计算机行业2024年4月投资策略

Kimi引领国产大模型群雄并起，国内AI应用迎发展良机

超配

核心观点

Kimi以长文本实现破圈，B、C两端应用均有突破。公司认为AI-Native产品核心价值在于个性化交互，Long Context可以解决90%的模型定制问题，因为足够长的上下文即可让模型定制，大幅减少fine-tune的成本。从全球各大模型迭代方向来看，提升长文本能力是全球大模型技术趋势；“大海捞针”测试验证了Kimi长文本能力，Kimi在中文领域对GPT-4 Turbo、Claude 2.1优势明显。24年3月下旬，Kimi进一步将上下文能力提升至200万汉字，同时Kimi应用火爆出圈，用户流量激增导致连续5次扩容。C端应用是公司主要发力点，目标打造超级应用，致力于成为AI原生交互入口。B端公司打造Moonshot AI开放平台，API与OpenAI兼容；内测期间已有法律、游戏、阅读等相关应用进行测试，并反馈较好。随着Kimi应用访问量持续提升，也将再次拉动算力需求的快速增长。

阶跃星辰多模态能力领先，积极布局算力和应用生态。公司在24年3月发布了3款Step系列通用大模型，包括Step-1千亿参数语言大模型、Step-1V千亿参数多模态大模型和Step-2万亿参数MoE语言大模型（预览版）；同时推出C端产品“跃问”、“冒泡鸭”。公司多模态能力领先，在OpenCompass的测试中，Step-1V多模态大模型性能已赶超ChatGPT-4、Qwen-VL等，位居榜首。应用端，公司参股财跃星辰，与中国知网、中文在线等开展合作；算力端，公司在“租用算力+自建厂房”外，还参股了上海智能算力科技公司。

Pixverse引领全球AI视频，目标3-6个月赶超Sora。Pixverse已经成为全球用户量最大的国产AI视频生成产品，24年2月用户访问量已突破124万次，环比增长120%。同样基于DiT的技术路线，PixVerse花了3个月的时间就做到了全球第一梯队的水平，迅速把视频内容做到了4K的分辨率，资源和资金的消耗比Runway、Pika至少小了一个数量级。公司采取To创作者和To消费者的双重策略，目标在24年底实现大规模C端应用。未来3-6个月，公司最重要的目标是技术上能够追平甚至赶超Sora。

投资建议：看好国产大模型持续突破，国内模型、应用、算力均迎来发展机会。2024年国内大模型新势力异军突起，产品力和应用体验快速追赶全球头部模型水平，部分领域已经接近，甚至达到了全球第一梯队。以Kimi为代表的长文本能力、阶跃星辰的多模态模型、Pixverse的AI视频生成，均验证了国内团队在一年多时间里取得的跨越式进步。随着模型能力持续迭代，国内在应用方面的创新性，算力国产化的进一步升级将带动AI生态进入正循环。重点关注金山办公、同花顺、海光信息。

风险提示：宏观经济低迷影响IT支出；国内大模型技术突破不及预期；AI相关商业化拓展不及预期；行业竞争加剧；相关政策进度不及预期。

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 2024E	EPS 2025E	PE 2024E	PE 2025E
688111	金山办公	买入	298.99	1381	3.68	4.89	81.25	61.14
300033	同花顺	买入	134.82	725	3.22	3.79	41.87	35.57
688041	海光信息	买入	76.79	1785	0.72	0.97	106.65	79.16

资料来源：Wind、国信证券经济研究所预测

行业研究·行业投资策略

计算机

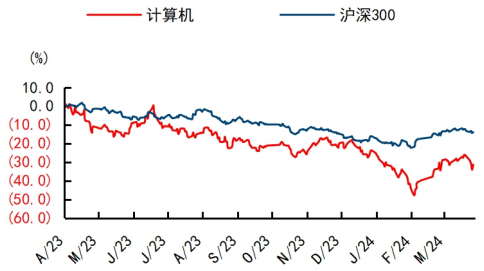
超配·维持评级

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3月投资策略：国产大模型Kimi带动产业链革新》——2024-03-21
《多模态AI大模型点评-OpenAI发布首款文生视频大模型Sora，训练算力需求大幅提升》——2024-02-17
《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
《计算机行业2024年1月投资策略-制造业提效降本，看好各工业软件龙头智能制造落地》——2024-01-07
《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

内容目录

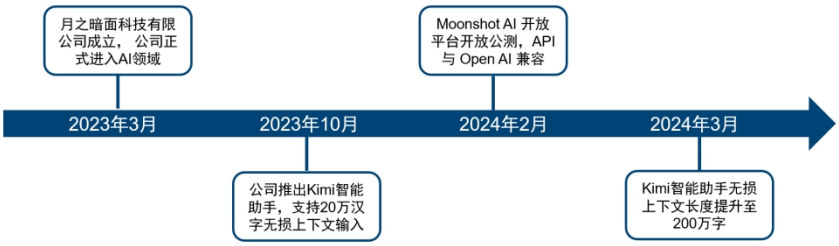
图表目录

Kimi 实现破圈，引领国产大模型新方向

月之暗面成为国产大模型新星

愿景宏大，Kimi 成为国内通用大模型头部应用。月之暗面科技有限公司成立于 2023 年 3 月 11 日，秉持“寻求将能源转换为智能的最优解”的愿景，致力于通过产品与用户共同创作，实现通用人工智能（AGI）目标。2023 年 10 月，公司正式推出第一款对话类产品 Kimi 智能助手，其基于千亿级模型参数构建并以长文本处理作为最核心能力，为用户提供高达 20 万汉字的输入与输出支持，实现了长上下文的无损记忆。公司产品迭代迅速，并于 2024 年 2 月将 Moonshot AI 开放平台启动公测，于 24 年 3 月进一步将上下文能力提升至 200 万汉字，不到半年提升 10 倍。当前，随着 Kimi 火爆出圈，已经成为国产 AI 头部应用。

