大模型进展2.0

行业研究·专题报告

投资评级：超配（维持评级）

报告摘要

- **AI大模型市场表现与竞争格局发生变化，Kimi成为国产大模型曙光。**市场上的大模型层出不穷，以Kimi为代表的产品凭借其在长文本处理领域的卓越能力，迅速成为用户访问量最高的产品，打破了现有竞争格局。Kimi在中文领域对GPT-4、Claude等国际大模型展现出明显优势，并通过不断的技术迭代和用户体验优化，实现了用户流量的激增和市场的快速扩张。公司认为，Kimi的AI-Native产品核心价值在于提供个性化交互，其长文本上下文处理能力(Long Context)能大幅减少模型定制成本，解决90%的模型定制问题。2024年3月下旬，Kimi进一步将上下文处理能力提升至200万汉字，随着用户流量的激增，服务连续进行了5次扩容。公司在C端致力于将Kimi打造成超级应用，成为AI原生交互的入口；在B端，通过Moonshot AI开放平台提供与OpenAI兼容的API，内测期间已有法律、游戏阅读等领域应用进行测试，反馈良好。随着Kimi应用访问量的持续增长，预计将再次拉动算力需求的快速增长，推动AI行业的算力基础设施发展。。

- **随着AI大模型技术的发展和应用场景的拓展，全球算力需求正面临重估。**Meta等科技巨头对AI算力的需求超出预期，预计到2024年底将拥有接近60万颗H100 GPU的等效算力。Sora模型的发布标志着AI视频生成领域的新突破，进一步推动了多模态大模型的发展，预示着未来对算力需求的大幅提升。同时，美国政府的限制措施可能促使中国等国家的企业自行购买算力卡或租赁国产AI算力，推动国产AI产业链的革新和发展。在此背景下，Kimi等国产大模型的成功，不仅带动了产业链的革新，还为内容创作、游戏互动、AI陪伴等领域带来了新的应用场景和创新机遇。此外，Step系列通用大模型的发布和Pixverse在AI视频生成领域的领先地位，进一步展示了国产AI技术的竞争力和市场潜力。

- **投资建议：**1）多模态大模型拉动全球算力需求快速增长，叠加美国将限制云厂商对华客户提供AI云服务，国产AI算力迎来发展机会；2）随着AI大模型成本下降与技术发展，AI应用产业将快速进步，建议关注AI应用相关个股。建议关注金山办公、科大讯飞、同花顺、海光信息。维持计算机行业超配评级。

- **风险提示：**宏观经济波动；下游需求不及预期；AI伦理风险；技术发展不及预期。

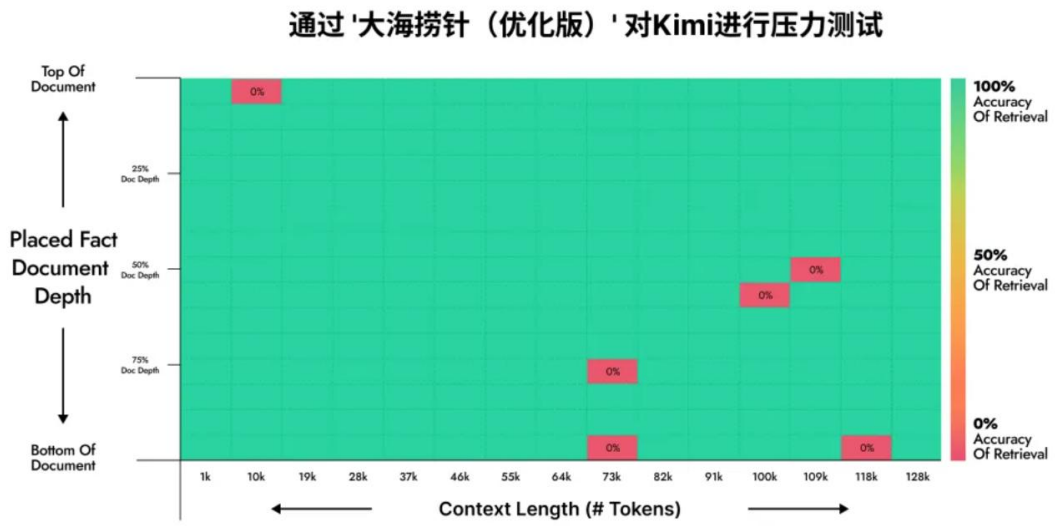
目录

月之暗面发布Kimi，长文本成为破局关键

- **月之暗面成为国产大模型新星。**2023 年10 月，清华大学杨植麟及其AI 团队“月之暗面”发布了Kimi，拥有优秀的长文本处理能力，可处理20万汉字输入。同时，利用Long Context 可以大幅减少 fine-tune 的成本，实现模型应用的“多、快、好、省”。例如可以先用5万字定制一个模型的能力，剩余还有大量文字窗口，也足够日常交互使用。而fine-tune需要构造数据并训练，时间较长且需要较高的复杂度，单位token的成本也更高。公司选择用Long Context方式来解决 90%的问题，更好向前向后兼容，也成为公司最高优先级的技术突破方向。

