

AI专题·从模型视角看端侧AI

模型技术持续演进，交互体验有望升级

西南证券研究发展中心
海外研究团队
2024年7月

核心观点

- ❑ **基础的构建：模型实现高效压缩是端侧AI的第一步。**模型尺寸变小、同时具备较好性能，是端侧AI的前提。目前，在10B参数规模以下的模型中，7B尺寸占据主流，3B及以下小模型仍在探索，部分小模型性能正逐步接近更大参数模型，如谷歌Gemini-Nano模型在部分测试基准上接近Gemini-Pro、Meta Llama-3-8B模型表现可与Llama-2-70B匹敌。模型厂商为兼顾模型尺寸与性能，在算法优化上进行积极探索，在模型压缩技术、稀疏注意力机制、多头注意力变体等领域取得持续进展，帮助模型减少参数、降低存算需求，同时保持较好的性能，为端侧AI奠定小模型的基础。
- ❑ **落地的关键：模型适配终端硬件是端侧AI的第二步。**小语言模型（SLM）不完全等于端侧模型，在模型实现高效压缩后，需要进一步与手机硬件进行适配，帮助小模型装进终端。从众多小模型论文中可以发现，当前主要存在内存、功耗、算力三大硬件瓶颈。其中，苹果在其论文《LLM in a flash》中指出，70亿半精度参数的语言模型，完全加载进终端需要超过14GB的DRAM空间；Meta在其MobileLLM模型论文中指出，一个约有5000焦耳满电能量的iPhone，仅支持7B模型在10 tokens/秒的AI生成速率下对话不足2小时。为解决以上问题，手机芯片厂商正加速推进AI芯片研发，在先进制程、内存容量及带宽、CPU和GPU性能、以及AI服务器上发力，手机品牌商也将配备更高性能的电池、散热元器件，提升整体终端硬件能力，更好地支持AI模型。
- ❑ **体验的突破：模型助力人机交互是端侧AI的第三步。**端侧模型通常能够支持用户完成AI初级任务，然而更丰富、更深度的交互体验需要UI模型、云端模型、以及系统级AI进行有力支撑。其中，UI模型可以提供手机UI界面理解的基础，云端模型能够帮助处理较为复杂的交互任务，系统级AI可以实现多种模型间的调用与协同。在AI时代下，模型的端侧意义不止于类似ChatGPT的聊天机器人软件，而在于赋能手机系统和应用交互的系统级AI，其带来的交互体验将成为影响用户换机的核心。从当前的海外合作阵营来看，可分为“苹果+OpenAI”和“谷歌+高通+三星”两大阵营。未来，随着端侧模型、配套硬件、AI系统的持续发展，终端市场有望呈现更多可能。
- ❑ **相关标的：**苹果(AAPL.O)、三星电子(005930.KS)、高通(QCOM.O)、谷歌(GOOGLE.O)等。
- ❑ **风险提示：**端侧AI技术进展不及预期风险；行业竞争加剧风险；应用开发不及预期风险等。

目 录

1 基础的构建：模型实现高效压缩是端侧AI的第一步

1.1 十亿级参数模型加速迭代，性能表现向百亿参数模型靠拢

1.2 模型压缩技术助力端侧部署，注意力优化机制降低存算需求

2 落地的关键：模型适配终端硬件是端侧AI的第二步

2.1 从小模型论文看端侧硬件瓶颈：内存/功耗/算力

2.2 从芯片厂商布局看硬件升级趋势：制程/内存/NPU/电池/散热

3 体验的突破：模型助力人机交互是端侧AI第三步

3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

1 模型实现高效压缩是端侧AI的第一步

海外小模型发展概况

模型
优化
技术

模型压缩：
知识蒸馏、量
化、剪枝等

稀疏注意力机制：
滑动窗口注意力机制、全局
注意力机制等

多头注意力变体：
分组查询注意力机制、多头
隐式注意力机制等

Flash attention等
...

技术支持

技术支持

模型

Gemma-2

Gemini-Nano

Llama-3.1

MobileLLM

Phi-3

OpenELM

Mistral

公司

Google

Meta

Microsoft

Apple

Mistral

训练
GPU
类型

TPUv4, TPUv5e

A100, H00

A100, H100

A100, H100

算力租赁等

特点

Gemma-2基于
Gemma-1优化模型
具体细节；
Gemini-Nano致力
于在终端设备上运
行；GQA由谷歌
创新提出

Llama追求数据上的
scaling law，Llama-
3.1加入多模态/多语
言/长文本/实用工具
等能力；MobileLLM
强调小模型的深度比
宽度更重要

Phi-1专注于
编码；Phi-2
开始学习推理；Phi-3擅
长编码和推理；强调数
据的小而精

核心目标在
于服务终端
设备及应用

欧洲LLM领
先独角兽

追求方向

追求方向

性能

将模型大小压缩至10B参数以下，性能向10B~100B级别参数的模型靠拢

1.1 小模型24H1加速迭代，模型性能持续提升

- ❑ **发展节奏**：24H1小模型加速推出，Meta Llama领先发布，微软、谷歌相继迭代，苹果厚积薄发。
- ❑ **模型参数**：7B模型占据主流；3B及以下小模型进一步探索，其中苹果小模型梯队分布明显。
- ❑ **训练数据**：Meta在有限参数下追求数据量上的scaling law；微软专注小而精的数据集；苹果旗下小模型的训练数据量与参数量的比值不低。
- ❑ **算力消耗**：23年GPU大多采用A100，24年主要采用H100；谷歌使用自研TPU；创企选择上云等。

23H2及24H1海外小模型版本迭代情况

公司	模型名称	发布日期	模型参数量 (B)	预训练数据量 (B Tokens)	预训练数据量与模型参数量的比值	GPU型号	预训练耗时
Google	Gemma-2-9B	2024年6月27日	9	8000	889	4096张TPUv4	/
	Gemma-2-2.6B	训练中	2.6	2000	769	512张TPUv5e	/
	Gemma-1-7B	2024年2月21日	7	6000	857	4096张TPUv5e	/
	Gemma-1-2B	2024年2月21日	2	3000	1500	512张TPUv5e	/
	Gemini-Nano-3.25B	2023年12月6日	3.25	/	/	TPUv5e or TPUv4	/
	Gemini-Nano-1.8B	2023年12月6日	1.8	/	/	TPUv5e or TPUv5	/
Meta	Llama-3-8B	2024年4月18日	8	15000	1875	H100	1300000小时
	Llama-2-7B	2023年7月18日	7	2000	286	A100	184320小时
	Llama-1-7B	2023年2月24日	7	1000	143	A100	82432小时
	MobileLLM-125M	2024年2月22日	0.125	250	2000	32张A100	/
	MobileLLM-350M	2024年2月22日	0.35	250	714	32张A100	/
微软	Phi-3-small-7B	2024年4月23日	7	4800	686	Phi-3系列模型中的Phi-3-medium (14B) 模型在512块H100上训练耗时42天	
	Phi-3-mini-3.8B	2024年4月23日	3.8	3300	868		
	Phi-2	2023年12月12日	2.7	1400	519	96块A100	14天
	Phi-1.5	2023年9月11日	1.3	30	23	A100	1500 小时
	Phi-1	2023年6月20日	1.3	7	5	4块A100	4天
苹果	OpenELM-0.27B	2024年4月25日	0.27	1500	5556	128块A100	3天
	OpenELM-0.45B	2024年4月25日	0.45	1500	3333	128块H100	3天
	OpenELM-1.08B	2024年4月25日	1.08	1500	1389	128块A100	11天
	OpenELM-3.04B	2024年4月25日	3.04	1500	493	128块H100	13天

1.1.1 谷歌Gemma系列模型：基于第一代模型架构对技术细节进行优化

- 基于千张TPU集群训练，模型性能在同类中较为领先。1) **Gemma-2-9B**：在4096张TPUv4上进行训练，在多数基准中得分超过Llama-3-8B和Mistral-7B等同类模型，MMLU 5-shot、GSM8K 5-shot的测试得分相较于前一代模型Gemma-1-7B分别有11%和32%的增长。2) **Gemma-2-2.6B**：在512张TPUv5e上进行训练，沿用第一代模型架构，对技术细节进一步优化，Gemma-2-2.6B模型较上一代Gemma-1-2.5B模型在参数量基本不变和数据集更小的情况下实现更优性能，MMLU 5-shot、GSM8K 5-shot的测试得分相较于上一代模型分别有21%和58%的增长。

谷歌Gemma系列模型性能情况

模型测试基准		Gemma-1-2.5B	Gemma-2-2.6B	Mistral-7B	LLaMA-3-8B	Gemma-1-7B	Gemma-2-9B
MMLU	5-shot	42.3	51.3	62.5	66.6	64.4	71.3
ARC-C	25-shot	48.5	55.4	60.5	59.2	61.1	68.4
GSM8K	5-shot	15.1	23.9	39.6	45.7	51.8	68.6
AGIEval	3-5-shot	24.2	30.6	44.0	45.9	44.9	52.8
DROP	3-shot,F1	48.5	52.0	63.8	58.4	56.3	69.4
BBH	3-shot,CoT	35.2	41.9	56.0	61.1	59.0	68.2
Winogrande	5-shot	66.8	70.9	78.5	76.1	79.0	80.6
HellaSwag	10-shot	71.7	73.0	83.0	82.0	82.3	81.9
MATH	4-shot	11.8	15.0	12.7		24.3	36.6
ARC-e	0-shot	73.2	80.1	80.5		81.5	88.0
PIQA	0-shot	77.3	77.8	82.2		81.2	81.7
SIQA	0-shot	49.7	51.9	47.0		51.8	53.4
Boolq	0-shot	69.4	72.5	83.2		83.2	84.2
TriviaQA	5-shot	53.2	59.4	62.5		63.4	76.6
NQ	5-shot	12.5	16.7	23.2		23.0	29.2
HumanEval	pass@1	22.0	17.7	26.2		32.3	40.2
MBPP	3-shot	29.2	29.6	40.2		44.4	52.4

1.1.2 谷歌Gemini-Nano系列模型：部分任务性能距Gemini Pro较小

- ❑ **专为设备部署而设计，擅长总结和阅读理解。** 2023年12月6日，谷歌发布Gemini系列自研大模型，参数规模从大至小分别为Gemini-Ultra、Gemini-Pro、Gemini-Nano，其中Gemini-Nano模型包括两种版本，Nano-1参数规模为1.8B，Nano-2为3.25B，旨在分别针对低内存和高内存的设备。
- **Gemini-Nano-1和Nano-2模型与参数规模更大的Gemini-Pro模型对比来看：**1) 根据BoolQ基准（主要用于衡量模型理解问题和回答问题的逻辑能力）得分，Gemini-Nano-1的准确率为71.6%，性能是Gemini-Pro的81%，Gemini-Nano-2的准确率为79.3%，是Gemini-Pro的90%，更接近Gemini-Pro的性能；2) TydiQA(GoldP)基准涉及回答复杂问题的能力，Gemini-Nano-1和Gemini-Nano-2的准确率为68.9%和74.2%，分别是Gemini-Pro的85%和91%，性能差距较小。
- **Gemini-Nano-1和Gemini-Nano-2模型对比来看：**随着模型参数规模从Nano-1的1.8B增加至Nano-2的3.25B，模型的性能表现在大多数任务性能均能得到提升。

谷歌Gemini-Nano系列模型性能情况

模型测试基准	Gemini-Nano-1 (1.8B)		Gemini-Nano-2	
	准确率 (%)	相对于Gemini Pro的比例	准确率 (%)	相对于Gemini Pro的比例
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%

1.1.3 Meta Llama系列模型：在有限参数下追求数据上的scaling law

- 同等参数情况下性能大幅提升，较小模型可以通过扩大训练数据量实现优秀性能。1) 对比同等参数模型来看，Llama-3的8B和70B模型相对于Llama-2的7B和70B模型性能均得到大幅提升。2) 对比Llama-3-8B和Llama-2-70B来看，在算力消耗基本持平的情况下，更好的模型性能可以通过在更大规模的数据集上训练实现，Llama-3-8B模型的参数量约为Llama-2-70B的1/9，但训练数据集是其7.5倍，最终的模型效果基本可与70B的模型相匹敌，且经过指令微调后，指令微调模型Llama-3-8B明显超过Llama 2 70B。

Meta Llama系列模型性能情况

指标			Llama 3		Llama 2	
模型阶段	类别	基准	Llama 3 70B	Llama 3 8B	Llama 2 70B	Llama 2 7B
预训练模型	General	MMLU (5-shot)	79.5	66.6	69.7	45.7
		AGIEval English (3-5 shot)	63.0	45.9	54.8	28.8
		CommonSenseQA (7-shot)	83.8	72.6	78.7	57.6
		Winogrande (5-shot)	83.1	76.1	81.8	73.3
		BIG-Bench Hard (3-shot, CoT)	81.3	61.1	65.7	38.1
		ARC-Challenge (25-shot)	93.0	78.6	85.3	53.7
	Knowledge reasoning	TriviaQA-Wiki (5-shot)	89.7	78.5	87.5	72.1
	Reading comprehensive	SQuAD (1-shot)	85.6	76.4	82.6	72.2
		QuAC (1-shot, F1)	51.1	44.4	49.4	39.6
		BoolQ (0-shot)	79.0	75.7	73.1	65.5
		DROP (3-shot, F1)	79.7	58.4	70.2	37.9
指令微调模型	多任务语言理解推理	MMLU (5-shot)	82.0	68.4	52.9	34.1
	专业知识推理能力	GPQA (0-shot)	39.5	34.2	21.0	21.7
	代码生成能力	HumanEval (0-shot)	81.7	62.2	25.6	7.9
	数学（小学数学问题）	GSM-8K (8-shot, CoT)	93.0	79.6	57.5	25.7
	数学（数学工具和函数）	MATH (4-shot, CoT)	50.4	30.0	11.6	3.8

1.1.4 Meta MobileLLM系列模型：强调小模型的深度比宽度更重要

- **模型参数进一步缩小，模型架构追求深而窄。** MobileLLM的模型参数仅为1.25亿和3.5亿，其技术报告聚焦于少于10亿参数的sub-billion（< 1B）模型，强调模型架构对小模型的重要性，**认为模型深度比宽度更重要**，并引入分组查询注意力机制等优化技巧，相较于同类125M/350M大小模型的基准测试得分相比，MobileLLM的平均分均有提高。**1) Zero-Shot常识推理任务方面：**在125M参数量级下，MobileLLM的模型性能显著优于OPT、GPT-Neo、Calaclafa等其他模型；在350M参数量级下，MobileLLM的各项测试得分均优于此前最先进的模型OPT-350M。**2) 问答和阅读理解任务方面：**根据在TQA问答的benchmark和RACE阅读理解的benchmark的测评结果，MobileLLM-125M和MobileLLM-350M模型的精度比同等量级的小模型要高出较多。

Meta MobileLLM系列模型性能情况

模型测试基准		MobileLLM-125M	Galactica-125M	OPT-125M	GPT-neo-125M	MobileLLM-350M	OPT-350M
ARC-e	0-shot	43.9	44.0	41.3	40.7	53.8	41.9
ARC-c	0-shot	27.1	26.2	25.2	24.8	33.5	25.7
BoolQ	0-shot	60.2	54.9	57.5	61.3	62.4	54.0
PIQA	0-shot	65.3	55.4	62.0	62.5	68.6	64.8
SIQA	0-shot	42.4	38.9	41.9	41.9	44.7	42.6
HellaSwag	0-shot	38.9	29.6	31.1	29.7	49.6	36.2
OBQA	0-shot	39.5	28.2	31.2	31.6	40.0	33.3
WinoGrande	0-shot	53.1	49.6	50.8	50.7	57.6	52.4
RACE	Acc, middle	39.7		34.7	34.7	45.6	37.1
RACE	Acc, high	28.9		27.5	27.0	33.8	28.0
TQA	F1 score, 1-shot	13.9		8.7	8.0	22.0	11.0
TQA	F1 score, 5-shot	14.3		9.6	7.9	23.9	12.3
TQA	F1 score, 64-shot	12.5		8.2	5.0	24.2	10.4

1.1.5 微软Phi系列模型：主要创新在于构建教科书质量的训练数据集

- **训练数据追求小而精，模型参数逐步扩大。**2023年6月，微软发布论文《Textbooks Are All You Need》，用规模仅为7B tokens的“教科书质量”的数据集，训练出1.3B参数、性能良好的Phi-1模型。此后，历代Phi模型沿用“Textbooks Are All You Need”的训练思想，进一步使用精挑细选的高质量内容和过滤的Web数据来增强训练语料库，以提升模型性能。在最新迭代的模型中，Phi-3-mini-3.8B通过3.3T tokens的训练，在学术基准和内部测试上可与经过15T tokens训练的Llama-3-In-8B模型相匹敌。

微软Phi系列模型性能情况

模型测试基准		Phi-3-mini-3.8b	Phi-3-small-7b	Phi-2-2.7b	Mistral-7b	Gemma-1-7b	Llama-3-In 8b
MMLU	5-Shot HKB*21	68.8	75.7	56.3	61.7	63.6	66.5
HellaSwag	5-Shot ZHB*19	76.7	77.0	53.6	58.5	49.8	71.1
ANLI	7-Shot NWD*20	52.8	58.1	42.5	47.1	48.7	57.3
GSM-8K	8-Shot, CoT CKB*21	82.5	89.6	61.1	46.4	59.8	77.4
MedQA	2-Shot JPO*20	53.8	65.4	40.9	50.0	49.6	60.5
AGIEval	0-Shot ZCG*23	37.5	45.1	29.8	35.1	42.1	42.0
TriviaQA	5-Shot JCWZ17	64.0	58.1	45.2	75.2	72.3	67.7
Arc-C	10-Shot CCE*18	84.9	90.7	75.9	78.6	78.3	82.8
Arc-E	10-Shot CCE*18	94.6	97.0	88.5	90.6	91.4	93.4
PIQA	5-Shot BZGC19	84.2	86.9	60.2	77.7	78.1	75.7
SociQA	5-Shot BZGC19	76.6	79.2	68.3	74.6	65.5	73.9
BigBench-Hard	3-Shot, CoT SRR*22 SSS*22	71.7	79.1	59.4	57.3	59.6	51.5
WinoGrande	5-Shot SLBBC19	70.8	81.5	54.7	54.2	55.6	65.0
OpenBookQA	10-Shot MCKS18	83.2	88.0	73.6	79.8	78.6	82.6
BoolQ	2-Shot CLC*19	77.2	84.8		72.2	66.0	80.9
CommonSenseQA	10-Shot THLB19	80.2	80.0	69.3	72.6	76.2	79.0
TruthfulQA	10-Shot, MC2 LHE22	65.0	70.2		53.0	52.1	63.2
HumanEval	0-Shot CTJ*21	58.5	61.0	59.0	28.0	34.1	60.4
MBPP	3-Shot AON*21	70.0	71.7	60.6	50.8	51.5	67.7

