

全球大模型将往何处去？

华泰研究

2024 年 7 月 01 日 | 中国内地

深度研究

大模型头部格局基本确定，AI Agent 将加速 AGI 进程

我们认为，海外闭源大模型已经形成 OpenAI 为首，Google、Anthropic 等紧随的格局。在头部闭源模型之下，Meta 引领开源模型生态，开源闭源模型差距逐步缩小。为了适配端侧需求，小参数模型也在快速发展。国内看，模型百花齐放，但技术辨识度不高，23 年头部互联网厂商和科技企业进展较快，24 年以来初创公司开始发力长文本、MoE 等领域。展望后续，Scaling Law+Transformer 仍将长期有效，合成数据或逐渐成为关键数据来源。此外，AI Agent 能够极大提高现有模型的表现，是实现 AGI 的重要推力。大模型技术是 AI 浪潮的软件“基础设施”，建议持续关注相关进展。

多模态+长文本+MoE 已成共识，大模型与小模型路线并驾齐驱

头部 GPT、Gemini、Claude 模型先后支持了多模态推理；Claude 较早实现了 200K 长文本，Gemini 将长文本推到 2M tokens；GPT-4、Mistral 展现了 MoE 架构的优势，Gemini 也在短期内更改为 MoE 架构。共识已经形成，国内大模型厂商均在跟进，Kimi 引领长文本趋势，MiniMax、阶跃星辰较早实践 MoE 模型。Mistral、微软、Meta、Google 的小模型性能不断突破，为端侧 AI 打下良好基础，成为与大模型并驾齐驱的另一条重要发展路线。

Scaling Law 未达边界，算力换智能仍然成立

OpenAI 在 Scaling Law 论文中，从理论上预测了边界递减的存在。但实际上，OpenAI、Google 和 Anthropic 仍在践行大参数等于高智能的路线。清华唐杰教授在 24 年 2 月北京人工智能产业创新发展大会上指出，Scaling Law 尽头远未到来，算力换智能继续成立。在参数持续变大的情况下，训练数据的需求量进一步提升，据 Epoch 预测，2030 年到 2050 年，将耗尽低质量语言数据的库存，未来训练数据的缺乏将可能减缓机器学习模型的规模扩展。因此，合成数据或成为关键。

AI Agent 是 AGI 的关键范式，具身智能是大模型重要落地场景

AI Agent 能够自主、全流程、多步骤的执行任务，大幅延展了大模型的能力范围，被认为是实现 AGI 的关键范式。斯坦福大学吴恩达教授在 24 年 3 月的红杉美国 AI 峰会上指出，如果用户围绕 GPT-3.5 使用一个 Agent 工作流程，其实际表现甚至好于 GPT-4。并且 AI Agent 的能力能够充分受益于大模型的演进。此外，大模型与机器人具身智能的结合（如 OpenAI 与 Figure），也有望随着模型能力的迭代快速发展。我们认为，24 年 AI Agent 和具身智能将成为新一代大模型的重要落地场景。

GPT-5 有望推动全球算力和应用的下一阶段发展

我们预期 GPT-5：1) MoE 架构将延续，专家参数和数量或变大；2) GPT-5 及之后模型的训练数据集质量更高、规模更大；3) 在思维链 CoT 的基础上，再加一层 AI 监督；4) 支持更多外部工具调用的端到端模型；5) 多种大小不同的参数，不排除推出端侧小模型；6) 从普通操作系统到 LLM 操作系统；7) 端侧 AI Agent 将更加实用和智能。我们认为，GPT-5 的发布有望推动全球算力和应用的下一阶段发展，建议关注：海外标的，AI 应用：微软、Adobe 等。国内标的，1) AI 服务器：浪潮信息等；2) AI 应用：金山办公、福昕软件、泛微网络等；3) 端侧：中科创达、网宿科技。

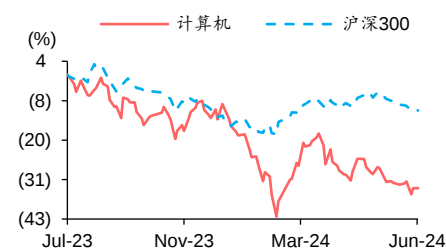
风险提示：宏观经济波动，技术进步不及预期，中美竞争加剧。本研报中涉及未上市公司或未覆盖个股内容，均系对其客观公开信息的整理，并不代表本研究团队对该公司、该股票的推荐或覆盖。

计算机

增持（维持）

研究员 谢春生
SAC No. S0570519080006 xiechunsheng@htsc.com
SFC No. BQZ938 +(86) 21 2987 2036
联系人 袁泽世，PhD
SAC No. S0570122080053 yuanzeshi@htsc.com
+(86) 21 2897 2228

行业走势图



资料来源：华泰研究

重点推荐

股票名称	股票代码	目标价 (当地币种)	投资评级
微软 (Microsoft)	MSFT US	477.33	买入
奥多比 (Adobe)	ADBE US	613.45	买入
浪潮信息	000977 CH	50.67	买入
金山办公	688111 CH	354.50	买入
福昕软件	688095 CH	73.96	买入
泛微网络	603039 CH	41.97	买入
中科创达	300496 CH	62.65	买入
网宿科技	300017 CH	12.08	买入

资料来源：华泰研究预测

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

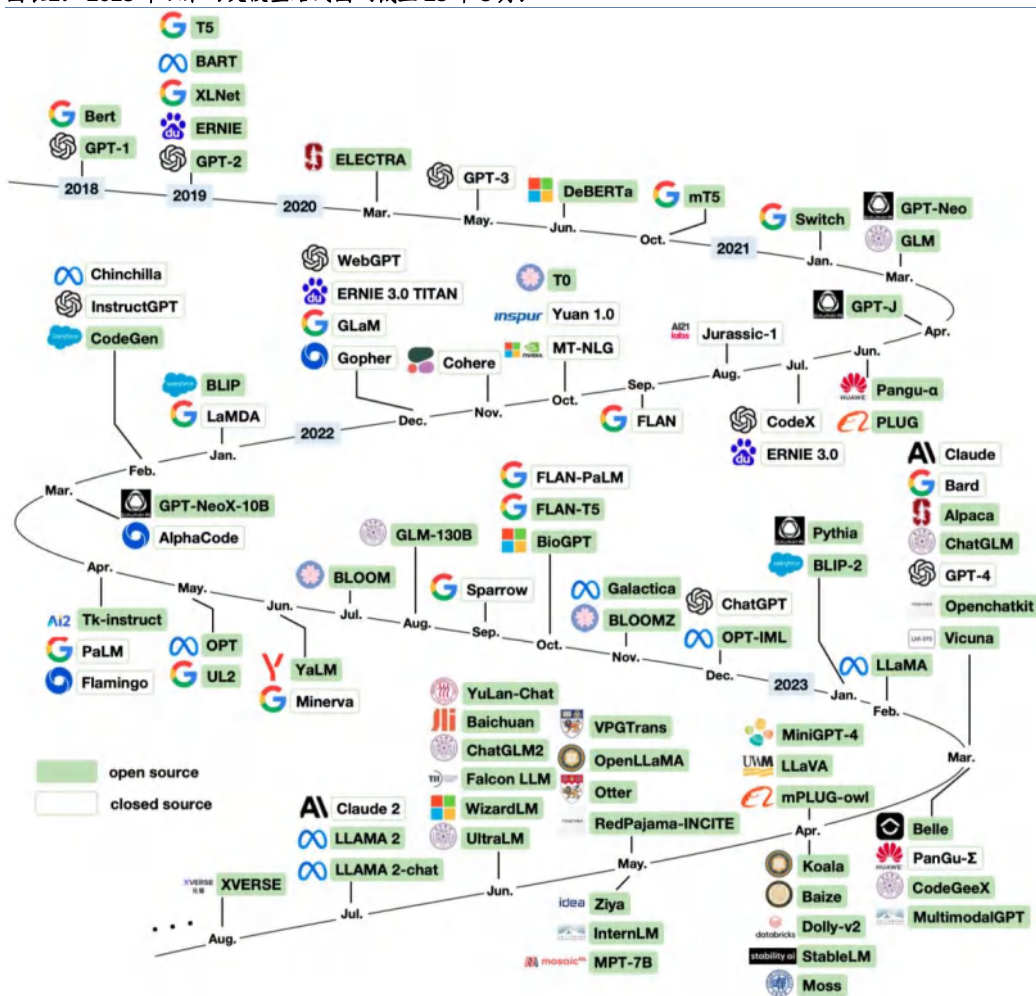
正文目录

大模型复盘：全球格局与模型特点基本明晰.....	3
全球格局：海外技术收敛，国内百花齐放.....	3
特点#1：大模型与小模型同步发展	7
特点#2：原生多模态逐步成为头部大模型的标配能力	11
特点#3：上下文作为 LLM 的内存，是实现模型通用化的关键.....	16
特点#4：MoE 是模型从千亿到万亿参数的关键架构.....	20
大模型展望：Scaling Law + AI Agent + 具身智能.....	23
展望#1：Scaling Law 理论上边界，但是目前仍未到达	23
展望#2：模型幻觉短期难消除但可抑制，CoT 是典型方法.....	24
展望#3：开源模型将在未来技术生态中占据一席之地	25
展望#4：数据将成为模型规模继续扩大的瓶颈，合成数据或是关键.....	27
展望#5：新的模型架构出现，但是 Transformer 仍是主流	29
展望#6：AI Agent 智能体是 AGI 的加速器	31
展望#7：具身智能与 LLM 结合落地加速	33
GPT-5 的几个预期	36
预期#1：MoE 架构将延续，专家参数和数量或变大.....	36
预期#2：GPT-5 及之后模型的训练数据集质量更高、规模更大	37
预期#3：在思维链 CoT 的基础上，再加一层 AI 监督.....	37
预期#4：支持更多外部工具调用的端到端模型	38
预期#5：多种大小不同的参数，不排除推出端侧小模型.....	39
预期#6：从普通操作系统到 LLM 操作系统.....	39
预期#7：端侧 AI Agent 将更加实用和智能	40
理想 vs 现实：从 AI+到+AI	42
大模型产业链相关公司及主要逻辑	46
相关产业链公司梳理	46
相关产业链公司逻辑	46
国产大模型初创公司投融资情况	54
风险提示.....	56

大模型复盘：全球格局与模型特点基本明晰

2023 年是大模型（LLM）技术和应用快速迭代的一年。重要催化剂是 22 年 11 月底发布的 ChatGPT。ChatGPT 虽然在技术基座上是之前已经问世的 GPT-3 和 InstructGPT，但它给了全球用户一个与 LLM 交互的自然语言界面，极大拉近了 LLM 与普通大众的距离，引起了资本的关注，成为大模型技术加速迭代的导火索。微软、Google、Meta、Nvidia 等龙头大厂，OpenAI、Anthropic、Mistral 等初创公司，以及斯坦福、清华、上交等学术机构，引领了 23 年的 LLM 发展。LLM 技术也从模型本身扩展到端侧、AI Agent、具身智能等更广泛的领域。此外，在大模型技术应用上，一方面，云 SaaS 厂商将 AI 赋能于传统 SaaS 软件，如微软 Copilot 和 Adobe Firefly；另一方面，以 AI 为核心的应用兴起，如 AI 搜索 Perplexity，文生图 Stable Diffusion、Midjourney、DALL-E，文生视频 Runway、Pika、Sora 等。

图表1: 2018 年以来的大模型路线图 (截至 23 年 8 月)



资料来源：《Examining User-Friendly and Open-Sourced Large GPT Models: A Survey on Language, Multimodal, and Scientific GPT Models》，Kaiyuan Gao（2023.8）、华泰研究

全球格局：海外技术收敛，国内百花齐放

海外开源大模型已经形成 OpenAI 为首，Google、Anthropic 等模型紧随的格局。开源模型中，虽然 Google Gemini 和 Anthropic 分别于 24 年 2 月和 3 月更新了 1.5 Pro(Gemini 1.0 是在 23 年 12 月)和 Claude 3，在上下文长度、数学、编码、专业领域等能力测评上超过了 GPT-4，但是考虑到：1) GPT-4 和 4 Turbo 实质上为 23 年 3 月 GPT-4 系列的迭代，比 Gemini 和 Claude 3 早推出近一年；2) ChatGPT 对多模态、App 语音交互、工具调用（联网、高级数据分析）、智能体(GPTs)等能力进行了有机整合；3)根据 UC 伯克利大学 Chatbot Arena 的榜单（该榜单为用户盲测模型评价的结果，较为客观），GPT-4 的用户体验仍是头部顶尖水平；4) GPT-5 已在训练中；5) GPT-4o 的端到端能力再次提升。因此，我们认为，OpenAI 的技术仍处于暂时领先。



图2： Gemini 1.0 在部分测评集上超越 GPT-4

	Gemini Ultra	Gemini Pro	GPT-4
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	90.04% CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)
GSM8K Grade-school math (Cobbe et al., 2021)	94.4% Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	53.2% 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	83.6% 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)
HumanEval Python coding tasks (Chen et al., 2021)	74.4% 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)
Natural2Code Python code generation. (New held-out set with no leakage on web)	74.9% 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	82.4 Variable shots	74.1 Variable shots	80.9 3-shot (reported)
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	95.3% 10-shot (reported)
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	74.4 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)

资料来源：Gemini 1.0 技术报告、华泰研究

图3： Claude 3 Opus 在部分测评集上超越 GPT-4

		Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4
MMLU General reasoning	5-shot	86.8%	79.0%	75.2%	86.4%
	5-shot CoT	88.2%	81.5%	76.7%	—
MATH Mathematical problem solving	4-shot	61%	40.5%	40.9%	52.9% ^[1]
	0-shot	60.1%	43.1%	38.9%	42.5% (from [2])
	Maj@32 4-shot	73.7%	55.1%	50.3%	—
GSM8K Grade school math		95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% SFT, 5-shot CoT
HumanEval Python coding tasks	0-shot	84.9%	73.0%	75.9%	67.0% ^[3]
GPQA (Diamond) Graduate level Q&A	0-shot CoT	50.4%	40.4%	33.3%	35.7% (from [4])
	Maj@32 5-shot CoT	59.5%	46.3%	40.1%	—
MGSM Multilingual math		90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% ^[5] 8-shot
DROP Reading comprehension, arithmetic	F1 Score	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot
BIG-Bench-Hard Mixed evaluations	3-shot CoT	86.8%	82.9%	73.7%	83.1% ^[6]
ARC-Challenge Common-sense reasoning	25-shot	96.4%	93.2%	89.2%	96.3%
HellaSwag Common-sense reasoning	10-shot	95.4%	89.0%	85.9%	95.3%
PubMedQA Biomedical questions	5-shot	75.8%	78.3%	76.0%	74.4%
	0-shot	74.9%	79.7%	78.5%	75.2%
WinoGrande Common-sense reasoning	5-shot	88.5%	75.1%	74.2%	87.5%
RACE-H Reading comprehension	5-shot	92.9%	88.8%	87.0%	—
APPS Python coding tasks	0-shot	70.2%	55.9%	54.8%	—
MBPP Code generation	Pass@1	86.4%	79.4%	80.4%	—

资料来源：Claude 3 技术报告、华泰研究

图4： UC 伯克利大学 Chatbot Arena 榜单（用户盲选评出）

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1287	+3/-4	51645	OpenAI	Proprietary	2023/10
2	Claude 3.5 Sonnet	1271	+4/-4	18007	Anthropic	Proprietary	2024/4
2	Gemini-Advanced-0514	1266	+3/-3	39732	Google	Proprietary	Online
3	Gemini-1.5-Pro-API-0514	1263	+2/-3	44131	Google	Proprietary	2023/11
5	GPT-4-Turbo-2024-04-09	1257	+3/-3	70880	OpenAI	Proprietary	2023/12
5	Gemini-1.5-Pro-API-0409-Preview	1257	+3/-3	56068	Google	Proprietary	2023/11
7	GPT-4-1106-preview	1250	+2/-3	85722	OpenAI	Proprietary	2023/4
7	Claude 3 Opus	1247	+2/-2	141415	Anthropic	Proprietary	2023/8
7	GPT-4-0125-preview	1245	+2/-3	79693	OpenAI	Proprietary	2023/12
9	Yi-Large-preview	1238	+4/-3	45584	01 AI	Proprietary	Unknown
11	Gemini-1.5-Flash-API-0514	1228	+3/-4	39983	Google	Proprietary	2023/11
12	Yi-Large	1216	+5/-6	10113	01 AI	Proprietary	Unknown
12	Gemma-2-27B-it	1214	+6/-7	5875	Google	Gemma license	2024/6
12	Bard (Gemini Pro)	1208	+5/-5	11897	Google	Proprietary	Online
12	GLM-4-0520	1207	+6/-5	8731	Zhipu AI	Proprietary	Unknown
13	Llama-3-70B-Instruct	1207	+3/-3	143625	Meta	Llama 3 Community	2023/12

注：截至 2024 年 6 月 30 日

资料来源：LMSYS Chatbot Arena、华泰研究

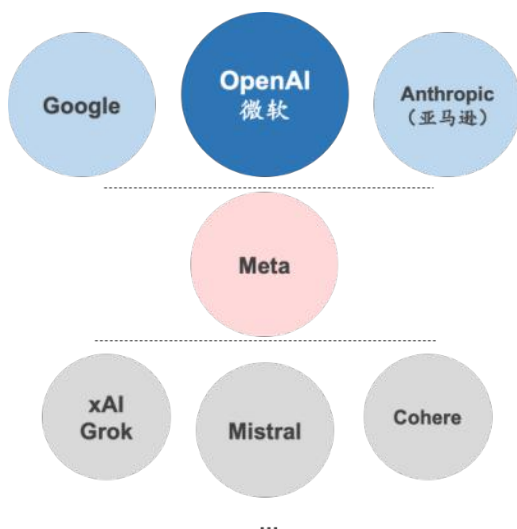
图表5: OpenAI vs Google vs Anthropic

公司	主要模型系列	输入模态	输出模态	上下文长度	架构特点	训练数据token数 (D)	参数 (N)	训练算力 (C)	微调支持
OpenAI (微软)	GPT-4o	文本、图片、音频、视频	文本、图片、音频	128K	Transformer解码器架构, MoE。实现了端到端多模态	GPT-4公开信息猜测是13T token	GPT-4公开信息猜测是1.8万亿, MoE	公开信息猜测是大约 2.15e25 的 FLOPS	可以
Google	Gemini 1.0 Nano / Pro / Ultra	文本、图片、视频、音频	文本、图片 (目前未安装)	Gemini 1.0为32K, 1.5已经上升到2M, 甚至内部实验中实现了100M	1.5开始采用MoE架构, 实现了端到端多模态	未公开。参考Gemma的2T和6T, 我们取个中间值4T来测算	Nano大小1.8B/3.25B, 其他未公开。假设Gemini Ultra和GPT-4差不多, 参数2T	根据公式 $C=6ND$, Nano 3.25B为1.3e22 FLOPS, Gemini Ultra约8e24 FLOPS	暂时不可微调
	Gemini 1.5 Pro / Flash								
	Gemma (开源)	文本	文本	8K	Transformer解码器架构	2B和7B分别在2T和6T tokens上训练	2B/7B	根据公式 $C=6ND$, 分别约4e21和4.22e22 FLOPS	开源, 可以自己微调
Anthropic (亚马逊)	Claude 3 Haiku / Sonnet / Opus	文本、图片	文本	200K, 可以向特定客户提供1M	未公开, 基于Claude 3技术文档中写的模型仍然是预测下一个token, 应该仍然是Transformer解码器概率更大	未公开。我们根据GPT和Gemma的情况, 大致认为是4T	未公开。但是Opus的性能看, 跟GPT-4可能差不多, 猜测1T	根据公式 $C=6ND$, Opus对应的是4e24 FLOPS	Claude 2.1需要填表申请。Claude 3暂时未披露

资料来源: 各公司官网、华泰研究

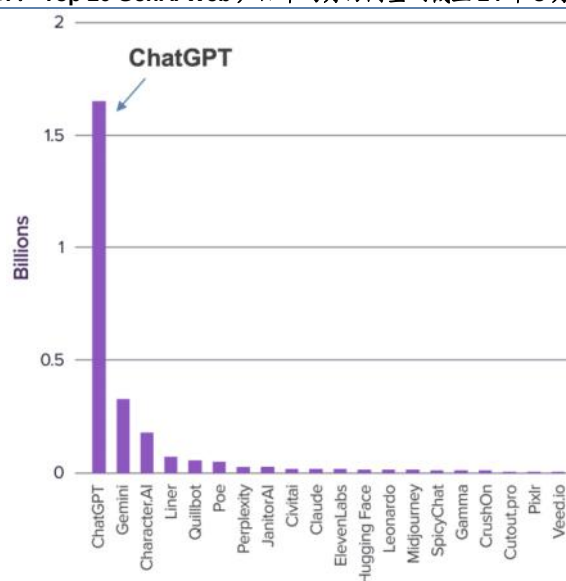
Meta 的 Llama 系列作为开源模型, 具有格局上的特殊性和分界性。海外模型厂商如果在模型性能上无法超越同代的开源 Llama 模型 (据 Meta 官网 4 月 18 日信息, Llama 3 的 8B 和 70B 先行版小模型已经发布, 最大的 400B 参数正在训练), 则很难在海外基础模型中占据一席之地, 除非模型具有差异化应用场景, 典型的如陪伴类应用 Character.ai。此外, 除了头部大参数模型, 能够超过同代 Llama 的较小参数或者有独特使用体验的模型, 也会得到用户青睐, 典型的如: 1) 马斯克旗下 xAI 的 Grok-1 (已开源)、Grok-1.5 (未开源), 能够独家使用 X 平台上的数据, 较好的响应用户实时信息查询需求; 2) 法国大模型初创公司 Mistral, 开源了 Mistral 7B、Mixtral 8x7B-MoE 小模型, 适配算力受限的端侧等平台, 随后又转入闭源模型, 更新了性能更强的 Mistral-medium 和 large, 并与微软合作, 在 Azure 上为用户提供 API。

图表6: Meta 在海外模型厂商中格局的特殊性



资料来源: 各公司官网、华泰研究

图表7: Top 20 GenAI Web 产品平均月访问量 (截至 24 年 3 月)



资料来源: a16z、华泰研究

图表8: Grok-1 开源信息



资料来源: xAI 官网、华泰研究

图表9: Mistral 的模型谱系和价格 (截至 2024 年 5 月 28 日)

开源		
	输入价格 \$/1M tokens	输出价格 \$/1M tokens
open-mistral-7b	0.25	0.25
open-mistral-8x7b	0.7	0.7
open-mistral-8x22b	2	6
闭源模型		
mistral-small	1	3
mistral-medium	2.7	8.1
mistral-large	4	12

资料来源: Mistral 官网、华泰研究

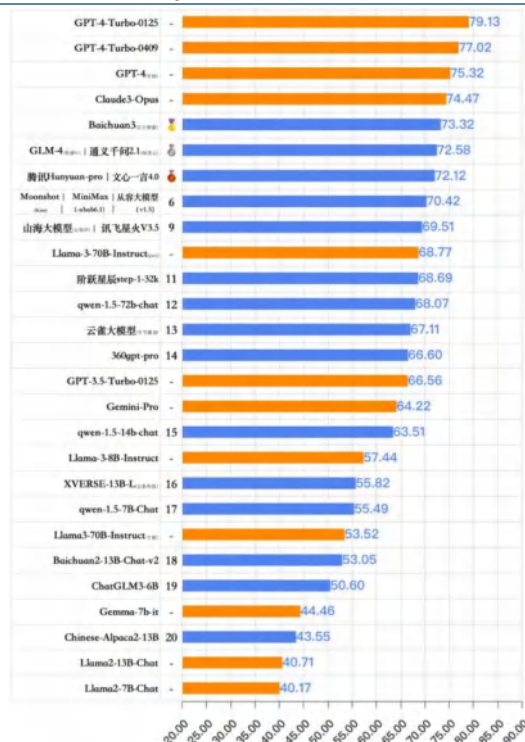
国内模型百花齐放, 互联网大厂、初创公司、科技企业均有代表性模型产品。国内模型技术辨识度不高, 据 SuperCLUE 测评结果榜单, 头部的国内模型在得分上相差并不显著。在国内主流的模型中, 互联网厂商和科技企业在大型模型上起步较早, 如百度在 GPT-4 发布的一天即 23 年 3 月 15 日发布文心一言, 23 年 3 月 29 日 360 智脑 1.0 发布, 23 年 4 月通义千问上线, 23 年 5 月 6 日讯飞星火 1.0 发布。进入 24 年, 初创公司的大模型产品得到了更广泛的关注, 例如 24 年 3 月月之暗面更新 Kimi 智能助手 200 万字的上下文支持能力, 直接引发了百度、360 等厂商对长上下文的适配。同月阶跃星辰 STEP 模型发布, 其 STEP 2 宣称万亿参数 MoE 模型, 直接对标 GPT-4 的参数 (一般认为是 1.8 T 参数的 MoE), 在大多数国内模型以千亿参数为主的环境下, 将参数量率先提升到万亿级别。4 月, MiniMax 也发布了万亿参数 MoE 架构的 abab 6.5。

图表10: 国内主流大模型格局



注: 绿色为互联网厂商, 黄色为初创公司, 灰色为科技公司
 资料来源: 各公司官网、华泰研究

图表11: 国内外模型 SuperCLUE 榜单



注: 截至 2024 年 4 月
 资料来源: CLUE 官网、华泰研究

图12：国内主流模型的对比（截至24年2月）

厂商类型	厂商产品	核心开源情况	用户情况	大模型平台	C端付费会员	API价格（最先进模型）	API调用	上下文长度	文件上传	多模态输入	多模态输出	插件能力	Agent能力	App	相关硬件
互联网科技厂商	文心一言		23年底用户超1亿	文心千帆	49.9元/月（23年11月）	ERNIE-Bot 4.0: 0.12元/千tokens	不支持	会员5000字，非会员2000字	支持	识图	图片、音频	有	有	有	
	HUAWEI		B端为主，暂无数据	华为云Stack											手机小艺
	通义千问	底座模型Qwen“全尺寸、全模态”开源，通义千问闭源	截至2024年2月，开源模型累计下载量超过300万	人工智能平台PAI		qwen-plus: 0.02元/1,000 tokens	定制API调用需申请	1000万字	支持	识图	图片			有	天猫精灵+AI
	腾讯混元		2023年12月超过50个腾讯业务和产品已经接入腾讯混元大模型测试	腾讯云 TI 平台		高级版: 0.10元/千tokens		16k			图片				小程序
人工智能企业	天工	Skywork-13B开源	2024年1月20日当天，天工全平台的日活跃用户数达到了34万			SkyChat-MegaVerse: 0.005元/千tokens		100K		识图			有	有	
	讯飞星火		截至今年1月，讯飞星火注册用户2400万			普通套餐: 0.3元/万tokens		最大8xx	仅支持图片	识图	图片	有	有	有	学习机、一体机
	SenseChat							32k							
大模型创业企业	ChatGLM	开源	ChatGLM过去一年与200多家企业进行了深度合作			GLM-3-Turbo: 0.005元/千tokens GLM-4/4V: 0.1元/千tokens	支持（需询价），或基于128k开源自己微调		支持图片和文档	识图	图片		有	有	
	MINIMAX		MiniMax 开放平台平均单日的 token 处理量达到了数亿			abab5.5: 0.015元/千tokens Abab6 (MoE): 0.1元/千tokens	支持, 0.06元/千tokens	16k					有专门的语音模型		
	百川智能	开源				0.016元/千tokens	开源模型支持	Baichuan2-Turbo-192k	仅支持文档						
	阶跃星辰		留存率大约在20%到30%之间					200k					有	有	
	Kimi智能助手		2024年2月，Kimi的月活跃用户数达到298.46万人		免费		支持	200万	支持				有	有	
高校研究院	BAI 智源研究院	开源，且悟道3.0为一组不同的模型					支持，开源的模型系列可做微调	Aquila 34B: 4k							
	紫东太初		截至23年11月，已吸引百余家单位和企业加入			0.0020元/千tokens		文本模型输入上限1000字	支持	视觉、文本		有		有	

资料来源：各公司官网、华泰研究

特点#1：大模型与小模型同步发展

根据 **Scaling Law**，更大参数、更多数据和更多算力能够得到更好的模型智能。2020年1月，OpenAI 发布论文《Scaling Laws for Neural Language Models》，奠定了 Scaling Law（缩放定律）的基础，为后续 GPT 的迭代指明了大参数、大算力方向。Scaling Laws 是一种经验性质的结论，并非完备的数学理论推导。OpenAI 在 decoder-only Transformer 架构的特定配置下进行了详尽的实验，摸清了模型性能（用模型 Loss 衡量，Loss 越小性能越好）与参数（N）、数据集 token（D）和投入训练算力（C）的关系——N、D、C 是影响 Loss 最显著的因素，三者增加将带来更好的模型性能。Transformer 架构中的层数、向量宽度等其它参数并不构成主要影响因素。

图13：Scaling Law 最主要的结论：训练时 N、D 和 C 增加将带来更好的模型性能

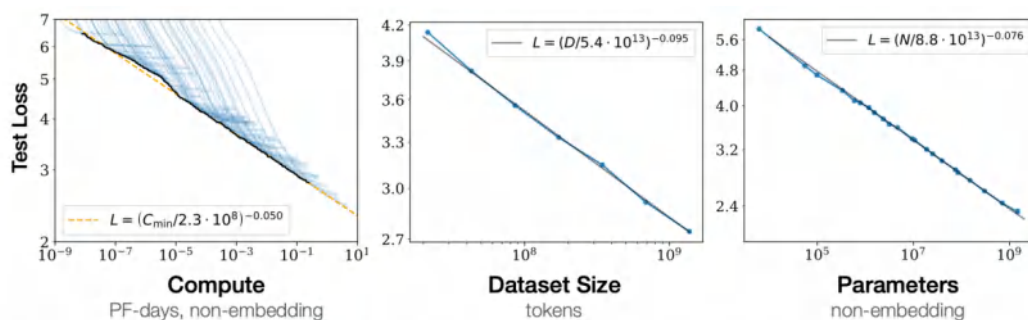


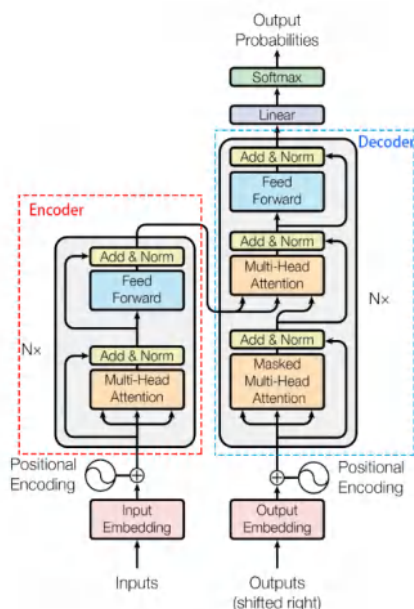
Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

