全球科技动态追踪●计算机行业　　2024 年 4 月 22 日

多重因素引致美股剧烈调整，Meta 发布最新开源大模型 Llama-3

计算机行业

推荐　　维持评级

摘要 | 04.12-04.19

- **股指动态** **美股三大指数、中概股及港股科技、A 股计算机指数全线下跌。**标普 500 指数-3.05%，纳斯达克综合指数-5.52%，费城半导体指数-9.23%；TAMAMA 科技指数-7.02%；纳斯达克中国金龙指数-2.09%；恒生科技指数-5.65%；计算机-3.73%。
- **个股表现** **热门科技股巨幅调整，集体伴随市值大幅缩水；英伟达市值跌破 2 万亿美元。**据统计，相比 4 月 12 日收盘价，4 月 19 日盘后，苹果合计-6.54%，英伟达-13.59%，特斯拉-14.03%，谷歌-2.31%，亚马逊-6.18%，META-6.02%，微软-5.40%，ARM-30.98%，英特尔-4.17%，高通-7.97%，AMD-10.19%。美股七大科技公司（苹果、英伟达、特斯拉、谷歌、亚马逊、Meta、微软）单周总市值合计蒸发 9661.66 亿美元（约合 6.9 万亿人民币），刷新历史记录。
- **10 年期国债及汇率** **美联储发言偏鹰派，10 年期美债持续走弱；美债利率再走高或使科技股阶段性再承压。**周内，美国 10 年期国债利率持续走高，累计上浮 12bps 至 4.62%；中国 10 年期国债利率下降至 2.25%，累计下降 2.97bps。4 月 19 日，美元兑人民币中间价报 7.10；较 4 月 12 日价累计调升 79 个基点。
- **核心观点**

  近期海外部分地区冲突导致地缘政治风险提升、避险情绪加剧，致使风险资产遭抛售。美联储发言偏鹰派，美债利率再走高或使科技股阶段性再承压。受手机及 PC 下游需求预期疲软影响，台积电下调预期，致使全球半导体资金大幅回撤；超微电脑未如期披露业绩预告，引发市场担忧 AI 科技股业绩不及预期风险，从而引发 AI 概念股剧烈回调。**我们认为，上述多重利空因素致使热门科技股巨幅调整。短期来看，四月下旬美股科技将陆续披露季报，在此前的高市场预期背景下，各公司业绩相对市场预期具较大不确定性。**

  Meta 发布最新开源大语言模型 Llama-3，性能较上一代大幅提升、赶超竞品。相较 Llama-2，在参数规模、训练数据集、模型架构及安全性方面均有不同程度提升，**被认为是“首个开源 GPT-4 级别的模型”。Llama-3 可大幅扩充 Meta AI 应用功能，实现与 Facebook、Instagram、WhatsApp 等 App 的无缝集成；此外，网页版 Meta AI 与 OpenAI ChatGPT-3.5 类似，无需注册/登陆，Llama-3 预计打开开源大模型新生态。**
- **风险提示**