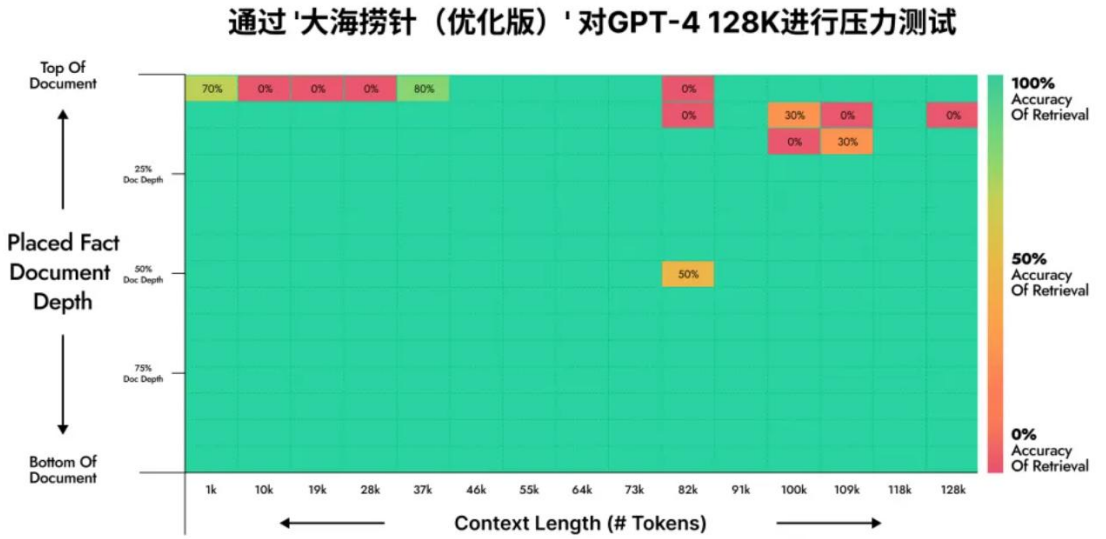
➢ **“大海捞针”测试验证了 Kimi 长文本能力。**长文本能力是实现人类与AI之间无损理解的基础，它使 AI 可以更准确地理解人类的复杂、感性思维，从而在多种应用场景中更有效地服务于人类。根据近一年全球各个大模型迭代方向，上下文窗口的“长文本”再持续升级。其中，在文本持续变长过程中，大型是否会忽略掉部分细节内容的问题一直是“长文本”能力的关键。因此有开发者进行了一项名为“大海捞针”的大模型长文本性能测试，即在文本中加入一句与该文本内容不相关的句子，测试大模型是否能通过Prompt把这句话准确提取出来。月之暗面的工程师在2023年12月也进行了测试，选取模型为Kimi chat(支持20万汉字输入)，GPT-4 Turbo(支持128K上下文窗口)，Claude 2.1(支持200K上下文窗口)。根据测试结果，Kimi chat在“大海捞针”中的表现明显好于GPT-4Turbo和Claude 2.1.

图：Kimi“大海捞针”实验表现



资料来源：Moonshot AI，国信证券经济研究所整理

图：GPT-4 Turbo“大海捞针”实验表现



资料来源：Moonshot AI，国信证券经济研究所整理

联合技术及服务壁垒，Kimi有望重塑竞争格局

- **Kimi通过以下几个核心策略实现了区别于市场的独特定位和快速增长：**

➢ 用户体验中心化：Kimi把用户体验作为产品开发和优化的核心，通过细致了解用户需求，提供流畅、直观的使用体验，提升用户满意度和忠诚度;

➢ 数据驱动的优化：利用用户行为数据，Kimi采用数据驱动的方法持续迭代产品功能，快速适应市场变化，保持技术和服务的领先优势;

➢ 创新的分享机制：引入分享功能增强用户互动，同时利用用户生成的数据和反馈优化模型，形成正向的数据循环，提高模型性能和用户体验。

➢ 专注核心功能优化：专注于提升核心功能如视频高清化等，满足用户特定需求，通过AI技术与用户体验的结合，打造差异化竞争优势。

➢ 避免过度扩张：Kimi选择专注于现有产品的持续优化，避免过度扩张产品线以确保产品和服务的高质量标准。

- **国产大模型在算力受限的背景下能表现如此优秀，主要是因为Kimi实现了AI产品发展中三个关键的scaling要素：模型、人才和用户。**

➢ 模型Scaling：Kimi通过持续优化其A1模型，不断增强模型的处理能力和应用范围，成功地提升了产品的核心竞争力。这种模型的scaling不仅涉及到算法的改进和优化，还包括对大数据的处理能力和学习效率的提升，确保模型能够处理更复杂的任务，满足更广泛的用户需求。

➢ 人才Scaling：注重人才的招聘和培养，扩展人才密度，这对快速推出产品至关重要。

➢ 用户Scaling：Kimi选择专注于c端市场，致力于开发能够覆盖广大用户需求的通用产品，而不是局限于某个B端的垂直领域。这种策略使Kimi能够吸引到足够大的用户规模，通过规模化的用户反馈进一步优化产品，形成了良好的用户增长和产品改进的正向循环。

图：Kimi 可以两分钟读完500份简历，筛选员工

资料来源：国信证券经济研究所整理

图：Kimi 可以读取英伟达报告，并分析财报历史

资料来源：国信证券经济研究所整理

Kimi 打破竞争格局，带动产业链发展

- **Kimi 优秀的性能可以带动多个产业的发展。**

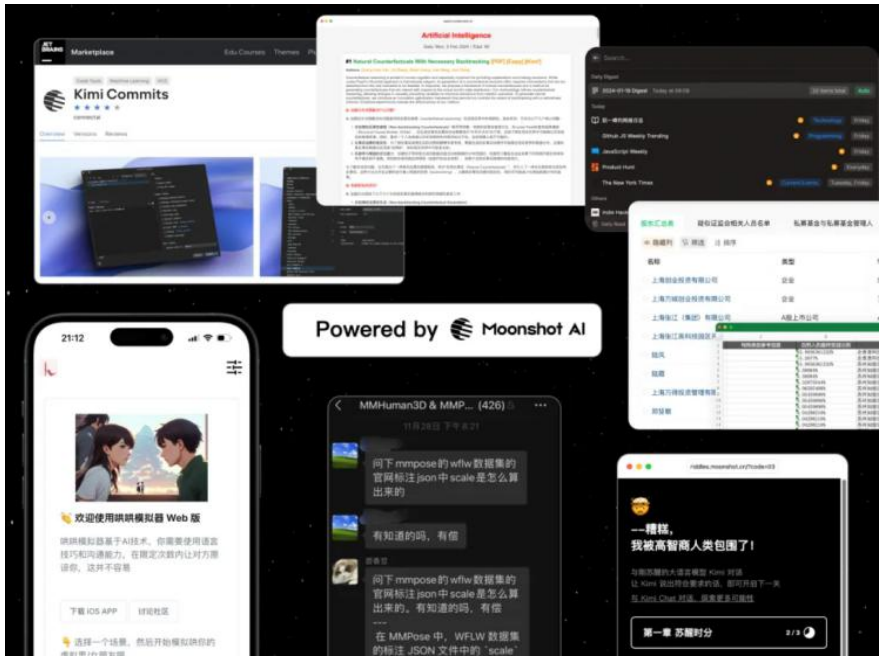
➢ 阅读和剧本创作中的应用：Kimi 的长文本处理能力在阅读和剧本创作领域展现出了深化内容与创新的潜力。它能够为小说和剧本等提供全书总结、剧本评估等高质量服务，这样不仅大幅提升了内容制作的效率，也极大丰富了用户的阅读体验。