资料来源：《Scaling Laws for Neural Language Models》，OpenAI（2020）、华泰研究

根据 **Scaling Law** 论文, 可以用 $6ND$ 来估算模型所需要的训练算力 (以 FLOPs 为单位)。Transformer 架构涉及了多种参数, 包括层数 (n_{layer})、残差流维数 (d_{model})、前馈层维数 (d_{ff})、注意力机制输出维数 (d_{attn})、每层注意力头数 (n_{head})、输入上下文 token 数 (n_{ctx}) 等。在训练数据进入 Transformer 解码器后, 每一步运算都会涉及相应的参数, 并对应有需求的算力。据 OpenAI 测算, 单个 token 训练时在 Transformer 解码器中正向传播, 所需 FLOPs (每秒浮点运算数) 为 $2N + 2n_{\text{layer}}n_{\text{ctx}}d_{\text{attn}}$ 。由于在论文写作于 2020 年, 当时模型上下文长度 n_{ctx} 并不长, 满足 $d_{\text{model}} > n_{\text{ctx}}/12$, 因此 $2N + 2n_{\text{layer}}n_{\text{ctx}}d_{\text{attn}}$ 可约等于 $2N$ 。在训练中反向传播时, 所需算力约为正向的 2 倍 (即 $4N$), 因此单个 token 训练全过程需要算力总共 $6N$ FLOPs, 考虑全部的训练 token 数 D , 共需算力近似 $6ND$ FLOPs。在推理时, 为了计算方便, 通常采用正向训练算力需求 $2ND$ 来计算所需 FLOPs。

值得注意的是, 目前 Claude 3、Gemini 1.5 Pro、Kimi 智能助手等大模型支持的上下文长度远超当年, $d_{\text{model}} > n_{\text{ctx}}/12$ 不再满足, 因此 $2n_{\text{layer}}n_{\text{ctx}}d_{\text{attn}}$ 应予以考虑。即上下文长度更长时, 训练需求的算力是高于 $6ND$ 的。

图表 14: Transformer 经典架构



图表 15: 单个 token 正向传播需要每一步的算力拆解

- n_{layer} — (number of layers)
- d_{model} — (dimension of the residual stream)
- d_{ff} — (dimension of the intermediate feed-forward layer)
- d_{attn} — (dimension of the attention output)
- n_{heads} — (number of attention heads per layer)
- n_{ctx} — (tokens in the input context)

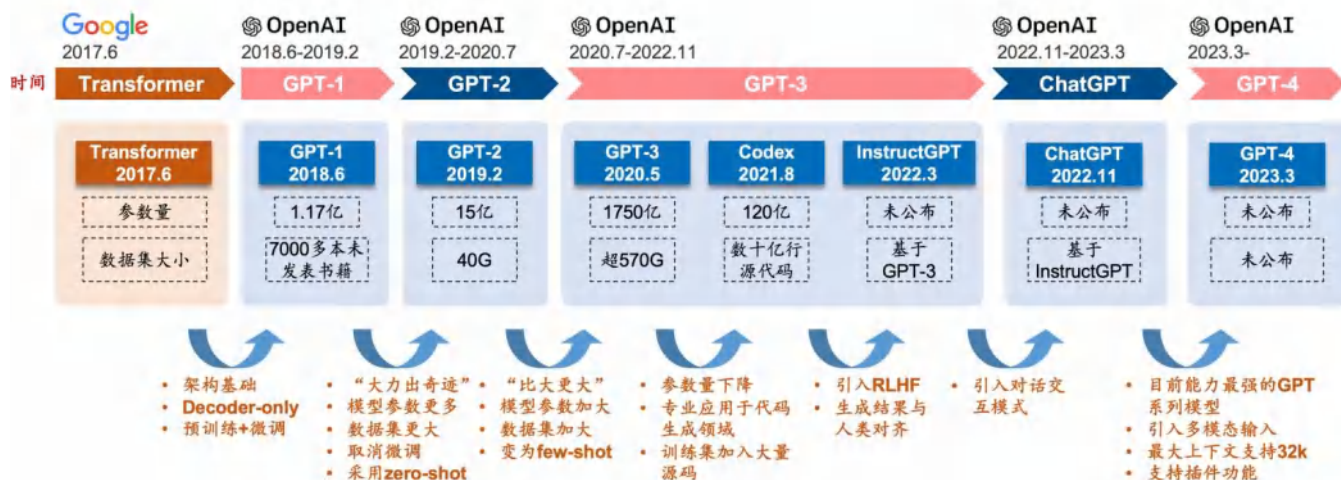
Operation	Parameters	FLOPs per Token
Embed	$(n_{\text{vocab}} + n_{\text{ctx}}) d_{\text{model}}$	$4d_{\text{model}}$
Attention: QKV	$n_{\text{layer}} d_{\text{model}} 3d_{\text{attn}}$	$2n_{\text{layer}} d_{\text{model}} 3d_{\text{attn}}$
Attention: Mask	—	$2n_{\text{layer}} n_{\text{ctx}} d_{\text{attn}}$
Attention: Project	$n_{\text{layer}} d_{\text{attn}} d_{\text{model}}$	$2n_{\text{layer}} d_{\text{attn}} d_{\text{model}}$
Feedforward	$n_{\text{layer}} 2d_{\text{model}} d_{\text{ff}}$	$2n_{\text{layer}} 2d_{\text{model}} d_{\text{ff}}$
De-embed	—	$2d_{\text{model}} n_{\text{vocab}}$
Total (Non-Embedding)	$N = 2d_{\text{model}} n_{\text{layer}} (2d_{\text{attn}} + d_{\text{ff}})$	$C_{\text{forward}} = 2N + 2n_{\text{layer}} n_{\text{ctx}} d_{\text{attn}}$

注: OpenAI Scaling Laws 论文中仅涉及右半部分解码器。其 GPT 系列模型同样仅涉及解码器部分

资料来源: 《Attention is all you need》, Ashish (2017)、华泰研究

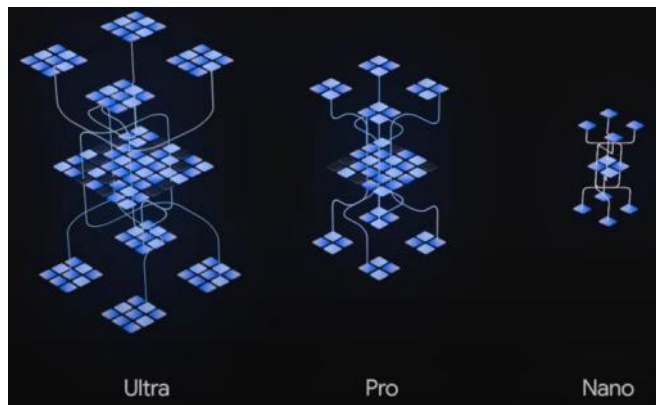
资料来源: 《Scaling Laws for Neural Language Models》, OpenAI (2020)、华泰研究

在 **Scaling Law** 指导下, OpenAI 延续了大参数模型的路线。2020 年 1 月 Scaling Laws 论文发表后不久, 2020 年 5 月 GPT-3 系列问世, 将参数从 GPT-2 的 15 亿提升到 1750 亿, 训练数据大小从 40G 提升到 570G (数据处理后, 处理前数据量更大), 分别提升了 100+ 倍和 14 倍。到了 GPT-4, 虽然 OpenAI 官方未公布参数大小, 但是根据 SemiAnalysis 的信息, 目前业界基本默认了 GPT-4 是 1.8 万亿参数的 MoE 模型, 训练数据集包含约 13 万亿个 token, 使用了约 25,000 个 A100 GPU, 训练了 90 到 100 天, 参数量、数据集和训练所需算力相比 GPT-3 又有数量级的提升。OpenAI 在不断践行 **Scaling Law**, 将模型的参数以及模型的智能提升到新的层级。

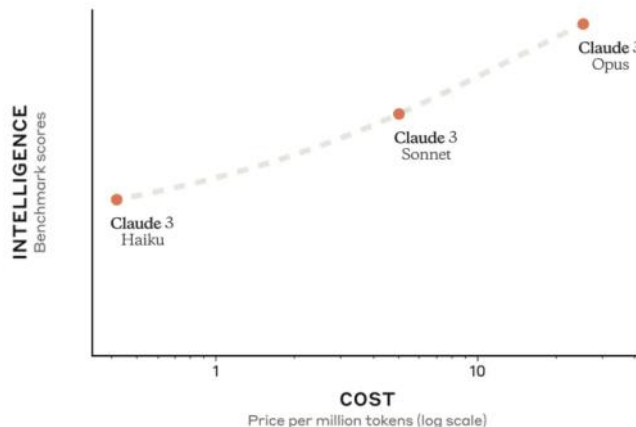
图表16: Scaling Laws 指导 OpenAI 将模型参数和智能提高到新的层级


资料来源: OpenAI GPT 1-4 系列论文、华泰研究

从 Google 和 Anthropic 的模型布局看, 印证了大参数能带来模型性能的提升。Google 的 Gemini 和 Anthropic 的 Claude 3 系列均分别提供了“大中小”三款模型, 虽然两家厂商并未给出模型参数、训练数据细节, 但是均表示更大的模型智能更强(参见图表 2 和图表 3), 推理速度相对较慢, 所需的算力和训练数据也相应更多, 是对 Scaling Law 的印证。此外, 我们梳理了全球主流模型厂商的参数情况, 同样发现旗舰模型的参数数量仍在变大。

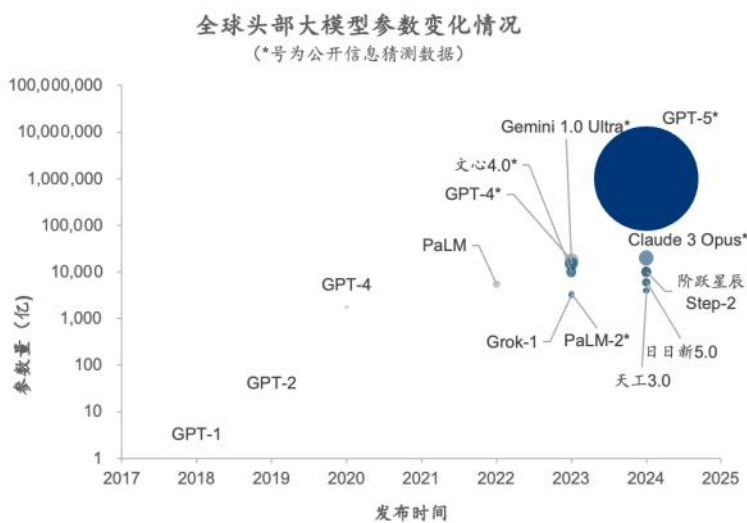
图表17: Gemini 系列: Ultra、Pro 和 Nano, 参数依次减少


资料来源: Google 官网、华泰研究

图表18: Claude 系列: Haiku、Sonnet、Opus, 智能依次升高


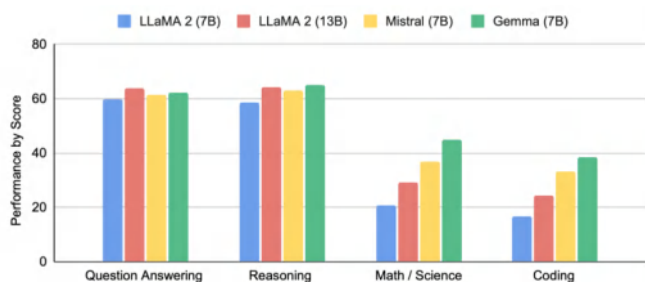
资料来源: Anthropic 官网、华泰研究

我们认为, 全球头部闭源模型的参数目前呈现的规律是: 跨代际更新, 模型参数进一步加大; 同代际更新, 随着模型技术架构优化和软硬件资源协同能力提高, 在模型性能不降的情况下, 参数或做的更小。Google 和 OpenAI 的最新模型都呈现了这个趋势。24 年 5 月 13 日, OpenAI 发布了 GPT-4o 模型, 在多模态端到端的架构基础上, 实现了更快的推理速度, 以及相比于 GPT-4 Turbo 50%的成本下降, 我们推测其模型参数或在下降。5 月 14 日 Google 发布了 Gemini 1.5 Flash, 官方明确指出 Flash 是在 Pro 的基础上, 通过在线蒸馏的方式得到, 即 Flash 的参数小于 Pro。

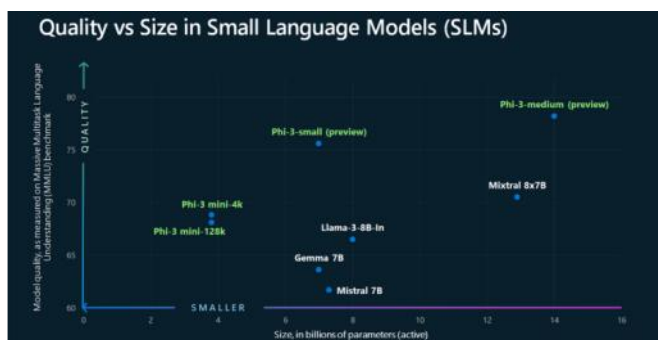
图表19：全球主流模型厂商旗舰模型的大参数发展趋势


资料来源：各公司官网、华泰研究

大参数并不是唯一选择，小参数模型更好适配了终端算力受限的场景。Google 的 Gemini 系列是典型代表，其最小的 Nano 包括 1.8B 和 3.25B 两个版本，并且已经在其 Pixel 8 Pro 和三星 Galaxy S24 上实现部署，取得了不错的终端 AI 效果。此外，Google 在 24 年 2 月开源了轻量级、高性能 Gemma (2B 和 7B 两种参数版本)，与 Gemini 模型技术同源，支持商用。Google 指出，预训练和指令调整的 Gemma 模型可以在笔记本电脑、工作站、物联网、移动设备或 Google Cloud 上运行。微软同样在 23 年 11 月的 Ignite 大会上提出了 SLM (小语言模型) 路线，并将旗下的 Phi 模型升级到 Phi-2，参数大小仅 2.7B，性能超过 7B 参数的 Llama 2。24 年 4 月 Phi-3 发布，最小参数仅 3.8B，其性能超过参数量大其两倍的模型，5 月微软 Build 大会上，Phi-3 系列参数为 7B 和 14B 的模型发布。

图表20：Gemma 的测试结果：超过同参数模型


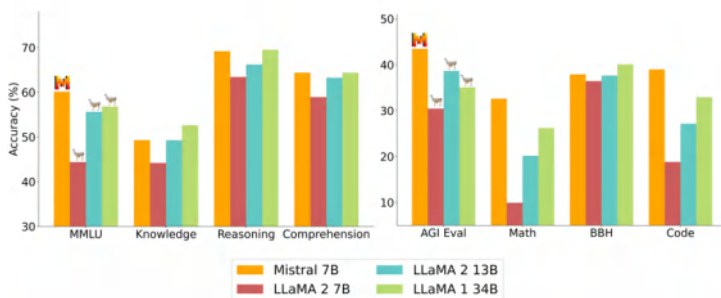
资料来源：Google 官网、华泰研究

图表21：Phi-3 mini 的测试结果：超过更大参数的 Llama 3


资料来源：微软官网、华泰研究

Mistral 发布的 7B 和 8x7B 模型也是开源小模型的典型代表。法国人工智能初创公司 Mistral AI 成立于 2023 年 5 月，其高管来自 DeepMind、Facebook 等核心 AI 团队。2023 年 9 月和 12 月，Mistral 分别开源了 Mistral-7B (73 亿参数) 和 Mixtral-8x7B-MoE (467 亿参数，8 个专家)。Mistral-7B 在多项测试基准中优于 130 亿参数的 Llama 2-13B。Mixtral-8x7B-MoE 在大多数测试基准上超过 Llama 2，且推理速度提高了 6 倍；与 GPT-3.5 相比，也能在多项测评基准上达到或超过 GPT-3.5 水平。在小参数开源模型中，Mistral 的竞争力很强。Mistral 推出的平台服务 La plateforme 也支持模型的 API 调用 (参见图表 9)。

图表22: Mistral-7B 的基准测试结果



资料来源: Mistral 官网、华泰研究

图表23: Mixtral-8x7B-MoE 的基准测试结果

	LLaMA 2 70B	GPT - 3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	70.6%
HellaSwag (10-shot)	87.1%	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	85.8%
WinoGrande (5-shot)	83.2%	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	60.7%
GSM-8K (5-shot)	53.6%	57.1%	58.4%
MT Bench (for Instruct Models)	6.86	8.32	8.30

资料来源: Mistral 官网、华泰研究

小参数模型的训练算力需求仍在变大, 定性看, 训推算力需求空间可观。虽然模型参数较小, 但是为了提高性能, 模型厂商均投入了大量的训练数据。如 Phi-2 有 1.4T 训练数据 tokens, Phi-3 为 3.3T tokens, Gemma 为 6T/2T tokens (分别对应 7B 和 2B 模型)。24 年 4 月 Meta 率先开源的两个 Llama 3 系列小模型 8B 和 70B, 对应的训练 token 已经达到了 15T, 并且 Meta 表示, 即使已经使用了 15T 的训练数据, 仍能看到模型性能的持续提升。我们认为, 虽然单个小模型相比于大模型训练算力需求并不大, 但是一方面小模型本身的训练数据集在不断增加, 另一方面, 未来在终端 AI PC 和手机, 甚至车机和机器人上, 都有可能部署终端模型, 因此定性看, 小模型总体的训练和推理算力需求仍然可观。

特点#2: 原生多模态逐步成为头部大模型的标配能力

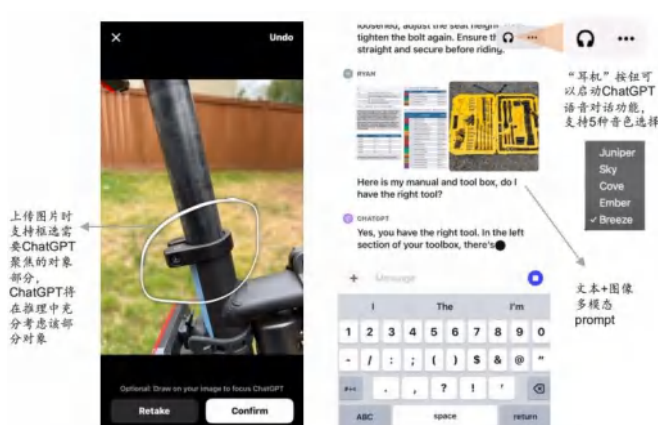
OpenAI 的 GPT 系列在全球闭源大语言模型厂商中率先适配多模态能力。抛开专门的多模态模型/产品, 如文生图 Stable Diffusion / Midjourney / DALL-E, 文生视频 Sora / Runway / Pika / Stable Video Diffusion 外, 在头部闭源 LLM 中, OpenAI 的 GPT-4 最先引入多模态能力。23 年 3 月, GPT-4 技术报告中即展示了 GPT-4 支持文本和图像两种模态作为输入。9 月 25 日, OpenAI 官方 Blog 宣布 GPT-4 的 Vision (视觉) 能力上线, 支持多图和文本的交错推理, 同时宣布 ChatGPT App 支持语音交互 (语音转文本模型为 Whisper, 文本转语音模型为 Voice Engine)。23 年 10 月 19 日, OpenAI 旗下新一代文生图模型 DALL-E 3 在 ChatGPT 中实装上线, 可以通过与 ChatGPT 对话来实现文生图。

图表24: GPT-4 技术报告率先宣布支持多模态图像输入



资料来源: GPT-4 技术报告、华泰研究

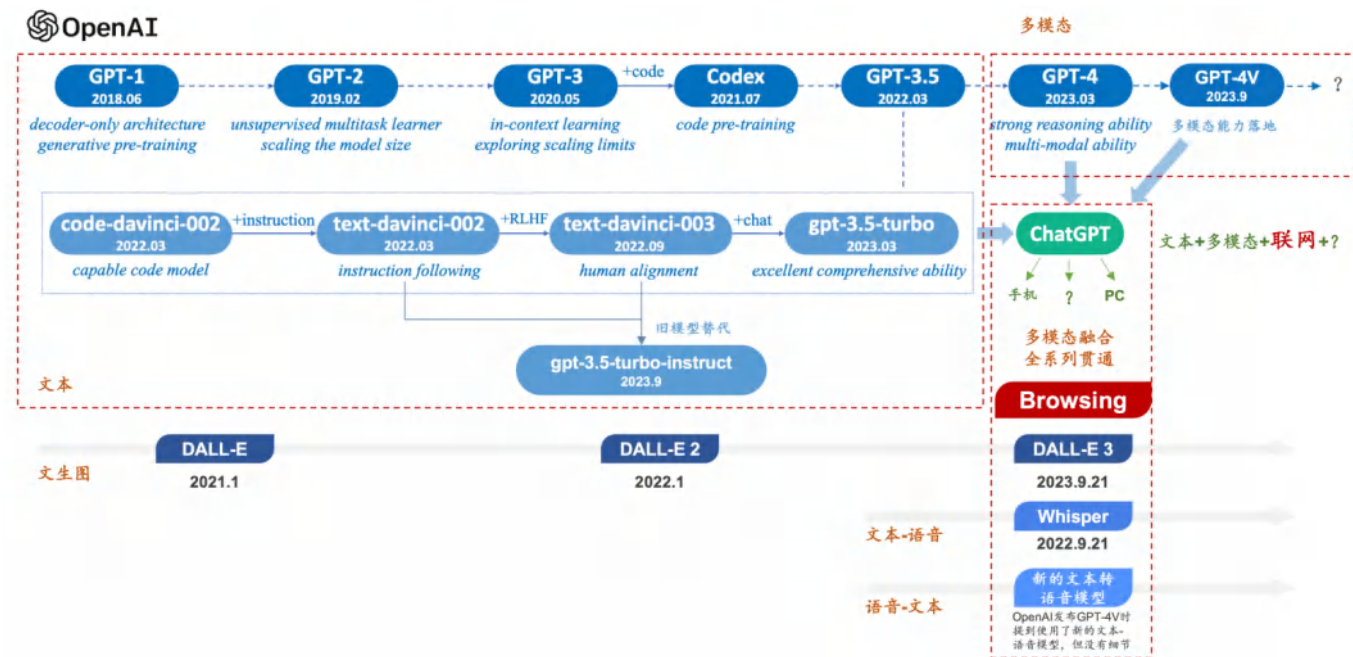
图表25: ChatGPT App 对图像和语音的支持



资料来源: OpenAI 官网、华泰研究

通过模型间非端到端协作，ChatGPT 网页端和 App 实现了完备的多模态能力支持。随着 OpenAI 的 GPT-4V、DALL-E 3、Whisper、Voice Engine 等模型的上线和更新，OpenAI 将所有的模型协同集成成 pipeline 形式，使得 ChatGPT 能够实现：1) 推理文本；2) 理解图像；3) 生成图像；4) 语音转文本；5) 文本转语音。ChatGPT 成为 2023 年支持模态最多的 LLM 产品。

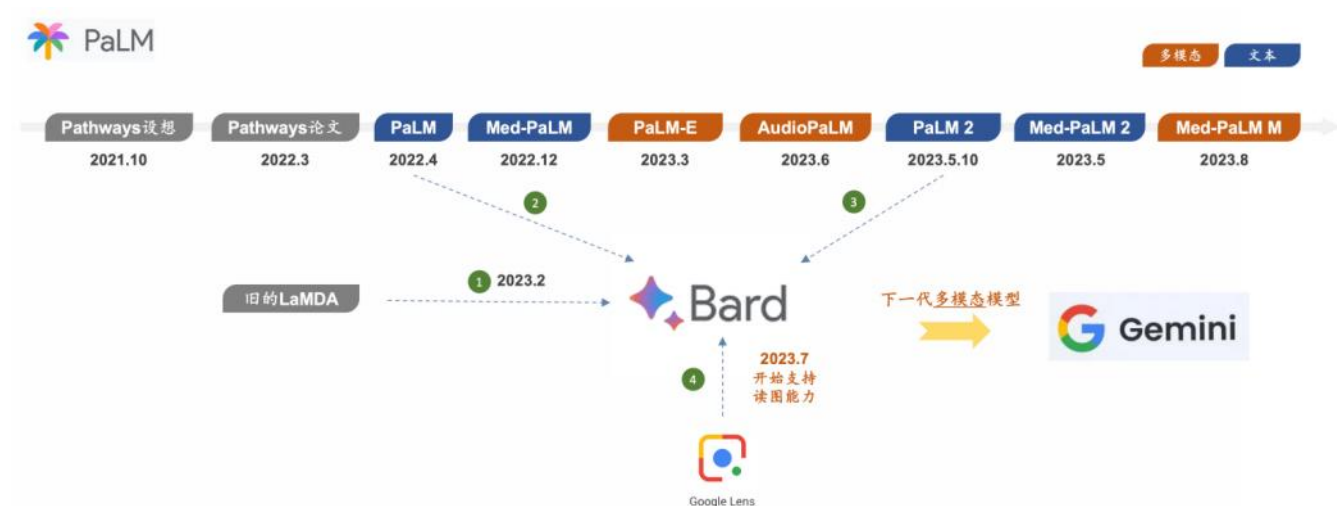
图表26：ChatGPT 多模态能力溯源



资料来源：OpenAI 官网、华泰研究

Google 从 PaLM 模型开始即在探索 LLM 向多模态领域的拓展。PaLM 是 Google Gemini 的前一代主要模型系列。2022 年 4 月，Google 的 PaLM 模型问世。PaLM 自身为大语言模型，仅支持文本模态，但是在 PaLM 的能力之上，Google 将图像、机器人具身数据转化为文本 token 形式，训练出多模态模型 PaLM-E。此外，还将音频模态与 PaLM 模型结合，发布 AudioPaLM。在医疗领域，Google 先基于 PaLM 训练出医疗语言模型 Med-PaLM，随后在 Med-PaLM 基础上将医疗图像知识增加到训练数据中，训练出医疗领域多模态模型 Med-PaLM M。

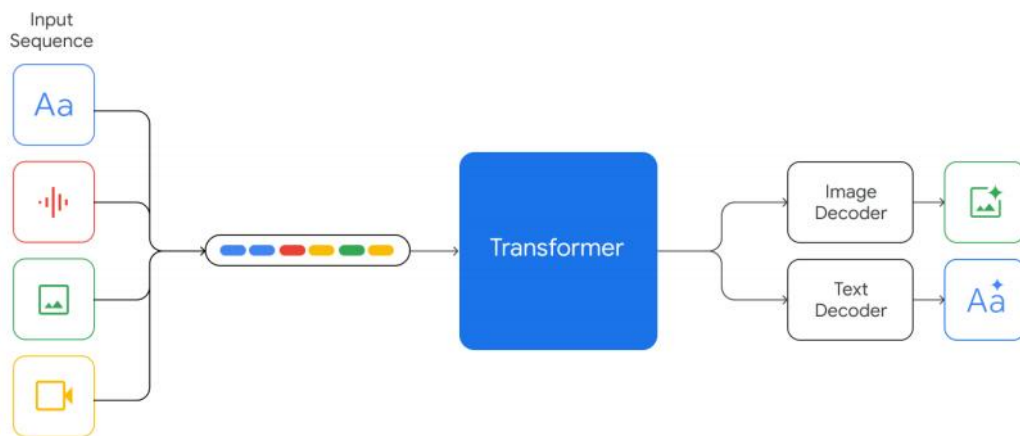
图表27：PaLM 模型从文本模态到多模态的演进路线



资料来源：Google 官网、华泰研究

Gemini 模型问世后，端到端原生多模态能力成为头部模型厂商的“标配”能力。2023 年 5 月的 I/O 大会上，Google 宣布了下一代模型 Gemini，但未透露细节。12 月，Gemini 1.0 模型发布，配备了 Ultra/Pro/Nano 三种参数大小依次递减的型号。Gemini 同样支持文本、图像、视频、音频等多模态，但是其范式和 OpenAI 的 ChatGPT 有很大区别：**ChatGPT 属于多种不同模型的集合，每个模型负责不同的模态，结果可以串联；而 Gemini 具备端到端的原生多模态能力，Gemini 模型自身可以处理全部支持的模态。**据 The Decoder 信息，23 年 OpenAI 内部已经在考虑一种代号为“Gobi”的新模型，该模型同样从一开始就被设计为原生多模态。我们认为，这种端到端的原生多模态范式将成为未来头部大模型厂商实现多模态的主流范式。

图表28： Gemini 1.0 技术报告中展示的端到端原生多模态架构



资料来源：Gemini 1.0 技术报告、华泰研究

Anthropic Claude 模型多模态能力“虽迟但到”，Claude 3 模型科研能力优异。Anthropic 的 Claude 系列模型在 2024 年 3 月更新到 Gen 3 后，全系适配了多模态图像识别能力，并在科学图表识别上大幅超越 GPT-4 和 Gemini 1.0 Ultra。此外，Claude 3 Haiku 有着优秀的成本控制和推理速度优势，据 Anthropic 官方，Haiku 的速度是同类产品的三倍，能够在一秒内处理约 30 页的内容（21K token），使企业能够快速分析大量文档，例如季度备案、合同或法律案件，且一美元就能分析 400 个最高法院案例或 2500 张图片。

