



计算机行业研究

买入（维持评级）

行业专题研究报告

证券研究报告

计算机组

分析师：孟灿（执业 S1130522050001）

mengcan@gjzq.com.cn

分析师：李忠宇（执业 S1130524100002）

lizhongyu01@gjzq.com.cn

联系人：赵彤

zhaotong3@gjzq.com.cn

联系人：孙恺祈

sunkaiqi@gjzq.com.cn

DeepSeek 推动“算力平权”，关注端侧 AI 和 Agent 投资机会

投资建议

- DeepSeek 近期推出 V3/R1/Janus-Pro 等多款大模型，凭借卓越性能和创新技术，在行业内引发强烈反响，效果显著“出圈”。其中 V3 正式训练成本仅为 557.6 万美元，约为 GPT-4 预算的 1/20；R1 性能比肩 OpenAI o1，但 API 服务定价为每百万输入 tokens 1 元（缓存命中）/ 4 元（缓存未命中），每百万输出 tokens 16 元，远低于 OpenAI o1。我们认为 DeepSeek 的突破标志着国产大模型技术迈入全球领先行列，其低成本特性将带动 AI 应用加速落地。
- 在探讨大模型成本优化与算力需求之间的关系时，我们发现这两者并不是简单的此消彼长，而是相互促进、共同发展的关系。“杰文斯悖论”认为，随着蒸汽机效率的提升，煤炭的消耗量反而增加了。模型算力效率的突破，虽然看似会影响单位计算的价格，进而压缩高性能芯片供应商的利润空间，但从长远来看，这将加速 AI 的普及和创新，带来算力需求的更大量级提升。定价的持续走低有望带来更快的商业化落地，进而衍生出更多的微调和推理需求，逐步激活全球 AI 应用及算力发展。
- 在推理侧算力成本降低之后，“算力平权”将有望加速端侧 AI 和 Agent 应用落地。同时落地应用也带来网络安全及信创建设需求，相关板块将迎来投资布局机会。
 - 端侧 AI 方面，DeepSeek 推动端侧推理成本压缩，AI 眼镜、AI 耳机、AI 学习机、AI 玩具有望四面开花：1）AI 眼镜：AI 眼镜在软件技术方面与智能手机有较多相似之处，在手机领域布局较深的产业链公司有望收益：比如雷鸟 V3 使用虹软科技的 AI 视觉算法猎鹰影像系统；闪极 AI 拍拍镜由云天励飞提供万物识别功能、由科大讯飞提供复杂环境语音识别、多语言翻译等核心技术。2）AI 耳机：科大讯飞相关团队自 2011 年起专注 AI+耳机领域，最新推出的 Pro2 提供三种录音转写、vaim ai 助理、多语种翻译等功能，产品定位商务人群价格较高。若后续将生活娱乐、AI 语音交互等功能延伸至耳机端侧、或与其他硬件融合，则相关产品销量有望大幅提升。3）AI 学习机：学而思、作业帮、科大讯飞为线上销售额 top3，市场份额分别为 28%、25%、13%，其中，价格在 6,000 元以上的高端学习机多搭载更优质的教学资源 and AI 功能，学而思和科大讯飞分别位列线上市场的销量和销售额首位。4）AI 玩具：目前发布的产品核心功能基本一致，包括语音识别、自然语言处理和机器学习，希望通过拟人、拟动物，以及拟 IP 的方式，与用户进行视/听/触多维度交互，萤石网络、汉王科技、汤姆猫、蜂助手均有布局。
 - Agent 方面，指一种能够独立执行任务、做出决策并与其他系统或用户交互的软件程序或算法。智能体可以执行各种任务，从简单的自动化任务到复杂的决策支持系统。Agent 能够接受用户的自然语言请求，具有主动性，能够自动拆解任务并在多个应用程序中无缝协作，目前在企业服务、教育、医疗健康、智能制造、创意软件领域均有布局。
 - 以 DeepSeek 为代表的国产模型迅速崛起，也因此遭受了网络恶意攻击，这警示大模型厂商提升网络安全能力刻不容缓。国内多家网络安全公司积极响应 AI 技术进步带来的业务需求，同时也有望借助大模型能力实现防护能力提升和降本增效。此外，AI 大模型自主可控已迫在眉睫，除了软件核心组件自主可控以外，硬件产业链的国产替代需求旺盛。2025 年是十四五收官之年，预期信创推进政策及配套财政支持政策快速落地，预计上半年信创产业链相关公司的订单回暖将会出现重要信号。

投资建议

建议关注国内生成式大模型龙头科大讯飞；AI 硬件有望成为应用落地的新载体，建议关注萤石网络、虹软科技等；AI 相关功能打磨能够带动 C 端应用月活量、付费率提升，建议关注金山办公、万兴科技等。

风险提示

底层大模型迭代发展不及预期，国际关系风险，应用落地不及预期，行业竞争加剧风险。

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费



内容目录

1. DeepSeek 强势崛起，国产大模型技术大幅跃升	3
2. 推理侧算力成本降低之后：算力普及化井喷+端侧/Agent 加速落地	6
2.1 Deepseek 对算力影响几何？—推理成本较 o1 降数量级，有望推动算力普及化井喷	6
2.2 端侧 AI: Deepseek 推动端侧推理成本压缩，眼镜+耳机+学习机+玩具四面开花	9
2.3 AI Agent: 强推理模型性能持续提升，Agent 应用落地有望全面提速	12
2.4 AI 安全：大模型厂商急需全方位加强防护，网安及信创板块迎来布局机会	14
3. 投资建议	15
4. 风险提示	17

图表目录

图表 1: DeepSeek-V3 性价比远超 GPT-4o 和 Claude 3.5 Sonnet	3
图表 2: DeepSeek-R1 在 V3 之上再提升	3
图表 3: DeepSeek-R1 推理 API 定价远低于 OpenAI-o1	3
图表 4: DeepSeek 通过多种方式实现训练成本极致压缩	4
图表 5: R1 在 V3 基础模型上进行微调和强化学习	5
图表 6: Janus-Pro 7B 在理解和生成两方面都超越了 LLaVA、Dalle 3 和 SD XL 等主流选手	6
图表 7: 通用人工智能有望开启第四次工业革命	7
图表 8: 1860 年起英国煤炭消耗量上行约一个世纪	8
图表 9: 1920 年起美国原油消耗量大幅上行	8
图表 10: Deepseek 推动 AI 算力效率拐点显现	8
图表 11: 技术革命的周期性示意图	9
图表 12: 各品牌代表性 AI 眼镜产品情况	10
图表 13: 手机、互联网及上市公司的代表性 AI 耳机产品情况	10
图表 14: 24 年我国学习平板线上销量情况	11
图表 15: 科大讯飞占学习平板线上销售额的 13%	11
图表 16: 海内外 AI 玩具产品情况	11
图表 17: LLM 驱动的 GUI Agent 的发展历程	12
图表 18: AI agents 应用场景	13
图表 19: 全球基于 AI 的网络安全市场规模快速增长	14
图表 20: 永信至诚春秋 AI 测评“数字风洞”平台	15
图表 21: 建议关注 DeepSeek 相关产业链投资机会	16