  地缘政治风险；技术迭代不及预期风险；科技巨头竞争加剧风险；下游需求不及预期风险。

分析师

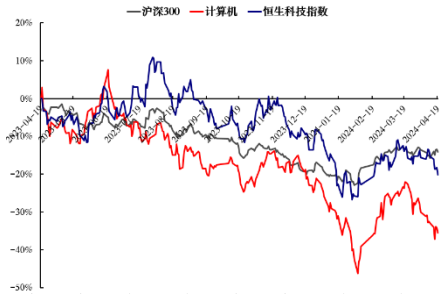
吴砚靖

☎：（8610）66568589

✉：wuyanjing@chinastock.com.cn

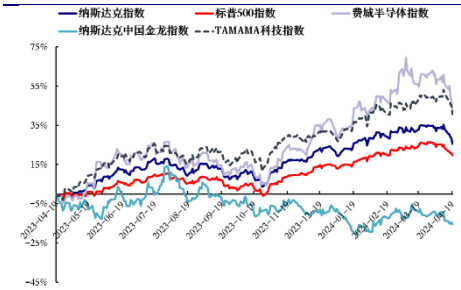
分析师登记编码：S0130519070001

国内表现 2024-04-19



资料来源：Wind，中国银河证券研究院

全球行情 2024-04-19



资料来源：Wind，中国银河证券研究院

相关研究

【银河计算机】全球科技动态追踪_Gaudi 3、MTIA v2 相继发布，AI 产业链加速升级

【银河计算机】全球科技动态追踪_ChatGPT 开放免注册使用，AI 聊天机器人或将颠覆传统搜索引擎

【银河计算机】全球科技动态追踪_OpenAI 公布模型 Voice Engine，AI 安全问题再受关注

目 录

一、全球市场表现

（一）股市动态

近期海外部分地区冲突导致地缘政治风险提升；4 月 19 日，标准普尔全球评级将以色列主权信用评级由 AA-下调至 A+。受此影响，避险情绪加剧，国际基准布伦特原油价格飙升至 90 美元/桶，致使风险资产遭抛售，股市承压。

美股三大指数、中概股及港股科技、A 股计算机指数全线下跌。标普 500 指数-3.05%，纳斯达克综合指数-5.52%，费城半导体指数-9.23%；TAMAMA 科技指数-7.02%；纳斯达克中国金龙指数-2.09%；恒生科技指数-5.65%；计算机-3.73%。

表 1：主要股指周变动

指数代码	指数简称	涨跌幅%					市盈率
		本周	上周	本月	本年度	2023	PE（TTM）
SPX.GI	标普500指数	-3.05	-1.56	-5.46	4.14	24.23	24.82
IXIC.GI	纳斯达克指数	-5.52	-0.45	-6.70	1.80	43.42	39.60
SOX.GI	费城半导体指数	-9.23	-1.54	-12.20	3.15	64.90	43.98
8884057.WI	TAMAMA科技指数	-7.02	1.03	-5.43	6.89	67.81	34.57
HXC.GI	纳斯达克中国金龙指数	-2.09	-3.22	-4.89	-10.17	-3.39	19.06
HSTECH.HI	恒生科技指数	-5.65	0.68	-5.74	-12.92	-8.83	20.40
CI005027.WI	计算机	-3.73	-4.22	-10.31	-18.20	8.90	78.10

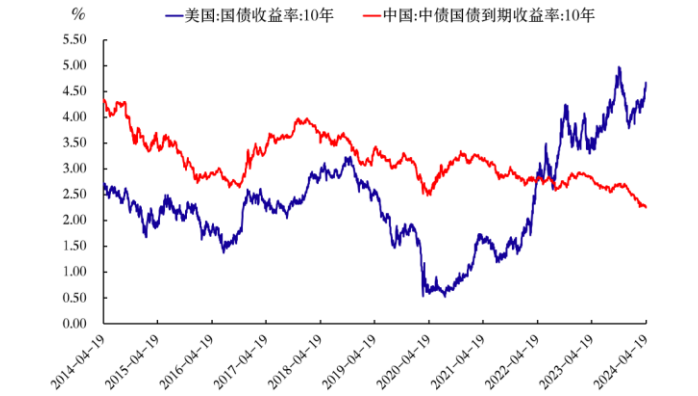
资料来源：Wind，中国银河证券研究院

（二）债市及汇率情况

美联储发言偏鹰派，市场通胀担忧反复，叠加美国经济数据强于预期，10 年期美债持续走弱；考虑利空因素边际影响预期收窄，预计美债利率上涨趋势放缓，但仍在底部区间内震荡。美债利率再走高或使科技股阶段性再承压。周内，美国 10 年期国债利率持续走高，累计上浮 12bps 至 4.62%；中国 10 年期国债利率下降至 2.25%，累计下降 2.97bps。

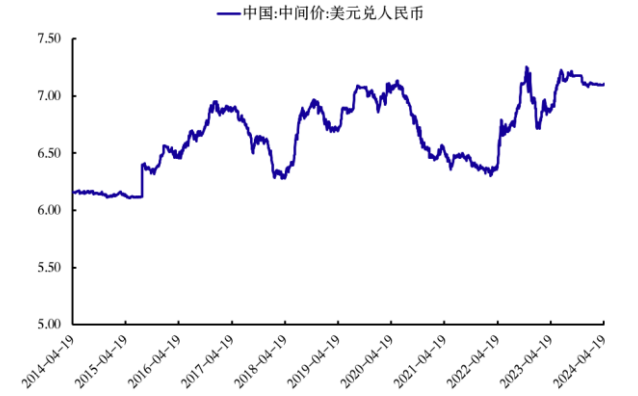
4 月 19 日，美元兑人民币中间价报 7.10；较 4 月 12 日价累计调升 79 个基点。

