电子行业深度报告

端侧 AI：模型创新快速迭代，看好苹果引领 AI 硬件起飞

增持（维持）

2025 年 02 月 23 日

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
证券分析师 陈妙杨
执业证书：S0600525020002
chenmy@dwzq.com.cn

投资要点

- **端侧 AI 革新人机交互，模型快速升级，巨头引领行业发展：**AI 自主化能力沿着“以指令为中心“到“以意图为中心”持续提升。LLM 从各个层面改造终端，其中 Agent 对开放式问题必不可少，背后是大模型带来的理解复杂输入、进行规划推理/合理使用工具的能力。据头豹，端侧 AI 市场规模 2023-2028 年预计 CAGR 高达 58%，2028 年超过 1.9 万亿元。从具体小模型性能表现上看，参数量对模型性能影响巨大，但受限于硬件，小模型的技术创新更加积极以提升有限参数量下的性能表现，其中量化/剪枝/蒸馏是最主要的模型压缩方式，各家小模型因数据集/压缩精度/量化混合方式等差异预计带来小模型的百花齐放。Agent 架构中，基础模型本身要引入新的输入类型，成为 VLA 模型，同时还增加了个性化和内存操作要求，均需要额外的优化。

- **硬件变革核心在内存，苹果发力内存创新应对内存瓶颈：**相比于云端模型，硬件是端侧模型的重要制约，需要升级以补齐短板。对比各家硬件，我们认为苹果在内存/电池/散热上提升空间巨大。我们认为内存及其操作带来的能耗是当前最短板，预计成为硬件核心变革方向，如半精度的 7B 模型仅参数加载占用 DRAM 就超过 14GB，同时 DRAM 耗能比 SRAM 和计算高出两个数量级。同时 iOS 和安卓内存利用效率差异巨大，我们认为安卓需要在 OS 层提供统一的 AI 基础模型，而 iOS 在模型压缩之外则需要提高硬件内存以克服硬件瓶颈。除了简单的增加内存容量外，苹果在内存结构、耗能、传输速度等方面创新密集，如与三星合作开发独立封装形式，以及推进全新的 WMCM 封装方式进一步提高芯片组合的灵活性和集成度。

- **多模态 UI 交互界面革命带来 Agent 的历史机遇：**根据交互的模式，任务执行方法可分为基于 API 和基于用户界面（UI）的方法。API 交互泛用性较弱。UI 界面方式在 Transformer 架构下较好克服了任务和 UI 元素之间的隐含关系，大幅提升了 GUI Agent 的可行性，有望成为主流。当前苹果和谷歌均发力 UI 交互模型，苹果的 Ferret-UI 和谷歌的 Screen AI 模型都采用读屏 AI 视觉语言模型，采用统一编码方式理解屏幕信息。从谷歌 UI 模型看，模型参数提升对性能影响较大，同时 5B 模型对性能提升尚未饱和，有必要继续提升模型性能。

- **风险提示：**底层创新不及预期，技术发展不及预期，竞争加剧风险

行业走势

相关研究

《信通院规范行业标准，AI 眼镜落地拐点年》
2025-02-20
《为什么人形机器人的终局是消费电子？》
2025-02-20

表 1：重点公司估值（2025/2/21）

代码	公司	总市值（亿元）	收盘价（元）	EPS 2023A	EPS 2024E	EPS 2025E	PE 2023A	PE 2024E	PE 2025E	投资评级
002475	立讯精密	3,255.57	44.98	1.51	1.89	2.36	29.72	23.80	19.06	买入
300433	蓝思科技	1,443.04	28.96	0.61	0.83	1.10	47.76	34.89	26.33	买入
002938	鹏鼎控股	990.49	42.72	1.42	1.53	1.77	30.13	27.92	24.14	买入
002600	领益智造	748.47	10.68	0.29	0.28	0.40	36.49	38.14	26.70	买入
002384	东山精密	563.98	33.06	1.15	1.01	1.58	28.71	32.73	20.92	买入

数据来源：Wind，东吴证券研究所
注：以上公司预测值来源于东吴证券研究所

内容目录

图表目录

1. AI 技术在端侧逐渐深化，引领全新人机协作方式

1.1. AI 自主化能力逐渐提高，与系统融合程度提升

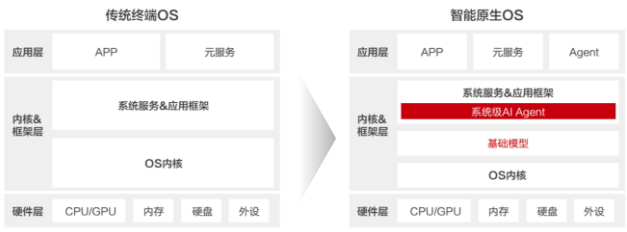
从指令到意图驱动，AI 自动化能力持续提升。根据自动化程度，可将个人智能助理（intelligent personal assistants,IPAs）分为 5 个级别。AI 自主化能力沿着“以指令为中心“到”以意图为中心“持续提升，用户只需要表达出需求，实现需求的过程将交由系统完成。从 L1- L3 级的智能体都在用户指令的被动驱动下工作，而 L4 级以上智能体能够理解用户的历史数据，感知当前状况，并在适当的时候主动提供个性化服务。

大模型和智能体驱动下一代终端操作系统。当前的操作系统依然是建立在静态规则和预定义的逻辑流程上，未来真正理解用、为用户量身定制的原生智能 OS 将进一步拓展终端 OS 的内涵。从 AI 技术在终端产品的落地上，一般经历：单点特定的 AI 增强实现**应用层集成 AI**→OS 智能化改造实现**系统层融合 AI（原子化控件化的 AI 能力）**→以 AI 为中心的全新 OS 并且系统级 AI Agent 出现。

图1：AI 智能化分级

数据来源：华为 AI 终端白皮书，东吴证券研究所

图2：AI OS 架构

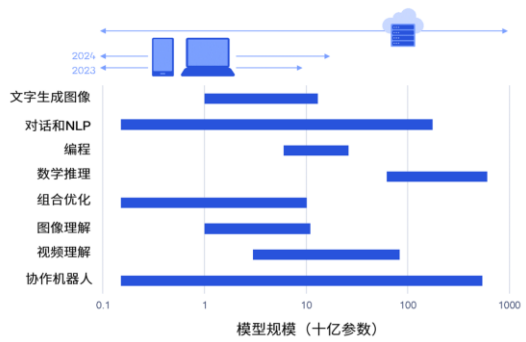


数据来源：华为 AI 终端白皮书，东吴证券研究所

1.2. 基于用户场景的端云混合架构

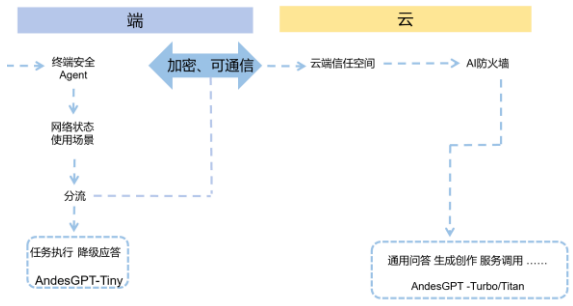
当前端侧模型性能和云端大模型依然有较大差距，因此基于用户场景的端云协同 AI 将构筑全局化智能。根据具体的场景（用户意图、网络情况、敏感隐私）以及所需的性能选择端侧或者云端执行，端侧适合无网、隐私要求高、响应要求高的场景。