1.1.6 苹果OpenELM系列模型：核心目标在于服务终端设备及应用

- ❑ **致力于服务终端设备，模型性能整体表现出色。** OpenELM的模型参数包括2700万、4500万、11亿和30亿四种大小，相较于市场主流的70亿参数模型，更加轻巧精悍，致力于让主流笔记本电脑和部分高性能智能手机也能承载和运行高性能模型。根据官方信息，OpenELM在同类模型表现较好：
- **OpenELM-1.08B：在使用较少预训练数据（仅为艾伦人工智能研究所AI2 Labs推出的先进开源模型—OLMo-1.18B模型的1/2）的情况下，性能超越OLMo，提升幅度达2.36%。**
- **OpenELM-3B：在衡量知识推理能力的ARC-C基准上，准确率为42.24%；在MMLU和HellaSwag两项基准测试中，得分分别为26.76%和73.28%，首批试用者反馈OpenELM模型表现稳定且一致性高，不易产生过于激进或不当内容的输出。**

苹果OpenELM系列小模型性能情况

模型测试基准		OpenELM-0.28B	MobiLlama-0.50B	OpenELM-0.45B	MobiLlama-0.80B	MobiLlama-1.26B	OLMo-1.18B	OpenELM-1.08B	OpenELM-3.04B
MMLU	5-shot	25.72	26.09	26.01	25.2	23.87	26.16	27.05	26.76
ARC-C	25-shot	27.65	29.52	30.2	30.63	34.64	34.47	36.69	42.24
CrowS-Pairs	25-shot	66.79	65.47	68.63	66.25	70.24	69.95	71.74	73.29
HellaSwag	10-shot	47.15	52.75	53.86	54.17	63.27	63.81	65.71	73.28
PIQA	0-shot	69.75	71.11	72.31	73.18	74.81	75.14	75.57	78.24
SciQ	0-shot	84.7	83.6	87.2	85.9	89.1	87	90.6	92.7
WinoGrande	5-shot	53.83	56.27	57.22	56.35	60.77	60.46	63.22	67.25
ARC-e	0-shot	45.08	46.04	48.06	49.62	56.65	57.28	55.43	59.89
BoolQ	0-shot	53.98	55.72	55.78	60.03	60.34	61.74	63.58	67.4
RACE	0-shot	30.91	32.15	33.11	33.68	35.02	36.75	36.46	38.76
TruthfulQA	0-shot	39.24	37.55	40.18	38.41	35.19	32.94	36.98	34.98
TruthfulQA-mc2	0-shot	39.24	37.55	40.18	38.41	35.19	32.94	36.98	34.98

1.2 模型架构持续优化，压缩技术不断创新

- 为压缩模型大小、在保持较小模型尺寸的同时实现高性能、以及能够支持较长的上下文，各海外模型厂商纷纷布局小模型，并在模型算法优化方面进行积极探索，于24H1呈现出多种技术创新方向，主要集中在**模型压缩技术**，**稀疏注意力机制**、**多头注意力变体**三大领域。

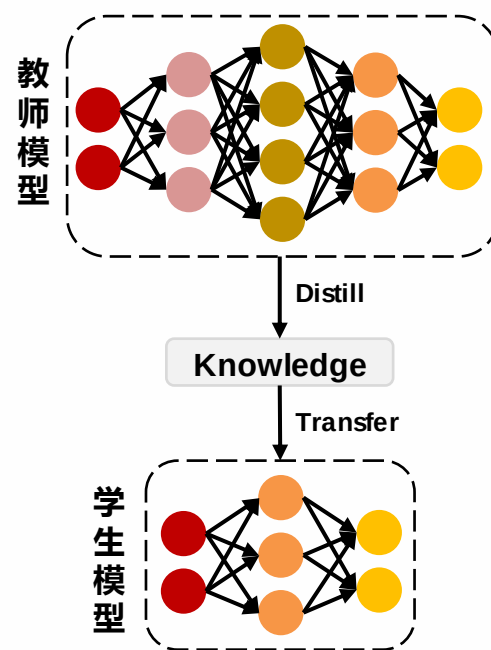
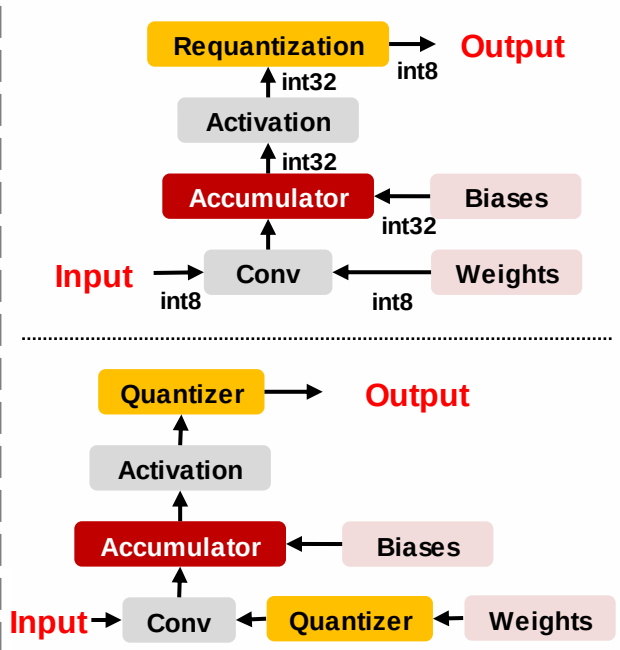
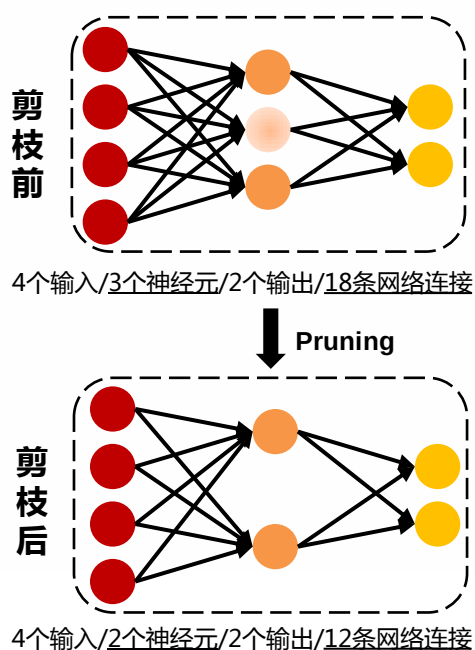
海外小模型架构优化及技术创新方向

公司	模型名称	发布日期	是否进行模型压缩？ 量化/剪枝/知识蒸馏	是否采用 稀疏注意力机制？	是否采用 FlashAttention？	是否采用 多头注意力变体？	支持的上下文长度 (tokens)
Google	Gemma-2-9B	2024年6月27日	知识蒸馏	滑动窗口&全局注意力	√	GQA	8,192
	Gemma-2-2.6B	训练中	/	滑动窗口&全局注意力	√	GQA	8,192
	Gemma-1-7B	2024年2月21日	/	/	√	MHA	8,192
	Gemma-1-2B	2024年2月21日	/	/	√	MQA	8,192
	Gemini-Nano-3.25B	2023年12月6日	量化、知识蒸馏	/	/	MQA	/
	Gemini-Nano-1.8B	2023年12月6日	量化、知识蒸馏	/	/	MQA	/
Meta	Llama-3-8B	2024年4月18日	/	/	/	GQA	8,192
	Llama-2-7B	2023年7月18日	知识蒸馏	/	/	GQA	4,096
	Llama-1-7B	2023年2月24日	/	/	/	MHA	2,048
	MobileLLM-125M	2024年2月22日	量化、知识蒸馏	/	/	GQA	/
	MobileLLM-350M	2024年2月22日	量化、知识蒸馏	/	/	GQA	/
微软	Phi-3-small-7B	2024年4月23日	/	局部块注意力	√	GQA	8,192
	Phi-3-mini-3.8B	2024年4月23日	量化	/	√	GQA	4,096
	Phi-2	2023年12月12日	/	/	√	MHA	2,048
	Phi-1.5	2023年9月11日	/	/	√	MHA	2,048
	Phi-1	2023年6月20日	/	/	√	MHA	2,048
苹果	OpenELM-0.27B	2024年4月25日	量化、知识蒸馏	/	√	GQA	2,048
	OpenELM-0.45B	2024年4月25日	量化、知识蒸馏	/	√	GQA	2,048
	OpenELM-1.08B	2024年4月25日	量化、知识蒸馏	/	√	GQA	2,048
	OpenELM-3.04B	2024年4月25日	量化、知识蒸馏	/	√	GQA	2,048

1.2.1 模型压缩技术：参数量化运用广泛，知识蒸馏热点较高

- **模型压缩技术持续发展，助力端侧部署。** 模型压缩技术旨在保持模型基本性能的情况下降低对推理算力的需求，主要包括三种方法：**1) 参数剪枝 (Pruning)**：删除部分权重参数、去除神经网络中的冗余通道、神经元节点等；**2) 参数量化 (Quantization)**：将浮点计算转成低比特定点计算，业内应用普遍；**3) 知识蒸馏 (Knowledge Distilling)**：将大模型作为教师模型，用其输出训练出一个性能接近、结构更简单的学生模型，由Geoffrey Hinton等人在2015年谷歌论文《Distilling the Knowledge in a Neural Network》中提出，目前关注较高，业内通常使用GPT-4和Claude-3作为教师模型。

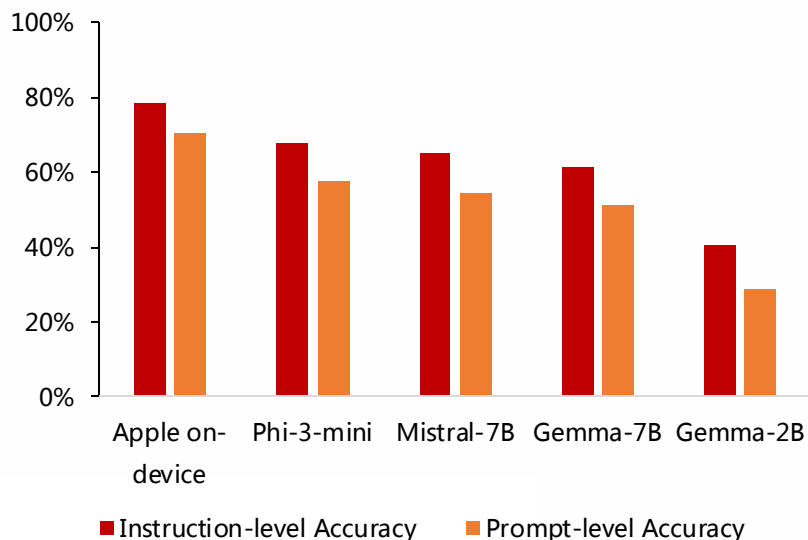
模型压缩的三种方法：剪枝/量化/知识蒸馏



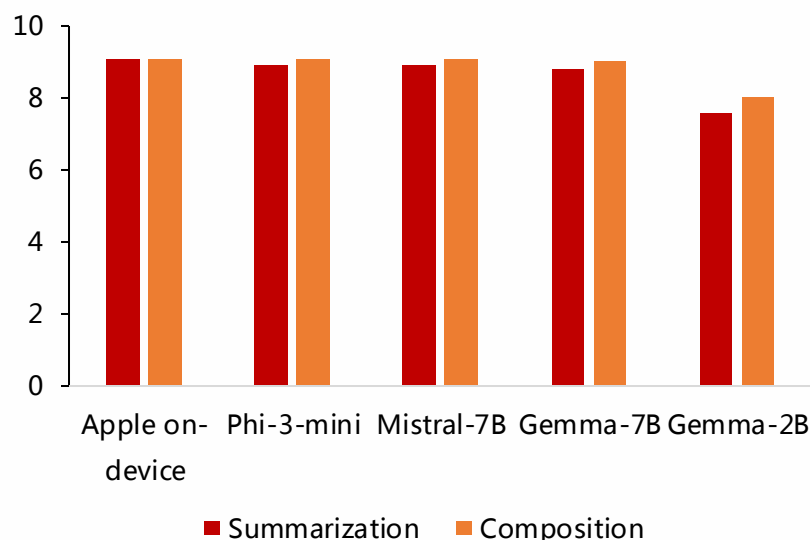
1.2.1 模型压缩技术：参数量化运用广泛，知识蒸馏热点较高

- ❑ **苹果OpenELM模型**：模型微调引入量化和知识蒸馏技术，提高模型泛化能力，帮助模型实现必要性能。根据2024年6月10日苹果发布的研究成果：
 - **1) 参数量化**：对于设备端推理，为保持模型质量，苹果采用混合2-bit和4-bit的配置策略，平均参数量化至3.5-bit，以实现与未压缩模型相同的准确性。
 - **2) 知识蒸馏**：苹果结合拒绝采样和知识蒸馏等多种技术，创新模型微调方法——a rejection sampling fine-tuning algorithm with teacher committee，其中，**Teacher Committee (教师委员会)**是指使用多个教师模型来指导学生模型的学习，每个教师模型可能具有不同的优势和专业领域，通过综合多个教师模型的知识，提供更全面、准确的指导，帮助学生模型更好地学习。

苹果端侧模型在指令遵循测评上得分更高



苹果端侧模型在指写作测评上得分更高



1.2.1 模型压缩技术：参数量化运用广泛，知识蒸馏热点较高

- ❑ **Meta MobileLLM模型**：采用量化和知识蒸馏技术，模型压缩后性能差距较小。根据2024年6月27日Meta发布的MobileLLM模型技术报告：
 - **1) 参数量化**：模型参数量化的消融实验分别对全精度BF16和量化后的W8A8（8位权重、8位激活）模型进行零样本常识推理任务测试，根据实验结果，量化后的模型效果相较于全精度BF16的模型，性能差距均在0.5以内，模型经过量化压缩后性能损失较小。
 - **2) 知识蒸馏**：在知识蒸馏的消融实验中，Meta将LLaMA-v2-7B作为教师模型，使用来自大型预训练教师模型（即LLaMA-v2-7B）和学生模型（MobileLLM-125M和350M模型）logits之间的交叉熵计算知识蒸馏损失（KD loss），再集成至小模型的预训练过程中。根据实验结果，MobileLLM-125M和350M模型经过教师模型的知识蒸馏后，性能误差分别分别仅为0.1和0.3。

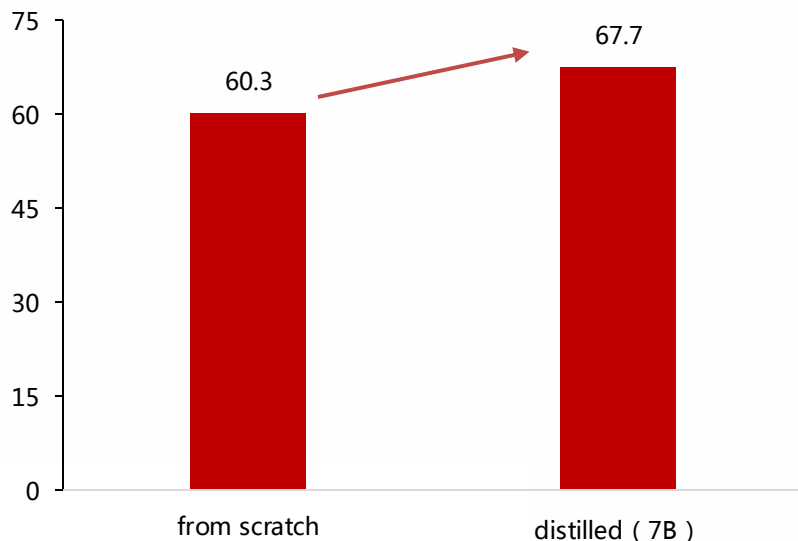
Meta MobileLLM模型关于参数量化和知识蒸馏的消融研究

消融研究	模型	精度	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.	Gap
量化	MobileLLM-125M	BF16	45.5	27.7	58.3	64.6	41.9	36.4	35.4	50.4	45.0	-
	MobileLLM-125M	W8A8	45.2	27.1	58.3	65.0	41.7	36.2	33.6	51.0	44.8	0.2
	MobileLLM-LS-125M	BF16	44.4	27.0	61.5	65.1	43.0	37.6	37.8	52.0	46.1	-
	MobileLLM-LS-125M	W8A8	44.0	27.5	60.9	64.6	43.1	37.7	37.7	51.0	45.8	0.3
	MobileLLM-350M	BF16	51.4	31.3	61.0	68.1	43.6	47.2	41.6	55.4	49.9	-
	MobileLLM-350M	W8A8	51.4	32.1	61.1	68.8	43.1	47.1	40.6	55.1	49.9	0.0
	MobileLLM-LS-350M	BF16	51.9	35.2	59.6	68.9	43.4	47.2	43.3	58.4	51.0	-
	MobileLLM-LS-350M	W8A8	51.3	33.8	59.5	69.1	43.7	47.2	43.0	57.0	50.6	0.4
知识蒸馏	125M model	Label	43.1	28.9	58.1	62.3	42.3	34.6	31.5	50.1	43.9	-
	125M model	Label+KD	41.8	28.5	58.5	61.6	41.1	34.5	32.7	51.6	43.8	0.1
	350M model	Label	50.2	31.8	56.9	67.7	44.3	45.8	40.8	55.5	49.1	-
	350M model	Label+KD	48.7	31.8	60.7	67.4	43.2	45.9	38.9	53.7	48.8	0.3

