 6 月 11 日苹果召开的 2024 年 WWDC（苹果开发者大会），其主要的核心特点如下：

1、苹果正式推出了属于自己的 AI 应用 Apple Intelligence。苹果 AI 不同于其他竞争对手的地方在于苹果强调的 AI 不是人工智能，而是个人智能（personal intelligence）。具体的实施方式为：在运算方式上，采用端侧（个人手机、笔记本等）算力+私有云端算力相结合的方式，个人隐私和安全数据的运算会放在端侧，非隐私数据的运算放苹果的私有云上。

2、全新的 Siri 升级。Siri 能够开始理解用户的自然语言，并具备情境感，还可以“读懂”用户 iPhone 的操作界面，更好的处理用户需求。当苹果自带模型无法解决用户需求时，可在用户同意下连接 Chatgpt-4o。

3、苹果对 iOS18、ipadOS、VisionOS、watchOS 和 MacOS 等系统进行全面 AI 升级，并基本实现了照片 Cleanup，AI 绘图用的 Image Playground 等类似在 Android 手机上的相关 AI 功能。

从近期公司股价走势，市场对苹果的 Apple intelligence 的整体策略和前景呈现乐观态度。Apple intelligence 对移动芯片的使用门槛要求达到 A17 pro 以及 M1 以上规格，对后继 iPhone 机型带来换机需求。

图 1：苹果股价 vs 费城半导体指数

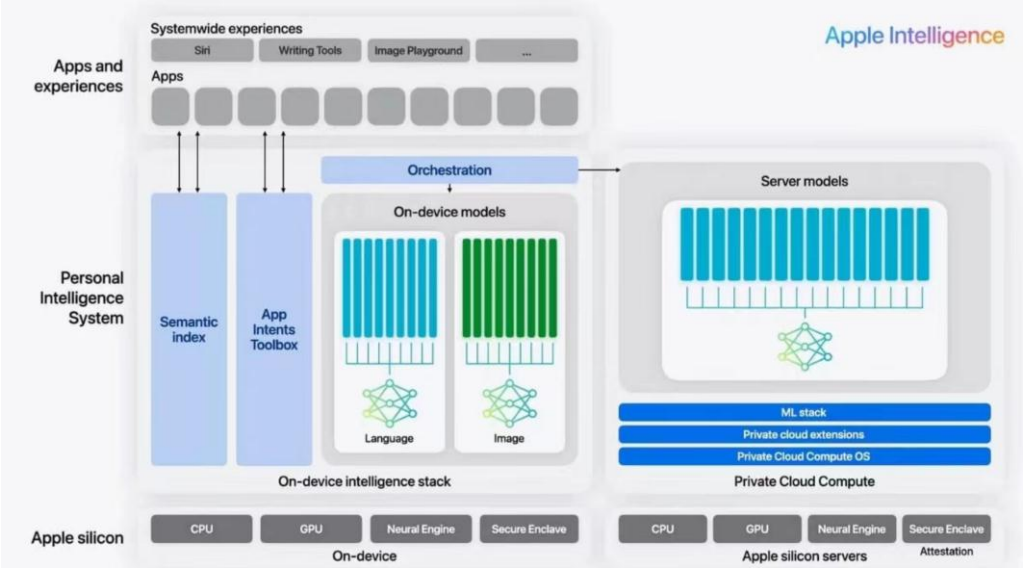
资料来源：Wind，国元证券研究所

2. 从 Apple intelligence 的技术路线看苹果 AI 发展思路

2.1 Apple intelligence 的技术路线

根据苹果在 WWDC 上展示的 Apple Intelligence 技术图谱以及公司在 Foundation model 方面上发布的官方文章，总体理解可以归纳为两点：1）端侧 2）隐私。

图 2：Apple intelligence 的技术路线



资料来源：namuwiki，国元证券研究所预测

1、苹果趟出了一条创新性的，能在移动端跑大模型同时兼顾效率和隐私的路。

2、苹果不需要自建独立大模型。通过 Orchestration 来判断和调动本地或者私有云上的模型，形成云端和本地的模型协同工作，构建自己的主体模型体系。这样在大模型上给自己最大的自由度，可以发展自己的 Ajax，可以合作 Chatgpt，也可以同 Gemini 或者文心一言等合作。

3、预计会大量倚重端侧的这些 3B 参数的 Foundational model。通过添加小型 LoRA adapters 适配器模块来提升模型的能力，兼具灵活和效率，这样在相对明确的功能和狭小范围上，实现对大模型效能的追赶或者超越。

适配器技术可以在本地设备上进行微调和适配，数据无需上传到服务器进行处理，这样可以更好地保护用户隐私。

4、On-device 的 Semantic index 可以协助整合用户在各应用程序之间的数据，以便生成式 AI 模型参照取用，能更好的理解分析用户的行为和信息。

App intents 不仅使得 OS 中 Siri 可以通过 App intent 来调用众多 App 的各种功能，同时能够实现 App 间的无缝对接，从事复杂的任务流程，真正能“懂”你的 AI 助手正在到来。

