计算机行业2024年3月投资策略

国产大模型Kimi带动产业链革新

超配

核心观点

Kimi成为国产大模型曙光，带动产业链革新。2023年10月，由清华大学杨植麟及其团队“月之暗面”推出的Kimi标志了国产AI大模型的重要进步。Kimi凭借其在处理长文本方面的卓越能力，能够处理高达20万字的输入，显示出其在无损阅读方面的巨大潜力。这一突破不仅提升了内容创作和整理的效率，还为小说、剧本创作等领域带来了深化和创新，同时在游戏互动、AI陪伴和专业领域任务执行等方面开辟了新的应用场景。

全球AI应用的增长普遍疲软，Kimi以其显著的用户增长和应用广度突围而出。从2023年10月到2024年3月，Kimi的日活用户从10万迅速增长至300多万，这一显著增长反映了Kimi在模型优化、人才扩展和用户吸引方面的成功策略。Kimi的成功不仅依赖于其技术优势，更在于其对用户体验的重视，包括通过数据驱动的持续产品优化、创新的分享机制以及对核心功能的精准打磨，这些因素共同提升了Kimi的市场竞争力。

Sora模型开创AI视频新纪元。OpenAI近期发布了其首款视频生成大模型Sora，标志着文生视频领域的一个重要突破。Sora拥有生成长达60秒、紧密贴合用户指令的视频的能力，这在视频内容的长度、多角度一致性以及对物理世界的理解方面展示了其显著优势。技术上，Sora采用了一种创新的方法，通过利用已知的图像片段（Patches）来推测接下来的片段，并将Transformer模型与Diffusion技术相结合，实现了复杂视频内容的高效生成。Sora的出现不仅代表了AI在视频生成领域的新里程碑，还对多模态大模型的发展产生了深远影响。随着这类模型对视频、图像、文本等不同类型数据处理能力的整合，对算力的需求也随之大幅提升。这一变化预示着未来在训练高效、功能强大的AI模型方面，将需要投入更多的计算资源。

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，建议关注国产AI算力龙头公司海光信息；2）大模型能力快速提升，多模态将进一步扩大AI的应用范围，此外，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，例如金山办公、同花顺。

风险提示：宏观经济复苏不及预期；云厂商资本开支不及预期；市场竞争加剧；产品研发不及预期；国产AI算力芯片导入不及预期等。

重点公司盈利预测及投资评级

公司代码	公司名称	投资评级	昨收盘（元）	总市值（亿元）	EPS 2023	EPS 2024E	PE 2023	PE 2024E
688041	海光信息	买入	80.19	1864	0.54	0.72	130.7	111.1
300033	同花顺	买入	138.84	746	2.61	3.22	60.13	43.09
688111	金山办公	买入	332.32	1535	2.86	3.87	110.69	85.86

资料来源：Wind、国信证券经济研究所预测

行业研究·行业月报

计算机

超配·维持评级

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

证券分析师：库宏垚
021-60875168
kuhongyao@guosen.com.cn
S0980520010001

联系人：艾宪
0755-22941051
aixian@guosen.com.cn

市场走势

资料来源：Wind、国信证券经济研究所整理

相关研究报告

《计算机行业2024年2月投资策略-全球AI训练算力重估，美方将限制对华AI云服务》——2024-02-03
《计算机2023年12月暨2024年度策略：大模型能力日新月异，AI将重塑各行各业》——2023-12-22
《GPTs更新（二）：视频应用凸起》——2023-12-03
《计算机行业2023年11月投资策略暨三季报总结-三季度基本面逐步复苏，关注AI产业创新机会》——2023-11-17
《OpenAI发布会解读：GPTs带来AI应用全面爆发》——2023-11-16