图表29： Claude 3 多模态评测结果对比

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4V	Gemini 1.0 Ultra	Gemini 1.0 Pro
Math & reasoning MMMU (val)	59.4%	53.1%	50.2%	56.8%	59.4%	47.9%
Document visual Q&A ANLS score, test	89.3%	89.5%	88.8%	88.4%	90.9%	88.1%
Math MathVista (testmini)	50.5% CoT	47.9% CoT	46.4% CoT	49.9%	53.0%	45.2%
Science diagrams AIZD, test	88.1%	88.7%	86.7%	78.2%	79.5%	73.9%
Chart Q&A Relaxed accuracy (test)	80.8% 0-shot CoT	81.1% 0-shot CoT	81.7% 0-shot CoT	78.5% 4-shot CoT	80.8%	74.1%

资料来源：Anthropic 官网、华泰研究

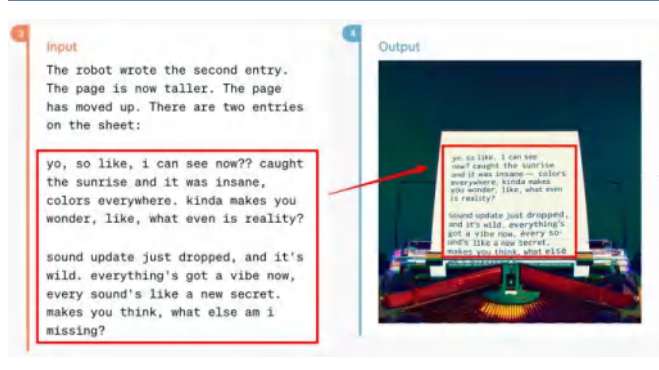
GPT-4o 在 GPT-5 发布之前实现了端到端的多模态支持，验证了原生多模态的技术趋势。 24 年 5 月 14 日 Google I/O 大会前夕，OpenAI 发布了新版模型 GPT-4o (omni)，弃用了之前 ChatGPT 拼接 GPT-4V、Whisper、DALL-E 的非端到端模式，统一了文本、图像、音频和视频模态，以端到端的方式，实现了输入文本、图像、音频和视频，输出文本、图像和音频，追上了 Google Gemini 的原生多模态进度，并且模态支持更加全面（4o 支持音频输出，Gemini 不支持）。4o 在文本、图像、音频等各项指标上均超越了同等级现有模型。

图表30：GPT-4o 实现了端到端的多模态支持



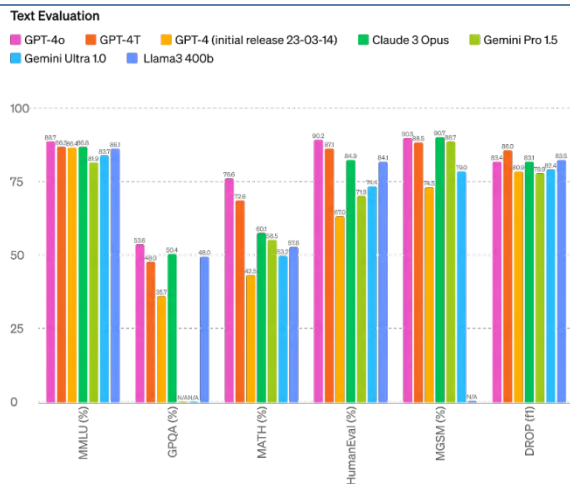
资料来源：OpenAI 官网、华泰研究

图表31：GPT-4o 的文生图能正确保留文字信息



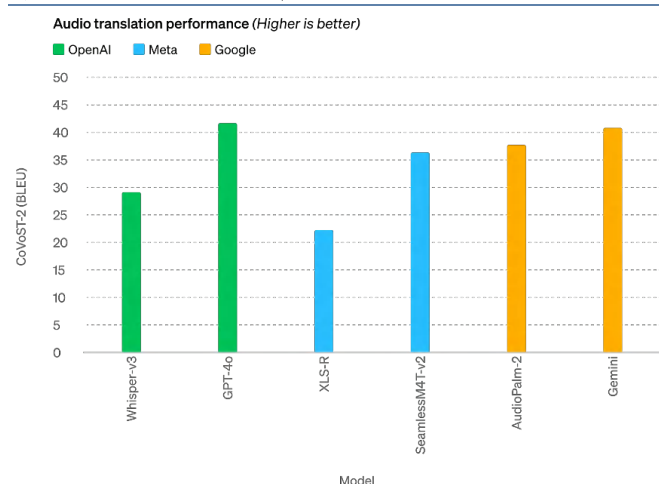
资料来源：OpenAI 官网、华泰研究

图表32：GPT-4o 在文本上的表现



资料来源：OpenAI 官网、华泰研究

图表33：GPT-4o 在语音翻译上的表现



资料来源：OpenAI 官网、华泰研究

Claude 3.5 Sonnet 增强了 UI 交互体验，与 GPT-4o 的语音交互相比朝着差异化路径发展。 6 月 21 日，Anthropic 宣布了 Claude 3.5 Sonnet 模型，在价格相比于 Claude 3 Sonnet 不变的情况下，在研究生水平推理、代码等能力（文本层面），以及视觉数学推理、图表问答等能力（视觉层面）上超过了 GPT-4o。Claude 3.5 Sonnet 另一个突出的性能是 UI 交互能力的增强，主要由 Artifacts 功能实现。当用户要求 Claude 生成代码片段、文本文档或网站设计等内容时，对话旁边的专用窗口中将实时出现相应的展示，例如编写的游戏、网页等。Anthropic 指出，Artifacts 交互方式未来将会从个人拓展到团队和整个组织协作，将知识、文档和正在进行的工作集中在一个共享空间中。我们认为，GPT-4o 和 Claude 3.5 Sonnet 均在优化用户交互上下功夫，但是两者的方向存在差异化，GPT-4o 更注重语音交互，而 Sonnet 更注重 UI 界面交互。

图表34: Claude 3.5 Sonnet 与其他模型在文本测评集的表现

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning GPQA, Diamond	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge MMLU	88.7%** 5-shot 88.3% 0-shot CoT	86.8% 5-shot 85.7% 0-shot CoT	— 0-shot CoT	85.9% 5-shot —	86.1% 5-shot —
Code HumanEval	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math MGSM	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text DROP, F1 score	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations BIG-Bench-Hard	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving MATH	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math GSM8K	96.4% 0-shot CoT	95.0% 0-shot CoT	—	90.8% 11-shot	94.1% 8-shot CoT

* Claude 3.5 Sonnet scores 67.2% on 5-shot CoT GPQA with maj@32
 ** Claude 3.5 Sonnet scores 90.4% on MMLU with 5-shot CoT prompting

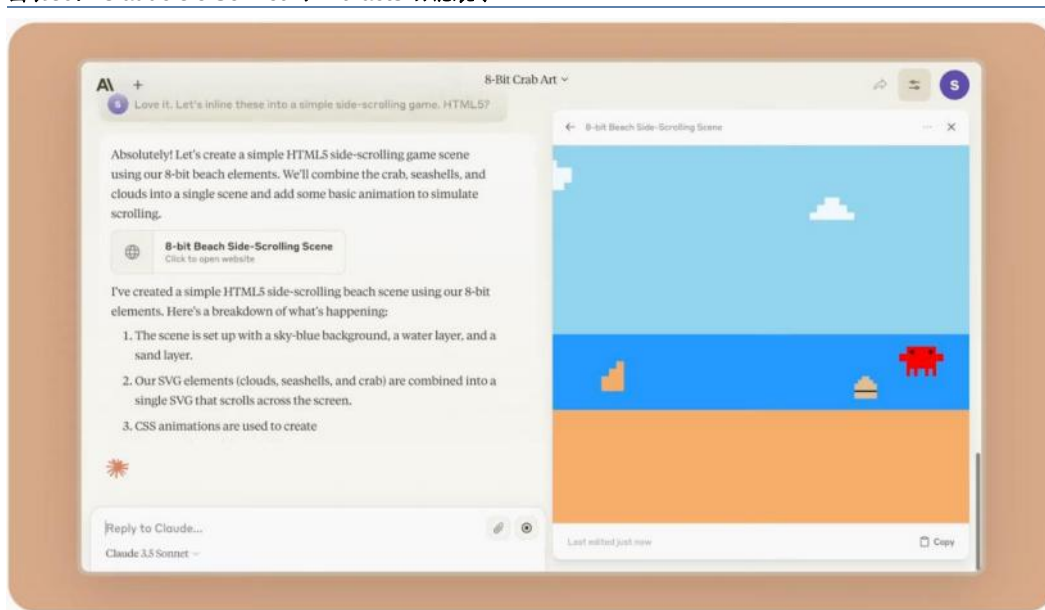
资料来源: Anthropic 官网、华泰研究

图表35: Claude 3.5 Sonnet 与其他模型在视觉测评集的表现

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
Visual math reasoning MathVista (testmini)	67.7% 0-shot CoT	50.5% 0-shot CoT	63.8% 0-shot CoT	63.9% 0-shot CoT
Science diagrams AIZD, test	94.7% 0-shot	88.1% 0-shot	94.2% 0-shot	94.4% 0-shot
Visual question answering MMMU (val)	68.3% 0-shot CoT	59.4% 0-shot CoT	69.1% 0-shot CoT	62.2% 0-shot CoT
Chart Q&A Reluend accuracy (test)	90.8% 0-shot CoT	80.8% 0-shot CoT	85.7% 0-shot CoT	87.2% 0-shot CoT
Document visual Q&A ANLS score, test	95.2% 0-shot	89.3% 0-shot	92.8% 0-shot	93.1% 0-shot

资料来源: Anthropic 官网、华泰研究

图表36: Claude 3.5 Sonnet 的 Artifacts 功能展示



注: UI 左侧为用户输入提示词区域, 右侧为 Artifacts 实时展示的编程结果

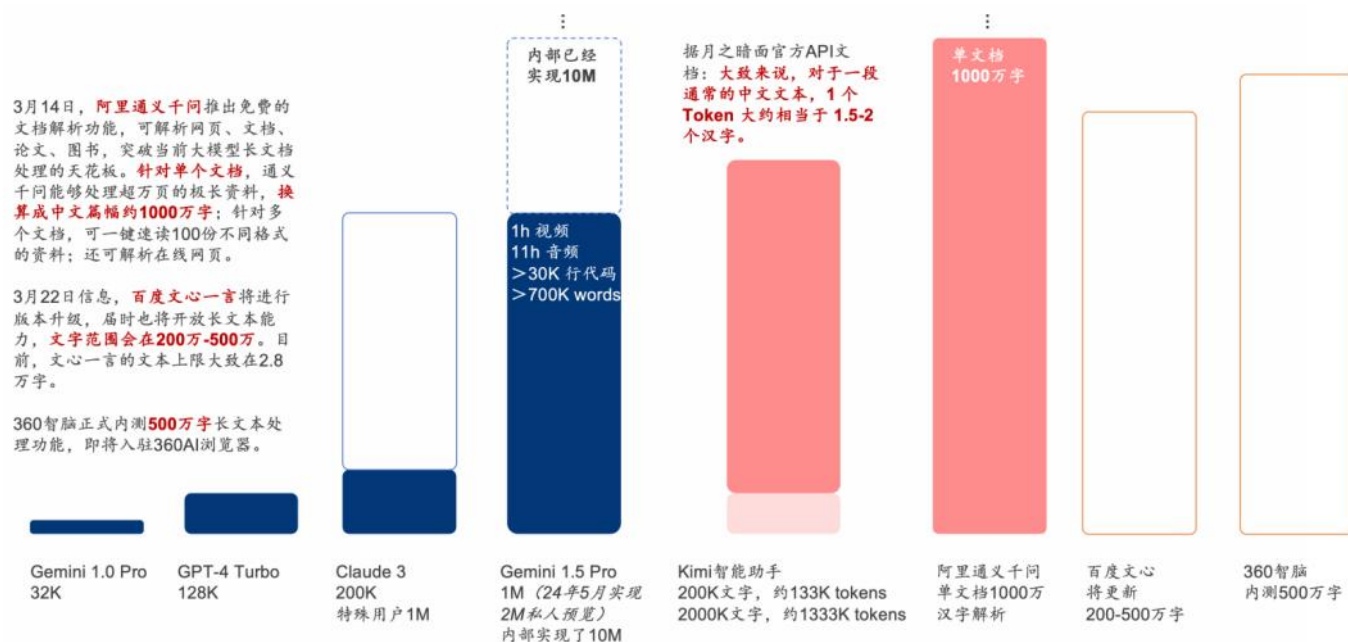
资料来源: Anthropic 官网、华泰研究

国内模型厂商积极适配多模态, 以图像理解能力为主。在 GPT-4 宣布支持多模态后, 国内厂商也积极适配多模态图片的识别、理解和推理。截至 2024 年 4 月, 国产主流模型多模态支持情况如下: 1) 百度文心一言, 说图解画支持单张图像推理, 支持图像生成。2) 阿里通义千问, 支持单张图片推理, 支持图像生成。阿里开源的模型 Qwen-VL 支持图像推理。3) 腾讯混元助手, 支持图像生成, 以及单张图像推理。3) 讯飞星火, 支持单张图像推理, 支持图像生成。4) 智谱 ChatGLM 4, 支持单张图像推理, 支持图像生成。5) 360 智脑, 支持图像生成。6) 字节豆包, 支持图像生成。7) Kimi 智能助手, 支持图片中的文字识别。月之暗面官方表示 24 年下半年将支持多模态推理。8) 阶跃星辰基于 Step 模型的助手跃问, 支持多图推理。

特点#3：上下文作为 LLM 的内存，是实现模型通用化的关键

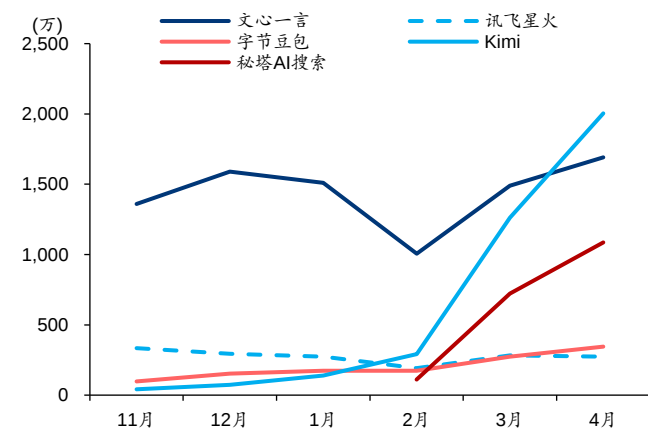
国外 LLM 厂商较早实现长上下文，国内厂商通过长上下文找到差异化竞争优势。国外较早实现长上下文的厂商是 Anthropic，旗下 Claude 模型在 23 年 11 月，将支持的上下文从 100K tokens 提升到 200K，同时期的 GPT-4 维持在 128K。24 年 2 月，Google 更新 Gemini 到 1.5 Pro 版本，将上下文长度扩展到 1M（5 月更新中扩展到 2M），并在内部实现了 10M，是目前已知最大上下文长度。国内方面，23 年 10 月由月之暗面发布的 Kimi 智能助手（原名 Kimi Chat），率先提供 20 万字的长上下文，并在 24 年迎来了用户访问量的大幅提升。24 年 3 月，阿里通义千问和 Kimi 先后宣布支持 1000 万字和 200 万字上下文，引发国内百度文心一言、360 智脑等厂商纷纷跟进长上下文能力迭代。我们认为，国内 LLM 厂商以长上下文为契机，寻找到了细分领域差异化的竞争路线，或有助于指导后续模型迭代。

图表37：全球主流模型厂商的长上下文布局（实线框代表暂未落地，实框代表已经落地）



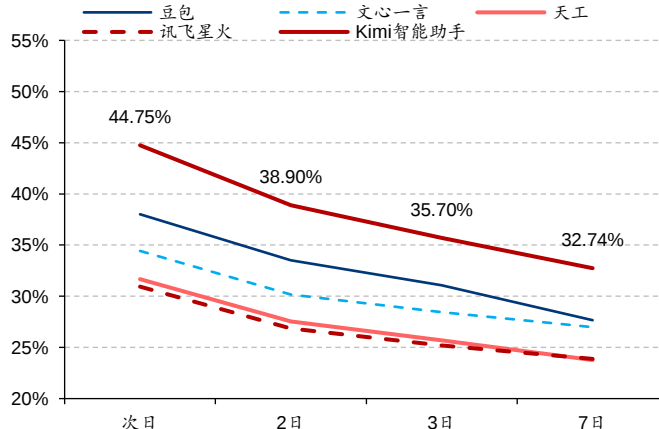
资料来源：各公司官网、华泰研究

图表38：国内主流大模型或产品访问量情况



资料来源：SimilarWeb、华泰研究

图表39：国内主流大模型 App 留存率情况（截至 23 年 3 月）



资料来源：QuestMobile、华泰研究

长上下文使得模型更加通用化。据月之暗面官方信息，长上下文能够解决 90% 的模型微调定制问题。对于短上下文模型，在执行具体的下游任务前，其已具备的能力往往仍有欠缺，需要针对下游任务进行微调。微调的基本步骤包括数据集的准备、微调训练等，中间可能还涉及微调结果不理想，需要重新梳理微调过程。而上下文长度足够的情况下，可以将数据作为提示词的一部分，直接用自然语言输入给大模型，让模型从上下文中学习，达到微调效果，使得模型本身更具有通用性。以 Google Gemini 1.5 Pro 为例，将 250K token 的 Kalamang 语（全球使用人数小于 200 人，几乎不存在于 LLM 的训练集中）直接作为上下文输入给模型，实现了接近人类的翻译水平。而 GPT-4 和 Claude 2.1 由于上下文支持长度不够，无法通过上下文学习到全部的知识。

图表40：Kalamang 语机器翻译评测结果对比（6 分为满分）

Model	kgv→eng	eng→kgv
	Human Evaluation (BLEURT)	Human Evaluation (chrF)
GPT-4 Turbo (0-shot)	0.24 (33.1)	0.1 (17.8)
GPT-4 Turbo (half book)	2.38 (51.6)	4.02 (48.3)
Claude 2.1 (0-shot)	0.14 (22.2)	0.00 (15.3)
Claude 2.1 (half book)	3.68 (57.1)	4.54 (52.5)
Gemini 1.5 Pro (0-shot)	0.24 (33.3)	0.08 (17.8)
Gemini 1.5 Pro (half book)	4.16 (63.4)	5.38 (58.3)
Gemini 1.5 Pro (full book)	4.36 (65.0)	5.52 (56.9)
Human language learner	5.52 (70.3)	5.60 (57.0)

资料来源：Gemini 1.5 Pro 技术报告、华泰研究

长上下文还能很好的适配虚拟角色、开发者、AI Agent、垂类场景等需求。

- 1) 虚拟角色 Chatbot:** 长文本能力帮助虚拟角色记住更多的重要用户信息，提高使用体验。
- 2) 开发者:** 基于大模型开发剧本杀等游戏或应用时，需要将数万字甚至超过十万字的剧情设定以及游戏规则作为 prompt 输入，对长上下文能力有着刚性需求。
- 3) AI Agent:** Agent 智能体运行需要自主进行多轮规划和决策，且每步行动都可能需要参考历史记忆信息才能完成。因此，短上下文会导致长流程中的信息遗忘，长上下文是 Agent 效果的重要保障。
- 4) 垂直场景客户需求:** 对于律师、分析师、咨询师等专业用户群体，有较多长文本内容分析需求，模型长上下文能力是关键。

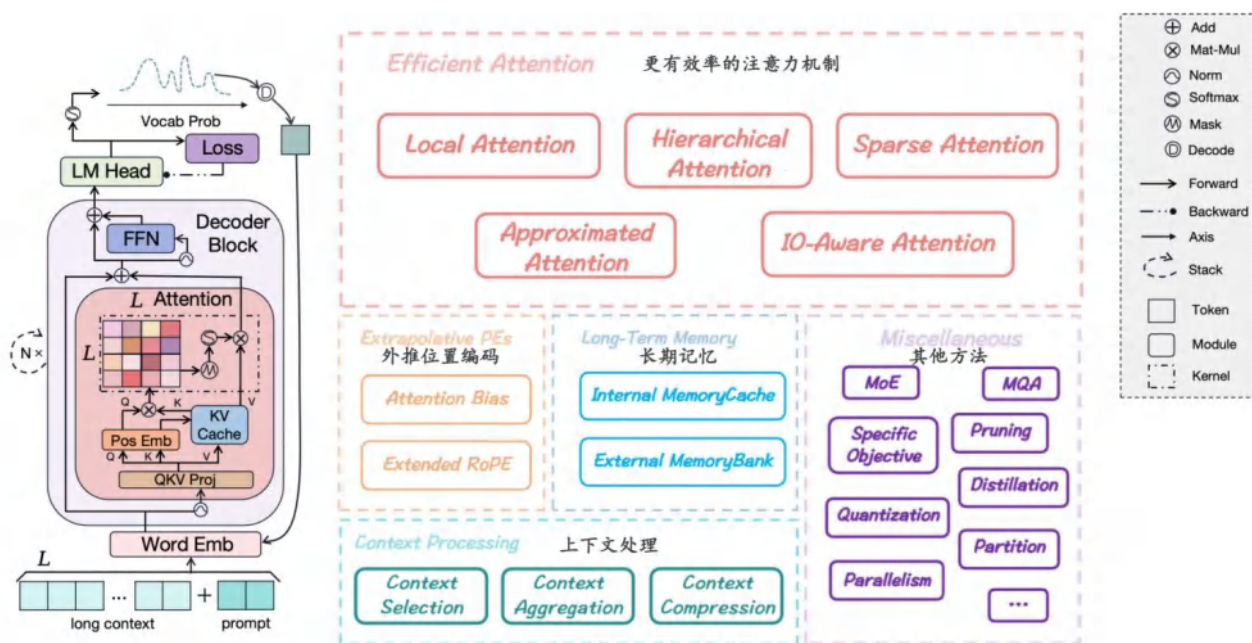
图表41：长上下文适配多种场景需求



资料来源：月之暗面官网、华泰研究

实现长上下文有多种方法,优化 Transformer 架构模块是核心。拆解 Transformer 解码器,可以通过改进架构中的各个模块来实现上下文长度的拓展。**1) 高效注意力机制:**高效的注意力机制能够降低计算成本,甚至实现线性时间复杂度。这样在训练时就可以实现更长的序列长度,相应的推理序列长度也会更长。**2) 实现长期记忆:**设计显式记忆机制,如给予外部存储,解决上下文记忆的局限性。**3) 改进位置编码 PE:**对现有的位置编码 PE 进行改进,实现上下文外推。**4) 对上下文进行处理:**用额外的上下文预/后处理,在已有的 LLM (视为黑盒) 上改进,确保每次调用中给 LLM 的输入始终满足最大长度要求。**5) 其他方法:**以更广泛的视角来增强 LLM 的有效上下文窗口,或优化使用现成 LLM 时的效率,例如 MoE (混合专家)、特殊的优化目标函数、并行策略、权重压缩等。更多细节可以参考华泰证券计算机团队 3 月 26 日研报《计算机:通过 Kimi,看长文本的实现》。

图表42: 从 Transformer 架构来拆解长上下文的改进方法



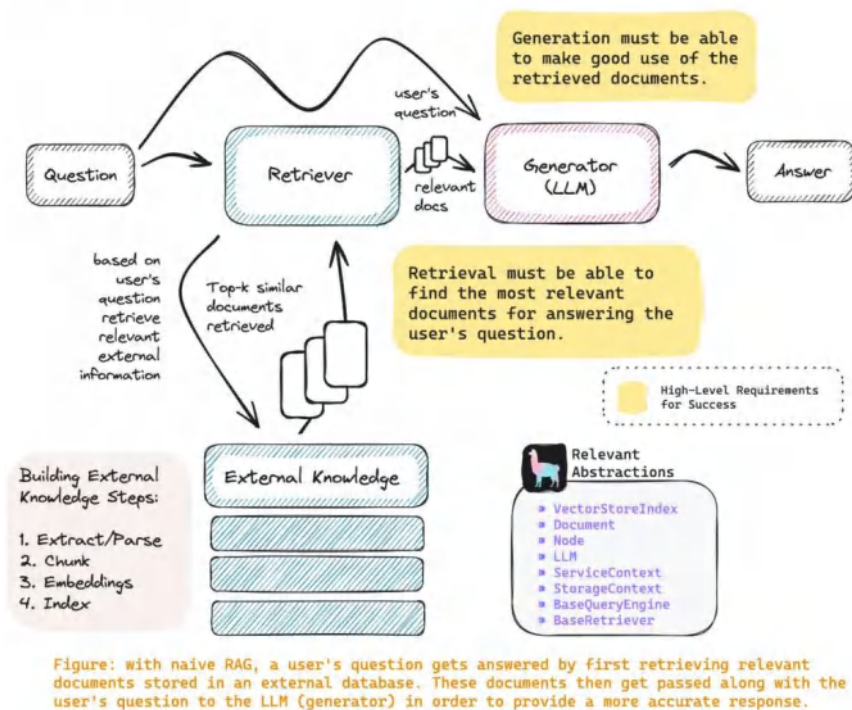
资料来源:《Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey》, Huang (2024)、华泰研究

RAG 与其他长文本实现方法相比,并没有显著的优劣之分,要结合场景进行选择。RAG 基本原理是,在用户提问时, retriever (检索器) 会从外部的知识库中检索最相关的信息传递给大模型,作为大模型推理所需知识的补充。RAG 更像是大模型本身的“外挂”帮手。而优化注意力机制等其他长上下文实现方法,则是大模型的“内生”能力,是模型本身能够支持输入更长的信息,并通过注意力机制掌握序列全局关系。“内生”似乎比“外挂”更高级,因为模型会捕捉到用户提出的所有历史信息,更适用于 C 端信息量有限场景。但是对于 B 端用户,其企业 Know-How 积累量巨大,且很多知识也是结构化的 QA (如客服),而模型上下文长度不可能无限延长(受制于算法、算力、推理时间等各种因素),因此 RAG 这种“外挂”的形式更加适合。例如,主要面向 B 端的大模型厂商 Cohere,将 RAG 作为模型重要能力以适配 B 端检索场景,其 Command R+模型本身上下文长度仅 128K。

我们认为,“内生”长文本技术是从根本上解决问题,是发展趋势,但是受制于算力等因素(未来或将逐步解决),短期内将与 RAG 共存,选择上取决于使用场景。

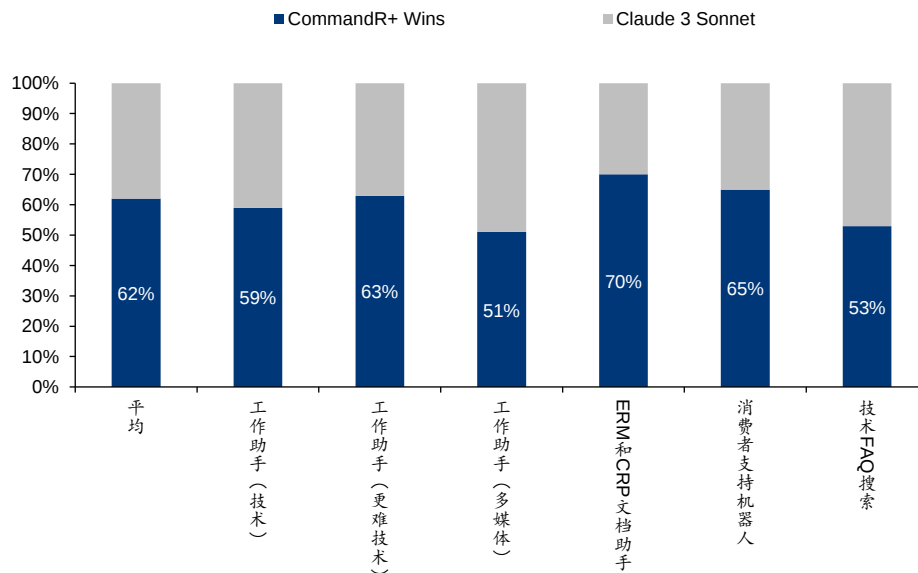
图表43：基础检索增强生成（RAG）的流程

Basic RAG



资料来源：CSDN、华泰研究

图表44：人类评判：企业 RAG 场景下 Command R+ 优于 Claude 3 Sonnet



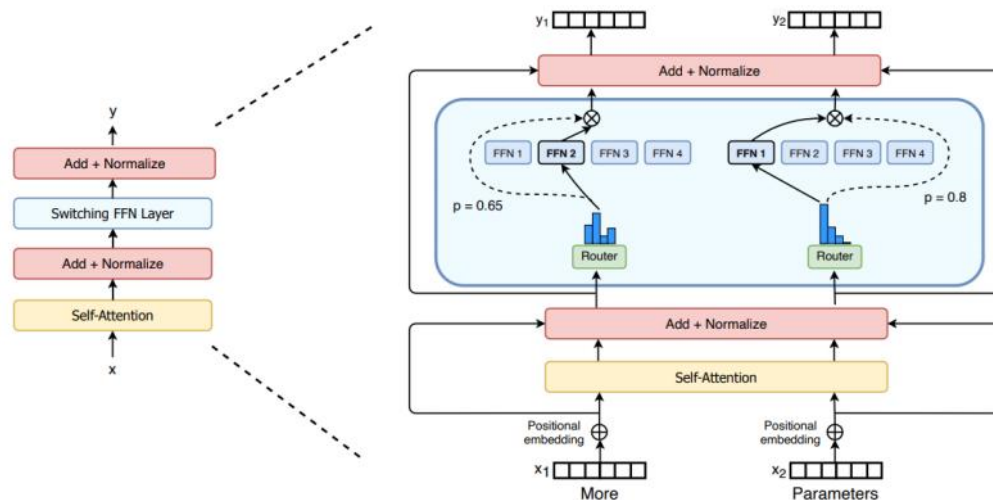
资料来源：Cohere 官网、华泰研究

特点#4: MoE 是模型从千亿到万亿参数的关键架构

MoE 架构有利于预训练和推理效率的提升,方便模型 **scale up** 到更大的参数。据 Hugging Face 信息,在有限的计算资源预算下,用更少的训练步数训练一个更大的模型,往往比用更多的步数训练一个较小的模型效果更佳。MoE 的一个显著优势是它们能够在远少于稠密模型所需的计算资源下进行有效的预训练,当计算资源有限时,MoE 可以显著扩大模型或数据集的规模,更快地达到稠密模型相同的质量水平。MoE 的引入使得训练具有数千亿甚至万亿参数的模型成为可能。MoE 特点在于:1)与稠密模型相比,预训练速度更快;2)与具有相同参数数量的模型相比,具有更快的推理速度(因为只需要调用部分参数);3)需要大量显存,因为所有专家系统都需要加载到内存中,而 MoE 架构的模型参数可达到上万亿;4)MoE 进行指令调优具有很大的潜力,方便做 Chatbot 类应用。