5、客户隐私和个性化要求，预计主体模型将主要由端侧 on-device model 推动。端侧算力越强，Apple intelligence 的体验就会越好。Apple silicon 所构成的端侧算力优势，将是未来苹果移动 AI 上的护城河。

2.2 苹果在端侧 AI 构建护城河

我们认为，苹果在移动端侧通过 CORE ML 架构和 Apple Silicon 相结合，形成自己独特的护城河。

苹果自己的 CORE ML 机器学习架构，最主要的特点包括：

- 可以让模型直接在苹果设备上所有 OS 系统运行而不需要依赖云端，保护使用者隐私。
- 通过把机器学习模型转换为 Core ML 格式，并利用优化技术如量化和剪枝，大幅压缩模型大小并提高推理速度，实现本地学习的超低能耗。
- 同苹果设备上的 CPU、GPU 和 Neural engine 无缝衔接，充分发挥端侧算力。
- OS 的 App 可以根据需求在端侧或云端调用 Foundational Model，实现强大的 App 应用体验。

市场仍有声音质疑苹果入局 AI PC 有些晚，但不要忘记的是，苹果是拥有自研操作系统和自研芯片的综合性厂商。

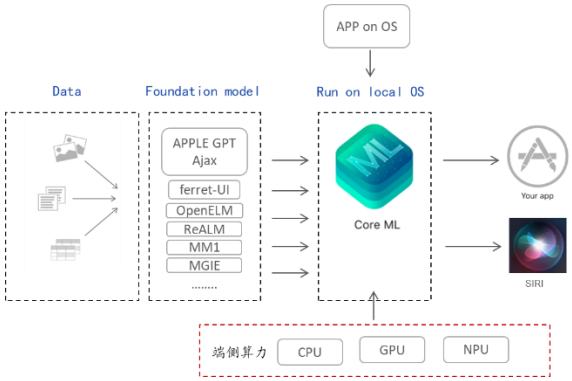
早在 2017 年苹果就开始引入 Core ML 到 iOS 11，发展到 2024 年，苹果基于自己的系统闭环，来实现端侧模型的性能和效率，多模态，软硬件无缝集成等方面，均已具备很强的能力。23 年 X86 架构尚未实现类似 ARM 架构中通过 NPU 做 AI 运算的时候，苹果 MacBook 上的 M2 中自研的 Neural engine 就能通过 Core ML 执行语音识别和图像处理等 AI 工作，而不是通过 CPU 或 GPU 去完成。

图 3：苹果 CORE ML 上的技术发展

年份	主要更新和功能
2017	**Core ML 引入：**首次推出Core ML框架，允许开发者在iOS设备上运行机器学习模型，支持图像识别、自然语言处理等任务。
2018	**Core ML 2：**引入模型压缩技术，使得机器学习模型在设备端运行速度更快，占用空间更小。**Create ML：**推出Create ML工具，允许开发者使用Swift在Mac上构建和训练机器学习模型。
2019	**Core ML 3：**支持更多类型的机器学习模型，增强了设备端训练功能。 **Create ML的改进：**进一步优化Create ML，使得开发者能够更容易地创建和训练复杂机器学习模型。
2020	**新的API和工具：**改进了机器学习API，使得开发者能够更方便地在应用中集成机器学习功能。 **隐私保护：**加强了数据隐私保护措施，确保机器学习任务主要在设备端完成。
2021	**性能优化：**进一步优化了模型运行的性能和效率。 **多模型类型支持：**增加了对新的模型类型和架构的支持，使得Core ML的应用范围更广。
2022	**模型部署简化：**简化了机器学习模型的部署过程，使得开发者能够更快速地将模型集成到应用中。 **增强现实（AR）支持：**改进了AR应用中的机器学习模型支持，提升了AR体验。
2023	**自监督学习：**引入了自监督学习，允许模型在没有标注数据的情况下进行训练，提升模型泛化能力。 **多模态支持：**增强了对多模态数据（如图像和文本结合）的支持，提高了复杂任务处理的能力。
2024	**强化学习：**增加了对强化学习模型的支持，使得开发者能够构建更复杂的AI系统。 **设备端和云端的智能调度：**引入了智能调度机制，根据任务需求在设备端和云端之间进行分配，优化资源利用和性能。 **无缝集成：**改进了与其他苹果生态系统服务的无缝集成，提高了整体用户体验和应用智能化水平。