图1：月之暗面发展历程



资料来源：公司官网，国信证券经济研究所整理

创 AI 融资新高，吸金能力强劲。2023 年 6 月，月之暗面获得来自红衫资本中国与真格基金的天使轮投资，投资金额超 2 亿美元。仅 4 月后，公司获得近 20 亿人民币的第二轮融资，主要投资机构包括红衫资本中国、砺思资本等。2024 年 2 月，公司完成超 10 亿美元的 A 轮融资，由阿里领投，红衫资本中国、小红书、美团等跟投，投后估值约 25 亿美元，打破国内 AI 领域最高单轮融资额度的记录。

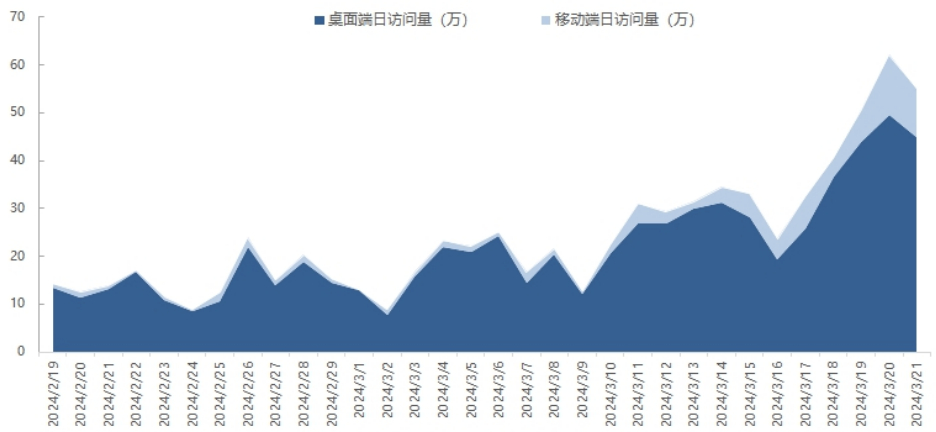
表1：月之暗面融资情况一览

轮次	时间	金额	投资方
天使轮	2023 年 6 月	超过 2 亿美元	红杉中国、真格基金、砺思资本
	2023 年 10 月	近 20 亿元人民币	红杉中国、今日资本、砺思资本等
A 轮	2024 年 2 月	超 10 亿美元	阿里巴巴、红杉中国、小红书、美团等

资料来源：澎湃新闻，企查查，国信证券经济研究所整理

Kimi 访问量迅速提升，产品得到市场广泛认可。随着 Kimi 在用户体验，长文本处理能力上口碑的日益积累，Kimi 的访问量整体呈现持续上涨。尤其是在 24 年 3 月后，公司开启了 200 万字的“长文本”输入的内测，产品能力进一步被市场认可。据 SimilarWeb 统计，3 月 20 日 Kimi 移动端/桌面端的访问量同比分别高增 332/987pct，尤其近一周访问量激增。根据 AI 产品榜 2 月数据，目前国内多数 AI 应用的访问量出现了下降，但 Kimi 仍保持了极高的增长。

图2：Kimi 智能助手自 3 月以来访问量持续增长

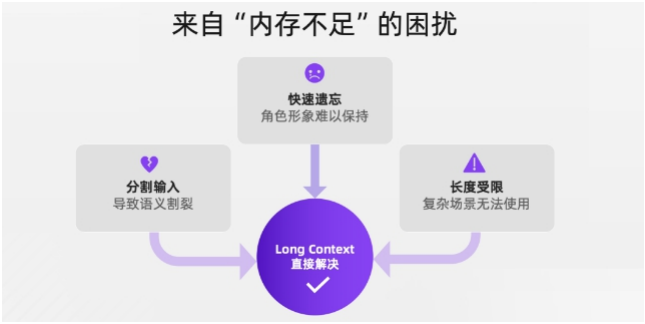


资料来源：SimilarWeb，国信证券经济研究所整理

Kimi 主打长文本能力，产品能力优异

上下文长度不足为传统大模型应用带来定制化和迭代问题。传统大模型应用中，由于较短的上下文，会出现分割输入（同一单词分段输入后，出现语义理解的歧义）、快速遗忘（角色扮演时，多轮对话后遗忘早期设定）、长度受限（ Agent 等场景下，复杂任务无法装载在 Context 中）等问题。在“记忆”有限的背景下，为了面对各类应用场景，传统大模型引入 fine-tune 实现定制化，而这样再模型迭代后又将面临再次 fine-tune 的困境。如此造成了较高的成本，同时也无法满足多数客户的需求。月之暗面公司认为，Long Context 是解锁模型定制与模型迭代之间矛盾的钥匙。

图3：长上下文能解决部分大模型的问题



资料来源：Moonshot AI，国信证券经济研究所整理

长上下文可以解决 90%的定制问题。通过 Long Context，上下文中可以承载足够多的信息，而这些信息足以让模型实现定制化。第一版的 Kimi Chat 就已支持长达 20 万字的上下文处理能力，能满足大多数场景应用，例如角色定制、客服交流、简历筛选等，在输入大多数客户要求的内容后（角色要求、产品手册、筛选标准），Long Context 仍能支持后续超长的互动空间。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。

