开源Llama 3.1发布：对端云AI的影响

行业研究 · 行业专题

计算机 · 人工智能

投资评级：优于大市（维持）

证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

摘要

国信证券 GUOSEN SECURITIES

- **Llama 3.1发布，开源大模型王者易主。**7月24日报道，美国科技巨头Meta推出迄今为止性能最强大的开源大模型——Llama 3.1 405B（4050亿参数），同时发布了全新升级的Llama 3.1 70B和8B模型版本；Meta评估了超150个基准数据集的性能，Llama 3.1 405B在代码生成和评估、数学推理、长上下文处理、工具使用和多语言支持等一系列任务中，可与GPT-4o、Claude 3.5 Sonnet和Gemini Ultra相媲美；在其他场景中，Llama 3.1 405B进行了与人工评估的比较，其总体表现优于GPT-4o和Claude 3.5 Sonnet。

- **开源引领，加速构建META生态。**与闭源模型不同，Llama 3.1是公开可用的模型，模型的权重可供下载；Llama 3.1开源使得更广泛的开发者及社区可以为其应用程序定制模型，并在任何环境中运行，包括本地服务器、云端、笔记本电脑、甚至手机等，同时无需将数据分享给Meta。同时，Meta透露，其更新了许可证，允许开发人员首次使用包括405B参数规模的Llama模型的输出来改进其他模型。

- **未来预期：转向MOE结构，落地三种商业模式。**MoE（混合专家模型）是一种基于Transformer架构的模型，旨在提高模型的计算效率和性能。其基本思想是通过多个“专家”网络（子模型）协同工作，根据输入数据的特征动态选择最合适的专家，从而优化计算资源的使用和模型的预测精度。基于Meta的商业模式，我们认为Llama 3.1在未来将有以下商业化落地模式： 1）云厂商使用费用：谷歌、亚马逊等下游云服务商提供基于Llama 3.1模型的服务，Meta将从中收取部分费用； 2）通过Meta生态间接变现：在Meta开发的Facebook、Instagram等软硬件产品上使用基于Llama 3.1模型的AI助手，从而吸引用户在软件内消费；3）广告服务：基于Llama 3.1模型提供广告开发以及精准投放服务，并收取费用。

- **风险提示**：大模型研发进展不及预期，AI应用落地不及预期，AI算力投入不及预期。

Llama 3.1发布：开源模型王者易主

国信证券 GUOSEN SECURITIES

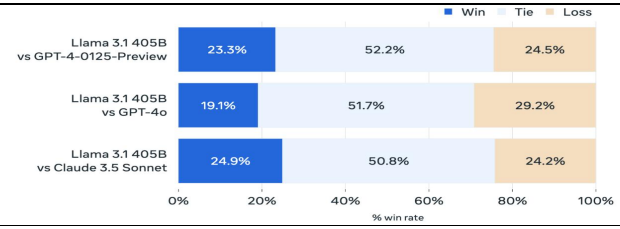
- **7月24日报道，美国科技巨头Meta推出迄今为止性能最强大的开源大模型——Llama 3.1 405B（4050亿参数），同时发布了全新升级的Llama 3.1 70B和8B模型版本。**
- Meta评估了超150个基准数据集的性能，Llama 3.1 405B在代码生成和评估、数学推理、长上下文处理、工具使用和多语言支持等一系列任务中，可与GPT-4o、Claude 3.5 Sonnet和Gemini Ultra相媲美。
- 在其他场景中，Llama 3.1 405B进行了与人工评估的比较，其总体表现优于GPT-4o和Claude 3.5 Sonnet。另外，升级后的Llama 3.1 8B和70B模型，相比于同样参数大小的模型性能表现也更好。
- Llama 3.1 405B支持上下文长度为128K Tokens，增加了对八种语言的支持，在基于15万亿个Tokens、超1.6万个H100 GPU上进行训练，这也是Meta有史以来第一个以这种规模进行训练的Llama模型。
- 与之前的Llama版本相比，Llama 3.1提高了用于训练前和训练后的数据数量和质量。这些改进包括为训练前数据开发更仔细的预处理和管理流程、开发更严格的质量保证以及训练后数据的过滤方法。
- 截至目前，已经有超过25个企业推出了基于Llama 3.1开源版本的新模型。其中，亚马逊AWS、Databricks和英伟达正在推出全套服务，AI芯片创企Groq等为Meta此次发布的所有新模型构建了低延迟、低成本的推理服务，Scale AI、戴尔等公司已准备好帮助企业采用Llama模型并使用自己的数据训练定制模型。国内方面，阿里云、腾讯云已上架 Llama 3.1模型，并支持精调和推理。