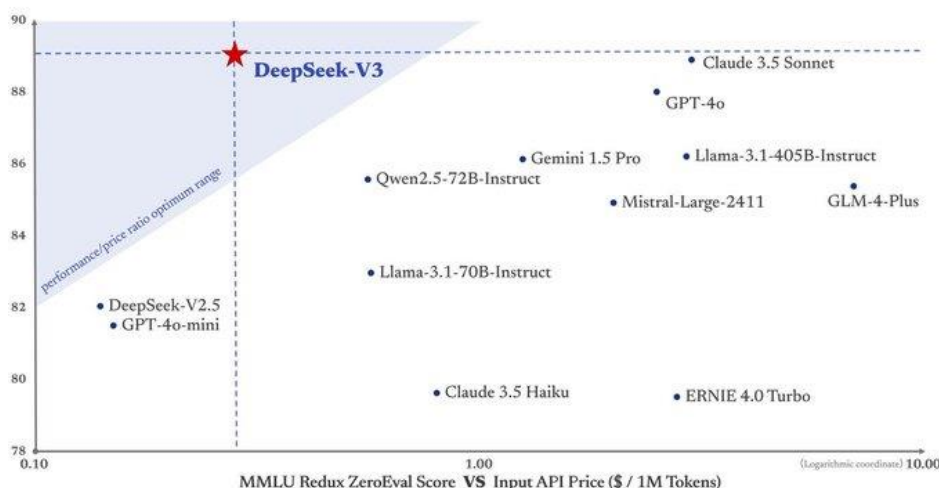
1. DeepSeek 强势崛起，国产大模型技术大幅跃升

DeepSeek 近期推出了多款大模型，凭借卓越性能和创新技术，迅速在行业内引发强烈反响，效果显著“出圈”。我们认为 DeepSeek 的突破标志着国产大模型技术迈入全球领先行列，其低成本特性将带动 AI 应用加速落地。

2024 年 12 月深度求索推出 DeepSeek-V3，以极低的训练成本，实现了与国际顶尖模型相媲美的性能。具体而言，DeepSeek-V3 在知识类任务（MMLU, MMLU-Pro, GPQA, SimpleQA）上的水平相比前代 DeepSeek-V2.5 显著提升，接近当前表现最好的模型 Claude-3.5-Sonnet-1022；长文本测评、代码生成、数学方面平均表现超过其他模型；中文能力与 Qwen2.5-72B 相近。

DeepSeek-V3 采用 MoE 架构，参数规模 671B，激活 37B，在 14.8T token 上进行了预训练。根据 DeepSeek-V3 论文，其正式训练成本仅为 557.6 万美元，约为 GPT-4 预算的 1/20。模型完全开源，API 价格为每百万输入 tokens 0.1 元（缓存命中）/ 1 元（缓存未命中），每百万输出 tokens 2 元，显著低于 Claude 3.5 Sonnet 和 GPT-4o。

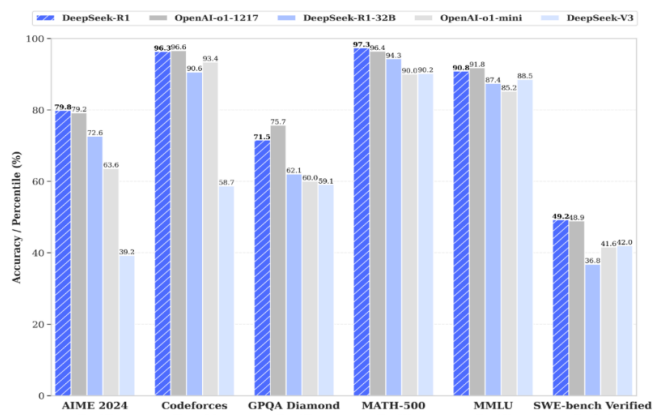
图表1：DeepSeek-V3 性价比远超 GPT-4o 和 Claude 3.5 Sonnet



来源：DeepSeek 公众号，国金证券研究所

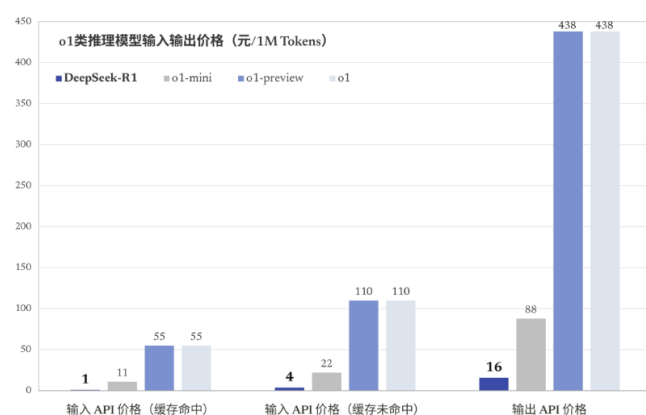
2025 年 1 月 20 日，深度求索发布正式版 DeepSeek-R1 模型，是基于 V3 基础模型进行微调和强化学习得到的，模型同步开源并实现性能再度提升，在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版。此外，深度求索还在 R1 的基础上蒸馏了 6 个开源小模型，其中 32B 和 70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果。DeepSeek-R1 API 服务定价为每百万输入 tokens 1 元（缓存命中）/ 4 元（缓存未命中），每百万输出 tokens 16 元，价格远低于 OpenAI-o1。

图表2：DeepSeek-R1 在 V3 之上再提升



来源：DeepSeek 公众号，国金证券研究所

图表3：DeepSeek-R1 推理 API 定价远低于 OpenAI-o1



来源：DeepSeek 公众号，国金证券研究所



DeepSeek-V3 和 R1 均具备低成本和高性能的特点，在使用经典 Transformer 解码器模块堆叠和 GPT-2/3 的架构范式基础上进行了多种算法创新。

➤ **低成本：**DeepSeek-V3/R1 模型降低训练成本主要依赖模型压缩、并行计算和提升硬件使用效率等方法。

- 模型压缩：包括使用多层注意力（Multi-Layer Attention, MLA）和 FP8 混合精度训练等。

1) MLA：传统 Transformer 中每一层都需要完整的计算和存储，其中的 Key 和 Value 矩阵往往占用大量内存空间，而 MLA 通过压缩和复用前序层的 K、V 来减少内存占用和计算量。DeepSeek 通过 Key/Value 的低秩压缩使得的训练内存占用减少了 20-30%，此外还通过对 Query 的低秩压缩，减少了计算过程中的激活内存占用。

2) FP8 混合精度训练框架：FP8 用 8 个二进制位来表示数字的格式，相比传统的 FP32 和 FP16 格式，精度有所降低，但是占用空间小、计算快。DeepSeek 在采用 FP8 格式时，采用了“混合精度”的方案，即在 8 位/16 位/32 位之间切换，相较原始方法计算速度提升一倍，并显著降低内存消耗。

- 并行计算：在系统架构层面，DeepSeek 使用了专家并行训练技术，通过将不同的专家模块分配到不同的计算设备上同时进行训练，提升了训练过程中的计算效率。在计算和通信重叠方面，DeepSeek 创新使用 DualPipe 方法进行多任务并行处理，能够让模型在进行计算的同时，在后台已经开始准备下一步需要的数据传输。这种设计确保了通信开销被很大程度地隐藏在计算过程中，极大提升了整体效率。根据 DeepSeek 的技术报告，DualPipe 算法减少了 50% 的计算气泡，有效隐藏了通信开销。

此外，DeepSeek-V3/R1 还进行无辅助损失负载均衡、跨节点通信优化，全面提升模型训练效率、降低训练成本。

图表4：DeepSeek 通过多种方式实现训练成本极致压缩

技术创新		训练效率影响	推理效率影响	模型能力影响	部署友好性影响
MLA 架构	Key/Value 低秩压缩	减少训练内存和计算开销	降低 KV 缓存内容	保持模型性能	降低部署内存需求
	Query 低秩压缩	降低激活内存占用	-	保持模型性能	-
FP8 混合精度训练		简化训练流程,提升训练稳定性	-	精度损失 <0.25%	-
DualPipe 算法		减少 50%流水线气泡,优化通信开销	-	-	-
无辅助损失负载均衡		提升训练稳定性,避免序列内平衡开销	-	增强专家分化效果	简化部署流程
跨节点通信优化		提升贷款利用率	-	-	支持大规模部署

来源：腾讯科技公众号，国金证券研究所

➤ **高性能：**DeepSeek-V3 模型的高性能源于参数量较大、训练数据质量高，并使用 MTP 技术进行架构革新，而 R1 在 V3 基础模型上进行蒸馏，进一步提升了模型能力。