图4：Long Context 实现功能案例

资料来源：Moonshot AI，国信证券经济研究所整理

图5：Long Context 解决了模型定制的 90%问题

资料来源：Moonshot AI，国信证券经济研究所整理

Kimi Chat 在长文本、代码生成方面能力表现出色。我们对其进行测试，发送了 8 篇有关货币政策的论文进行解析，在响应速度方面，对于 30 页以内的文档，上传后即可迅速完成解析，仅需 20 秒综述相关文本即输出完毕；在长文本概括能力方面，模型能兼顾各文章核心观点，并能在处理多篇论文时，有效地比较和融合相似的观点，同时保证文本输出的可读性和层次性。在代码复现方面，Kimi 能够运用 Python、R 等编程语言，提供精准且高效的代码实现，其不仅能根据论文的模型和关键变量迅速编写出清晰、结构良好的代码，还能逐行解释代码的功能与内在逻辑，确保用户能够充分理解其工作原理。

图6：Kimi 阅读 8 篇指定论文并进行文献综述

资料来源：Kimi 智能助手，国信证券经济研究所整理

图7：Kimi 用指定语言复现学术论文代码

资料来源：Kimi 智能助手，国信证券经济研究所整理

长文本能力成为产业共识，Kimi 取得领先并成为破局关键

提升长文本能力是全球大模型技术趋势。目前，海内外的主要大模型玩家正在积极扩展上下文窗口，以进一步强化模型处理复杂信息、理解并生成有逻辑的长文本能力。目前 Kimi 升级到 200 万汉字，在全球范围内保持领先。而国内多家大模型也在积极迭代，近期阿里通义千问宣布将向所有人免费开放 1000 万字的长文档处理功能；百度文心也将进行升级，长文本能力将达到 200-500 万字；360 智脑也在内测 500 万字长文本处理功能。长文本能力成为大模型技术主要突破方向。

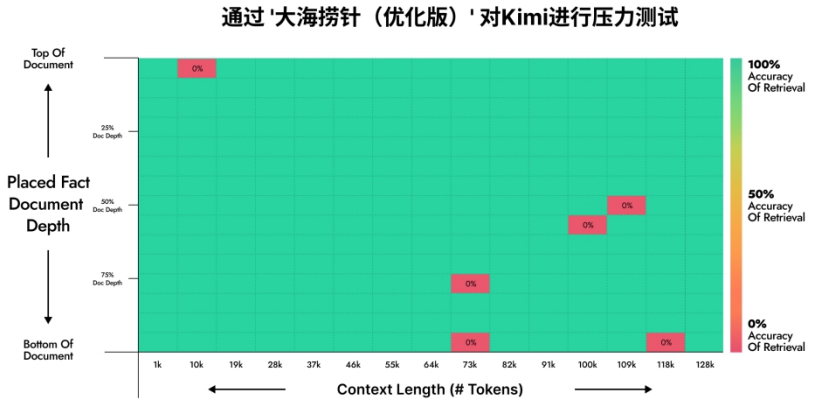
表2：全球玩家不断提升上下文窗口大小

公司	模型名称	发布时间	上下文长度(token 数)
OpenAI	GPT-4 Turbo	2023 年 11 月	128k
Google	Gemini 1.5 Pro	2024 年 2 月	1M（内测版本 10M）
Anthropic	Claude 3	2024 年 3 月	200k(内侧版本 1M)
华为	PanGu-Σ	2023 年 4 月	32k
百川智能	Baichuan 2	2023 年 10 月	192k（约 35 万汉字）
百度	文心一言	2023 年 10 月	2.8 万汉字
零一万物	Yi-34B	2023 年 11 月	200k
科大讯飞	星火 3.5	2024 年 1 月	8k
月之暗面	Kimi	2024 年 3 月	200 万汉字
阿里巴巴	Qwen-72B	2024 年 3 月	32k

资料来源：各公司官网、新浪科技、国信证券经济研究所整理

“大海捞针”测试验证了 Kimi 长文本能力。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，大模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-4 Turbo 和 Claude 2.1。

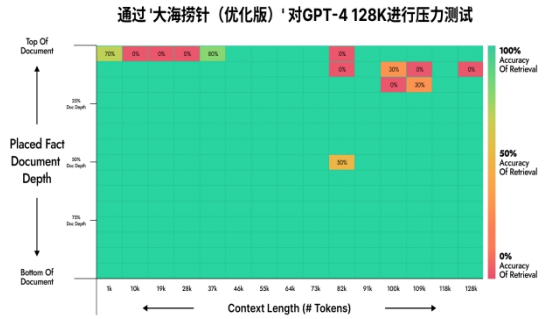
图8：Kimi“大海捞针”实验表现



资料来源：Moonshot AI，国信证券经济研究所整理

该测试在为英文语料，中文语料 Kimi 优势更为明显。月之暗面的工程师进行了多次实验，在以上英文版测试中，Kimi 已经证明其领先的长文本能力。同时在中文版的“大海捞针”实验中，Kimi 优势更为显著。该测试主要体现了大模型本身的长文本记忆能力和指令遵循能力；Kimi 在技术突破上选择了长下文方向，并取得了全球领先的水准。

