计算机

证券研究报告
2024 年 09 月 01 日

海外科技巨头季报回顾：Scaling law 不变，变化在推荐算法在内的 AI 场景有望跑通与巨头开始考虑 NV 外的第二选择

微软、Google、Meta 与 Google 等企业发布了新一季度财报，在电话会中对 AI 的表现与 AI 算力需求持续乐观

微软、Google、Meta 与 Google 四家企业陆续发布了新一期的财报，我们认为这 4 家企业在生成式 AI 上取得了较大的进展。微软在 Copilot 与 Azure 云表现出色，目前拥有超 60,000 名 Azure AI 用户，有超过 77,000 家组织采用了 Copilot；Google 计划推出第六代 AI 加速器 Trillium，与 TPUv5e 相比，它的单芯片峰值计算性能提高近 5 倍，能效提高 67%，同时在 Gemini 与 Google Cloud 上也进步较大；Meta 预计在下一代大语言模型相较于 LLaMA3 有 10 倍的算力提升，且 Meta AI、推荐算法与广告系统也取得了进展；Amazon 持续在 Rufus AI 助手等产品上迭代，也在自研芯片上不断加码。在芯片的上游，台积电也对未来 AI 的需求保持了积极乐观的展望。

AI 芯片或迎来更多参与者与行业竞争

苹果在最新用于 Apple Intelligence 的端侧与云端大模型训练中主要使用了 Google 的 TPU，Anthropic 也使用谷歌 Cloud TPU v5e 芯片为其大语言模型 Claude 提供硬件支持；此外，AMD 的 MI300 芯片销量持续超预期，最新季度预测全年收入 45 亿美金，相较于上个季度的 40 亿美金持续上调，微软积极使用了这一款芯片。我们认为除 Nvidia 之外的 AI 芯片或许也会迎来市场机会。

展望未来，训练与推理端的算力需求有望持续增长

训练侧，Meta 推出 LLaMA3.1 405B 模型，在 1.6 万张 H100 基础上完成训练，而 Meta 预计 LLaMA4 模型有望比上一代模型提升 10 倍训练量，我们预计以 Meta 为首的海外互联网公司依然在积极追逐新一代模型。此外我们观察到以 Meta 为首的互联网厂商，有望迎来推荐算法的升级，Meta 在推荐算法上逐步完成了 CPU 到 GPU 的转变，我们预计这也有望带来大量的推理算力需求。

建议关注

（1）四小龙：寒武纪、海光信息、神州数码、中科曙光
（2）华为： 软通动力、烽火通信、广电运通、拓维信息
（3）英伟达：浪潮信息、智微智能[1]

风险提示：AI 算力景气度下降的风险、AI 芯片竞争加剧的风险、政策落地不及预期风险

投资评级

行业评级	强于大市(维持评级)
上次评级	强于大市

作者

缪欣君 分析师
SAC 执业证书编号：S1110517080003
miaoxinjun@tfzq.com
刘鉴 联系人
liujianb@tfzq.com

行业走势图

资料来源：聚源数据

相关报告

1 《计算机-行业深度研究:国产算力 G 端篇：地方智算接力运营商，高景气度预计持续至 25 年底》 2024-08-12
2 《计算机-行业专题研究:算力知识普惠系列一：AI 芯片的基础关键参数》 2024-08-07
3 《计算机-行业点评：“微软蓝屏”引发全球“大宕机”，自主可控有望加速》 2024-07-21

[1] 与通信组联合覆盖

内容目录

1. 海外科技公司季报回顾：Capex 维持高增长，应用与内部提效并举 4
1.1. Google 在 Gemini、谷歌云和 AI 芯片等领域持续发力 4
1.2. 微软 Azure AI 表现优秀，Copilot 继续保持高增长 4
1.3. Meta 在推荐算法与广告系统继续投入、Meta AI 崭露头角 5
1.4. 亚马逊 AWS 自研芯片加速，AI 辅助购物等场景逐步落地 6
1.5. AI 推动大厂 Capex 持续上行，台积电对 AI 景气度中长期乐观 6
2. AI 芯片行业或迎来更多参与者 8
2.1. 苹果训练 Apple Intelligence 大模型积极拥抱 Google TPU 8
2.2. AMD 的 MI300 销量持续超预期 9
3. 展望未来：下一代大模型算力需求将 10 倍增长，推荐算法有望迎来生成式重大革新 9
3.1. Meta 正式发布 LLaMA3.1，正在迈向 LLaMA4 9
3.2. 推荐算法或迎来大升级，从深度学习推荐迈向生成式推荐 11
3.2.1. Advantage+与 Meta Lattice 等颇见成效 11
3.2.2. 从 DLRM 到 GR，生成式推荐算法或迎来突破 12
4. 建议关注 14
5. 风险提示 14

图表目录

图 1：近 9 个季度微软、谷歌、亚马逊及 Meta 资本支出（单位：亿美元）及 CAPEX 总和同比 7
图 2：2024 财年第二季度各平台收入占比 8
图 3：近十季度智能手机与高性能计算营收及同比（单位：亿台币） 8
图 4：2024 财年第二季度各制程工艺收入占比 8
图 5：近十季度各制程工艺收入（单位：亿台币） 8
图 6：AFM-on-device 论文表述 9
图 7：AFM-server 论文表述 9
图 8：405B LLaMA 模型的训练过程 10
图 9： 405B LLaMA 模型的训练集群 GPU 利用效率 10
图 10：LLaMA 3 与其余模型在不同任务下性能对比 10
图 11：Meta Lattice 广告推荐设计简介 11
图 12：深度学习推荐模型与生成推荐系统在特征和训练的对比 13
图 13：推荐算法过去几年训练量逐步增加 13
图 14： 深度学习推荐算法与生成式推荐算法的优劣比较 13

表 1：谷歌旗下产品近期表现 4
表 2：微软旗下 AI 产品近期表现 5
表 3：Meta 旗下 AI 产品近期表现 5
表 4：亚马逊旗下产品 AI 近期表现 6

表 5：Alphabet、微软、Meta 和亚马逊对未来 CAPEX 值的预测 7
表 6：近期 AMD 有关 MI300 的论述 9
表 7：Meta Lattice 特性表述 12