➢ 游戏行业的互动升级：Kimi的长文本能力可用于生成复杂剧情和长篇人机对话极大丰富了游戏的互动性和沉浸感。

➢ 此外，Kimi 的长文本技术突破使得其应用场景从长文章分析扩展至AI陪伴和 AI Agent，如扮演小说中的角色或完成专业领域的特定任务。

- 这一变化为 AI在娱乐教育、专业服务等领域的深入应用开辟了新的可能性。Kimi 的发展吸引了多方企业的合作，涉及内容审核、数据训练和行业应用等多个环节。这些合作促进了 AI技术的实际应用，同时为各合作方带来了增值机会。

图：接入Moonshot AI 开放平台内测的应用



资料来源：国信证券经济研究所整理

Sora 开创AI 视频生成新纪元

- **OpenAI发布Sora大模型，通过Patches和Scaling Transformers革新视频生成技术。**

➢ **多模态融合与Patches技术：**Open AI通过将视觉数据转换为Patches的方法，仿照语言模型中token的应用，实现了文本多模态的统一，涵盖了代码、数学和自然语言等多种形式。Patches作为一种高效且可扩展的表示方法，在生成视频和图像的模型训练中展现了其独特价值。

➢ **通过时空Patches高效生成视频：**OpenAI创新性地开发了一套减少视觉数据维度的网络技术，这项技术可以把原始视频变成一个既在时间上也在空间上被压缩的潜在格式。Sora模型正是在这个压缩后的潜在空间中接受训练，从而能够生成新视频。为了将这些潜在的视频表示重新转化为清晰的图像，OpenAI还专门训练了一个解码器模型。通过对输入视频进行压缩并将其分解为一系列的时空Patches，这些Patches 便成了Transformer 模型的输入单位。这种方法使得Sora模型能够处理不同分辨率，持续时间和宽高比的视觉内容。在生成视频时，OpenAI能够通过在特定的网格中排列这些随机初始化的 Patches，从而有效控制生成视频的大小和形状。这一策略同样适用于图像处理，因为可以将图像看作是静态的单帧视频。

➢ **Sora采用 Scaling Transformer 提升模型效率：**OpenAI 通过应用Scaling Transformers的技术，成功地扩展了视频生成模型的能力。Scaling Transformers是指一系列旨在提高 Transformer 模型规模和效率的技术和方法，以便处理更大的数据集、更复杂的任务或在更大规模上运行，同时提高性能。在使用固定的初始条件(种子)和输入数据进行视频样本的训练过程中，OpenAI展示了通过增加训练过程中的计算量(例如，使用更多的计算资源或进行更多次的训练迭代)可以显著提高生成的视频样本的质量。

图：Sora 根据提示词生成视频

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.
提示词：一位时尚的女人走在东京的街道上，街道上到处都是温暖的发光霓虹灯和动画城市标志。她身穿黑色皮夹克，红色长裙，黑色靴子，背着一个黑色钱包。她戴着墨镜，涂着红色口红。她自信而随意地走路。街道潮湿而反光，营造出五颜六色的灯光的镜面效果。许多行人四处走动。

资料来源：OpenAI，国信证券经济研究所整理

图：Sora 根据提示词生成视频

资料来源：OpenAI，国信证券经济研究所整理

Sora 核心优势：强大的语言理解能力和一致性

国信证券 GUOSEN SECURITIES

- **Sora 核心优势在于强大的语言理解能力和一致性。**

➢ **强大的语言理解：**Sora引入了先进的字幕生成技术，借鉴 DALL·E3的重字幕(re-captioning)方法，为视频自动生成富有描述性的字幕。这一步骤不仅提升了视频与文字之间的匹配度，还极大改善了视频的整体品质。此外，通过 GPT将简短的用户指令 prompt 转化为详尽的描述，Sora能够精确地按照用户的需求创造视频，显著提高了生成视频的准确度和质量。

➢ **以图像和视频作为提示生成视频：**Sora的功能不限于将文字提示转换成视频它还能够处理图像或已有视频等多种类型的输入。这种能力让 Sora 成为一个应用广泛的编辑工具，能够轻松完成包括制作无缝循环视频、将静止图片变为生动动画，以及对视频进行前后时间轴的扩展等多项任务。0penAl 通过展示基于 DALL·E2 和 DALL·E3技术生成的示例视频，展现了Sora 在图像和视频编辑方面的强大能力和广阔应用前景。

➢ **灵活的视频扩展技术：**Sora 使用了基于Transformer 架构的扩散模型，可处理多种类型的输入数据，并能够在视频时间线上添加或修改内容。Sora能利用如SDEdit 这样的技术，在没有任何预设样本的情况下，改变视频中的风格或背景环境。这意味着用户可以更自由地定制他们的视频内容，不仅限于内容的创建，还包括对视频风格和环境的个性化调整，增强了视频编辑的灵活性和创造性。

➢ **出色的适应能力：**Sora拥有强大的视频生成和调整能力，能够应对不同分辨率和屏幕比例的需求。无论是宽屏格式(1920x1080像素)还是竖屏格式(1080x1920 像素)，Sora都能够自如地处理，确保生成的视频内容能够完美匹配不同设备的显示需求。此外，在进行高清视频内容创作前，Sora 能够迅速制作出低分辨率的视频原型，这一点对于加速创作过程和优化内容设计来说非常有用。简而言之，Sora使得视频制作变得更加灵活和高效，可以根据不同的显示设备和内容需求灵活调整视频规格。

➢ **场景和物体的一致性和连续性：**Sora能制作出视角多变的视频，使得角色和场景的三维移动看起来更自然。它还能有效解决物体被遮挡的问题。传统模型在追踪视野外物体时常常遇到困难，但 Sora 通过同时预测多帧内容，可以保证即使主体暂时消失在画面中也不会影响其一致性。

OpenAI新一代模型能力有望大幅提升

- **Sam Altman透露新一代大模型相关进展，模型能力大幅提升**。2024年1月，OpenAI首席执行官Sam Altman先后受邀参加了《Unconfuse Me》、达沃斯经济论坛，透露新一代大模型相关进展：1）大模型进展：目前OpenAI首要任务是推出下一代大模型，可能不命名为GPT-5，展望未来两年，人工智能有望在推理能力和可靠性、多模态（语音输入/输出、图像、视频）、可定制化和个性化三个领域大幅提升，其认为至少在未来5-10年内，AI大模型技术将处于一个非常陡峭的成长曲线上。2）新一代大模型架构和能力提升：OpenAI新一代模型将是一个多模态大模型，支撑语音、图像、代码和视频，并在个性化和定制化方面实现重大更新，具备更强的推理能力和更高的准确性；Sam Altman认为如果GPT-4解决了人类任务的10%，则新一代大模型有望解决人类任务的15%或20%；同时，AI大模型的幻觉问题有望在新一代大模型中解决。3）通往AGI之路：大模型能力提升不在于解决具体的问题，而是广泛意义的通用性在逐步增强。