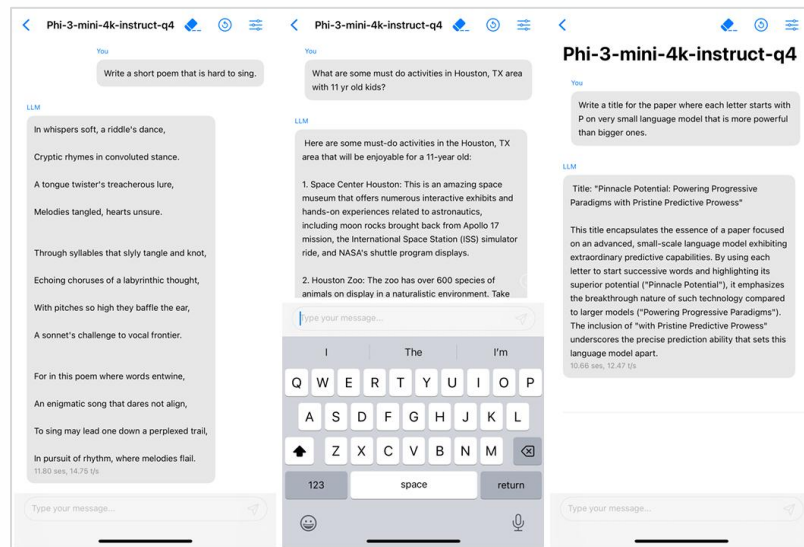
1.2.1 模型压缩技术：参数量化运用广泛，知识蒸馏热点较高

- ❑ **谷歌Gemini-Nano模型：知识蒸馏+量化压缩。** Nano的1.8B和3.25B模型由更大的Gemini模型知识蒸馏训练而来，并将其量化至4-bit，以便在低内存和高内存的设备上部署运行。
- ❑ **谷歌Gemma-2模型：通过大模型压缩蒸馏出一个小模型，再用数据去训练，比从头训练小模型的效果更好。** 根据谷歌技术报告，Gemma-2的9B和2.6B模型在训练策略上均采用知识蒸馏技术，使模型能够在相同训练数据体量下达到更好的效果。根据消融实验，基于500B tokens的训练数据集，由7B模型知识蒸馏后的2.6B模型，相较于一个从零开始训练的2.6B模型，三项基准测试均分更高。
- ❑ **微软Phi-3模型：量化压缩。** 微软将Phi-3-mini模型量化至4-bit，内存占用约为1.8GB，根据其端侧部署测试，该模型可在带有苹果A16仿生芯片的iPhone 14上以原生方式运行并完全离线。

Gemma-2-2.6B知识蒸馏后三项测试均分更高



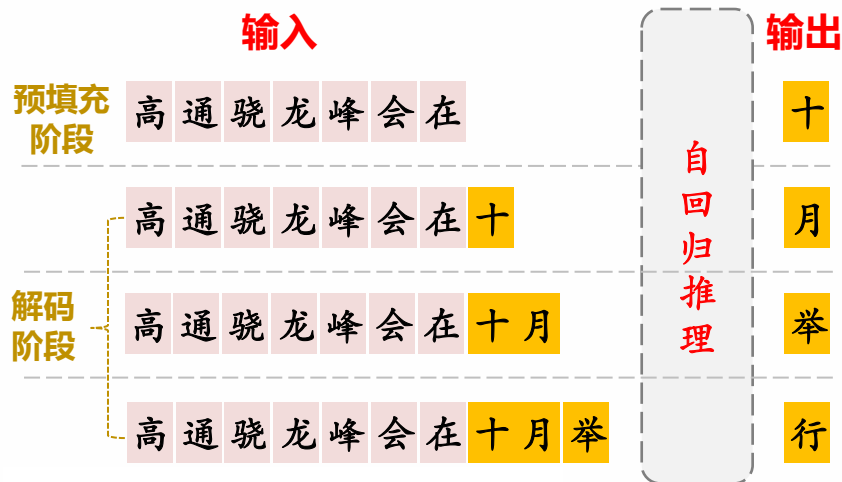
Phi-3-mini量化后在A16仿生芯片iPhone上运行



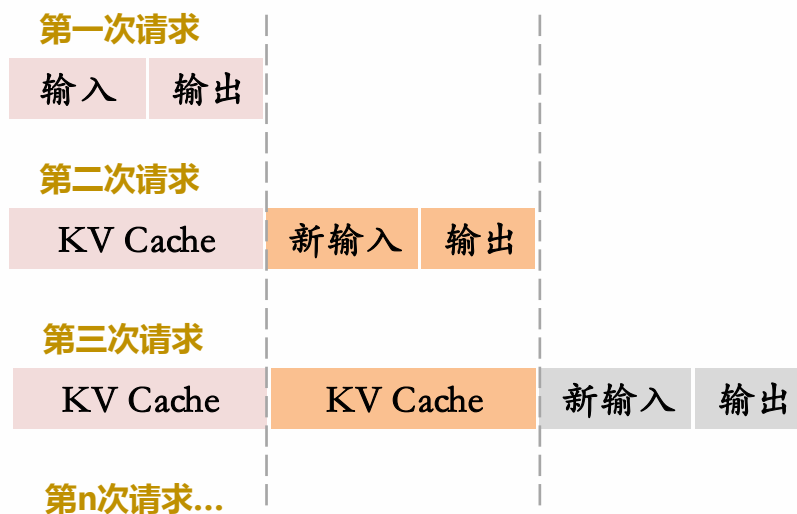
1.2.2 多头注意力变体：减少注意力头数量，降低内存占用

- ❑ **KV cache**：通过缓存中间计算结果，以“内存空间”换“计算时间”。当前，主流的大语言模型基本采用Transformer decoder-only架构，其推理过程主要包括预填充和解码阶段。**1) 预填充阶段**：根据用户提出的prompt，生成第一个token；**2) 解码阶段**：在生成第一个token之后，开始采用自回归方式逐个生成后续的token，每个token的生成均需要依赖并attention此前的token，因此，随着解码过程的进行，需要向此前生成的token的关注会越来越多，计算量也逐渐增大。
- ❑ 为减少解码过程中的重复计算，可以通过引入KV Cache，即缓存中间结果、在后续计算中直接从Cache中读取而非重新计算，从而实现“以空间换时间”，使显存占用增加、但计算需求减少。

LLM自回归推理过程示意图



LLM在多轮对话场景中引入KV Cache

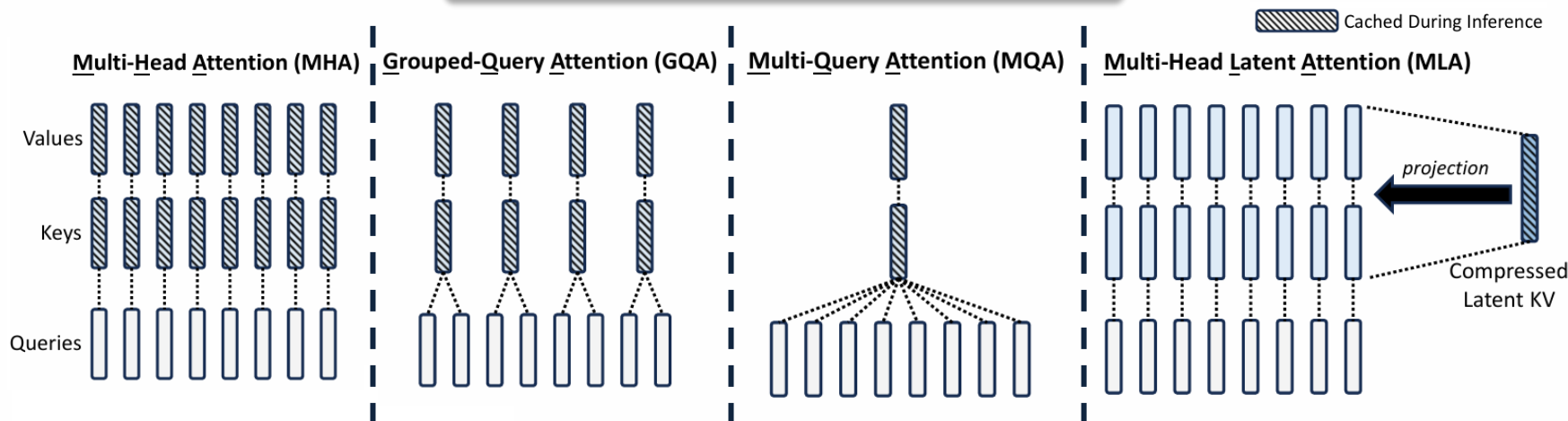


1.2.2 多头注意力变体：减少注意力头数量，降低内存占用

□ 为平衡模型性能与存算成本，产生多种注意力变体。对比各注意力变体的特征来看：

- ① **多头注意力机制 (MHA)**：1个Query Head对应1个KV Head，模型效果更好，但随着模型参数增长、以及更长的上下文，会形成过大的KV cache，从而带来明显的访存瓶颈。
- ② **多查询注意力机制 (MQA)**：只保留一个KV Head，通过多个Query Heads共享相同的KV Head，使模型内存占用减少、推理速度更快，但是性能损失较大。
- ③ **分组查询注意力机制 (GQA)**：将Query Heads进行分组，每组Query Heads对应一个KV Head，介于MHA和MQA之间，由多个Query共享一组KV，在减少内存占用的同时，提升数据处理速度，保持模型处理下游任务的性能。
- ④ **多头隐式注意力机制 (MLA)**：将KV值压缩至低维空间，减少模型推理的内存占用和计算需求。

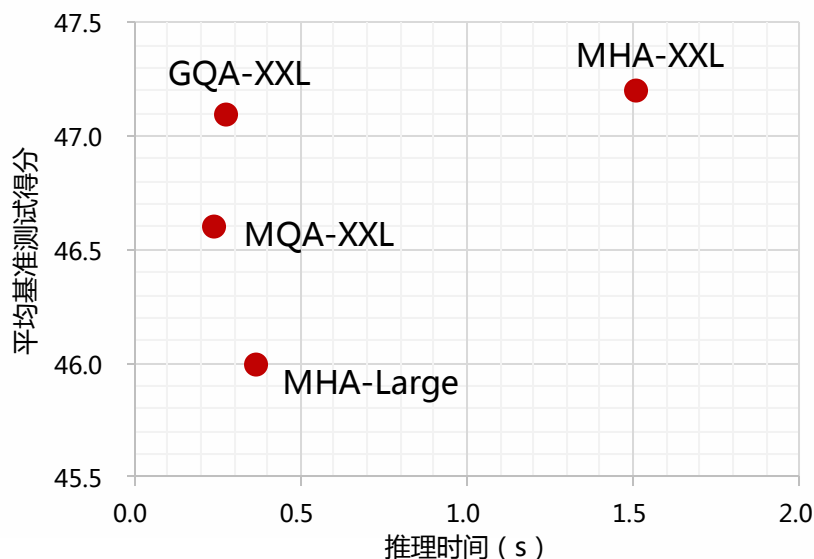
LLM推理中有关KV Cache的注意力机制及改进



1.2.2 多头注意力变体：减少注意力头数量，降低内存占用

- ❑ **GQA由谷歌率先提出，成为当前主流注意力变体。** GQA技术由Google Research团队于2023年12月提出，根据论文《GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints》中关于各种注意力变体的表现来看，MHA基准测试均分最高、但推理耗时较长，MQA推理时间最短，但模型性能略差，而**GQA能够平衡模型性能和推理速度**，在较短的推理时间内取得较好的表现性能。从模型当前采用程度来看，**截至24H1，GQA仅提出约半年时间，便在主流小模型中得到广泛采用**，谷歌的Gemma-2，微软的Phi-3、Meta的Llama-3和MobileLLM模型、苹果的端侧模型OpenELM，以及法国创企Mistral-7B更新版本均采用分组查询注意力机制。

MHA、GQA、MQA对比



采用GQA技术的主流小模型

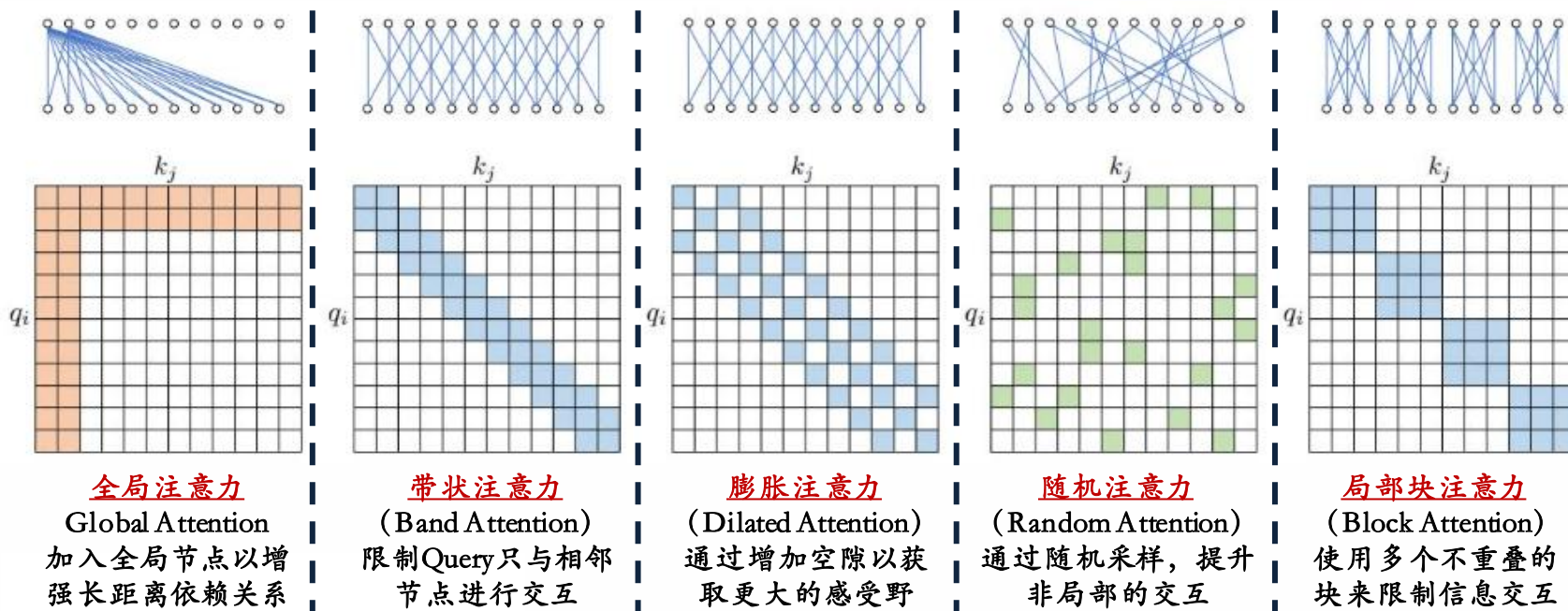
公司	模型名称	发布日期	注意力机制 (Attention variant)
Google	Gemma-2-9B	2024年6月27日	GQA
	Gemma-2-2.6B	训练中	GQA
Meta	Llama-3-8B	2024年4月18日	GQA
	Llama-2-7B	2023年7月18日	GQA
	MobileLLM-125M	2024年2月22日	GQA
	MobileLLM-350M	2024年2月22日	GQA
微软	Phi-3-small-7B	2024年4月23日	GQA
	Phi-3-mini-3.8B	2024年4月23日	GQA
苹果	OpenELM-0.27B	2024年4月25日	GQA
	OpenELM-0.45B	2024年4月25日	GQA
	OpenELM-1.08B	2024年4月25日	GQA
	OpenELM-3.04B	2024年4月25日	GQA
Mistral	Mistral-7B-v0.3	2024年5月22日	GQA
	Mistral-7B-v0.2	2024年3月24日	GQA

资料来源：《GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints》，西南证券整理

1.2.3 稀疏注意力机制：选择性处理信息，降低计算需求

- ❑ **稀疏注意力(Sparse Attention)机制**：选取一部分信息进行交互，节省注意力机制成本。在当前主流模型架构Transformer中，注意力矩阵可以通过限制Query-Key对的数量来减少计算复杂度，即将注意力机制稀疏化。稀疏注意力机制主要采用基于位置信息和基于内容的稀疏化方法，其中，基于位置信息的稀疏注意力方法更加主流，主要包括全局/带状/膨胀/随机/局部块五种类型。近年来，随着大语言模型的加速发展，计算和存储压力增大，使得稀疏注意力机制不断优化，逐步衍生出基于以上稀疏注意力机制的复合模式，涌现出Longformer等稀疏注意力模型。