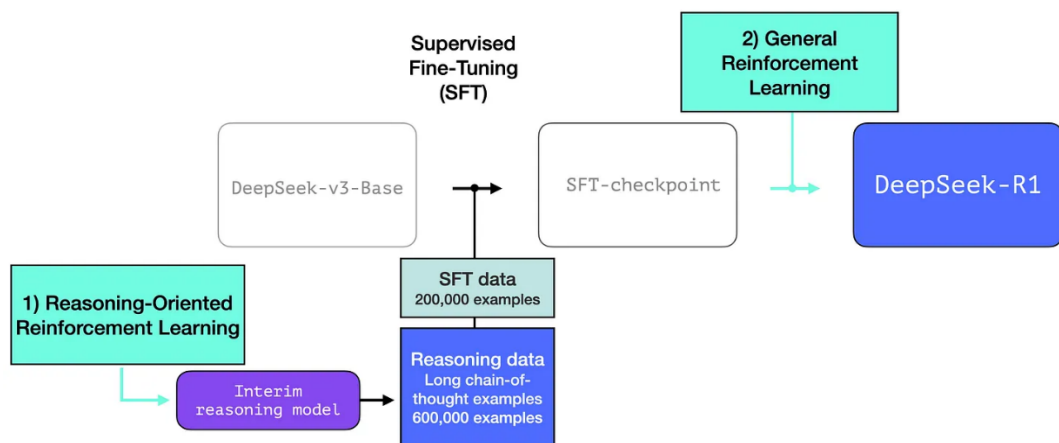
- 参数量：DeepSeek-V3 的总参数量有 671B，每个 token 激活 37B 参数。整体参数总量比 Llama 3.1 的 405B 高，也远超 Qwen 2.5 的 72B。在 Scaling Law 尚未碰壁的情况下，较高的模型参数量利于性能提升。
- 数据精筛：DeepSeek-V3 采用了更多元化的数据获取策略。基础训练数据来源于经过严格筛选的 CommonCrawl 语料库，研发团队还引入大规模的代码数据集、数学推理数据、科学文献等。在数据清洗环节，DeepSeek 采用了专有的数据过滤算法，删除低质量、不完整的数据。团队还通过数据混合采样策略和课程学习方法，进一步优化了训练过程中的数据使用效率。
- 多 token 预测（MTP）：传统语言模型一次只预测一个 token 的范式，使用 MTP 后模型会同时预测多个连续位置的 token。这种并行预测机制不仅提高了训练效率，还让模型能够更好地捕捉 token 之间的依赖关系。在保持输出质量的同时，



模型整体性能提升 2-3%。

- SFT 微调和大规模强化学习: R1 在 V3 的基础模型上进行 SFT 微调 and 强化学习, 继承了 V3 的低成本特性, 并在推理方面表现更优。R1 的训练分为两个阶段:
 - 1) 有监督微调 (SFT): 这一环节使用 2 类数据, 一是来自中间推理模型生成的 60 万个长思维链推理示例, 二是 20 万个来自 V3 模型生成的与非推理训练样本相关的数据。其中, 中间模型是使用强化学习方法对 V3 的基础版本进行训练得到的, 其推理能力已经与 OpenAI o1 相近, 在训练过程中模型思维链长度逐步增加。在微调过程中, DeepSeek 还引入数千条高质量、包含长思维链的冷启动数据, 显著提升了模型的可读性和多语言处理能力。
 - 2) 通用强化学习: 针对非推理类应用场景进行了扩展优化——通过引入实用性奖励模型与安全性奖励模型 (类似 Llama 模型的机制) 对相关应用的提示进行多维评估。这种复合奖励机制确保模型在扩展应用边界时, 既能保持输出有效性, 又能遵循安全伦理规范。

图表5: R1 在 V3 基础模型上进行微调和强化学习



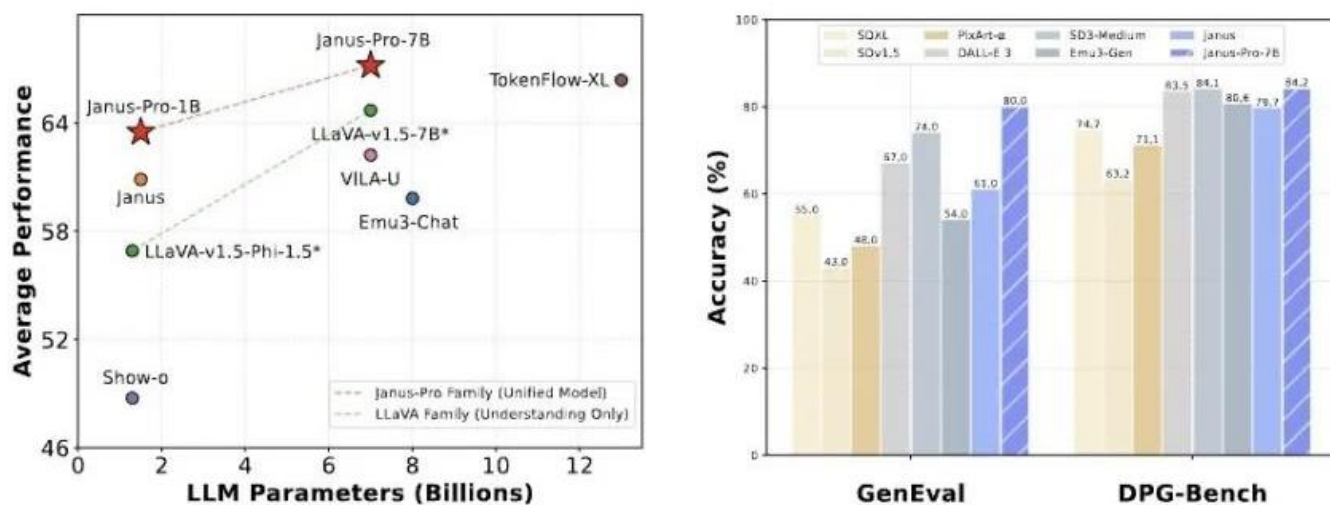
来源: Language Models & Co., 国金证券研究所

除 V3/R1 外, DeepSeek 在 2025 年 1 月 28 日凌晨, 再发两款多模态框架——Janus-Pro 和 JanusFlow。图像理解和图像生成统一的多模态模型训练难度较高, 而 DeepSeek 通过为模型配置双头编码器, 分别单独负责理解和创作, 成功突破训练瓶颈; 规避了 ImageNet 数据集不真实的问题, 直接使用真实的文生图数据进行训练, 使得训练时间减少 40%、生成质量提升 35%、模型对真实场景的适应性大幅提升。

- Janus-Pro 是一款统一多模态理解与生成的创新框架, 包括 1.5B 和 7B 两个模型, Janus-Pro 运用 Transformer 架构, 将文本、图像、音频等多种模态的数据进行统一处理, 实现对不同模态信息的理解与生成。在图像生成领域, Janus-Pro 在基准测试中表现优于 OpenAI 的 DALL-E 3 以及 Stable Diffusion 3 Medium 等; 在多模态理解领域, 其在 MMBench 评分超过此前最佳水平。
- JanusFlow-1.3B 是多模态理解模型, 参数量仅为 1.3B, 将基于视觉编码器和 LLM 的理解框架与基于 Rectified Flow 的生成框架直接融合, 实现了两者在单一 LLM 中的端到端训练。JanusFlow-1.3B 在视觉理解和生成任务上均超过此前同规模的统一多模态模型。