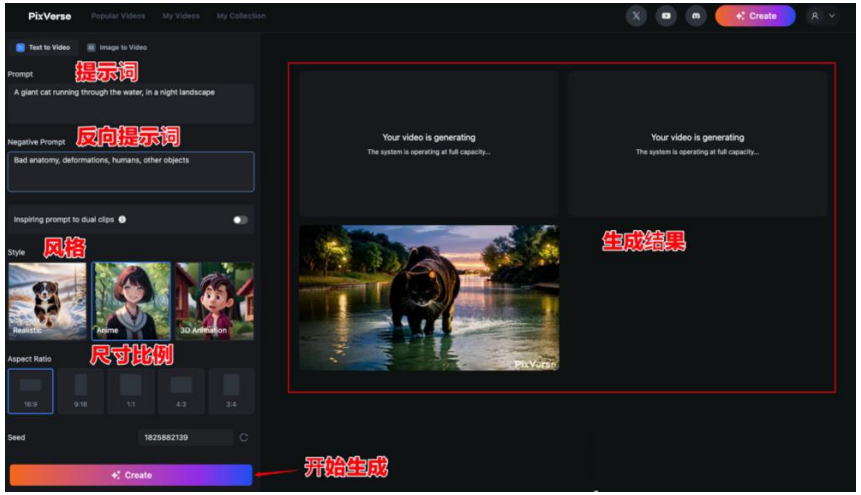
图：Sam Altman透露GPT-5相关进展

资料来源：达沃斯经济论坛，国信证券经济研究所整理

PixVerse 定位全球视频多模态应用，引领AI 创新潮流

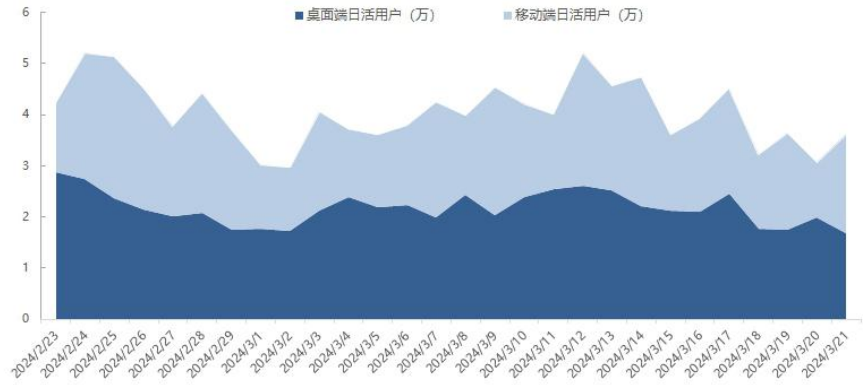
- 爱诗科技有限公司成立于2023年，是一家迅速崛起的AI视频生成大模型及应用企业。2024年1月，公司推出海外产品Pixverse，具备文生视频、图生视频等多种功能，目前已在海外AI视频生成领域占据一席之地，成为全球用户量最大的国产AI视频生成产品。
- **To创作者和To消费者的双重策略，目标在2024年底做到大规模的C端应用落地。**公司认为AI视频生成产品的第一阶段是To创作者，理解创作者动机；第二阶段将直面消费者。公司希望打通To C市场的AI视频生成全链路，持续推进国内外产品迭代，目标在24年底实现大规模C端应用。
- **访问量快速增长，PixVerse成国产AI视频之光。**目前PixVerse已初步搭建了稳定的创作者生态，并根据用户反馈进行模型迭代，在未来有望成为现象级、端到端的AI Native应用。据Similarweb统计，PixVerse在24年2月用户访问量已突破124万次，环比增长120%;2月访问量增速超越海外竞争对手Pika、Runway等，跻身全球AI视频生成工具第一梯队。

图：Pixverse 视频生成界面



资料来源：爱诗科技，国信证券经济研究所整理

图：Pixverse 日访问量



资料来源：Similarweb，国信证券经济研究所整理

目录

Kimi 火爆拉动算力需求增长

- Kimi大模型推理算力测算推理过程：主要包括分词(Tokenize)、嵌入(Embedding)、位置编码(PositionalEncoding)、Transformer 层、Softmax。推理主要计算量在Transfomer解码层，对于每个token、每个模型参数，需要进行2x1Flops=2次浮点运算，则单词推理算力消耗为 模型参数量 x (提问 Tokens + 回答 Tokens) x 2。推理算力计算假设及结果：

- **模型参数量：**如上文所述，假设Kimi大模型参数量为2000亿。
- **推理单次 Token量：**正常用户对话通常在1000 Token左右，假设推理单次 Token量为 1000。
- **推理算力需求：**根据 AI 大模型推理算力公式，单次Kimi大模型推理所需算力=2*Kimi 大模型参数量*(提问Tokens+回答Tokens)=2*(2000亿)*(1000)=8.0e^{14}Flops =800 TFlops。假设Kimi 日活为10万，单日活用户每天调用Kimi频率为30次，则Kimi单日推理调用总次数为300万次，则单日推理算力需求为2. 4e'TFlops。所需推理卡数及时间:考虑英伟达 A10卡目前国内储备量较大、成本较低，假设使用英伟达 A10卡进行Kimi模型推理，英伟达A10卡在FP16精度下算力为125TFlops，假设芯片利用率为 30%，同时考虑自天高并发因素(即夜间用户并不会使用 Kimi)，所以假设 Kimi 推理算力需求会集中在一天12个小时内，则2.4 e^{9} TFlops/(125 TFLOPs x 30% x 3600s x 12h/天)=1481张A10，即满足10万日活用户推理需求，需要1481张A10算力芯片作为支撑。

图：Kimi推理算力测算

Kimi 模型参数量（亿）		
2000		
推理单次调用 Token 量		
1000		
Kimi 日活（万）	**单日活用户每天调用 Kimi 频率（次）**	**Kimi 单日推理调用总次数（万次）**
10	30	300
单日推理算力需求（TFlops）		
2400000000		
推理算力卡需求（假设采用英伟达 A10 芯片）	**A10 FP16 精度下算力（TFlops）**	**芯片利用率**
	125	30%
推理卡需求（张）	1481	（假设 Kimi 用户使用集中在一天 12 小时以内）