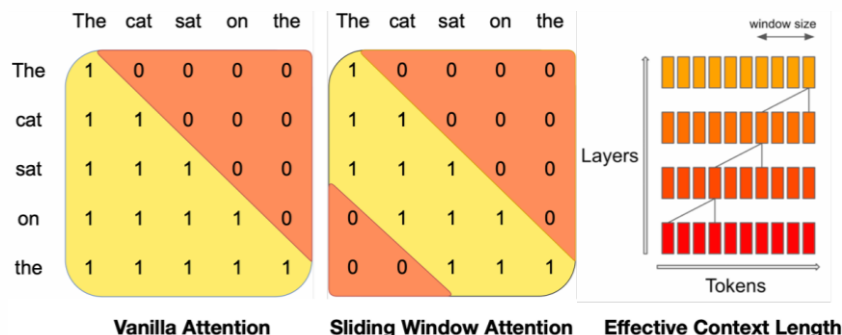
基于位置信息的注意力机制稀疏化方法



1.2.3 稀疏注意力机制：选择性处理信息，降低计算需求

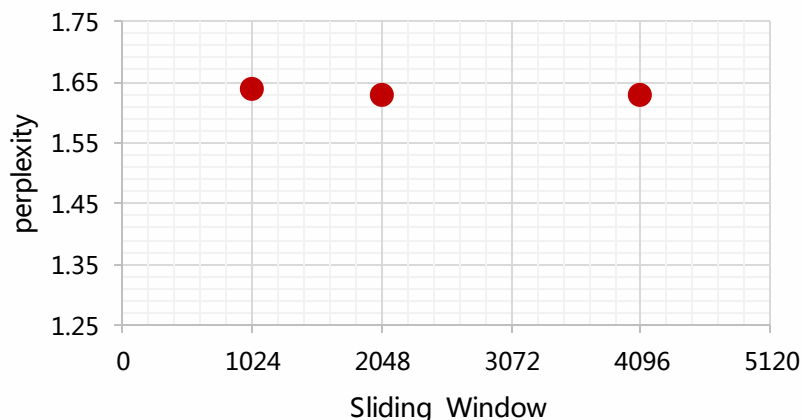
- ❑ 滑动窗口注意力 (Sliding Window Attention-SWA) 机制：关注临近位置信息，简化计算步骤。
- 1) Mistral-7B：创新使用SWA机制，解决长文本问题。SWA作为一种稀疏注意力机制，在输入序列中的每个token周围使用一个固定大小的窗口，其计算复杂度为 $O(s \times w)$ （其中 s 是输入序列的长度， w 是固定的窗口大小，且 $w < s$ ），相较于计算复杂度为 $O(s \times s)$ 的完全自注意力机制，会更加高效。在长文本情况下，一般相邻tokens的相关性更大，因此，在文本生成时并不需要对所有tokens计算注意力值，只需计算每个token前的 n 个tokens的注意力值，从而在更长的上下文情况下不增加KV Cache缓存的大小。
- 2) Gemma-2：交替使用局部滑动窗口和全局注意力，捕捉细节的同时保证全局理解。Gemma-2在架构上基本沿用第一代模型设计，在注意力机制上进行细节优化，实现局部滑动窗口和全局注意力的交替使用，其中，滑动窗口大小设置为4096 tokens，而全局注意力窗口为8192 tokens，滑动窗口注意力机制可以确保模型能够精确捕捉文本细节，全局注意力机制有助于保持模型对上下文的正确理解。

Mistral-7B：采用SWA机制解决长文本问题



滑动窗口注意力机制

Gemma-2：调整滑动窗口大小对困惑度影响较小



目 录

1 基础的构建：模型实现高效压缩是端侧AI的第一步

1.1 十亿级参数模型加速迭代，性能表现向百亿参数模型靠拢

1.2 模型压缩技术助力端侧部署，注意力优化机制降低存算需求

2 落地的关键：模型适配终端硬件是端侧AI的第二步

2.1 从小模型论文看端侧硬件瓶颈：内存/功耗/算力

2.2 从芯片厂商布局看硬件升级趋势：制程/内存/NPU/电池/散热

3 体验的突破：模型助力人机交互是端侧AI第三步

3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

2 小模型能上终端是端侧AI的第二步

手机终端硬件发展概况

从小模型论文
看硬件瓶颈

苹果论文《LLM in a flash》指出：7B参数、半精度的LLM，完全加载进终端所需的DRAM空间超过14GB

Meta MobileLLM论文指出：一个约有5000焦耳满电能量的iPhone，可支持7B模型在10tokens/秒的AI生成速率下进行对话不到2小时

Meta MobileLLM论文指出：用于计算的SRAM通常限制在约20MB左右，一般只能容纳一个单独的Transformer块

硬件瓶颈

硬件瓶颈

当前手机
硬件配置

	先进制程	最大显存	最大内存	最大带宽	L2-Cache	L3-Cache	AI算力	TDP
苹果 A17 Pro	TSMC 3nm	6 GB	8 GB	51.2 GB/s	20 MB	24 MB	35 TOPS	11W
联发科 天玑 9300	TSMC 4nm	/	16 GB	76.8 GB/s	/	18 MB	33 TOPS	12.5W
高通 骁龙8 Gen 3	TSMC 4nm	6 GB	24 GB	76.6 GB/s	/	12 MB	34 TOPS	12.5W

硬件升级

硬件升级

未来硬件
升级方向

先进制程：从4nm向3nm、从3nm向2nm升级

存储：从8+6GB向16+12GB扩容、从50GB/s、80GB/s向更高带宽升级

算力：从35TOPS向更高算力升级

电池：钢壳+硅碳负极+叠片工艺

散热：VC散热板+石墨烯

2.1 从小模型论文看端侧硬件瓶颈——内存容量

❑ 将LLM装进终端要求手机内存有多少DRAM容量？

- 苹果在其发布的论文《LLM in a flash》中指出：在通常的LLM推理阶段，LLM直接加载至DRAM中，一个7B参数、半精度的LLM，完全加载进DRAM所需的存储空间超过14GB。考虑到目前主流手机的DRAM最高也就16GB的水平，在端侧直接使用DRAM来加载7B LLM面临巨大挑战。

❑ 通常一个应用最多可以占用多少DRAM内存？

- Meta在其MobileLLM模型论文中指出：将8-bit量化权重下的LLaMA-2-7B模型整合至手机，内存代价过高，手机目前DRAM容量从iPhone 15的6GB到Google Pixel 8 Pro的12GB不等，由于DRAM需要与操作系统和其他应用程序共享，一个移动应用不应超过DRAM的10%（即1~2GB）。
- 微软在其Phi-3模型技术报告中指出，Phi-3-mini可在手机上实现本地推理，在3.8B尺寸、在量化为4-bit权重下，大约**占用1.8GB的内存**。

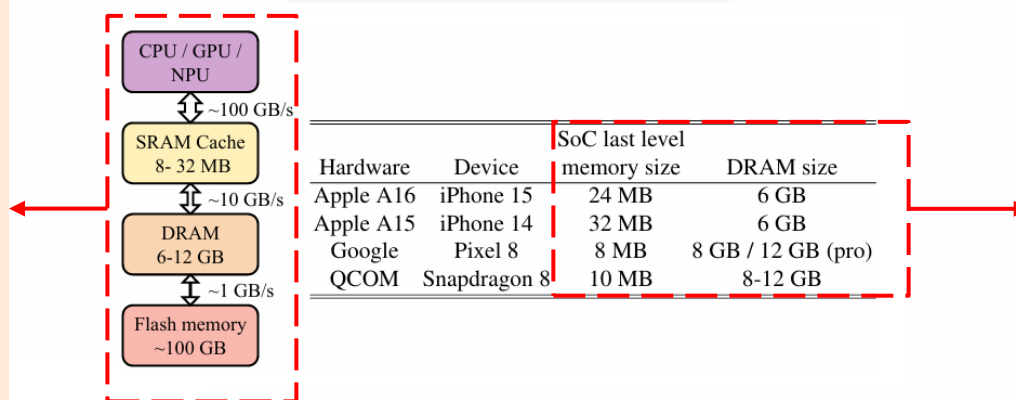
1) 闪存 (Flash Memory) 特点：

- ① **大存储**：可存储的内容多，如图所示的100G；
- ② **低带宽**：数据传输速率低，如图所示的1GB/s。

2) DRAM特点：

- ① **小存储**；
- ② **高带宽**。

移动设备中的存储层次结构



1) 用于**执行高速应用程序的操作内存**主要位于**DRAM**中，通常限制在**6-12GB**；

2) 用于**计算的SRAM**通常限制在**20M**左右。

2.1 从小模型论文看端侧硬件瓶颈——内存容量

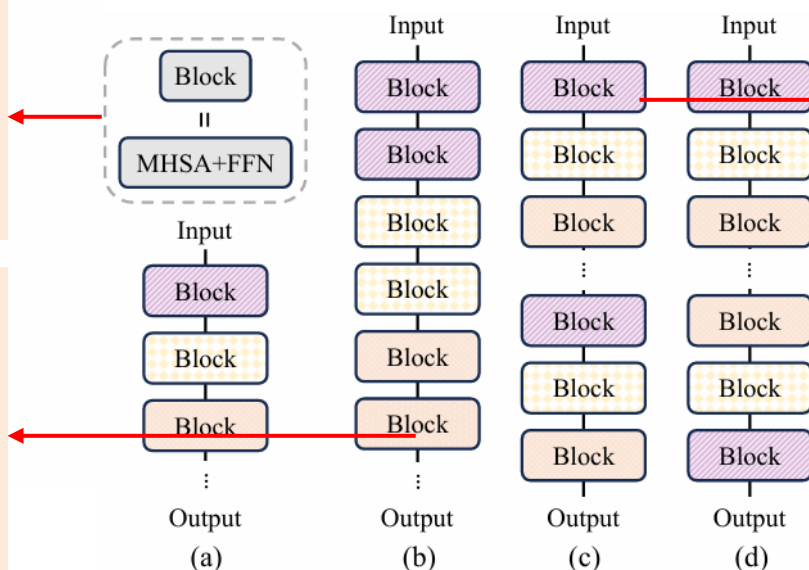
❑ 如何解决当前手机DRAM内存容量空间有限与LLM需求之间的矛盾？

- **Meta MobileLLM采用方法：**由操作系统和其他应用程序需要共享DRAM容量，一个移动应用不应超过DRAM的10%，因此，Meta选择研究并部署一个小于10亿参数的LLM，推出仅有125M和350M参数大小的MobileLLM模型，模型优化方法包括但不限于前文所提及的量化、知识蒸馏、GQA等方法，并**采取“共享层”策略，即通过增加隐藏层的数量而不增加额外的模型存储成本**。通常在手机内存层次结构中，**用于计算的SRAM通常限制在约20MB左右，一般只能容纳一个单独的Transformer块**，在“层共享”策略下，Meta MobileLLM将共享的权重放入缓存中，在SRAM和DRAM之间实现数据共享，从而提高自回归推理的整体执行速度。

Meta MobileLLM模型提出“层共享”方法

(a) 没有层共享的基准模型：通常一个transformer块包含多头自注意力（MHSA）和前馈网络（FFN）；

(b) 立即以块为单位共享 (Immediate block-wise sharing)：能够最好地利用缓存，因为共享权重可以保留在缓存中，并立即两次计算。



(c) 全局重复共享 (Repeat-all-over sharing)：该方法下，模型在零样本常识推理测试中具备更高的性能。

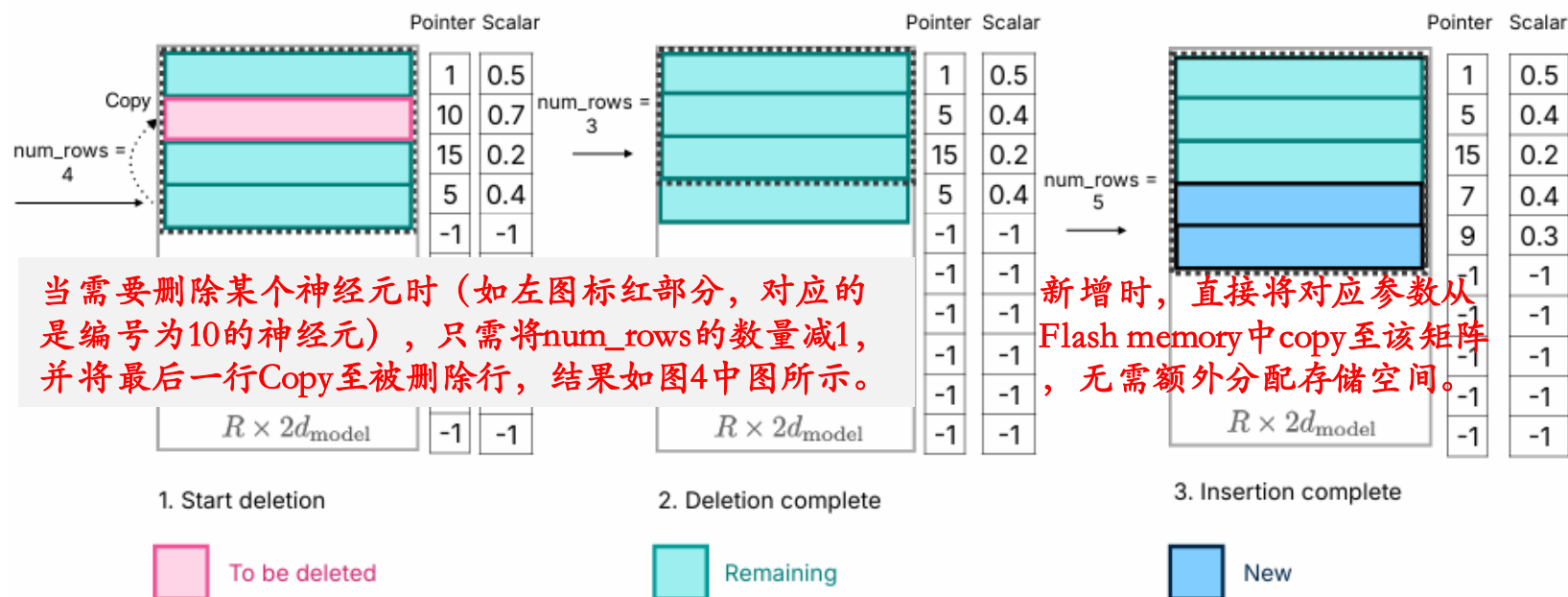
(d) 逆共享 (Reverse sharing)

2.1 从小模型论文看端侧硬件瓶颈——内存容量

❑ 如何解决当前手机DRAM内存容量空间有限与LLM需求之间的矛盾？

- **苹果《LLM in a flash》解决思路：**由于LLM所需内存大小显著大于实际DRAM容量，因此，苹果尝试将LLM放在Flash Memory中，在每次需要进行推理时，仅将部分必要菜单参数加载到DRAM中。在该方案中，需解决两个问题：①如何快速识别出模型的哪些参数是必要的；②由于Flash memory到DRAM的带宽较低，如何加快由Flash memory到DRAM的传输效率。针对以上问题，苹果提出三种解决思路：①**减少数据传输量**；②**提高传输吞吐量**；③**优化DRAM数据管理**。

苹果对DRAM中的数据进行精细化管理



2.1 从小模型论文看端侧硬件瓶颈——算力

- **GPU算力影响首个token的推理延迟，内存带宽影响后续每个token的推理延迟。** LLM推理过程主要包括预填充（并行处理输入prompt的所有tokens，并生成第一个token）和解码阶段（逐个生成后续的token），其中，预填充所需要的时间 = 模型浮点计算量(FLOPS) / GPU半精度浮点算力，根据该公式可以看出，**预训练阶段的性能瓶颈主要在于GPU算力，即GPU算力影响首个token的推理延迟**；而解码阶段每个token所需的生成时间 = 模型参数量所占字节数(bytes) / 内存带宽(GB/s)，根据公式可以看出，**解码阶段的主要性能瓶颈是内存带宽，即内存带宽影响后续每个token的推理延迟**。与此同时，GPU算力的有效利用率和内存带宽的有效利用率的高低也会影响模型的推理速度。

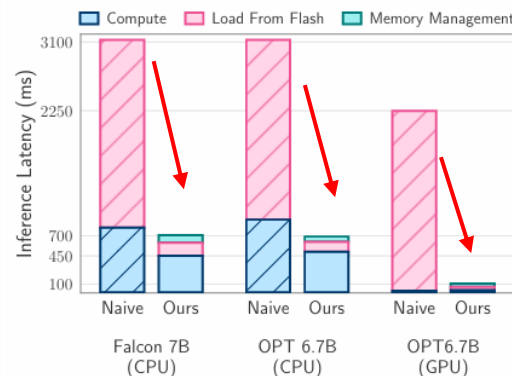
Meta MobileLLM推理延迟表现

Meta指出：MobileLLM-125M模型能够以每秒 50 个令牌的速度运行，而最先进的 iPhone App MLC Chat 使用 LLaMA7B 模型以每秒 3~6 个令牌的速度运行。

Table 7: Latency analysis of MobileLLM-125M (30 layers), MobileLLM-LS-125M (2×30 layers, adjacent blocks sharing weights), and a 60-layer non-shared weight model, with consistent configurations in all other aspects.