图表6: Janus-Pro 7B 在理解和生成两方面都超越了 LLaVA、Dalle 3 和 SD XL 等主流选手



来源：腾讯科技公众号，国金证券研究所

2. 推理侧算力成本降低之后：算力普及化并喷+端侧/Agent 加速落地

2.1 Deepseek 对算力影响几何？—推理成本较 o1 降数量级，有望推动算力普及化并喷

Deepseek 性能对齐 OpenAI 最新模型水平，推理成本相比 o1 降数量级。Deepseek 的推理成本低至每百万 Token 0.14 美元，相比 OpenAI-o1-mini 低了一个数量级。Deepseek 的 LLM（如 V3、R1）基于混合专家（MOE）和多重潜在注意力（MLA）架构，显著降低训练和推理成本（训练成本仅为 Meta Llama 的 10%，推理成本为 OpenAI 的 1/7）。

在探讨大模型成本优化与算力需求之间的关系时，我们发现这两者并不是简单的此消彼长，而是相互促进、共同发展的关系。微软首席执行官曾引用“杰文斯悖论”来阐释这一现象，杰文斯在其著作《煤炭问题》中指出，随着蒸汽机效率的提升，煤炭的消耗量反而增加了。这一悖论的核心在于：技术进步提高了资源的使用效率，而效率的提高降低了资源的使用成本，进而刺激了资源需求的增长。这种增长有时甚至超过了效率提升所带来的节约，最终导致资源总消耗量的增加。

我们认为，工业革命的本质是以新一代生产力工具的大范围应用为核心，以主体能源为基础动力的全局性经济范式转型。历史上共发生过三次工业革命，我们认为，本轮通用型人工智能发展浪潮有望启动第四次工业革命：

- 第一次技术革命：生产力工具：蒸汽机；主体能源类型：煤炭。
- 第二次技术革命：生产力工具：内燃机；主体能源类型：石油。
- 第三次技术革命：生产力工具：计算机；主体能源类型：电力。
- 第四次技术革命：生产力工具：通用人工智能；主体能源类型：算力。



图表7：通用人工智能有望开启第四次工业革命

第 X 次工业革命	生产力变革	能源动力变革	关键事件驱动
第一次	蒸汽机	煤炭	1763~1775 年，瓦特在马修·博尔顿的支持下着手改良蒸汽机；1776 年，瓦特设计的蒸汽机进行商业化应用。1829 年，蒸汽动力机车“火箭号”在利物浦到曼彻斯特的铁路上试验成功。全国性铁路网络的建设促进了物流效率的提升与社会分工的加速，进而推动了标准零部件生产的规模经济雏形诞生。
第二次	内燃机	石油	1860 年，比利时工程师艾蒂安·勒努瓦以蒸汽机为蓝本，制成了首台以天然气为燃料的燃气发动机。1862~1876 年，德国工程师尼古拉斯·奥托将内燃机的热效率提升至 10%；1876 年，奥托制成了四冲程循环的内燃机。1885 年，戈特利布·戴姆勒制成了第一台汽油机，并于次年造出第一辆用汽油机驱动的汽车。1908 年，第一辆 T 型车从密歇根州底特律的福特工厂下线，福特汽车开启了大规模流水线装配的生产时代，为规模经济的工厂组织形态奠基。1924 年，意大利建成了世界上第一条高速公路 A8，连接米兰和瓦雷泽，全长约 40 公里，标志着高速公路时代的开始。1956 年，美国通过《联邦援助公路法案》，开工建设彼时世界上最大的公路网络。
第三次	计算机	电力	1943-1945 年，美国的 ENIAC（电子数值积分计算机）完成，这是第一台完全电子的通用计算机，能够重新编程以执行各种任务。1971 年，全球第一款商用微处理器英特尔微处理器 4004 问世。1974 年，罗伯特·卡恩和文顿·瑟夫提出 TCP/IP，定义了电脑网络之间传送报文的方法。1983 年，苹果电脑公司推出全球第一款搭载图形界面的个人电脑 Apple Lisa。1990 年，蒂姆·伯纳斯-李创建了第一个网页浏览器、第一个网页服务器和第一个网站，开启 PC 互联网时代。2007 年，初代 iPhone 面世，开启移动互联网时代。两轮互联网革命将实现了全球范围内的信息交流与远程通讯，降低沟通成本，促使全球人员协作成为可能。
第四次	通用人工智能	算力	1997 年，IBM 深蓝战胜国际象棋世界冠军加里·卡斯帕罗夫。2016 年，AlphaGo 战胜围棋世界冠军李世石。2017 年，《Attention Is All You Need》发表，Transformer 模型问世，后续成为 AI 大模型的主流选择。2022 年，OpenAI 旗下大语言模型产品 ChatGPT 问世。2023 年，OpenAI 旗下多模态大模型 GPT4 与 Google 旗下多模态大模型 Gemini 相继问世。2024 年，OpenAI 旗下 Text to Video/Image 大模型 Sora 问世。大模型时代动辄千亿参数的训练需要巨量的算力资源支撑，以 GPU 为代表的芯片或将成为通用人工智能时代的新能源。AI 与软件结合有望替换文书、设计与管理工作，AI 与硬件结合以机器人形态有望取代体力劳动工作。

来源：Wikipedia, Carlota Perez 《技术革命与金融资本——泡沫与黄金时代的动力学》，中国科学院《人工智能的历史、现状和未来》，国金证券研究所

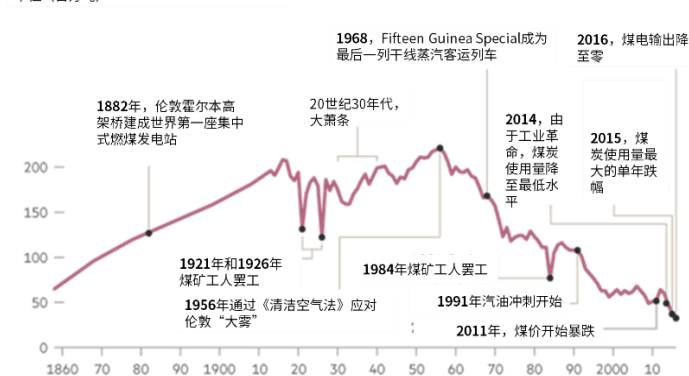
历史上历次技术革命所涉及的主体能源都会随新一代生产力工具广泛应用而呈现能源需求持续上行的阶段，以煤炭和石油两种一次化石能源为例说明：

- 以瓦特蒸汽机的大范围应用为代表的第二次技术革命最早在英国展开，后续逐渐扩散至欧洲大陆和美国。1860 年起，英国煤炭消耗量上行约一个世纪，达峰之后逐渐回落，1860~1920 年英国煤炭消耗量的快速爬升主要受益于蒸汽机在铁路运输、纺织、制造、采矿等各行业的广泛应用（1882 年以后同时受益于煤电厂的兴建）。
- 以内燃机的大范围应用为代表的第三次技术革命最早在美国和德国展开，后续逐渐扩散至欧陆各国。1920 年起，美国原油生产量持续保持高速爬升，1970s，美国国内石油开采量放缓，对海外原油进口飙升，国内石油需求持续增长。



图表8：1860年起英国煤炭消耗量上行约一个世纪

英国历史煤炭使用情况
单位：(百万吨)



来源：FT《Britain passes historic milestone with first days of coal-free power》，
国金证券研究所