资料来源：英伟达，国信证券经济研究所整理

Meta算力需求超预期，算力卡采购数量大幅增长

- **Meta将于24年底拥有接近60万颗H100 GPU的等效算力**。2024年1月19日，Meta董事长兼CEO扎克伯格在Facebook上发表视频，详细介绍了Meta在人工智能领域的最新进展和未来规划，聚焦于Meta通用人工智能（AGI）的追求，以及Meta做出了相关战略调整。

➢ 从硬件侧，Meta正在积极部署英伟达H100 GPU，计划至24年底部署接近35万颗H100 GPU，叠加英伟达A100和其他AI芯片，将拥有接近60万颗H100 GPU的等效算力，以支撑下一代AI大模型LIama 3的训练；

➢ 从组织架构侧，Meta将其两大AI研究团队（FAIR和GenAI）合并，共同致力于通用人工智能（AGI）的构建；

➢ 从智能产品侧，提到了Ray-Ban Meta智能眼镜，关注元宇宙未来的发展。

- **24年Meta算力卡采购数量同比大幅增长，算力需求超预期。**根据Omdia Research统计数据，23年全球大厂纷纷采购H100 GPU，其中Meta和微软采购15万颗，位居第一；其次，谷歌、亚马逊、Oracle、腾讯采购5万颗，主要用于AI云业务的建设以及自研AI大模型的训练需要。根据扎克伯格公布的24年算力卡采购预期：

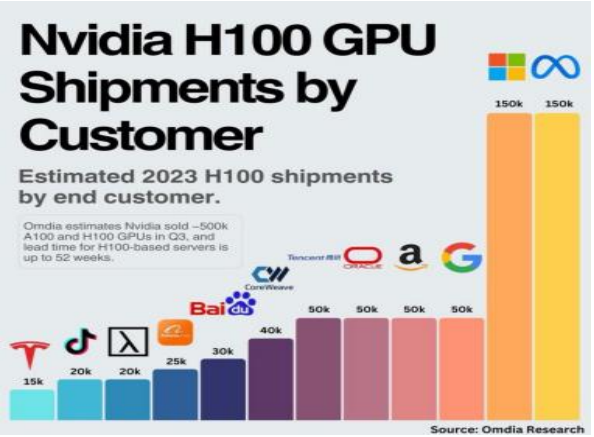
➢ H100 GPU：23年公司采购15万颗，24年预计采购20万颗，同比+33.33%，合计24年底在手35万颗H100 GPU，对应增量资本支出12.5亿美金（假设单科H100 GPU 2.5万美金，增量为5万颗）；

➢ 其他等效H100 GPU：24年底等效H100 GPU数量达到25万颗，包括A100以及将要出货的英伟达H200、AMD MI300X等AI芯片，由于H100 GPU的性价比优于A100，23年全年Meta A100采购数量相对较少，若24年底达到25万颗的等效H100算力，我们认为Meta将大量采购英伟达H200、AMD MI300X等高性价比芯片。

图：扎克伯格介绍Meta在AI领域的最新进展和规划

资料来源：Meta，国信证券经济研究所整理

图：3年全年Meta采购15万颗H100 GPU



资料来源：Omdia Research，国信证券经济研究所整理

多模态大模型拉动AI训练、推理算力需求增长

- **大模型训练算力测算：**训练过程可分前向传播（Forward Pass）和反向传播（Backward Pass）。

➢ 前向传播：输入数据（例如图像、文本等）通过神经网络的各层进行传递，以得到输出结果，包含输入数据与权重矩阵相乘、应用激活函数等操作，目的为将计算网络预测输出，并将其与实际目标值比较，计算损失函数（Loss Function）的值。

➢ 反向传播：一种高效计算梯度算法，从输出层开始，沿着网络层次结构向输入层反向传播，计算每个权重的梯度（注：梯度表示权重对损失函数贡献的大小）；同时，在计算出所有权重的梯度后，使用优化算法更新权重，达到减小损失函数值的目的。

➢ 计算次数：一次前向传播需要一次计算，一次反向传播需要两次计算（计算梯度+权重更新），则完成一次神经网络迭代需要对所有输入的数据和模型参数进行3次计算；每一次计算就是矩阵运算，对于一次矩阵运算需要进行一次乘法及加法（共计2次浮点运算），即对于每个Token、每个模型参数，需要进行2 × 3 Flops = 6次浮点运算。以GPT-3大模型训练为例，模型参数量为175B，训练Token数量为300B，采用稠密（Dense）模型，其需要的训练总算力为175B × 300B × 6 = 3.15e^{23} FLOPs。

➢ 所需算力卡数量及时间：假设使用业内FLOPS最大的利用率来测算（此处取46.2%），单卡A100 FP16精度下算力为312 TFLOPS，则3.15 e^{23} FLOPs/(312 TFLOPs × 46.2% × 3600s × 24h/天)= 2.53万张A100/天，即若使用1000张A100，大约训练一遍GPT-3需要25.3天。

- **大模型推理算力测算：**推理过程主要包括分词（Tokenize）、嵌入（Embedding）、位置编码（Positional Encoding）、Transformer层、Softmax。推理主要计算量在Transfomer解码层，对于每个token、每个模型参数，需要进行2 × 1 Flops = 2次浮点运算，则单词推理算力消耗为模型参数量 × （提问Tokens + 回答Tokens） × 2。

- 以GPT-3单次推理为例，假设用户每次提问20 Tokens，ChatGPT回答300 Tokens，模型参数量为175B，则单次推理算力需求为175B × （20 Tokens + 300 Tokens） ×2 = 1.12e^{14} FLOPs，若使用单张A100 GPU进行推理，假设芯片利用率为46.2%，则完成单次所需时间为1.12 e^{14} FLOPs/(312 TFLOPs × 46.2%)=0.78s

表：芯片利用率情况

Model	# of Parameters (in billions)	Accelerator Chips	Model FLOPS Utilization
GPT-3	175B	V100	21.3%
Gopher	280B	4096 TPU v3	32.5%
Megatron-Turing NLG	530B	2240 A100	30.2%
PaLM	540B	6144 TPU v4	46.2%

资料来源：Aakanksha Chowdhery等著-《PaLM:Scaling Language Modeling with Pathways》-arXiv(2022)-P9，国信证券经济研究所整理