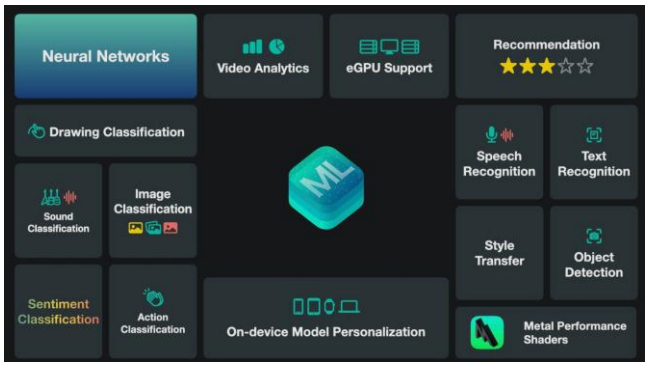
资料来源：Infoq，CSDN，diglog，机器之心，智东西，国元证券研究所

图 4：CORE ML 是端侧 AI 运算的软体核心



资料来源：apple 官网，国元证券研究所

图 5：苹果的 CORE ML 功能架构



资料来源：电子工程专辑，国元证券研究所

3. AI 功能带动换机潮，端侧 AI 应用促使 BOM 提升

苹果的 AI 侧重端侧运算对终端设备的硬件规格提出更高要求，若要使用 iOS 18 的 AI 功能则要求搭载 A17 pro 以及 M1 以上规格芯片。

Apple intelligence 的功能体验需求有望推动和加快部分换机需求，我们预计2024Q2 iPhone 出货量在 3900 万部左右，达到相对谷底，Q3 和 Q4 出货量或能达到 6260 万部以及 7600 万部，全年达到 2.229 亿部，较 23 年略下滑 640 万部。而 25 年有望达到 2.39 亿部，同比增长 7.2%。

图 6：iPhone 整体出货量预估（百万部）

mn units	23Q1e	2Q23e	3Q23e	4Q23e	2023e	24Q1e	2Q24e	3Q24e	4Q24e	2024e	2025e
iPhone 13 (6.1")	1.5	1.0	0.3		**2.8**						
iPhone 13 Pro (6.1")	1.3	1.3	0.5		**3.1**						
iPhone 13 Pro Max (6.7")	2.0	1.0	0.4		**3.4**						
iPhone 14 (6.1")	10.0	10.0	8.0	6.0	**34.0**	2.1	1.0	0.3		**3.4**	
iPhone 14 Plus (6.7")	4.0	3.0	2.5	2.0	**11.5**	0.8	0.6	0.3	0.1	**1.8**	
iPhone 14 Pro (6.1")	16.0	13.0	7.0	5.5	**41.5**	2.5	2.2	0.8	0.3	**5.8**	
iPhone 14 Pro Max (6.7")	19.0	10.0	8.0	6.0	**43.0**	1.5	1.0	0.2	0.1	**2.8**	
iPhone 15			9.0	15.0	**24.0**	11.0	10.5	6.0	3.0	**30.5**	**2.0**
iPhone 15 plus			4.0	7.0	**11.0**	5.0	5.0	3.0	2.0	**15.0**	**1.0**
iPhone 15 pro			8.0	15.0	**23.0**	11.0	9.0	9.0	6.0	**35.0**	**3.0**
iPhone 15 ultra			10.0	12.0	**22.0**	8.6	8.5	6.0	3.5	**26.6**	**2.0**
iphone16								9.5	15.0	**24.5**	**25.0**
iphone16 plus								5.0	12.0	**17.0**	**15.0**
iphone16 pro								10.0	18.0	**28.0**	**26.0**
iphone 16 promax								10.0	13.5	**23.5**	**22.0**
iphone17											**30.0**
iphone17 plus											**20.0**
iphone17 pro											**42.0**
iphone 17 promax											**41.0**
整体出货量	**57.8**	**41.3**	**59.7**	**70.5**	**229.3**	**44.5**	**39.8**	**62.6**	**76.0**	**222.9**	**239.0**

资料来源：国元证券研究所预测

从产业链调研来看，预计今年的 iPhone 16 全系列会搭载 A18 芯片，Pro 系列搭载 A18 Pro 芯片。A18 和 A18 Pro 预计都会采用 TSMC 的 N3E 工艺。A18 Pro 则有可能搭载更多的 NPU 核心，算力更强大，芯片的面积也相应扩大。

图 7：苹果移动芯片制程变化

芯片	P-core	E-core	GPU核心数	NPU核心数	NPU算力	制程（nm）
A11 Bionic	2	4	3	2	0.6 TOPS	10nm
A12 Bionic	2	4	4	8	5 TOPS	7nm
A13 Bionic	2	4	4	8	5.5 TOPS	7nm
A14 Bionic	2	4	4	16	11 TOPS	5nm
A15 Bionic	2	4	5	16	15.8 TOPS	5nm
A16 Bionic	2	4	5	16	17 TOPS	4nm
A17 Pro	2	4	6	16	35 TOPS	3nm-N3B
A18（预测）	2	4	6	16	40-50TOPS	3nm-N3E
A18 Pro（预测）	2	4	6	16+	50TOPS+	3nm-N3E

资料来源：机械之心，ofweek，eefocus，至顶网，wallstreetcn，国元证券研究所预测

其他硬件方面，Baseband 部分可能升级到高通的 X75；射频方面升级到 WIFI 7；Pro Max 机型上，长焦镜头升级到 48M，即后摄部分升级为两颗 48M 摄像头；屏幕尺寸从 6.7 寸升级到 6.9 寸。