1. 海外科技公司季报回顾：Capex 维持高增长，应用与内部提效并举

1.1. Google 在 Gemini、谷歌云和 AI 芯片等领域持续发力

2024 财年 Q2，Google 在 Gemini、AI 芯片、AI 手机与云服务侧表现亮眼。目前有超 150 万开发人员使用 Gemini，并有超过 20 亿的谷歌直接与间接使用量，该产品得到了广泛的市场认可，Uber 和 WPP 在客户体验和营销等领域使用 Gemini Pro 1.5 和 Gemini Flash 1.5。芯片侧，谷歌推出第六代 AI 加速器 Trillium，与 TPUv5e 相比，它的单芯片峰值计算性能提高近 5 倍，能效提高 67%。手机侧，谷歌推出了新款 Pixel 8a，搭载谷歌最新的 Google Tensor G3 芯片。谷歌云服务侧，谷歌为云客户提供的 AI 基础设施和生成式 AI 解决方案已创造了数十亿美元的收入，并广泛被开发人员使用。此外，谷歌在 Vertex、AI 概览、Circle to search 等业务中均取得了良好进展，正在逐步打开市场。Circle to search 经推出以来已在 1 亿台安卓设备上使用。

谷歌围绕组织效率和结构，以及产品和流程优先级，进而设计成本基础，这反映在其员工人数逐年下降。其将设备和服务产品领域与平台和生态系统产品领域结合起来，以提高速度和效率。近期谷歌已预订 40 万颗 GB200 芯片，布局其 AI 蓝图。

表 1：谷歌旗下产品近期表现

产品	近期表现
Gemini	（1）凭借 200 万个 token，Gemini 拥有最长的上下文支持功能。 目前超过 150 万开发人员使用 Gemini，有超过 20 亿的谷歌直接或间接使用 Gemini。 （2）Uber 和 WPP 在客户体验和营销等领域使用 Gemini Pro 1.5 和 Gemini Flash 1.5。 Gemini for Workspace 帮助 Click Therapy 分析患者反馈，帮助他们构建有针对性的数字治疗方案。 （3）谷歌正使用 Gemini 保护其客户免受网络安全威胁。Wipro 的软件工程师正在使用 Gemini 代码辅助来更快地开发、测试和记录软件。 （4）谷歌将使用 Gemini 在 Gmail 和 Google Photo 推出图片问答功能。
AI 芯片	谷歌推出了第六代 AI 加速器 Trillium，与 TPUv5e 相比，它的单芯片峰值计算性能提高近 5 倍，能效提高 67%。
AI 手机	谷歌推出了新款 Pixel 8a，搭载谷歌最新的 Google Tensor G3 芯片。它提供了出色的 AI 体验，例如 Circle to Search、Best Take 和 Gemini 驱动的 AI 助手。
Google Cloud	谷歌为云客户提供的 AI 基础设施和生成式 AI 解决方案已经创造了数十亿美元的收入，目前存在 200 多万开发人员使用该产品。
Vertex	Vertex 帮助德意志银行、美国空军等客户构建强大的 AI 代理。
AI 概览	（1）AI 概览提升了搜索查询的响应能力，并引入了新的搜索方式。 （2）AI 概览在视觉搜索中也得到了应用，如通过 Lens 进行视频提问。
Circle to search	推出后已在 1 亿台安卓设备上使用。

资料来源：Seeking Alpha，天风证券研究所

1.2. 微软 Azure AI 表现优秀，Copilot 继续保持高增长

2024 财年 Q4，微软在 Copilot 与 Azure 云表现出色，目前拥有超 60,000 名 Azure AI 用户，且已有超过 77,000 家组织采用了 Copilot。微软通过如 Copilot 等 AI 驱动提升了开发者和企业用户的工作效率，从而间接提升公司的整体运营效率，其中已有超过 77,000 家组织采用了 Github Copilot，同比增长 180%。此外，微软继续扩大数据中心覆盖范围，通过推动市场份额的增长，提升了未来的工作效率，Azure 使用了来自 AMD 和英伟达的 AI 加速器以及自研的 Azure Maia 和 Cobalt 100，拥有超过 60000 名 Azure AI 用户，同比增长 60%。Microsoft Fabric 侧，本季度引入了实时智能功能，客户可以对大量时间敏感的数据进行洞察，并累计拥有超 14,000 名付费客户；同时，这个季度微软新推出了全新的

Copilot+电脑。

表 2：微软旗下 AI 产品近期表现

产品	近期具体表现
Copilot	（1）Copilot 为 GitHub 今年的收入增长贡献了 40% 以上，其规模已经超过微软收购 GitHub 时的收入规模。已有超过 77,000 家组织（包括 BBVA、FedEx、H&M、Infosys 和 Paytm）采用了 Copilot，同比增长 180%。 （2）每天在工作中使用 Copilot 的人数比上一季度增长了近一倍。Copilot 客户数量环比增长超过 60%。拥有大于等于 10,000 席位的客户数量环比增长了一倍多，其中包括 Capital Group、迪士尼、陶氏、Kyndryl、诺华。仅安永一家就将为其 150,000 名员工部署 Copilot。 （3）包括嘉年华邮轮公司、Cognizant、伊顿、毕马威、Majesco 和麦肯锡在内的 50,000 家组织已经使用了 Copilot Studio，环比增长超过 70%。将通过 DAX Copilot 将 Copilot 扩展到医疗保健等特定行业。到目前为止，包括 Community Health Network、Intermountain、Northwestern Memorial Healthcare 和俄亥俄州立大学韦克斯纳医学中心在内的 400 多家医疗保健组织已经购买了 DAX Copilot，环比增长 40%，AI 生成的临床报告数量增加了两倍多。
Azure	（1）Azure 使用了来自 AMD 和英伟达的 AI 加速器，以及自研的 Azure Maia 和 Cobalt 100，拥有超过 60000 名 Azure AI 用户，同比增长 60%。 （2）Azure OpenAI 服务提供一流的前沿模型，包括本季度的 GPT-4o 和 GPT-4o mini。 （3）目前拥有 36,000 名 Azure Arc 客户，同比增长 90%。它被各行各业的领先公司所采用，包括 H&R Block、铃木、瑞士再保险公司、Telstra 以及 Freshworks、Meesho 和 Zomato 等数字原生企业。
Microsoft Fabric	（1）Microsoft Fabric 目前拥有超过 14,000 名付费客户，包括埃森哲、Kroger、罗克韦尔自动化和蔡司等各行各业的领导者，环比增长 20%。 （2）本季度， Fabric 引入了全新的实时智能功能，客户可以对大量时间敏感的数据进行洞察。
AI PC	推出了全新 Copilot+ 电脑。