MoE 由稀疏 MoE 层和门控网络/路由组成。MoE 模型仍然基于 Transformer 架构,组成部分包括:1)稀疏 MoE 层:这些层代替了传统 Transformer 模型中的稠密前馈网络层,包含若干“专家”(例如 8、16、32 个),每个专家本身是一个独立的神经网络。这些专家甚至可以是 MoE 层本身,形成层级式的 MoE 结构。稀疏性体现在模型推理时,并非所有参数都会在处理每个输入时被激活或使用,而是根据输入的特定特征或需求,只有部分参数集合被调用和运行。2)门控网络/路由:决定将用户输入的 tokens 发送到哪个具体的专家。例如下图中,“More”对应的 token 被发送到第二个专家处理,而“Parameters”送到第一个专家。一个 token 也可以被发送到多个专家进行处理。路由器中的参数需要学习,将与网络的其他部分一同进行预训练。

图表45: Switch Transformers 论文中的典型 MoE 架构

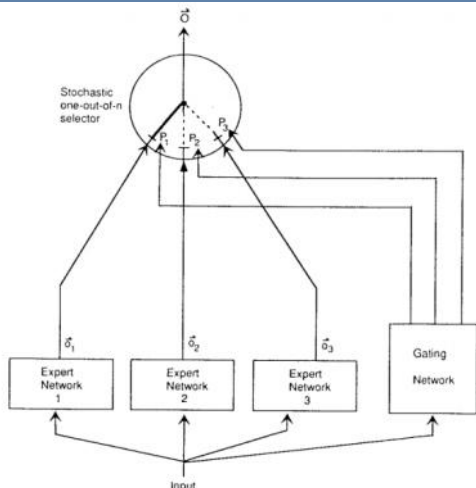


资料来源:《Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity》, Google (2021)、华泰研究

专家数量存在边际递减效应,MoE 的选择也要考虑模型的具体应用场景。据 Hugging Face 信息,增加更多专家可以加速模型的运算速度和推理效率,但这一提升随着专家数量的增加而边际递减,尤其是当专家数量达到 256 或 512 之后更为明显。另外,虽然推理时只需要激活部分参数,但是推理前仍然需要将全量的模型参数加载到显存中。据 Switch Transformers 的研究结果,以上特性在小规模 MoE 模型下也同样适用。在架构的选择上,MoE 适用于拥有多台机器(分布式)且要求高吞吐量的场景,在固定的预训练计算资源下,稀疏模型往往能够实现更优的效果。在显存较少且吞吐量要求不高的场景,传统的稠密模型则是更合适的选择。

Google 是 MoE 架构的早期探索者之一，OpenAI 实现了 MoE 的商业化落地。MoE 的理念起源于 1991 年的论文《Adaptive Mixture of Local Experts》。在 ChatGPT 问世之前，Google 已经有了较深入的 MoE 研究，典型代表是 20 年的 GShard 和 21 年的开源 1.6 万亿 Switch-Transformer 模型。23 年 3 月 GPT-4 问世，OpenAI 继续走了闭源路线，没有公布模型参数。但是据 SemiAnalysis 信息，GPT-4 的参数约 1.8 万亿，采用 MoE 架构，专家数为 16，每次推理调用两个专家，生成 1 个 token 约激活 2800 亿参数（GPT-3 为 1750 亿参数），消耗 560 TFLOPs 算力。在 GTC 2024 演讲上，黄仁勋展示了 GB200 训练 GPT 模型示意图，给出的参数也是 GPT-MoE-1.8T，交叉印证。

图表46: MoE 的理念起源（专家和门控网络系统）



资料来源：《Adaptive Mixture of Local Experts》，Jacobs（2021）、华泰研究

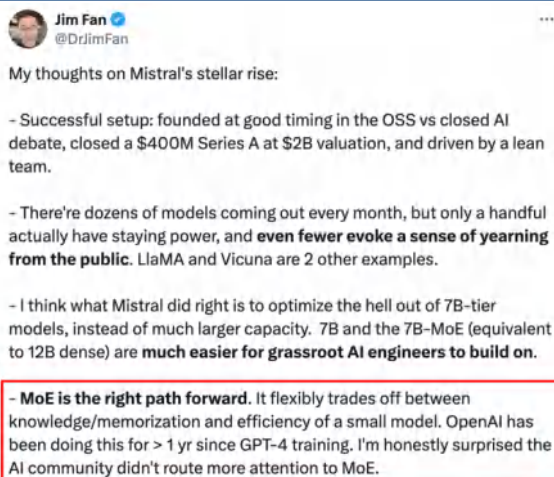
图表47: GTC 2024 对 GPT 模型参数的描述



资料来源：GTC 2024 Keynote、华泰研究

Mistral 引发 MoE 关注，Google 掀起 MoE 浪潮，国内厂商跟随发布 MoE 模型。23 年 12 月，Mistral 开源 Mixtral-8x7B-MoE，以近 47 亿的参数在多项测评基准上达到或超过 1750 亿参数的 GPT-3.5 水平，引发了全球开发者对 MoE 架构的再次关注。英伟达的研究主管 Jim Fan 指出 MoE 将成为未来模型发展的重要趋势。24 年 2 月，Google 将其最先进模型系列 Gemini 更新到 1.5 Pro，并指出架构上从稠密架构切换到 MoE 架构，实现了 1.5 Pro 模型性能的大幅提升，核心能力超过 Gemini 1.0 Ultra。国内外模型厂商随即跟进发布 MoE 相关模型，包括 xAI 开源的 Grok-1（23 年 10 月已实现 MoE，24 年开源）、MiniMax abab6、Databricks DBRX、AI21 Jamba、阿里 Qwen-1.5 MoE、昆仑万维天工 3.0、阶跃星辰 STEP 2、商汤日日新 5.0 等。

图表48: Nvidia Jim Fan 看好 MoE 对 AI 发展的重要性



资料来源：Nvidia Jim Fan、华泰研究

图表49: Gemini 1.5 Pro 核心能力超过 1.0 Ultra

Gemini 1.5 Pro	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 87.1% (27/31 benchmarks)	Win-rate: 54.8% (17/31 benchmarks)
Text	Win-rate: 100% (13/13 benchmarks)	Win-rate: 77% (10/13 benchmarks)
Vision	Win-rate: 77% (10/13 benchmarks)	Win-rate: 46% (6/13 benchmarks)
Audio	Win-rate: 60% (3/5 benchmarks)	Win-rate: 20% (1/5 benchmarks)

资料来源：Gemini 1.5 Pro 技术报告、华泰研究



图表50：国内外典型 MoE 模型比较

序号	公司名称	模型名称	MoE 参数等信息
1	OpenAI	GPT-4	1.8T, 16 个专家, 每个 token 激活 2 个专家
2	Mistral AI	Mixtral 8x7B	46.7B, 8 个专家, 每个 token 激活 2 个专家
3	Google	Gemini 1.5 Pro	官方未透露
4	xAI	Grok-1	314B, 8 个专家, 每个 token 激活 2 个专家
5	Databricks	DBRX	132B, 16 个专家, 每个 token 激活 4 个专家
6	AI21	Jamba	52B, 16 个专家, 每个 token 激活 4 个专家
7	MiniMax	abab 6.5	万亿参数
8	阿里	Qwen-1.5 MoE	14.3B, 64 个专家, 每个 token 激活 4 个专家
9	昆仑万维	天工 3.0	4000 亿参数
10	阶跃星辰	Step 2	万亿参数
11	商汤	日日新 5.0	6000 亿参数
12	腾讯	混元	万亿参数

资料来源：各公司官网、华泰研究

大模型展望：Scaling Law + AI Agent + 具身智能

展望 24 年及之后的大模型发展方向，我们认为，1) Scaling Law 虽然理论上有边界，但是实际上仍远未达到；2) 虽然有 Mamba、KAN 等新的架构挑战 Transformer，但是 Transformer 仍是主流，短期内预期不会改变；3) 以 Meta Llama 为首的开源模型阵营日益强大，占据了整个基础模型的超半数比重，且与闭源模型差距缩短；4) AI Agent 是实现 AGI 的重要加速器。5) 具身智能随着与 LLM 技术的融合，将变得更加可用。

展望#1: Scaling Law 理论上有边界，但是目前仍未到达

Scaling Law 的趋势终将会趋于平缓，但是目前公开信息看离该边界尚远。OpenAI 在 2020 年 1 月的 Scaling Law 论文中明确指出，整个研究过程中 OpenAI 在大算力、大参数和大训练数据情况下，并没有发现 Scaling Law 出现边界递减的现象。但也提到，这个趋势终将趋于平缓（level off），因为自然语言具有非零熵。但是实际上，根据斯坦福大学 2023 年的 AI Index 报告，2012-2023 年，头部模型训练消耗的算力仍然在持续增大。

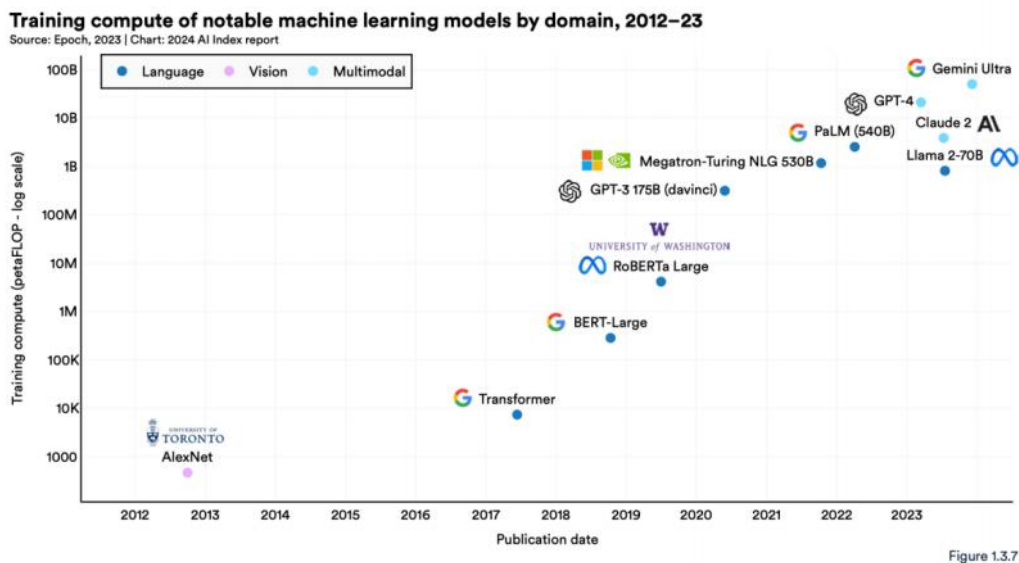
图表51：OpenAI Scaling Law 论文中对于边界的描述

We observe no signs of deviation from straight power-law trends at large values of compute, data, or model size. Our trends must eventually level off, though, since natural language has non-zero entropy.

在较大的计算、数据或模型参数下，我们没有观察到偏离直线幂律趋势的迹象。然而，我们的趋势最终将稳定下来，因为自然语言具有非零熵。

资料来源：《Scaling Laws for Neural Language Models》，OpenAI（2020）、华泰研究

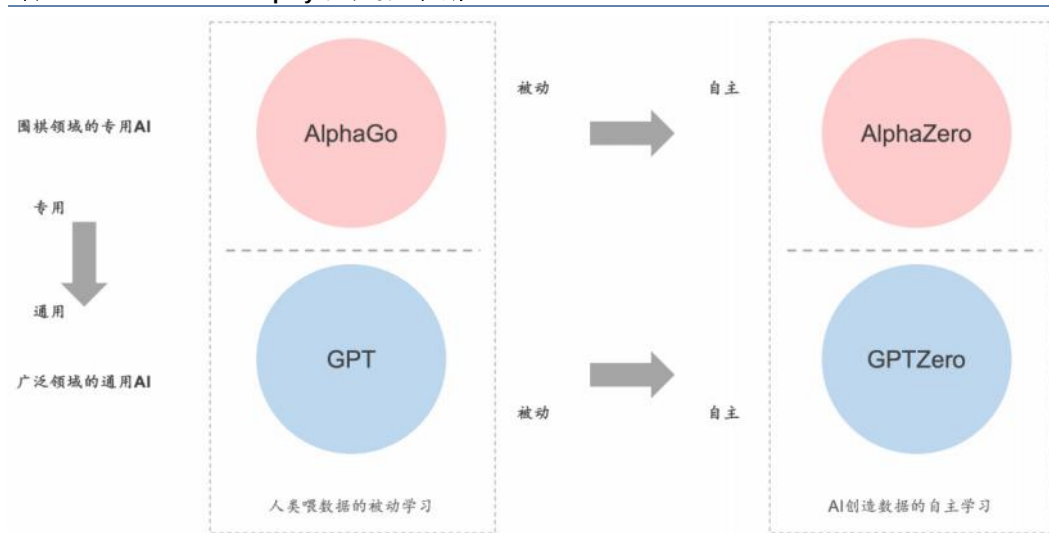
图表52：2012-2023 年头部典型大模型的训练算力消耗情况



资料来源：2023 AI Index 报告、华泰研究

可预期的时间内, **Scaling Law** 的上限尚未看到, **self-play** 是趋势。我们认为, 虽然 OpenAI 从理论上预测了 **Scaling Law** 的趋势会区域平缓, 但是目前全球头部模型厂商依然遵循更大的参数等于更高的智能。Gemini 和 Claude 3 发布的模型产品矩阵即验证了这一观点, 例如更小的 Claude 3 Haiku 输出速度快于最大的 Claude 3 Opus, 价格更低, 智能情况和测评得分也更低, 参见图表 17-图表 18。清华大学教授、智谱 AI 的技术牵头人唐杰教授在 24 年 2 月北京人工智能产业创新发展大会上发表演讲《ChatGLM: 从大模型到 AGI 的一点思考》, 也指出了目前很多大模型还在 1000 亿参数左右, “我们还远未到 **Scaling law** 的尽头, 数据量、计算量、参数量还远远不够。未来的 **Scaling law** 还有很长远的路要走。”此外, 唐杰教授还认为, “今年的阶段性成果, 是实现 GPT 到 GPT Zero 的进阶, 即大模型可以自己教自己”, 类似于 AlphaGo 到 Alphazero 的转变, 实现模型 self-play。

图表53: GPT-zero 的 self-play 能力是重要趋势



资料来源:《ChatGLM: 从大模型到 AGI 的一点思考》, 唐杰 (2024)、华泰研究

展望#2: 模型幻觉短期难消除但可抑制, CoT 是典型方法

大模型的幻觉来源包括数据、训练过程、推理过程等。LLM 的幻觉 (hallucination), 即 LLM 输出内容与现实世界的事实或用户输入不一致, 通俗说就是“一本正经胡说”。幻觉的来源主要分为 3 类: 1) 与训练数据相关的幻觉; 2) 与训练过程相关的幻觉; 3) 与推理过程相关的幻觉。

根据幻觉来源的不同, 针对性的有各种解决方法。1) 数据相关的幻觉: 可以在准备数据时, 减少错误信息和偏见, 扩展数据知识边界, 减少训练数据中的虚假相关性, 或者增强 LLM 知识回忆能力, 如使用思维链 (CoT)。2) 训练过程相关的幻觉: 可以避免有缺陷的模型架构, 例如改进模型架构或优化注意力机制; 也可以通过改进人类偏好, 减轻模型与人类对齐时的奉承性。3) 推理过程相关的幻觉: 主要是在解码过程中, 增强解码的事实性和忠诚性, 例如保证上下文和逻辑的一致等。

图表54：大模型幻觉的类别和解决方法

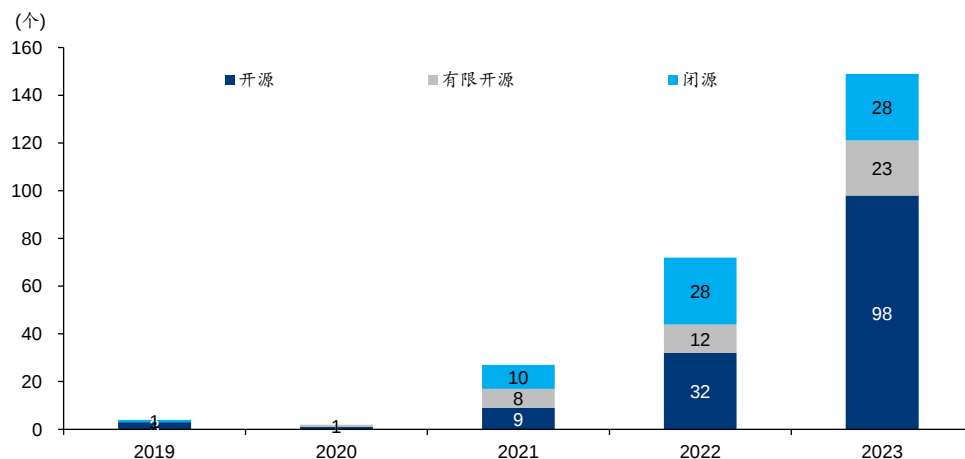
幻觉类别	解决方法	详细方法
消除模型幻觉	消除数据相关的幻觉	减少错误信息和偏见
		增强真实数据。保持训练数据的事实正确性，收集高质量的数据
		剔除偏差。1) 对于重复偏差，删除重复数据；2) 社会偏差，选择多样化、平衡的数据
		知识编辑。纳入更多知识，可通过修改模型参数或使用外部模型插件实现
	消除训练相关的幻觉	检索增强，即RAG。从外部知识检索到相关内容，增强模型输出
		减少训练数据中的虚假相关性。可通过无偏见数据集微调来实现
	消除推理相关的幻觉	减轻知识召回失败
		增强LLM知识回忆能力。例如通过思维链CoT
		改进模型底层架构或改进注意力机制。比如用双向自回归架构和注意力锐化正则化器
		优化预训练目标。避免传统的目标可能会导致模型输出中的零散表述和不一致
		减少模型的奉承性。通过改进人类的偏好判断，进而改进偏好模型来实现
		独立解码。例如有人提出事实核采样算法，平衡生成的事实内容和输出的多样性
		后期编辑解码。利用LLM的自我校正能力，来完善输出
		上下文一致性。例如上下文感知解码，通过减少对先验知识的依赖，促进模型对上下文信息关注
		逻辑一致性。例如使用对比解码，可以防止LLM错过推理步骤

资料来源：《A Survey on Hallucination in Large Language Models》，Lei（2023）、华泰研究

展望#3：开源模型将在未来技术生态中占据一席之地

2023 年开源模型在全球基础模型中所占的比重大幅提高。根据斯坦福大学 2023 年的 AI Index 报告, 2021-2023 年全球发布的基础模型数量持续增多, 且开源模型的占比大幅提高, 21-23 年占比分别为 33.3%、44.4%和 65.7%。此外, 4 月 OpenAI CEO 和 COO 在接受访谈时, 指出“开源模型无疑将在未来的技术生态中占据一席之地。有些人会倾向于使用开源模型, 有些人则更偏好于托管服务, 当然, 也会有许多人选择同时使用这两种方式。”

图表55：2023 年开源模型在基础模型中的比重大幅提升



资料来源：2023 AI Index 报告、华泰研究

Meta 持续开源 Llama 系列模型，证明了开源模型与闭源模型差距持续缩小。4 月 19 日，Llama 3-8B 和 70B 小模型发布，支持文本输入和输出，架构和 Llama 2 基本类似（Transformer decoder），上下文长度 8K，15T 训练 token（Llama 2 是 2T）。评测结果看，Llama-70B 与 Gemini 1.5 Pro 和 Claude 3 Sonnet 相比（这两个闭源模型参数都预期远大于 70B），在多语言理解、代码、小学数学等方面领先。Llama 3 继续坚持开源，可商用，但在月活超 7 亿时需向 Meta 报备。根据 Meta 官方信息，Llama 3 将开源 4000 亿参数版本，支持多模态，能力或是 GPT-4 级别。目前训练的阶段性 Llama 3-400B 已经在 MMLU 测评集（多任务语言理解能力）上得分 85 左右，GPT-4 Turbo 得分是 86.4，差距很小，且 Llama 3 400B 仍将在未来几个月的训练中持续提升能力。基于 Llama 1 和 2 带来的繁荣开源模型生态，我们认为，正式版 Llama 3 发布后，或将进一步缩小开源模型与闭源模型的差距，甚至在某些方面继续赶超。

图表56: Llama 3-70B 在多个测评集超越最新闭源模型

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

资料来源：Meta 官网、华泰研究

图表57: 开源模型社区 Hugging Face 上 Llama 2 变体超 16000 个



资料来源：Hugging Face 官网（截至 24 年 4 月）、华泰研究

图表58: 根据 UC 伯克利大学 Chatbot Arena 分数统计, 开源闭源模型间差距在缩小



资料来源：LMSYS Chatbot Arena、华泰研究

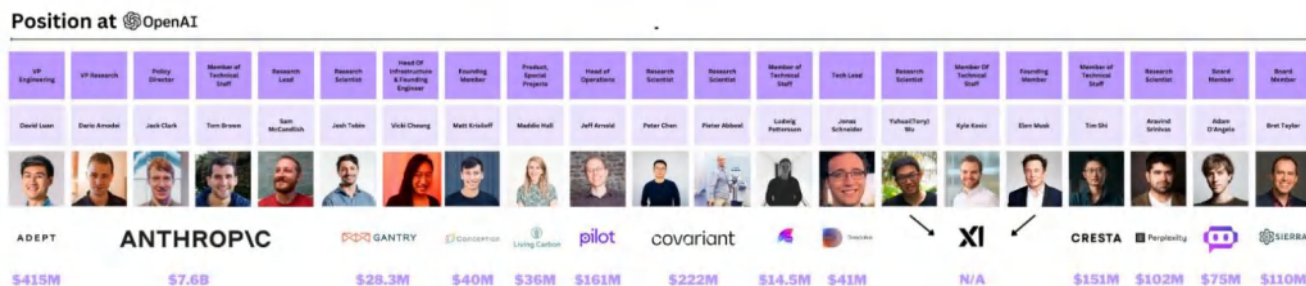
大模型的开源闭源之争尚未有定论。开源和闭源在各个领域中谁占主导，并没有定数。复盘来看，闭源在操作系统、浏览器、云基础设施、数据库等领域占据了主导地位，开源在内容管理系统、网络服务器等领域优势地位明显。反观大模型领域，开源闭源谁将最终胜出尚未有定论。当下，闭源模型的优势在于：**1) 资源集中**：大模型训练属于计算资源密集型行业，在当前各大云厂商算力储备爬坡阶段，只有闭源才能实现万卡级别的大规模分布式集群；**2) 人才集中**：OpenAI、Google、Anthropic、Mata 等大模型头部厂商，集中了目前全球为数不多的大模型训练人才，快速形成了头部效应。那我们的问题是，这种优势持续性有多长？资源方面，未来随着算力基础设施的逐步完善、单位算力成本的下降、推理占比逐步超过训练，大厂的资源密集优势是否还会显著？人才方面，全球已经看准了 LLM 的方向，相关人才也在加速培养，OpenAI 的相关人才也在快速流失和迭代，人才壁垒是否也在降低？

图表59：各个领域的开源与闭源之争



资料来源：LinkedIn、华泰研究

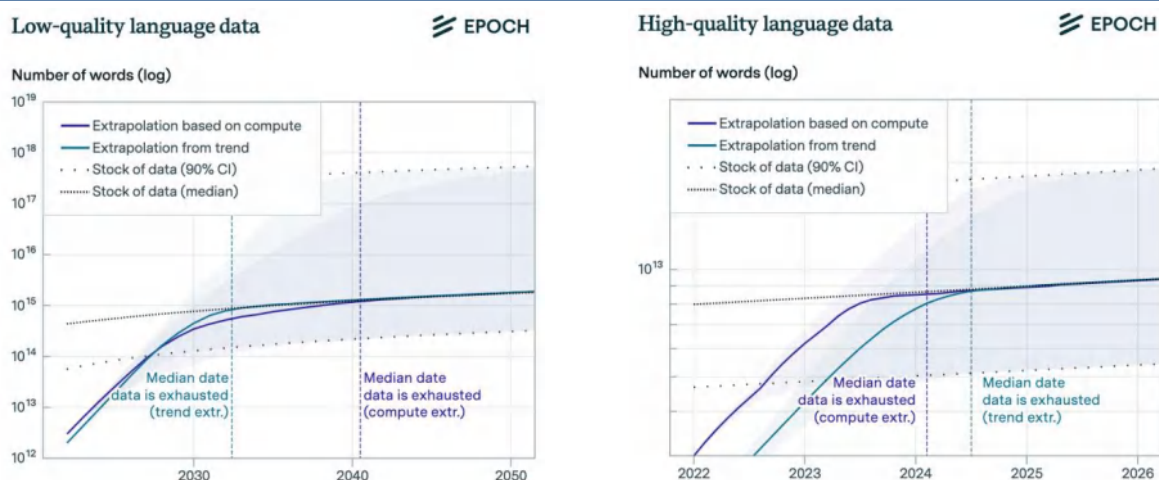
图表60：OpenAI 很多大模型人才开始离职创业



资料来源：chiefaioffice、华泰研究

展望#4：数据将成为模型规模继续扩大的瓶颈，合成数据或是关键

Epoch 预测，未来训练数据的缺乏将可能减缓机器学习模型的规模扩展。据 Epoch 预测，2030 年到 2050 年，将耗尽低质量语言数据的库存；到 2026 年，将耗尽高质量语言数据的库存；2030 年到 2060 年，将耗尽视觉数据的库存。由于大参数模型对数据量需求的增长，到 2040 年，由于缺乏训练数据，机器学习模型的扩展大约有 20% 的可能性将显著减慢。值得注意的是，以上结论的前提假设是，机器学习数据使用和生产的当前趋势将持续下去，并且数据效率不会有重大创新（这个前提未来可能被新合成技术打破）。

图表61: Epoch 预测低质量和高质量数据耗尽的时间图


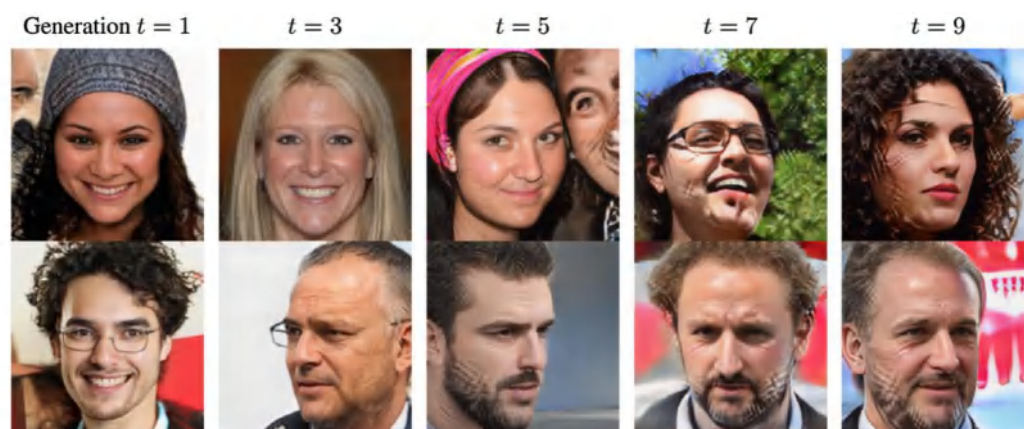
资料来源: Epoch 官网、华泰研究

合成数据是解决数据缺乏的重要途径,但目前相关技术仍需要持续改进。理论上,数据缺乏可以通过合成数据来解决,即 AI 模型自己生成训练数据,例如可以使用一个 LLM 生成的文本来训练另一个 LLM。在 Anthropic 的 Claude 3 技术报告中,已经明确提出在训练数据中使用了内部生成的数据。但是目前为止,使用合成数据来训练生成性人工智能系统的可行性和有效性仍有待研究,有结果表明合成数据上的训练模型存在局限性。例如 Alemohammad 发现在生成式图像模型中,如果在仅有合成数据或者真实人类数据不足的情况下,将出现输出图像质量的显著下降,即模型自噬障碍 (MAD)。我们认为,合成数据是解决高质量训练数据短缺的重要方向,随着技术演进,目前面临的合成数据效果边际递减问题或逐步解决。

图表62: Claude 3 在训练时使用了内部生成的数据

Claude 3 models are trained on a proprietary mix of publicly available information on the Internet as of August 2023, as well as non-public data from third parties, data provided by data labeling services and paid contractors, and data we generate internally. We employ several data cleaning and filtering methods, including deduplication and classification. The Claude 3 suite of models have not been trained on any user prompt or output data submitted to us by users or customers, including free users, Claude Pro users, and API customers.