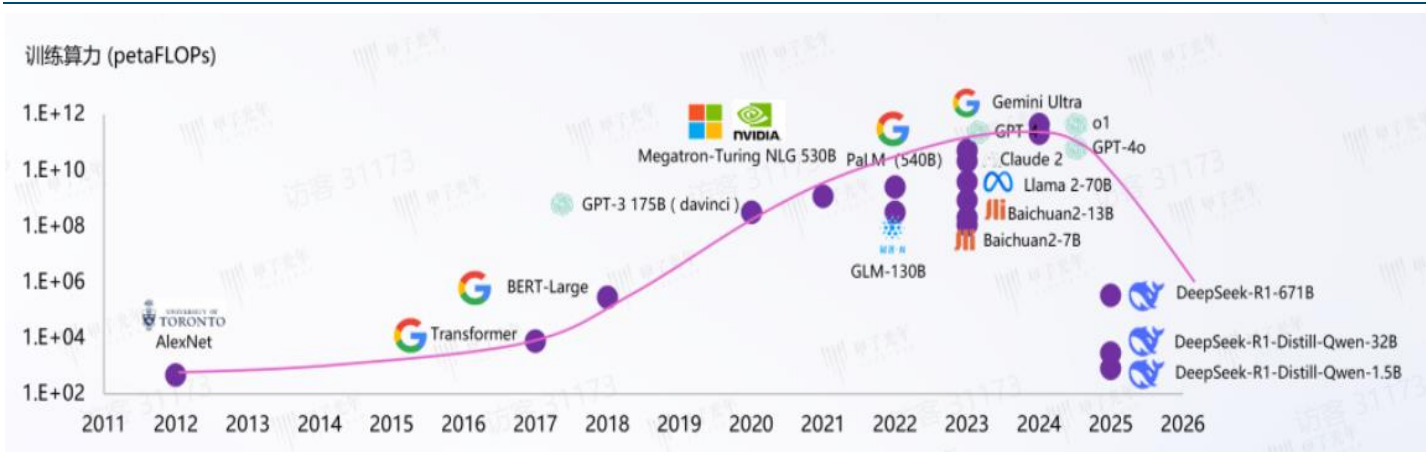
图表9：1920年起美国原油消耗量大幅上行



来源：U.S. Energy Information Administration, Annual Energy Review 2009，
国金证券研究所

在当今数字经济中，算力的重要性堪比“水电煤”。Deepseek 的技术方向和开源定位，正是推动算力普及化革命的关键因素。我们认为，高资源使用效率不仅不会减少算力的消耗，反而可能因为使用成本的降低，刺激了更大的算力需求，最终导致算力总消耗量的增加。

图表10：Deepseek 推动 AI 算力效率拐点显现

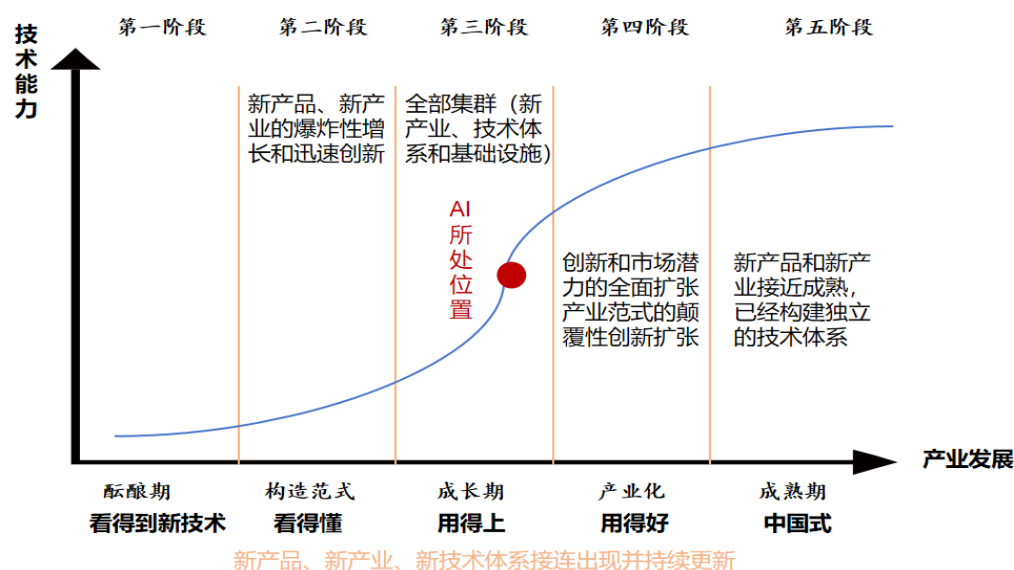


来源：甲子光年智库，国金证券研究所

模型算力效率的突破，虽然看似会影响单位计算的价格，进而压缩高性能芯片供应商的利润空间，但从长远来看，这将加速人工智能的普及和创新，带来算力需求的更大量级提升。定价的持续走低有望带来更快的商业化落地，进而衍生出更多的微调和推理需求，逐步激活全球 AI 应用及算力发展。因此，技术进步提高了资源使用的效率，不仅没有减少这种资源的消耗，反而因为使用成本降低，刺激了更大的需求，最终导致资源使用总量反而上升。



图表11：技术革命的周期性示意图



来源：甲子光年智库，国金证券研究所

2.2 端侧 AI：Deepseek 推动端侧推理成本压缩，眼镜+耳机+学习机+玩具四面开花

Deepseek 对推理算力的压缩能够降低端侧 AI 成本。近日，英特尔公布，Deepseek 目前能够在英特尔产品上运行，也可以在 AI PC 上实现离线使用。在其最新发布的酷睿 Ultra 200H（Arrow Lake H）平台上，Deepseek-R1-1.5B 模型能够顺利运行，在因式分解演示中，能够迅速演绎逻辑思维，并最终解决数学难题，突破难关。小米、比亚迪等企业迅速接入 API，特斯拉中国区客服宣布测试 Deepseek 语音助手，称其“中文理解能力优于现有系统”。OPPO、荣耀宣布年内推出搭载 Deepseek 轻量版模型的智能眼镜，可实现实时同声传译和 AR 导航。

- AI 眼镜：**有望成为最能承载 AI 应用的新硬件产品，参与玩家众多，深耕手机算法的计算机厂商有望受益。目前该赛道的玩家包括 AR 眼镜厂商、互联网大厂、手机厂商、AI 创业公司等，在 2025 CES 中参展的 AI 及 AR 眼镜数量达到 47 个，提供的主要功能包括拍摄、AR 显示、语音交互、翻译等，竞争较为激烈。根据 Wellseenn 预测，2025 年全球 AI 智能眼镜将进入新品密集发布期，出货量有望达 400 万副，到 2030 年出货量有望增长至 8000 万副。AI 眼镜在软件技术方面与智能手机有较多相似之处，在手机领域布局较深的产业链公司有望收益：比如雷鸟 V3 使用虹软科技的 AI 视觉算法猎鹰影像系统；闪极 AI 拍拍镜由云天励飞提供万物识别功能、由科大讯飞提供复杂环境语音识别、多语言翻译等核心技术。


图表12：各品牌代表性 AI 眼镜产品情况

发布时间	品牌	产品名称	价格	主要功能
2012	谷歌	Google Glass	1500 美元	拍摄、AR 显示、语音交互、AR 导航、语音播报
2022.8	小米	小米家眼镜相机	2699 元	AR 显示、拍摄、翻译、语音交互
2022.9.21	亮亮视野	听语音 AR 字幕眼镜	2999 元	AR 显示、语音交互、AI 实时字幕
2022.9.30	雷朋&Meta	Meta RayBan	299 美元	拍摄、语音交互、AI 问答、AI 翻译
2023.9	华为	华为智能眼镜 2	1699 元	语音播报、语音交互
2024.2.26	OPPO	OPPO Air Glass 3		AR 显示、语音交互、AI 能力
2024.4.26	李未可	Meta Lens Chat	699 元	AI 语音交互、AI 问答、翻译、语音导航
2024.11.12	百度	小度 AI 眼镜		拍摄、语音交互、AI 能力
2024.11.18	Rokid	Rokid Glasses	2499 元	拍摄、AR 显示、语音交互、AI 识物、AI 问答、AI 助手、AI 导航
2024.11.29	影目	INMO GO2	3299 元	AR 显示、语音交互、AI 同传翻译
2024.12.19	闪极	AI 拍拍镜	1499 元	拍摄、语音交互、AI 助手、录音
2025.1.7	雷鸟	雷鸟 V3	1699 元	拍摄、AI 问答、AI 搜索、AI 识物、AI 实时翻译、AI 记录提醒