资料来源：Seeking Alpha，天风证券研究所

1.3. Meta 在推荐算法与广告系统继续投入、Meta AI 崭露头角

Meta 在内部推荐算法和广告系统发力并取得一定效果，大模型变现有望实现。2024 财年 Q2，Facebook 方面，Meta 推出了全屏视频播放器和统一视频推荐服务，将 Reels、长视频和直播整合为一，拓展了 Meta 的人工智能系统，并且相较于最初从 CPU 至 GPU 的转变，该举措已经增加了 Facebook Reel 的用户粘性。Meta 希望朝着一个统一的推荐系统发展，这个系统支持提供所有的内容，包括用户可能认识的人。6 月推出的新视频播放器和排名系统表现出色，预计将提高视频推荐的相关性。LLaMA 模型侧，Meta 计划将 LLaMA 4 的计算量提升至 LLaMA 3 的十倍，同时继续开源后续版本。

此外，Meta 正在优化广告插入时间，增加用户会话期间的广告转化率，而不增加广告数量。广告系统方面，Meta Lattice 人工智能架构使广告性能和效率得以提升，且美国的广告客户在使用 Advantage+购物广告系列后，广告支出回报率提升了 22%。最后，Meta AI 自推出以来，用户已进行了数十亿次查询，展示了其广泛的使用和认可。**Meta 通过人工智能上取得的进步缩短产品的开发周期，并通过供应链标准化，节省项目资金花费，提高了资金回报率。**

表 3：Meta 旗下 AI 产品近期表现

产品	近期表现
LLaMA 模型	训练 LLaMA 4 所需的计算量将接近训练 LLaMA 3 的十倍，Meta 将继续开源后续几代 LLaMA
推荐算法	在 Facebook 推出了全屏视频播放器和统一的视频推荐服务，将 Reels、长视频和直播整合为一。这拓展了 Meta 的统一人工智能系统，并且相较于最初 CPU 转变为 GPU，该举措已经增加了 Facebook reel 的用户粘性。Meta 希望朝着统一的推荐系统发展，这个系统支持提供所有的内容，包括用户可能认识的人。
广告系统	（1）广告采用 Meta Lattice 人工智能架构，继续使广告性能和效率提升。美国广告客户在采用 Advantage+ 购物广告系列后，广告支出回报率提高了 22%。从 Advantage+ Creative 使用 LLaMA 3 以来，响应质量得到很大提高。

	（2）在 Facebook 和 Instagram 的广告上，Meta 正在计算用户会话期间的最佳广告插入时间。该项功能可以在不增加广告数量和不减少广告负载下推动收入增长转化。
Meta AI	自首次推出 Meta AI 以来，用户已经使用 Meta AI 进行了数十亿次查询。

资料来源：Seeking Alpha，天风证券研究所

1.4. 亚马逊 AWS 自研芯片加速，AI 辅助购物等场景逐步落地

2024 财年 Q2，亚马逊旗下产品 AWS 在生成式 AI 领域取得显著进展，其提供多层次的解决方案并投资于自家定制芯片。AWS 推出了 Amazon SageMaker，简化了模型构建和部署的过程，并提供了 Amazon Bedrock 中间层服务，Amazon Q 通过其代码转换功能，成功将 30,000 多个 Java JDK 应用程序在几个月内完成迁移，为公司节省了 2.6 亿美元和 4,500 名开发人员的时间。芯片方面，AWS 开发了 Graviton 系列定制硅片，为客户节省了约 30%至 40%的成本，进一步增强了其市场竞争力。Rufus AI 助手则通过辅助购物决策，模拟试穿服装并改善购买体验，提升了用户的购物体验。

表 4：亚马逊旗下产品 AI 近期表现

产品	近期表现
AWS	（1）AWS 在生成式 AI 领域提供多层次解决方案并投资于自家定制芯片（如 Trainium 和 Inferentia）来降低计算成本。 （2）推出 Amazon SageMaker 简化模型构建和部署。AWS 提供了 Amazon Bedrock 中间层服务，拥有广泛的模型选择和强大的生成式 AI 功能。
Amazon Q	Amazon Q 生成式 AI 助手的代码转换功能成功将 30,000 多个 Java JDK 应用程序在几个月内完成迁移，为公司节省了 2.6 亿美元和 4,500 名开发人员的时间。
芯片	开发了 Graviton 系列定制硅片，为客户节省了约 30%至 40%的成本，成为 AWS 的另一个差异化优势。
Rufus AI 助手	Rufus AI 助手辅助购物决策，AI 工具模拟试穿服装和改善购买体验。