DRAM 部分，考虑苹果在边缘算力上具备较高的能耗控制能力，以及高效的内存带宽技术，今年可能会继续保持 8GB 的配置。

在其他方向上，电源管理和散热要求的提升，还有电池容量需求增加带动了钢壳锂电池需求和电源管理部分的半导体价值量提升，也带动了快充需求。

图 8：iPhone16 BOM 成本拆分（美元）

	iphone15 promax			iphone 16 promax		
	规格	BOM 成本$	供应商	预测技术升级趋势	预测BOM 成本	供应商
主芯片		**75**			**89**	
AP	A17Pro	60	APPLE/TSMC 3nm/N3b	A18将采用TSMC N3e,更强的GPU和NPU神经网络算力	72	APPLE/TSMC 3nm/N3e
BB	X70	15	Qualcomm/Samsung 4nm	升级到X75	17	Qualcomm/TSMC 4nm
射频前端		**48**			**53**	
RF Tranciver		10	Qualcomm		11	Qualcomm
5G/Wifi 6e FEM		20	Avago，Skyworks，Qorvo	可能升级到WIFI7	23	Avago，Skyworks，Qorvo
前端射频模组		8	Murata，Skyworks		9	Murata，Skyworks
其他芯片		**36**	TI,Bosch，Alps，NXP，ADI，Dialog	支持40W充电和20W无线充电PMIC和controller规格提升	**39**	TI,Bosch，Alps，NXP，ADI，Dialog
镜头模组		**127**			**150**	
48MP 主摄		40	SONY,LG innotek, 舜宇，Largon，Genius，Alps，Mitsumi	至少有一颗12MP升级到48MP，提升iphone 空间照片能力，和立体摄像功能	40	SONY,LG innotek, 舜宇，Largon，Genius，Alps，Mitsumi
12MP 长焦+潜望式		35	SONY,LG innotek, Largon，Johwa，水晶，蓝特		32	SONY,LG innotek, Largon，Johwa，水晶，蓝特
12MP 超广角		10	SONY,LG innotek, Largon，Genius	主摄可能采用modling glass	35	SONY,LG innotek, Largon，Genius
后置LiDAR TOF		10	SONY, IIVI/Lumentum		12	SONY, IIVI/Lumentum
FACE ID		15	Lumentum，IIVI,AMS Osram, LG innotek, STM ,采钰，精材，TSMC		16	Lumentum，IIVI,AMS Osram, LG innotek, STM ,采钰，精材，TSMC
存储		**36**			**43**	
DDR5	8GB	20	SK Hynix	8GB	25	SK Hynix
FLASH	256GB	16	Kioxia	256GB	18	Kioxia
显示屏		**85**			**90.5**	
Display	6.7inch，460 ppi 2796 x 1290，120Hz 自适应刷新	75	Samsung ， LG Display	尺寸从6.7inch升到6.9inch	80	Samsung ， LG Display
盖板玻璃		7	Corning		7.5	Corning
架构		**70**			**67.5**	
太合金边框和housing		58	富士康, Jabil		55	富士康, Jabil，比亚迪
玻璃后盖		6	蓝思，伯恩		5.5	蓝思，伯恩
电池		**10**		磨砂金属外壳，叠片电池，散热材料	**12**	
总计		559			619.5	

资料来源：国元证券研究所测算

4. 行业重点公司推荐

随着苹果在 WWDC 展示其在端侧加速对 AI 方面的布局，将进一步带动行业发展，随着产品逐步落地，硬件产品有望从中受益，我们持续看好苹果产业链和相关公司，建议关注：

组装：立讯精密、歌尔股份；**模拟芯片**：圣邦股份；**微型传动**：兆威机电；**快充**：安克创新；**电池**：珠海冠宇；**散热**：中石科技；**显示面板**：京东方；**光学零件**：水晶光电