图1：Llama 3.1与主流大模型测试对比

Category / Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0125)	GPT-4 Omni	Claude 3.5 Sonnet
General MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	88.7	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	77.0
IFEval	88.6	85.1	84.3	85.6	88.0
Code HumanEval (0-shot)	89.0	73.2	86.6	90.2	92.0
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	90.5
Math GSM8K (8-shot, CoT)	96.8	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	76.6	71.1
Reasoning ARC Challenge (0-shot)	96.9	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	59.4
Tool use BFCL	88.5	86.5	88.3	80.5	90.2
Nexus	58.7	-	50.3	56.1	45.7
Long context ZeroSCROLLS/QuALITY	95.2	-	95.2	90.5	90.5
InfiniteBench/En.MC	83.4	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	100.0	100.0	90.8
Multilingual Multilingual MGSM (0-shot)	91.6	-	85.9	90.5	91.6

资料来源：Meta官网，国信证券经济研究所整理

图2：Llama 3.1 405B模型人类评估测试

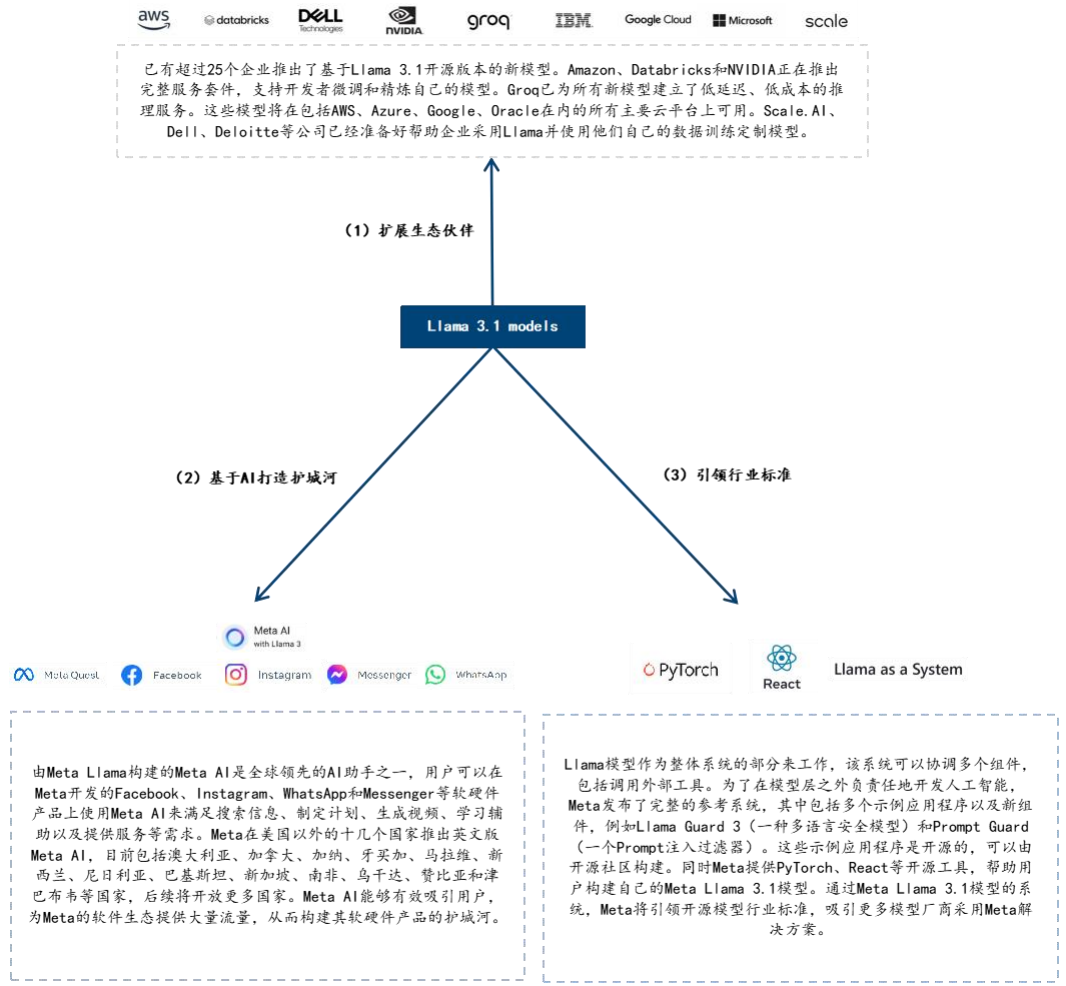


资料来源：Meta官网，国信证券经济研究所整理

开源引领：加速构建META生态

- 与闭源模型不同，Llama 3.1是公开可用的模型，模型的权重可供下载。Meta在llama.meta.com以及Hugging Face上提供下载途径，开发者可以完全根据他们的需求和应用定制这些模型，能够在新的数据集上进行训练，并进行额外的微调。
- Llama 3.1开源使得更广泛的开发者及社区可以为其应用程序定制模型，并在任何环境中运行，包括本地服务器、云端、笔记本电脑、甚至手机等，同时无需将数据分享给Meta。
- 同时，Meta透露，其更新了许可证，允许开发人员首次使用包括405B参数规模的Llama模型的输出来改进其他模型。Meta的商业模式基于为客户打造体验和服务，基于Meta的商业模式，我们认为本次Llama 3.1开源主要由于以下原因：
- 1）不同于闭源模型厂商，Meta的商业模式主要通过在生态里的应用、广告盈利，因此公开发布Llama不会影响Meta的收入、可持续性或研究投资能力，而这些对闭源模型厂商则会有影响；
- 2）Meta的商业模式决定了其必须确保不被锁定在竞争对手的封闭生态系统中，以免限制自身的开发。通过开源吸引大量开发者使用，Llama将发展成完整的生态系统，包括工具创新、效率改进、硬件优化和其他集成，基于Llama开发的AI助手将部署在Meta的软件当中，为用户带来全新体验，从而增加用户粘性，为自身其他产品打造护城河；
- 3）Meta有着长期开源项目的成功经验。曾通过开源数据中心设计从而引领行业标准，从而在建设数据中心时节省数十亿美元，Meta同样希望Llama将成为开源大模型行业的标准，使自身生态系统在未来受益。