资料来源: 2023 AI Index 报告、华泰研究

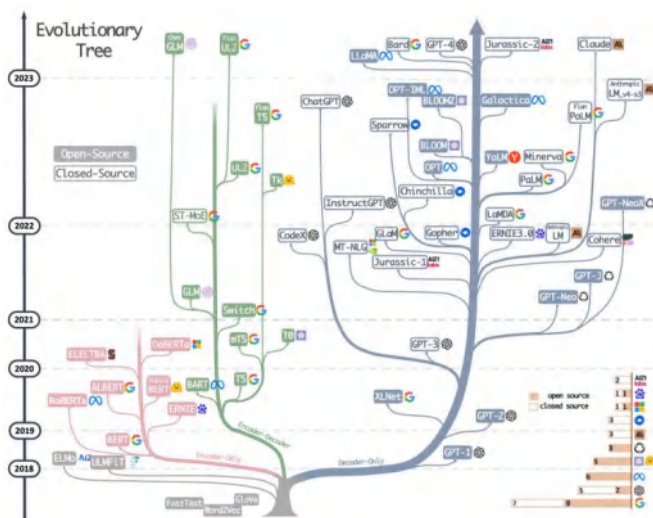
图表63: 在图像生成模型中由于合成数据引入而出现的模型输出效果递减


资料来源: 《Self-Consuming Generative Models Go MAD》, Alemohammad (2023)、华泰研究

展望#5：新的模型架构出现，但是 Transformer 仍是主流

Transformer 架构主流地位未被撼动。截止 23 年 5 月，LLM 绝大部分仍然以 Transformer 为基础架构，包括当前最先进的 GPT-4 系列、Google Gemini 系列、Meta Llama 系列，均是以 Transformer 的解码器架构为主。虽然有研究者提出了 Mamba 等基于状态空间模型（SSM）的新模型架构，实现了：1）推理时的吞吐量为 Transformer 的 5 倍；2）序列长度可以线性扩展到百万级别；3）支持多模态；4）测试集结果优于同等参数规模的 Transformer 模型。但从工程实现来看，暂时未得到大范围的使用。Google 也探索了循环神经网络的递归机制与局部注意力机制的结合；KAN 的提出也从底层替换了 Transformer 的基础单元 MLP（多层感知机）。但我们认为，以上方法都缺乏大量的工程实践和成熟的工程工具，短期内替换掉 Transformer 可能性不大。

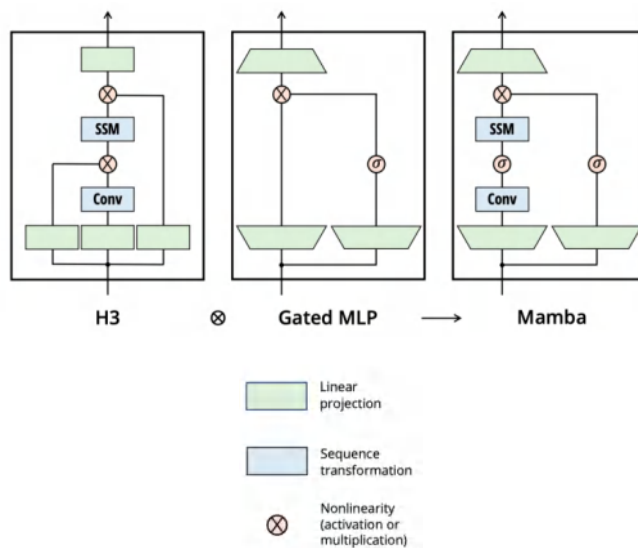
图表64：Transformer 架构是 LLM 主流架构



注：灰色以外的分支均为 Transformer 架构，粉色为编码器架构，绿色为解码器架构，蓝色为解码器架构

资料来源：《Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond》，Jingfeng Yang（2023.4）、华泰研究

图表65：Mamba 架构



资料来源：《Mamba: Linear-Time Sequence Modeling with Selective State Spaces》，Albert Gu（2023.12）、华泰研究

全球首个基于 Mamba 架构的生产级模型发布，Mamba 开始得到落地验证。24 年 3 月，AI21 发布世界首个 Mamba 的生产级模型 Jamba，融合了 Mamba+Transformer+MoE 等不同类型的技术。Jamba 基本信息如下：1）共 52B 参数，其中 12B 在推理时处于激活状态；2）共 16 位专家，推理过程中仅 4 个专家处于活跃状态；3）模型基于 Mamba，采用 SSM-Transformer 混合的架构；4）支持 256K 上下文长度；5）单个 A100 80GB 最多可支持 140K 上下文；6）与 Mixtral 8x7B 相比，长上下文的吞吐量提高了 3 倍。从测评结果看，Jamba 在推理能力上优于 Llama 2 70B、Gemma 7B 和 Mixtral 8x7B。Mamba 架构开始得到验证。

图表66: Jamba 模型的测评结果

	推理			综合评估				GSM8K (CoT)
	HellaSwag	Arc Challenge	WinoGrande	PIQA	MMLU	BBH	TruthfulQA	
Lama2 13B	80.7%	59.4%	72.8%	80.5%	54.8%	39.4%	37.4%	34.7%
Lama2 70B	85.3%	67.3%	80.2%	82.8%	69.8%	51.2%	44.9%	55.3%
Gemma 7B	81.2%	53.2%	72.3%	81.2%	64.3%	55.1%	44.8%	44.5%
Mixtral 8x7B	86.7%	66.0%	81.2%	83.0%	70.6%	50.3%	46.8%	60.4%
Jamba	87.1%	64.4%	82.5%	83.2%	67.4%	45.4%	46.4%	59.9%

资料来源: AI21 官网、华泰研究

图表67: Jamba 架构 (左) 吸收了 Mamba+Transformer+MoE 多种技术

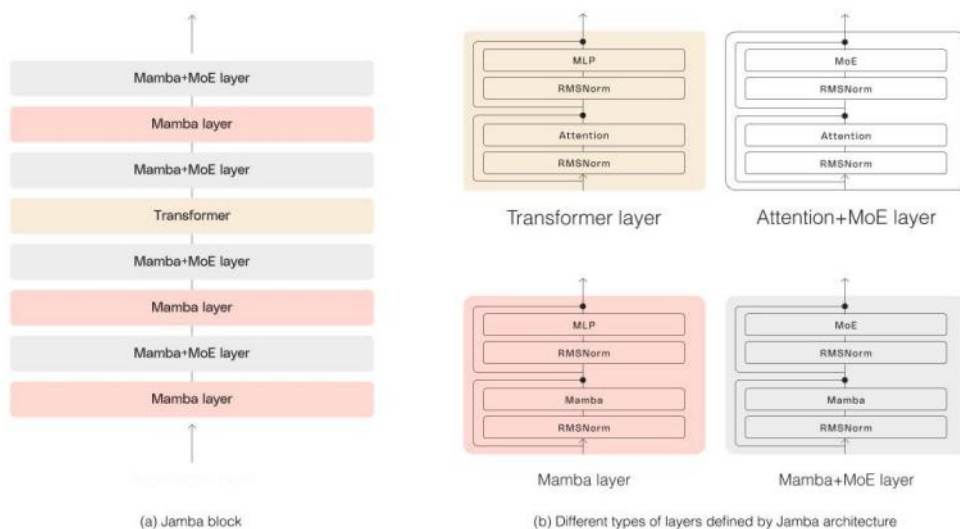
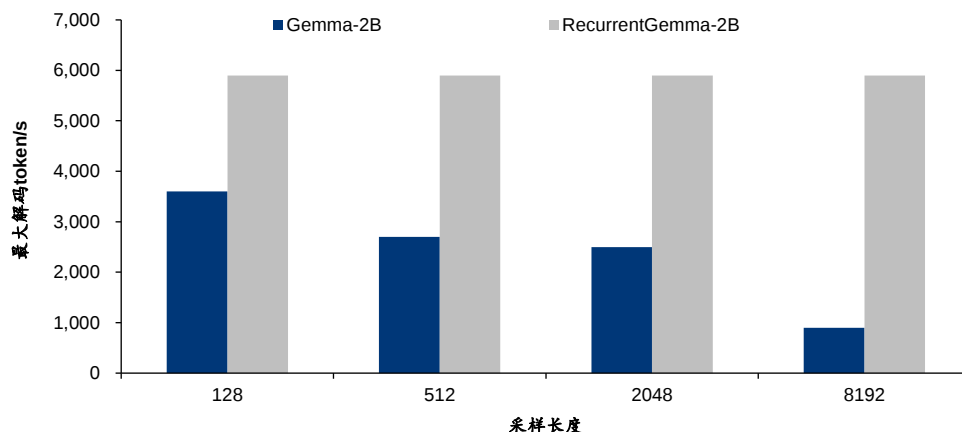


Diagram showing (a) a single Jamba block, (b) Different types of layers. Jamba implementation includes 4 Jamba blocks, each containing 8 layers, a 1/7 ratio of attention/Mamba layers, and MoE applied every 2 layers

资料来源: AI21 官网、华泰研究

Google RecurrentGemma 架构也与 Transformer 不同，是另一种新的路线探索。 RecurrentGemma 基于 Google 开源的小模型 Gemma，在此基础上，引入了循环神经网络 (RNN) 和局部注意力机制来提高记忆效率。由于传统的 Transformer 架构中，需要计算两两 token 之间的注意力机制，因此时间和空间复杂度均随着 token 的增加而平方级增长。由于 RNN 引入的线性递归机制避免了平方级复杂度，RecurrentGemma 带来了以下几个优势：1) 内存使用减少：在内存有限的设备（例如单个 XPU）上生成更长的样本。2) 更高的吞吐量：由于内存使用量减少，RecurrentGemma 可以以显著更高的 batch 大小执行推理，从而每秒生成更多的 token（尤其是在生成长序列时）。更重要的是，RecurrentGemma 展示了一种实现高性能的非 Transformer 模型，是架构革新的重要探索。

图表68: RecurrentGemma 的吞吐量远高于 Gemma



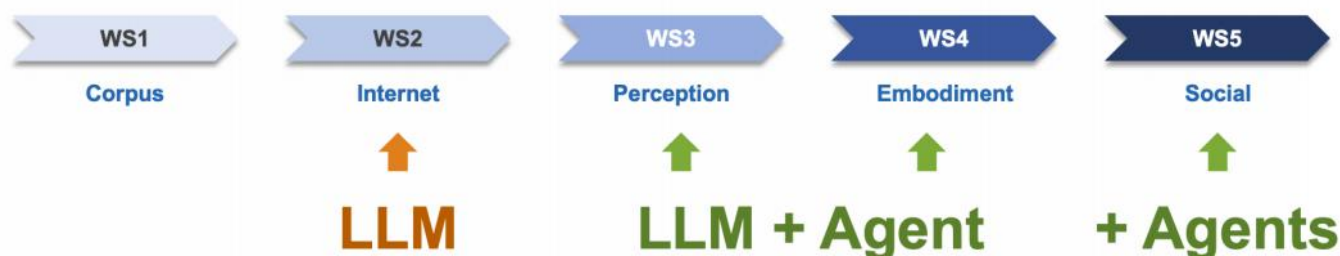
资料来源: Google 官网、华泰研究

展望#6: AI Agent 智能体是 AGI 的加速器

计算机科学中 **Agent** 指计算机能够理解用户的意愿并能自主地代表用户执行任务。Agent（中文翻译智能体、代理等）概念起源于哲学，描述了一种拥有欲望、信念、意图和采取行动能力的实体。将这个概念迁移到计算机科学中，即意指计算机能够理解用户的意愿并能自主地代表用户执行任务。随着 AI 的发展，AI Agent 用来描述表现出智能行为并具有自主性、反应性、主动性和社交能力的人工实体，能够使用传感器感知周围环境、做出决策，然后使用执行器采取行动。

AI Agent 是实现人工通用智能 (AGI) 的关键一步，包含了广泛的智能活动潜力。2020 年，Yonatan Bisk 在《Experience Grounds Language》中提出 World Scope (WS)，来描述自然语言处理到 AGI 的研究进展，包括 5 个层级：WS1. Corpus (our past); WS2. Internet (most of current NLP); WS3. Perception (multimodal NLP); WS4. Embodiment; WS5. Social。据复旦大学 NLP 团队，纯 LLM 建立在第二个层次上，即具有互联网规模的文本输入和输出。将 LLM 与 Agent 技术架构结合，并配备扩展的感知空间和行动空间，就有可能达到 WS 的第三和第四层。多个 Agent 可以通过合作或竞争来处理更复杂的任务，甚至观察到涌现的社会现象，潜在地达到第五 WS 级别。

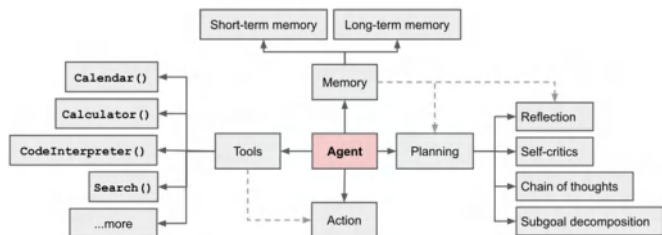
图表69: Agent 能将 LLM 抬升到更接近 AGI 的层级



资料来源: 《The Rise and Potential of Large Language Model Based Agents: A Survey》, Zhiheng Xi (2023.9)、华泰研究

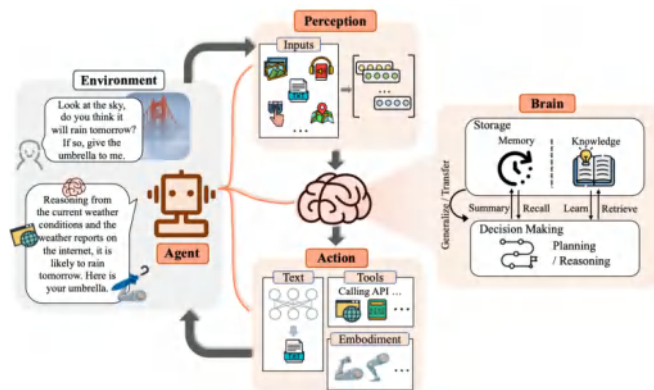
AI Agent 主要由 LLM 大脑、规划单元、记忆单元、工具和行动单元组成。不同研究中的 AI Agent 框架组成略有差别。比较官方的定义是 OpenAI 安全系统负责人 Lilian 提出的，她将 Agent 定义为 LLM、记忆（Memory）、任务规划（Planning Skills）以及工具使用（Tool Use）的集合，其中 LLM 是核心大脑，Memory、Planning Skills 以及 Tool Use 等则是 Agents 系统实现的三个关键组件。此外，复旦大学 NLP 团队也提出了由大脑、感知和动作三部分组成的 AI Agent 框架。

图表70：OpenAI Lillian 提出的 Agent 框架



资料来源：OpenAI 官网、华泰研究

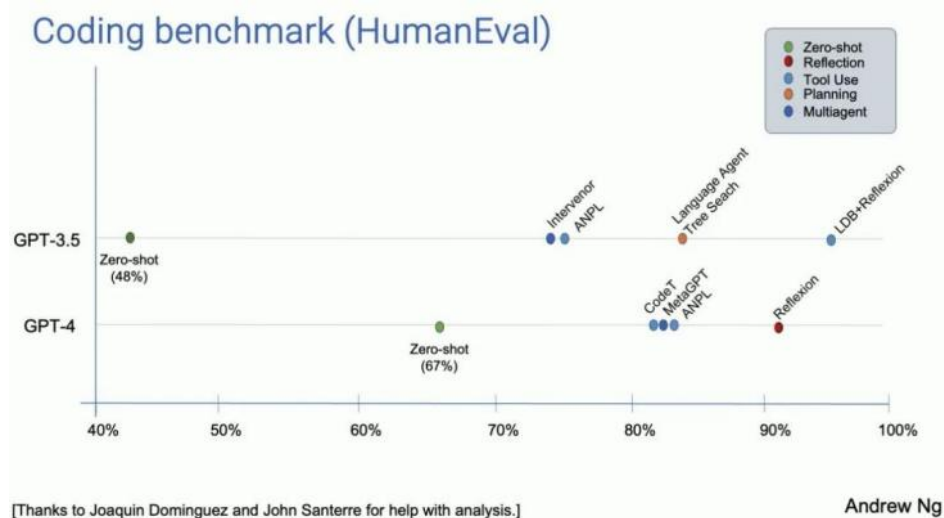
图表71：复旦大学 NLP 团队提出的 Agent 框架



资料来源：《The Rise and Potential of Large Language Model Based Agents: A Survey》，Zhiheng Xi (2023.9)、华泰研究

吴恩达教授指出，LLM 加上反思、工具使用、规划、多智能体等能力后，表现大幅提升。斯坦福大学教授、Amazon 董事会成员吴恩达在红杉美国 AI Ascent 2024 提出，如果用户围绕 GPT-3.5 使用一个 Agent 工作流程，其实际表现甚至好于 GPT-4。其中，反思指的是让模型重新思考其生成的答案是否正确，往往会带来输出结果的改进；工具使用包括调用外部的联网搜索、日历、云存储、代码解释器等工具，补充模型的能力欠缺；多智能体协作指的是多种智能体互相搭配来完成一个复杂任务，每种智能体会负责自己所擅长的一个领域，类似人类社会之间的协作，实现超越单个智能体能达到的效果。

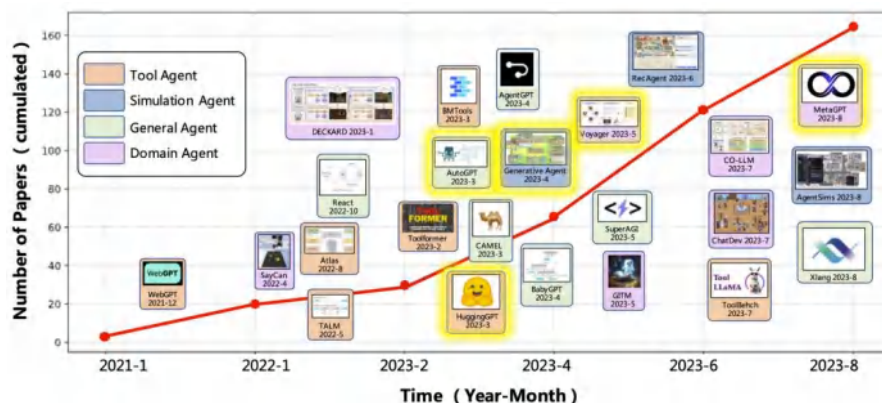
图表72：GPT-3.5 和 4 在反思+使用工具+规划+多智能体时能大幅提高模型模型表现



资料来源：吴恩达演讲、华泰研究

Agent 相关研究处于爆发期。伴随 LLM 的快速迭代发展，基于 LLM 的 AI Agent 涌现，典型的如 Auto-GPT、微软的 HuggingGPT、斯坦福小镇 Generative Agent、Nvidia Voyager 等。24 年 3 月，AI 初创公司 Cognition 发布第一个 AI 软件工程师自主智能体 Devin，能够使用自己的 shell、代码编辑器和 Web 浏览器来解决工程任务，并在 SWE-Bench 基准测试上正确解决了 13.86% 的问题，远超之前方法的正确率。我们认为，2024 年基于 AI Agent 的应用和产品仍将会继续涌现，其效果也将持续受益于大模型能力的提升，AI Agent 将成为实现 AGI 的重要助推器。

图表73：2023 年伴随 LLM 的发展 Agent 项目也处于爆发期



资料来源：《A Survey on Large Language Model based Autonomous Agents》，Wang Lei (2023.8)、华泰研究

展望#7：具身智能与 LLM 结合落地加速

AI 龙头公司在具身智能领域有模型、框架层面的丰富研究成果。23 年 5 月，Nvidia CEO 黄仁勋指出，AI 的下一个浪潮将是具身智能。各个 AI 头部厂商均有相关的研究成果。23 年年初，微软的 ChatGPT for Robotics 初次探讨了 LLM 代替人工编程，来对机器人实现控制。Google 延续了 2022 年的具身智能成果，将 RT 系列模型升级到视觉动作语言模型 RT-2，将 Gato 升级到能自我迭代的 RoboCat，并开源了迄今最大的真实机器人具身智能数据集 Open X-Embodiment。Nvidia 也有 VIMA 和 OPTIMUS 等具身智能研究，并在 24 年 2 月成立了专门研究具身智能的小组 GEAR。斯坦福李飞飞教授的 VoxPoser 结合视觉模型和语言模型优势，建模了空间 Value Map 来对机器人轨迹进行规划。Meta 也发布 RoboAgent，并在训练数据集收集上利用了自家的 CV 大模型 SAM。

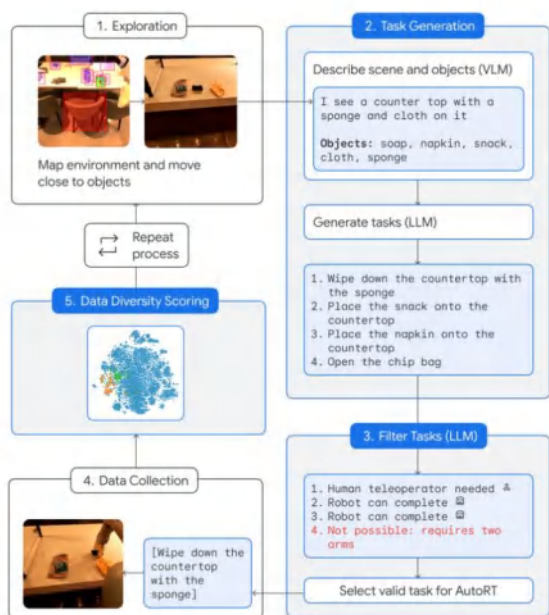
图74：2023 年 AI 龙头公司为具身智能贡献了众多成果



资料来源：各公司官网、华泰研究

2024 年，具身智能仍是 LLM 重要的终端落地场景，技术仍在持续迭代。1) 24 年 1 月，斯坦福大学发布 Mobile ALOHA 机器人，利用模仿学习，在人类做出 50 个示例后，机器人即能自行执行下游任务。2) 同月，Google 一次性发布了三项具身智能成果。Auto-RT 解决机器人数据来源问题，通过 LLM 和 VLM（视觉语言模型）扩展数据收集；SARA-RT 显著加快了 Robot Transformers 的推理速度；RT-Trajectory 将视频转换为机器人轨迹，为机器人泛化引入了以运动为中心的目标。3) AI 机器人公司 Figure 推出了 Figure 01，采用端到端 AI 神经网络，仅通过观察人类煮咖啡即可在 10 小时内完成训练。4) 从目前 Tesla Optimus 发布视频情况看，Optimus 的神经网络已经能够指导机器人进行物品分拣等动作，且控制能力进一步提高。

图75：Google AutoRT 流程图



资料来源：Google 官网、华泰研究

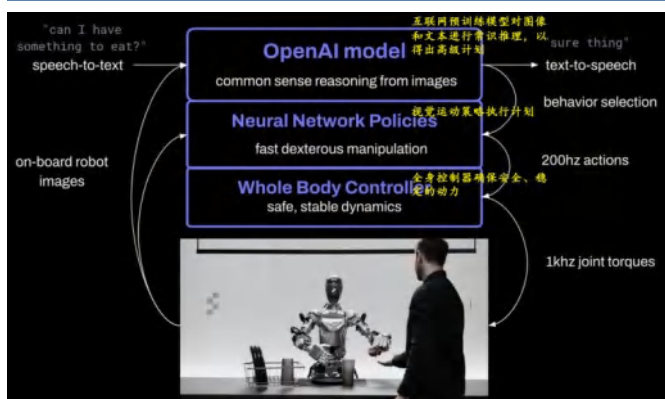
图76：Figure 公司的 Figure 01



资料来源：Figure 官网、华泰研究

OpenAI 与 Figure AI 率先合作，实现了大模型对具身智能的赋能。24 年 3 月，OpenAI 官方宣布与 Figure AI 机器人公司合作，将多模态模型扩展到机器人感知、推理和交互。宣布合作 13 天后，Figure 01 已经与 OpenAI 的视觉语言模型结合，并发布了演示视频。ChatGPT 从顶层负责用户交互、环境感知（依靠 vision 视觉能力）、复杂问题拆解，而 Figure 01 自身的神经网络和控制系统负责底层的自主任务执行，实现了强交互的自主任务执行。随后，国内大模型厂商百度与机器人整机厂商优必选也宣布合作，“复刻”了 OpenAI+Figure 的合作路线，由文心大模型负责交互推理、优必选 Walker X 负责底层任务实现。我们认为，多模态大模型和机器人结合的路线已经走通，随着 24 年模型能力持续迭代（GPT-4o 的出现），以及人形机器人自主和控制能力的加强，LLM+具身智能落地加速，并将更加可用、好用。

图表77：OpenAI 与 Figure 合作的技术架构



资料来源：Figure AI 官网、华泰研究

图表78：百度+优必选复刻 OpenAI+Figure 路线



资料来源：优必选官网、华泰研究

GPT-5 的几个预期

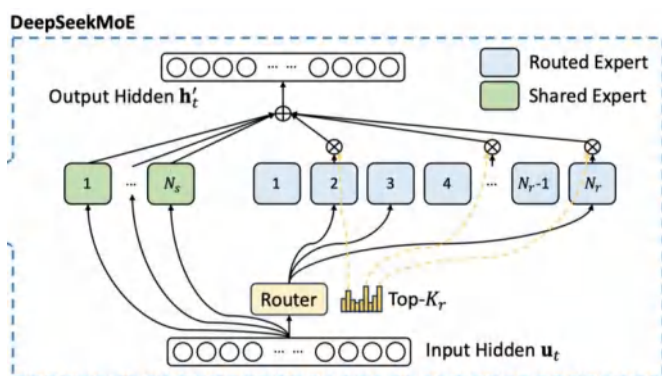
OpenAI 从 GPT-3 开始实行闭源商业化路线，相关的模型技术几乎不再公布细节。我们基于对全球大模型发展趋势的研究和把握，提出几个 GPT-5 可能的预期和展望，并给出相应的推测逻辑。

预期#1: MoE 架构将延续，专家参数和数量或变大

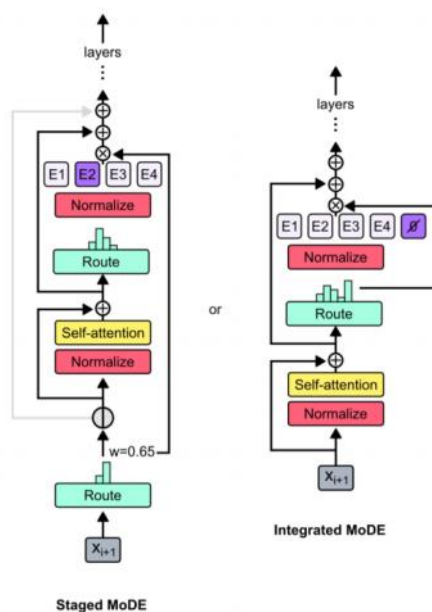
MoE 是现阶段实现模型性能、推理成本、模型参数三者优化的最佳架构方案。1) MoE 将各种专家通过路由 (router) 机制有机整合，在各种下游任务上，能够充分利用每个专家的专业能力，提高模型表现；2) MoE 天然的稀疏架构，使得 MoE 模型与同参数稠密模型在推理成本上有较大节省；3) 同理，在推理成本固定的情况下，MoE 模型相比稠密模型，能够把模型参数堆到更大，同样能够提升模型性能。

我们认为, OpenAI 在 GPT-5 模型迭代时仍将采用 MoE 架构, 或有部分改进。相比 GPT-4, GPT-5 的 MoE 架构或将有以下改进: 1) 每个专家的参数更大, 例如每个专家大小与 GPT-4 相同, 近 2T 参数。即使 OpenAI 无法将单个 2T 参数专家做成稠密架构, 也可以使用 MoE 嵌套 MoE 的方式实现。2) 专家数量变多, 例如幻方旗下 DeepSeek V2 模型即使用改进的 DeepSeekMoE 架构, 采取了更细粒度的专家结构, 将专家数扩展到 160+, 以适应更加丰富和专业的下游任务。3) MoE 架构本身可能有改进, 例如 Google DeepMind 提出了 Mixture of Depth (MoD) 架构, 向 Transformer 的不同层 (layer) 引入类似 MoE 的路由机制, 对 token 进行选择性地处理, 以减少推理成本。MoD 可以和 MoE 混合使用, 相当于对 MoE 进行了改进。OpenAI 或也会有类似的改进技术。

图表79: DeepSeek v2 采取的细粒度专家 MoE 架构



图表80: Google DeepMind 混合了 MoD 和 MoE 架构



资料来源:《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》, DeepSeek AI (2024)、华泰研究

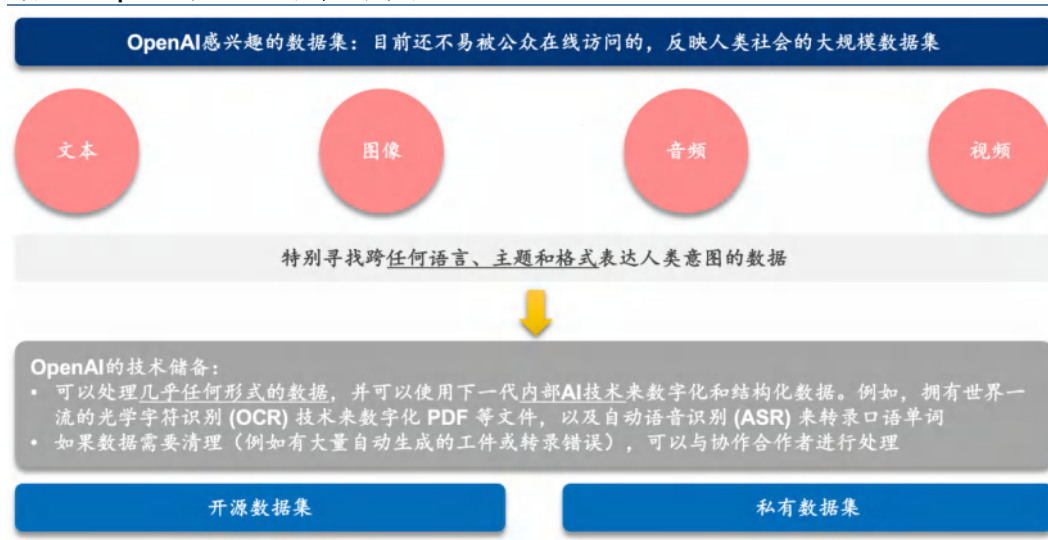
资料来源:《Mixture-of-Depths: Dynamically allocating compute in transformer-based language models》, Google DeepMind (2024)、华泰研究