来源：智东西公众号，国金证券研究所

- AI 耳机：**主要提供办公效率辅助功能，多款产品接入第三方大模型。目前 AI 耳机赛道中的玩家包括：1) 传统手机厂商如华为、小米、三星等，AI 耳机通常与手机绑定，可以用于唤醒手机端智能体；2) 传统耳机制造商，如飞利浦、纽曼等，使用腾讯、百度等第三方大模型，可唤醒对应的手机 APP；3) 互联网及 AI 厂商，科大讯飞相关团队自 2011 年起专注 AI+耳机领域，最新推出的 Pro2 提供三种录音转写、viam ai 助理、多语种翻译等功能，产品定位商务人群价格较高。各家 AI 耳机均定位为生产力工具，提供同传翻译、语音转文字等功能，功能较为局限，若后续能够拓展至生活娱乐场景、AI 语音交互等功能延伸至耳机端侧、或与其他硬件融合，则相关产品销量有望大幅提升。

图表13：手机、互联网及上市公司的代表性 AI 耳机产品情况

品牌	产品名称	价格	主要功能	上市时间
华为	FreeBuds Pro4	1499 元	同传、小艺智能体	2024.11
小米	Xiaomi Bus 5	699 元	语音转文字、生成会议纪要	2024.7
三星	Galaxy Buds 3	999 元	同传、多语种翻译	2024.7
飞利浦	飞利浦 8 号	1199 元	接入腾讯混元大模型、录音转文字、AI 会议总结	2024.5
字节跳动	Ola Friend	1199 元	接入豆包智能助手、旅行导游、英语陪练	2024.10
百度	小度 G108 耳机	499 元	同传、连接小度 APP	2024.10
科大讯飞	IFLYBUDS Pro2	1969 元	录音转写、同传听译、AI 会议助理	2024.5

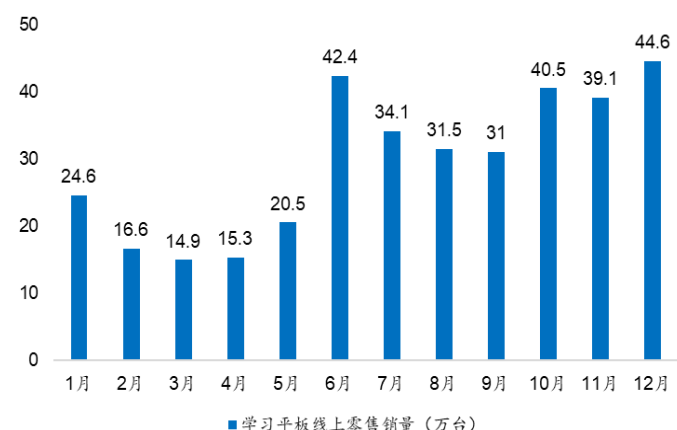
来源：定焦 One 公众号，国金证券研究所

- AI 学习机：**学习平板销量稳步增长，融合 AI 功能的高端学习机占比提升明显。根据



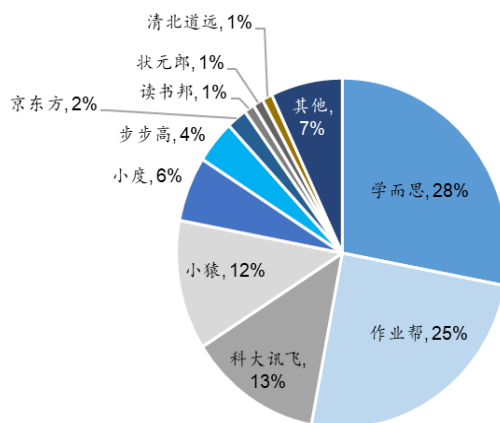
洛图科技数据，2024 年，中国学习平板市场全渠道销量为 592.3 万台，同比 2023 年增长 25.5%，受到学生寒暑假影响，6 月、12 月销量较高；全年销售额为 190.6 亿元，同比 2023 年增长 37.6%。学习平板线上市场的零售量达 393.7 万台，在全渠道中的占比为 66.5%；学而思、作业帮、科大讯飞为线上销售额 top3，市场份额分别为 28%、25%、13%。从价格区间来看，价格在 6,000 元以上的高端学习机多搭载更优质的教学资源 and AI 功能，24 年其线上市场份额达 12%，较 23 年增加 5.4pct，在这一细分市场中，学而思和科大讯飞的贡献较大，分别位列线上市场的销量和销售额首位。

图表14：24 年我国学习平板线上销量情况



来源：奥维睿沃，国金证券研究所

图表15：科大讯飞占学习平板线上销售额的13%



来源：奥维睿沃，国金证券研究所

- **AI 玩具：**目前发布的产品核心功能基本一致，包括语音识别、自然语言处理和机器学习，希望通过拟人、拟动物，以及拟 IP 的方式，与用户进行视/听/触多维度交互。受到功能和 IP 溢价影响，AI 玩具的价格差异较大。相较于 AI 眼镜、AI 学习机、智能音箱等产品，AI 玩具依托玩偶、宠物和知名 IP 为载体，能够提供更多陪伴和情感支持，该领域更容易出现爆款产品，比如官方售价 449 元的 AIGC 玩具挂件 BubblePal，开售首月 GMV 突破 400 万。

图表16：海内外 AI 玩具产品情况

品牌	产品名称	价格	主要功能	上市时间
字节跳动	显眼包	未公布	内嵌豆包大模型、扣子专业版等 AI 技术，可以与人进行对话交互和情感交流	2024.9
Vanguard	Moflin	2779 元	聊天，与主人交流后能以多种方式改变自己的情绪，包括愤怒、悲伤、快乐，并像真实的生物一样做出反应	2024.11
Folotoy	Fofu、Catus、Kola、Magicbox 等	10-1299 元	可接入大模型的 AI 玩具，内置 7 种角色，家长可设计角色对话和克隆声音	2024.4-11
亿家亿伴	爱小伴	899-1399 元	基于专有儿童语料，提供信息交互、外部知识、长期记忆，可以陪孩子讲故事、过家家、主动情绪互动等	2024.12
奥飞娱乐	喜羊羊超能铃铛娃娃	399 元	结合“喜羊羊”形象和小冰大语言模型，提供 AI 智慧畅聊	2024.5
实丰文化	AI 魔法星	未公布	支持面向家庭提供 AI 故事共创、定制爸爸妈妈的声音讲故事、AI 全场景问答和连续对话等功能	2025.1
Haivivi 跃然创新	BubblePal	399 元	魔法灯泡挂件，可让玩具开口说话，通过 App 控制，结合多种模型保证响应和对话质量	2024.7
汉王科技	Go Go Bird	未公布	主要面向青少年人群，主攻 STEAM 教育和航模方向，是一款寓教于乐的智能性玩具。产品模块化的	2021.9



			设计，可以使孩子在产品组装过程中对机械组装有系统化的思考以及实质性的操练。	
萤石网络	RK3	未公布	RK3 家庭助理机器人作为 AI 代理的载体，基于多模态大模型，可以化身家庭管家、保安、医生以及教师，将萤石“悉心关爱每位家庭成员”的善意蕴藏其中。它拥有“老人专属看护服务”，可以为老人提供吃药提醒、跌倒检测、线上视频问诊等特色功能；同时还有“儿童专属服务”，让小朋友可以畅学全国各地名师课堂，为他们进行绘本阅读、作业批改，并帮助家长敦促孩子的坐姿等。	2024.4
汤姆猫	AI 童伴机器人	1499 元起	为实现汤姆猫的“说话”功能，汤姆猫 AI 童伴机器人，内置了专门研发的汤姆猫情感陪伴垂直模型。该模型由汤姆猫和其投资的大模型科技公司西湖心辰联合推出。研发团队训练了大量针对汤姆猫这一 IP 角色的人物设定、声音数据以及儿童情感陪伴领域的语料。	2024.12
蜂助手	云手机	未公布	云手机通过“AI+RPA”赋能银行客户运营模式，实现银行客户全生命周期服务策略，升级客户服务体验。	2024.8