图1：国债收益率（10 年期）



资料来源：Wind，中国银河证券研究院

图2：美元兑人民币汇率(中间价)



资料来源：Wind，中国银河证券研究院

（三）重点公司表现

受手机及 PC 下游需求预期疲软影响，台积电下调预期。4 月 18 日，台积电下调全年不含存储器在内的半导体行业增长预期至 10%，下调代工行业的增长预期至 14%~19%，同时车用领域芯片预估从上季度 “增长”调为“衰退”，致使全球半导体资金大幅回撤。

超微电脑未如期披露业绩预告，拖累英伟达及一众科技股股价。超微电脑于 4 月 19 日新闻稿中表示，公司将于 4 月 30 日公布 2024Q3 财季业绩，但未提供业绩预告，而被视为负面信号，引发市场恐慌情绪。据悉，美股科技预计于 4 月下旬起陆续披露季报，超微电脑此举引发市场担忧 AI 科技股业绩不及预期风险，从而引发 AI 概念股剧烈回调；AI 概念股下跌波及 A 股相关概念股，包括光模块、AI 服务器、AI 芯片、交换机及液冷等板块承压。消息当日，超微电脑跌幅达 23.14%；拖累英伟达单日股价暴跌 10.00%，创 2020 年 3 月 16 日以来最大单日跌幅，刷新最大单日跌幅纪录。