我们预计, GPT-5 或将达到百万亿参数, 即 2T 参数/专家*64 专家 (或更多)。

预期#2: GPT-5 及之后模型的训练数据集质量更高、规模更大

OpenAI 不断加速与私有高质量数据公司的合作进度，为训练大模型做数据储备。2023 年 11 月，OpenAI 即官宣推出数据合作伙伴计划，将与各类组织合作生成用于训练 AI 模型的公共和私有数据集，包括冰岛政府、非营利法律组织“Free Law Project”等。2024 年，OpenAI 在 4-5 月先后与英国金融时报、程序员交流网站 Stack Overflow、论坛网站 Reddit 宣布合作，相关数据覆盖了新闻、代码、论坛交流等场景。我们认为，OpenAI 在早期的数据储备中，已经将网络公开可获得的数据进行了充分的开发，根据 OpenAI 的 Scaling Law 和 Google Chinchilla 的结论，随着模型参数的增大，想要充分训练模型，必须增大训练数据规模，这也从 OpenAI 的广泛数据合作关系中得到印证。我们认为，GPT-5 及之后模型的训练数据集，将有望吸纳更多高质量的私域数据，数据规模也将变得更大。

图表81: OpenAI 推出的数据合作伙伴计划



资料来源：OpenAI 官网、华泰研究

图表82: OpenAI 加快与私有数据公司的合作



资料来源：OpenAI 官网、华泰研究

图表83: Chinchilla 给出的最优训练条件：数据集≈20 倍参数

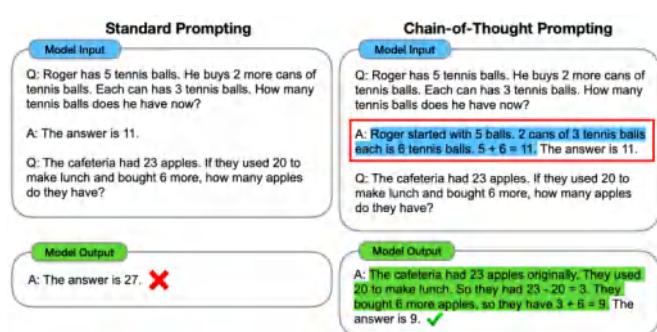
Parameters	FLOPs	FLOPs (in Gopher unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

资料来源：Training Compute-Optimal Large Language Models, Google DeepMind (2022)、华泰研究

预期#3: 在思维链 CoT 的基础上，再加一层 AI 监督

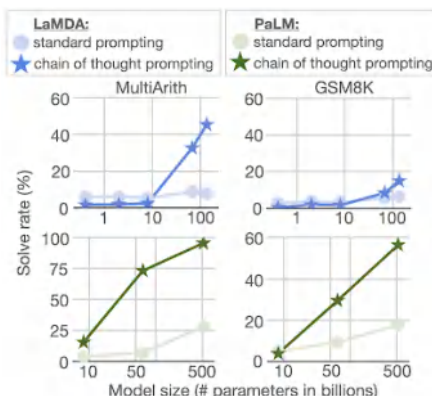
思维链能够在不改变模型的情况下提高其表现性能。2022 年，Jason Wei 在《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》中首次提出思维链 (chain of thought, CoT) 概念，使模型能够将多步骤问题分解为中间步骤。通过思维链提示，足够规模 (~100B 参数) 的语言模型可以解决标准提示方法无法解决的复杂推理问题，提高各种推理任务的表现。以算数推理 MultiArith 和 GSM8K 为例，当使用思维链提示时，增加 LaMDA 和 PaLM 模型参数可以显著提高性能，且性能大大优于标准提示。此外，思维链对于模型的常识推理任务（如 CommonsenseQA、StrategyQA 和 Date Understanding 等）同样有明显的性能提升作用。

图84：标准提示与思维链提示的示例



资料来源：Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Jason (2022)、华泰研究

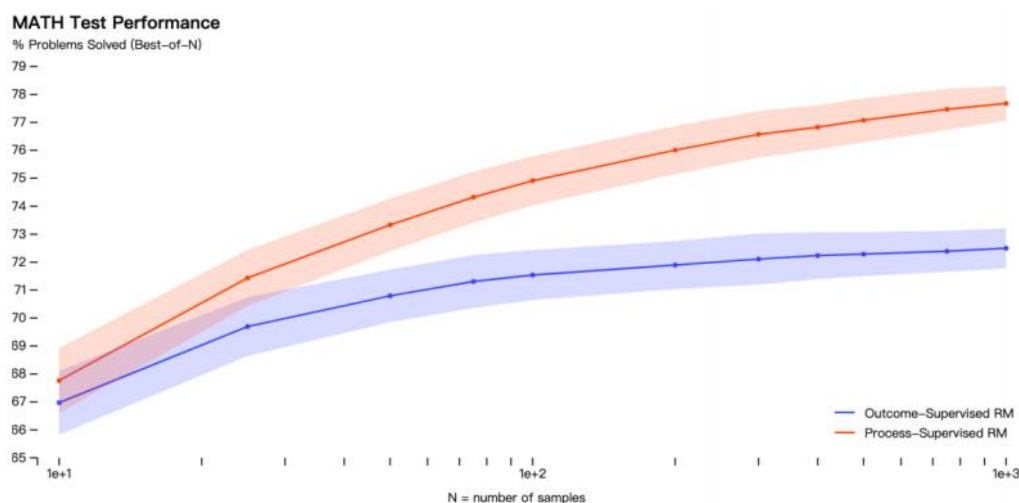
图85：算数推理中思维链大大提高模型性能表现



资料来源：Google Research 官网、华泰研究

OpenAI 探索了过程监督对模型的性能提升，有望与 CoT 结合，进一步提高推理能力。23 年 5 月，OpenAI 官方 blog 宣布训练了一个奖励模型，通过奖励推理的每个正确步骤（“过程监督”），而不是简单地奖励正确的最终答案（“结果监督”），来更好的解决模型的数学推理能力和问题解决能力。与结果监督相比，过程监督有优势：1）过程监督相当于直接奖励了模型遵循对齐的 CoT，流程中的每个步骤都受到精确的监督；2）过程监督更有可能产生可解释的推理，因为它鼓励模型遵循人类思考的过程。最终的 MATH 测试集结果中，过程监督能够提升相对于结果监督 5pct 以上的正确率。我们认为，这种基于 CoT 的过程监督方法，有可能帮助 GPT-5 进一步提高模型推理的正确性，压制模型幻觉。

图86：MATH 测评下过程监督结果好于结果监督

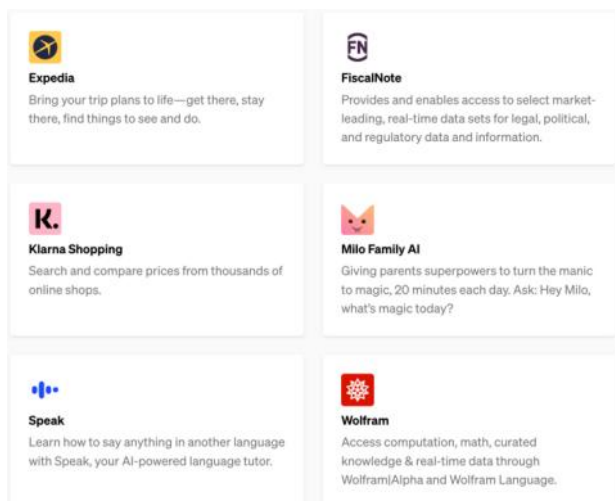


资料来源：OpenAI 官网、华泰研究

预期#4：支持更多外部工具调用的端到端模型

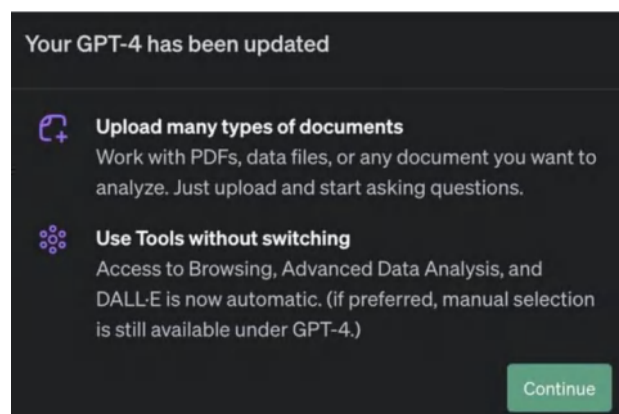
GPT-5 有望在 GPT-4 少量的外部工具基础上，增加更多的可调用工具，扩展能力边界。目前基于 GPT-4 系列的 ChatGPT，能够调用 Bing 搜索、高级数据分析（源代码解释器）、DALL-E 文生图等外部工具，并且在 23 年 11 月推出 All Tools 能力，让 ChatGPT 在与用户对话时自主选择以上三种工具。外部工具调用使得模型在性能基本保持不变的情况下，能力边界得到扩展，其实质与 Agent 调用工具类似。此外，曾在 23 年 3 月推出的 ChatGPT Plugins 功能，本质也是外部工具，但是由于 GPT-4 能力的有限，导致能够在单个对话中使用的 Plugins 只有三个，因此 Plugins 逐渐被 GPTs 智能体取代。我们认为，随着 GPT-5 推理能力的进一步提高，将有能力更好的自主分析用户需求，以更合理的方式，调用更多的外部工具（100-200 个），如计算器、云存储等，从而进一步扩展 GPT-5 的模型能力边界。

图表87：最早推出的部分 Plugins



资料来源：OpenAI 官网、华泰研究

图表88：All Tools 功能整合了可调用的工具



资料来源：OpenAI 官网、华泰研究

GPT-4o 已经打下多模态端到端的基础，GPT-5 将延续。我们认为，GPT-4o 验证了头部厂商大模型原生多模态的发展趋势，这一趋势不会轻易改变，因为端到端的原生多模态，很好的解决了模型延时（如 GPT-4 非端到端语音响应平均时间超 5s，而 4o 端到端语音响应时间平均仅 320ms）、模型误差（由于误差是不可避免的，级联的模型越多，误差累计越大，端到端仅 1 份误差）等问题，因此 GPT-5 将延续端到端多模态结构，或将有部分改进。如进一步降低端到端的响应延迟，优化用户使用体验；加入更多的模态支持，如深度、惯性测量单位（IMU）、热红外辐射等信息，以支持更复杂的如具身智能等场景。

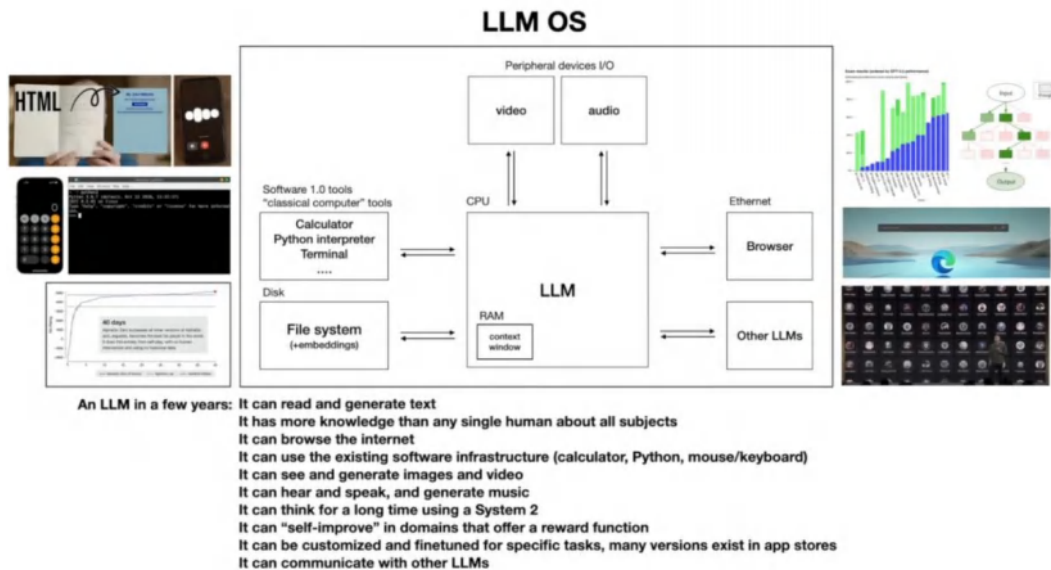
预期#5：多种大小不同的参数，不排除推出端侧小模型

Google 和 Anthropic 均在同代模型中推出参数大小不同的版本，GPT-5 有望跟进。如图表 17-18 所示，Google 和 Anthropic 均采取了同代模型、不同大小的产品发布策略，以平衡用户的成本和性能体验。据海外开发者 Tibor Blaho 信息，ChatGPT 安卓版安装包 1.2024.122 版本中发现了三个新的模型名称：gpt-4l, gpt-4l-auto, gpt-4-auto，其中 l 代表“lite”（轻量），或是 OpenAI 开始考虑布局大小不同的模型矩阵。由于 Google 官方已经实现了最小参数的 Gemini Nano 模型在 Pixel 8 Pro 和三星 Galaxy S24 系列实装，且据 Bloomberg 信息，OpenAI 与 Apple 正在探索端侧模型上的合作，我们预测，GPT-5 也有可能推出端侧的小参数模型版本。

预期#6：从普通操作系统到 LLM 操作系统

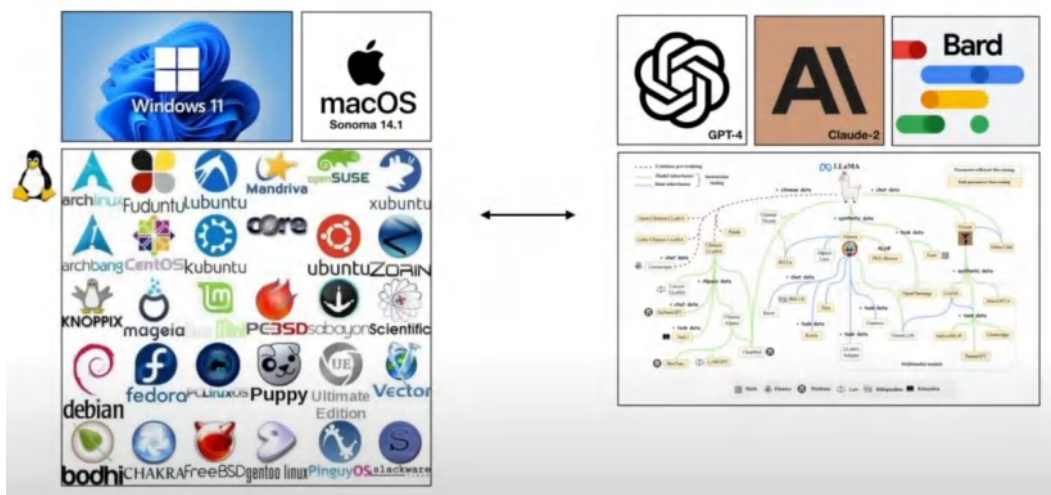
LLM 操作系统是 Agent 在系统层面的具象化。LLM OS 是前 OpenAI 科学家 Andrej Karpathy 提出的设想，其中 LLM 将替代 CPU 作为操作系统核心，LLM 的上下文窗口是 RAM，接受用户指令并输出控制指令，在 LLM 核心外部有存储、工具、网络等各种“外设”供 LLM 调用。我们认为，从结构上看，LLM OS 和图表 67 所示的 Agent 架构十分相似，可以看做 Agent 在操作系统领域的具象化。LLM OS 的核心就是模型能力，随着 GPT-5 推理性能的不断提升，我们认为 LLM 和 OS 结合的范式将更有可能实现，届时人类和 OS 的交互方式将不再以键鼠操作为主，而会转向基于 LLM 的自然语言或语音操作，进一步解放人类双手，实现交互方式的升级。

图表89：前 OpenAI 科学家 Andrej Karpathy 提出的 LLM OS 构想



资料来源：OpenAI 官网、华泰研究

图表90：操作系统与 LLM 的相似性



资料来源：各公司官网、华泰研究

预期#7：端侧 AI Agent 将更加实用和智能

OpenAI 和 Google 已经将模型的重点使用场景定位到端侧 AI Agent。24 年 5 月 13-14 日，OpenAI 和 Google 分别召开发布会和开发者大会，其中最值得关注和最亮的部分就是端侧 AI Agent。OpenAI 基于最新的端到端 GPT-4o 模型打造了新的 Voice Mode，实现了更拟人、更个性化、可打断、可实时交互的 AI 助手，并能够使用 4o 的视觉能力，让助手针对用户看到的周围环境和 PC 场景进行推理；Google 的 Project Astra 也实现了类似的效果，并且能够根据模型“看到”的场景进行 recall。我们认为，头部模型厂商遵循了模型边迭代、应用边解锁的发展路径，目前已经将模型的使用场景聚焦到了端侧。结合 OpenAI 与 Apple 的合作进展看，端侧 AI 或将在 24 年下半年成为重点。

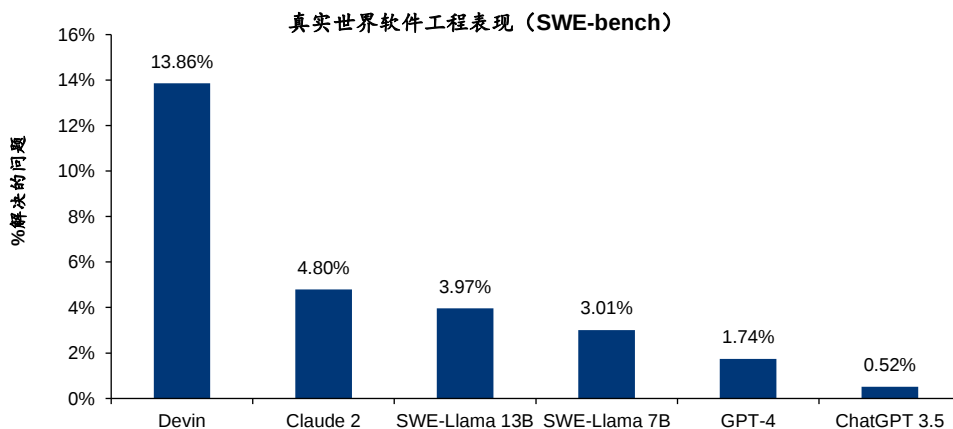
图表91: OpenAI 新 Voice Mode 与 Google Project Astra 对比

	产品名	定位	端侧支持	支持模态	是否支持实时打断	系统层面支持	是否支持记忆	推出时间	和 Apple 合作可能性
OpenAI	New Voice Mode	—	手机和 PC	文本、语音、是 图像和视频	是	OpenAI 本身没有 OS, 目前 LLM OS 也未成熟, 智能依附于 iOS 或 Android, 能获得多少权限, 得它们说的算	未展示	几周或几个月后推出	目前公开信息看, 合作可能性较大
Google	Project Astra	AI 通用智能体	手机和 PC	文本、语音、是 图像和视频	未展示, 推测不行	深度控制 Android 系统, 有系统权限	支持, 但不清楚能够保留记忆的时长	其中一些功能将于今年晚些时候出现在 Google 产品中	虽然 Apple 最先和 Google Gemini 谈合作, 但后续进展不如 OpenAI 快

资料来源: 各公司官网、华泰研究

更加智能的 GPT-5 能够将 AI Agent 能力推上新的台阶。我们认为, OpenAI 在第四代 GPT 的大版本下, 已经通过端到端的 4o 实现了 AI Agent 更实时、更智能的多模态交互。但是基于目前模型的推理性能, AI Agent 在实现多任务、多步骤的自主任务执行时成功率仍不够高。以 PC 端基于 GPT-4 的 AI 软件工程师智能体 Devin 为例, 在 SWE-Bench 基准测试(要求 AI 解决 GitHub 上现实世界开源项目问题)上进行评估时, Devin 在没有人类协助的情况下能正确解决 13.86% 的问题, 远远超过了之前最好方法对应的 1.96% 正确率, 即使给出了要编辑的确切文件, Claude 2 也只能成功解决 4.80% 的问题。但是 13.86% 的成功率, 仍然距离实用较远, 究其原因还是模型的智能能力“不够”。我们认为, 随着 GPT-5 核心推理能力进一步提高, 或能将“类 Devin”产品正确率提升到 80% 以上, AI Agent 将变得更加实用和智能。

图表92: Devin 解决真实工程问题正确率显著提高但仍离实用较远

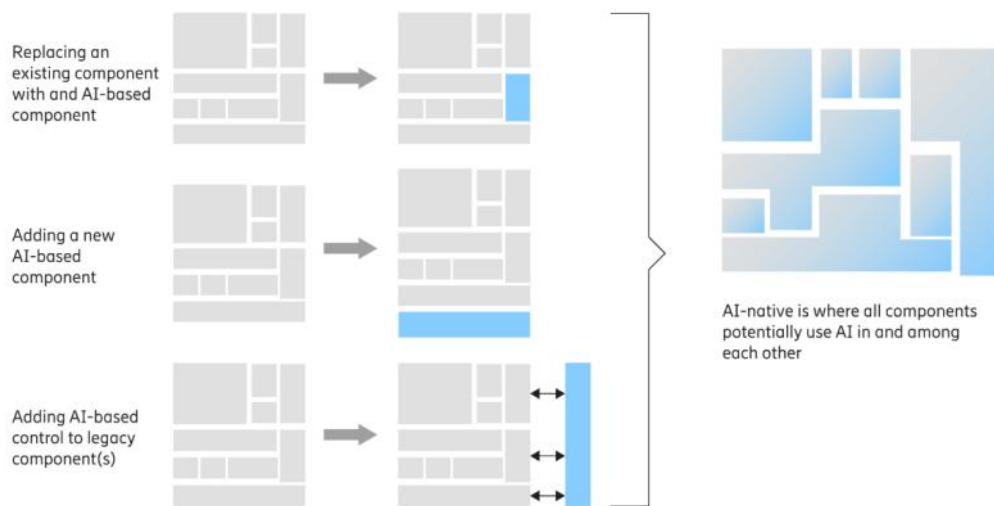


资料来源: Cognition 官网、华泰研究

理想 vs 现实：从 AI+ 到 +AI

据 Ericsson 白皮书《Defining AI native》，AI 与系统可以分为非原生和原生两类。对于非 AI 原生（None AI-native）系统，又可根据 AI 组件的部署方式细分为：1）替换已有部件。即在现有的系统组件中，将其中的一部分用基于 AI 的组件进行替换或增强。2）增加新的部件。即不改变现有系统中组件的情况下，增加一部分基于 AI 的组件。3）增加 AI 控制。同样不改变现有系统的组件，增加基于 AI 的控制组件部分，来对已有组件进行控制，在传统功能之上提供自动化、优化和额外功能。对于 AI 原生（AI-native）系统，系统中所有的组件均基于 AI 能力构建，整个 AI 原生系统拥有内在的、值得信赖的 AI 功能，AI 是设计、部署、操作和维护等功能的自然组成部分。

图表93：Devin 解决真实工程问题正确率显著提高但仍离实用较远

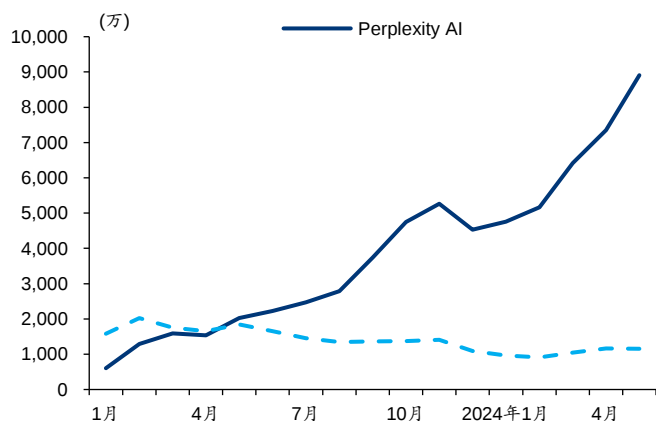


资料来源：Ericsson 官网、华泰研究

AI+指的是 AI 原生形式，是理想的 AI 应用和硬件构建方法，但是目前的大模型能力还无法很好的支持这一实现。

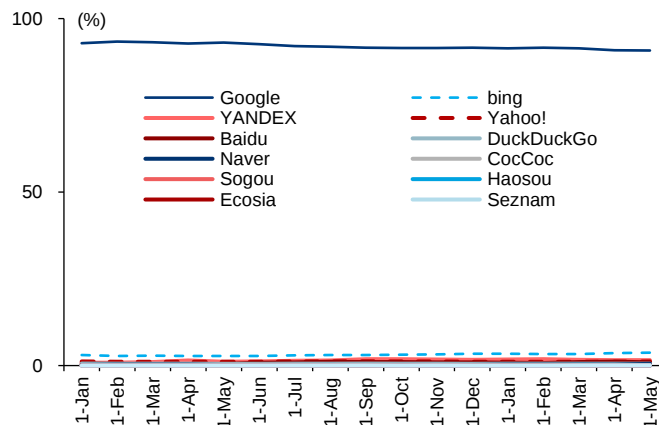
在 AI+应用方面，典型的如 AI 原生搜索类应用 Perplexity。据 SimilarWeb 数据，2023 年 1 月-2024 年 5 月，Perplexity 每月的网站访问量不断提升，截至 24 年 5 月，月网站访问量已经达到了近 9000 万，较大幅度领先于同样做 AI 原生搜索的 You.com。但是从搜索引擎的全球市占率看，据 Statcounter 数据，Google 的搜索引擎市占率从 23 年 1 月的 92.9% 仅微降到 24 年 5 月的 90.8%，Bing 的市占率从 23 年 1 月的 3.03% 微升到 24 年 5 月的 3.72%。我们认为，目前为止，AI 原生的搜索应用并未对传统搜索产生本质影响。

图表94：Perplexity 和 You.com 的网站访问情况



资料来源：SimilarWeb 官网、华泰研究

图表95：搜索引擎市占率情况



资料来源：Statcounter 官网、华泰研究

在 AI+硬件方面,代表产品为 Ai Pin 和 Rabbit R1。23 年 11 月,智能穿戴设备公司 Humane 发布基于 AI 的智能硬件 Ai Pin, 由 GPT 等 AI 模型驱动, 为 AI 原生的硬件, 支持激光屏、手势、语音等操作。24 年 4 月, Rabbit 推出 AI 驱动硬件 R1, 大小约为 iPhone 的一半。R1 用户无需应用程序和登录, 只需简单提问, 就能实现查询、播音乐、打车、购物、发信息等操作。R1 内部运行 Rabbit OS 操作系统, 基于“大型动作模型”(Large Action Model, LAM) 打造, 而非类似于 ChatGPT 的大型语言模型。LAM 可以在计算机上理解人类的意图, 借助专门的 Teach Mode, 用户可以在计算机上演示操作, R1 将进行模仿学习。但是以上两款产品发布后, 据 BBC 和 Inc 等信息, 产品的用户体验一般, 问题主要包括 AI 模型响应过慢、对网络通畅性要求过高、无法端侧推理、电池发热严重等。

图表96: Ai Pin 的交互输出方式和交互逻辑



资料来源: Humane 官网、华泰研究

图表97: Rabbit 基本参数及多模态识别展示



资料来源: Rabbit 官网、华泰研究

+AI 指的是非原生 AI 形式, 在成熟的软硬件系统上叠加一定的 AI 功能, 更符合当前模型的能力, 或成为近期的迭代重点。

在+AI 应用方面, 微软的 Copilot 系列是典型的成熟 SaaS+AI 应用。从功能覆盖来看, 微软基于成熟的操作系统、企业办公、客户关系管理、资源管理、员工管理、低代码开发等业务环节, 上线了 Copilot 相关功能, 并初步实现各应用间的 Copilot 联动。据微软 24Q1 财报数据, Github Copilot 用户数已超 50000 家, 付费用户人数 180 万人, ows 系统层面的 Copilot 装机量约 2.3 亿台。更具体细节可以参考华泰计算机 3 月 7 日报告《全球软件龙头, 拥抱 AI 时代》。

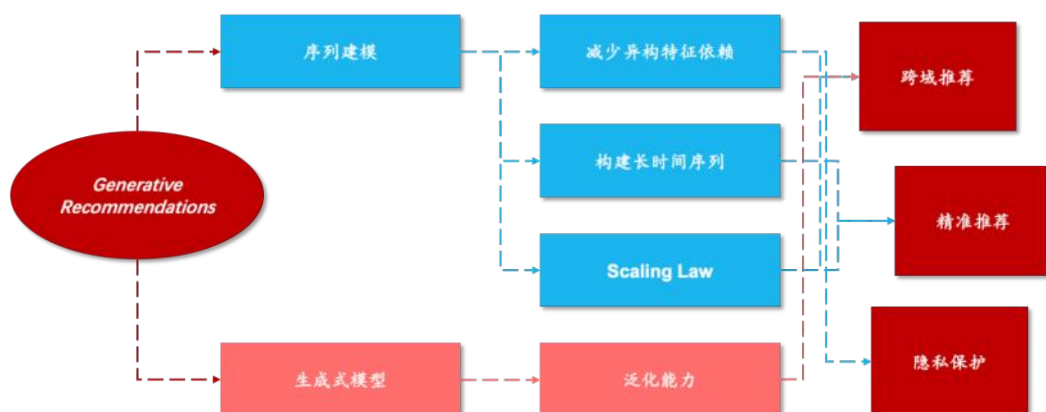
图表98: 微软在财报中公布的 Copilot 产品矩阵用户数情况

生产力	23Q2	23Q3	23Q4	24Q1	
Github Copilot用户数 (家)	27,000	37,000	50,000	N/A	超 90% 的财富100 强企业是 GitHub 客户, 收入同比+45%
qoq		37%	35%	N/A	
Github Copilot付费用户数 (万人)		100	130	180	
qoq		N/A	30%	38%	Power Apps MAU达2,500万, yoy +40%
Power Platform AI用户数 (家)	63,000	126,000	230,000	330,000	
qoq		100%	83%	43%	
Copilot Studio用户数 (家)			10,909	30,000	Cineplex 为客户服务代理构建了 Copilot, 将查询处理时间从 15 分钟缩短到 30 秒
qoq			N/A	175%	
Sales Copilot用户数 (家)		15,000	30,000	N/A	
qoq		N/A	100%	N/A	
Windows Copilot (台)			75,000,000	225,000,000	
qoq			N/A	200%	