	Load	Init	Execute
MobileLLM	39.2 ms	1361.7 ms	15.6 ms
MobileLLM-LS	43.6 ms	1388.2 ms	16.0 ms
60-layer non-shared	68.6 ms	3347.7 ms	29.0 ms

苹果通过筛选模型参数后实现延时减少



苹果根据每个令牌生成步骤有选择地加载参数

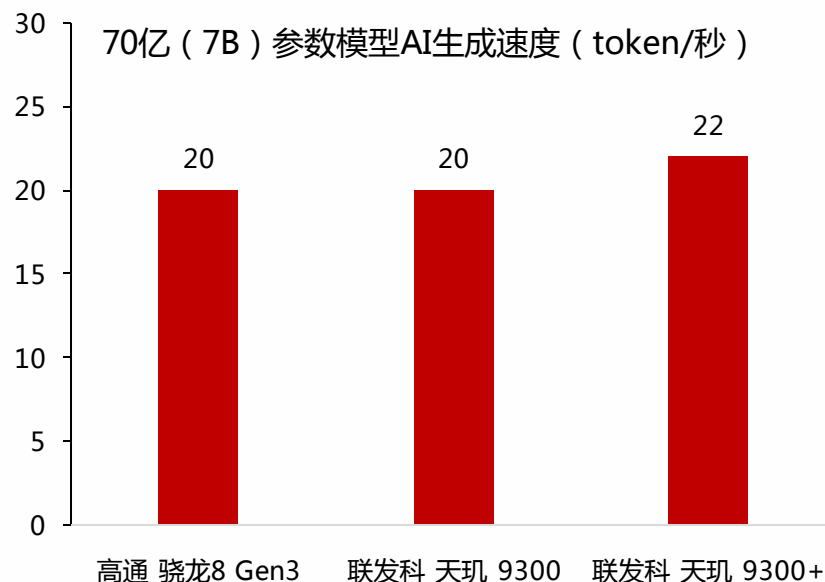
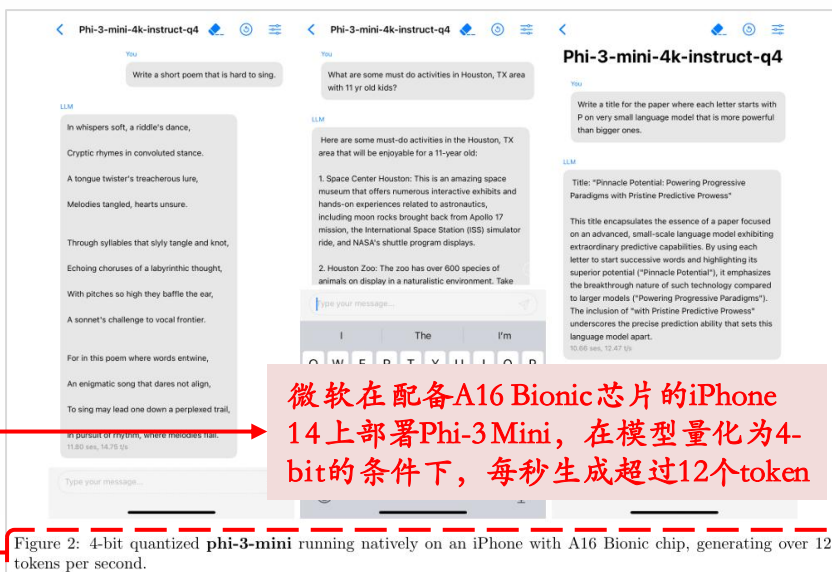
Figure 1: Inference latency of 1 token when half the memory of the model is available. Our method selectively loads parameters on demand per token generation step. The latency is the time needed to load from flash multiple times back and forth during the generation of all tokens and the time needed for the computations, averaged over all generated tokens.

2.1 从小模型论文看端侧硬件瓶颈——内存带宽

- ❑ GPU算力影响首个token的推理时延，内存带宽影响后续每个token的推理延迟。
- 根据微软Phi-3模型技术报告，微软通过在配备A16 Bionic芯片的iPhone 14上部署Phi-3 Mini，进行了量化模型的测试，该设备完全离线运行，每秒生成超过12个token。
- 根据苹果官网，小模型在iPhone 15 Pro上的测试结果来看，端侧AI的延迟约为0.6毫秒，生成速率为每秒30个token，性能较为合理。
- ❑ 目前，主流AI手机芯片对于7B参数模型的AI生成速度一般在每秒20个tokens左右。

iPhone 14每秒生成速率超过12个tokens

主要AI手机芯片AI生成速度 (token/秒)



2.1 从小模型论文看端侧硬件瓶颈——功耗

❑ 模型产生更多功耗，电池性能有待提高，散热能力仍需加强。

Meta MobileLLM中的功耗瓶颈

Furthermore, considerations of *portability* and *computa-*

- LLM能耗：模型每十亿参数下，每个token消耗0.1焦耳。因此，对于一个7B参数的模型，每个token消耗0.7焦耳。
- 一个满电的iPhone，大约有5000焦耳能量，可以支持7B模型在10 tokens/秒的AI生成速率下进行对话不到2小时，每64个tokens消耗0.2%的电量。

a mobile app should not exceed 10% of the DRAM, since DRAM is shared with the operating system and other applications (Malladi et al., 2012). This motivates deploying sub-billion parameter LLMs. Additionally, factoring in LLM energy consumption (0.1 J/token per billion in model parameters (Han et al., 2016; Malladi et al., 2012)), a 7B-parameter LLM consumes 0.7 J/token. A fully charged iPhone, with approximately 50kJ of energy, can sustain this model in conversation for less than 2 hours at a rate of 10 tokens/s, with every 64 tokens draining 0.2% of the battery.

These demands converge on a singular imperative: the adoption of compact models for on-device execution. By utilizing a sub-billion model, such as a 350M 8-bit model consuming only 0.035 J/token, an iPhone can support conversational use an entire day. Moreover, the decoding speed can be significantly enhanced, as exemplified by the benchmark results of the 125M model, capable of operating at 50 tokens/s, compared to the state-of-the-art iPhone App MLC Chat utilizing the LLaMA 7B model at 3~6 tokens/second⁵. In light of these considerations, our paper is motivated by the design and implementation of LLMs with parameters

MobileLLM-350M在8-bit权重下的模型，每个token仅消耗0.035焦耳，iPhone可以支持全天的会话使用。

2.2 从芯片厂商布局看硬件升级趋势——先进制程

- **手机芯片采用先进制程，工艺有望向3nm迈进。** 23Q4，高通和联发科分别在其10月和11月峰会上发布旗下手机芯片骁龙8Gen3和天玑9300，两者均采用台积电4nm制程工艺。根据高通和联发科历年一年一迭代的发布节奏，骁龙8Gen4和天玑9400手机处理器可能于24Q4推出，并有望基于台积电3nm工艺打造。而苹果相较于其他手机芯片厂商工艺更为领先，于23Q3率先推出采用3nm制程的iPhone芯片A17 Pro，未来有望在先进制程上保持领先。

主流手机芯片厂商Roadmap

公司	芯片	21Q4	22Q1	22Q2	22Q3	22Q4	23Q1	23Q2	23Q3	23Q4	24Q1	24Q2
苹果	A系列				A16 @TSMC N4P				A17/ A17 Pro @TSMC N3			
高通	骁龙8系	骁龙8 Gen 1 @TSMC 4nm				骁龙8 Gen 2 @TSMC 4nm				骁龙8 Gen 3 @TSMC N4P		
联发科	天玑9000系	天玑 9000 @TSMC 4nm		天玑 9000+ @TSMC 4nm		天玑 9200 @TSMC 4nm		天玑 9200+ @TSMC 4nm		天玑 9300 @TSMC 4nm		天玑 9300+ @TSMC 4nm
三星	Exynos系		Exynos 2200 @三星 4nm							Exynos 2400 @三星 4nm		

2.2 从芯片厂商布局看硬件升级趋势——内存容量及带宽

- **内存容量仍需扩大，带宽需求持续升级。**2024年初，联发科陈立忠博士在腾讯科技的采访中提到，手机运行百亿参数的AI模型，需要至少13GB的内存和130GB/s的带宽，而2023年旗舰手机的配置，内存通常为16GB，带宽为50GB/s，使得手机终端难以运行大模型。**1) 内存容量方面：**对比目前市场上三大主流手机芯片来看，苹果A17 Pro芯片在内存容量明显低于联发科的天玑9300和高通的骁龙8Gen3。随着AI模型应用的加速，苹果手机的内存容量有望从iPhone15的6+8GB，向8+8GB或12+16GB等更高配置提升。**2) 内存带宽方面：**当前主流手机芯片的最大带宽在50GB/s或80GB/s级别，距离130GB/s的带宽需求仍有较大差距，根据前文对小模型上终端的硬件瓶颈分析，更高的带宽及带宽利用率，将有效提升AI模型的推理速度、优化用户体验。

主要手机芯片存储参数对比

配置	指标	苹果 A17 Pro	联发科 天玑 9300	高通 骁龙8 Gen 3
发布时间		23Q3	23Q4	23Q4
终端应用		iPhone 15 Pro/Pro Max	vivo	三星Galaxy Z Fold/Flip 6、小米14系列、OPPO Find X7
GPU	型号	苹果 A17 Pro	ARM Immortals-G720 MC12	Qualcomm Adreno 750
	制程	TSMC 3nm	TSMC 4nm	TSMC 4nm
	运算单元及频率	6核，24运算单元，频率1.40GHz	12运算单元，频率1.00GHz	频率0.90GHz
	最大显存	6 GB	/	6 GB
内存	内存类型	LPDDR5-6400Mbps	LPDDR5T-9600Mbps	LPDDR5X-9600Mbps
	内存通道	1 (Single Channel)	4 (Quad Channel)	4 (Quad Channel)
	最大内存	8 GB	16 GB	24 GB
	最大带宽	51.2 GB/s	76.8 GB/s	76.6 GB/s
缓存	L2-Cache	20 MB	/	/
	L3-Cache	24 MB	18 MB	12 MB

2.2 从手机芯片厂商布局看硬件升级趋势——AI处理器

- **提高AI处理器配置，支持更强AI模型。** 1) **苹果A17 Pro**：集成16个神经网络核心，支持35TOPS的AI计算能力，相较于上一代手机芯片A16（17TOPS）的AI算力实现显著提升；2) **联发科天玑9300**：搭载联发科第七代AI处理器APU 790，AI算力为33TOPS，支持终端运行10/70/130亿、最高可达330亿参数的端侧模型。3) **高通骁龙8Gen3**：通过升级AI引擎Hexagon NPU的微架构，性能相较上一代提升98%，能效比提升40%，AI算力为34TOPS，最大支持100亿参数的端侧模型。

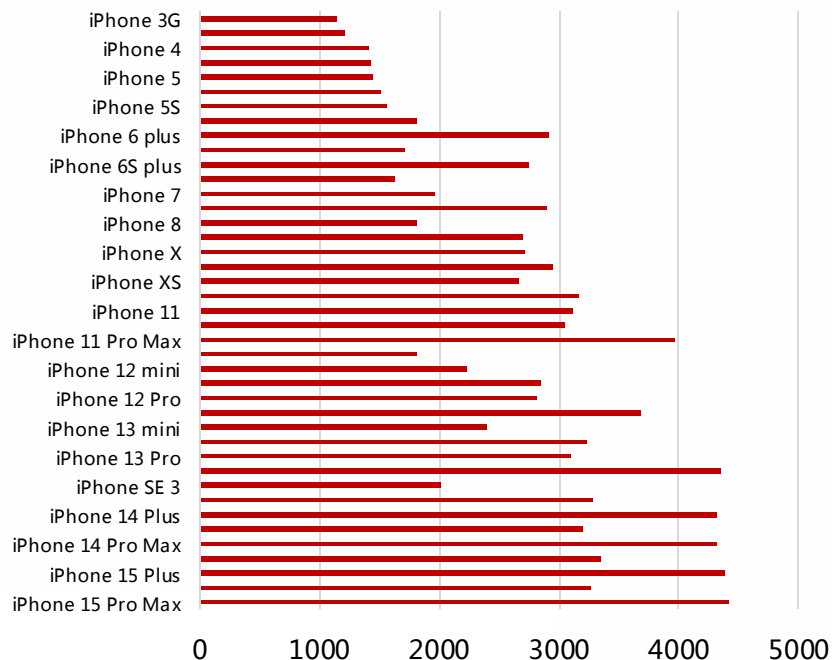
主要AI手机芯片参数对比

配置	指标	苹果 A17 Pro	联发科 天玑 9300	高通 骁龙8 Gen 3
发布时间		23Q3	23Q4	23Q4
终端应用		iPhone 15 Pro/Pro Max	vivo	三星Galaxy Z Fold/Flip 6、小米14系列、OPPO Find X7
CPU	架构	hybrid架构	hybrid架构	hybrid架构
	内核	6核： 2个A-Core 频率3.78 GHz； 4个B-Core 频率2.11 GHz。	8核： 1个A-Core 频率3.25 GHz； 3个B-Core 频率 2.85 GHz； 4个C-Core 频率2.00 GHz。	8核： 1个A-Core 频率3.40 GHz； 5个B-Core 频率2.96 GHz； 2个C-Core 频率2.27 GHz。
GPU	型号	苹果 A17 Pro	ARM Immortals-G720 MC12	Qualcomm Adreno 750
	制程	TSMC 3nm	TSMC 4nm	TSMC 4nm
	运算单元及频率	6核，24运算单元，频率1.40GHz	12运算单元，频率1.00GHz	频率0.90GHz
	最大显存	6 GB	/	6 GB
AI	处理器	16 Neural cores	MediaTek APU 790	Hexagon NPU
	AI算力	35 TOPS	33 TOPS	34 TOPS
其他	指令集	Armv8.6-A (64 bit)	Armv9-A (64 bit)	Armv9-A (64 bit)
	操作系统	iOS	Android	Android, Windows 10/11 (ARM)

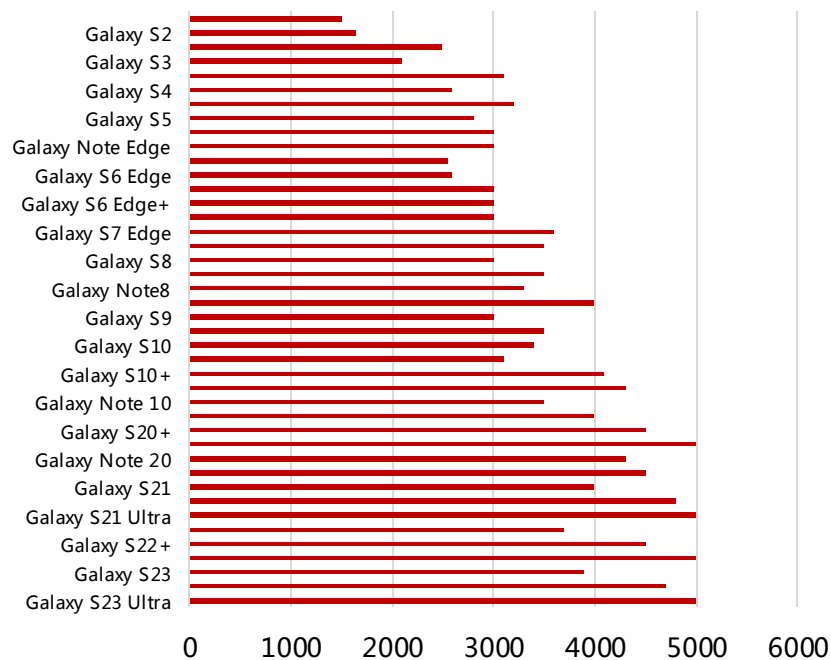
2.2 从手机芯片厂商布局看硬件升级趋势——电池

- 端侧AI要求单机带电量提升，手机电池向“钢壳+硅碳负极+叠片工艺”三方向发力。1) 钢壳替换铝塑膜：不锈钢外壳可帮助手机电池提升使用寿命、增加循环次数，同时满足欧洲可拆卸要求，具备相对优势。2) 硅碳负极掺硅比例提高：目前高端消费类电池多使用硅碳负极，在AI模型上终端的背景下，电池能量密度要求进一步提高，掺硅比例有望持续提升。3) 叠片工艺代替卷绕工艺：苹果、三星等头部厂商的高端机型已率使用叠片工艺，以满足高充电速率、轻薄化、异形化等的需求，未来，随着AI手机等高端机型的加速推出，叠片工艺有望被广泛运用。

苹果手机电池容量变化 (mAh)



三星手机电池容量变化 (mAh)



2.2 从手机芯片厂商布局看硬件升级趋势——散热

- ❑ **手机散热需求增加，引入VC散热板+石墨烯。** AI手机在模型运行时需要处理大量数据、带来更大功耗，从而对手机的散热性能进一步提高。目前，高端手机逐渐采用“超薄VC均热板+石墨及石墨烯”散热方案，其中，VC均热板具备轻薄、散热面积大等优势，石墨烯散热材料具备高导热性、重量轻、柔韧性好等特点，从而能够形成散热效率更高的导热组合方案。未来，AI手机有望基于该方案，进一步扩大VC散热面积、增加石墨片的用量等。此外，相较于铝塑膜，钢壳电池同样具有更好的散热性能，有助于提升手机的散热表现。

手机散热技术原理及优劣势对比

散热技术	原理	导热系数	优点	缺点
石墨烯	具有独特的晶体结构，表面可与电子产品内部发热器件贴合，能够以最大的有效表面积，将电子产品发热器件表面上热力均匀的分布在二维平面，有效实现热量转移。	5300 W/m·K	高导热性、机械性能好、重量轻，材料薄、柔韧性好	制备工艺复杂、价格昂贵
热管	利用工作流体的蒸发与冷凝来传递热量。将铜管内部抽真空后充入工作流体，流体以蒸发-冷凝的相变过程在内部反复循环，不断将热端的热量传至冷却端，从而形成将热量从管子的一端传至另一端的传热过程。	1500~50000 W/m·K	高导热性、可长距离传输热量、可塑性强	弯曲会降低性能、一般不耐压力
均热板	发热源运行时产生的热量传导至均热板的蒸发端，内部的冷凝液会迅速吸收这些热量并转化为蒸汽，从而带走大量的热能。由于水蒸气的潜热性，均热板的热蒸汽会由高压区扩散到低压区（冷凝端），当蒸汽接触温度较低的内壁时会迅速凝结为液体并释放热能。最后，这些液体会利用毛细作用流回蒸发端，最终形成一个水气并存的双相循环系统。	> 2000 W/m·K	轻薄、散热面积大	价格较高