图3：终端 AI 能力主线增强但当前依然不足



数据来源：高通，东吴证券研究所

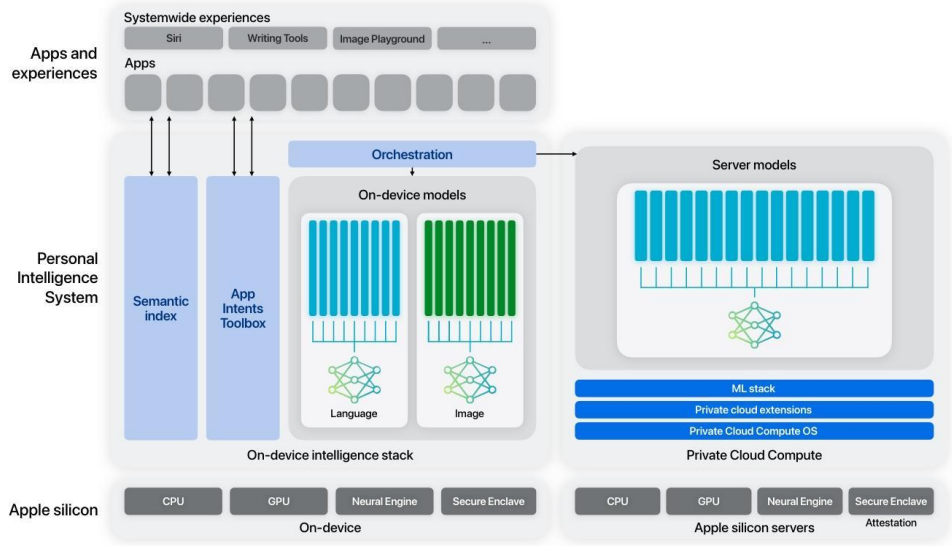
图4：端侧和云端 AI 处理的分工示意



数据来源：Oppo，东吴证券研究所

苹果通过小型本地模型+私有云模型+第三方大模型三层架构实现系统级 AI。苹果在设备端部署一个参数量为 30 亿的语言模型和一个图像模型。同时具备一个编排层（Orchestration）协调多个模型，根据用户请求调用对应能力的模型。无论是端侧还是私有云模型都基于 Apple 芯片底座以提供计算和安全支持，应用层都以 Siri、Writing Tools 等形式以实现用户体验的一致性。同时在外部第三方大模型上合作 ChatGPT 以响应更加开放和复杂的需求。

图5：苹果的 AI 生态架构



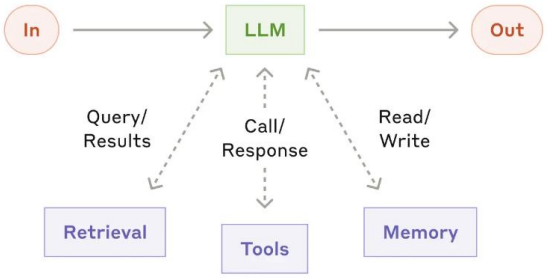
数据来源：苹果，东吴证券研究所

1.3. LLM 从各个层面改造终端

大模型在各个层面改造终端交互体验，预计分为三种形式。最开始增强型 LLM 架构，可以增加单点的功能体验，比如在翻译/图片处理/语音转换中使用 AI；阶段二主要以工作流形式，比如生成营销文案并翻译等；最后阶段是针对开放问题的 Agent 形态。这三种形态出现形式有先后，但是最终预计并存以针对不同特性的场景。

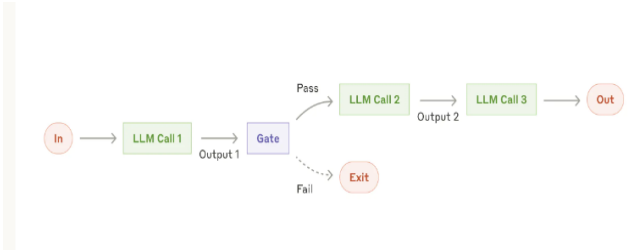
最基础的构建块（增强型 LLM）——工作流——自主的 Agent，复杂性持续提升。同时虽然非 Agent 在自动化水平上提升有限，但是**由于调用的是 LLM，在智能化水平、可处理任务等方面也有巨大的提升。**

图6：增强型 LLM 架构



数据来源：《Building effective agents》，东吴证券研究所

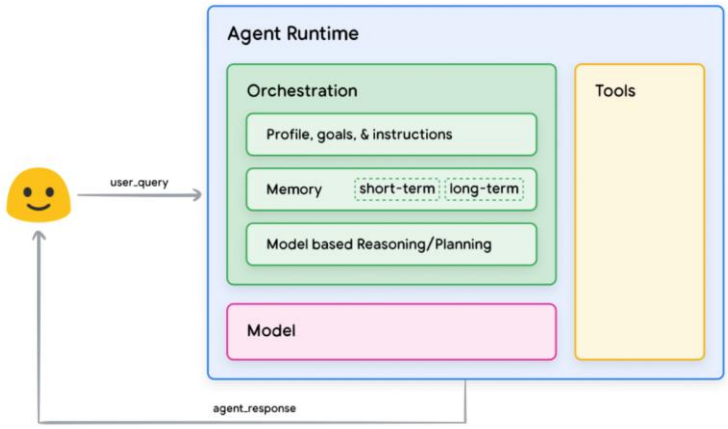
图7：提示链工作流



数据来源《Building effective agents》，东吴证券研究所
注：工作流包括提示链/路由/并行化/工作程序/评估器等，此次仅以提示链举例

对于开放式问题 Agent 必不可少，要具备理解复杂的输入、进行推理和规划、可靠地使用工具的能力。Agent 可用于开放式问题，这些问题难以预测所需的步骤，无法硬编码固定路径。Agent 需要独立规划和操作，并可能返回用户那里获取更多信息或判断。在执行工作中， Agent 在每一步从环境中获取“真实情况”，以评估进展，在遇到响应节点或者障碍的时候暂停以获取人类反馈。**人类只需要提出需求+监管成果，AI 将进行任务的分解、选择工具、监控进度。**