资料来源: 公司公告、华泰研究

另一个+AI的典型应用是Meta的推荐算法+AI大模型赋能。据4月19日扎克伯格访谈，Meta从22年即开始购入H100 GPU，当时ChatGPT尚未问世，Meta主要利用这些算力开发短视频应用Reels以对抗Tiktok，其中最核心的就是推荐算法的改进。2024年4月，Meta发布生成式推荐系统论文《Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations》，开创性提出了基于Transformer的生成式推荐（Generative Recommenders, GRs）架构（更具体细节可以参考华泰计算机5月23日报告《云厂AI算力自用需求或超预期》）。据Meta 24Q1电话会，截至24Q1，Facebook上约有30%的帖子是通过AI推荐系统发布的，在Instagram上看到的内容中有超过50%是AI推荐的，已经实现了推荐引擎+AI对推荐和广告业务的赋能。

图表99：Meta提出首个生成式推荐系统模型



资料来源：《Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations》（2024）、华泰研究

在+AI硬件方面，在成熟的PC和手机上已经探索出了硬件+AI的演进道路。虽然原生的AI硬件如Ai Pin和Rabbit R1并未取得巨大成功，但是微软、联想的AI PC布局，以及Apple的AI手机布局已经清晰。从目前各厂商终端侧模型布局看，有以下特点：

1) 端侧模型参数量普遍在100亿参数以下。端侧能够支持的模型参数大小，重要的取决因素是NPU（神经处理单元）的算力多少，以及内存DRAM的大小。端侧最先进的芯片NPU算力基本在40TOPS左右，支持的参数一般在百亿级别。

图表100：AI手机主要芯片的参数情况

	发布时间	吞吐量	文生图速度	支持精度	支持模型	工艺制程
 100亿参数 ~73 TOPS	2023年10月	7B参数模型 20 tokens/s	运行Stable Diffusion和ControlNet图像生成时间小于1秒	支持INT4、INT8、INT16、FP16等整数和浮点格式，支持INT8+INT16混合精度	支持微软、百川智能、安毕、百度、智谱、OpenAI、Meta、有道、stability.ai、BLOOM等AI大模型	台积电第二代4nm
 330亿参数 68 TOPS	2024年5月 9300系列是23年11月	7B参数模型 20 tokens/s	运行Stable Diffusion文生图时间小于1秒	支持INT4、INT8、INT16、FP16等整数和浮点格式，开发混合精度INT4量化技术	阿里通义、百川、百度文心、Google Gemini Nano、零一万物、Meta Llama 2/3	台积电第三代4nm

注：x TOPS 指的是NPU提供的算力。

资料来源：高通官网、联发科官网、华泰研究

图表101: AI PC 主要芯片的参数情况

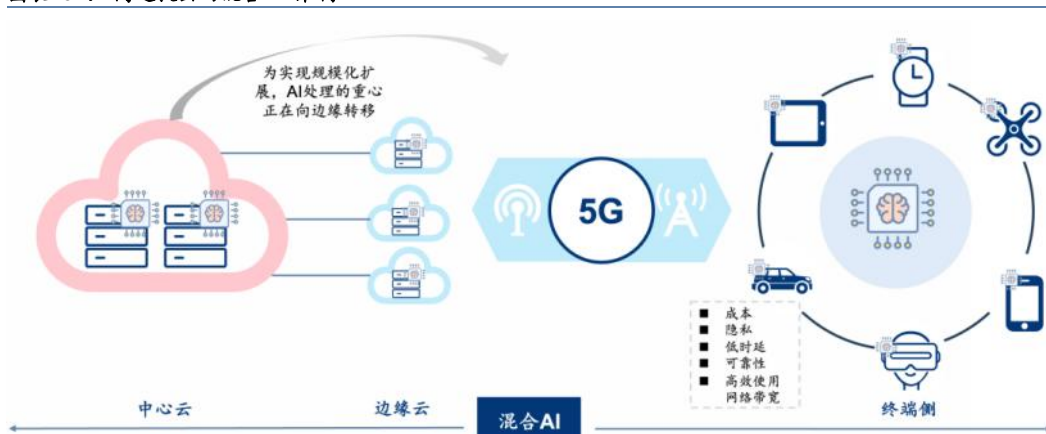
	发布时间	吞吐量	架构	CPU核	GPU核	工艺制程	内存带宽	其他
 130亿 参数 45 TOPS 全部75 TOPS	2023年10月	7B参数模型 30 tokens/s	Arm	12核, 最高 3.8 GHz	最高高达 4.6 TFLOP	台积电第二代4nm	136GB/s LPDDR5x	
 ? 亿 参数 38 TOPS	2024年5月	—	Arm	10核	10核	第二代3nm	120GB/s 统一存储带宽	
 ? 亿 参数 50 TOPS 全部80 TOPS	2024年6月	—	x86	12核/24线程, 最高5.1GHz	16计算单元	TSMC 4nm FinFET	120.0 GB/s	
 ? 亿 参数 48 TOPS 全部120 TOPS	24Q3	—	x86	—	—	—	—	优化了 500+AI模型

注: x TOPS 指的是 NPU 提供的算力。

资料来源: 高通官网、Apple 官网、AMD 官网、Intel 官网、华泰研究

2) 端云协同模式将长期存在。由于端侧模型参数量有限, 导致无法处理较复杂的任务, 因此还需要依赖云端或服务器端的模型配合。高通于 23 年 5 月发布白皮书《混合 AI 是 AI 的未来》, 指出 AI 处理能力持续向边缘转移, 越来越多的 AI 推理工作负载在手机、笔记本电脑、XR 头显、汽车和其他边缘终端上运行。终端侧 AI 能力是赋能端云混合 AI 并让生成式 AI 实现全球规模化扩展的关键。此外, 以 Apple Intelligence 的模型布局为例, 其中的编排层 (Orchestration) 会根据任务难易决定推理使用终端模型还是云端模型。我们认为, 这种端云协同的方式在端侧+AI 的形式下有望长期存在。

图表102: 高通提出的混合 AI 架构



资料来源: 《混合 AI 是 AI 的未来》, 高通 (2023)、华泰研究

3) Arm 架构芯片布局略快于 x86 架构。微软的第一批 Copilot+ PC 搭载的高通骁龙 X Elite 芯片和 Apple 自研的 M 系列芯片, 均是基于 Arm 架构打造。AMD 和 Intel 的 x86 架构 AI PC 芯片在时间上略有落后。我们认为, Arm 架构有望在终端+AI 领域提高市场份额, 但是最终 Arm 和 x86 的格局尚需观察。

大模型产业链相关公司及主要逻辑

相关产业链公司梳理

建议关注：

海外标的，AI 应用：微软、Adobe 等。国内标的，1) AI 服务器：浪潮信息等；2) AI 应用：金山办公、福昕软件、泛微网络等；3) 端侧：中科创达、网宿科技。

其他产业链相关公司：

算力相关产业链公司，海外标的：1) 芯片：Nvidia, AMD 等；2) AI 服务器：超威电脑、Dell 等；3) 定制芯片：博通等；4) 云服务：微软，Google, Amazon, Oracle；5) 存储：美光。国内标的：1) 光模块：中际旭创、天孚通信、新易盛等；2) AI 服务器：工业富联等；3) PCB：沪电股份等；4) 国产算力：海光信息、寒武纪、神州数码等。

模型相关产业链公司，主要是全球头部的模型厂商，包括微软，Google，Meta 和 Amazon。

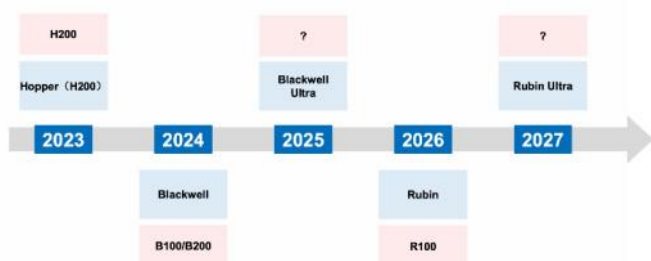
端侧 AI 相关产业链公司，海外标的：1) 手机端：Apple 及其产业链公司；2) PC 端：微软、AMD 等；3) 手机+PC 端：Arm、Qualcomm 等。4) 车+机器人端：Tesla。国内标的：1) AI PC：联想集团等。2) Apple 官方发布的 23 年度供应商名单 (Apple Supplier List 2023) 中的 AH 股公司。

相关产业链公司逻辑

一看海外算力：头部厂商训练大参数模型，需要大规模算力集群，将带动算力芯片、服务器、云服务等需求增长。

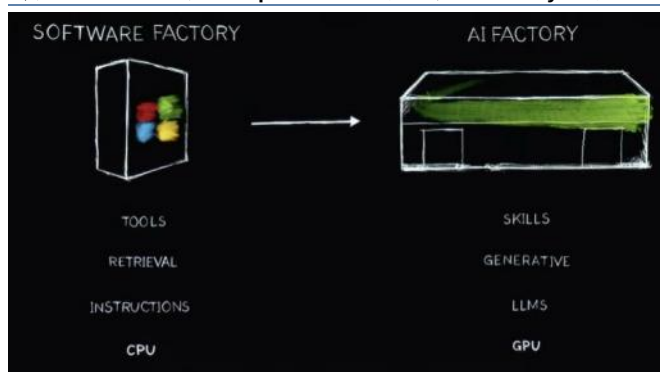
顶尖模型参数一般都在万亿级别以上（参考上文论述的 GPT-4 参数为 1.8 万亿），其训练不仅需要庞大的训练数据，更需要大规模的分布式数据中心计算集群。目前头部厂商中，Google 使用自研的 TPU 来训练和推理 Gemini 模型，Anthropic 在训练 Claude 3 时同时使用了 Google（可能使用了 TPU）和 Amazon（Nvidia GPU 为主）的 AWS 云服务（因为 Google 和 Amazon 分别对其投资了 20 亿美元和 40 亿美元），微软（OpenAI）、Meta、xAI 等头部厂商大多采用 Nvidia 的 GPU 进行训练和推理。因此，Nvidia 成为目前全球 AI 算力重要的提供商，根据其 FY25Q1 财报，Nvidia 数据中心营收为 225.63 亿美元（同比+426.7%，环比+22.6%）。在 24 年的 Computex 大会上，Nvidia 宣布了未来的芯片布局：2025 年发布 Blackwell Ultra GPU，2026-2027 年发布下一代 Rubin 和 Rubin Ultra GPU 芯片，实现一年一迭代。同时，Nvidia 强化了 AI Factory 概念，实现大模型 token 的“量产”。

图表103: Nvidia 在 Computex 大会上宣布了 GPU 迭代计划



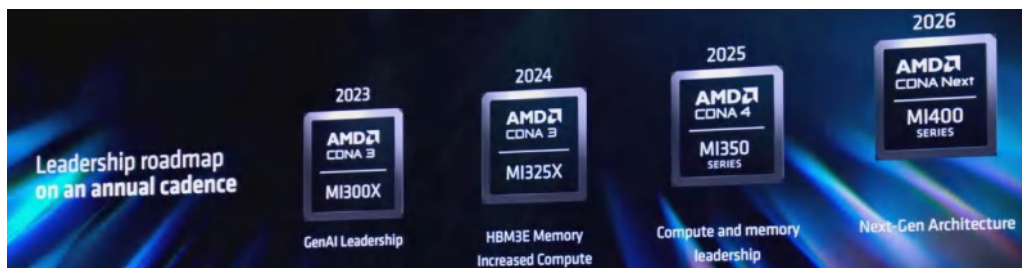
资料来源：Nvidia Computex 大会演讲、华泰研究

图表104: Nvidia 在 Computex 大会上强化了 AI Factory 概念



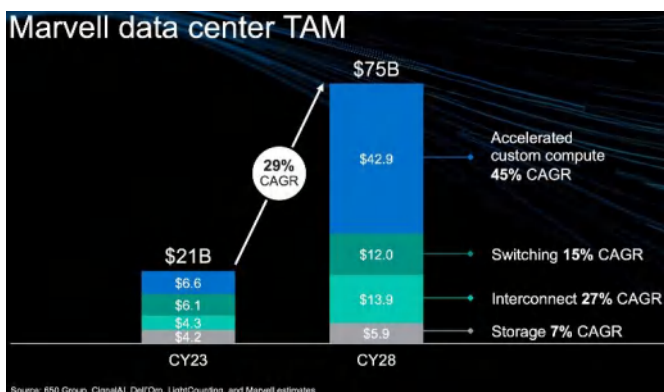
资料来源：Nvidia Computex 大会演讲、华泰研究

据 tom's HARDWARE 信息，同样作为数据中的芯片的提供商，AMD 的 MI300 系列也获得了微软、Meta 等部分客户的订单，同样受益于头部厂商大参数模型的训练和推理。AMD 在 Computex 大会上，同样宣布了未来的芯片迭代计划，与 Nvidia 一样实现了一年一迭代，23-24 年分别发布 MI300X、MI325X、MI350 和 MI400。

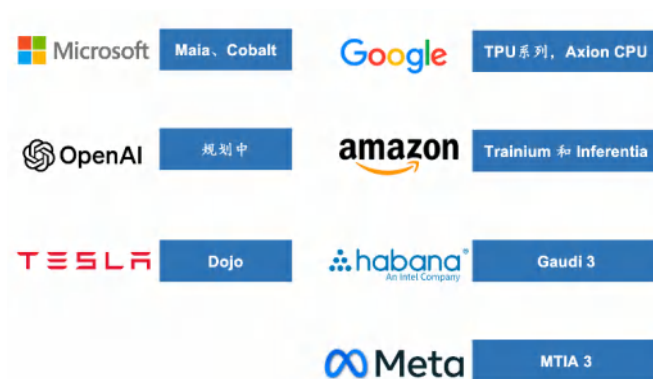
图表105：AMD 在 Computex 大会上宣布的 GPU 迭代路径


资料来源：AMD Computex 大会演讲、华泰研究

Google TPU 在大模型训推的成功应用，以及头部云厂商加速研发定制芯片，带动了定制加速芯片的崛起。据 Marvell 在 24 年 4 月的 AI Era Event 数据，23 财年 Marvell 数据中心 TAM 为 210 亿美元，其中加速定制计算为 66 亿美元；到 28 财年，加速定制计算 TAM 将达到 429 亿美元，5 年 CAGR 为 45%。据 24 年 Computex 大会联发科的数据，2023 年定制加速计算 TAM 为 61 亿美元，到 2028 年将达到 400 亿美元，5 年 CAGR 为 43%，加速计算市场成长迅速。据 Business Insider 信息，博通是 Google TPU 最大的代工厂商，Meta 也已成为博通的大客户，并有望在 2024 年下半年和 2025 年推出 Meta 的第三代 AI ASIC 芯片（3nm 制程的 MTIA 3）。此外，博通还布局了 AI Ethernet 芯片和高速网络组件。24Q2 财报，博通给出指引，2024 年 AI 相关收入比例将达到 35% 以上，收入体量将在 110 亿美元左右。我们认为，加速定制芯片将随着大模型推理需求的铺开而加速渗透。

图表106：Marvell 在 AI Era Event 上给出的加速定制计算 TAM 预测


资料来源：Marvell 官网、华泰研究

图表107：头部厂商开始研发定制加速芯片


资料来源：各公司官网、华泰研究

若以 GPU 芯片为上游，那么产业链下一环节则是服务器。超威电脑（SMCI）是 Nvidia 的重要合作伙伴，为 Nvidia H 系列芯片适配了十余款服务器，其产品成为下游客户的重要选择之一。在 FY24Q3 财报中，超威电脑上调 FY24Q4 收入指引到 51-55 亿美元，24 年全年指引为 147-151 亿美元，并指出今年年底马来西亚工厂有望投产，进一步扩大产能。

除超威电脑，在服务器领域，Dell 凭借着在 B 端客户的资源积累，以及对 Nvidia 的产品适配，相关服务器产品（PowerEdge XE9680/XE9680L 等）也在放量。据 Dell FY25Q1 财报，Dell AI 服务器收入 17 亿美元，环比增长 112.5%；积压订单 38 亿美元，环比增长 31%；新增订单 26 亿美元。据 SemiAnalysis 信息，Dell 相比超威电脑的一个优势在于，Dell 利用旗下 Dell Financial Services（DFS，戴尔金融服务），以比正常市场更低的利率向较小的第三方新型云厂商（NeoCloud，如 Coreweave 等）提供融资，来购买其服务器产品。虽然 Dell 的服务器毛利率略高于超威电脑，但是由于低廉的融资利率，导致第三方厂商最终部署 Dell 服务器产品的成本低于超威电脑。

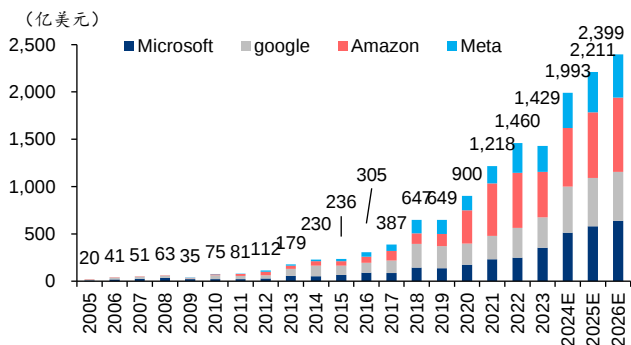
图表108: Dell AI 服务器积压订单、收入、新增订单情况

财年	财报或公告时间	AI 服务器积压订单 (亿美元)	AI 服务器收入 (亿美元)	AI 服务器新增订单 (亿美元)
24Q2	2023.7.31	8		
	2023.8.31	20		12
24Q3	2023.10.31	16	5	1
24Q4	2024.1.31	29	8	21
25Q1	2024.5.3	38	17	26

注：当前积压订单=上期积压订单-当期收入+当期新增订单

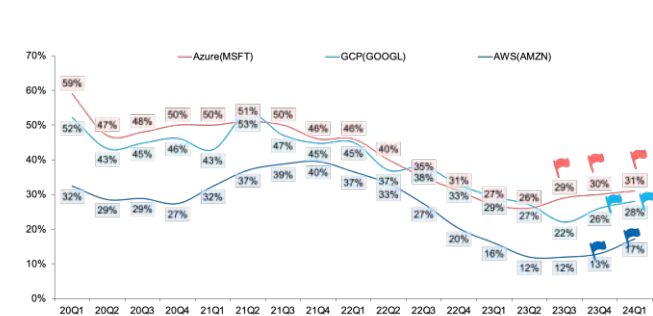
资料来源：公司公告、华泰研究

头部云厂商在购买 AI 算力、构建大规模集群、训练大参数模型的同时，也提供：1) 闭源模型和开源模型的云托管服务。按照模型调用 token 数、或企业的模型定制套餐用量，实现对大模型 API 调用的收费，并为开发者提供 AI 一站式开发工具（参照前文各家模型的定价）。2) 针对传统业务的改造。根据华泰计算机 5 月 23 日的报告《云厂 AI 算力自用需求或超预期》，Meta 基于 Transformer 大模型技术的第三代推荐系统，构建了具备泛化能力的统一推荐模型，同时提升用户意图理解与项目推荐效果。据 Meta 24Q1 财报，AI 已经赋能了广告业务，截至 24Q1，Facebook 上约有 30% 的帖子是通过 AI 推荐系统发布的，在过去几年里增长了 2 倍，用户在 Instagram 上看到的内容中有超过 50% 由 AI 推荐。Google 也在搜索引擎中逐步扩展 SGE（搜索生成式体验）和 AI Overviews 服务，让大模型赋能搜索业务。从效果来看，Microsoft、Google、Amazon 云收入增速自 23Q3 逐步企稳上升，24Q1 三家公司云收入同比增速分别为 31%、28%、17%，分别环比提升 1、2、4pct。

图表109: 2005-2026 年云厂商+Meta 的 CapEx 情况


注：各家公司均为自然年口径

资料来源：各公司官网、Visible Alpha 预测、华泰研究

图表110: 2020Q1-2024Q1 云业务收入同比增速变化


注：各家公司均为自然年口径

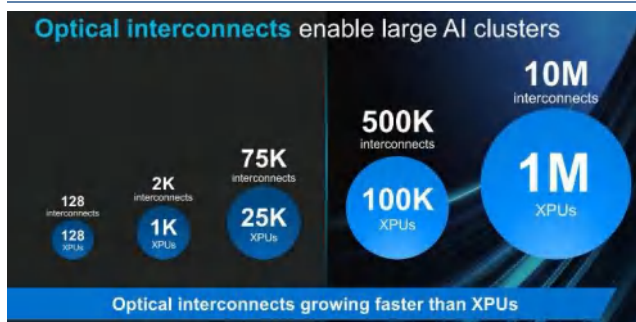
资料来源：各公司官网、华泰研究

二看国内算力：国内相关厂商能够参与到全球算力产业链。此外，美国商务部工业和安全局（BIS）对先进计算芯片的出口限制，加速了国产替代需求。

Nvidia 作为目前全球 AI 算力龙头，能参与到 Nvidia 产业链的国内相关标的包括：1) **光模块相关公司**，包括中际旭创、天孚通信、新易盛（均为华泰通信覆盖）。据 Marvell AI Era Event，GPT-3 在 1000 张卡的 GPU 集群上训练，使用约 2000 个光互连（比例 1:2），未来对于 100 万张卡的 GPU 集群，需求的光互连可能达到 1000 万，比例为 1:10。2) **PCB 相关公司**，包括沪电股份（华泰电子覆盖）等。

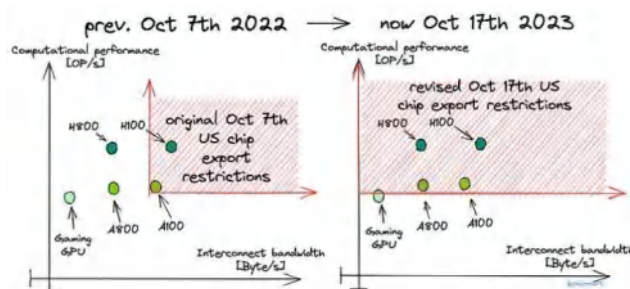
类似的，国内 AI 服务器需求也受国产大模型训练和推理的提振，相关公司包括工业富联和浪潮信息。此外，23 年 10 月，美国商务部工业和安全局（BIS）修改和强化了 2022 年 10 月发布的“先进计算芯片和半导体制造设备出口管制规则”，改用总处理性能（TPP）和性能密度（PD）两个参数，进一步限制国内能购买到的高端算力芯片，A800、H800 在列。因此，国产算力替代需求紧迫。在三期国家大基金等国家政策的支持下，国产算力芯片有望迎来快速迭代，相关标的包括海光信息、寒武纪、神州数码等。

图表111: Marvell 在 AI Era Event 上给出的光互联增长预测



资料来源: Marvell 官网、华泰研究

图表112: 美国商务部工业和安全局 2023 年更新的芯片限制法案图解



资料来源: SemiAnalysis、华泰研究

三看大模型进展: 大模型是 AI 的软件“基础设施”, 是当前 AI 能力的主要来源, 头部模型的进展和迭代情况, 或对整个 AI 产业链产生重要影响。

微软作为 OpenAI 最大的投资人和关系密切的合作伙伴, 较早享受到了 OpenAI GPT 系列模型的最新迭代成果。同时前文也提到, 微软正在自研 Phi 系列小模型, 进一步降低推理成本, 并适配终端 Copilot+ PC 的使用场景。Google 是 Transformer 的提出者 (2017 年), 也是 AI 领域的重要理论创新者, 虽然 23 年早期的 PaLM 1/2 模型性能不及同时期的 GPT 系列, 但是 23 年底 Gemini 系列正式问世, 达到了 GPT-4 同级别的水平, 并且在后续更新的 Gemini Pro 1.5 版本中, 充分发挥了长文本的优势, 最长支持到 2M 上下文长度 (参见前文分析)。随着 Gemini 系列模型的持续迭代, 或将继续缩小与 GPT 系列模型的差距。

Meta 是大模型开源领域的倡导者和践行者, 23-24 年先后开源了 Llama 1-3 系列, 打造了良好的开源生态, 并且 Llama 性能与闭源头部厂商模型的差距也呈缩小态势 (参见图表 58)。基于 Llama 模型的 Meta AI 助手, 帮助 Meta 推进大模型技术在 App (Facebook、Instagram 等)/智能可穿戴设备 (雷朋眼镜) 上的落地, 以更好的实现用户数据驱动的模式迭代, Meta 后续模型的迭代值得关注。Amazon 自身的大模型关注度一般, 但是 Amazon 于 24 年 5 月底完成了对 Anthropic 的 40 亿美元投资, 而据之前的分析, Anthropic Claude 模型属于全球 Top 3 性能的闭源模型, 或为 Amazon 的云业务带来增量。

四看端侧 AI 进展: 小模型性能的提升为其在终端部署提供了条件。基于成熟的 PC 和手机硬件, 端侧 AI 的进展值得关注。端侧可以从三个部分拆解, 即芯片、AI PC 和 AI 手机。

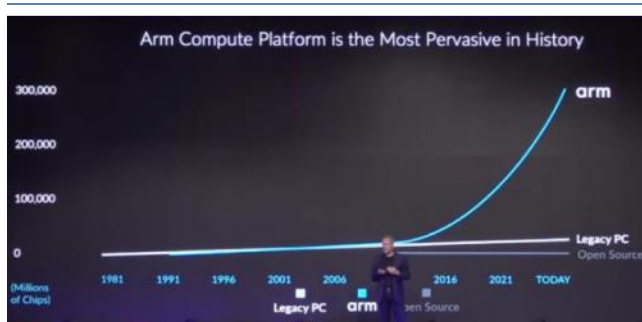
从芯片底层看, 1) Qualcomm 在 AI PC+手机两个端侧均有布局。Qualcomm 在移动平台上的骁龙系列芯片为安卓阵营的高端芯片。据 Statcounter 数据, 截至 24 年 5 月, 安卓 OS 的市占率在移动 OS 中为 71.8%, 在全部 (包含桌面 OS) 市占率为 43.9%, 大幅领先于其他 OS。此外, Qualcomm 在 23 年 10 月发布了 PC 端的芯片 X Elite 系列, 早于竞争对手 AMD 和 Intel。X Elite 搭载的 NPU 算力超 45 TOPs, 并率先在微软发布的 Copilot+ PC 上搭载, 6 月 18 日开售。2) ARM 架构抢先布局 AI 端侧。安卓系统、Qualcomm X Elite 芯片均是基于 Arm 架构搭建。与采用复杂指令集 (CISC) 的 x86 架构相比, 采用精简指令集 (RISC) 的 Arm 架构能够更好的控制功耗, 非常适合终端场景。在 24 年 Computex 大会上, Arm 宣布 Arm 架构已经成为历史上最受欢迎的计算平台, 基于 Arm 的芯片数量达到 3000 亿, 覆盖移动+PC+车+云, Arm 认为其在 ows 中的市场份额 5 年内将达到 50%。

图113: Qualcomm X Elite 芯片重要参数



资料来源: Qualcomm 官网、华泰研究

图114: Arm 架构已经成为最受欢迎的计算平台



资料来源: Arm Computex 大会演讲、华泰研究

从AI PC看,微软和联想较早定义并发布了AI PC产品。4月18日,2024 Lenovo Tech World 发布了内置个人智能体“联想小天”的AI PC系列产品,其中内嵌了个人大模型,能与用户自然交互,在工作、学习和生活等诸多场景中为用户带来全新的AI体验。24年5月微软 Build 大会上,微软正式发布 Copilot+ PC,首批选用 Qualcomm 的 X Elite 芯片,并在 PC 上部署了微软自研的模型 Phi-Silica,能够实现 Recall (回忆用户曾经使用过的界面内容)、图像生成、文本处理等功能,6月18日正式开售。在24年 Computex 大会上,AMD 也发布了 Ryzen AI 300 系列芯片,并预告将与多个 OEM (acer、ASUS、HP、Lenovo、msi 等) 合作 100+款 AI PC,最早7月上市。

图115: 首批微软 Copilot+ PC



资料来源: Qualcomm 官网、华泰研究

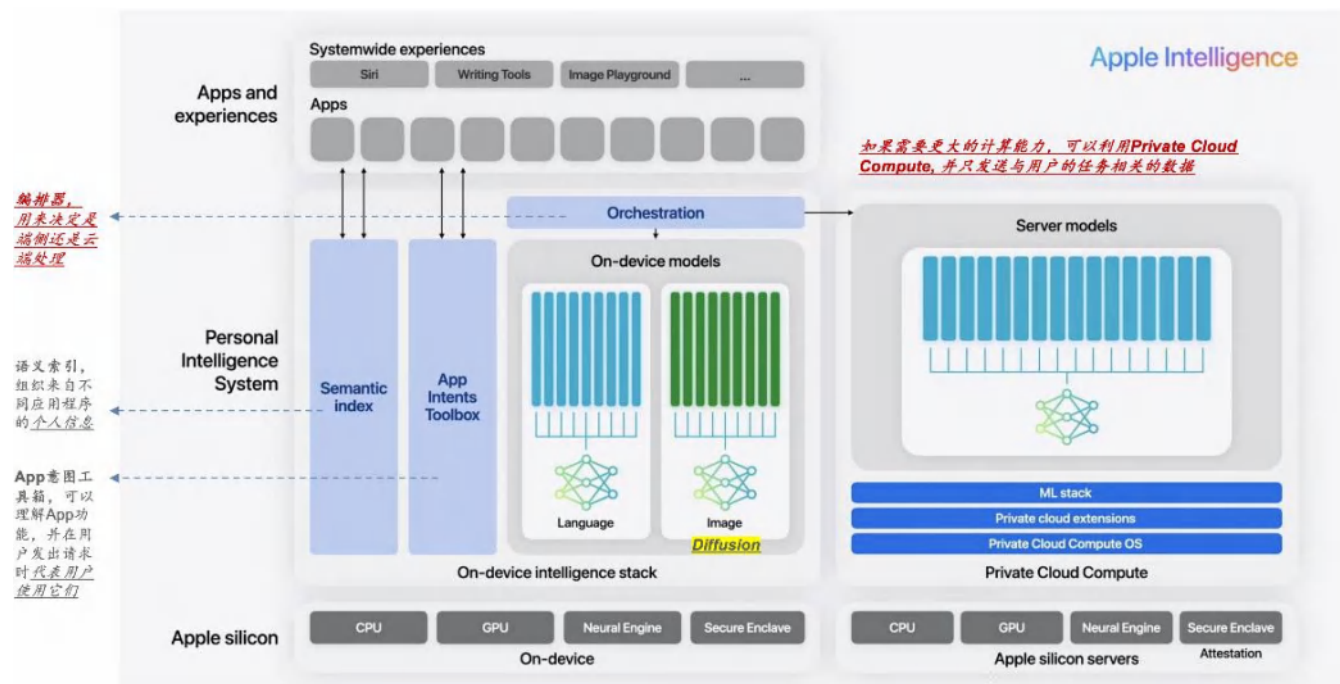
图116: AMD 将在7月发布 Ryzen AI 芯片驱动的 AI PC



资料来源: AMD Computex 演讲、华泰研究

从AI手机看,虽然 Google 在23年12月已经将 Gemini Nano 模型部署到了 Google Pixel 8 Pro 上,但是 Pixel 用户量较少,并未对AI手机产生较大影响。24年6月 Apple WWDC 上,发布了 Apple 自研全套的AI系统——Apple Intelligence,基本“复现”了已有的端侧能力,包括文本处理、图像生成、Siri 智能语音交互等。同时 Apple 在端侧和云端均部署了自研的大模型(端侧参数大小 3B,云端未知参数),并能够以第三方模型的形式接入 OpenAI 的 GPT-4o 及其他模型。据 Statista 数据,截至2023年,全球 iPhone 手机用户高达 13.8 亿。大量用户叠加 Apple 完善的软件(iOS 和 App Store)和硬件(A系列芯片)生态,或对AI手机的演进产生重大影响。相关标的包括 Apple 的产业链公司,涉及检测、面板、电池、光学、声学、结构件等环节。