图 9：重点关注公司

2024/6/15						销售增速		PE倍数		P/B		ROE	
单位：RMB	股票代码	国元评级	目标股价（元）	当前股价（元）	市值（亿元）	2024E	2025E	2024E	2025E	2024E	2025E	2024E	2025E
立讯精密	**002475.SZ**	**买入**	**38.22**	**34.35**	**2466.35**	**20.1%**	**10.1%**	**18.0**	**14.0**	**3.52**	**2.81**	**20.0%**	**20.0%**
兆威机电	**003021.SZ**	**买入**	**85.30**	**52.37**	**125.38**	**34.4%**	**30.2%**	**38.5**	**27.0**	**2.70**	**2.45**	**7.0%**	**9.1%**
圣邦股份	**300661.SZ**	**增持**	**96.00**	**80.00**	**376.58**	**24.1%**	**24.8%**	**83.5**	**58.3**	**8.84**	**7.79**	**10.6%**	**13.4%**
水晶光电	002273.SZ	-	17.70	16.18	225.00	27.3%	19.6%	28.1	22.8	2.61	2.41	8.7%	14.0%
珠海冠宇	688772.SH	-	16.57	13.65	153.87	18.5%	19.9%	20.0	13.5	2.04	1.80	10.1%	13.2%
领益智造	002600.SZ	-	6.92	5.12	358.82	26.0%	18.9%	15.3	11.5	1.67	1.48	11.3%	13.2%
鹏鼎控股	002938.SZ	-	29.13	34.50	799.90	11.4%	11.1%	21.0	18.2	2.44	2.21	11.7%	12.3%
歌尔股份	002241.SZ	-	18.55	17.56	600.05	6.3%	13.3%	28.6	21.1	1.85	1.71	6.4%	8.1%
京东方A	000725.SZ	-	5.14	4.07	1520.63	17.2%	11.4%	30.4	16.9	1.14	1.08	3.6%	5.9%
中石科技	300684.SZ	-	16.51	17.35	51.96	26.4%	26.2%	36.2	24.2	2.56	2.38	7.2%	10.0%
安克创新	300866.SZ	-	82.09	72.78	384.63	20.2%	16.9%	20.1	17.0	4.00	3.40	20.3%	20.4%
中位数					376.58	20.2%	18.9%	28.1	18.2	2.56	2.38	10.1%	13.2%

资料来源：Wind，国元证券研究所预测　注：股价为 2024 年 6 月 14 日收盘价；未覆盖公司采用 Wind 一致预期

图 10：重点关注公司营收增速

营收增速	2021	2022	2023	2024E	2025E	21-23CAGR%	24-25CAGR%
立讯精密	66.4%	39.0%	8.4%	20.1%	10.1%	35.8%	15.0%
兆威机电	-4.6%	1.1%	4.6%	34.4%	30.2%	0.3%	32.3%
圣邦股份	87.1%	42.4%	-17.9%	24.1%	24.8%	29.8%	24.5%
水晶光电	18.2%	14.9%	16.0%	27.3%	19.6%	16.3%	23.4%
珠海冠宇	48.5%	6.1%	4.3%	18.5%	19.9%	18.0%	19.2%
领益智造	8.0%	13.5%	-1.0%	26.0%	18.9%	6.6%	22.4%
鹏鼎控股	11.6%	8.7%	-11.4%	11.4%	11.1%	2.4%	11.3%
歌尔股份	35.5%	34.1%	-6.0%	6.3%	13.3%	19.5%	9.8%
京东方A	61.8%	-18.6%	-2.2%	17.2%	11.4%	8.8%	14.3%
中石科技	8.6%	27.6%	-21.0%	26.4%	26.2%	3.1%	26.3%
安克创新	34.4%	13.3%	22.9%	20.2%	16.9%	23.2%	18.5%

资料来源：Wind，国元证券研究所预测 注：未覆盖公司采用 Wind 一致预期

5. 风险提示

上行风险：

消费电子景气度加速提升：AI 主要搭载于消费电子，景气度上行将对业绩产生利好。

AI 设备渗透率加速：市场对 AI 产品接受度高将带动公司加快 AI 产品落地。

AI 应用加速落地：AI 技术在端侧的应用加速，将有望带动消费电子换机潮，将利好公司业绩。

下行风险：

智能手机需求不及预期：若苹果手机销量下滑，对苹果和其供应链的厂商的营利能力带来压力。

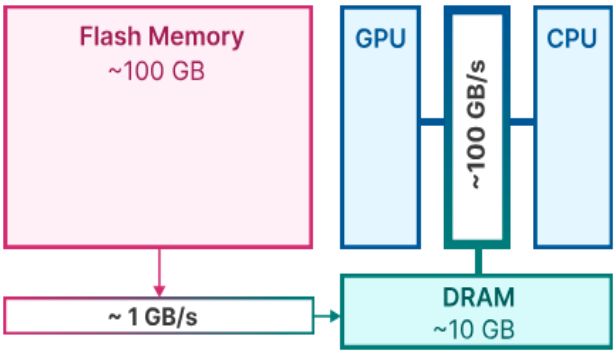
AI 应用落地不及预期：苹果重点布局端侧 AI 模型，倘若模型落地慢于预期，则对公司的产品竞争力产生不利影响。

附录：苹果大模型内容介绍

LLM in flash：DRAM 带宽高但容量小，限制模型大小，LLM in flash 提出将模型存储于闪存，通过增加数据块的大小和使用多线程读取，弥补带宽的不足，减少闪存到 DRAM 的数据传输时间，在闪存上实现高效的 LLMs 推理。