表 2：重点公司周数据

所属板块	地区	证券代码	公司名称	股价（美元/港元/新台币）		区间涨跌幅（%）	总市值（亿美元/亿港元/亿新台币）[2024-04-19]	市销率 PS（TTM）	市盈率 PE（TTM）	PE(2022)	PE(2023)	PE(2024E)
				[2024-04-12]	[2024-04-19]							
算力及终端	美股	AAPL.O	苹果(Apple)	176.55	165.00	-6.54	25,479.10	6.61	25.25	20.71	30.87	0.00
		NVDA.O	英伟达(NVIDIA)	881.86	762.00	-13.59	19,050.00	31.27	64.01	36.86	280.04	67.34
		TSLA.O	特斯拉(TESLA)	171.05	147.05	-14.03	4,683.23	4.84	31.23	30.98	52.67	43.45
		HPQ.N	惠普(HP)	28.53	27.81	-2.52	272.12	0.51	7.97	8.24	9.14	0.00
		CSCO.O	思科(CISCO)	48.49	48.32	-0.35	1,956.57	3.42	14.56	16.57	16.28	0.00
		ASML.O	阿斯麦	961.84	859.54	-10.64	3,391.65	12.01	44.11	36.70	34.34	42.57
		AMD.O	超威半导体(AMD)	163.28	146.64	-10.19	2,369.91	10.45	277.51	79.12	278.85	75.95
		INTC.O	英特尔(INTEL)	35.69	34.20	-4.17	1,455.85	2.68	86.20	13.61	125.43	35.57
		QCOM.O	高通(QUALCOMM)	171.29	157.63	-7.97	1,759.15	4.85	22.66	9.51	22.36	18.27
		ARM.O	ARM	126.33	87.19	-30.98	896.38	30.51	1,054.56	0.00	147.03	0.00
		ON.O	安森美半导体(ON SEMICONDUCTOR)	66.64	60.71	-8.90	260.96	3.16	11.95	14.18	16.47	14.00
	港股	0909.HK	明源云	2.35	1.98	-15.74	38.46	2.10	-5.95	-10.65	-8.66	-11.77
		9698.HK	万国数据-SW	6.32	6.02	-4.75	91.77	0.84	-1.91	-21.99	-2.92	-4.96
		1686.HK	新意网集团	2.46	2.42	-1.63	56.61	2.23	6.24	11.66	7.98	0.00
	台股	2330.TW	台积电	818.00	750.00	-8.31	211,085.96	8.66	22.35	11.71	18.05	19.31
		2454.TW	联发科	1,180.00	993.00	-15.85	15,884.26	3.66	20.63	8.46	21.09	0.00
互联网	美股	GOOGL.O	谷歌(ALPHABET)-C	157.73	154.09	-2.31	19,251.99	6.23	25.96	19.04	23.69	22.64
		AMZN.O	亚马逊(AMAZON)	186.13	174.63	-6.18	18,156.61	3.16	59.68	-314.82	51.61	42.53
		META.O	脸书(META PLATFORMS)	511.90	481.07	-6.02	12,264.42	9.09	31.37	13.75	23.27	24.97
		NFLX.O	奈飞(NETFLIX)	622.83	555.04	-10.88	2,391.67	6.85	37.17	29.21	39.40	33.79
		PDD.O	拼多多	115.50	113.49	-1.74	1,507.81	4.31	17.79	22.77	22.94	12.29
		NTES.O	网易	92.15	93.51	1.48	587.75	4.13	14.52	16.34	14.46	14.23
		BIDU.O	百度	97.54	95.05	-2.55	331.31	1.75	11.62	36.42	14.52	10.82
		TCOM.O	携程网	48.85	48.00	-1.74	331.11	5.22	23.43	116.72	17.58	25.86
		BABA.N	阿里巴巴	71.29	69.07	-3.11	1,735.11	1.34	12.46	23.78	18.64	12.81
	港股	9988.HK	阿里巴巴-SW	71.65	66.70	-6.91	13,589.62	1.33	12.32	23.80	18.53	12.63
		0700.HK	腾讯控股	309.60	303.80	-1.87	28,694.98	4.27	22.57	15.17	21.90	16.85
		80700.HK	腾讯控股-R	287.20	282.00	-1.81	26,029.50	4.37	23.12	0.00	21.86	17.24
		9999.HK	网易-S	151.10	142.80	-5.49	4,603.33	4.03	14.18	16.52	13.96	13.88
		9888.HK	百度集团-SW	97.55	92.50	-5.18	2,594.83	1.75	11.58	36.49	14.48	10.76
		89888.HK	百度集团-SWR	90.45	85.75	-5.20	2,353.79	1.79	11.84	0.00	14.50	10.99
		1024.HK	快手-W	49.15	44.35	-9.77	1,926.55	1.53	27.30	-19.96	32.60	14.48
		81024.HK	快手-WR	45.50	41.00	-9.89	1,747.59	1.56	27.85	0.00	32.58	14.76
		9626.HK	哔哩哔哩-W	94.60	85.25	-9.88	359.13	1.44	-6.75	-8.90	-7.41	-17.85
		2518.HK	汽车之家-S	53.05	47.75	-9.99	243.23	3.07	11.72	14.79	12.86	11.12
		9898.HK	微博-SW	64.00	61.45	-3.98	144.57	1.05	5.40	52.13	7.52	5.44
软件及应用	美股	MSFT.O	微软(MICROSOFT)	421.90	399.12	-5.40	29,656.36	13.03	35.93	24.58	38.62	33.75
		SNOW.N	SNOWFLAKE	158.56	145.45	-8.27	486.09	17.32	-58.14	-67.89	-82.25	0.00
		ORCL.N	甲骨文(ORACLE)	121.11	114.88	-5.14	3,157.49	6.01	29.67	32.81	34.08	23.86
		CRM.N	赛富时(SALESFORCE)	294.32	270.37	-8.14	2,622.59	7.52	63.41	91.82	1224.61	0.00
		ADBE.O	奥多比(ADOBE)	474.09	465.02	-1.91	2,083.29	10.45	43.39	32.69	50.01	36.72
		INTU.O	财捷(INTUIT)	621.11	605.17	-2.57	1,694.35	11.23	61.17	52.92	73.39	61.00
		SNPS.O	新思科技(SYNOPSYS)	557.16	510.71	-8.34	779.06	12.71	55.35	49.43	63.63	50.17
		CDNS.O	铿腾电子(CADENCE)	305.85	280.25	-8.37	763.96	18.68	73.38	51.91	71.17	0.00
		ADSK.O	欧特克(AUTODESK)	236.98	216.50	-8.64	463.13	8.43	51.12	81.13	63.29	0.00
		U.N	Unity	25.31	22.89	-9.56	89.14	4.08	-10.79	-9.35	-18.80	0.00
	港股	0020.HK	商汤-W	0.66	0.58	-12.12	194.12	4.76	-2.73	-10.98	-5.46	-4.67
		80020.HK	商汤-WR	0.62	0.54	-12.90	176.09	4.89	-2.81	0.00	-5.46	-4.79
		3888.HK	金山软件	25.70	23.50	-8.56	314.70	3.20	58.99	-5.25	61.69	25.97
		0268.HK	金蝶国际	8.24	7.28	-11.65	261.63	4.10	-112.96	-133.49	-178.41	-835.92
		9878.HK	汇通达网络	30.30	28.95	-4.46	162.86	0.18	32.92	70.68	34.57	24.87
		3650.HK	KEEP	6.33	6.73	6.32	35.38	1.47	2.90	0.00	6.00	-13.18
		0354.HK	中国软件国际	4.63	4.40	-4.97	127.91	0.66	16.25	24.18	22.39	11.48
		1357.HK	美图公司	3.18	2.92	-8.18	132.41	4.34	31.72	57.75	38.62	23.36
		3896.HK	金山云	1.49	1.39	-6.71	52.89	0.68	-2.20	-2.69	-3.18	3.68
		2013.HK	微盟集团	1.37	1.12	-18.25	31.30	1.27	-3.74	-8.29	-9.62	-51.05
		1675.HK	亚信科技	7.20	6.79	-5.69	63.51	0.73	10.80	13.73	13.69	6.40
		2121.HK	创新奇智	4.85	4.11	-15.26	23.22	1.20	-3.61	-31.65	-8.15	-5.81
		2400.HK	心动公司	14.20	13.84	-2.54	66.51	1.78	-72.59	-16.75	-53.48	18.49
		0777.HK	网龙	11.12	10.50	-5.58	55.78	0.69	9.19	9.71	10.84	5.26