内容目录

图表目录

国产大模型曙光 Kimi 带动产业链革新

月之暗面发布 Kimi，成为国产大模型曙光

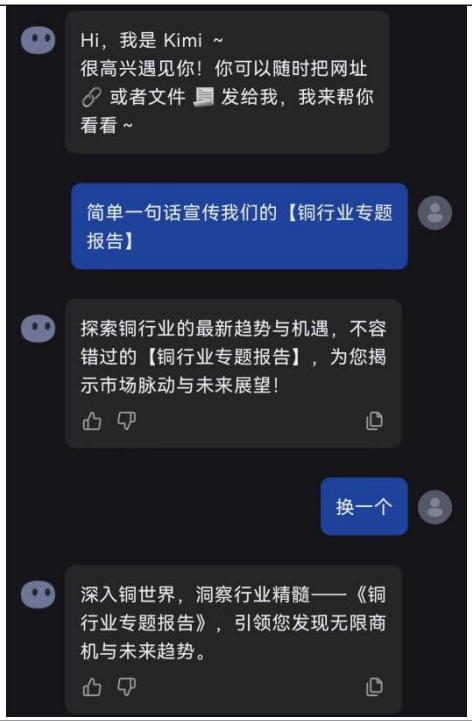
2023 年 10 月，清华大学杨植麟及其 AI 团队“月之暗面”发布了 Kimi，是国产大模型的代表作之一，拥有优秀的长文本处理能力，可处理 20 万汉字输入，得到业界高度关注。依赖于优秀的长文本处理能力，用户可在 Kimi 上传各种资料并进行整理，包括 word、pdf、txt 以及图片；此外，Kimi 也支持访问互联网链接，读取文本信息并归纳总结。2024 年 3 月 18 日，Kim 公布了支持 200 万文字输入的内测邀请，测试者可使用 Kimi 获得 200 万字长度的无损阅读能力。

Kimi 虽然参数量尚未公开，但其强大的性能已引起广泛关注。其显著优势在于：1）高效处理长文本，可处理 20 万汉字的数据；2）改善用户体验、提升用户留存度。如提供个性化体验：了解用户偏好、兴趣和行为模式，提供定制化内容或服务，增强客户满意度和忠诚度；利用长文本理解能力，在多轮对话后也能准确回应客户需求，可构建连贯、有意义的长对话，提高交流自然性和效率。

图1：Kimi 可以阅读英文论文并整理

资料来源：Kimi，国信证券经济研究所整理

图2：Kimi 可以根据提示词生成宣传文本



资料来源：Kimi，国信证券经济研究所整理

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。

1）模型 Scaling：Kimi 通过持续优化其 AI 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。