图9：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

图10：Claude 2.1“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

应用体验得到广泛认可，B、C 两端均有突破

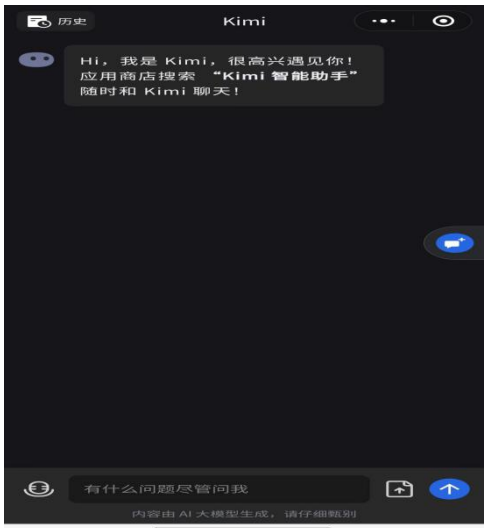
C 端：目标打造超级应用，致力于成为 AI 原生交互入口。月之暗面以实现通用人工智能（AGI）为目标，主要聚焦和发力在 C 端，以“通用性”打造超级 APP。公司以 C 端找到产品、技术、市场方向为最高优先级任务。Kimi 一经推出即受到市场广泛追捧，在应用端获得行业领先的用户体验。除了长文本优势之外，Kimi 联网搜索总结能力，一方面让信息交互获得最好的实时性，一方面有据可依的“参考资料”一定程度上能够解决用户的“幻觉焦虑”。因此 Kimi 也更容易拓宽应用场景，产品受到学术科研、互联网从业者、程序员、内容创作者、教育工作者、职场白领等人群的广泛认可。同时，公司认为 AI-Native 产品核心价值在于个性化交互，因此基于 Long Context 可以实现用户更多信息的保存，也更能对用户进行画像，并形成定制化交互；随着用户信息积累越多，模型也更容易实现精准推送，用户留存也会逐步提升。以 C 端入口来看，一旦 AI 形成用户粘性，有望对传统搜索、内容推荐等各类应用产生影响。

互联网打法快速占领用户心智，已经过多次扩容。在产品和运营领域，公司也配备了一批操盘过数亿 DAU 产品的产品经理和运营专家等人才。除了网页端和 APP 端之外，公司也推出了小程序；配合市场导流，公司小程序凭借极其便捷的交互体验，也形成了用户快速裂变。24 年 3 月下旬，随着用户访问量快速提升，Kimi 流量异常增高，Kimi 智能助手的 APP 和小程序一度无法正常使用，公司也进行了 5 次扩容。

图11：Kimi 尝试解决实时性和幻觉焦虑

资料来源：Kimi Chat，国信证券经济研究所整理

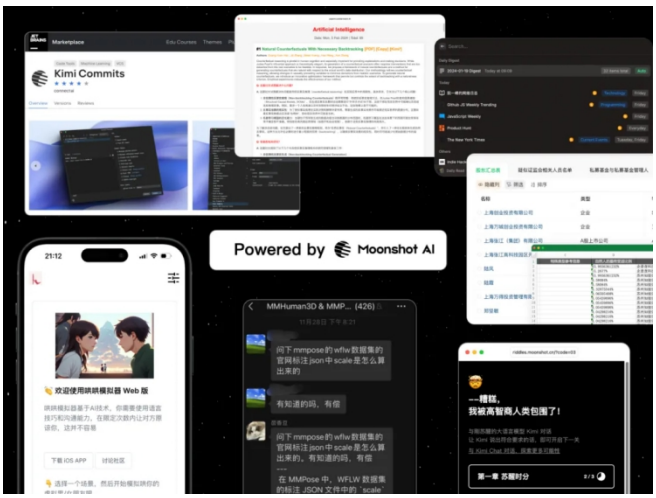
图12：Kimi 小程序



资料来源：Kimi Chat，国信证券经济研究所整理

B 端：打造 Moonshot AI 开放平台，API 与 OpenAI 兼容。月之暗面于 2024 年 2 月将 Moonshot AI 开放平台进行公测，为了方便快速接入 Kimi 同等能力的大模型，Moonshot AI 开放平台的 API 于 OpenAI 进行兼容，开发者无需对代码做参数意外的修改即可平滑千亿。在公测之前，已有数百位开发者在 Moonshot AI 开放平台内测打造了不少应用，这些均是未来 B 端拓展的可能性。例如自动生成代码修改记录的 Kimi Commits、做企业法律问题尽调的案牍 AutoDocs、高效阅读论文的 Cool Papers 等。其中“哄哄模拟器”最为火爆，其是一款角色扮演的小游戏，最高同时在线人数超过 1 万，消耗的 tokens 也很快过亿。该产品一开始接入的 GPT3.5，随后由于成本压力，切换为 Moonshot AI，但用户反馈依然很好。随着产品能力进一步认可，未来越来越多应用有望接入 Moonshot AI。

图13：接入 Moonshot AI 开放平台内测的应用



资料来源：Moonshot AI，国信证券经济研究所整理

Moonshot AI 定价优于 OpenAI，长文本场景是应用落地方向。目前公司提供三个基础模型 moonshot-v1-8k/32k/128k，每百万个 token 的定价为 12、24、60 元。与 GPT-4 Turbo 对比，其价格为每百万个 token 输入 10 美金、输出 30 美金，Moonshot AI 价格优势明显。结合公司最大优势在于“超长文本”的上下文，Moonshot AI 有望在长文本场景大放异彩，例如法律助手、科研助手、AI 阅读等是较好的落地场景。

表3：Moonshot AI 的 API 定价

模型	计费单位	价格（人民币）
moonshot-v1-8k	1M tokens	12.00
moonshot-v1-32k	1M tokens	24.00
moonshot-v1-128k	1M tokens	60.00

资料来源：Moonshot AI，国信证券经济研究所整理

Kimi 火爆再次拉动算力需求增长

训练算力测算：

训练过程：前向传播（Forward Pass）和反向传播（Backward Pass）。1）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。**2）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。**3）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。