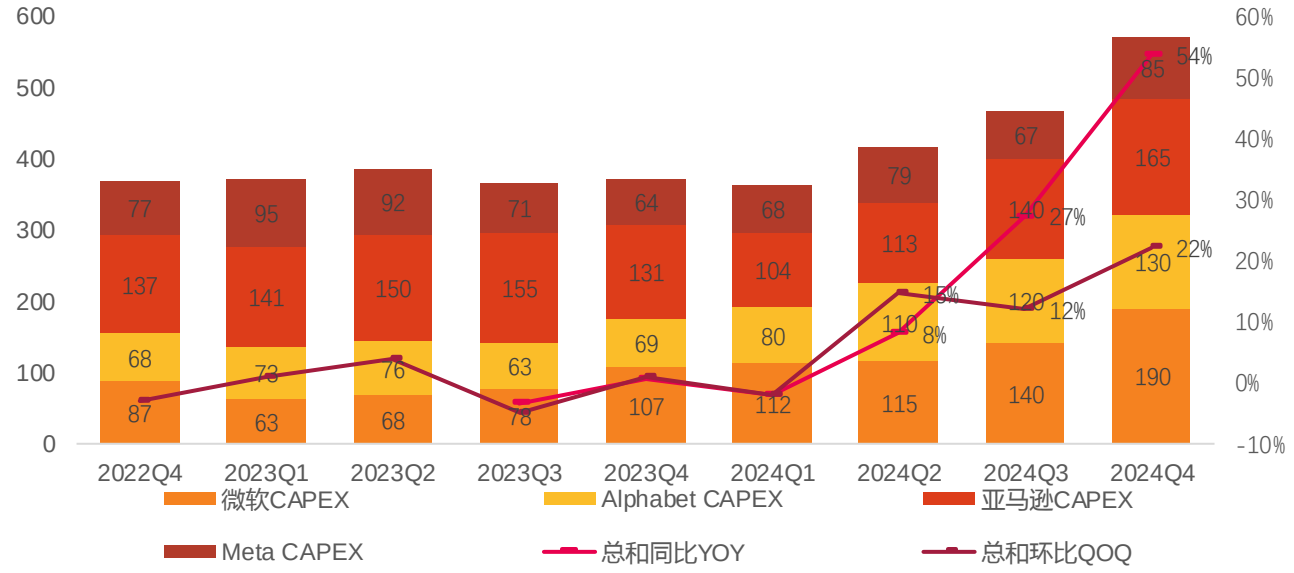
资料来源：Seeking Alpha，天风证券研究所

1.5. AI 推动大厂 Capex 持续上行，台积电对 AI 景气度中长期乐观

在大模型推动下，海外互联网大厂的 Capex 支出同比明显加速增长。从刚刚发布的季报和交流可以看到，最新一个季度谷歌、微软、Meta 和亚马逊 4 家公司的 Capex 总和为 570 亿美金，同比增长 54%，环比增长 22%，且我们观察到无论是同比还是环比增速都处于加速状态。具体来看，Google 最新季度的 CAPEX 为 130 亿美元，预计未来三四季度的 capex 不低于 120 亿美元；微软的 CAPEX 在 2024 年第二季度大幅上涨到 190 亿美元，预计 2025 财年将高于 2024 财年，且 capex 呈环比增长。亚马逊的 CAPEX 先上升再下降后上升，在 2024 财年迅猛上升，在 2024 年第二季度达到 165 亿美元，且预计下半年高于上半年。Meta 的 CAPEX 在 2024 二季度达到 85 亿美元，在二季度预计 24 财年该值处于 370-400 亿水平，而先前预期值为 350-400 亿。

四家互联网公司都在积极的投资 AI 基础设施，近期谷歌继续预订 40 万颗 GB200 芯片，布局其 AI 蓝图， Meta 预计 2025 年的资本支出将大幅增长，主要投资支持人工智能研究和产品开发工作。

图 1：近 9 个季度微软、谷歌、亚马逊及 Meta 资本支出（单位：亿美元）及 CAPEX 总和同比



资料来源：Seeking alpha，天风证券研究所

注：四家公司财年计算方式各异，此处以微软的财年为基准

表 5：Alphabet、微软、Meta 和亚马逊对未来 CAPEX 值的预测

企业名称	Capex 值预估
谷歌	最新一季度 CAPEX 约 130 亿美金，预计未来三四季度的 capex 不低于 120 亿美元
微软	最新一季度 CAPEX 达 190 亿美金，预计 2025 财年将高于 2024 财年，且 capex 呈环比增长
Meta	最新季度的 CAPEX 约 85 亿美金，预计 24 财年该值处于 370-400 亿水平（先前预期为 350-400 亿），预计 2025 年的资本支出将大幅增长，主要投资支持人工智能研究和产品开发工作。
Amazon	上半年资本支出为 305 亿美元，最新季度的 CAPEX 约 172 亿美元，预计下半年高于上半年

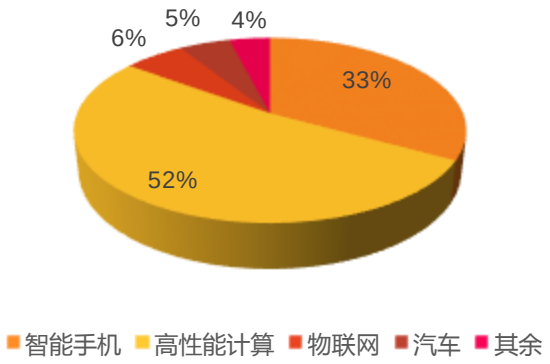
资料来源：Seeking Alpha，公司季报，天风证券研究所

从 AI 芯片上游看，台积电也对中期 AI 的景气度展望乐观。整体业绩上看，得益于市场对台积电领先的 3 纳米和 5 纳米技术的需求，二季度台积电营业收入额为 6735 亿台币，环比增长 13.6%，同比增长 40.1%。此外，台积电上调四年预期，并预计以美元计算下的 2024 年收入增长将略高于 20%。**行业收入侧，2024 第二季度，高性能计算平台收入颇丰，营业收入达 3502 亿台币，同比增长 66%，占比 52%。**回顾近十季度发展，高性能计算业务收入稳步增长；除个别季度外，同比均保持增长态势；在 2024 财年的一二季度同比增长尤其显著。此外，智能手机业务收入可观，营业收入达 2223 亿台币，同比增长 40%，占比 33%

制程工艺侧，2024 第二季度 5 纳米工艺技术贡献 35%的晶圆收入，而 3 纳米和 7 纳米工艺收入分别占总营业收入额的 15%和 17%。近十季度，5nm 制程保持稳定增长，占比逐渐提升，逐步成为台积电收入的主要贡献者。3nm 制程于 2023 年三季度贡献收入，但增长迅速，展现出较强的市场需求和潜力。反观 7nm 制程，其收入和占比均呈现下降趋势。

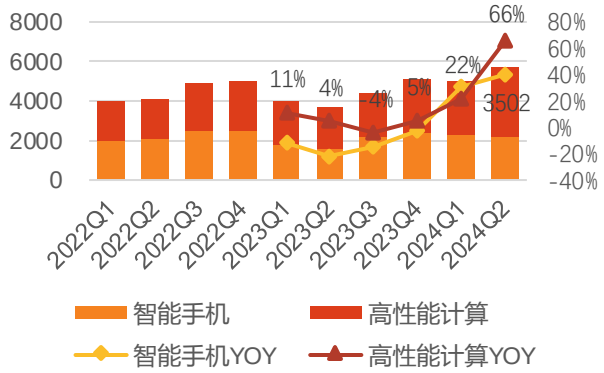
展望未来，台积电预计到 2024 年第三季度，公司整体业务将受到智能手机和人工智能相关需求的强劲支持，展望 2024 年全年，公司预测不包括内存的整体半导体市场将增长约 10%。