2）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。

3）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。

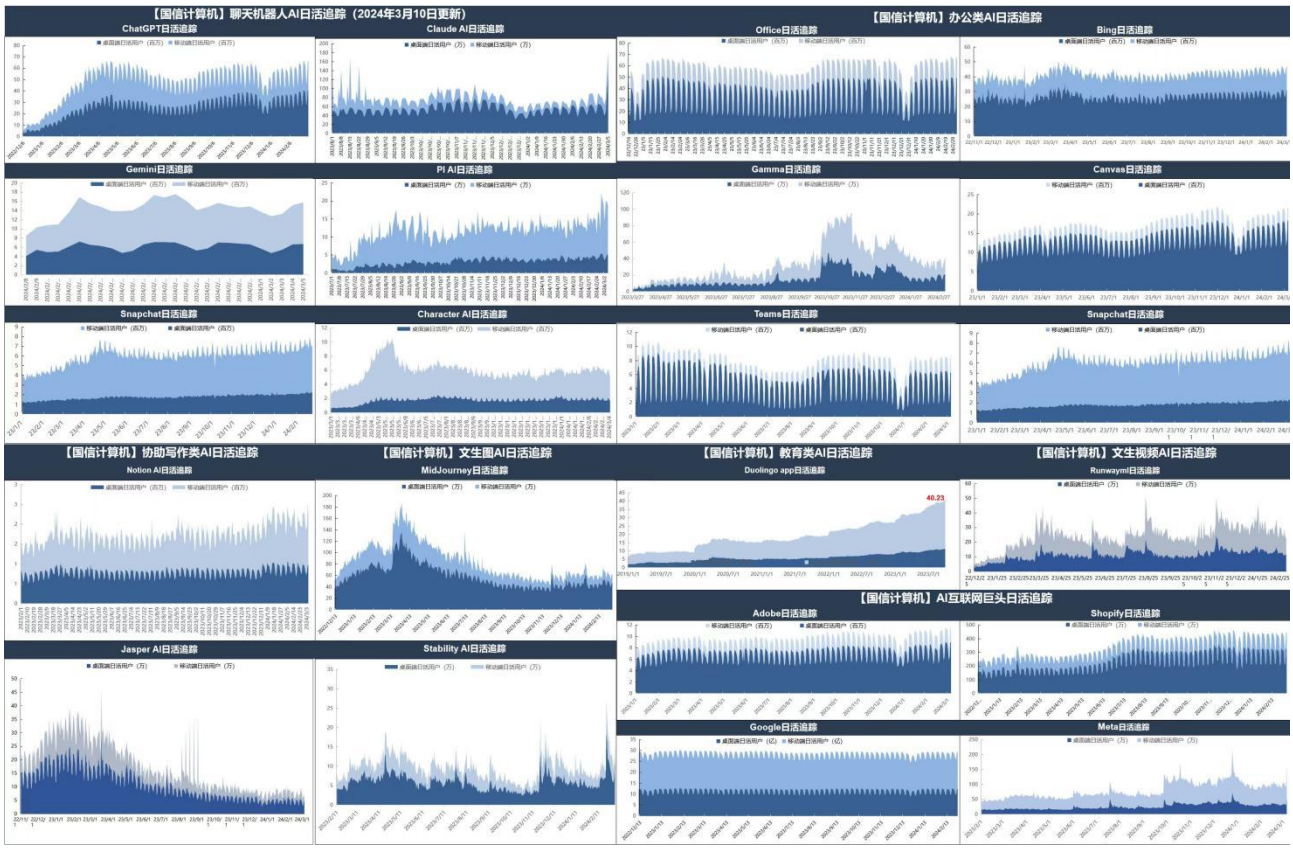
Kimi 突破 AI 应用现状，保持环比高速增长

全球 AI 应用增长均比较疲软。在对大约 1 万款 AI 产品进行统计分析后，我们得到了海外及国内 AI 应用的访问量数据：

海外：AI 应用增速较低，大部分应用环比增长仅为个位数，其中 ChatGPT 月访问量达到 16 亿，环比增长了 1.08%;Bing 月访问量下降了 7.87%，至 13.4 亿，日活略有下降;Google 的 Germini 月访问量增加至 3.26 亿,主要系 2 月份谷歌将 Bard 更名为 Gemini，并同步开放 Gemini Advanced 订阅和上线 Gemini App 版。

国内：2 月份总体访问量环比下降约 20%。主要是受到春节假期的影响，用户的互联网使用习惯可能发生了变化，导致访问量暂时下降。此外，这也说明了当前大部分 AI 产品尚未完全融入到用户的日常生活场景中。

图3：近期海外 AI 应用增长疲软，大部分应用日活、月活增速仅有个位数



资料来源：SimilarWeb，国信证券经济研究所整理

Kimi 访问量保持环比高速增长。在众多 AI 产品中，Kimi 环比持续高增，日活从 23 年 10 月份的 10 万迅速提升至目前 300 多万，预计下月可能超过阿里、通义的千万用户量级。

图4：Kimi 保持环比高增，并有望超越文心一言、通义千问

国内排名	产品名 AI产品榜	分类 aicpb.com	2月上榜访问量	2月上榜变化
1	百度文心一言	AI ChatBots	10.06M	-33.43%
2	阿里通义千问	AI ChatBots	3.65M	-45.05%
3	Kimi（Moonshot）	AI ChatBots	3.05M	107.60%
4	稿定AI	AI Design Tool	2.47M	-11.88%
5	魔音工坊	AI Audio Editing Tools	2.09M	-4.26%
6	讯飞星火	AI ChatBots	1.92M	-29.16%
7	火山方舟	Model Training & Deplc	1.83M	-23.22%
8	抖音豆包	AI ChatBots	1.73M	-2.18%
9	清华智谱清言	AI ChatBots	1.71M	27.47%
10	ProcessOn	AI Mind Map Generato	1.42M	-29.31%

资料来源：aicpb.com，国信证券经济研究所整理

长文本处理能力，是人类与 AI 交流无损理解的基础。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。

图5：长文本建模是自然语言处理(NLP)领域的一项重要技术

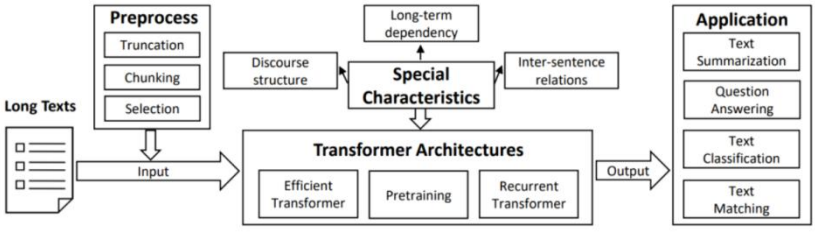


Figure 1: An illustrative process of modeling long texts with Transformers.