图8：AI Agent 架构图

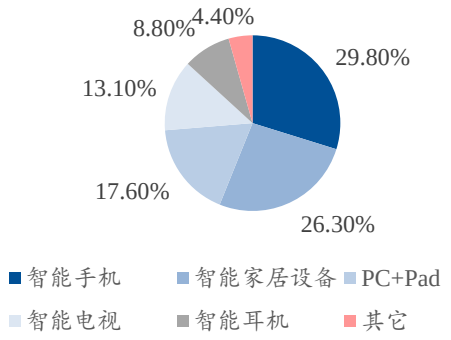


数据来源：谷歌《Agent》，东吴证券研究所

1.4. 消费级终端带动端侧 AI 高速发展

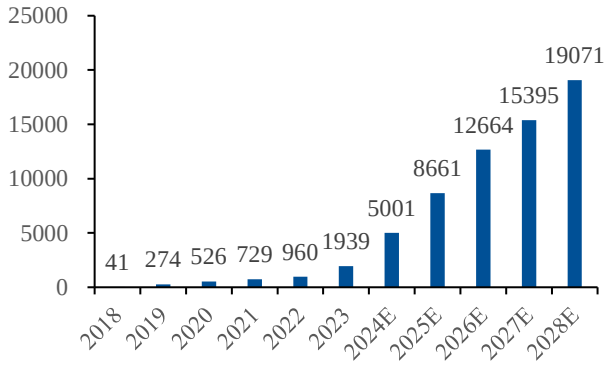
AI 可以改造多种终端，端侧 AI 市场规模 2023-2028 年预计 CAGR 高达 58%，2028 年超过 1.9 万亿。2023 年全球存量消费终端设备达 228 亿台，其中智能手机占 29.8%、智能家居设备（不含 TV）占 26.3%，PC 和 PAD 占 17.6%。2023 年以前端侧 AI 技术已经在智能安防和车载设备两个重要领域应用，快速发展但规模不大。从 2023 年开始，随着亿级出货量的 PC 和手机开始 AI 化，两者庞大的市场将在未来支撑端侧 AI 行业迅速发展，2023 年中国端侧 AI 市场规模不到 2000 亿，预计 2028 年超过 1.9 万亿，2023-2028 年 CAGR 为 58%。

图9：存量消费终端设备结构（%，2023 年）



数据来源：华为《AI 终端白皮书》，东吴证券研究所

图10：中国端侧 AI 行业市场规模，2018-2028E（亿元）



数据来源：头豹，东吴证券研究所

2. 端侧模型逐步迭代，巨头引领行业发展

2.1. 基础的小模型迭代加速，技术创新大于参数量提升

通过技术进步，相同规模模型性能持续提升。Gemma2 是谷歌 24 年 6 月发布的模型，通过架构和技术的改进提升了在规模相当的情况下的性能。如 Gemma2 和 Gemma1 分别为 2T 和 3T tokens 上训练，但是各类测试机的表现都大幅提升，包括评估语言理解能力的 MMLU 提升约 10pct，评估数学能力的 GSM8K 表现提升 9pct 等。同时横向比较看，Gemma-2 2B 到 Gemma-2 9B 的表现提升远高于 9B 到 27B 的提升。

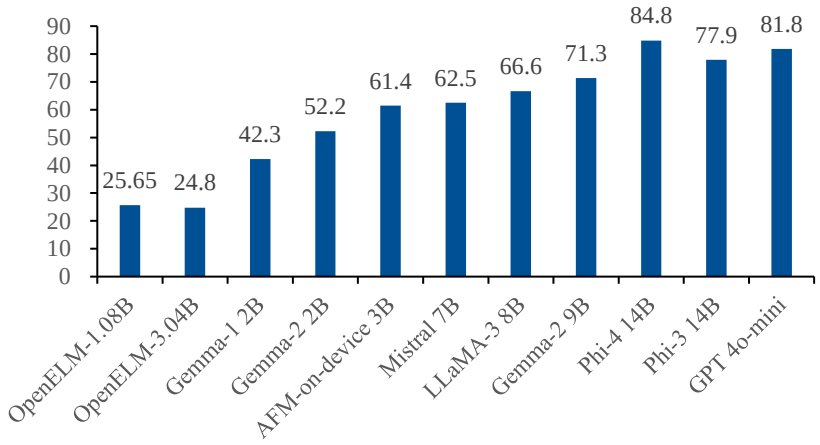
若选用 MMLU 数据集进行端侧模型的比较，首先我们认为随着新技术的出现模型性能会快速提升，如 Gemma1 和 Gemma2。其次整体的模型性能和模型参数呈现明显的正向关系。另外数据集的差异、蒸馏技术的使用等会带来相同模型巨大的性能差异，如均为苹果发布的 3B 模型，开源的 OpenELM-3B 性能远低于 AFM-on-device。时间相近的各家发布的相同参数的模型性能基本处在同一档次。

图11：小型模型比较

Benchmark	metric	Gemma-1 2B	Gemma-2 2B	Mistral 7B	LLaMA-3 8B	Gemma-1 7B	Gemma-2 9B	Gemma-2 27B
MMLU	5-shot	42.3	52.2	62.5	66.6	64.4	71.3	75.2
ARC-C	25-shot	48.5	55.7	60.5	59.2	61.1	68.4	71.4
GSM8K	5-shot	15.1	24.3	39.6	45.7	51.8	68.6	74
AGIEval	3-5-shot	24.2	31.5	44.0+	45.9*	44.9+	52.8	55.1
DROP	3-shot, F1	48.5	51.2	63.8*	58.4	56.3	69.4	74.2
BBH	3-shot, CoT	35.2	41.9	56.0°	61.1	59.0°	68.2	74.9
Winogrande	5-shot	66.8	71.3	78.5	76.1	79	80.6	83.7
HellaSwag	10-shot	71.7	72.9	83	82	82.3	81.9	86.4
MATH	4-shot	11.8	16	12.7	-	24.3	36.6	42.3
ARC-e	0-shot	73.2	80.6	80.5	-	81.5	88	88.6
PIQA	0-shot	77.3	78.4	82.2	-	81.2	81.7	83.2
SIQA	0-shot	49.7	51.9	47.0*	-	51.8	53.4	53.7
Boolq	0-shot	69.4	72.7	83.2*	-	83.2	84.2	84.8
TriviaQA	5-shot	53.2	60.4	62.5	-	63.4	76.6	83.7
NQ	5-shot	12.5	17.1	23.2	-	23	29.2	34.5
HumanEval	pass@1	22	20.1	26.2	-	32.3	40.2	51.8
MBPP	3-shot	29.2	30.2	40.2*	-	44.4	52.4	62.6
Average (8)		44	50	61	61.9	62.4	70.2	74.4
Average (all)		44.2	48.7	55.6	-	57.9	64.9	69.4