图：公开模型的算力数据

Model Name	Model Size (# parameters)	Training Data (# tokens)	Training Compute (FLOPs)	GPU Resource
GPT-4	1.8T	13T	2.15E+25	222.6万张A100/天，25000张A100，需要训练时间90天
GPT-3	175B	300B	3.1E+23	2.53万A100/天，1000张A100，需要训练时间接近一个月
Baichuan	7B，13B	1.4TB	5.88E+22-1.09E+23	>4720 A100/天
Llama 2	7B，13B，70B	2000B	8.4E+22-8.4E+23	>6744 A100/天
Falcon	40B	1TB	2.4E+23	19267 A100/天
Chat-GLM2	6B，130B	1TB	3.6E+22-7.8E+23	>2890 A100/天
文心一言	>100B	>1TB	>6.0E+23	>4.82万 A100/天
盘古	110B	40TB	2.64E+25	212万 A100/天

资料来源：腾讯云，国信证券经济研究所整理

图像模态拉动AI算力需求增长

国信证券 GUOSEN SECURITIES

- **图像训练数据大幅提升训练Token量。**以BEIT方法为例，单一图片训练素材可以有两种表达形式，即Image Patches和Visual Tokens。

➢ **Image Patches：**将图片分成N=HW/P2个展平的2D块，每个image patches会被展平成向量，并对其进行线性变换操作，进而得到一系列展平的2D块的序列；随后使用类BERT的子监督训练方式（Masked Image Modeling），即随机隐藏部分Image Patches，让模型对隐藏的部分进行预期，进而不断计算预测的Patches和真实的Patches之间的差异，并将该差异作为Loss函数进行反向传播来更新参数。

➢ **Visual Tokens：**BEIT通过DVAE（Discrete Variational Autoencoder，核心原理是试图构建一个从隐变量Z生成目标数据X的模型）中的Image Tokenizer，将单一图片训练素材转化为离散的Tokens（即隐变量），再通过生成器（Decoder）重建原图。

➢ **图片对训练数据量的提升：**以Image Patches方法为例，1张图片可以分割为N=HW/P2个2D块（即视为输入的Tokens），其中（H，W）为输入图片的分辨率，（P，P）是2D块的大小，在《BEIT: BERT Pre-Training of Image Transformers》实际操作中，有1张224*224大小的图片分割成16*16大小的2D小块，即单一图片相当于（224*224）/（16*16）=196个Tokens。而在纯文本训练素材中，单一单词约为4/3个Token，则1张图片（分辨率224*224）约等于147个单词。根据上文所述，AI训练算力需求 = 模型参数量 × 训练Token量 × 6，图片训练素材的加入，拉动训练Token量的大幅增长，进而大幅提升AI训练算力需求。

➢ **增量测算：**a）数据量：根据《Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning（Pablo等著，2022年）》披露数据，2022年全球图片数量在$5e^{10}$-$2e^{11}$个，我们取中间值（即1 e^{11}个），选取常用图片分辨率（1024×768），则单张图片对应（1024*768）/（16*16）=3072个Tokens，则全部图片对应3.072 e^{14}个Tokens。b）算力需求：假设使用这些图片数据对一个5000亿参数模型进行训练，则对应的AI训练算力需求 = 500B × 3.072 e^{14} × 6 = 9.216e26 FLOPs。 c）训练卡需求：以英伟达H100为例，在FP16精度下算力为1979 TFLOPS，仍假设芯片利用率为46.2%，则9.216 e^{26} FLOPs/（1979 TFLOPs × 46.2% × 3600s × 24h/天 × 30天/月）= 38.89万张H100/月，即完成对图片数据的训练需使用38.89万张H100训练一个月（针对单一模型），假设全球有5家厂商使用图片素材进行自研大模型训练，则需要194.45万张H100训练一个月。

视频模态拉动AI算力需求增长

- **视频训练数据大幅提升训练Token量。**以字节跳动最新提出《MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation》方法为例，该模型是一个多阶段端到端视频生成模型，具体可分为以下4个关键模块：

➢ Text-to-Image模块（文本到图像）：从给定的文本提示，生成概括所描述场景的高分辨率图像（分辨率为1024*1024）；

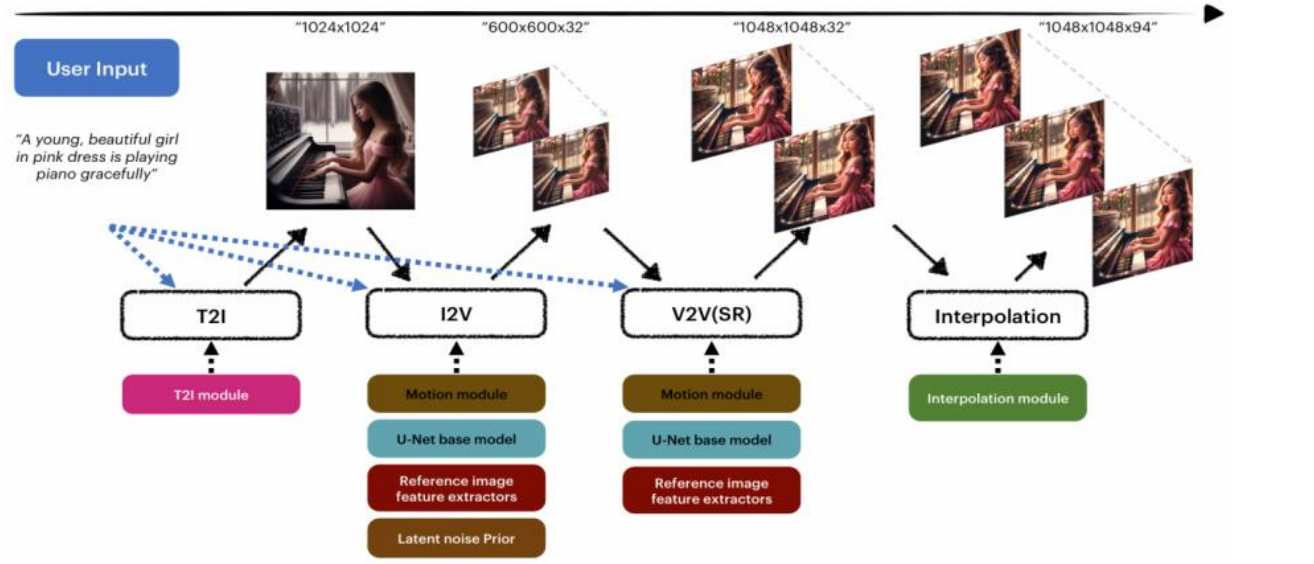
➢ Image-to-Video模块（图像到视频）：通过文本提示和生成的图像创建关键帧（32帧），使得静态图像动态化（分辨率为600*600）；