图3：Llama 3.1对Meta生态的影响

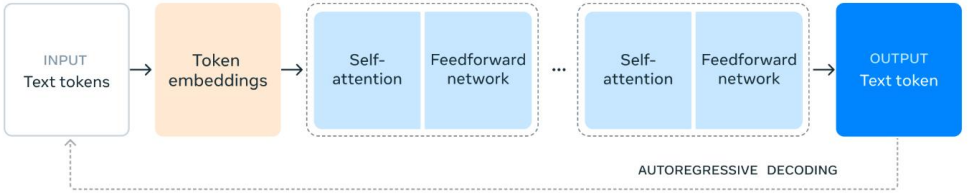


资料来源：Meta官网，Github官网，国信证券经济研究所整理

未来预期：转向MOE结构，落地三种商业模式

- 为训练Llama 3.1模型，Meta做出了以下设计：
- 1）选择了标准的仅解码器的Transformer模型架构，并进行了一些微小调整，而不是使用专家混合模型，以最大化训练的稳定性；
- 2）采用了迭代的后训练程序，每轮使用监督微调和直接偏好优化。这使Meta能够为每轮创建最高质量的合成数据，并提高每项能力的性能；
- 3）与之前的Llama版本相比，Meta改进了用于前训练和后训练的数据的数量和质量，包括开发更仔细的前训练数据预处理和策划管道，开发更严格的质量保证和后训练数据的过滤方法等。
- MoE（混合专家模型）是一种基于Transformer架构的模型，旨在提高模型的计算效率和性能。其基本思想是通过多个“专家”网络（子模型）协同工作，根据输入数据的特征动态选择最合适的专家，从而优化计算资源的使用和模型的预测精度。目前Meta正在准备Llama 4模型，我们认为，随着Scaling Law持续，模型训练参数将持续增加，大幅提高模型训练的硬件需求，未来Meta会更多关注于MoE架构，从而在控制训练成本的前提下获得更强的模型能力。
- 基于Meta的商业模式，我们认为Llama 3.1在未来将有以下商业化落地模式：
- 1）云厂商使用费用：谷歌、亚马逊等下游云服务商提供基于Llama 3.1模型的服务，Meta将从中收取部分费用；
- 2）通过Meta生态间接变现：在Meta开发的Facebook、Instagram等软硬件产品上使用基于Llama 3.1模型的AI助手，从而吸引用户在软件内消费；
- 3）广告服务：基于Llama 3.1模型提供广告开发以及精准投放服务，并收取费用。