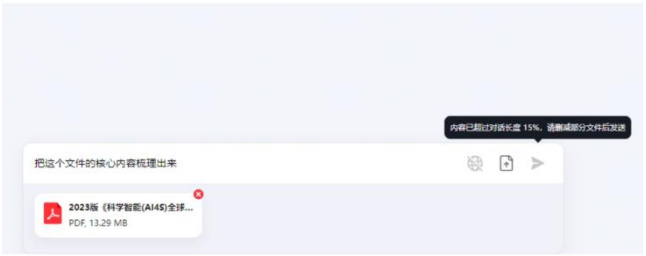
资料来源：A Survey on Long Text Modeling with Transformers，国信证券经济研究所整理

以 Claude 3 的进化和其开发团队提供的 Prompt library 为例，我们可以看到 AI 开发者是如何通过精细化的提示库来指引用户以更高效、更精准的方式与 AI 进行交流。这不仅减少了沟通的冗余，也提高了交流的效率和效果，使得 AI 能够更好地服务于人类的需求。然而，目前市场上各种 AI 大模型在处理长文本方面的能力还存在限制，如 Claude 对文件大小的限制等。

图6：AI 开发者通过精细化的提示库来指引用户以更高效、更精准的方式与 AI 进行交流

资料来源：Kimi，国信证券经济研究所整理

图7：目前市场上各种 AI 大模型在处理长文本方面的能力还存在限制

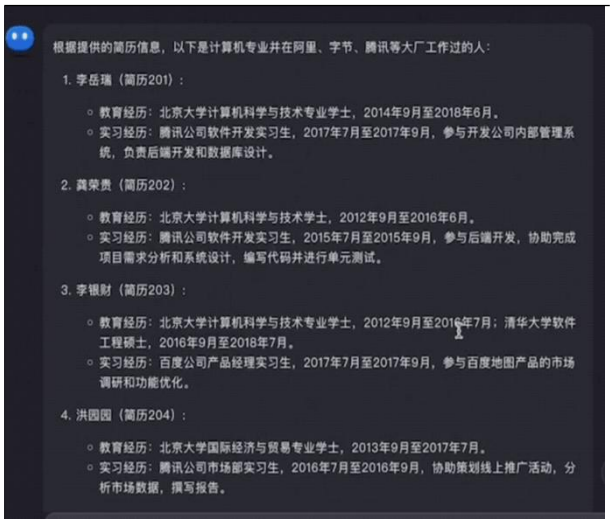


资料来源：Kimi，国信证券经济研究所整理

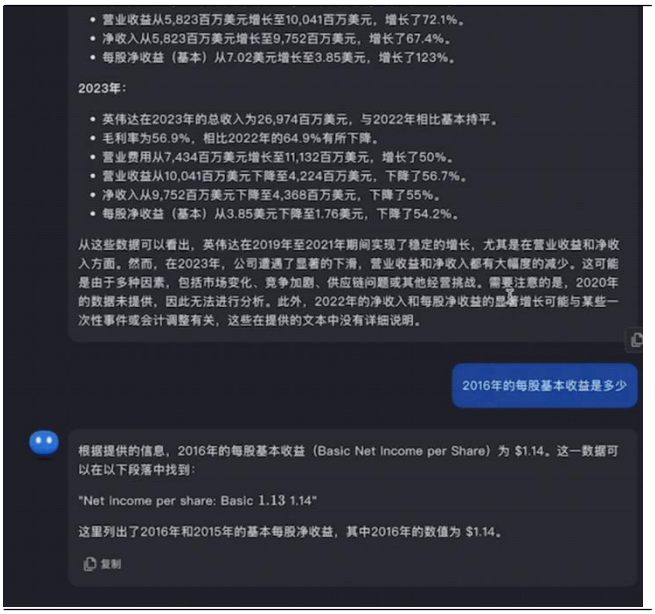
Kimi 的长文本能力是解决复杂问题的关键，并可优化用户体验。1）解决复杂问题：长文本模型通过处理大量信息，特别适合执行如企业知识库整合等需要广泛知识和深入理解的任务。这使得 AI 能在更广泛的上下文中提供精确、全面的解决方案；2）模型性能提升：处理更长的文本能让 AI 模型捕获更深层次的语言特征和复杂关系，使模型在执行任务时更接近人类的思维方式；3）忠诚度与指令遵循：长文本模型必须准确保留原文意义（忠诚度）并能根据任务指令执行操作（指令遵循能力），确保其在处理复杂任务时的可靠性和有效性；4）用户体验优化和数据飞轮效应：通过优化用户体验和实施分享机制，长文本模型不仅促进用户增长，还通过用户生成数据进一步优化模型，形成数据飞轮效应。

图8：Kimi 可以两分钟读完 500 份简历，筛选员工

图9：Kimi 可以读取英伟达报告，并分析财报历史



资料来源：Kimi，国信证券经济研究所整理



资料来源：Kimi，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

联合技术壁垒和服务壁垒，Kimi 有望重塑竞争格局

AI 产品的竞争壁垒为技术壁垒、服务壁垒，除了提供较强的技术实力之外，满足用户需求并提供优质体验也非常重要，特别是未来 AI 产品性能可能趋同，因此短期内的市场颠覆能力成为关键竞争点。创新产品如 Kimi 有望重塑竞争格局，通过提供个性化体验和优越性能显示出巨大的发展潜力。