目 录

1 基础的构建：模型实现高效压缩是端侧AI的第一步

1.1 十亿级参数模型加速迭代，性能表现向百亿参数模型靠拢

1.2 模型压缩技术助力端侧部署，注意力优化机制降低存算需求

2 落地的关键：模型适配终端硬件是端侧AI的第二步

2.1 从小模型论文看端侧硬件瓶颈：内存/功耗/算力

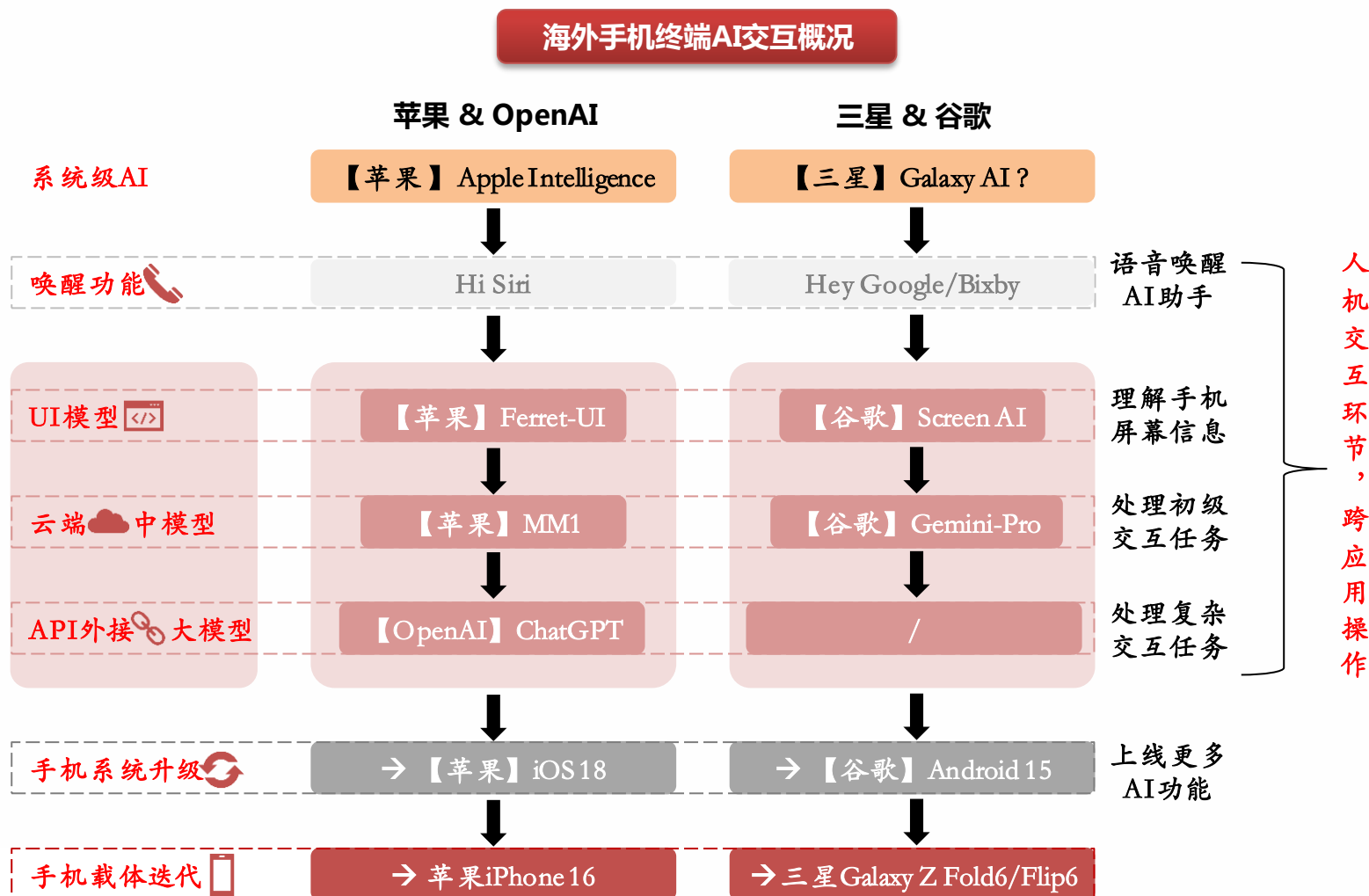
2.2 从芯片厂商布局看硬件升级趋势：制程/内存/NPU/电池/散热

3 体验的突破：模型助力人机交互是端侧AI第三步

3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

3 模型带来应用交互是端侧AI的第三步



3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

❑ **发展节奏对比：苹果推进加速，谷歌尝试较早。**1) **苹果**：面对日益激烈的AI竞争，苹果于23年10月推出第一代可理解指代和定位的多模态大模型Ferret，24年4月提出可以处理更高分辨率图像的改进版本Ferret-v2，以及专门针对手机用户界面的多模态大模型Ferret-UI，赋予AI理解和操作应用界面的能力。2) **谷歌**：由于LLM作为大语言模型，通常无法直接理解UI界面，谷歌针对这一痛点，于2022年9月和10月发表Spotlight和Pix2Struct相关论文，对终端UI界面的理解进行初步探索；2024年2月，谷歌推出UI模型Screen AI，完善相关布局。**目前，大语言模型（LLM）难以直接理解手机UI界面，应用形态以chatbot为主，而谷歌Screen AI和苹果Ferret-UI采用多模态视觉语言模型（VLM），实现对屏幕信息的理解，为用户与终端设备之间的交互提供UI理解基础。**

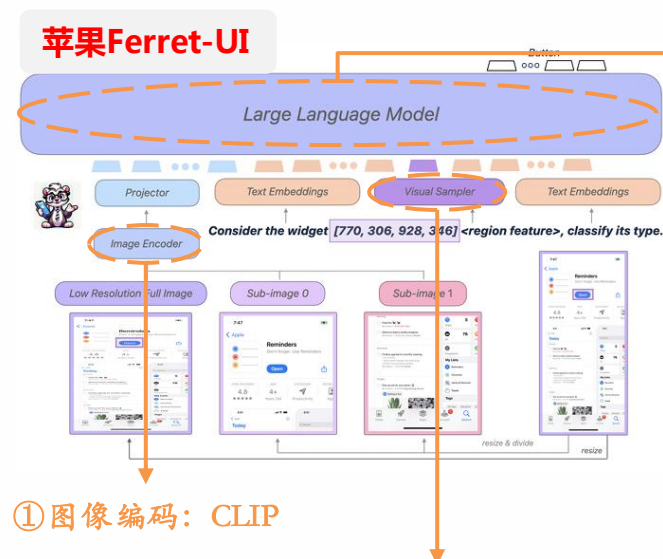
苹果和谷歌UI模型相关布局

公司	模型名称	发布日期	模型参数量（B）	屏幕截图处理方法	UI模型训练任务设计
苹果	Ferret-v2-13B	2024年4月11日	13	anyres：“任意分辨率”方法，将横屏或竖屏一分为二	1）指代任务： OCR、图标识别、控件分类 2）定位任务： 控件列表、查找文本、查找图标、查找控件 3）高级任务： 详细描述、对话感知、对话交互、功能推理
	Ferret-v2-7B	2024年4月11日	7		
	Ferret-UI-anyres	2024年4月8日	/		
	Ferret-UI-base	2024年4月8日	/	/	
	Ferret-v1-13B	2023年10月11日	13		
	Ferret-v1-7B	2023年10月11日	7		
谷歌	Screen AI-5B	2024年2月7日	4.62	Pix2Struct：将屏幕截图解析为简化的HTML	1）屏幕标注： 检测和识别屏幕上存在的用户界面元素； 2）屏幕问答；3）屏幕导航： 解释导航指令并识别要交互的适当用户界面元素； 4）屏幕摘要： 简洁地总结屏幕内容
	Screen AI-2B	2024年2月7日	1.88		
	Screen AI-670M	2024年2月7日	0.675		
	Pix2Struct	2022年10月7日	/		
	Spotlight	2022年9月29日	/		

3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

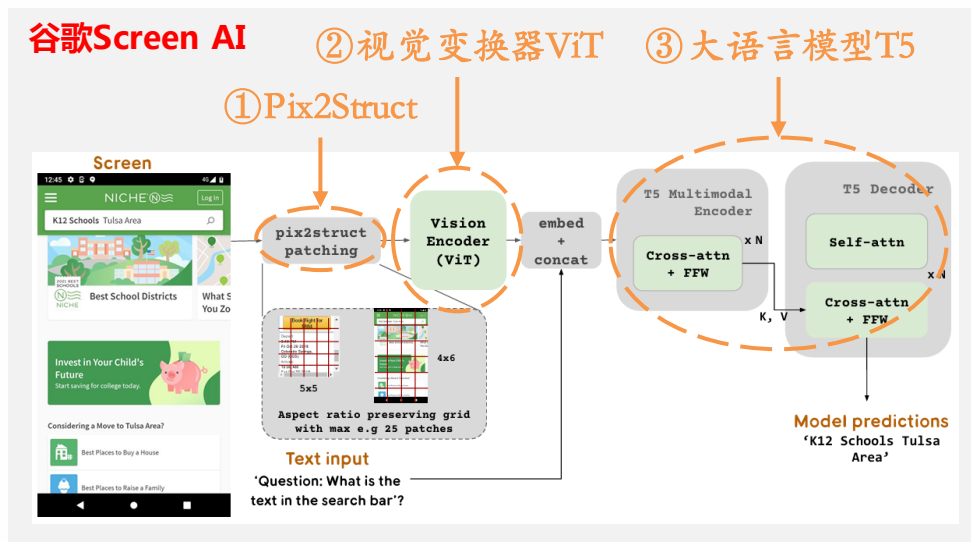
- ❑ **模型架构对比：多模态视觉语言模型为基，专为理解UI屏幕设计。** 1) 苹果：anyres+CLIP+LLM。Ferret-UI模型通过CLIP进行图像编码，采用混合区域表示方法和任意分辨率（anyres）技术进行图片特征提取，再使用Transformer解码器对图像特征和文本指令进行联合编码，理解UI元素的语义信息并进行推理，最后根据任务需求输出相应的文本或指令。2) 谷歌：Pix2Struct+ViT+T5。Screen AI基于Spotlight研究成果，借鉴PaLI模型架构，结合视觉变换器ViT与大语言模型T5，同时修改ViT的图像修补步骤，加入Pix2Struct策略，将视觉任务转化为文本到文本或图像到文本的问题。

苹果Ferret-UI和谷歌Screen AI模型架构对比



- ② 区域特征提取：空间感知视觉采样器（visual sampler）根据文本指令精确定位屏幕上的目标区域或元素，并将该区域的图像特征提取出来

- ③ 语义理解：使用LLM对图像特征和文本指令进行联合编码，理解UI元素的语义信息，并进行推理。



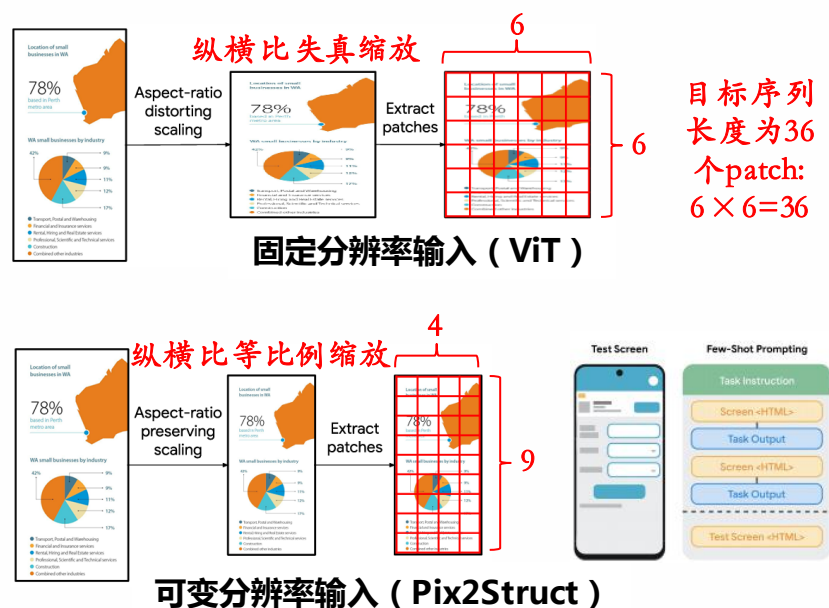
3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

- **屏幕截图处理技术对比：**1) **苹果：**将屏幕截图一分为二，对区域图像进行特征提取。Ferret-UI中的任意分辨率技术专注于处理手机界面，由于手机UI屏幕通常比自然图像具备更长的长宽比，且屏幕截图主要分为竖屏和横屏，因此，苹果选择将屏幕截图在水平或垂直方向上一分为二，将完整图片分为子图像，以增强细节识别、进行特征提取、适应各种屏幕纵横比。2) **谷歌：****纵横比等比例缩放，UI信息转换为HTML语言。**针对手机界面、电脑界面等具有各种分辨率和纵横比屏幕截图，谷歌采用Pix2struct技术，保留屏幕截图原有纵横比，实现可变分辨率输入，并将UI界面信息转换成HTML语言，让LLM可以理解和操作界面。

苹果Ferret-UI屏幕截图处理方法anyres



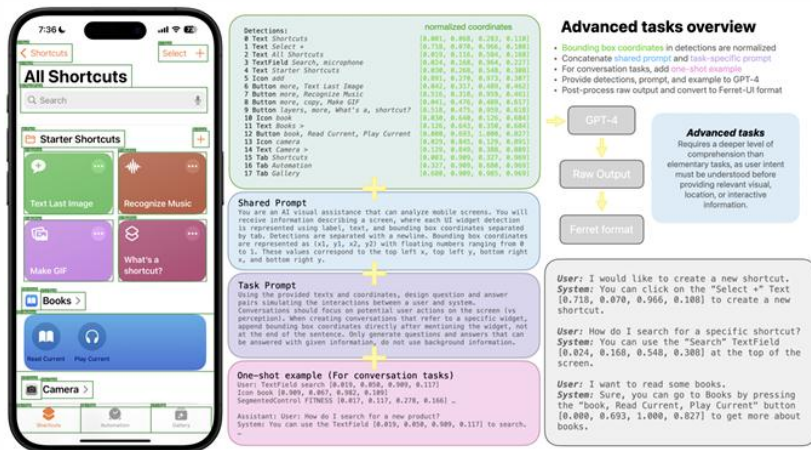
谷歌Screen AI屏幕截图处理方法Pix2Struct



3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

- ❑ **训练任务对比：1) 苹果：针对手机屏幕信息进行训练，强调对UI元素的精细理解和复杂任务的执行能力。**训练数据集方面，苹果通过收集iPhone和安卓手机的UI数据，对UI元素进行精细标注，实现更好的空间理解；训练任务方面，苹果设计出涵盖从基础到高级的14种不同手机UI任务，包括图标识别/文本查找/控件列表生成/控件分类/图标识别/OCR等基础任务，以及详细描述/感知/交互对话/功能推理等高级任务，加深用户与应用的交互。**2) 谷歌：训练数据涵盖电脑、手机截屏等多样数据，有望提供更广泛的应用场景。**训练数据集方面，谷歌Screen AI不仅关注手机UI界面，还关注电脑等其他终端屏幕信息，并加入图表、文档等数据，为模型提供更广泛的应用场景；训练任务方面，Screen AI针对识别屏幕用户界面元素/复杂问题解答/屏幕导航/内容摘要四大任务进行设计，以加深对屏幕截图的理解。