图 2：2024 财年第二季度各平台收入占比



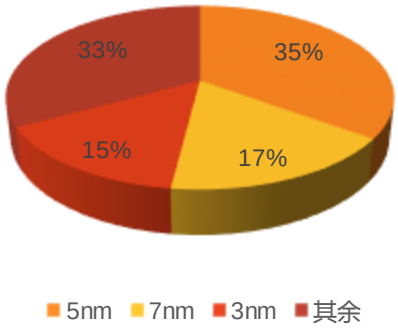
资料来源：台积电公司公告，天风证券研究所

图 3：近十季度智能手机与高性能计算营收及同比（单位：亿台币）



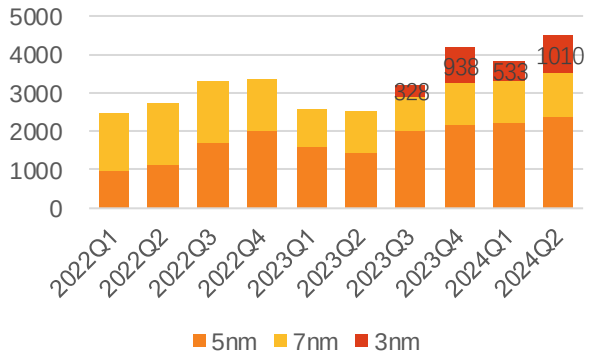
资料来源：台积电公司公告，天风证券研究所

图 4：2024 财年第二季度各制程工艺收入占比



资料来源：台积电公司公告，天风证券研究所

图 5：近十季度各制程工艺收入（单位：亿台币）



资料来源：台积电公司公告，天风证券研究所

2. AI 芯片行业或迎来更多参与者

2.1. 苹果训练 Apple Intelligence 大模型积极拥抱 Google TPU

Apple 公司发布了一篇论文《Apple Intelligence Foundation Language Models》。文中描述了苹果为支持 Apple Intelligence 功能而开发的基础语言模型，包括一个 30 亿个参数的设备端的模型，以及一个为云端大型语言模型。根据论文，其并未采用英伟达 H100 等 GPU 训练 Apple Intelligence 基础模型，转而选择谷歌自研的 TPU。苹果在 TPUv4 和 TPUv5p 集群的硬件上训练两个基础模型：一个是参数规模达到 30 亿的设备端模型 AFM-on-device，使用 2048 块 TPU v5p 训练而成，本地运行在苹果设备上；一个是参数规模更大的服务器端模型 AFM-server，使用 8192 块 TPU v4 芯片训练，运行在苹果自有数据中心里。

此外，Anthropic、Midjourney、Salesforce、Hugging Face 和 AssemblyAI 等知名 AI 创企在大量使用 Cloud TPU。其中，Anthropic 使用谷歌 Cloud TPU v5e 芯片为其大语言模型 Claude 提供硬件支持，以加速模型的训练和推理过程。此外，许多科研、教育机构等也在使用谷歌 TPU 芯片来支持其 AI 相关的研究项目。这些机构可以利用 TPU 芯片的高性能计算能力来加速实验过程，从而推动前沿科研和教育进展。

图 6：AFM-on-device 论文表述

AFM-on-device: For the on-device model, we found that knowledge distillation [Hinton et al., 2015] and structural pruning are effective ways to improve model performance and training efficiency. These two methods are complementary to each other and work in different ways. More specifically, before training AFM-on-device, we initialize it from a pruned 6.4B model (trained from scratch using the same recipe as AFM-server), using pruning masks that are learned through a method similar to what is described in [Wang et al., 2020; Xia et al., 2023]. The key differences are: (1) we only prune the hidden dimension in the feed-forward layers; (2) we use Soft-Top-K masking [Lei et al., 2023] instead of HardConcrete masking [Louizos et al., 2018]; (3) we employ the same pre-training data mixture as the core phase to learn the mask, training for 188B tokens. Then, during the core pre-training of AFM-on-device, a distillation loss is used by replacing the target labels with a convex combination of the true labels and the teacher model's top-1 predictions, (with 0.9 weight assigned to the teacher's labels), training for a full 6.3T tokens. We observe

资料来源：《Apple Intelligence Foundation Language Models》（Apple），天风证券研究所

图 7：AFM-server 论文表述

AFM-server: We train AFM-server from scratch for 6.3T tokens on 8192 TPUv4 chips, using a sequence length of 4096 and a batch-size of 4096 sequences. The batch size was determined using a scaling law fit to model size and compute budget, however we find that downstream results are relatively insensitive to a fairly wide range of batch sizes, and expect that any value between 0.5× and 2× the predicted batch size would have yielded similar results (the predicted optimum was in fact ∼3072, but 4096 allowed for better chip utilization). We perform a learning rate sweep using a proxy model with a model dimension of 768, finding that the optimum learning rate spans 0.01-0.02, so we choose 0.01 to be conservative. Linear layers will have an effective learning rate scaled by ∼0.1 due to the use of μParam (simple).[2]

资料来源：《Apple Intelligence Foundation Language Models》（Apple），天风证券研究所

2.2. AMD 的 MI300 销量持续超预期

AMD 于本季度上调数据中心收入，从四月预测的 40 亿美元增加至 45 亿美元。MI300 在 AMD 的 2024 财年 Q2 季度收入首次超过 10 亿美元，其中微软扩大对 MI300X 加速器的使用，为 GPT-4 Turbo 和多个产品服务（包括 Microsoft 365 Chat、Word 和 Teams）提供支持。微软成为本季度第一家宣布全面推出公共 MI300X 实的大型超大规模企业。

AMD 将持续迭代和推出新 AI 芯片产品，MI325 与 MI350 系列规划表明确。AMD 将于今年末尾推出 MI325X，MI325X 使用了与 MI300 相同的基础设施，并通过提供双倍的内存容量和 1.3 倍的峰值计算性能实现在 Gen-AI 领域的领先。AMD 计划在 2025 年推出基于新 CDNA 4 架构的 MI350 系列，该架构有望将推理性能提高 35 倍。

表 6：近期 AMD 有关 MI300 的论述

季度	关于 MI300 的表述
2023Q4	AMD 为 HPE、戴尔、联想、Supermicro 和其他服务器供应商推出差异化的 MI300 平台；部署 El Capitan 超级计算机购买了大量的 AMD Instinct MI300A 加速器；微软使用 MI300X 训练 GPT-4 预计 MI300X 的产量增长稍缓慢，但现实情况是客户需求强劲，计划于 2024Q1 增加产量；与竞争对手相比，AMD 于 MI300X 提供更多的带宽和内存容量。
2024Q1	MI300 成为 AMD 历史上增长最快的产品，在两个季度内累计总销售额就超 10 亿美元，与 H100 相比，MI300x GPU 提供了领先的推理性能和显着的 TCO 优势。 在云端，MI300x 生产部署已扩展到 Microsoft、Meta 和 Oracle，为内部工作和公开产品提供生成式 AI 训练和推理。
2024Q2	MI300 季度收入首次超过 10 亿美元；微软扩大了对 MI300X 加速器的使用，成为本季度第一家宣布全面推出公共 MI300X 实例的大型超大规模企业。 在 Computex 上宣布 MI300 为首款支持最新 SD 3.0 图像生成 LLM 的 GPU。