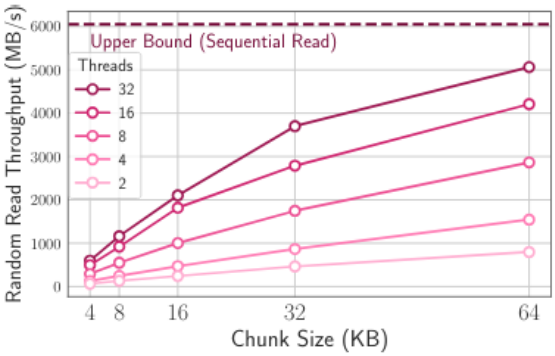
LLM in flash 可以实现在 DRAM 有限的设备中运行比 DRAM 容量大 2 倍的模型，通过此技术的优化，CPU 推理速度比标准 LLM 模型提高 4-5 倍，GPU 的推理速度提高 20-25 倍。LLM in Flash 可帮助苹果在设备端搭载更大端侧模型的同时提升模型的推理速度。

图 11：统一内存架构下的带宽



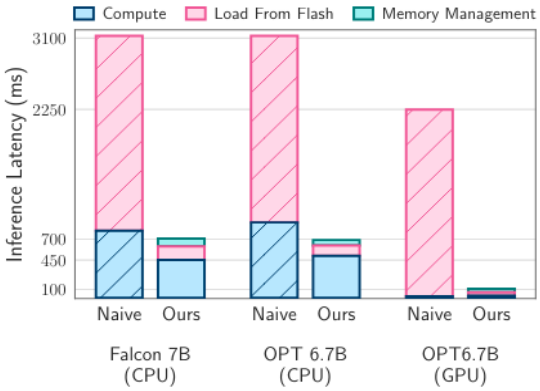
资料来源：《LLM in a flash: Efficient Large Language Model Inference with Limited Memory》，国元证券研究所

图 12：闪存随机读取吞吐量



资料来源：《LLM in a flash: Efficient Large Language Model Inference with Limited Memory》，国元证券研究所

图 13：不同模型 1 Token 的延迟



资料来源：《LLM in a flash: Efficient Large Language Model Inference with Limited Memory》，国元证券研究所

Ajax 语言模型框架：Ajax 是苹果独有的语言模型框架，苹果通过 Ajax 优化 iOS18，对 Siri、Spotlight 搜寻、Safari 等现有功能与应用在端侧进行 AI 升级。现阶段 Siri 可接入云端 Chatgpt 为用户提供 AI 服务，未来苹果预计通过自研框架 Ajax 开发聊天机器人，后续有望替换 Chatgpt 在苹果云端的应用。

ReALM：ReALM 用于解决 LLM 与非对话实体之间的指代消解（Reference Resolution）问题。人类语言常包括“他们”或“那个”的模糊，所以让 LLM 理解上下文至关重要。ReALM 模型会重建平台上的屏幕关键信息，对屏幕指代位置进行标记，以便大语言模型能够在上下文中了解指代出现的位置以及周围的文本内容。

ReALM 的性能全方位超越同类模型 MARRS，最小的模型与 GPT-4 性能相当，较大的模型则优于 GPT-4。ReALM 可以让 Siri 更容易理解用户的对话和屏幕上显示的信息，并做出更加精准的操作。

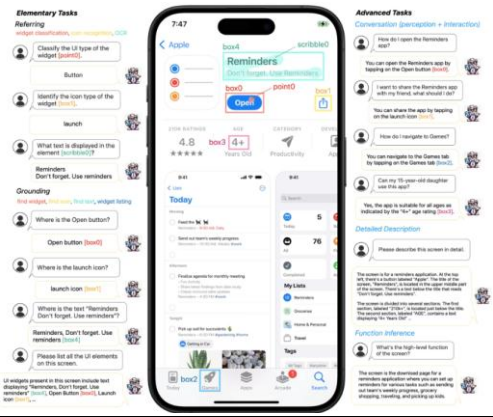
图 14：ReALM 对比其他模型

Model	Conversational	Synthetic	On-screen	Unseen
MARRS	92.1	99.4	83.5	84.5
GPT-3.5	84.1	34.2	74.1	67.5
GPT-4	97.0	58.7	90.1	98.4
ReALM-80M	96.7	99.5	88.9	99.3
ReALM-250M	97.8	99.8	90.6	97.2
ReALM-1B	97.9	99.7	91.4	94.8
ReALM-3B	97.9	99.8	93.0	97.8

资料来源：《ReALM: Reference Resolution As Language Modeling》，国元证券研究所

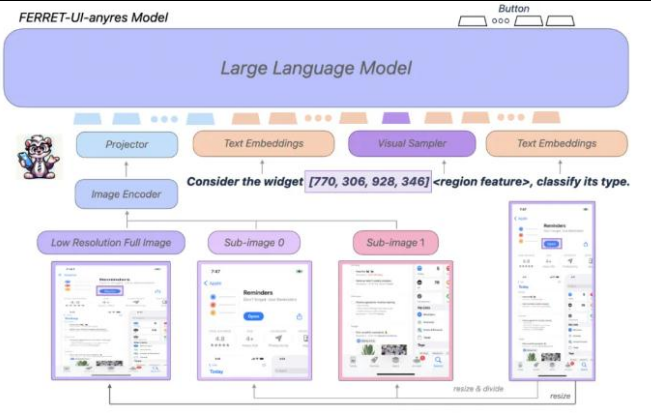
Ferret-UI 模型：苹果开发出 Ferret-UI 模型，是首个专门针对 UI 屏幕设计的用于精确引述和定基任务的 MLLM。Ferret-UI 通过引入任何分辨率（anyres）概念，其加入了额外的细粒度图像特征，将屏幕分割成多个子图像，并对每个子图像放大，从而捕捉到更多细节。Ferret-UI 为 Apple intelligence 跨 APP 操作打下基础，使其可以更精准的识别 APP 界面。