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。

1）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；

2）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；

3）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。

4）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。

5）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。

Kimi 为多个行业带来了潜在发展机遇

Kimi 优秀的性能可以带动多个产业的发展。如：

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。

Sora 开创 AI 视频生成新纪元

OpenAI 发布 Sora 大模型，革新文生视频技术

2024 年 2 月 16 日，OpenAI 推出文生视频大模型 Sora，引发业界高度关注。该模型能根据提示词生成不同分辨率、时长和宽高比的视频，包括全高清视频，时长可达 1 分钟。Sora 模型能够生成包含多个角色、特定类型运动和主体及背景精确细节的复杂场景。该模型不仅能理解用户在提示中所要求的内容，还能理解这些事物在现实世界中的存在方式。

图10：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11：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 的重要意义在于它再次推动了 AIGC 在 AI 驱动内容创作方面的上限。在此之前，ChatGPT 等文本类模型已经开始辅助内容创作，包括插图和画面的生成，甚至使用虚拟人制作短视频。而 Sora 则是一款专注于视频生成的大模型，通过输入文本或图片，以多种方式编辑视频，包括生成、连接和扩展，属于多模态大模型的范畴。这类模型在 GPT 等语言模型的基础上进行了延伸和拓展。

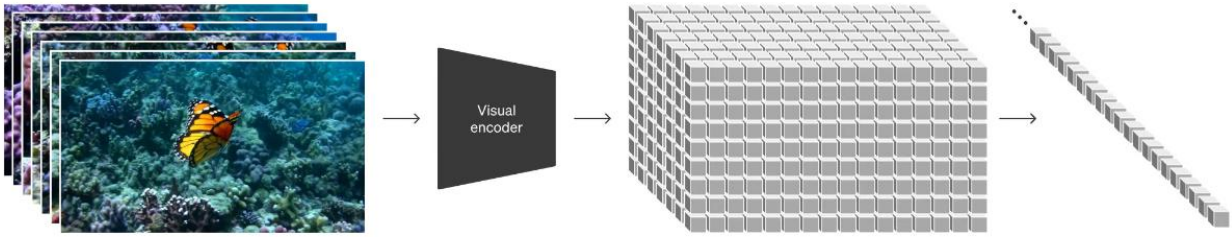
Sora 模型：通过 Patches 和 Scaling Transformers 革新视频生成技术

1）多模态融合与 Patches 技术：OpenAI 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。

2）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。

图12：将视频数据转换为 patches



资料来源：OpenAI，国信证券经济研究所整理

3）**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13：将 Scaling transformers 用于视频生成

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

1）强大的语言理解：Sora 引入了先进的字幕生成技术，借鉴 DALL·E 3 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。

2）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。OpenAI 通过展示基于 DALL·E 2 和 DALL·E 3 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。

3）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。

4）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。。

5）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。

Sora 或可对影视制作及传媒游戏等行业产生深远影响

使用 Sora 技术或可降低视频、动画和游戏制作的整体成本。其原因主要在于：1）Sora 能够自动将文本、图像甚至视频提示转换成复杂的视频内容，减少了人工编辑和制作所需的时间和劳力；2）减少专业需求：通过 Sora 生成视频内容，可以减少对专业视频制作人员、动画师和后期处理专家的依赖，尤其是在初期内容创作和原型设计阶段；3）节约制作资源：对于广告制作、短视频创作和游戏开发等领域，Sora 可以生成高质量的视频和动画内容，无需昂贵的拍摄设备、场地租赁或复杂的后期制作；4）快速原型和迭代：Sora 允许快速创建内容原型，并易于修改和迭代，这有助于在项目早期阶段发现并解决潜在问题，避免了成本高昂的后期更改。

在短视频创作领域，Sora 有望降低短剧制作的综合成本，解决“重制作而轻创作”的问题。这将使短剧制作的重心回归高质量剧本内容创作，对创作者的构思能力提出更高要求。为企业降低成本、提高效益，广告制作公司可以通过 Sora 生成符合品牌需求的广告视频，从而显著减少拍摄和后期制作成本。游戏和动画公司也能够利用 Sora 直接生成游戏场景和角色动画，降低 3D 模型和动画制作的成本。通过节省下来的成本，企业可以提升产品和服务质量，或进行技术创新，从而推