图表117: Apple Intelligence 的模型布局

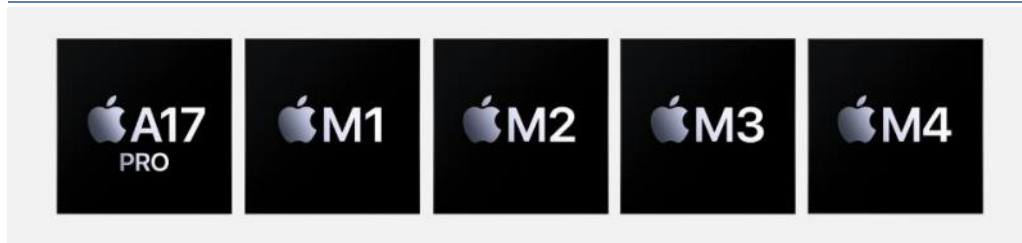


资料来源: Apple 官网、华泰研究

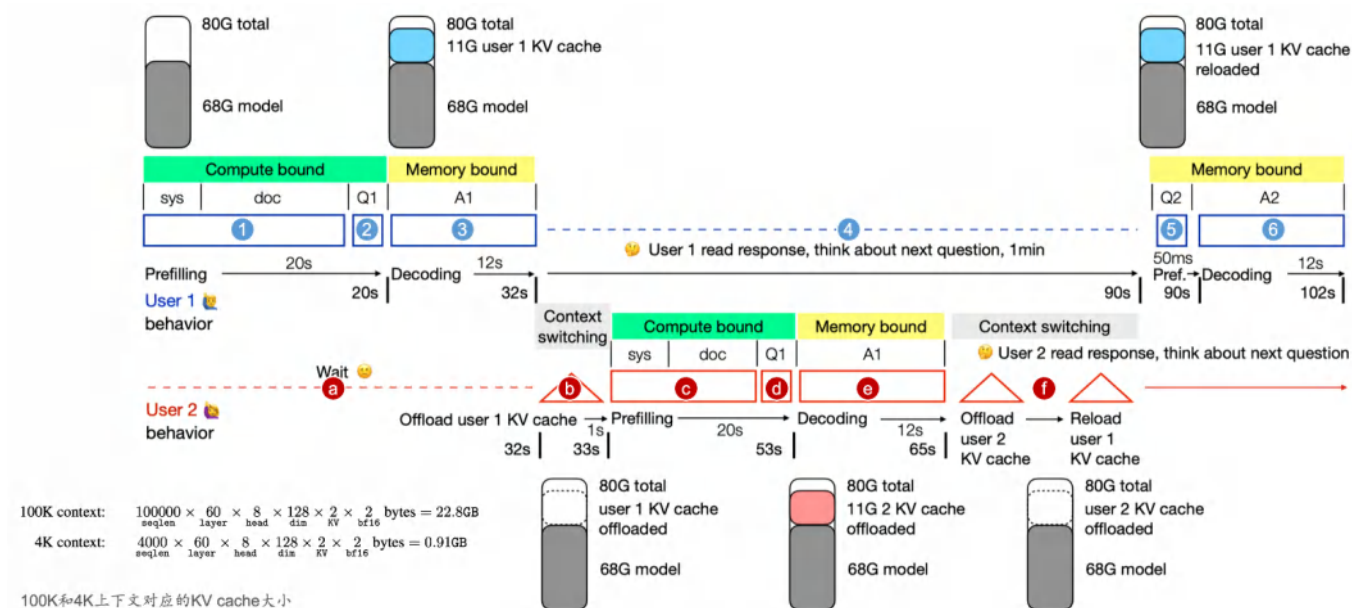
端侧 AI 还有望带动存储的需求增长。从 Apple Intelligence 支持的设备看, 仅支持 A17 Pro 及以上规格芯片的手机, 以及 M1 及以上规格芯片的 PC。然而根据 Apple 官方数据, M1 系列芯片的神经引擎 (Neural Engine) 算力表现可以达到 11 TOPS, 而 A16 芯片可以达到 17 TOPS, 即仅从算力上看旧款的 A16 是有能力支持 Apple Intelligence 的。然而从内存的角度看, M1 的 DRAM 最低为 8GB, 高于 A16 配备的 6GB。我们认为, 这缺少的 2G DRAM 可能就是旧款 Apple 手机端侧设备不支持 Apple Intelligence 的重要原因。

该推论可以交叉验证, 据华泰计算机 6 月 10 日研报《国产大模型“凭”什么降价? 》, 以 A100 80G 推理 34B 参数模型为例, 34B 模型的参数数据加载到 A100 的 HBM (高性能 DRAM) 中需要占用 68GB 的内存空间 (相当于全部空间的 85%), 类比 Apple 端侧 3B 模型常驻 iPhone DRAM 中, 也需要消耗一定量的内存。这种模型常驻内存的形式, 消耗了有限 DRAM 的部分空间, 因此导致 A16 配置的 6GB 无法在满足日常 App 运行的情况下, 常驻 3B 大模型进行推理。我们预测, 未来为了保证 Apple Intelligence 的正常推理, Apple 或将加大手机的内存起始配置, DRAM 存储等相关标的包括美光等。

图表118: 支持 Apple Intelligence 的芯片类型



资料来源: Apple 官网、华泰研究

图表119：大模型参数常驻 DRAM 内存，将占据相当的一部分存储空间


注：大模型参数占据的空间为灰色块表示部分

资料来源：《Challenges in Deploying Long-Context Transformers: A Theoretical Peak Performance Analysis》，Yao（2024）、华泰研究

五看 AI 应用进展：AI 大模型最终要落地到应用上，实现商业化，值得持续关注。截止 24 年 6 月，仍然没有出现所谓的 Super App（超级应用），但是海外面向 B 端客户的商业化产品在不断成熟。

微软的 Microsoft 365 Copilot 是较早结合大模型技术的应用,于 23 年 3 月发布,基于 GPT-4 的能力,实现了在 Word、Excel、PowerPoint、Outlook、Teams 等应用中使用自然语言进行创作、编辑、摘要、生成 PPT 等功能。随后,微软持续拓展 Copilot 品类,推出了 GitHub Copilot、Copilot for Sales、Copilot for Service 等产品,并先后实现了定价。在 FY23Q4-FY24Q3 连续 4 个季度,微软宣布 AI 相关收入分别占 Azure 云收入的 1%、3%、6%和 7%,体现了 AI 对于云业务的持续拉动。(参见华泰计算机 3 月 7 日报告:《微软:全球软件龙头,拥抱 AI 时代》)。

图表120: Microsoft Copilot 产品发布时间线梳理（标星代表具备该项能力）

	Copilot	Copilot Pro	Copilot for Microsoft 365	Copilot for Sales	Copilot for Service
定价	免费	\$ 20/人/月	\$ 30/人/月	\$ 50/人/月	\$ 50/人/月
基础能力 Foundational capabilities	★	★	★	★	★
联网检索 Web grounding	★	★	★	★	★
商业数据保护 Commercial Data Protection	★	★	★	★	★
优先模型访问 Priority Model Access		★	★	★	★
Microsoft 365 应用程序（不含 Teams）		★	★	★	★
Teams 助手 Copilot in Teams			★	★	★
Graph 接地 Microsoft Graph grounding			★	★	★
企业级数据保护 Enterprise-Grade Data Protection			★	★	★
Copilot 自定义		Copilot GPT Builder	Copilot Studio	★	★
特定能力 Role-specific capabilities				★	★

资料来源：Microsoft Blog 官网、华泰研究

另一个 AI 应用领域较为领先的厂商是 Adobe。Adobe 将 AI 与自身产品结合同样很早，在 2023 年 3 月发布生成式 AI 核心产品 Firefly 以来，Adobe 陆续将生成式 AI 技术与 Pr、PS、Express 等多个产品结合。23 年 9 月，Adobe 宣布近半年的测试完成，正式将 AI 能力集成于 Creative Cloud 产品线中，并于 23 年 11 月 1 日提高订阅价格，根据产品线的不同，提价幅度大约为 6%-10%，实现了 AI 对 ARPU 的抬升。24 年 4 月 15 日，Adobe 一方面宣布正在开发新的 Firefly 视频模型，另一方面表示第三方生成式 AI 工具和模型将能够与 Premiere Pro 产品相结合，包括 OpenAI 的 Sora、Runway Gen 系列和 Pika。Adobe 意在打造创意式平台，吸引更多的多模态模型接入。根据 Adobe 24Q2 财报，Adobe 宣布 Firefly 自问世以来，已经完成了超过 90 亿次请求，在 iOS 和 Android 上推出的全新 Express 应用程序月活跃用户比上季度翻了一番。在 AI 的赋能下，Adobe 提高了 24 年度数字媒体业务的 ARR、数字体验订阅的收入，实现了 AI 对业务增长的促进作用。（参见华泰计算机 1 月 17 日报告：《奥多比：创意+营销 SaaS 龙头拥抱 AI 时代》）

图表121: Adobe 系列产品矩阵提价情况

订阅方案	付费方式	2023.11 价格变化（美元）	提价幅度	每月附赠的 fast credits	2023.11 更新变化
Creative Cloud for individual Single App Plans	annual billed monthly option	20.99→22.99 (+2)	9.5%	500 或 250 或 100 或 25, 由订阅的 App 决定	•新增 Generative Credits、Adobe Express 和 Adobe Firefly •在 2023 年 11 月 1 日之前，Creative Cloud、Adobe Firefly、Adobe Express 和 Adobe Stock 付费订阅者将不受生成式 credit 额度限制。从 2023 年 11 月 1 日起，将适用 credit 额度
	month-to-month option	31.49→34.49 (+3)	9.5%		
	annual prepaid option	239.88→263.88 (+24)	10%		
Creative Cloud for Individual All Apps Plan	annual billed monthly option	54.99→59.99 (+5)	9%	1000	•Credit 消费标准：Generative Fill, Generative Expand, Text to Image, Generative Recolor 为 1credit/次；Text Effects 11 月 1 日前免费，之后 1credit/次 •免费用户每月送 25credits •目前生成图片为 2000 x 2000 像素，未来提供更高分辨率的图像、动画、视频和 3D 生成式 AI 功能，消耗点数将更多
	month-to-month option	82.49→89.99 (+7.5)	9%	1000	
	annual prepaid option	599.88→659.88 (+60)	10%	1000	
Creative Cloud for Teams	Teams single app plans / month / license	35.99→37.99 (+2)	5.6%	500	
	Teams All Apps plan / month / license	84.99→89.99 (+5)	5.9%	1000	
Creative Cloud for Enterprise and Reseller Plans	需联系 Adobe 帐户代表或经销商	单独定价	—	需要根据企业实际情况定制	

资料来源：Adobe 官网、华泰研究



国产大模型初创公司投融资情况

国内大模型初创公司得到 VC 青睐，头部公司均是估值在 10 亿美元以上的独角兽。国内最主要的大模型初创公司有 5 个，智谱、月之暗面、MiniMax（稀宇科技）、百川智能和零一万物。从投资方来看，主要资金来自互联网公司和美元基金。其中，阿里在投资上最为积极，已经向全部 5 家初创公司投资。从估值上看，据已公开的信息，智谱、月之暗面估值最高，为 30 亿美元左右；MiniMax 其次，估值 25 亿美元左右；百川智能和零一万物估值在 10 亿-20 亿美元之间。

图表122：国产大模型初创公司估值及主要投资方（截至 24 年 6 月）

	估值	互联网公司						美元基金			
		阿里	腾讯	百度	小米	美团	米哈游	小红书	红杉	真格	高瓴 IDG
智谱	30 亿美元	✓	✓		✓	✓			✓		✓
百川	130 亿元	✓	✓		✓						
月之暗面	30 亿美元	✓				✓		✓	✓	✓	
MiniMax	25 亿美元	✓	✓				✓				✓
零一万物	10 亿美元	✓									✓

资料来源：企查查、华泰研究

图表123: 提及公司列表

公司代码	公司简称
MSFT US	微软
未上市	OpenAI
GOOG US	谷歌
ADBE US	奥多比
NVDA US	Nvidia
AMD US	AMD
SMCI US	超威电脑
DELL US	Dell
META US	Meta
AAPL US	Apple
ARM US	Arm
QCOM US	Qualcomm
TSLA US	Tesla
AMZN US	Amazon
ORCL US	Oracle
3396 HK	联想
未上市	Anthropic
未上市	xAI
未上市	Mistral AI
未上市	Inflection
未上市	Cohere
未上市	Character.ai
未上市	Databricks
未上市	AI21
BIDU US	百度
BABA US	阿里
0700 HK	腾讯
601360 CH	360
002230 CH	科大讯飞
300418 CH	昆仑万维
0020 HK	商汤
未上市	智谱
未上市	MiniMax
未上市	月之暗面
未上市	百川智能
未上市	零一万物
未上市	阶跃星辰
300308 CH	中际旭创
300394 CH	天孚通信
300502 CH	新易盛
000977 CH	浪潮信息
601138 CH	工业富联
688041 CH	海光信息
000034 CH	神州数码
688256 CH	寒武纪-U
688111 CH	金山办公
688095 CH	福昕软件
603039 CH	泛微网络
300496 CH	中科创达
300017 CH	网宿科技
002463 CH	沪电股份

资料来源: Bloomberg、华泰研究



风险提示

宏观经济波动。若宏观经济波动,可能对 AI 产业资本投入产生负面影响,导致 AI 产业变革、新技术落地节奏、整体行业增长不及预期。

技术进步不及预期。若 AI 技术、大模型技术、AI 应用进展不及预期,或对行业落地情况产生不利影响。

中美竞争加剧。中美竞争加剧,或影响国内算力基础设施布局,导致国内 AI 大模型技术迭代速度放缓。

研报中涉及到未上市公司或未覆盖个股内容,均系对其客观公开信息的整理,并不代表本研究团队对该公司、该股票的推荐或覆盖。

图表124：重点推荐公司一览表

股票名称	股票代码	投资评级 (当地币种)	最新收盘价	目标价	市值 (百万)	EPS (元)				PE (倍)			
			(当地币种)	(当地币种)		2023	2024E	2025E	2026E	2023	2024E	2025E	2026E
微软 (Microsoft)	MSFT US	买入	446.95	477.33	3,321,869	9.70	12.24	14.14	16.51	46.08	36.52	31.61	27.07
奥多比 (Adobe)	ADBE US	买入	555.54	613.45	246,326	11.50	13.05	15.07	17.55	48.31	42.57	36.86	31.65
浪潮信息	000977 CH	买入	37.28	50.67	54,881	1.21	1.45	1.72	1.99	30.81	25.71	21.67	18.73
金山办公	688111 CH	买入	205.00	354.50	94,722	2.85	3.63	4.85	6.44	71.93	56.47	42.27	31.83
福昕软件	688095 CH	买入	43.75	73.96	4,003	-0.99	-0.03	0.29	0.72	-44.19	-1,458.33	150.86	60.76
泛微网络	603039 CH	买入	31.71	41.97	8,264	0.69	0.97	1.11	1.29	45.96	32.69	28.57	24.58
中科创达	300496 CH	买入	46.56	62.65	21,380	1.01	1.33	1.75	2.26	46.10	35.01	26.61	20.60
网宿科技	300017 CH	买入	8.01	12.08	19,551	0.25	0.25	0.31	0.35	32.04	32.04	25.84	22.89

资料来源：Bloomberg，华泰研究预测

图表125：重点推荐公司最新观点

股票名称	最新观点
微软(Microsoft) (MSFT US)	智能云业务为重要增长动力，维持“买入”评级 微软发布 FY24Q3 季报，FY24Q3 总收入 619 亿美元(yoy+17%)，营业利润 276 亿美元(yoy+23%)；净利润 219 亿美元(yoy+20%)。生产力业务收入 196 亿美元(yoy+12%)，智能云收入 267 亿美元(yoy+21%)，个人计算收入 156 亿美元(yoy+18%)。公司指引 FY24Q4 生产力业务收入同增 9%-10%；智能云业务收入同增 18%-20%；个人计算业务收入同增 9%-12%，智能云仍为重要增长动力。预计 FY24-26 公司 EPS 为 12.24/14.14/16.51 美元，可比公司 FY24 31.8xPE，考虑公司优势卡位与 AI 领先布局，给予 FY24E 39xPE，目标价 477.33 美元，“买入”。 风险提示：宏观经济波动；技术落地不及预期。 报告发布日期：2024 年 04 月 26 日 点击下载全文：微软(Microsoft)(MSFT US,买入)：智能云受益 AI，Copilot 顺利推进
奥多比(Adobe) (ADBE US)	FY24Q2：收入/ARR/利润超指引，关注 AI 商业化进展 公司 FY24Q2 单季度收入 53.1 亿美元，同增 10.2%（此前指引 52.5-53.0 亿美元）；RPO 178.6 亿美元，同增 17.3%。盈利（Non-GAAP）：经营利润 24.4 亿美元，OPM 46.0%；净利润 20.2 亿美元，NPM 38.1%，摊薄 EPS 4.48 美元（此前指引\$4.35-4.40 美元）；公司上调全年业绩指引。我们看好公司 AI 商业化前景，预测 FY24-26 公司总收入 214.8/239.2/268.7 亿美元，2024 年可比公司平均估值 10.4xPS，考虑公司在创意领域较强的市场地位，给予 FY24 12.8xPS，目标价 613.45 美元/股，维持“买入”评级。 风险提示：宏观经济波动；技术落地不及预期；市场竞争加剧。 报告发布日期：2024 年 06 月 15 日 点击下载全文：奥多比(Adobe)(ADBE US,买入)：业绩超指引，AI 商业化或加速
浪潮信息 (000977 CH)	24Q1 收入同增 85%，受益算力产业高景气度 浪潮信息发布一季报，2024 年 Q1 实现营收 176.07 亿元（重述 yoy+85.32%），归母净利 3.06 亿元（重述 yoy+64.39%），扣非净利 2.40 亿元（重述 yoy+40.15%）。我们认为公司收入高速增长，反映了算力产业较高的景气度，在 AI 驱动下，大厂或将保持较高的资本开支水平，从而拉动算力产业高增，公司作为 AI 服务器龙头有望持续受益。我们预计公司 2024-2026 年 EPS 分别为 1.45、1.72、1.99 元。可比公司 2024 年一致预期 PE 均值为 23.4 倍，考虑公司是全球 AI 服务器龙头，具备深厚技术、渠道积累，给予 2024 年 35 倍 PE，目标价 50.67 元，“买入”。 风险提示：宏观经济波动风险；市场竞争加剧；供应链风险。 报告发布日期：2024 年 04 月 30 日 点击下载全文：浪潮信息(000977 CH,买入)：24Q1 收入同增 85%，AI 或带动收入起量
金山办公 (688111 CH)	AI 商业化有望加快，维持“买入”评级 金山办公发布一季报，24Q1 营收 12.25 亿元（yoy+16.54%），归母净利 3.67 亿元（yoy+37.31%），扣非净利 3.52 亿元（yoy+40.56%）。销售/管理/研发费用率为 17.62%/9.22%/33.03%，同比-6.11pct/-0.72pct/-0.75pct，收入稳健增长+费用管控推动利润快速增长。我们认为随着 C 端定价推进，B 端产品落地，AI 商业化有望加快。我们预计 24-26 年公司 EPS 为 3.63/4.85/6.44 元，收入 58.47/77.36/102.64 亿元，可比公司平均 24E 17.2xPS()，考虑公司 AI 产品快速迭代，给予 24 年 28xPS，目标价 354.50 元，维持“买入”。 风险提示：付费用户增长不及预期、AI 商业化不及预期。 报告发布日期：2024 年 04 月 23 日 点击下载全文：金山办公(688111 CH,买入)：AI 商业化有望加快
福昕软件 (688095 CH)	24Q1 收入加速增长+亏损同比收窄，维持“买入”评级 福昕软件发布年报及一季报，2023 年实现营收 6.11 亿元（yoy+5.33%），归母净利-9094 万元（22 年为-174 万元），扣非净利-1.79 亿元（22 年为-0.78 亿元）。24Q1 营收 1.69 亿元（yoy+16.87%、qoq+0.33%），归母净利-1060.90 万元（同比收窄 6.47%），扣非净利-2065.62 万元（同比收窄 21.55%）。亏损主因订阅转型影响收入增速+费用投入力度加大。随着转型逐步深入，收入有望逐步加速增长。我们预计公司 2024-2026 年 EPS 分别为-0.03、0.29、0.72 元。可比公司 24 年 ifind 一致预期 PS 均值为 9.1 倍，给予 2024 年 9.1 倍 PS，目标价 73.96 元，维持“买入”评级。 风险提示：新应用拓展不及预期；市场竞争风险；联营企业业绩波动风险。 报告发布日期：2024 年 04 月 29 日 点击下载全文：福昕软件(688095 CH,买入)：24Q1 收入加速增长，关注 AI 商业化



股票名称	最新观点
泛微网络 (603039 CH)	关注下游需求恢复情况，维持“买入”评级 泛微网络发布一季报，2024 年 Q1 实现营收 3.43 亿元（yoy+3.13%、qoq-65.51%），归母净利 2799.47 万元（yoy+4764.41%、qoq-83.27%），扣非净利 1737.72 万元（yoy+343.16%）。公司收入边际改善，同比增速较 23Q4 提升 2.27pct，降本增效推动利润恢复。关注下游需求恢复情况与 AI 商业化推进情况。我们预计公司 2024-2026 年 EPS 分别为 0.97、1.11、1.29 元，可比公司 24 年一致预期 PE 均值 43.4x，给予公司 24 年 43.4 倍 PE，目标价 41.97 元，维持“买入”评级。 风险提示：产品拓展不及预期，下游需求不及预期，市场竞争加剧。 报告发布日期：2024 年 04 月 26 日 点击下载全文：泛微网络(603039 CH,买入)：关注下游需求修复情况
中科创达 (300496 CH)	研发投入加大利润承压，维持“买入”评级 中科创达发布一季报，2024 年 Q1 实现营收 11.78 亿元（yoy+1.01%、qoq-13.75%），归母净利 9075.91 万元（yoy-46.10%、qoq+164.92%），扣非净利 8540.87 万元（yoy-46.23%）。经营性净现金流为 1.64 亿元，同比-50.21%。利润下滑主因研发投入力度加大，经营净现金流下滑主因销售回款减少、职工相关支出及其他经营支付的现金增加。我们预计公司 2024-2026 年 EPS 分别为 1.33、1.75、2.26 元，可比公司 2024E 平均估值 47.2xPE（一致预期），给予公司 2024 年 47.2xPE，对应目标价 62.65 元，维持“买入”评级。 风险提示：下游需求波动；技术落地不及预期。 报告发布日期：2024 年 04 月 24 日 点击下载全文：中科创达(300496 CH,买入)：研发投入加大利润承压，关注端侧智能机遇
网宿科技 (300017 CH)	24Q1 利润高增，卡位 AI 算力基础设施增长可期 网宿科技发布一季报，2024 年 Q1 实现营收 11.20 亿元（yoy-4.13%），归母净利 1.38 亿元（yoy+45.96%），扣非净利 9834.84 万元（yoy+183.45%），经营性净现金流为 2.44 亿元，同比+354.97%。我们认为短期看公司受益于海外市场流量增长推动收入结构优化，长期看卡位 AI 算力基础设施环节，有望持续受益于 AI 产业趋势演进。我们预计公司 2024-2026 年 EPS 分别为 0.25、0.31、0.35 元。可比公司 24E 平均 PE 47.7x（一致预期），给予 24 年 47.7xPE，目标价 12.08 元，维持“买入”评级。 风险提示：海外流量增长不及预期、CDN 行业竞争格局恶化。 报告发布日期：2024 年 04 月 25 日 点击下载全文：网宿科技(300017 CH,买入)：24Q1 利润高增，受益 AI 产业趋势

资料来源：Bloomberg，华泰研究预测

免责声明

分析师声明

本人，谢春生，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并未就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构（以下统称为“华泰”）对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰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。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，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。对依据或者使用本报告所造成的一切后果，华泰及作者均不承担任何法律责任。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。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

本报告版权仅为本公司所有。未经本公司书面许可，任何机构或个人不得以翻版、复制、发表、引用或再次分发他人（无论整份或部分）等任何形式侵犯本公司版权。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并需在使用前获取独立的法律意见，以确定该引用、刊发符合当地适用法规的要求，同时注明出处为“华泰证券研究所”，且不得对本报告进行任何有悖原意的引用、删节和修改。本公司保留追究相关责任的权利。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

中国香港

本报告由华泰证券股份有限公司制作，在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题，请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 天孚通信（300394 CH）、新易盛（300502 CH）、中际旭创（300308 CH）：华泰金融控股（香港）有限公司、其子公司和/或其关联公司实益持有标的公司的市场资本值的 1%或以上。
- 腾讯控股（700 HK）、寒武纪（688256 CH）、海光信息（688041 CH）、金山办公（688111 CH）：华泰金融控股（香港）有限公司、其子公司和/或其关联公司在本报告发布日担任标的公司证券做市商或者证券流动性提供者。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure 其他信息请参见下方“美国-重要监管披露”。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师谢春生本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 神州数码（000034 CH）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日之前的 12 个月内担任了标的证券公开发行或 144A 条款发行的经办人或联席经办人。
- 神州数码（000034 CH）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日之前 12 个月内曾向标的公司提供投资银行服务并收取报酬。
- 天孚通信（300394 CH）、新易盛（300502 CH）、中际旭创（300308 CH）：华泰证券股份有限公司、其子公司和/或其联营公司实益持有标的公司某一类普通股证券的比例达 1%或以上。
- 腾讯控股（700 HK）、寒武纪（688256 CH）、海光信息（688041 CH）、金山办公（688111 CH）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日担任标的公司证券做市商或者证券流动性提供者。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

新加坡

华泰证券（新加坡）有限公司持有新加坡金融管理局颁发的资本市场服务许可证，可从事资本市场产品交易，包括证券、集体投资计划中的单位、交易所交易的衍生品合约和场外衍生品合约，并且是《财务顾问法》规定的豁免财务顾问，就投资产品向他人提供建议，包括发布或公布研究分析或研究报告。华泰证券（新加坡）有限公司可能会根据《财务顾问条例》第 32C 条的规定分发其在华泰内的外国附属公司各自制作的信息/研究。认可投资者、专家投资者或机构投资者使用，华泰证券（新加坡）有限公司不对本报告内容承担法律责任。如果您是非预期接收者，请您立即通知并直接将本报告返回给华泰证券（新加坡）有限公司。本报告的新加坡接收者应联系您的华泰证券（新加坡）有限公司关系经理或客户主管，了解来自或所分发的信息相关的事宜。

评级说明

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（含此期间的股息回报）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，台湾市场基准为台湾加权指数，日本市场基准为日经 225 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准

中性：预计行业股票指数基本与基准持平

减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15% 以上

增持：预计股价超越基准 5%~15%

持有：预计股价相对基准波动在 -15%~5% 之间

卖出：预计股价弱于基准 15% 以上

暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策

无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

新加坡：华泰证券（新加坡）有限公司具有新加坡金融管理局颁发的资本市场服务许可证，并且是豁免财务顾问。公司注册号：202233398E

华泰证券股份有限公司

南京

南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521

电子邮件：ht-rd@htsc.com

深圳

深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017

电话：86 755 82493932/传真：86 755 82492062

电子邮件：ht-rd@htsc.com

北京

北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/

邮政编码：100032

电话：86 10 63211166/传真：86 10 63211275

电子邮件：ht-rd@htsc.com

上海

上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120

电话：86 21 28972098/传真：86 21 28972068

电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司

香港中环皇后大道中 99 号中环中心 53 楼

电话：+852-3658-6000/传真：+852-2567-6123

电子邮件：research@htsc.com

http://www.htsc.com.hk

华泰证券（美国）有限公司

美国纽约公园大道 280 号 21 楼东（纽约 10017）

电话：+212-763-8160/传真：+917-725-9702

电子邮件：Huatai@htsc-us.com

http://www.htsc-us.com

华泰证券（新加坡）有限公司

滨海湾金融中心 1 号大厦，#08-02，新加坡 018981

电话：+65 68603600

传真：+65 65091183

©版权所有 2024 年华泰证券股份有限公司