训练算力计算假设及结果：

1）**模型参数量：**大模型参数量普遍在千亿参数量以上，但同时考虑到月之暗面作为初创公司，在公司发展初期，受资金成本、在手训练数据量、开发周期等多方面因素制约，我们认为其参数量应同 GPT-3（参数量为 1750 亿）相近，故假设 kimi 大模型参数量为 2000 亿。

2）**训练数据量：**训练数据量和模型参数量呈现正相关关系，即若模型参数量过大，训练数据量不足，会出现“训练不足”情况；同时，若模型参数量一定，训练数据量过大，会出现“训练过拟合“情况。我们已知 GPT-3 的训练数据量约 45TB（即 1750 亿参数的模型同 45TB 训练数据相配合），但考虑到公司对海外数据的可获得性受限，我们假设训练数据量为 30TB。同时，我们将 30TB 的训练数据量转化为 Token 数量，即约 4.9 万亿个 Tokens

3）**训练算力需求：**根据 AI 大模型训练算力公式，Kimi 大模型训练所需算力= 6 * Kimi 大模型参数量 * 训练数据量 = 6 *（2000 亿）*（4.9 万亿）= $5.9e^{24}$ Flops = 5.9 万亿 TFlops。

所需训练卡数及时间：考虑到国内英伟达 H100/H800 算力卡相对紧缺，我们假设

Kimi 使用英伟达 A100/A800 算力卡进行训练。单卡 A100/A800 在 TF32 精度下算力为 156 TFlops，假设芯片利用率为 30.2%，则 5.9 万亿 TFLOPs/(156 TFLOPs × 30.2% × 3600s × 24h/天 × 90 天)= 1.38 万张 A100/三个月，即使用 A100/A800 在 3 个月内完成一轮训练需要 1.38 万张卡。

表4：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和预测

表5：Kimi 训练算力测算

Kimi 模型参数量（亿）
2000

Kimi 训练数据量（TB）	Kimi 训练数据量（kb）	英文字母/字节（Unicode 规则）	英文字母数量（个）	平均 word 长度
30	37580963840	2	1.88E+13	5.1

对应 word 数量（个）	word/tokens 比例	对应 Tokens 数量（个）
3684408219608	0.75	4912544292810

对应训练需求测算（Flops）	训练算力需求（TFlops）
5.90E+24	5895053151373

训练算力卡需求（假设采用英伟达 A100 芯片）	A100 TF32 精度下算力（TFlops）	芯片利用率
	156	30.2%

3 个月内完成一轮训练所需卡数（张）
13793

资料来源：英伟达，国信证券经济研究所测算

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（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。

推理算力计算假设及结果：

1）**模型参数量：**如上文所述，假设 kimi 大模型参数量为 2000 亿。

2）**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。

3）**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 $2.4e^{9}$ 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 125 TFlops，假设芯片利用率为 30%，同时考虑白天高并发因素（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 $2.4e^9$ TFlops/(125 TFLOPs × 30% × 3600s × 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。

表6：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测算

阶跃星辰发布大模型，多模态能力领先

阶跃星辰多款大模型问世，参战国内大模型市场

阶跃星辰是通用大模型新势力，当前密集发布各类模型和产品。阶跃星辰自 2023 年 4 月成立以来，在算力、数据等领域深度布局，并以“智能阶跃，十倍每一个人的可能”为宗旨，自主研发人工智能通用（AGI）大模型。2024 年 3 月，公司正式对外宣布重要阶段性成果，发布了 3 款 Step 系列通用大模型，包括 Step-1 千亿参数语言大模型、Step-1V 千亿参数多模态大模型和 Step-2 万亿参数 MoE 语言大模型（预览版），并基于 Step-1、Step-1V 大模型推出 C 端产品“跃问”、“冒泡鸭”，旨在提高用户工作、学习效率并丰富生活娱乐体验。

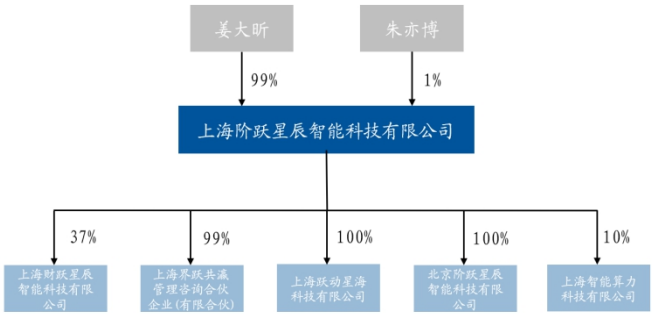
表7：阶跃星辰大模型及产品介绍

名字	类别	图例	定位
Step-1 大模型	大模型	Step-1	**千亿参数语言大模型**，专注于自然语言处理任务，如文本生成、语义理解、问答系统等。
Step-1V 大模型	大模型	Step-1V 千亿参数多模态大模型	**千亿参数多模态大模型，**结合了自然语言处理和计算机视觉能力，能够处理图像和文本的多模态任务。
Step-2 大模型（预览版）	大模型	Step-2 万亿参数语言大模型	**万亿参数 MoE 语言大模型**，参数规模翻倍扩大，进一步提升语言理解和生成能力。
跃问	To C 产品	跃问 个人效率助手 | 跃问越爱问	基于 Step 系列大模型的效率工具，**旨在提高用户专业领域的工作和学习效率**。
冒泡鸭	To C 产品	冒泡鸭 AI开放世界 | 探索故事，创造角色	基于 Step 系列大模型的 AI 开放世界平台，**旨在为用户提供沉浸式的 AI 交互体验**。