动生产力的进一步提升。

Sora 或可于年末发布，并用于影视行业。OpenAI 的首席技术官（CTO）穆拉蒂最近介绍了 Sora 项目的最新进展和未来发展方向。他指出，Sora 能够在几分钟内生成时长为 20 秒、分辨率为 720P 的视频，这一能力大大提升了生成视频的效率。目前，Sora 生成的视频还不包含声音，但 OpenAI 计划未来加入语音功能，以使内容更加丰富和完整。此外，为了确保 Sora 的安全性和可靠性，OpenAI 正在进行红队测试，并可能在 2024 年末发布该技术。穆拉蒂还宣称，OpenAI 可能将与电影行业合作，旨在帮助大幅降低布景等方面的成本，展现了 Sora 在未来电影制作中的潜在应用价值。

附：近期 AI 事件

表1：近期 AI 事件汇总

事件	日期	地区	领域	简述
字节跳动版 GPTS 扣子上线	2.1	国内	文本生成	字节跳动正式推出「Coze 扣子」AI Bot 开发平台。Coze 是一款应用程序编辑平台，用于开发下一代 AI 聊天机器人。任何用户都可以快速、低门槛地搭建自己的 Chatbot，且平台支持用户将其一键发布到飞书、微信公众号、豆包等渠道。
Meta 第二代自研 AI 芯片正式投产	2.2	国外	其他	Meta 计划今年在数据中心部署第二代 Artemis AI 芯片，主要性能集中在推理领域，将与 Meta 购买的现成的英伟达 GPU 协同以增强 AI 算力，共同夯实该公司的 AI 基础设施能力。
阿里通义千问 Qwen1.5 发布	2.6	国内	文本生成	Qwen1.5 是通义千问系列的最新迭代版本。QWen 1.5 提供从 0.5 亿参数到 720 亿参数不等的六个模型尺寸。性能评测结果显示，尤其是小模型在某些任务上的表现优于同类产品。
谷歌 Bard 更名为 Gemini	2.8	国外	文本生成	Bard 全新升级，更名 Gemini，参数较低的版本将继续免费供用户使用，新增付费计划(19.99 刀每月)，付费用户可使用 Gemini Ultra 1.0，其性能与 GPT-4 接近。同时 Gemini Pro 上线移动端，在 Google App 中可以使用，目前只对美国、日本、韩国开放。相比 GPT-4，Gemini Ultra 的最大优势之一是与 Google 服务的集成，包括金融、地图、文档、邮件等。
Meta 发布 MetaVoice-1B	2.9	国外	音频生成	MetaVoice-1B 由 MetaVoice 公司研发的，拥有 1.2B（12 亿）参数，在 10 万小时的语音数据上进行训练，这意味着它能够学习到丰富的语音模式和细节，从而生成更加自然、逼真的语音输出。朗读带感情，支持 30s 声音克隆，支持长文本合成
英伟达发布 GPU RTX 2000 Ada	2.12	国外	其他	英伟达最低端专业显卡，Ada Lovelace 架构专业家族的入门级成员，桌面显卡最低端型号 RTX 4060 一样的 AD107 小核心，但是 CUDA 核心数更少仅为 2816 个，同时有 88 个第四代 Tensor 核心、22 个第三代 RT 核心。单精度浮点性能 12.0TFlops，RT 核心性能 27.7TFlops，Tensor 核心性能 191.9TFlops，号称对比 Ampere 架构的上一代 RTX 2000 分别提升 1.5 倍、1.7 倍、1.8 倍，此外 VR 性能提升最多 3 倍。要价为 625 美元，约合人民币 4500 元
OpenAI 发布文生视频模型 Sora	2.16	国外	视频生成	Sora 可以根据用户的文本提示创建最长 60 秒的逼真视频，该模型了解这些物体在物理世界中的存在方式，可以深度模拟真实物理世界，能生成具有多个角色、包含特定运动的复杂场景。继承了 DALL-E 3 的画质和遵循指令能力，能理解用户在提示中提出的要求.
谷歌发布 Gemini 1.5 Pro	2.16	国外	文本生成	Gemini 1.5 Pro 是谷歌最新发布的一款 Gemini 系列模型，采用 MoE 架构，上下文窗口高达 100 万个 tokens（对比目前 GPT-4 Turbo 的上下文窗口长度为 128K，即 12.8 万个 tokens），支持跨模态理解、分析和推理，即支持对图片和视频的分析。
StabilityAI 发布 Stable Diffusion 3	2.21	国外	图像生成	一款类似于 OpenAI 的 Sora 的 Diffusion Transformer 文生图模型。SD3 包含从 800m 到 8B 参数的一系列模型，与现有图像生成模型相比，规模和图像质量有了很大飞跃，对提示词的理解、文字生成和人物的手部和面部生成效果有了大幅提升。3 月 5 日，Stable Diffusion 3 技术报告新鲜出炉：该模型在视觉美观度、文本遵循和排版等方面均超越了 DALLE 3、Midjourney v6 和 Ideogram v1 等先进的文本到图像生成系统。
谷歌发布 Gemma 系列模型	2.21	国外	文本生成	Gemma 2b 和 7b 是 Google 基于 Gemini 技术推出一款轻量级、经过微调的开源小模型，它非常适合各种文本生成任务，包括问答、摘要和推理。尤其是 Gemma 2b，它的参数量相对较小，因此可以部署在资源有限的环境中，如笔记本电脑、台式机或个人