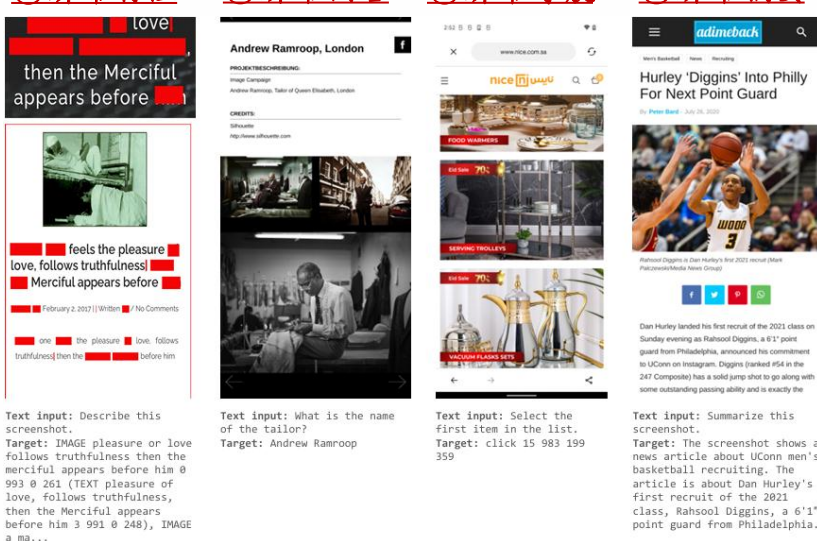
苹果Ferret-UI任务设计



高级任务：模型识别图标、控件等元素，输出对应元素的标注界限框，并将识别结果、prompt、零样本一起发送至GPT-4

谷歌Screen AI任务设计

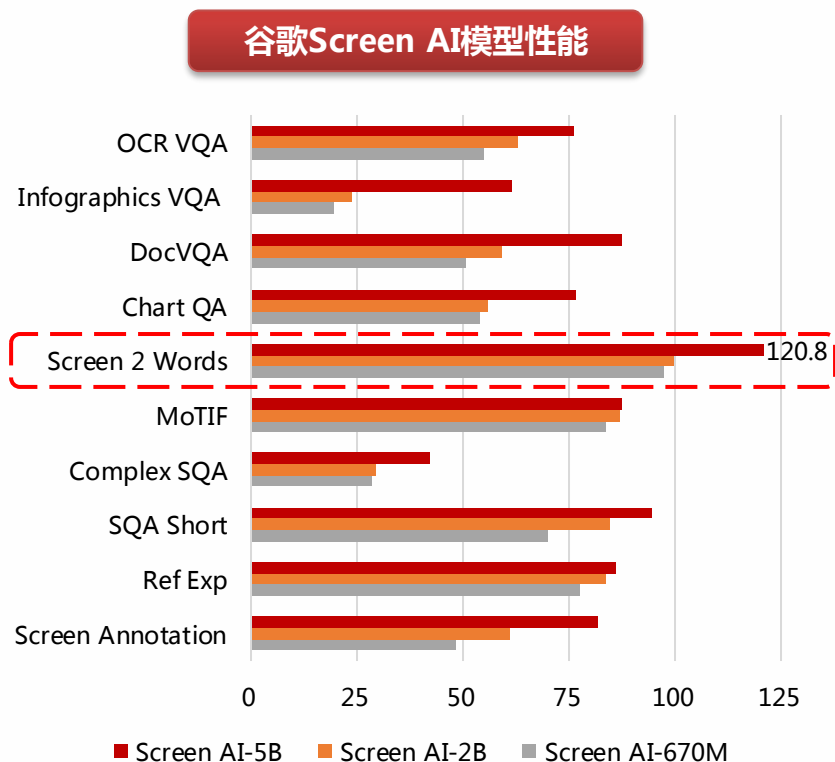
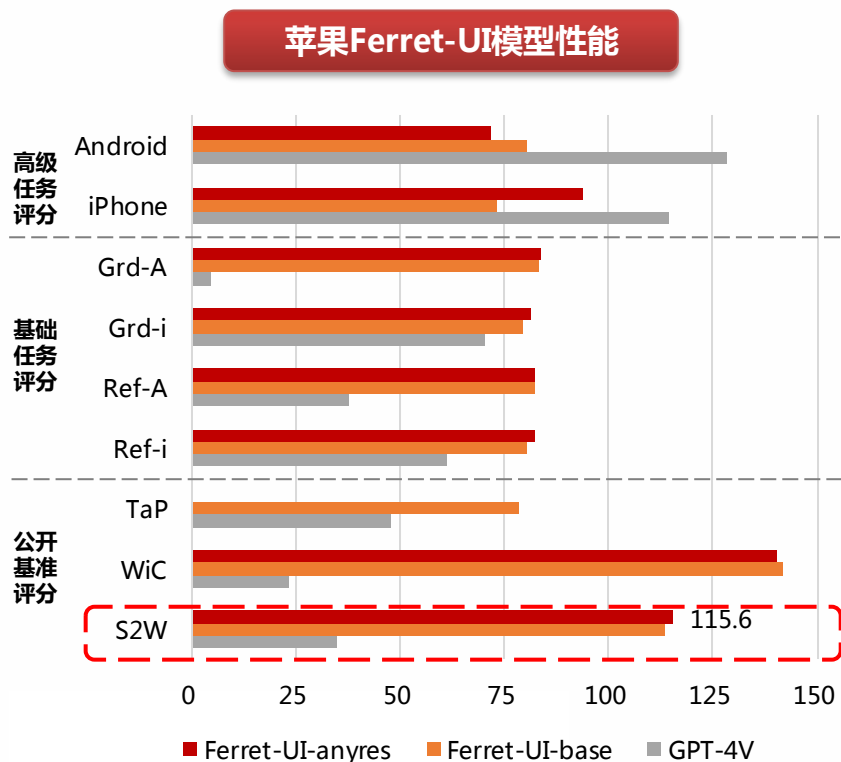
① 屏幕标注 ② 屏幕问答 ③ 屏幕导航 ④ 屏幕摘要



资料来源：谷歌，西南证券整理

3.1 UI模型：手机界面理解能力提升，任务设计为人机交互奠定基础

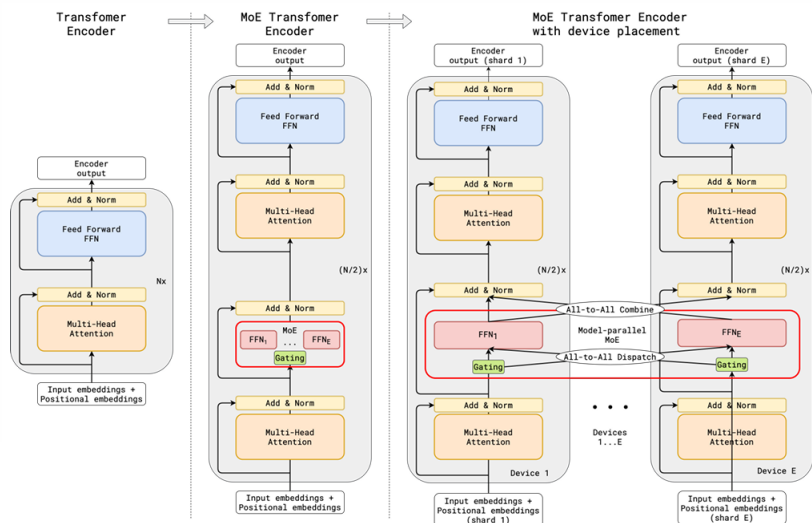
- ❑ 模型性能对比：苹果Ferret-UI可帮助实现高级任务，谷歌Screen AI图片理解能力突出。
- 1) 苹果Ferret-UI：Ferret-UI在公开基准和基础任务上的评分上表现较好，在引用、定位和推理等方面展现出较强的能力，功能的增强有助于赋能更多下游UI应用。
- 2) 谷歌Screen AI：Screen AI图片理解能力优秀，在多种QA测试集上表现出较好性能；相较于苹果Ferret-UI，Screen AI在公开测试基准Screen2Words上的得分更高。



3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

- ❑ **MoE架构集成多个专家模型，处理任务专业高效。** MoE模型（mixture-of-experts model）即混合专家模型，作为大模型架构的一种，**MoE基于“术业有专攻”的设计思想**，由多个子模型（专家模型）组成，通过将任务分门别类、再分配给不同的专家模型，使处理问题更加灵活高效；而稠密模型（Dense model），偏向于“通才”模型，能够处理众多任务，但专业能力可能不及MoE。
- ❑ **云端模型补充端侧AI能力，MoE架构可适配于多种场景。** 近半年来，海外模型厂商纷纷采用MoE架构，其中，谷歌基于过去的领先研究在Gemini-1.5-Pro模型中采用MoE架构，苹果基于MoE架构对MM1模型进行扩展，且两者分别应用于苹果和三星手机的AI系统中。此外，由于MoE模型中的每个子模型能够专注处理某类特定任务，因此，手机厂商可以针对端侧AI的应用场景，对混合专家模型进一步微调优化，使其与多种端侧任务适配，推动端侧AI的推理效率。

MoE模型架构示意图



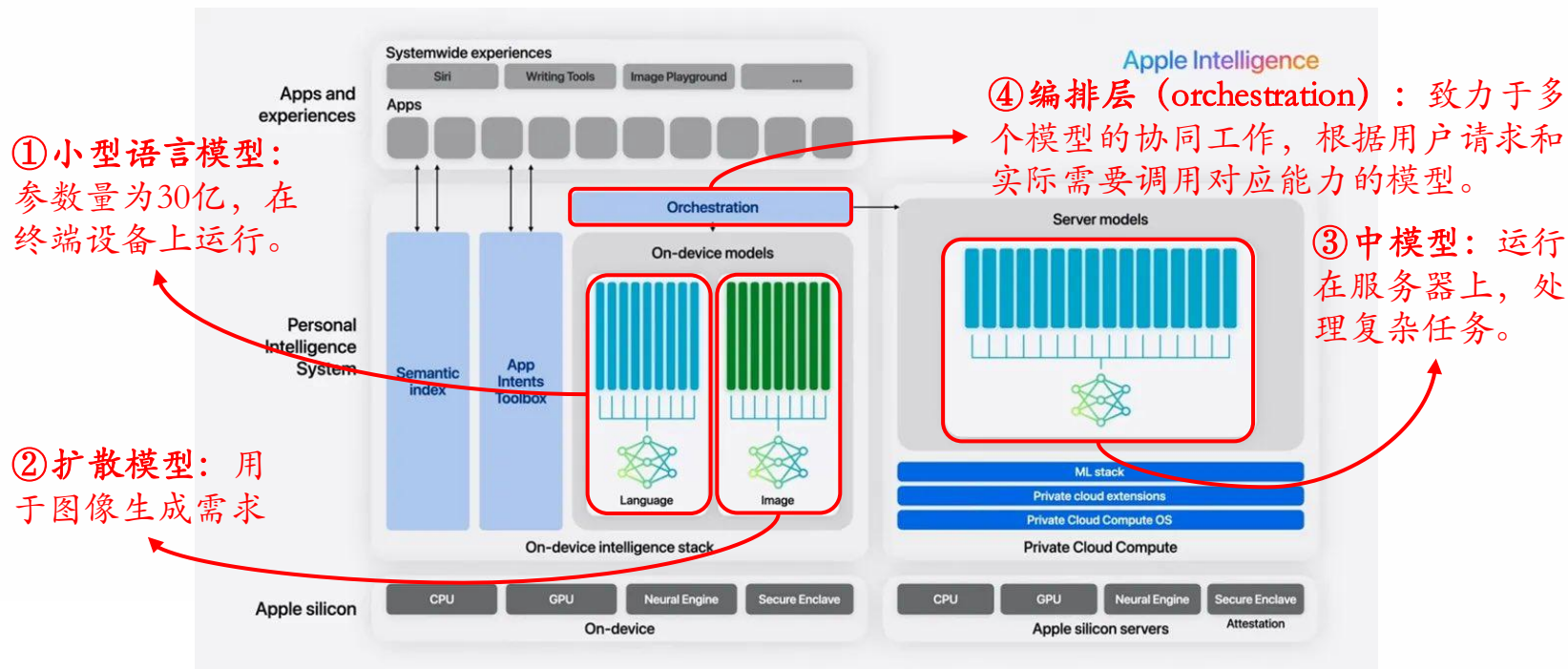
海外MoE模型发展情况

公司	模型名称	发布日期	MoE架构
苹果	MM1	2024.03	3B-MoE：68B参数，64个专家； 7B-MoE：47B参数，32个专家
谷歌	Gemini1.5 Pro	2024.02	/
微软	WizardLM 2 8×22B	2024.04	8个专家
Mistral	Mistral 8×7B	2023.12	46.7B参数，8个专家，每个token 激活2个专家
xAI	Grok-1	2024.03	314B参数，8个专家，每个token 激活2个专家
AI21	Jamba	2024.03	52B参数，16个专家，每个token 激活4个专家

3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

- ❑ **苹果Apple Intelligence系统**：采用“本地模型+私有云模型+第三方大模型”的三层AI架构设计，编排层助力实现灵活调用。苹果在端侧采用一个本地小语言模型和扩散模型，以满足用户基础AI功能；云端搭载Cloud MM1模型，实现用户与手机的初步交互；为处理更加复杂的任务，苹果给出第三方入口，支持接入OpenAI ChatGPT等大语言模型；编排层则可根据用户请求和实际需要，调用相对应能力的语言模型，帮助模型间的协同和灵活调用。

苹果Apple Intelligence系统级AI



3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

- ❑ **三星Galaxy AI系统：与谷歌积极展开合作，通过“本地模型+云端模型”实现推理。**2024年1月，三星推出Galaxy AI，迈入移动AI时代。2024年7月，三星举行Galaxy全球新品发布会，与谷歌推进Gemini在AI手机上的合作，其中，Gemini-Nano用于本地端侧推理，Gemini-Pro和Imagen2作为云端推理，相关AI功能将首批搭载在Galaxy Z Fold6和Galaxy Z Flip6机型上，并与折叠屏手机形态进行结合，致力于为用户带来更丰富的交互体验，抢占AI流量入口。

三星Galaxy AI部分功能



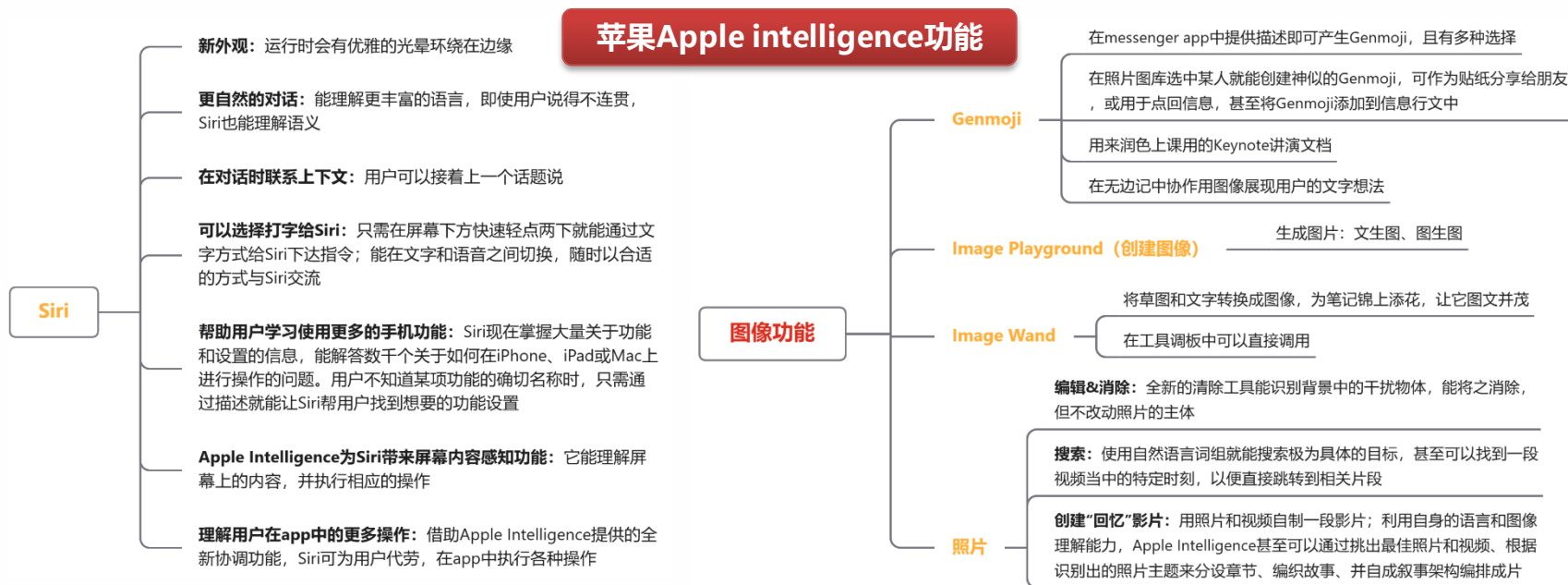
3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

- ❑ **智能及个性化服务增强，跨APP交互有望提升。** 1) 从苹果Apple intelligence与三星Galaxy AI的主要功能对比来看：图像功能方面两者相当，苹果在语音助手、写作工具、图像及照片方面的AI功能更强，三星在翻译和画圈搜索方面更具竞争力。2) 从当前AI功能发展情况来看：随着AI技术与移动设备的紧密结合，模型能够通过提高对UI元素的理解能力，融入系统级AI，使得AI手机能够提供更智能、更直观的交互体验；此外，AI手机能够更好地学习用户行为和偏好，为用户提供更加个性化的服务和推荐；从跨APP调用来看，AI交互功能占比及交互深度有望逐步提升。

苹果Apple intelligence与三星Galaxy AI主要功能对比

系统级AI	苹果 Apple Intelligence	VS	三星 Galaxy AI
图画功能	Image Wand ：根据官方演示，可以将草图变成图画，甚至可以将草图同一页面上的任何文本考虑在内，以便在生成AI图像之前更好地理解用户的草图	≈	Sketch to Image ：可以选择五种风格来呈现绘图，包括水彩、插画、素描、波普艺术和3D卡通，草图更简单，绘图准确率越高
语音助手	Siri ：更自然的语言处理功能；能够从照片、日历、信息和其他应用程序中提取数据，以便更多地了解用户	>	Hey Google/Bixby ：与Gmail和日历等谷歌应用程序集成，以便更多地了解用户
写作工具	集成写作工具主要针对用户已经完成的文本，作为助手可以帮助变换风格、重写改写，或者提取关键内容、进行校对等	>	邮件应用程序中内置 Composer 等功能，可以根据提示生成电子邮件，但AI功能尚未深度集成至Gmail或邮件应用程序中
图像&照片	Genmojis 可以根据提示生成表情符号，可以访问照片库，并根据真实照片来创建人物形象； 清理/编辑照片和移除不需要的对象 等AI功能； Image Playground 功能可以创建有趣图像，能够识别照片库中的人物，创建完成后可在提示中使用	>	Galaxy AI可以根据照片生成人物的3D图像
翻译	可以处理语音输入，但输出仅限于文本	<	Interpreter 可以让手机实时翻译不同的语言；手机能够读出文本，并在电话中使用
搜索	针对图片中的特定内容暂时不能进行联网搜索	<	Circle to Search 可以针对用户圈出的内容，别出物体并进行搜索，并具备联网功能

3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景



3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

苹果Apple intelligence功能



书写工具

写作

整理草草记下的课堂笔记

让邮件的表述句句到位

校对

把关语法和用词，确保句式结构清晰

查看建议修改，以及修改的释义，点击全部接受可迅速进行修改

摘要

将要点提炼出来放在邮箱/文件开头

提取摘要：将摘要显示在邮件顶部，从而直达重点

智能回复：当需要回复一个活动邀请时，用户会看到基于该邮件的回复建议；如果回答是选择是否参加，邮件就会识别邀请函中包含的问题，并提供智能选项，以便快捷回复

邮件处理

判断优先级：将Priority Messages移至上方，Apple Intelligence能读懂收到的邮件内容，判断哪封邮件最紧急，并将其置顶，比如今天的晚餐邀请，或是当天下午航班的登机牌，具备对语言的深刻理解

提取摘要：会显示通知的摘要，以便用户能快速扫视

判断优先级：Priority Notifications会显示在叠放内容的顶端，可以一览注意事项

通知：

“减少干扰”模式：Apple Intelligence可以实现全新的专注模式，以减少干扰，在模式期间能理解通知的内容，选出急待注意的通知，尽可能帮助用户保持专注，免受边缘信息的干扰

备忘录APP：可以帮助记详细的笔记；录音完成后，Apple Intelligence可以生成摘要

录音=>转写=>摘要

电话APP：录音、转写，和Apple Intelligence加持的摘要功能，也适用于电话APP；在实时通话中开始录音，参与者都会自动收到通知

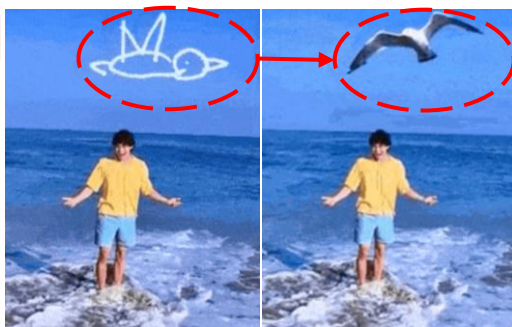
3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