资料来源：公司官网，国信证券经济研究所整理

公司创始团队均是业界专家。创始人和 CEO 姜大昕博士是 NLP 领域全球知名专家，曾任微软全球副总裁、微软亚洲互联网工程院首席科学家。核心团队还包括系统负责人朱亦博博士和数据负责人焦斌星博士，朱亦博拥有多次单集群万卡以上的系统建设与管理实践经验；焦斌星曾担任微软必应引擎核心搜索团队负责人。截至 2024 年 3 月 27 日，公司创始人兼实际控制人姜大昕直接持有 99%的股份，股权高度集中。此外，公司系统负责人朱亦博作为公司联合创始人持有 1%的股份。同时公司在 B 端应用（参股财跃星辰）、算力等方面已经开始积极布局。

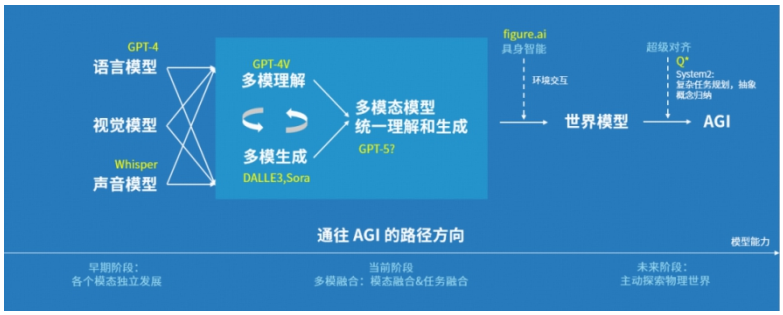
图14：阶跃星辰股权穿透图（截至 2024 年 3 月 27 日）



资料来源：企查查，国信证券经济研究所整理

多模态融合与生成统一，是 AGI 必经之路。公司认为要实现 AGI 的终极目标，模型必须经历“单模——>多模——>世界模型”的演化。目前，专注于语言、视觉和声音等单一模态的大模型正逐渐向多模态过渡，多模态输入通过先进的编码技术等实现在同一表征空间的整合。但是，大模型真正的具身智能还需要多模理解和生成的统一，即生成和理解统一在一个模型，然后在主动环境交互的加持下成为世界模型。最终，在其基础上加入 System2（复杂任务的规划能力和抽象概念的归纳）能力，模型就真正演化到了 AGI 阶段。

图15：阶跃星辰 AGI 发展路径



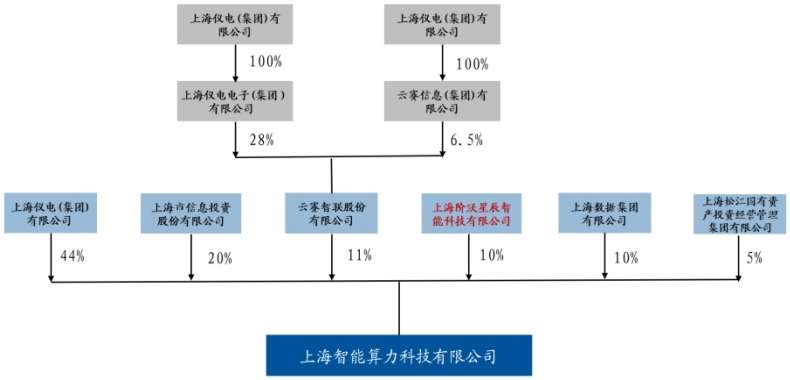
资料来源：阶跃星辰，国信证券经济研究所整理

应用和算力积极布局，AI 生态逐步构建

应用、算力等多维度持续布局。公司在算力、系统、数据和算法四项工程上持续突破。算力方面，公司通过“租用算力+自建厂房”的形式积极进行算力储备；系统方面，公司训练千亿模型的 MFU（有效算力输出）达 57%；数据方面，公司对全球互联网高质量语料的分布有深入了解，并建立强大的数据处理和知识图谱流

水线；算法方面，公司构建了万亿参数的 MoE 架构。公司同时在股权合作方面布局，公司与上海仪电、上海信投、云赛智联等共同参股上海智能算力科技有限公司，其中，阶跃星辰、云赛智联持股分别持股 10%、11%。拥有丰富的算力建设经验背书，该平台或将成为上海算力供给重要输出来源。

图16：上海智能算力科技有限公司股权结构（截至 2024 年 3 月 27 日）



资料来源：企查查，国信证券经济研究所整理

B 端数据筑基，多领域深度合作。数据是大模型训练的关键要素，其质量和规模直接影响大模型的语义理解与泛化能力，多样化的数据集则是对各种输入有出色鲁棒性和适应性的保障。公司创始人及骨干团队曾就职于必应搜索引擎，有多年的互联网语料分析经验，深谙如何利用全球互联网上高质量的语料来弥补中文语料的不足，并建立了完善的数据处理和知识图谱流水线。同时，公司在 B 端开展广泛的合作以优化数据质量、扩大数据规模。目前，公司已和中国知网、中文在线等头部知识服务领域的语料提供者达成战略合作，获取各个垂直领域丰富的数据资源。此外，公司与财联社联合创办公司财跃星辰并发布国内首个金融大模型平台——“财跃 F1”，聚焦智能运营、智能风控、智能投顾等多个应用场景，通过强大的通用图像处理和图表理解能力为金融机构提供优质的信息服务。

表8：阶跃星辰合作项目

领域	合作对象	内容
金融	界面财联社	联合创办财跃星辰，围绕财经资讯、智能投研、智能投顾等领域推进大模型的应用落地，发布国内首个千亿参数多模态金融大模型——“**财跃 F1 金融大模型**”。
网络文学	中文在线	探索大模型在灵感激发、内容创作等网络文学创作领域的应用。
知识服务	中国知网、中文在线	围绕大众知识服务等场景研究和推进大模型的应用。