				手机等移动设备。目前各个主流推理框架或工具都已适配完成，包括 llama.cpp，mlc 等。
字节跳动发布 SDXL-Lightning	2.23	国内	图像生成	SDXL-Lightning 在 SDXL 基础上，利用一种渐进式对抗蒸馏的技术，实现了前所未有的图像生成速度，可以在 4-8 步内生成 1024 分辨率的高质量图像，计算成本为基础模型的十分之一。SDXL-Lightning 可以作为增速插件无缝集成到 SDXL 模型中，目前主流的 SDXL 都开始适配。
JuggernautXL-V9-Lighting 发布	2.26	国外	图像生成	JuggernautXL 是目前最好的开源文生图模型之一，也是全球下载量最大的 AI 绘画大模型，基于 SDXL 微调，可以生成摄影级别的图片。本次更新使用了字节跳动的 Lighting 技术，只需 4 个 step 就可以生成高质量图像，相比 Turbo 和 LCM 等优化技术，Lighting 生成的图像质量更好，速度更快。
Phind 发布 70B 代码生成模型	2.23	国外	文本生成	Phind-70B 在 HumanEval 上得分为 83%，超越 GPT-4，成为最强代码生成模型。相比 GPT-4，Phind-70B 不会出现懒惰或者拒绝回答的问题，而且推理速度可以达到每秒 80+token，远高于 GPT-4 的每秒 20+token。Phind-70B 基于 CodeLlama-70B 训练，在 500 亿 Token 上进行了微调，支持 32K 上下文窗口。另外 Phind-34B 也同时发布。
MistralAI 发布旗舰模型 Mistral Large	2.26	国外	文本生成	顶级的推理能力，多语言支持，内置函数调用功能，32K 上下文窗口，MMLU 得分 81.2，Mistral Large 成为仅次于 GPT-4 的商业模型。MistralAI 与微软合作，将 Mistral Large 集成到了 Azure AI Studio。
微软提出 BitNet 1.58	2.27	国外	文本生成	大语言模型或将迎来 1-bit 时代。全新的 BitNet b1.58 神经网络，将大模型的权重参数从浮点数替换成三元组（-1，0，1），原本的矩阵计算从乘法变成了整数加法，在保持推理精度的同时，极大地减少了模型所需的存储空间和计算资源。以 3B 模型为例，相比 llama 模型，推理速度提高了 2.71 倍，GPU 使用率只有 llama 的四分之一，随着参数量的增大，效果更加明显。
Playground 2.5 发布	2.28	国外	图像生成	Playground 是基于 SDXL 架构重新训练的模型，V2.5 版提高了图片的美学质量，画面的颜色、对比度、光影效果、图像细节等方面都有大幅提升。根据 1000+个用户的人工评测，Playground V2.5 是目前最符合人类偏好的文生图模型。
阿里巴巴发表视频生成模型 EMO	2.28	国内	视频生成	EMO 能够通过单一参考图像和音频输入，生成具有丰富表情和多样头部姿势的虚拟角色视频。EMO 利用先进的注意力机制和去噪网络，支持多语言和多种肖像风格的动态表现，为内容创作和虚拟角色动画制作提供了新工具。
苹果取消造车，加码生成 AI	2.28	国外	其他	苹果公司搁置并取消了自动驾驶电动汽车的所有开发计划，汽车团队众多成员将被调往人工智能部门，由高管约翰·詹南德里亚领导。这些员工将专注于推动生成式人工智能项目。蒂姆·库克线上股东大会上表示，公司认为“生成式人工智能具有令人难以置信的突破潜力”，承诺今年晚些时候向大家分享“在 AIGC 方面开创新局面的方式”，这是另一项“能重新定义未来的技术”。
Adobe 发布 AI 音乐创作工具原型 Project Music GenAI Control	2.29	国外	音频生成	无需专业的音乐知识，只需利用文本提示就可以生成和编辑音乐。未来 Adobe 全家桶都会向 AI 方向进化。
英伟达市值破 2 万亿美元	3.1	国外	其他	2.21 凌晨英伟达公布 FY24Q4 业绩，营收同比大增 265%，每股收益同比暴增 765%，连续三个季度创纪录，美股走高，3 月 1 日的涨势使得英伟达市值达到 2.06 万亿美元，使其成为仅次于微软和苹果的华尔街第三大价值公司。
Anthropic 发布 Claude-3	3.4	国外	文本生成	Claude-3 的多模态和语言能力指标全面碾压 GPT4，在推理、数学、编码、多语言理解和视觉方面树立行业新基准。Cl 包括三款模型：小杯-Haiku、中杯-Sonnet、大杯-Opus，最强的 Opus 需付费订阅，价格为 20 美元/月。从官方发布的测试结果来看，Claude 3 Opus 在所有评估基准上碾压 GPT-4，数学能力和逻辑推理能力尤其突出。作为首个多模态 GenAI，用户可以上传照片、图表、文档等非结构化数据，由 AI 模型进行分析和回答。
谷歌发布最新「读屏」AI	3.4	国外	音频生成	ScreenAI 是一种理解用户界面和信息图表的全新视觉语言模型，能够完成各种屏幕 QA 问答、总结摘要等任务，核心是一种新的屏幕截图文本表示方法，可以识别 UI 元素的类型和位置。研究人员使用谷歌语言模型 PaLM 2-S 生成了合成训练数据，以训练模型回答关屏幕信息、屏幕导航和屏幕内容摘要的问题。
政府工作报告首提（人工智能+）	3.5	国内	其他	报告指出，制定支持数字经济高质量发展政策，积极推进数字产业化、产业数字化，促进数字技术和实体经济深度融合。深化大数据、人工智能等研发应用，开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。报告还指出，适度超前建设数字基础设施，加快形成全国一体化算力体系。
零一万物发布 Yi-9B	3.6	国内	文本生成	号称 Yi 系列中的“理科状元”，代码和数学能力同级别最优，具有 90 亿参数，消费级显卡良好兼容，为广大开发者和研究人