数据来源：谷歌，东吴证券研究所

图12：各小型模型 MMLU 基准测试得分



数据来源：谷歌、苹果、微软等，东吴证券研究所

2.2. 端侧模型需要进一步提升参数量以提高性能

端侧模型需要提高参数量以提高各类任务能力。Gemini Nano 是谷歌 Gemini 系列中专用于设备端部署的小模型，其中 Nano 1 和 2 的参数量分别为 1.8B 和 3.25B，针对低/高内存设备。从对应的测试集看，1）参数对模型表现影响巨大，1.8B 参数的 Nano 1 全面弱于 Nano 2。2）缩减参数后模型在回答真实性方面依然保持相当的准确率，但是在推理/编码/数学方面准确率较低，相比 Gemini Pro 准确率大幅下降。

图13：Gemini Nano 与 Pro 性能比较

	Gemini Nano 1 1.8B		Gemini Nano 2 3.25B	
	准确率	相对于Gemini Pro比例	准确率	相对于Gemini Pro比例
BoolQ	71.6	81%	79.3	90%
TydiQA（GoldP）	68.9	85%	74.2	91%
NaturalQuestions(Retrieved)	38.6	69%	46.5	83%
NaturalQuestions (Closed-book)	18.8	43%	24.8	56%
BIG-Bench-Hard (3-shot)	34.8	47%	42.4	58%
MBPP	20	33%	27.2	45%
MATH (4-shot)	13.5	41%	22.8	70%
MMLU (5-shot)	45.9	64%	55.8	78%

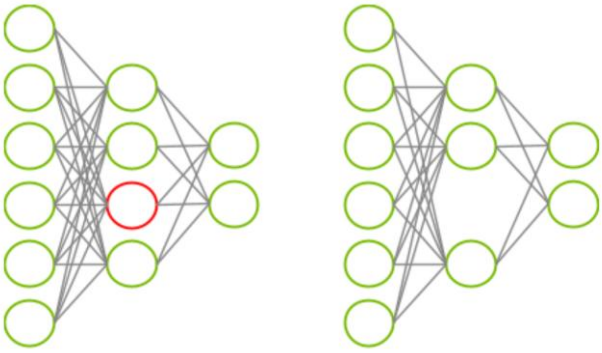
数据来源：谷歌，东吴证券研究所

2.3. 量化/剪枝/蒸馏技术压缩模型以降低硬件要求

量化/剪枝/蒸馏是主要的模型压缩方式。为在有限的硬件资源上部署更多参数的模型（或在模型保持基本性能的情况下降低对硬件的需求），需要对模型进行压缩。其中量化/剪枝/蒸馏是主要方法。**量化**：用低精度数值表示参数，可以减少模型的内存占用和计算开销，比如从 32bits 转化为 8bits，内存开销为原来的 1/4，计算成本仅为原来的 1/16。**剪枝**：删除不必要的神经元/权重参数/节点等。如下图所示，修剪前需求执行 32 次乘法累加和 32 个参数（权重）存储在内存中，修剪后只需要 24 次乘法累加和 24 个参数，计算复杂性和内存都降低了 25%。**蒸馏**：将大模型作为教师模型，用其输出训练一个性能接近但更轻量化的学生模型。如 Gemma 的 2B 和 9B 模型由 27B 的模型蒸

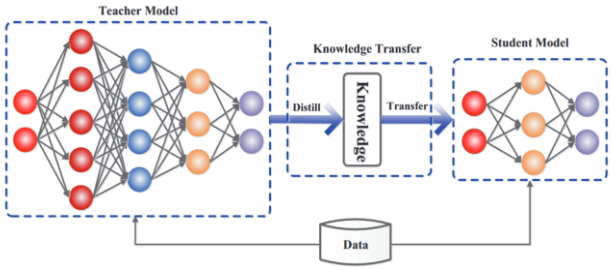
馏而来，苹果 3B 的 AFM-on-device 也是由剪枝后的 6.4B 模型蒸馏而来。

图14：小型化技术-剪枝原理示意图



数据来源：NVIDIA，东吴证券研究所

图15：小型化技术-蒸馏原理示意图



数据来源：《Knowledge Distillation：A Survey》，东吴证券研究所

参数量-性能的取舍带来更多样的创新方向以提升效率。多重制约带来各类方法百花齐放，如高质量的数据集，高效的训练方式、先进的压缩方法等创新方向。如苹果使用混合精度量化实现 3.7bits 的量化水平后，针对量化后的质量损失，苹果加入准确率恢复适配器实现了近乎无损的量化压缩。

图16：量化与精度恢复对模型性能的影响（以未量化的模型表现为 1）

量化精度	模型	IFEval Instruction-Level	AlpacaEval 2.0 LC	GSM8K (8-shot CoT)
16	AFM-on-device	100.0%	100.0%	100.0%
3.5	quantized	98.4%	76.7%	82.2%
	acc-recovered（rank 16）	98.8%	94.7%	92.1%
3.7	quantized	97.9%	87.3%	91.3%
	acc-recovered（rank 16）	100.6%	94.8%	96.0%

数据来源：苹果，东吴证券研究所

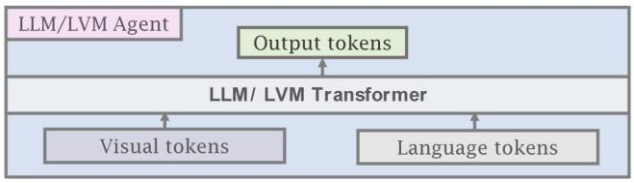
2.4. Agent 架构差异带来数据困境，Transformer 是转折点

基础模型本身：

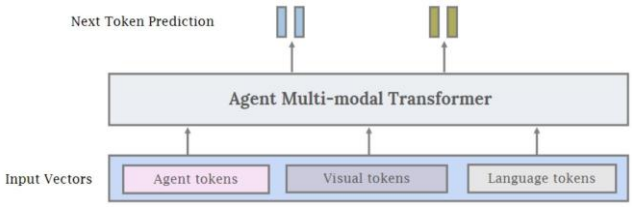
端侧是训练环节采用新的 Agent Transformer 架构，需要引入新的输入类型，成为 VLA 模型。视觉语言模型（VLMs）和大语言模型（LLMs）在任务规划（做什么）方面展现出了颇具潜力的能力。但是 Agent 的每项任务都需要底层控制策略（怎么做），才能在与环境的交互中取得成功，因此需要引入第三种通用数据类型——Agent tokens。即 Agent 模型是一个 VLA 模型（Vision-Language-Action Model，即视觉-语言-动作模型），是一个融合了视觉、语言和动作的多模态大模型范式。