图4：Llama 3.1模型架构



资料来源：Meta官网，Github官网，国信证券经济研究所整理

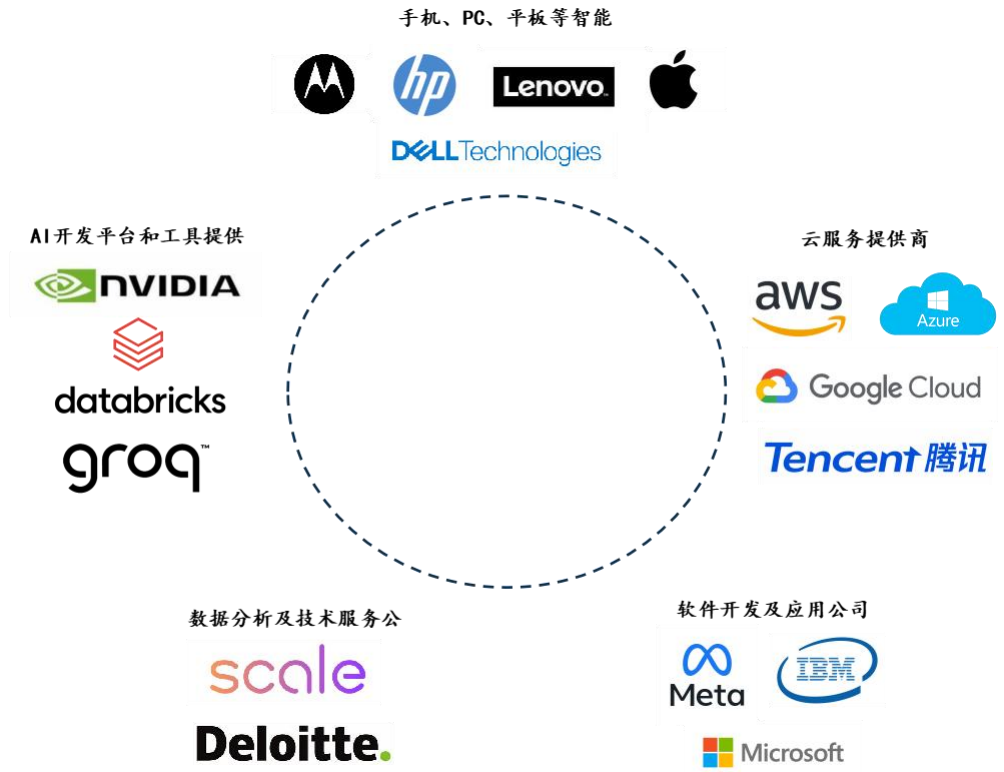
图5：MoE相关研究增长强劲

资料来源：《A Survey on Mixture of Experts》，国信证券经济研究所整理

利好产业：硬件+云+软件全方位受益

- 我们认为，Llama 3.1模型开源将为多个产业提供自主训练AI模型的案例及工具，将首先利好以下产业：
- 1）手机、PC、平板电脑等智能终端厂商：Llama 3.1模型开源为智能终端厂商提供了根据自身需求开发及训练AI智能体的全套工具，有助于打破OpenAI及谷歌等闭源厂商的封锁。受硬件水平限制，预计PC厂商将首先受益，预计明年手机等其他终端达到7-8B的运行环境后也将受益，目前戴尔等终端厂商已准备好采用Llama并使用自己的数据训练自定义模型；
- 2）云服务提供商：Llama 3.1模型将在AWS、Azure、Google Cloud、Oracle、腾讯云以及阿里云等主要云服务提供商的平台可用，云服务商能够利用Llama 3.1模型来增强其 AI 服务，吸引更多的企业客户使用他们的云平台进行AI开发和部署；
- 3）AI开发平台和工具提供商：Amazon、Databricks和NVIDIA等公司正推出完整的服务套件，支持开发者微调和优化Llama 3.1模型。开源模型将刺激下游AI训练需求，并使这些平台能够提供更强大的工具和服务，吸引更多的开发者和企业客户；
- 4）数据分析及技术服务公司：Scale.AI等数据分析公司将帮助企业使用Llama 3.1模型进行数据标注、清洗和分析，提升数据处理和应用的效率。Deloitte等其他咨询公司将利用Llama 3.1模型为企业提供定制化的AI解决方案和技术服务，帮助企业在各自的领域中实现智能化转型；
- 5）软件开发及应用公司：Meta本身以及其他软件公司可通过Llama 3.1模型开发软件内部的AI服务功能，从而增强自身竞争力及用户粘性。