- ❑ **三星 Galaxy AI功能**：聚焦多模态、情境式特点，建立端侧个人化知识图谱，在隐私保护情境下对用户数据进行学习。2024年1月首次推出Galaxy AI，2024年7月进行升级，新增Sketch to Image（将手绘涂鸦变成更精美图片）、Interpreter（实时翻译电话或面对面交流语言）、Composer（用于生成电子邮件）等新功能。

三星 Galaxy AI 主要功能

① 图像生成

涂鸦生成：
支持用户在
照片上手绘
涂鸦，Galaxy
AI可智能生
成相关图像

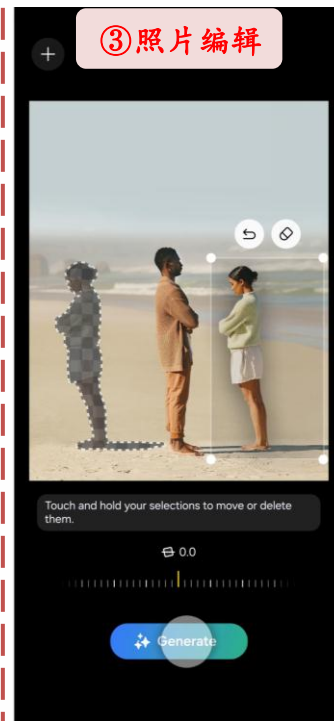


② 图片搜索

圈出图像某块
区域，Galaxy
AI可以识别图
像或文字，并
进行联网搜索

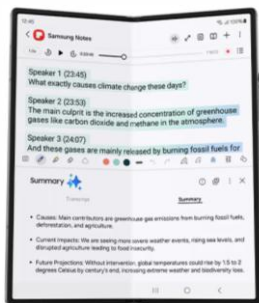


③ 照片编辑



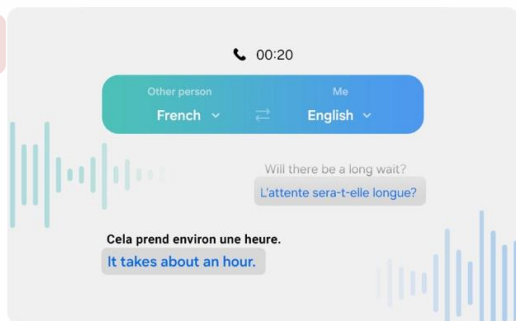
④ 写作工具

编辑邮件：手
机本地的邮件
应用程序中内
置Composer功
能，可根据提
示生成邮件



⑤ 通话转译

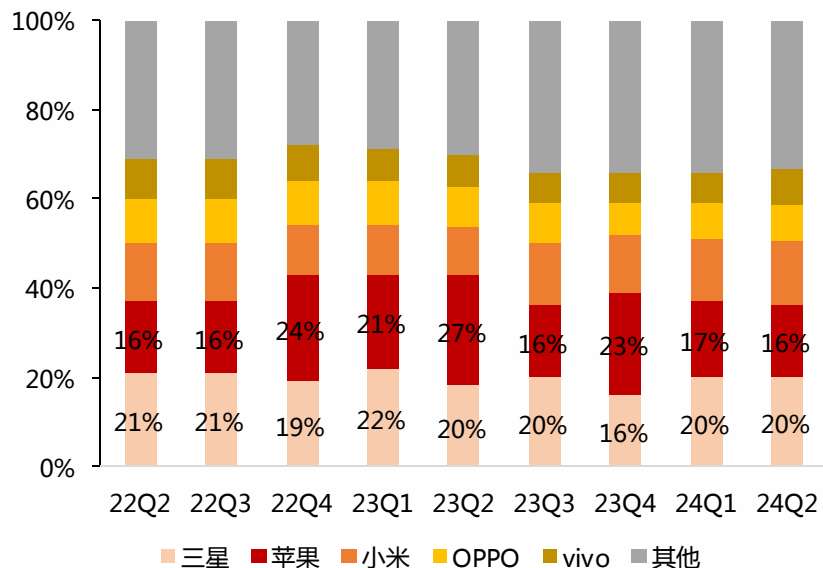
Interpreter：
在用户对话
式，可以让
手机实时翻
译呈不同的
语言。



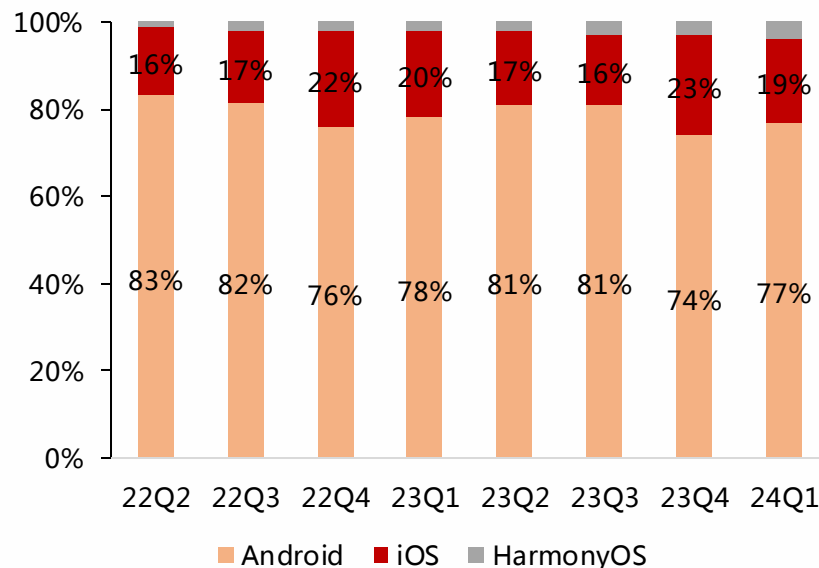
3.2 系统级AI：云端模型补充交互体验，系统升级支持更多AI场景

- ❑ **操作系统持续更新，AI功能多方位升级。**苹果和谷歌作为全球智能手机操作系统的两大巨头，分别主导着iOS和Android平台的生态发展，两大巨头在操作领域积极推进，力求在AI时代推动人机交互方式革新。**1) 苹果：**2024年7月16日，**苹果iOS 18** Public Beta 1（公测版）发布，更新内容与Beta 3开发者测试版一致，根据苹果官网信息，Apple Intelligence将于秋季进入Beta开发者测试版，首发支持美国地区英语用户，25年将推出更多功能并支持多种语言和平台。**2) 谷歌：**2024年5月16日，谷歌在Google I/O大会上宣布**Android 15** Beta2版本，测试版（Beta Release）阶段进入尾声，未来将进入平台稳定性（Platform Stability）和最终版本确定（Final Release）阶段。

全球智能手机销量市场份额



全球智能手机操作系统市场份额



3.2 系统级AI：云端模型补充人机交互能力，系统升级支持更多AI场景

- 从海外科技厂商模型进展来看：苹果和谷歌在端侧AI模型方面布局较为完善。从小语言模型、端侧模型、再到助力人机交互的UI模型、云端大语言模型等，均表明AI与终端设备相结合的发展方向。苹果和谷歌作为全球智能手机操作系统的两大巨头，有望凭借其领先的端侧AI技术，为人机交互带来更加智能、高效、自然、个性化的体验，推动AI时代平台生态的持续发展。
- 从海外端侧AI合作阵营来看：一是“苹果+OpenAI”，二是“谷歌+高通+三星”。未来，随着端侧模型、手机芯片、终端设备、操作系统等各环节的持续发展，终端市场有望展现更多合作可能。

从模型视角看手机端侧AI

①基础的构建：

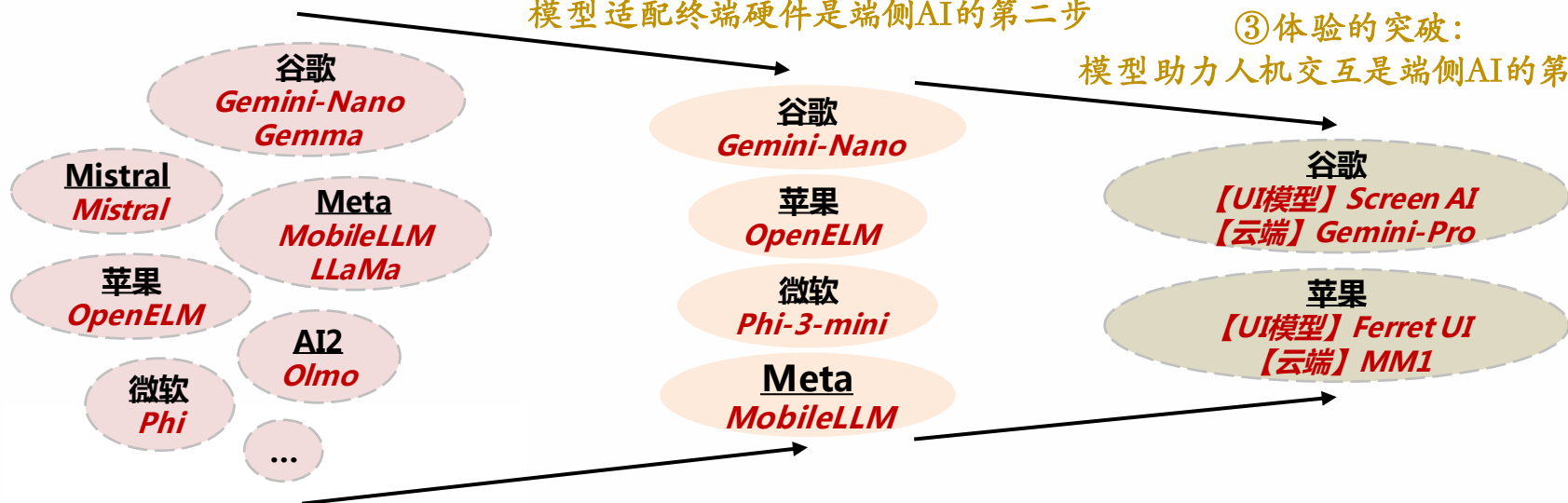
模型实现高效压缩是端侧AI的第一步

②落地的关键：

模型适配终端硬件是端侧AI的第二步

③体验的突破：

模型助力人机交互是端侧AI的第三步



相关标的：苹果（APPL.O）

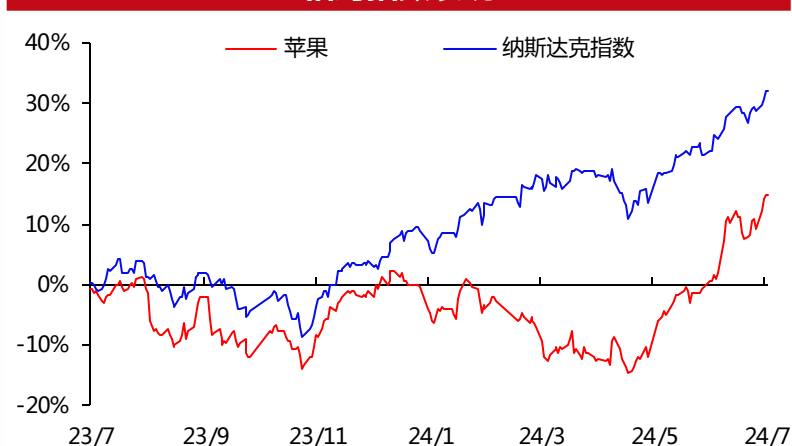
- ❑ **投资逻辑：**Apple Intelligence驱动苹果新一轮创新成长周期。
 - ✓ 基于系统级AI、跨应用的信息整合能力、端云结合和私有云服务的差异化部署方式等优势，Apple Intelligence将为苹果带来新一轮创新和成长周期；
 - ✓ Apple Intelligence或带动苹果硬件产品的换机需求，预计2025-2026财年iPhone销售复合增长10%、Mac销售复合增长7.5%；
 - ✓ AI融合或提升软件服务业务价值，未来三年复合增长10%。
- ❑ **业绩预测与投资建议：**预计公司2025-2026财年净利润复合增速为11%，2024/2025财年对应估值为33xPE、30xPE。维持“买入”评级。
- ❑ **风险提示：**消费电子需求或不及预期；AI进展或不及预期；反垄断政策风险。

业绩预测和估值指标

指标	FY2023A	FY2024E	FY2025E	FY2026E
营业收入（百万美元）	383285.00	387532.68	418154.10	454173.01
营业收入增长率	-2.80%	1.11%	7.90%	8.61%
GAAP净利润（百万美元）	96995.00	102015.30	113032.65	125603.57
GAAP净利润增长率	-2.81%	5.18%	10.80%	11.12%
EPS（美元）	6.33	6.65	7.37	8.19
P/E（GAAP）	34.8	32.8	29.6	26.6

数据来源：Wind、西南证券（PE截至2024年7月26日）

相对指数表现



数据来源：Wind、西南证券整理

风险提示

- ❑ 端侧AI技术进展不及预期风险；
- ❑ 行业竞争加剧风险；
- ❑ 应用开发不及预期风险等。

附：相关资料

发布主体	论文或报告名称
苹果	OpenELM : An Efficient Language Model Family with Open Training and Inference Framework
	MM1 : Methods, Analysis & Insights from Multimodal LLM Pre-training
	Ferret-UI : Grounded Mobile UI Understanding with Multimodal LLMs
	Ferret-v2 : An Improved Baseline for Referring and Grounding with Large Language Models
	FERRET : REFER AND GROUND ANYTHING ANY WHERE AT ANY GRANULARITY
	LLM in a flash : Efficient Large Language Model Inference with Limited Memory
微软	Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone
	Phi-2 : The surprising power of small language models
	Textbooks Are All You Need II: phi-1.5 technical report
	Textbooks Are All You Need
谷歌	Gemma 2 : Improving Open Language Models at a Practical Size
	Gemma : Open Models Based on Gemini Research and Technology
	Gemini 1.5 : Unlocking multimodal understanding across millions of tokens of context
	Gemini : A Family of Highly Capable Multimodal Models
	ScreenAI : A Vision-Language Model for UI and Infographics Understanding
	Pix2Struct : Screenshot Parsing as Pretraining for Visual Language Understanding
	Enabling Conversational Interaction with Mobile UI using Large Language Models
	SPOTLIGHT : MOBILE UI UNDERSTANDING USING VISION-LANGUAGE MODELS WITH A FOCUS
	GShard : Scaling Giant Models with Conditional Computation and Automatic Sharding
Meta	MobileLLM : Optimizing Sub-billion Parameter Language Models for On-Device Use Cases
	Introducing Meta Llama 3 : The most capable openly available LLM to date
	Llama 2 : OpenFoundation and Fine-Tuned Chat Models
	LLaMA : Open and Efficient Foundation Language Models

资料来源：各公司官网，西南证券整理



分析师：王湘杰
执业证号：S1250521120002
电话：0755-26671517
邮箱：wxj@swsc.com.cn

联系人：尤品柯
电话：15370010930
邮箱 ypk@swsc.com.cn

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司 评级

买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下

行业 评级

强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

重要声明

西南证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会核准的证券投资咨询业务资格。

本公司与作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供本公司签约客户使用，若您并非本公司签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司也不会因接收人收到、阅读或关注自媒体推送本报告中的内容而视其为客户。本公司或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或争取提供投资银行或财务顾问服务。

本报告中的信息均来源于公开资料，本公司对这些信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据。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，本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，本公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

本报告及附录版权为西南证券所有，，任何机构和个人不得以任何形式翻版、复制和发布。如引用须注明出处为“西南证券”，且不得对本报告及附录进行有悖原意的引用、删节和修改。未经授权刊载或者转发本报告及附录的，本公司将保留向其追究法律责任的权利。



西南证券研究发展中心

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	魏晓阳	销售经理	15026480118	wxyang@swsc.com.cn
	崔露文	销售副总监	15642960315	clw@swsc.com.cn	欧若诗	销售经理	18223769969	ors@swsc.com.cn
	谭世泽	高级销售经理	13122900886	tsz@swsc.com.cn	李嘉隆	销售经理	15800507223	ljlong@swsc.com.cn
	李煜	高级销售经理	18801732511	yfliyu@swsc.com.cn	龚怡芸	销售经理	13524211935	gongyy@swsc.com.cn
	卞黎旸	高级销售经理	13262983309	bly@swsc.com.cn	孙启迪	销售经理	19946297109	sqdi@swsc.com.cn
	田婧雯	高级销售经理	18817337408	tjw@swsc.com.cn	蒋宇洁	销售经理	15905851569	jyj@swsc.com.c
	张玉梅	销售经理	18957157330	zymyf@swsc.com.cn				
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	王一菲	销售经理	18040060359	wyf@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	王宇飞	销售经理	18500981866	wangyuf@swsc.com
	杨薇	资深销售经理	15652285702	yangwei@swsc.com.cn	路漫天	销售经理	18610741553	lmtyf@swsc.com.cn
	姚航	高级销售经理	15652026677	yhang@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
	张鑫	高级销售经理	15981953220	zhxin@swsc.com.cn				
广深	郑龔	广深销售负责人	18825189744	zhengyan@swsc.com.cn	丁凡	销售经理	15559989681	dingfyf@swsc.com.cn
	杨新意	广深销售联席负责人	17628609919	yxy@swsc.com.cn	陈紫琳	销售经理	13266723634	chzlyf@swsc.com.cn
	张文锋	高级销售经理	13642639789	zwf@swsc.com.cn	陈韵然	销售经理	18208801355	cyrif@swsc.com.cn
	龚之涵	销售经理	15808001926	gongzh@swsc.com.cn	林哲睿	销售经理	15602268757	lzh@swsc.com.cn