资料来源：Seeking Alpha，天风证券研究所

3. 展望未来：下一代大模型算力需求将 10 倍增长，推荐算法有望迎来生成式重大革新

3.1. Meta 正式发布 LLaMA3.1，正在迈向 LLaMA4

Meta 正式发布了最新的开源模型 LLaMA3.1，推出 405B 参数的稠密模型。模型训练侧，Meta 使用了约 15.6T 的 tokens，且这些 tokens 有 8K 的上下文长度，这远超 LLaMA 2 使用的 1.8T 个语料库，LLaMA 3 旗舰模型具有 4050 亿个参数，并使用了 3.8 ×10^25 次浮点运算（FLOPs）进行预训练，超 LLaMA 2 的最大版本近 50 倍。Pre-training 完成后，模型进行了监督微调和 Direct Preference Optimization，在这个过程中模型集成了包括使用工具和代码、推理能力上的增强。在训练侧，405B 模型是在 1.6 万张 H100 的 GPU 上完成，每张 GPU 大概 TDP 为 700W，配备了 80G 的 HBM3,训练时使用了 Meta 的 Grand Teton AI server 平台。

图 8：405B LLaMA 模型的训练过程

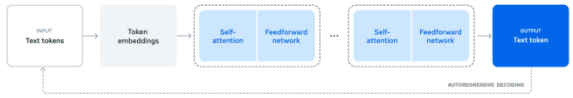


Figure 1 Illustration of the overall architecture and training of Llama 3. Llama 3 is a Transformer language model trained to predict the next token of a textual sequence. See text for details.

资料来源：《The LLaMA 3 Herd of Models》（LLaMA Team、AI@Meta），天风证券研究所

图 9：405B LLaMA 模型的训练集群 GPU 利用效率

GPUs	TP	CP	PP	DP	Seq. Len.	Batch size/DP	Tokens/Batch	TFLOPs/GPU	BF16 MFU
8,192	8	1	16	64	8,192	32	16M	430	43%
16,384	8	1	16	128	8,192	16	16M	400	41%
16,384	8	16	16	4	131,072	16	16M	380	38%

Table 4 Scaling configurations and MFU for each stage of Llama 3 405B pre-training. See text and Figure 5 for descriptions of each type of parallelism.

资料来源：《The LLaMA 3 Herd of Models》（LLaMA Team、AI@Meta），天风证券研究所

LLaMA3.1 是目前最领先的开源大语言模型之一。模型功能侧，其支持多种语言，并在大量任务中的完成质量与先进的语言模型(如 GPT-4)相当。LLaMA 3.1 可以支持 8 种语言（英语，德语、法语、意大利语、葡萄牙语、印地语、西班牙语和泰语），适用于多语言对话智能体和翻译场景等用途；在上下文长度上，比起 LLaMA 2、LLaMA 3，LLaMA 3.1 系列模型中所有上下文增加了 16 倍，为 128K；Meta 强调，LLaMA 3.1 还在工具使用方面得到了改进，支持零样本工具使用，包括网络搜索、数学运算和代码执行基于长上下文，模型不仅知道何时使用工具，还能理解如何使用以及如何解释结果。此外， 通过微调，LLaMA 3.1 在调用自定义工具方面提供了强大的灵活性。

图 10：LLaMA 3 与其余模型在不同任务下性能对比

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**

资料来源：《The LLaMA 3 Herd of Models》（LLaMA Team、AI@Meta），天风证券研究所

在 Meta 最新的业绩交流会上，Meta 预计 LLaMA 4 训练将加倍消耗算力，约为 LLaMA 3 的训练所消耗算力的十倍。鉴于 LLaMA3.1 是基于 16k 的 H100 集群训练而来，预计下一代 LLaMA 4 模型需要的训练集群扩大接近 10 倍。

3.2. 推荐算法或迎来大升级，从深度学习推荐迈向生成式推荐

Meta 在推荐算法和广告系统上不断升级，推出了包括 Advantage+、广告排名架构 Meta Lattice 等在内的应用，此外 Meta 在推荐算法上也在发布了新的论文，推出了生成式推荐并已经正式使用。

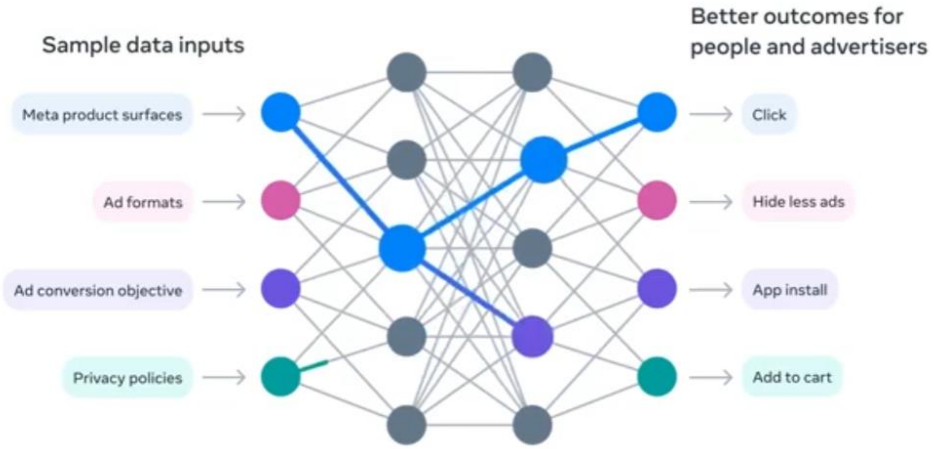
3.2.1. Advantage+与 Meta Lattice 等颇见成效

Meta 为广告主们提供更加自动化的服务。通过 Advantage+ 产品组合，广告商可以自动执行广告系列设置流程的一个步骤。在今年第二季度，Meta 进一步更新 Advantage+ Shopping，广告客户在一个广告中可以上传多张图片和视频，该产品根据广告投放场景与上传资料自动生成广告样式，以达到最佳宣传效果。5 月，Meta 开始在 Advantage+ Creative 中推出完整的图像生成功能。Meta 继续投资 Conversion API，帮助企业与他们的营销数据关联，使其广告效果更佳。

Meta 不断深入理解用户观看广告的偏好,以对正确的对象在合适的时间与场景下有效地展示广告。例如，Meta 正在优化在 Facebook 和 Instagram 的用户会话期间展示广告的内容与时机。