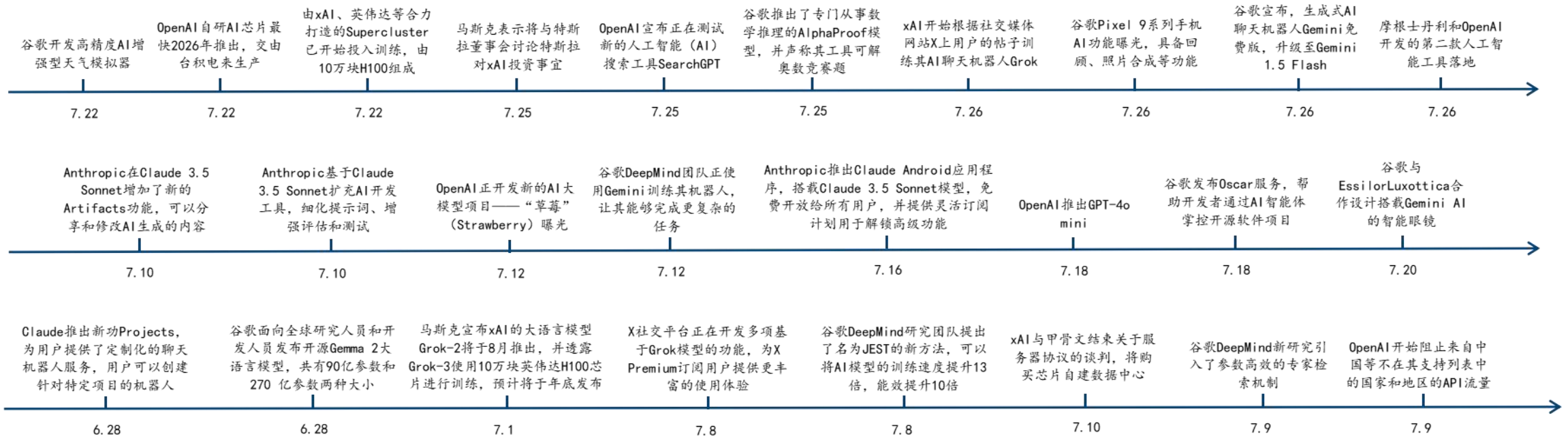
图6：Llama 3.1模型利好多个产业



资料来源：Meta官网，国信证券经济研究所整理

大模型厂商进展：AI竞赛白热化，大模型厂商各有侧重

图7：近期大模型厂商进展



资料来源：OpenAI官网，谷歌官网，xAI官网，Authropic官网，国信证券经济研究所整理

大模型厂商进展：AI竞赛白热化，大模型厂商各有侧重

国信证券 GUOSEN SECURITIES

- OpenAI：当地时间7月25日，OpenAI发布AI搜索产品SearchGPT的原型，目前SearchGPT还未开放公测，仅有1万名用户被邀请。与传统搜索引擎不同，SearchGPT不仅仅集成了实时网络信息，应该也包括类似多步推理的功能，能够按照问题总结相关信息并回答，而不需要用户自己去浏览网页。7月18日，OpenAI推出GPT-3.5 Turbo替代品—GPT-4o mini，即GPT-4o更小参数量的简化版本。ChatGPT的免费用户、Plus用户和Team用户都能够使用GPT-4o mini。GPT-4o mini价格也大幅下降，GPT-4o mini每100万输入token价格为15美分，每100万输出token价格为60美分，比GPT-3.5 Turbo便宜超60%。我们认为，OpenAI目前侧重技术突破及行业引领，重点在办公侧AI应用落地，并通过不断优化价格及模型提高自身竞争能力。

- 谷歌：当地时间7月26日，Android Headline展示了Pixel 9系列手机中的诸多Gemini AI功能，改善用户交互体验，包括手机版的微软Recall功能，可以保存设备屏幕截图，从而满足用户的AI需求。7月26日，谷歌推出了专门从事数学推理的AlphaProof模型，并使用AI拿下IMO奥数银牌。我们认为，谷歌目前侧重于技术商业化及市场占有率，重点在搭载AI的智能终端落地，多款搭载AI模型的产品已在进程当中。

- xAI：当地时间7月26日，马斯克宣布xAI开始根据社交媒体网站X上用户的帖子训练其AI聊天机器人Grok。7月22日，由xAI、X、英伟达等合力打造的由10万块H100组成孟菲斯超级集群（Memphis Supercluster）已经开始启用，为世界上最强大的人工智能训练集群，xAI团队、X团队、英伟达以及其他一些支持公司已开始在该集群上进行训练。我们认为，目前xAI侧重于推进模型训练以及应用落地，重点在于：1）硬件端数据中心建设，为后续模型训练提供基础；2）软件端加速模型迭代，Grok-2以及Grok-3将于本年内推出；3）推进模型应用落地，X平台正开发多项基于Grok模型的功能，包括生成式AI聊天机器人的Grok侧面板、账户总结和高亮文本搜索等功能。