来源：定焦 One 公众号，爱小伴 AliPal 公众号，FounderPark 公众号，贝壳财经公众号，无人机网公众号，子弹财经公众号，iFind，国金证券研究所

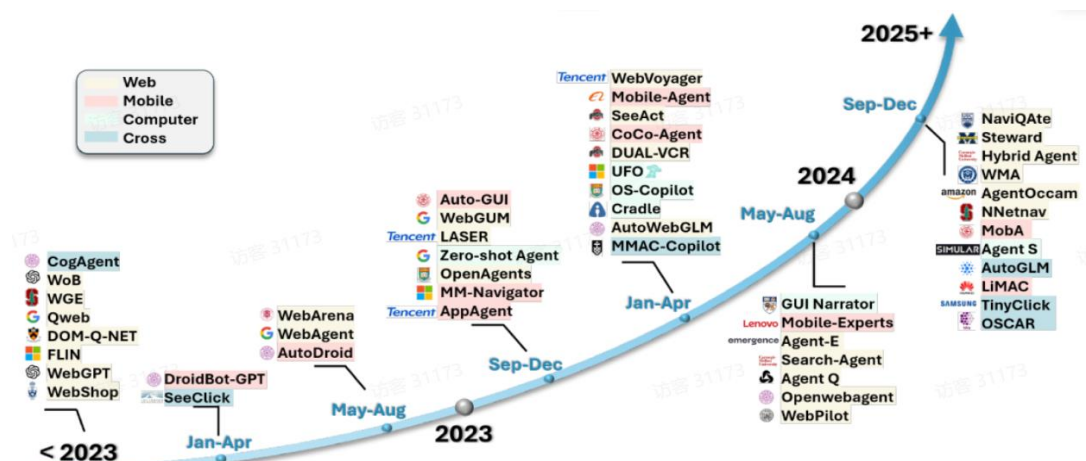
2.3 AI Agent：强推理模型性能持续提升，Agent 应用落地有望全面提速

智能体（Agent）指一种能够独立执行任务、做出决策并与其他系统或用户交互的软件程序或算法。智能体可以执行各种任务，从简单的自动化任务到复杂的决策支持系统。Agent 能够接受用户的自然语言请求，具有主动性，能够自动拆解任务并在多个应用程序中无缝协作。例如，它可以从 Word 中提取信息、总结网页、处理 PDF 文件，并在 PowerPoint 中创建幻灯片，最终通过 Teams 发送。

在图形用户界面（GUI）的背景下，Agent 可以通过屏幕截图和小部件树（widget trees）来感知 GUI 状态，并执行操作以模仿用户行为，如鼠标点击、键盘输入或手机上的触摸手势。

在 2023 年 LLM 模型出现之前，GUI Agent 的工作范围和能力均受到较大限制。自那时起，基于 LLM 的方法在网络、移动和桌面环境等不同平台上蓬勃发展，推动了该领域众多创新成果的诞生。

图表17：LLM 驱动的 GUI Agent 的发展历程





图表17: LLM 驱动的 GUI Agent 的发展历程

来源：小蜜记机器学习微信公众号，国金证券研究所

Deepseek R1 显著降低了大语言模型的技术门槛，AI Agents 的应用前景变得更加广阔。

图表18: AI agents 应用场景

应用领域	应用场景	主要内容	代表公司
企业服务	智能客服	AI Agents 能够实时响应客户需求，提供更加个性化的服务。例如，电商平台可通过 Agents 自动处理客户咨询、订单跟踪和退换货申请，减少人工客服的工作量。	致远互联： 公司推出五大类 AI 产品，包括知识服务类、内容创作类、数据分析类、流程自动化类和对话服务类，以多种方式为用户数智赋能，以及在公文管理、会议管理、知识管理、合同管理等应用和解决方案中融合智能化能力等。在原生 AI 应用方面，公司推出 AI 智能表单产品—iForm。
	办公自动化	企业可以部署 AI Agents 协助完成数据分析、文档处理以及任务分配。这种应用不仅提高了工作效率，还能帮助员工更专注于创造性工作。	
教育	个性化辅导	AI Agents 能够根据学生的学习进度、兴趣和能力定制教学内容，为每个学生打造专属的学习路径。	竞业达： 2024 年，基于星空大模型，公司已实现多级知识图谱 AI 智能生成，将专业、课程、知识智能关联，能够自动生成 AIGC 课堂质量诊断报告。 科大讯飞： 讯飞 AI 学习机发布 AI 作业过滤器，为学生做习题时提供“优先级”选项，以实现“练得更少、学得更好”目标。
	学术研究助手	研究人员可以通过 Agents 快速整理文献、生成数据分析报告，甚至探索跨学科的研究方向，大幅提升科研效率。	
医疗健康	辅助诊断	通过分析患者病史和实时数据，AI Agents 能够协助医生进行初步诊断，甚至提出优化治疗方案。	卫宁健康： 公司自研医疗大模型 WiNGPT，目前 WiNGPT 在语音生成查房记录、Agent 增强式问答、CDSS 创新方案、患者血液管理等众多实践领域已有应用案例。
智能制造	生产流程优化	通过实时监控生产线数据，AI Agents 可以快速发现问题并给出调整建议，从而提高生产效率。	中控技术： 基于 TPT 流程工业时序大模型的工业 App。 赛意信息： “基于善谋 GPT 的 PCB 行业大模型”已经能够准确识别和提取 PCB 工程图纸中的关键的工程参数，如尺寸、材料、层数等，将原本需要 4~6 小时的参数提取工作缩短至几分钟，同时确保了 95% 以上的准确率，大大提升了从订单接收到生产准备的效率，显著降低了因人工错误导致的质量问题。 鼎捷数智： 融合汽车企业行业知识经验与 GPT 大模型，运用 AI 和 NLP 技术，构建车企 ChatFile 企业知识管理赋能平台，处理海量车辆零散知识、自动化语种转译、协助团队跨文档交流、降低学习成本和培训周期；
	供应链管理	AI Agents 能够动态分析市场需求，优化库存管理，并自动与供应商进行交互，大幅降低企业的运营成本。	



			实现企业知识智能化交互，提升知识获取的体验和效率，减少时间和沟通成本。
创意领域	自动生成内容	AI Agents 可以生成高质量的文本、图像、视频，协助设计师、作家等创作者快速完成作品原型。	万兴科技： 公司旗下视频创意软件万兴播爆，支持数字人口播短视频制作；文档创意软件亿图脑图，支持 AI 生成思维导图和大纲。
	交互式体验	游戏开发中，AI Agents 可作为智能 NPC（非玩家角色），与玩家进行更具沉浸感的交互，提升娱乐体验。	

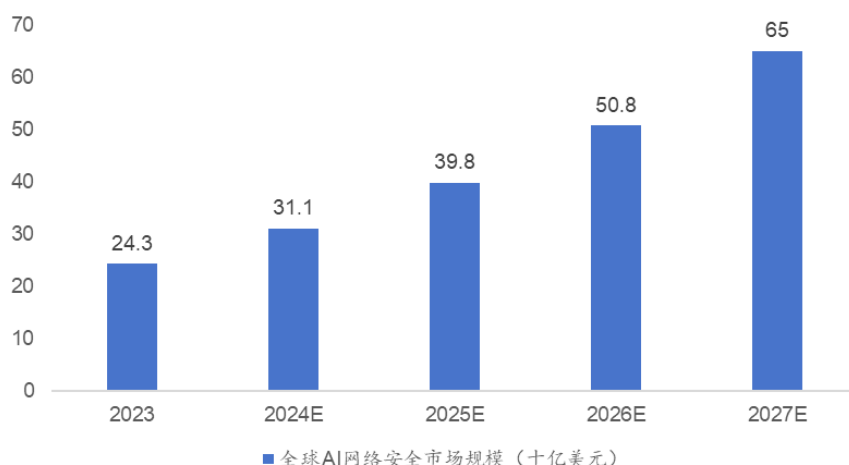
来源：智用人工智能应用研究院微信公众号，国金证券研究所

2.4 AI 安全：大模型厂商急需全方位加强防护，网安及信创板块迎来布局机会

在人工智能时代，AI 大模型得到了广泛应用，但随之而来的安全问题也日益凸显，安全需求变得极为迫切，这背后蕴藏着巨大的市场空间与机会。以 DeepSeek 为代表的国产模型，凭借技术优势在竞争激烈的 AI 领域迅速崛起，然而也因此遭受了网络恶意攻击，这警示着大模型厂商提升网络安全能力已刻不容缓。面对不断升级的网络攻击手段，以及 AI 本身可能被用于发动更复杂攻击的风险，国产模型急需全方位加强网络安全防护，构建完善的安全体系。