资料来源：Wind，中国银河证券研究院

多重利空因素致使热门科技股巨幅调整。据统计，相比 4 月 12 日收盘价，4 月 19 日盘后，苹果合计-6.54%，英伟达-13.59%，特斯拉-14.03%，谷歌-2.31%，亚马逊-6.18%，Meta-6.02%，微软-5.40%，ARM-30.98%，英特尔-4.17%，高通-7.97%，AMD-10.19%。

美股科技集体伴随市值大幅缩水；英伟达市值跌破 2 万亿美元。美股七大科技公司（苹果、英伟达、特斯拉、谷歌、亚马逊、Meta、微软），单周总市值合计蒸发 9661.66 亿美元（约合 6.9 万亿人民币），刷新历史记录。周内，英伟达市值缩水近 3 千亿美元，市值跌破 2 万亿美元。

短期来看，四月下旬美股科技将陆续披露季报，在此前的高市场预期背景下，各公司业绩相对市场预期具较大不确定性。

表 3：美股七大科技公司总市值变化情况（相对前一周）

公司	总市值(亿美元)		绝对值变化（亿美元）
	2024-04-12	2024-04-19	
苹果	27262.64	25479.10	-1783.54
英伟达	22046.50	19050.00	-2996.50
特斯拉	5447.58	4683.23	-764.35
谷歌	19694.94	19251.99	-442.96
亚马逊	19352.28	18156.61	-1195.68
META	13050.40	12264.42	-785.98
微软	31349.01	29656.36	-1692.65
总计			**-9661.66**

资料来源：Wind，中国银河证券研究院

二、行业要闻

（一）算力及终端

【SK 海力士与台积电达成战略合作，共同开发 HBM4 芯片】

SK 海力士与 2023 年 4 月 19 日宣布公司就下一代 HBM 产品生产和加强整合 HBM 与逻辑层的先进封装技术，将与台积电公司密切合作。双方近期签署了谅解备忘录。公司计划与台积电合作开发预计在 2026 年投产的 HBM4，即第六代 HBM 产品。

两家公司将首先将致力于针对搭载于 HBM 封装内最底层的基础裸片（Base Die）进行性能改善。HBM 是将多个 DRAM 裸片（Core Die）堆叠在基础裸片上，并通过 TSV1 技术（在 DRAM 芯片打上数千个细微的孔，并通过垂直贯通的电极连接上下芯片的技术）进行垂直连接而成。基础裸片也连接至 GPU，起着对 HBM 进行控制的作用。**此外，双方将协力优化 SK 海力士的 HBM 产品和台积电的 CoWoS2（**台积电独有的制程工艺，是一种在称为硅中阶层（Interposer）的特殊基板上搭载并连接 GPU、xPU 等逻辑芯片和 HBM 的封装方式。其技术在 2D 封装基板上集成逻辑芯片和垂直堆叠（3D）的 HBM，并整合成一个模组，因此也被称为 2.5D 封装技术**）技术融合，共同应对 HBM 相关客户的要求。**