图17：多模态大模型架构

图18：Agent Transformer 架构



数据来源：《AGENT AI》，东吴证券研究所



数据来源：《AGENT AI》，东吴证券研究所

Transformer 带来状态-动作交互策略学习机遇，有望大幅提升泛化能力。由于互联网的数据以静态数据组成，无法捕捉人类决策和环境交互作用，因此 Agent tokens 非常不足。动作数据收集成本过高，传统的训练模式下存在泛化能力弱，自我容错率低，成本过高的问题。但是由于大模型具备模糊匹配能力，在预训练过程中，通过模糊匹配而非精准映射，学习策略进而提高泛化能力。**同时我们更看好消费电子 GUI 成为最先落地的产品形态。**

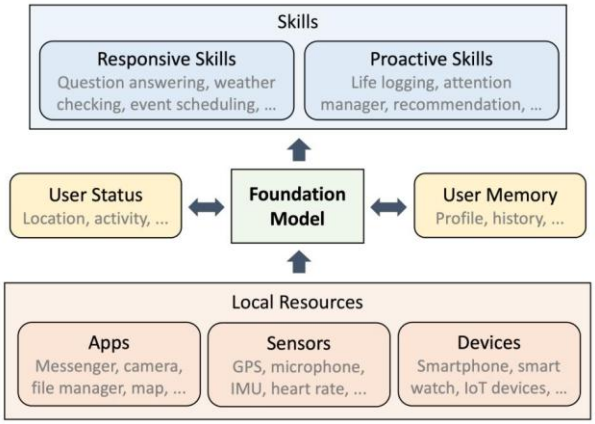
基础模型之外：Agent 还增加了个性化和内存操作要求，

需要基础大模型基础上，个人 LLM 大模型还需要具备任务执行、情境感知和记忆能力，考验终端厂商要求极高。个人 LLM 大模型需要具备：任务执行（将用户的指令或主动感知到的任务转化为针对个人资源的操作行动）、情境感知（感知用户及环境的当前状态，为任务执行提供全面的信息）、记忆（记录用户数据，使智能体能够回顾过往事件、总结知识并实现自我进化）。

图19：AI 终端关键技术特征

数据来源：华为《AI 终端白皮书》，东吴证券研究所

图20：端侧 Agent 的组成部分



数据来源：《PERSONAL LLM AGENTS》，东吴证券研究所

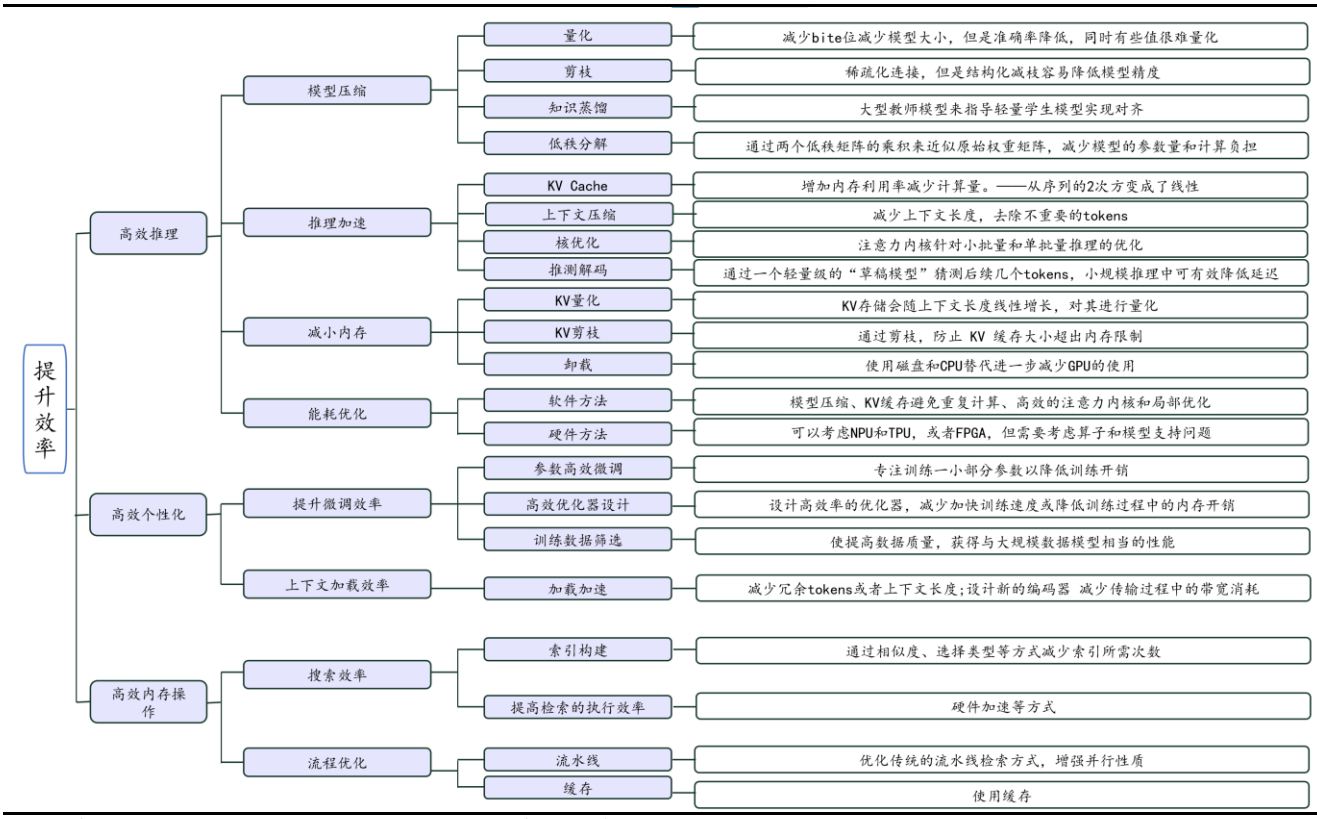
推理能力之外，端侧模型对个性化和内存操作要求较高，需要其它额外的优化。情境感知和记忆能力由更基础的流程制程，主要包括智能体的推理、个性化和记忆检索。比如任务执行过程中需要任务分解/规划，背后是推理能力；需要切换任务，需要依靠个性化来调整任务的权重；同时执行过程中需要检索历史数据，背后是内存操作能力。

图21：AI Agent 底层能力和高级能力之间的映射关系

底层能力 / 高级功能	推理	个性化	内存操作
任务执行	任务分解 决策制定	任务切换	检索历史数据
情境感知	环境感知 用户感知		存储感知结果
记忆		通过微调进行自我进化	增加和维护内存