此外，Meta 利用 AI 提高广告营销效果。其正在不断改进广告模型，为广告客户提供更好的效果。具体来说，Meta 正广泛推广其广告排名架构 Meta Lattice，其可以访问颗粒度较小的数据，推动广告客户的表现。此外，该模型可以跨目标和界面概括学习，而非仅针对单一目标和界面进行优化。这说明即使可供学习的数据较少，用户们也可以在界面上收到更相关的广告推荐。

图 11：Meta Lattice 广告推荐设计简介



资料来源：Meta 官网，天风证券研究所

Meta Lattice 可处理延迟性反馈。Meta Lattice 不仅可以从新鲜信号中捕捉到人的实时意图，还可以从缓慢、稀疏和延迟的信号中捕捉到长期兴趣。

Meta Lattice 可以平衡多个领域和目标。其能够平衡多个领域和目标之间的性能，并达到在不损害其他目标的情况下无法进一步改进任何目标的状态（即帕累托最优）。

Meta Lattice 具有高级模型扩展性。Meta Lattice 拥有数万亿个参数，经过数千亿个示例的训练，这些示例来自数千个数据域，包括 Meta 的平台界面和面向广告商的产品。

Meta Lattice 可最大化 AI 资本支出效率。以前数百个模型需要单独训练、服务和优化。目前 Meta 引入了两个级别的资源共享：（1）通过联合优化实现跨领域、跨目标、跨排名阶段的横向共享；（2）从大型高容量上游模型到轻量级下游垂直模型的分层共享。通过资源共享增强，可以显著减少计算需求量。

表 7：Meta Lattice 特性表述

特性	具体描述
跨目标和界面概括学习	模型可以跨目标和界面概括学习，而非仅针对单一目标和界面进行优化。即使可供学习的数据较少，用户们也可以在界面上收到更相关的广告推荐。
处理延迟性反馈	广告与观看广告的人之间的互动时间可以从几秒钟（例如点击、点赞）到几天（例如考虑购买、添加到购物车，然后从网站或应用进行购买）。通过具有时间感知的多分布建模，Meta Lattice 不仅可以从新鲜信号中捕捉到人的实时意图，还可以从缓慢、稀疏和延迟的信号中捕捉到长期兴趣
多领域和多目标平衡	其能够平衡多个领域和目标之间的性能，并达到在不损害其他目标的情况下无法进一步改进任何目标的状态（即帕累托最优）。
高级模型扩展性	Meta Lattice 拥有数万亿个参数，经过数千亿个示例的训练，这些示例来自数千个数据域，包括 Meta 的平台界面和面向广告商的产品。定制深度分层集成网络模型建立在 Transformers 主干之上，在 GPU 上具有高度的可扩展性。
最大化 AI 资本支出效率	以前数百个模型需要单独训练、服务和优化。目前 Meta 引入了两个级别的资源共享：（1）通过联合优化实现跨领域、跨目标、跨排名阶段的横向共享；（2）从大型高容量上游模型到轻量级下游垂直模型的分层共享。通过资源共享增强，可以显著减少计算需求量。

资料来源：Meta 官网，天风证券研究所

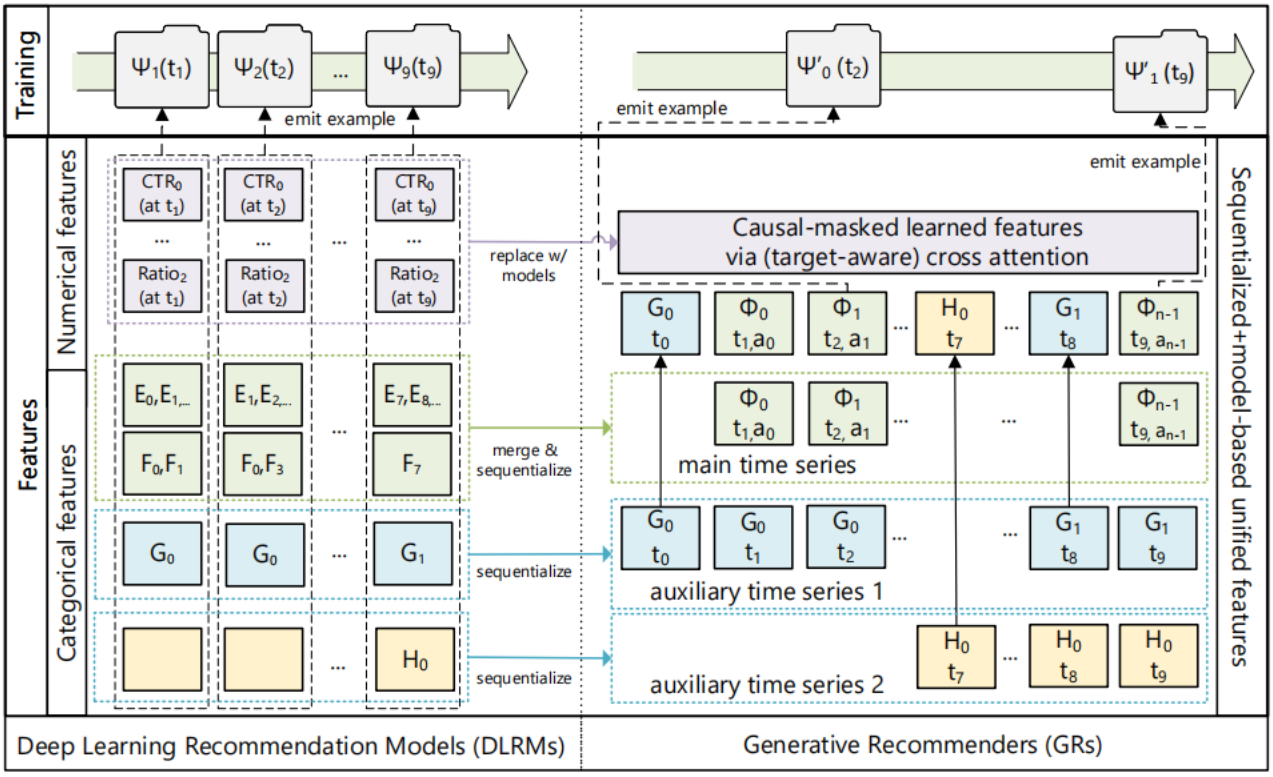
Instagram 上部署 Meta Lattice 的早期结果表明，在 Instagram 的不同界面（例如 Feed、Story 和 Reels）以及各种广告客户目标（例如点击次数、视频观看次数和转化次数）上进行知识共享，提高了约 8%的广告效果。