➢ Video-to-Video模块（视频到视频）：增强并细化视频帧的内容，并拓展至更高的分辨率（分辨率为1048*1048）；

➢ Video Frame Interpolation（VFI，帧插值）模块：在关键帧之间插入帧以平滑视频运动（94帧），确保动作流畅和时间一致性。

- 其中，Text-to-Image模块（文本到图像）的训练同前文图像模态训练相似，除了BEIT方法外，OFA等方法亦可得到不错的Text-to-Image模型。

图：MagicVideo-V2模型结构

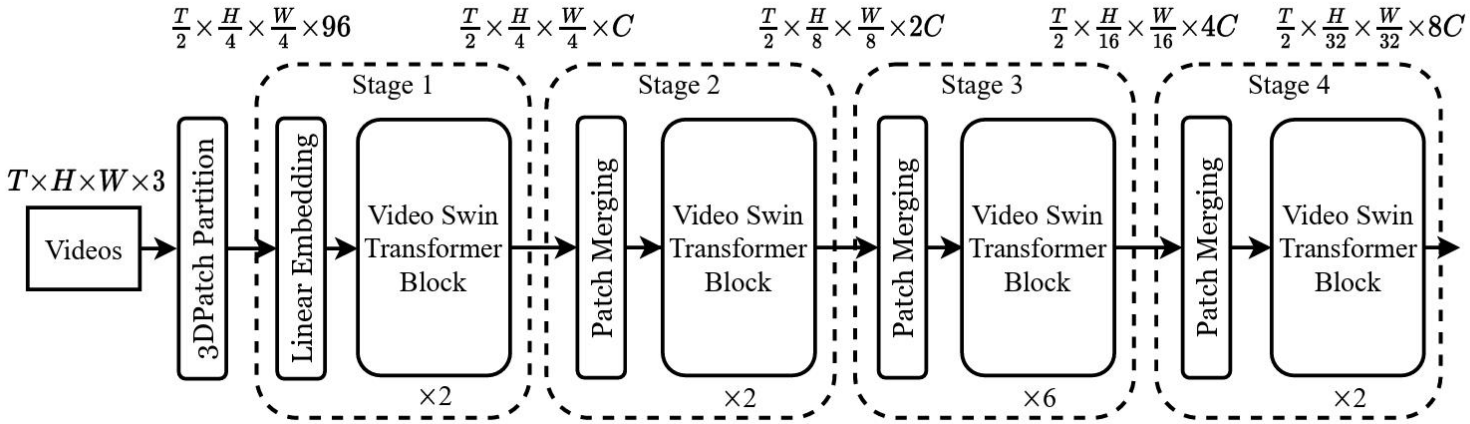


资料来源：Wang等著-《MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation》-arXiv（2024）-p2
国信证券经济研究所整理

视频模态拉动AI算力需求增长

- Image-to-Video模块需要视频数据进行训练。根据Ze Liu等著《Video Swin Transformer（2021）》，输入一个尺寸为T×H×W×3的视频（此处T选取32，代表从视频中采样得到32帧，采样方法可自行选择，通常为等间隔采样，视频长度通常约10s；每帧包含H×W×3个像素），通过3D Patch Partition可以得到(T/2)*(H/4)*(W/4)个3D Patch（尺寸为2*4*4*3），即为Tokens，之后再经过Video Swin Transformer和Patch Merging获得多帧数据的高维特征，完成视频数据训练。根据《Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning（Pablo等著，2022年）》披露数据，Youtube每分钟大约上传500小时视频，则我们可以得到Youtube一年增量视频数据为500×3600×24×365=157.68亿秒。通常分类任务视频为10s左右，对应采样帧数为32，假设每帧图片分辨率为1024×768，则10s视频对应的Token数量为 (32/2)*(1024/4)*(768/4)=78.64万个Tokens，则Youtube一年增量视频数据为1.24 e^{15}个Tokens，假设使用Youtube一年增量视频数据对5000亿大模型完成一遍训练对应的算力需求为500B × 1.24e15 × 6 = 3.72 e^{27} FLOPs。以英伟达H100为例，在FP16精度下算力为1979 TFLOPS，仍假设芯片利用率为46.2%，则3.72 e^{27} FLOPs/(1979 TFLOPs × 46.2% × 3600s × 24h/天 × 30天/月)= 156.98万张H100/月，即完成对视频数据的训练需使用156.98万张H100训练一个月（针对单一模型，仅计算Youtube一年增量视频数据）；且后续Video-to-Video模块（视频到视频）、Video Frame Interpolation（VFI，帧插值）模块仍需要算力支撑。

图：对视频素材划分3D Patch Partition



资料来源：Ze Liu等著-《Video Swin Transformer》-arXiv（2021）-p3，国信证券经济研究所整理

美国限制对华云服务，看好国产算力需求提升

- **美国将限制云厂商对华客户提供AI云服务。**美国商务部部长Gina Raimondo宣布，美国政府正推出一项提案，阻止外国实体，特别是来自中国的实体，使用美国的云计算进行AI大模型的训练。美方认为这是保障国家安全和美国技术优势的一项努力。根据2024年1月29日美国BIS部门发布的相关文件，提到“requiring U.S. Infrastructure as a Service(IaaS) providers of IaaS products to verify the identity of their foreign customers， along with procedures for the Secretary to grant exemptions.(要求提供IaaS产品的IaaS厂商确认其外国客户身份，遵循安全部门豁免程序)”。

- **国内领先大模型厂商影响有限，看好国产算力需求提升。**国内领先大模型厂商大多自建智算中心，使用自有的AI算力训练大模型，该政策对国内领先大模型厂商影响有限。国内AI大模型初创公司受制于创业初期资金不足，部分厂商租赁海外云厂商AI算力进行自研AI大模型训练；同时，国内训练垂类模型的部分AI应用厂商亦会租赁海外云厂商AI算力进行调优；该政策发布后，部分国内AI大模型初创公司和国内训练垂类模型的AI应用公司有望自行购买算力卡或租赁国产AI算力进行模型的训练和后续的推理，看好国产算力需求提升。

图：美国BIS文件

5698 Federal Register / Vol. 89, No. 19 / Monday, January 29, 2024 / Proposed Rules

DEPARTMENT OF COMMERCE

15 CFR Part 7

[Docket No. 240119–0020]

RIN 0694–AJ35

Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities

AGENCY: Bureau of Industry and Security, Department of Commerce.