图 15：Ferret-UI 模型



资料来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，国元证券研究所

图 16：Ferret-UI-anyres 架构



资料来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，国元证券研究所

Ferret-UI 优化明显，性能表现整体超过 Spotlight 和 GPT-4V。基础版 Ferret-UI

在公共基准和初级任务方面均超过 Spotlight 和 GPT-4V，在高级任务中略逊色于 GPT-4V，相较于未优化的 Ferret，性能提升明显。

图 17：Ferret-UI 对比其他模型

	Public Benchmark			Elementary Tasks				Advanced Tasks	
	S2W	WiC	TaP	Ref-i	Ref-A	Grd-i	Grd-A	iPhone	Android
Spotlight [30]	106.7	141.8	**88.4**	-	-	-	-	-	-
Ferret [53]	17.6	1.2	46.2	13.3	13.9	8.6	12.9	20.0	20.7
Ferret-UI-base	113.4	**142.0**	78.4	80.5	**82.4**	79.4	83.5	73.4	80.5
Ferret-UI-anyres	**115.6**	140.3	72.9	**82.4**	**82.4**	**81.4**	**83.8**	93.9	71.7
GPT-4V [1]	34.8	23.5	47.6	61.3	37.7	70.3	4.7	**114.3**	**128.2**

资料来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，国元证券研究所

OpenELM 是一款苹果专门为手机和电脑等终端设备而设计的小模型，发布了 2.7 亿、4.5 亿、11 亿和 30 亿四个参数版本，其使用了“分层缩放”策略，有效分配 Transformer 模型的每一层参数，从而提高准确率。在约 10 亿参数规模下，与 OLMo 相比，OpenELM 的准确率提升了 2.36%，需要的预训练 Token 减少了 50%。

图 18：OpenELM 对比 LLMs

Model	Public dataset	Open-source		Model size	Pre-training tokens	Average acc. (in %)
		Code	Weights			
OPT [55]	✗	✓	✓	1.3 B	0.2 T	41.49
PyThia [5]	✓	✓	✓	1.4 B	0.3 T	41.83
MobiLlama [44]	✓	✓	✓	1.3 B	1.3 T	43.55
OLMo [17]	✓	✓	✓	1.2 B	3.0 T	43.57
OpenELM (Ours)	✓	✓	✓	1.1 B	1.5 T	**45.93**

资料来源：《OpenELM: An Efficient Language Model Family with Open Training and Inference Framework》，国元证券研究所

多模态大模型 MM1 具备高达 300 亿参数，采用 MoE 架构优化性能，拥有强大的多模态学习和推理能力，在上下文预测、多图像和思维链推理等方面具有较好的表现，擅长在用户输入和文本中寻找规则。MM1 让 Siri 变得更自然化、情景化和个人化，更好的理解用户需求。

与其他模型在图像描述和视觉问题回答方面相比，MM1-3B 和 7B 在 8 样本情况下表现优于其他模型，MM1-30B 则在 8 样本和 16 样本情况下性能高于其他模型。