3.2.2. 从 DLRM 到 GR，生成式推荐算法或迎来突破

Meta 发表了名为《Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations》的论文，提出了一个已经在企业内大规模使用的新一代推荐算法。根据 Meta 公开论文，其研发团队设计基于生成式推荐系统的全新广告推荐架构，将推荐任务重新表述为生成式建模框架内的顺序传导问题。其提出 HSTU（分层顺序传导单元）；HSTU 修改传统的注意力机制，针对常见的大型、非固定的推荐数据进行优化。该模型已成功部署于大型互联网平台的多个层面，并将在线 A/B 测试的关键指标提升 12.4%。基于 HSTU 的模型比现有模型速度更快、效率更高，在合成数据集与公共数据集上，其性能远超传统深度学习推荐模型，在 NDCG 值中高达 65.8%。在传统模型还具有可扩展性，其质量随着训练计算的而呈现幂律分布，而非线性关系。

论文提出 DLRM（深度学习推荐模型）通常将交互视为独立事件，而不是序列的一部分。这使得在处理复杂的用户操作序列方面存在局限性，可能难以应对用户偏好和商品受欢迎程度随时间变化的动态特性。而在 GR（生成推荐系统）中使用的顺序传导方法使模型能够更好地捕捉用户行为的时间动态，并以更结构化的方式对项目之间的依赖关系进行建模。GR 利用 Transformers 等架构，这些架构处理序列表现优异，并且可以生成具有上下文感知的推荐。此外，GR 统一了特征空间，在顺序框架内处理分类和数值特征。这与 DLRM 中对这些特征的传统分离不同。

图 12：深度学习推荐模型与生成推荐系统在特征和训练的对比

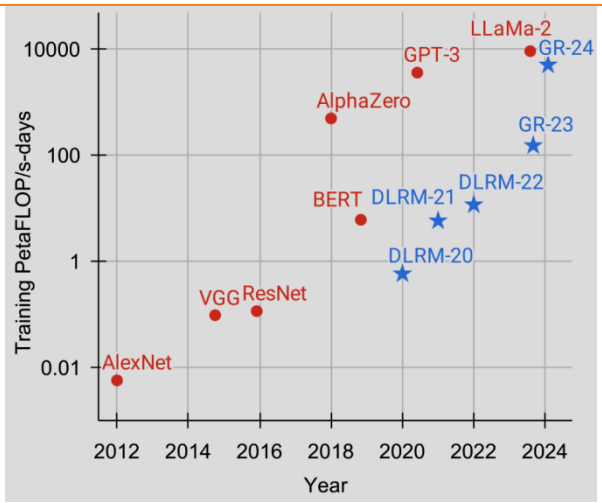


资料来源：《Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations》（Jiaqi Zhai、Lucy Liao 等），天风证券研究所

论文结果指出在低计算状态下，DLRM 可能会优于 GR，这证实了特征工程在传统 DLRM 中的重要性。然而，GR 表现出更好的可扩展性，而 DLRM 则出现瓶颈，与以前的工作中的调查结果一致。且 GR 在嵌入参数与非嵌入参数上都表现出了较好的可拓展性，GR 能够到 1.5 万亿个参数模型，而 DLRM 性能在大约 2000 亿个参数处饱和。

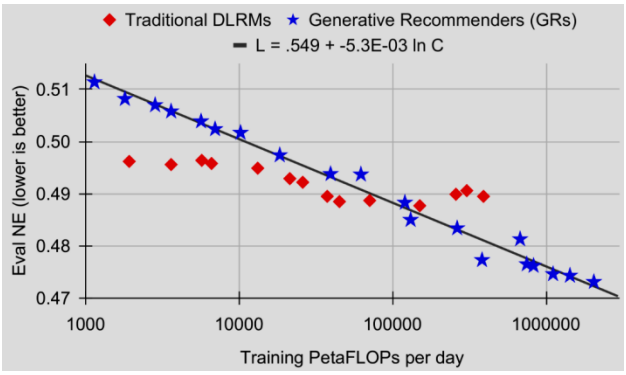
此外，论文指出，在一个合理的范围内，与应用的训练计算总量相比，精确的模型超参数起的作用较小。序列长度在 GR 中起着明显更重要的作用，并且重要的是串联放大序列长度和其他参数。这也许是论文提出的方法最重要的优点，第一次证明了 LLM 的缩放定律也可以应用于大规模推荐系统。

图 13：推荐算法过去几年训练量逐步增加



资料来源：《Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations》（Jiaqi Zhai、Lucy Liao 等），天风证券研究所

图 14：深度学习推荐算法与生成式推荐算法的优劣比较



资料来源：《Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations》（Jiaqi Zhai、Lucy Liao 等），天风证券研究所

4. 建议关注

（1）四小龙：寒武纪、海光信息、神州数码、中科曙光

（2）华为： 软通动力、烽火通信、广电运通、拓维信息

（3）英伟达：浪潮信息、智微智能[2]

5. 风险提示

（1）AI 算力景气度下降的风险

算力支出与下游应用息息相关，若 AI 应用需要更长期才能突破，则算力支出的高景气可能不可持续

（2）AI 芯片竞争加剧的风险

AI 芯片领域有较多参与者，未来市场竞争可能加剧

（3）政策落地不及预期风险

地方政府智算中心主要依靠各地政策推动，如政策落地不及预期，则可能影响智算中心算力建设相关公司

[2] 与通信组联合覆盖

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

本报告是机密的，仅供我们的客户使用，天风证券不因收件人收到本报告而视其为天风证券的客户。本报告中的信息均来源于我们认为可靠的已公开资料，但天风证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，天风证券及/或其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，天风证券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。

天风证券的销售人员、交易人员以及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。天风证券没有将此意见及建议向报告所有接收者进行更新的义务。天风证券的资产管理部门、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。

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到天风证券及/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一参考依据。

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	海口	上海	深圳
北京市西城区德胜国际中心 B 座 11 层	海南省海口市美兰区国兴大道 3 号互联网金融大厦	上海市虹口区北外滩国际客运中心 6 号楼 4 层	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100088	A 栋 23 层 2301 房	邮编：200086	邮编：518000
邮箱：research@tfzq.com	邮编：570102	电话：(8621)-65055515	电话：(86755)-23915663
	电话：(0898)-65365390	传真：(8621)-61069806	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com