【甲骨文宣布在未来 10 年在日本投资超 80 亿美元建设人工智能基础设施】

据路透社 4 月 17 日报道，甲骨文表示将在未来 10 年内投资超过 80 亿美元，以满足日本对云计算和人工智能基础设施的需求。甲骨文在一份声明中表示，最新的投

资将加强该公司在日本的云计算服务—甲骨文云基础设施（OCI）的服务能力和覆盖范围。甲骨文还将扩大其运营范围，并通过雇佣日本人员支持工程团队。

【微软计划到 2024 年底积累 180 万枚人工智能芯片】

据 Business Insider 4 月 18 日消息，**微软计划到 2024 年底积累 180 万枚人工智能芯片，将其拥有的 GPU 数量增加两倍。**

Business Insider 称，从当前财年到 2027 财年，微软预计将在 GPU 和数据中心上花费约 1000 亿美元。据 Business Insider 援引的 DA Davidson 的分析师估计，微软去年在英伟达芯片上花费了 45 亿美元，一位微软高管表示，这个数字与微软的实际支出大致相符。

微软内部正在努力设计自己的人工智能芯片，以减少对英伟达的依赖，但一些员工持怀疑态度，因为微软落后英伟达多年，而且技术进步如此之快。

【AMD 发布 AI PC 专用处理器：Ryzen Pro 8040 和 Ryzen Pro 8000】

芯片设计商 AMD 于 4 月 16 日推出了一系列用于人工智能商务笔记本电脑和台式机（AI PC）的新处理器：Ryzen Pro 8040 系列（笔记本处理器）、Ryzen Pro 8000 系列系列（台式机处理器）。AMD 称之为商用 PC 迄今为止最强大的芯片，这两个系列的芯片都采用了 4nm 的生产工艺。

据悉，这些新的 AMD 芯片将从 2024 年第二季度开始，为包括惠普和联想等品牌的 PC 机型提供支持。据 CNBC 援引美国咨询公司 Gartner 估计，到 2024 年，AI PC 的出货量将占所有 PC 的 22%。该研究公司预计，到今年年底，AI PC 的出货量将达到 5450 万台。

【Mentee Robotics 展示了旗下首款人形机器人的原型 Menteebot】

据 The Robot Report2024 年 4 月 17 日报道，初创公司 Mentee Robotics 展示了旗下首款人形机器人的原型 Menteebot，号称在所有操作层都接入了 AI，是“可以被指导的”个性化 AI 机器人。

Menteebot 能够使用人工智能来理解自然语言命令。该机器人的运动基于一种新的机器学习方法，称为模拟到现实（Sim2Real）。Mentee Robotics 声称，该机器人的目标市场有两个：一是是家庭应用市场，够执行一系列任务，包括餐桌布置、餐桌清理、衣物处理；二是仓库应用，仓库自动化机器人能够有效地定位、检索和运输物品，并能够处理重达 25 公斤（55 磅）的负载。

Mentee Robotics 表示，它计划 2025 年第一季度之前发布生产完毕的原型机器人。

图3：Menteebot 机器人

资料来源：Mentee Robotics 官网，中国银河证券研究院

（二）大模型及云应用

【Meta 公布其最新大语言模型 Llama-3】

Meta 公布其最新大语言模型 Llama 3，该模型被认为是“首个开源 GPT-4 级别的模型”。Llama-3 目前有两个参数型号，分别为 8B 和 70B：1）Llama-3-8B：拥有 80 亿个参数，专门为需要快速推理和较少计算资源的应用场景设计，同时也能保持较高的性能标准；**2）Llama-3-70B：**拥有 700 亿个参数，能够处理更为复杂的任务，提供更深入的语言理解，有更强的生成能力。Meta 声称再未来几个月会推出具有新功能、更长上下文窗口和增强的性能的模型，并且会发布 Llama-3 的技术报告。

Llama-3 与 Llama-2 相比在多个方面有显著提升。1）在参数规模上，两种不同规模的参数模型能够提供更丰富更精确的应用场景；2）Llama-3 的训练数据集比 Llama-2 扩大了 7 倍，超过 15Ttoken，其中包含比 Llama-2 多 4 倍的代码数据，使得 Llama-3 再代码方面表现出色，且为了即将到来的多语言应用，Llama-3 预训练数据集的 5% 以上由涵盖 30 多种语言的高质量非英语数据组成；3）在模型架构上，Llama-3 采用更高效的分词器和分组查询注意力（Grouped Query Attention，GQA）技术，提高了模型的推理效率和处理长文本的能力；4）在安全性上，Llama-3 引入了 Llama Guard 2，以及 Code Shield 和 CyberSec Eval 2 等安全工具，增强了模型的安全可靠性。