数据来源：《PERSONAL LLM AGENTS》，东吴证券研究所

图22：提升 Agent 效率的技术及其简要说明



数据来源：《PERSONAL LLM AGENTS》，东吴证券研究所

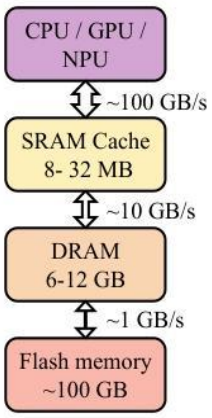
3. 硬件升级满足高性能需求，适配核心在内存

3.1. 内存是端侧硬件 AI 推理能力的短板

内存和由内存操作带来的能耗是端侧设备 AI 能力最短板。模型的优化和压缩的最终目的是为了提升端侧 AI 推理能力，硬件是端侧模型第二个必要的组成部分，硬件也将补齐短板。内存及其操作带来的能耗是当前最短板。Meta 指出，由于 SRAM 缓存通常在 20MB，仅能容纳一个 Transformer 块，而 Flash 容量足够但速度较慢，DRAM 是要和系统和其它应用共享，留给模型的 DRAM 空间更小，单个应用程序不应超过 DRAM 的 10%。苹果论文《LLM in a flash》指出，在 LLM 推理阶段，仅将 7B 参数，半精度的 LLM 的参数加载入 DRAM 所需空间就超过 14GB。微软的 3.8B 参数的 Phi-3-mini 模型在 4 bits 量化水平下需要占用 1.8GBDRAM。

同时根据论文《EIE》，在相同的精度下，DRAM 耗能比 SRAM 耗能高两个数量级（640：5），比计算耗能高更多。内存读取是主要的耗能项。根据每 10 亿参数生成 1token 需要耗能 0.1J，一个满电的 iPhone 大约 50KJ，以 10Token/s 的速度仅支持 7B 的模型运行不到 2 小时。同时能耗增加带来散热问题突出。

图23：常见移动设备中的内存层级结构



数据来源：《MobileLLM》，东吴证券研究所

图24：计算和存储操作的能耗

Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit 32KB SRAM	5	50
32 bit DRAM	**640**	**6400**

数据来源：《EIE》，东吴证券研究所
注：该能耗为在 LPDDR2、45nm 工艺情况下计算

对比各家 SoC，苹果在内存/电池/散热上提升空间巨大。芯片制程上看，苹果一般会优先拿到最先进的制程，但是苹果的 DRAM 容量较安卓相差较大，目前最新手机仅为 8GB，相比安卓旗舰最高 24GB 差异明显。苹果的设备功耗极限潜力很高但受限于散热无法全部发挥。同时苹果和安卓手机的电池续航相差极大，苹果增加电池容量的空间巨大，作为对比，iPhone 16 Pro Max 电池容量为 4685mAh，小米 15 Pro 为 6100mAh。

图25：主流手机 SoC 对比

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
NPU	45 TOPS	80 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（最高）	24GB	24GB	24GB		8GB	8GB
TDP	6.3W	8.2W	7W	-	8W	8W
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、OPPO Find X7 Ultra、一加12/Ace 3 Pro/Ace 5、Redmi K70 Pro/K80、三星Galaxy S24/S24+/S24 Ultra/Galaxy Z Flip6/Galaxy Z Fold6/W25/W25 Flip等	小米15/小米15 Pro/小米15 Ultra、REDMI K80 Pro、红魔10 Pro、一加13/一加Ace 5 Pro、realme GT7 Pro、iQOO 13、荣耀Magic7/Magic7 Pro/Magic7	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：苹果、高通等，东吴证券研究所

3.2. 安卓和 iOS 内存利用效率差异大

内存作为端侧 AI 提升速度、规模和效率的短板，预计成为硬件核心变革方向。从当前 iOS 和安卓的内存占用来看，普遍 iOS 内存利用效率更高尤其是在应用程序上。根据 Android authority，相同的 App 在 iOS 的内存占用远低于安卓。主要原因是安卓为消除硬件差异性代码运行在虚拟机中，需要编译为中间语言，而 iOS 则是原生编写因此内存占用较小，而游戏基本用游戏引擎均为原生编写差异较小。同时在国内由于安卓监管较少，各 App 为保活留存后台，集成功能等也增加了 App 的内存占用。因此我们认为安卓需要在 OS 层提供统一的 AI 基础模型，而 iOS 在模型压缩之外则需要提高硬件内存以克服单个功能占用大量内存的硬件瓶颈。

图26：安卓和 iOS 运行 App 时内存占用对比

App name	iOS(MB)	Android(MB)
Acrobat Reader	117	390
Booking.com	73	330
Gmail	190	259
Google Maps	224	300
Youtube	176	282
eBay	69	300
Google Photos	136	281
Twitter	100	366

数据来源：Androidauthority，东吴证券研究所

图27：安卓和 iOS 运行游戏时内存占用对比

Game	iOS(MB)	Android(MB)
Subway Surfers	500	761
1945 Airforce	550	852
Candy Crush	219	289
Brawl Stars	572	507
Minecraft	462	803
Asphalts 9	749	803
Shadowgun Legends	1130	899
Elder Scrolls Blade	1030	952

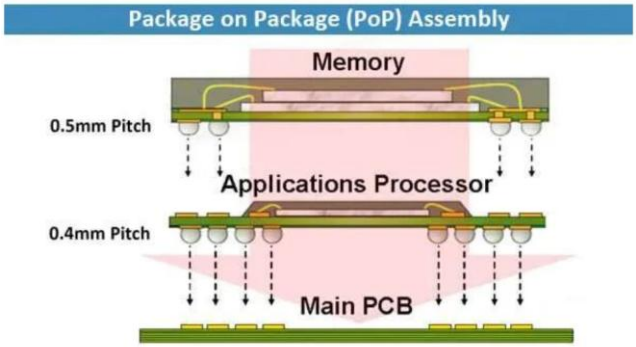
数据来源：Androidauthority，东吴证券研究所

3.3. 苹果硬件积极应变，创新方案集中内存方向

除了增加内存容量之外，苹果在内存结构、耗能、传输速度等方面创新密集。

苹果合作三星研发独立封装形式。韩媒 The Elec 报道，三星应苹果要求，开始研究新的 LPDDR DRAM 封装方式——独立封装，改变 2010 年起，iPhone 沿用至今的堆叠式封装（POP）方案——内存直接叠在 SoC 上通过 Pitch 连接以最大限度减少设备体积。**但是 PoP 技术由于芯片过于靠近，端侧 AI 负载要求下带宽和散热问题严重。**通过分开封装 DRAM 和 SoC，可以增加 I/O 引脚数量，提高数据传输速率和并行数据通道数量，并改善散热性能，显著提高内存带宽并增强 iPhone 的 AI 能力。同时三星还尝试在 iPhone DRAM 中应用 LPDDR6-PIM（内存内置存储器）技术，该技术的数据传输速度和带宽是 LPDDR5X 的两到三倍，专为设备端 AI 设计。

图28：PoP 封装的形式



数据来源：IT 之家，东吴证券研究所

全新的 WMCM 封装方式进一步提高芯片组合的灵活性和集成度。预计 2026 年苹果芯片先进封装方式将从现在的 InFO（集成扇出型）改为 WMCM 的封装方式（Wafer-Level Chip-Scale Packaging）。相比于现在的单芯片封装形式，WMCM 可将多芯片集成在同一封装中，可以开发更复杂的芯片，将 CPI、GPU、DRAM、NPU 等灵活的排列集成在一个封装中。WMCM 在信号传输方面表现出色，能减少信号延迟和干扰，对需要高速数据处理的设备尤为重要。