				员提供了前所未有的便利性和强大功能。
马斯克 xAI 开源大模型 Grok	3.11	国外	文本生成	Grok 基于 Grok-1，是马斯克 x.AI 公司的第一个大语言模型，开发大约花了四个月的时间（包括两个月的训练），上下文长度为 8192，Grok-1 的实力与 GPT-3.5 相当。开发者社区对 Grok 开源版本的反馈和改进可能有助于 xAI 加速开发新版本，这些新版本 xAI 可以选择开放源代码或保留专有权。
OpenAI 开源 Transformer Debugger	3.12	国外	文本生成	OpenAI 机器学习研究员 Jan Leike 宣布，OpenAI 要开源内部一直使用的大杀器——Transformer 调试器，Transformer Debugger 是 OpenAI 对齐团队（Superalignment）开发的一种工具，旨在支持对小体量语言模型的特定行为进行检查，该工具把自动可解释性技术与稀疏自动编码器进行了结合。
Meta 推出 2 个 24K GPU 集群	3.12	国外	其他	Meta 推出 2 个 24K GPU 集群，为训练 Llama3 构建超强资源池，声称到 2024 年底将完成 350,000 个 NVIDIA H100 GPU 集群的构建。届时，其整个资源池计算能力将相当于近 600,000 个 H100。

资料来源：腾讯科技、凤凰科技，国信证券经济研究所整理

投资建议

大模型应用拉动全球算力需求快速增长，关注国产 AI 算力侧机会。随着长文本模型应用的落地及多模态大模型的发展，全球对算力的需求正快速上升。特别是在美国对华提供 AI 云服务的限制背景下，国产 AI 算力在训练和推理领域均迎来发展机遇。看好国产算力需求提升，建议关注国产 AI 算力龙头公司海光信息。

大模型能力快速提升，关注 AI 应用侧机会。Sam Altman 预测，在未来 5-10 年，AI 大模型技术将迎来急剧增长，特别是在推理能力、多模态交互、以及定制化和个性化方面。OpenAI 即将推出的新一代多模态大模型，将支持语音、图像、代码和视频，预计解决人类任务的比例将从 10%提升至 15%或 20%，同时解决幻觉问题。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，例如金山办公、同花顺。

风险提示

宏观经济复苏不及预期；云厂商资本开支不及预期；市场竞争加剧；产品研发不及预期；国产 AI 算力芯片导入不及预期等。

免责声明

分析师声明

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。

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

重要声明

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 层
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