Llama-3 在性能上优于同等级参数模型。经过指令微调后的 Llama-3-8B 模型在 MMLU、GPQA、HumanEval、GSM-K 和 MATH 等测试中优于 Gemma 7B、Mistral 7B。微调后的 Llama-3-70B 在 MMLU、GPQA、HumanEval、GSM-K 和 MATH 等测试中优于 Gemini Pro 1.5 和 Claude 3 Sonnet。

图4：Llama-3 性能测试对比图

Meta Llama 3 Instruct model performance

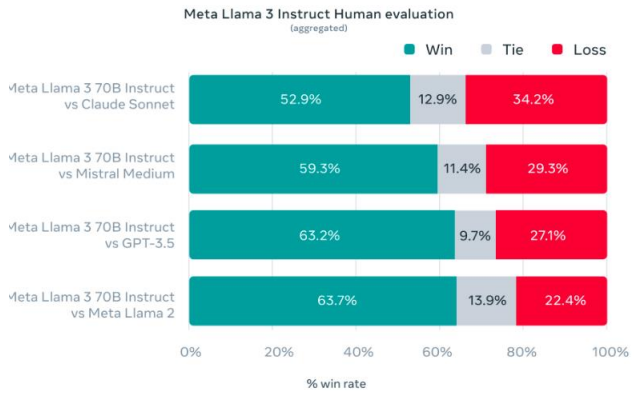
	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

	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，中国银河证券研究院

Meta 开发了一套高质量评估集。该评估集包含 1,800 个提示，涵盖 12 个关键用例：寻求建议、头脑风暴、分类、封闭式问答、编码、创意写作、提取、栖息角色/角色、开放式问答、推理、重写和总结。为了防止模型在这个评估集上意外过拟合，建模团队也无法访问。**通过对 Claude Sonnet、Mistral Medium 和 GPT-3.5 上测试，Llama-3 表现优异，均已超过 50%的胜率胜出其他模型。**

图5：Llama-3 和同类模型在人类评估集上的测试结果



资料来源：Meta 官网，中国银河证券研究院

Llama-3 的技术架构基本情况。1）采用解码器（decoder-only）架构，标准 Transformer 模型架构；2）Llama-3 采用了具有 128K 个 token 的分词器，使其拥有更强的性能；3）采用 GQA 技术，该技术将注意力机制中的查询分组，减少了计算量的同时保持了性能；4）Llama-3 支持 8192 个 token 的序列，并且使用掩码技术确保注意力不会跨越文档边界；5）Llama-3 可以使用指令微调提升模型在特定任务上的表现。

Llama-3 可大幅扩充 Meta AI 应用功能，实现与 Facebook、Instagram、WhatsApp 等 App 的无缝集成；此外，网页版 Meta AI 与 OpenAI ChatGPT-3.5 类似，无需注册（登陆），Llama-3 预计打开开源大模型新生态。

【Meta、USC、CMU 和 UCSD 推出无限上下文架构：Megalodon】

本周，Meta、USC（南加州大学）、CMU 和 UCSD 的研究人员突出了具有无限上下文长度的神经网络架构：Megalodon（巨齿鲨）。

Megalodon 有望解决 transformer 有限上下文的难题。现在普遍使用的 transformer 架构在处理上下文时，会受到二次复杂度，以及长度外推能力的限制。虽然现已有次二次方解决方案，但它们在预训练效率和下游任务准确性方面的经验表现还不如 Transformer。**Megalodon 基于 Mega 的架构进行了改进，增加了数个技术组件：**1）复杂指数移动（CEMA）组件，可以增强模型处理复杂数据的能力；2）时间步长归一化层：将传统的组归一化技术扩展到自回归序列建模任务中，允许模型在处理序列数据时，进行有效的归一化；3）为了增强大规模 LLM 预训练的稳定性，提出了将归一化注意力，和带有两跳残差的预归一化相结合的配置。

在具体表现上。在与 Llama2 的受控正面比较中，Megalodon 在 70 亿个参数和 2 万亿个训练令牌的规模上实现了比 Transformer 更好的效率。Megalodon 的训练损失为 1.70，处于 Llama2-7B（1.75）和 13B（1.67）之间。