资料来源：财联社，国信证券经济研究所整理

公司多模态大模型领先。跟据 OpenCompass 的测试，阶跃星辰的 Step-1V 多模态大模型性能已赶超 ChatGPT-4、Qwen-VL 等，位居榜单首位，大模型支持语音、图像、视频等多种输入，具备极强的推理能力与精确的超长文本理解能力。公司也提供开放平台，支持 200K 上下文，打造个人开发者和 B 端接入生态。

图17：Step-1V 多个测试集表现领先

	Method	Avg. Score	MMBench Test	MMBench-CN Test
1	Step-1V Verified StepFun	67.1	80.7	79.9
2	Qwen-VL-Max Verified Alibaba Group	65.8	77.6	75.7
3	GeminiProVision Verified Google	63.8	73.6	74.3
4	GPT-4v Verified OpenAI	63.2	77	74.4
5	InternLM-XComposer2-VL Verified Shanghai AI Lab	63	80.7	79.4
6	LLaVA-Next-Yi-34B Verified University of Wisconsin-Madison	62.3	81.1	79
7	Qwen-VL-Plus Verified Alibaba Group	60.2	67	70.7
8	DeepSeek-VL-7B Verified DeepSeek	55.6	73.8	71.4

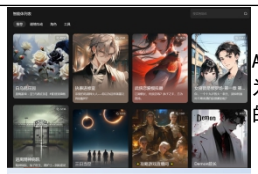
资料来源：OpenCompass，国信证券经济研究所整理

图18：Step-1V 大模型具备数学和推理能力

资料来源：公司官网，国信证券经济研究所整理

C 端探索差异化竞争路线。在此基础上，C 端产品“冒泡鸭”AI 聊天机器人与效率工具“跃问”实现差异化的产品定位。“冒泡鸭”用户可以在设定剧情下与 AI 角色进行互动，并创建智能体满足个性化需求，享受更加丰富和有趣的 AI 应用交互体验；“跃问”则面向专业领域的学生及工作人士，具备数据分析、文字推理等能力，并可通过自然语言交互实现信息查询、任务管理等服务。

表9：跃问、冒泡鸭差异化产品定位

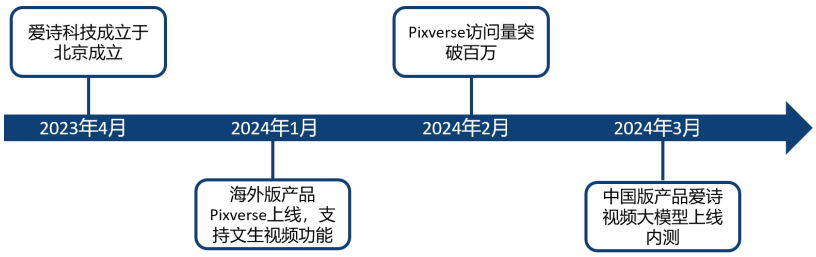
产品	界面	产品定位	目标用户群体	主要功能
冒泡鸭		AI 开放世界平台，为用户提供沉浸式的 AI 交互体验	对 AI 技术感兴趣的普通消费者、通过 AI 技术丰富自己娱乐体验的游戏玩家和内容创作者	提供丰富的 AI 角色和场景，用户可以与这些角色进行对话、完成任务、探索世界等互动；支持用户自定义角色和场景，满足个性化的创造需求
跃问		基于大模型的效率工具，帮助用户提高工作和学习效率	需要处理大量信息和知识的工作人士、需要高效学习的学生群体	提供强大的信息检索和知识图谱功能，用户可以快速获取所需的信息和知识；支持任务管理、日程安排等功能，帮助用户更好地规划和管理自己的工作和学习

资料来源：产品介绍，公司官网，国信证券经济研究所整理

Pixverse 引领全球 AI 视频，国内开启测试

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 4 月，是一家迅速崛起的 AI 视频生成大模型及应用企业。公司通过创新的“双融合”技术策略，即结合内容理解与生成，并整合文字、图片、视频等多种模态，致力于开发出全球领先的视觉多模态算法大模型，为视觉创意与智能生成带来新可能。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，已经成为全球用户量最大的国产 AI 视频生成产品。

图19：爱诗科技发展历程



资料来源：公司官网，国信证券经济研究所整理

公司团队豪华，主打视频领域。爱诗科技创始人王长虎深耕计算机视觉、人工智能领域 20 年；曾任字节跳动视觉技术负责人，参与了抖音和 TikTok 等产品从 0 到 1 的建设和发展，主导了字节跳动视觉大模型从 0 到 1 的建设；曾任微软亚洲研究院主管研究员。团队成员也多来自字节、微软亚洲研究院、快手、腾讯等头部机构的核心技术团队。2023 年 8 月，公司对外宣布完成数千万人民币的天使轮融资，剧报道快手创始人宿华参与投资。2024 年 3 月 11 日，公司再次获得来自达晨财智的亿级人民币 A 轮融资。