图29：苹果芯片封装形式改变

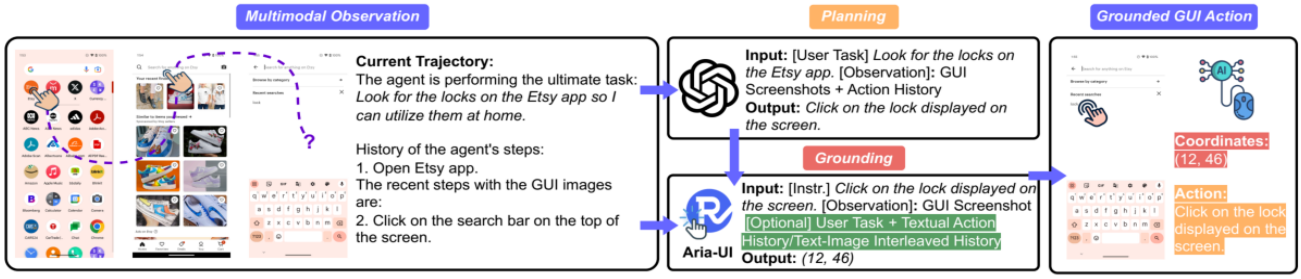
数据来源：IT 之家，东吴证券研究所

4. 多模态 UI 交互界面革命带来 Agent 的历史机遇

4.1. Transformer 架构带来 UI 交互的机遇

AI 智能助理的任务执行过程先后分为两个环节，包括规划和定位。其中规划包括使用哪些工具、哪些步骤进行执行等。定位可理解为执行过程中具体的工具（Tools）、技能或者坐标进行对外部环境的操作。

图30：Agent 任务执行过程（以 GUI 界面为例）



数据来源：《Aria-UI: Visual Grounding for GUI Instructions》，东吴证券研究所

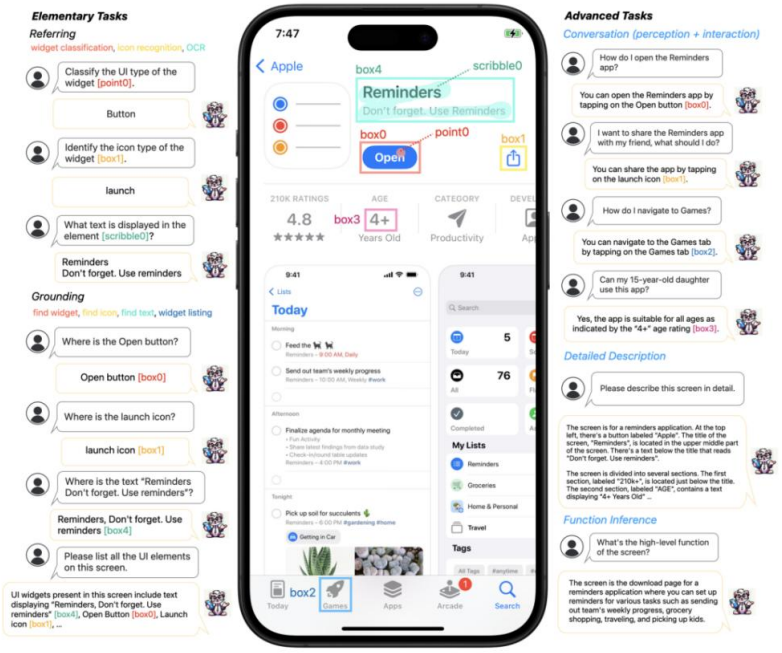
根据交互的模式，任务执行方法可分为基于 API 和基于用户界面（UI）的方法，GUI 有望在 Transformer 加持下成为主流。API 方式在运行时需要利用模型和内置的示例等方式来选择合适的 API 进行交互。但是 API 交互方式依赖公开可用的 API，存在应用程序支持不够全面、同时存在人类可以轻松执行但是 API 难以实现的任务、拓展性和自动化能力较弱等不足。UI 界面方式在 Transformer 架构下较好克服了任务和 UI 元素之间的隐含关系，大幅提升了 GUI Agent 的可行性。

图31：智能体和 API 之间的调用方式示例（Extension 方式）

数据来源：Google《Agent》，东吴证券研究所

GUI 交互可进一步分为基于文本的 GUI 交互和多模态的 GUI 交互，当前多模态 GUI 交互方式中 Grounding 难度比 Planning 更大。视觉定位阶段（Grounding）需要将规划好的指令精准映射到实际界面元素上，确保操作的准确执行。其中 Grounding（精准定位）由于设备兼容性/指令多样性/场景复杂性等原因，当前难度甚至比规划决策阶段更大，元素识别错误也是当前 LMMs GUI 模型的主要错误来源。多模态 GUI 交互方式数据不足、存在屏幕录制要求高以及当前 LMMs 模型推理能力有限的困扰。

图32：UI 界面复杂，元素识别困难

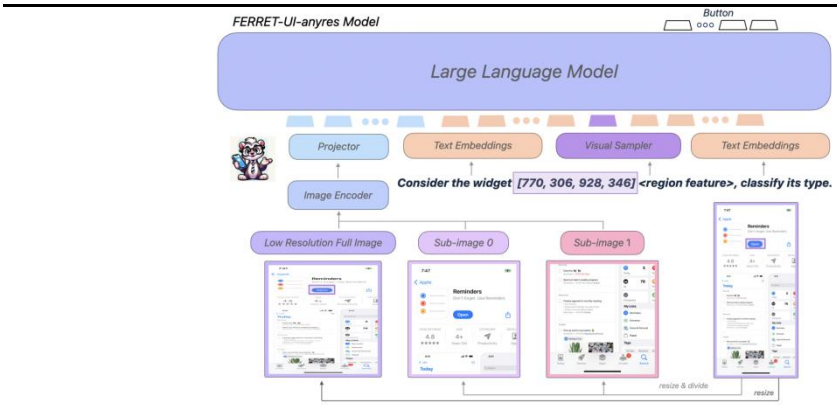


数据来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，东吴证券研究所

4.2. 苹果和谷歌均发力 UI 交互模型

苹果在 24 年 4 月推出 Ferret-UI 模型，Ferret-UI 通过一个视觉编码器对图像进行编码，苹果 Ferret-UI 加入任意分辨率（Anyres）技术切割放大子图像来解决 UI 交互中的小型对象识别问题，同时采用混合区域表示的方法进一步提高图像的特征提取能力，将图像特征和文本指令联合编码后进行推理，并精准根据任务要求执行。从测试看，模型针对定位任务/识别任务和高级任务表现均较为优秀。