【OpenAI 开设东京中心，添加针对日本优化的 GPT-4 模型】

OpenAI 在 2024 年 4 月 14 日宣布在东京开设了第一家亚洲办事处，在日本开展业务。

Techcrunch 在报道中强调 OpenAI 此举意义重大。1）在扩张过程中，OpenAI 可能需要将其技术本地化，适配不同的语言环境；2）随着人工智能的利弊成为政府、监管机构以及公众讨论的焦点，对 OpenAI 来说，切实了解并影响这些对其有利的趋势变得尤为重要。

针对日本优化后的 GPT-4 模型相比之前性能更佳，运行速度是之前的 3 倍，且增强了对日语中细微差别的理解，包括文化理解等，这对于 GPT-4 的商业应用很重要。目前，OpenAI 正在向一些日本本地企业提供对 GPT-4 自定义模型的早期访问权限，并且“在未来几个月内”通过 OpenAI API 逐步开放访问权限。

三、风险提示

地缘政治风险；技术迭代不及预期风险；科技巨头竞争加剧风险；下游需求不及预期风险。

图表目录

表格目录

分析师承诺及简介

本人承诺以勤勉的执业态度，独立、客观地出具本报告，本报告清晰准确地反映本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告的具体推荐或观点直接或间接相关。

吴砚靖，TMT/科创板研究负责人。北京大学软件项目管理硕士，10 年证券分析从业经验，历任中银国际证券首席分析师，国内大型知名 PE 机构研究部执行总经理。具备一二级市场经验，长期专注科技公司研究。

免责声明

本报告由中国银河证券股份有限公司（以下简称银河证券）向其客户提供。银河证券无需因接收人收到本报告而视其为客户。若您并非银河证券客户中的专业投资者，为保证服务质量、控制投资风险、应首先联系银河证券机构销售部门或客户经理，完成投资者适当性匹配，并充分了解该项服务的性质、特点、使用的注意事项以及若不当使用可能带来的风险或损失。

本报告所载的全部内容只提供给客户做参考之用，并不构成对客户的投资咨询建议，并非作为买卖、认购证券或其它金融工具的邀请或保证。客户不应单纯依靠本报告而取代自我独立判断。银河证券认为本报告资料来源是可靠的，所载内容及观点客观公正，但不担保其准确性或完整性。本报告所载内容反映的是银河证券在最初发表本报告日期当日的判断，银河证券可发出其它与本报告所载内容不一致或有不同结论的报告，但银河证券没有义务和责任去及时更新本报告涉及的内容并通知客户。银河证券不对因客户使用本报告而导致的损失负任何责任。

本报告可能附带其它网站的地址或超级链接，对于可能涉及的银河证券网站以外的地址或超级链接，银河证券不对其内容负责。链接网站的内容不构成本报告的任何部分，客户需自行承担浏览这些网站的费用或风险。

银河证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。银河证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

评级标准

评级标准		评级	说明
评级标准为报告发布日后的 6 到 12 个月行业指数（或公司股价）相对市场表现，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，北交所市场以北证 50 指数为基准，香港市场以摩根士丹利中国指数为基准。	行业评级	推荐：	相对基准指数涨幅 10%以上
		中性：	相对基准指数涨幅在-5%～10%之间
		回避：	相对基准指数跌幅 5%以上
	公司评级	推荐：	相对基准指数涨幅 20%以上
		谨慎推荐：	相对基准指数涨幅在 5%～20%之间
		中性：	相对基准指数涨幅在-5%～5%之间
		回避：	相对基准指数跌幅 5%以上

联系

中国银河证券股份有限公司研究院

深圳市福田区金田路 3088 号中洲大厦 20 层

上海浦东新区富城路 99 号震旦大厦 31 层

北京市丰台区西营街 8 号院 1 号楼青海金融大厦

公司网址：www.chinastock.com.cn

机构请致电：

深广地区：程曦 0755-83471683chengxi_yj@chinastock.com.cn
苏一耘 0755-83479312suyiyun_yj@chinastock.com.cn

上海地区：陆韵如 021-60387901luyunru_yj@chinastock.com.cn
李洋洋 021-20252671liyangyang_yj@chinastock.com.cn

北京地区：田薇 010-80927721tianwei@chinastock.com.cn
唐嫚羚 010-80927722tangmanling_bj@chinastock.com.cn