表10：爱诗科技融资历程

融资轮次	时间	投资金额	投资方
A 轮	2024 年 3 月	亿级人民币	达晨财智
天使轮	2023 年 8 月	数千万人民币	宿华（快手联合创始）等

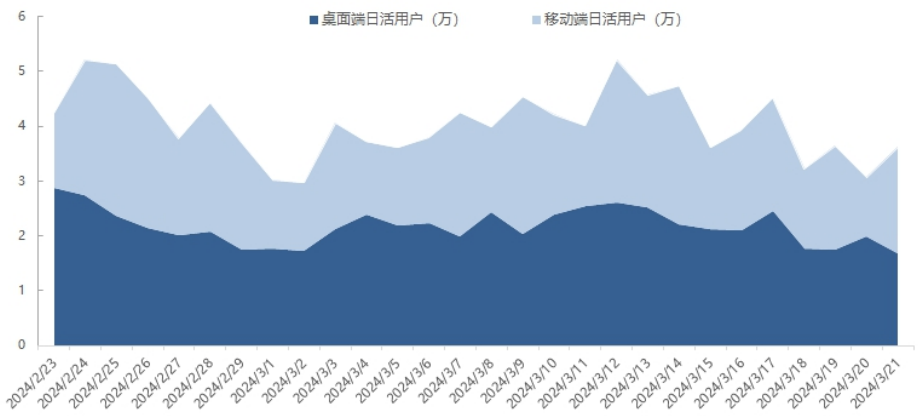
资料来源：天眼查，国信证券经济研究所整理

公司技术进步快速，目标 3-6 个月赶超 Sora。Sora 的出现引爆了文生视频的新内容业态，其采用了 DiT（Diffusion Transformer）的技术路线已得到业内认可。该技术架构灵活，显著提升了视频生成质量。而爱诗科技在创立之初就选择了这条路线，在 2023 年 10 月，PixVerse 就把生成的视频内容做到了 4K 的分辨率。得益于正确技术路线的选择，PixVerse 花了 3 个月的时间就做到了全球第一梯队的水平，资源和资金的消耗比 Runway、Pika 至少小了一个数量级。未来 3-6 个月，公司最重要的目标是技术上能够追平甚至赶超 Sora。

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。

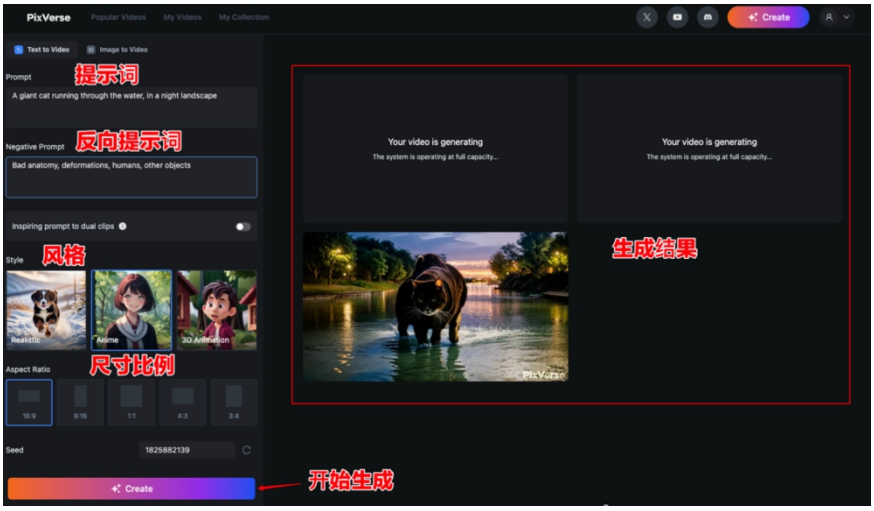
图20：Pixverse 日访问量



资料来源：Similarweb，国信证券经济研究所整理

扎根 AI 视频生成，实现多元模态融合。在 AI 大语言模型和图片生成领域，目前已有 ChatGPT、Midjourney 等产品依靠先发优势占据领先地位，而视频生成产品成熟度较低，尚在商业化落地早期阶段。公司也认为 Sora 的视频大模型技术相当于在 GPT2 和 GPT3 之间，还没到 GPT-4 的水平，市场仍有足够空间。一方面，相较于图文领域有 GPT、Stable-Diffusion 等通用开源模型，视频领域玩家仍在研究模型并筑建各自的算法优势；另一方面，如何突破 1 分钟的时长上限并满足一致性与准确性的要求，对公司技术升级、模型训练、算法创新等提出了不可忽视的挑战。爱诗科技瞄准这一细分领域，利用国内短视频数据积累、动态建模技术及团队骨干操盘上万块 GPU 的工程能力等优势，打造了多方面的产品护城河。在输入端，Pixverse 支持文本、图像、音频等模态，灵活性高；在输出端，Pixverse 支持 3D、动漫、现实等多种风格，同时向创作者提供提升视频分辨率、去噪等服务，多样性强。

图21：Pixverse 视频生成界面



资料来源：公司官网，国信证券经济研究所整理

投资建议

看好国产大模型持续突破，国内模型、应用、算力均迎来发展机会。2024 年国内大模型新势力异军突起，产品力和应用体验快速追赶全球头部模型水平，部分领域已经接近，甚至达到了全球第一梯队。以 Kimi 为代表的长文本能力、阶跃星辰的多模态模型、Pixverse 的 AI 视频生成，均验证了国内团队在一年多时间里取得的跨越式进步。随着模型能力持续迭代，国内在应用方面的创新性，算力国产化的进一步升级将带动 AI 生态进入正循环。例如，金山办公已经发布 WPS AI 的灰度定价，比 WPS 超级会员提价超过 120%，功能涵盖 AI 内容创作、AI PPT 等。国内 AI 应用商业模式进入落地期，国产算力也积极进入新一轮产品发布期，重点关注金山办公、同花顺、海光信息。

风险提示

宏观经济低迷影响 IT 支出；国内大模型技术突破不及预期；AI 相关商业化拓展不及预期；行业竞争加剧；相关政策进度不及预期。

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