图33：Ferret-UI-Anyres 架构



数据来源：苹果，东吴证券研究所

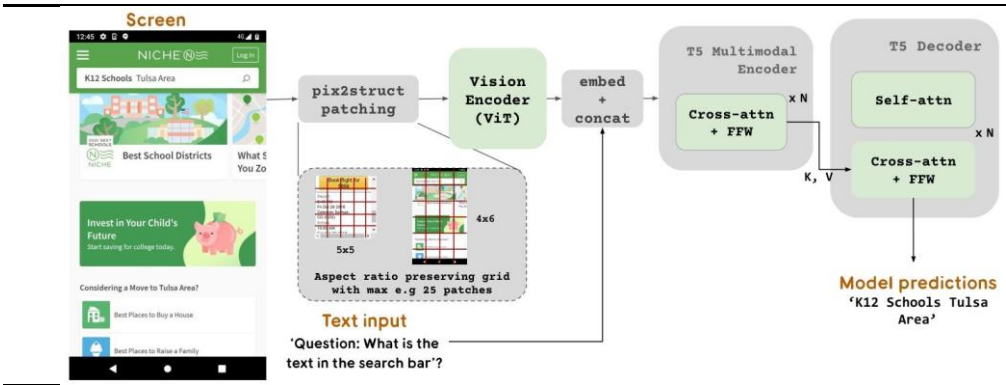
图34：Ferret-UI-anyres 准确率表现优异

	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**

数据来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，东吴证券研究所
注 1： S2W : screen2words, WiC : widget captions, TaP: taperception
注 2： “i”: iPhone, “A”: Android, “Ref”: Referring, “Grd”: Grounding

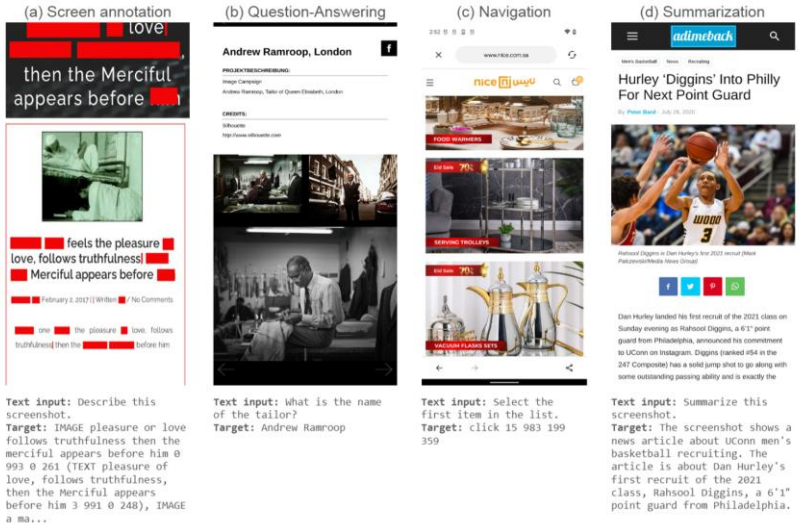
Screen AI 是谷歌 24 年 2 月推出的可读屏 AI 视觉语言模型，专门设计用于理解和处理用户界面（UI）和信息图标，能够理解和生成与屏幕元素相关的文本，如屏幕信息理解、问题回答、UI 导航指令、内容摘要等。从架构上，ScreenAI 使用视觉编码器和语言编码器组成的多模态编码器。视觉编码器基于 Vision Transformer（ViT）架构，将输入的屏幕截图作为一系列图像嵌入，语言编码器则处理截图相关的文本信息，如 UI 元素标签和描述等。编码器的输出被传递给一个解码器 T5,负责生成文本输出，能够根据输入的图像和文本嵌入生成自然语言响应。同时与苹果类似，谷歌采用更灵活的屏幕分割策略以适用于不同的手机和电脑屏幕。

图35：谷歌 Screen AI 模型架构



数据来源：谷歌，东吴证券研究所

图36：Screen AI 模型的屏幕理解/问答/导航/摘要四类任务示例



数据来源：谷歌，东吴证券研究所

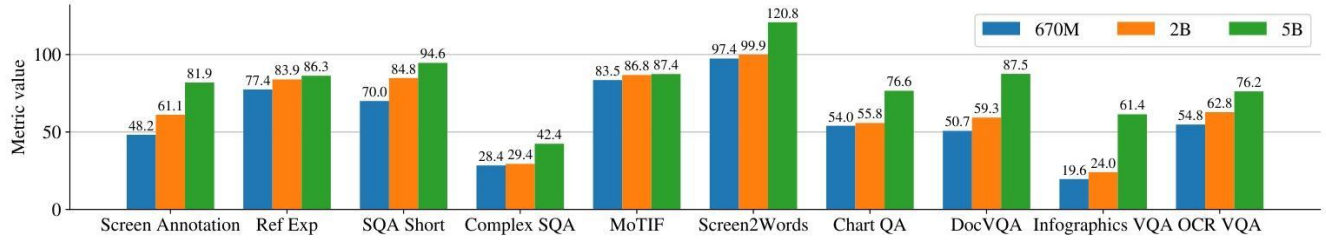
模型参数提升对性能影响较大，尤其是在复杂任务中，5B 模型对性能提升尚未饱和。谷歌发布了三个 670M/2B 和 5B 三个大小的模型，参数差异主要在图像和文本编解码器的大小，不同的编码器支持不同的输入，如正方形下 670M 模型最大输入分辨率为 720 × 720，5B 模型则支持 812 × 812 的最大输入分辨率。从最后的模型表现中，所有的任务中增加模型规模都会提升性能，并且在最大的 5B 模型时性能提升尚未饱和。同时对于复杂任务，2B 和 5B 模型之间的性能提升远大于 670M 模型到 2B 模型之间的提升。

图37：谷歌不同参数 Screen AI 模型参数量及其分配细节

模型	图形编码器	文本编解码器	参数量	最大输入分辨率
670M	B16（92M）	mT5 base（583M）	675M	720×720
2B	H14（653M）	mT5 Large（1.23B）	1.88B	756×756
5B	G14（1.69B）	UL2-3B（2.93B）	4.62B	812×812

数据来源：谷歌，东吴证券研究所

图38：不同参数的 Screen UI 模型的表现



数据来源：谷歌，东吴证券研究所

5. 风险提示

底层创新不及预期：端侧 AI 模型的发展需要底层技术架构和模型优化技术、硬件技术的创新，若创新不及预期，模型体验较差会影响用户体验。

技术发展不及预期：除了底层创新，新技术如新存储方式等的落地也需要后端制造环节的技术落地和成本优化，若技术发展不及预期，产品成本或过高无法普及。

竞争加剧风险：当前互联网厂商/消费电子品牌厂商等均在抢占入口环节，竞争或加剧。

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