图 19：MM1 模型

资料来源：《MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training》，国元证券研究所

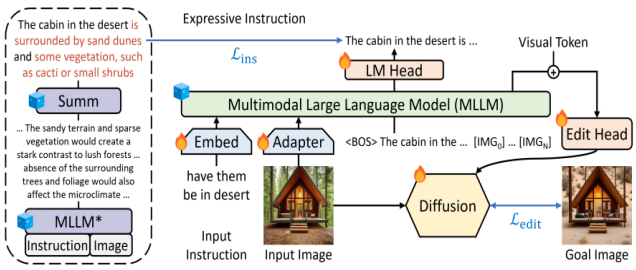
图 20：多模态预训练评估

Model	Shot	Captioning			Visual Question Answering			
		COCO	NoCaps	TextCaps	VQAv2	TextVQA	VizWiz	OKVQA
MM1-3B Model Comparisons								
Flamingo-3B [3]	0†	73.0	–	–	49.2	30.1	28.9	41.2
	8	90.6	–	–	55.4	32.4	38.4	44.6
MM1-3B	0	73.5	55.6	63.3	46.2	29.4	15.6	26.1
	8	**114.6**	**104.7**	**88.8**	**63.6**	**44.6**	**46.4**	**48.4**
MM1-7B Model Comparisons								
IDEFICS-9B [58]	0†	46.0*	36.8	25.4	50.9	25.9	35.5	38.4
	8	97.0*	86.8	63.2	56.4	27.5	40.4	47.7
Flamingo-9B [3]	0†	79.4	–	–	51.8	31.8	28.8	44.7
	8	99.0	–	–	58.0	33.6	39.4	50.0
Emu2-14B [105]	0†	–	–	–	52.9	–	34.4	42.8
	8	–	–	–	59.0	–	43.9	–
MM1-7B	0	76.3	61.0	64.2	47.8	28.8	15.6	22.6
	8	**116.3**	**106.6**	**88.2**	**63.6**	**46.3**	**45.3**	**51.4**
MM1-30B Model Comparisons								
IDEFICS-80B [58]	0†	91.8*	65.0	56.8	60.0	30.9	36.0	45.2
	8	114.3*	105.7	77.6	64.8	35.7	46.1	55.1
	16	116.6*	107.0	81.4	65.4	36.3	48.3	56.8
Flamingo-80B [3]	0†	84.3	–	–	56.3	35.0	31.6	50.6
	8	108.8	–	–	65.6	37.3	44.8	57.5
	16	110.5	–	–	66.8	37.6	48.4	57.8
Emu2-37B [105]	0	–	–	–	33.3	26.2	40.4	26.7
	8	–	–	–	67.8	49.3	54.7	54.1
	16	–	–	–	68.8	50.3	57.0	57.1
MM1-30B	0	70.3	54.6	64.9	48.9	28.2	14.5	24.1
	8	123.1	111.6	92.9	70.9	49.4	49.9	58.3
	16	**125.3**	**116.0**	**97.6**	**71.9**	**50.6**	**57.9**	**59.3**

资料来源：《MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training》，国元证券研究所

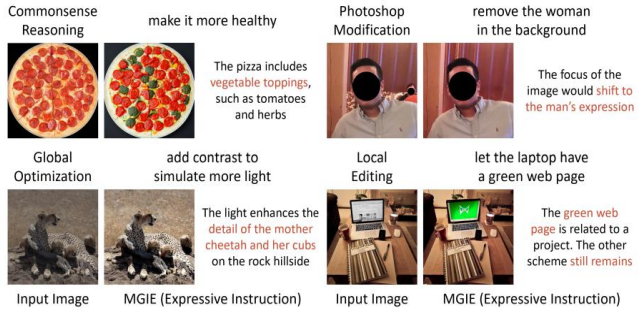
MGIE（MLLM-Guided Image Editing）：MGIE 可以用于解决图像编辑指令模糊的问题，MLLM 会补充更详细的指令，利用预期目标的潜在想象力执行图像编辑，MGIE 就可以从固有的视觉推导中实现合理编辑。例如，当“健康”的定义过于模糊时，MGIE 可以将“蔬菜配料”与披萨精确地联系起来。MGIE 让 Apple intelligence 在图像创作中更具准确性和创造性。

图 21：MGIE 原理



资料来源：《GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS》，国元证券研究所

图 22：MLLM 补充详细指令



资料来源：《GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS》，国元证券研究所

声学模型融合（AMF）：利用外部声学模型的优势补充 E2E 系统的固有功能，达到完善语音识别的作用。声学模型融合解决了域不匹配的问题，单词错误率也有明显降低，在测试中错误率降低了 14.3%，同时增强了对命名实体和稀有词的识别。AMF 可减少用户与 Siri 交互时因表达错误或稀有词而产生的偏差。

国元证券股份有限公司
GUOYUAN SECURITIES CO.,LTD.

投资评级说明：

（1）公司评级定义		（2）行业评级定义	
买入	预计未来 6 个月内，股价涨跌幅优于上证指数 20%以上	推荐	预计未来 6 个月内，行业指数表现优于市场指数 10%以上
增持	预计未来 6 个月内，股价涨跌幅优于上证指数 5-20%之间	中性	预计未来 6 个月内，行业指数表现介于市场指数±10%之间
持有	预计未来 6 个月内，股价涨跌幅介于上证指数±5%之间	回避	预计未来 6 个月内，行业指数表现劣于市场指数 10%以上
卖出	预计未来 6 个月内，股价涨跌幅劣于上证指数 5%以上		

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告。本人承诺报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业操守和专业能力，本报告清晰准确地反映了本人的研究观点并通过合理判断得出结论，结论不受任何第三方的授意、影响。

证券投资咨询业务的说明

根据中国证监会颁发的《经营证券业务许可证》（Z23834000），国元证券股份有限公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询业务是指取得监管部门颁发的相关资格的机构及其咨询人员为证券投资者或客户提供证券投资的相关信息、分析、预测或建议，并直接或间接收取服务费用的活动。证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

一般性声明

免责条款

国元证券研究所

合肥	上海
地址：安徽省合肥市梅山路 18 号安徽国际金融中心 A 座国元证券	地址：上海市浦东新区民生路 1199 号证大五道口广场 16 楼国元证券
邮编： 230000	邮编： 200135
传真： （0551）62207952	传真： （021）68869125
	电话： （021）51097188