- Authropic：当地时间7月16日，Anthropic推出Claude Android应用程序，搭载Claude 3.5 Sonnet模型，用户可免费访问Anthropic最佳的AI模型Claude 3.5 Sonnet，并通过Anthropic的Pro和Team订阅升级计划。用户将能够在设备间同步他们与Claude的对话，并可以将照片或文件上传到应用程序进行实时图像分析，Claude Android应用程序还包括实时语言翻译功能。7月10日，Anthropic在Claude 3.5 Sonnet增加了新的Artifacts功能，可以分享和修改AI生成的内容。Artifacts功能允许用户将自己制作的游戏或者银承程序发布储蓄，还可以从共享平台上下载其他人制作的内容，并利用AI进行修改。Artifacts并不局限在Claude平台内部，用户可以轻松地将它们分享到任何地方。我们认为，Authropic侧重于提升其AI模型的易用性，通过提供Claude iOS应用程序以及Claude Android应用程序，Authropic目前重点在于提升其应用程序的竞争力，从而吸引更多消费者使用其平台。

风险提示

国信证券 GUOSEN SECURITIES

- **大模型研发进展不及预期**：大模型的发展受制于模型架构、参数量和训练数据量的提升，合法合规的高质量数据获取愈发困难，大模型研发进展可能不及预期；
- **AI应用落地不及预期**：受制于模型成熟度、用户习惯的培养等，AI应用落地可能不及预期，进而反向制约产业对大模型的投入；
- **AI算力投入不及预期**：AI算力是支撑大模型迭代的基石，但其需要大量的资本投入，AI算力投入可能不及预期。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、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意见或建议不一致的投资决策。

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

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032