ACTION: Proposed rule; request for comments.

SUMMARY: The Executive order of January 19, 2021, ''Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities,'' directs the Secretary of Commerce (Secretary) to propose regulations requiring U.S. Infrastructure as a Service (IaaS) providers of IaaS products to verify the identity of their foreign customers, along with procedures for the Secretary to grant exemptions; and authorize special measures to deter foreign malicious cyber actors' use of U.S. IaaS products. The Executive order of October 30, 2023, ''Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,'' further directs the Secretary to propose regulations that require providers of certain IaaS products to submit a report to the Secretary when a foreign person transacts with that provider or reseller to train a large Artificial Intelligence (AI) model with potential capabilities that could be used in malicious cyber-enabled activity. The Department of Commerce (Department) issues this notice of proposed rulemaking (NPRM) to solicit comment on proposed regulations to implement those Executive orders.

DATES: Comments must be received April 29, 2024.

ADDRESSES: All comments must be submitted by one of the following methods:

- *By the Federal eRulemaking Portal: https://www.regulations.gov* at docket number DOC–2021–0007.
- *By email directly to: IaaScomments@bis.doc.gov.* Include ''E.O. 13984/E.O. 14110: NPRM'' in the subject line.
- *Instructions:* Comments sent by any other method or to any other address or individual, or received after the end of the comment period, may not be considered. For those seeking to submit confidential business information (CBI), please clearly mark such submissions as CBI and submit by email or via the Federal eRulemaking Portal, as instructed above. Each CBI submission must also contain a summary of the CBI, clearly marked as public, in sufficient detail to permit a reasonable understanding of the substance of the information for public consumption. Such summary information will be posted on *regulations.gov*.

FOR FURTHER INFORMATION CONTACT: Kellen Moriarty, U.S. Department of Commerce, telephone: (202) 482–1329, email: *IaaScomments@bis.doc.gov*. For media inquiries: Jeremy Horan, Office of Congressional and Public Affairs, Bureau of Industry and Security, U.S. Department of Commerce: *OCPA@bis.doc.gov*.

SUPPLEMENTARY INFORMATION:

I. Background

IaaS products offer customers the ability to run software and store data on servers offered for rent or lease without having to assume the direct maintenance and operating costs of those servers. Foreign malicious cyber actors have utilized U.S. IaaS products to commit intellectual property and sensitive data theft, to engage in covert espionage activities, and to threaten national security by targeting U.S. critical infrastructure. After carrying out such illicit activity, these actors can quickly move to replacement infrastructure offered by U.S. IaaS providers of U.S. IaaS products (''U.S. IaaS providers''). The temporary registration and ease of replacement for such services makes it more difficult for the government to track malicious actors. Additionally, the ability of malicious actors to use foreign-person resellers of U.S. IaaS products (''foreign resellers''), who might not track identity, hinders law enforcement's ability to obtain identifying information about malicious actors through service of compulsory legal process. This shift in adversary tradecraft also challenges the U.S. Government's ability to identify victims of malicious cyber activity and enable specific network defense and remediation efforts. Furthermore, the emergence of large-scale computing infrastructure—to which U.S. IaaS providers and foreign resellers provide access as a service, and which foreign malicious actors could use to train large AI models that can assist or automate their malicious cyber activity—has raised considerable concern about the identities of entities that transact with providers to engage in certain AI training runs.

To address these threats, the President issued E.O. 13984, ''Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities,'' which provides the Department with authority to require U.S. IaaS providers to verify the identity of foreign users of U.S. IaaS products, to issue standards and procedures that the Department may use to make a finding to exempt IaaS providers from such a requirement, to impose recordkeeping obligations with respect to foreign users of U.S. IaaS products, and to limit certain foreign actors' access to U.S. IaaS products in appropriate circumstances. The President subsequently issued E.O. 14110, ''Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,'' which calls for the Department to require U.S. IaaS providers to ensure that their foreign resellers verify the identity of foreign users. E.O. 14110 also provides the Department with authority to require U.S. IaaS providers submit a report to the Department whenever a foreign person transacts with them to train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity.

II. Introduction

E.O. 13984 and E.O. 14110 draw upon the President's authority from the Constitution and laws of the United States, including the International Emergency Economic Powers Act (IEEPA) (50 U.S.C. 1701 *et seq.*), the National Emergencies Act (NEA) (50 U.S.C. 1601, *et seq.*), and 3 U.S.C. 301. Section 1 of E.O. 13984 requires the Secretary to propose, for notice and comment, regulations that mandate that U.S. IaaS providers verify the identity of foreign persons that sign up for or maintain accounts that access or utilize U.S. IaaS providers' IaaS products or services (Accounts or Account)—that is, a know-your-customer program or Customer Identification Program (CIP). Under E.O. 13984, such a program must set forth the minimum standards for IaaS providers to verify the identity of a foreign person connected with the opening of an Account or the maintenance of an existing Account. The proposed regulations must include the types of documentation and procedures required to verify the identity of any foreign persons acting as a lessee or sub-lessee of these products or services; the records that IaaS providers must securely maintain regarding a foreign person that obtains an Account; and methods of limiting all third-party access to this collected information, except insofar as such access is otherwise consistent with E.O. 13984 and allowed under applicable law. Moreover, the proposed regulations

资料来源：BIS，国信证券经济研究所整理

第一，宏观经济下行风险。若宏观经济波动，产业变革及新技术的落地节奏或将受到影响，宏观经济波动导致下游需求不及预期，可能对 IT 投资产生负面影响，从而导致整体行业增长不及预期。

第二，行业竞争加剧。国内各厂商纷纷加大 AI 相关投入，导致产品陷入同质化竞争。

第三，国内 AI 大模型、算力等技术发展不及预期，影响AI 在各行业应用进度。

第四，相关政策推进不及预期，如生成式 AI 应用需面临相关政策要求等。

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	买入	股价表现优于市场代表性指数20%以上
		增持	股价表现优于市场代表性指数10%-20%之间
		中性	股价表现介于市场代表性指数±10%之间
		卖出	股价表现弱于市场代表性指数10%以上
	行业投资评级	超配	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
		低配	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