这种严峻的形势为网安和信创产业链带来了广阔的市场需求与技术挑战。AI 企业需要加大网络安全和自主可控相关投入、引进专业人才、优化技术，网络安全公司恰好能凭借专业能力与丰富经验，为 AI 企业提供定制化的安全解决方案，助力其提高系统的安全性与稳定性。政府与行业协会制定相关法律法规和行业标准，也需要网络安全公司提供专业支持与技术参考。根据 Techopedia 数据，2023 年全球基于 AI 的网络安全市场规模为 243 亿美元，预计 2027 年将达 650 亿美元，2023-2027 年 CAGR 约 27.9%。

图表19：全球基于 AI 的网络安全市场规模快速增长



来源：Techopedia，智领安平行业网公众号，国金证券研究所

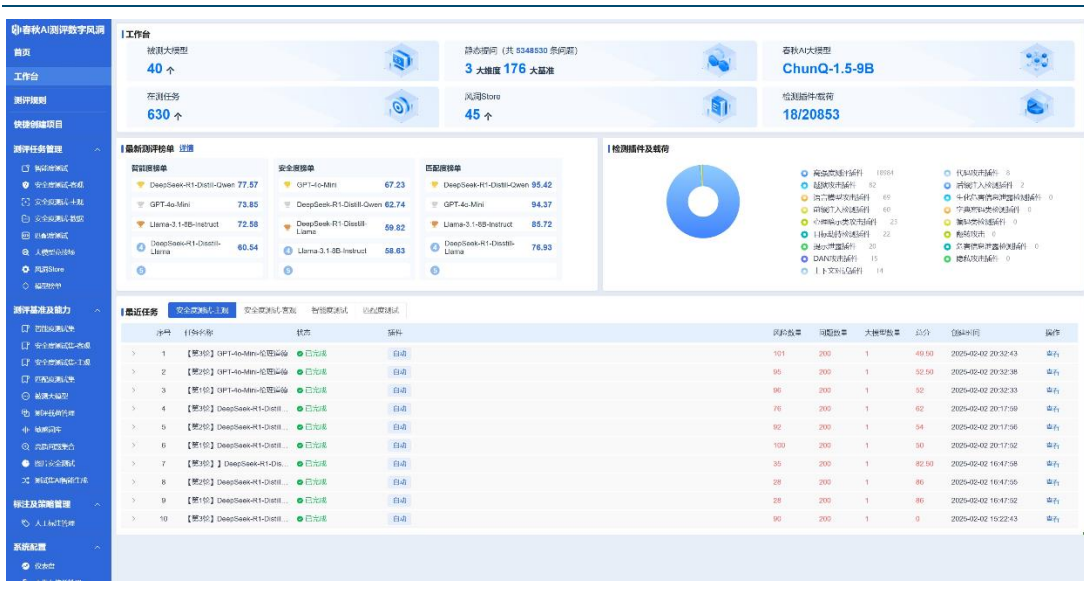
未来，随着 AI 技术持续发展，一方面，网络安全公司在保障 AI 大模型安全方面将扮演愈发关键的角色，带来业绩增量；另一方面，网络安全公司也亟需降本增效的路径，而以 DeepSeek 为代表的优秀国产化大模型将无疑是网安公司的优先合作伙伴，发展国产化安全垂类模型工具能力，从而实现安全检测与防护能力提升的同时实现降本增效。

国内多家网络安全公司积极拥抱 AI 技术进步带来的业务需求，接入 DeepSeek 优化自身安全产品。其中，奇安信自主研发的 QAX 安全大模型已全面完成了 DeepSeek 的深度接入，率先将其引入到威胁研判、安全运营、渗透测试和漏洞管理、身份与访问管理、网络钓鱼防护、恶意软件和勒索软件防护、数据泄露防护、安全培训、供应链安全等场景之中，并展现出了卓越表现，其中安全专业问答整体性能分数提升约 16%，极大提升了智能威胁分析和决策的准确度。另外，永信至诚依托在网络靶场和数字安全测评领域的深厚技术积累与业务实践成果，构建春秋 AI 测评「数字风洞」平台，基于标准化测评数据和海量



业务场景模版，实现对 AI 智能产品智能度、安全度和匹配度的综合测评，通过以模测模、以模强模，简化测评流程，提高测评效率。平台内置了覆盖 18 个知识领域和 300 万余测评题目的智能评估体系，能够测评从基本知识应用到复杂推理能力的具体表现。帮助企业了解产品的实际认知能力，确保在业务场景中被准确应用。

图表20：永信至诚春秋 AI 测评“数字风洞”平台



来源：永信至诚公众号，国金证券研究所

另外，以 deepseek 为主的 AI 大模型的自主可控迫在眉睫，除了软件核心组件自主可控以外，硬件产业链的国产替代需求旺盛。2025 年是十四五收官之年，预期信创推进政策和配套财政支持政策快速落地，预计上半年信创产业链相关公司的订单回暖将会出现重要信号。

3. 投资建议

DeepSeek 的低成本、高性能特点有望带动 AI 应用和端侧硬件的加速落地，建议关注相关领域投资机会：

- 1) 端侧智能硬件：契合中国优势+需求，落地场景分为娱乐、工作两大类，产品包括耳机/眼镜/手机/电脑/智能家居等，看好萤石网络、虹软科技、科大讯飞，关注汉王科技。
- 2) 软件应用：C 端重点关注高用户基数/出海两类，看好金山办公、万兴科技、福昕软件、彩讯股份。B 端重点关注增收能力/制造业配套两类，看好焦点科技、新致软件、鼎捷数智、汉得信息等。G 端重点关注契合中国国情的落地能力，看好科大讯飞、第四范式、中科星图、航天宏图。
- 3) 网络安全：AI 对网安行业带来新机遇和新挑战，叠加行业见底反转，关注奇安信，永信至诚，启明星辰，安恒信息，亚信安全，深信服，天融信。
- 4) 信创：AI 产业链自主可控需求迫切，十四五投入进度明确，关注中科曙光，神州数码，软通动力，拓维信息，中国软件，达梦数据。


图表21：建议关注 DeepSeek 相关产业链投资机会

类型	企业	内容
应用	金山办公	金山办公正在与 Deepseek 进行对接测试，未来是否合作要看后续发展。
应用	远光软件	依托 DeepSeek “高效推理能力与低延迟响应” 特性，极大地提升了远光 DAP 等产品在私域知识分析、自然语言交互和场景理解等的流畅度和精准度，进一步增强了核心产品智能化能力。集成 DeepSeek 后，所有远光软件的客户可升级体验全新 AI 服务。
应用	泛微网络	在接入 DeepSeek 大模型后，泛微将借助 DeepSeek 强大的自然语言处理、机器学习、推理等能力，显著提升泛微各项产品智能化效果以及智能体构建能力：泛微产品的各个应用模块能够快速调用 DeepSeek 的智能化能力；用户可在泛微智能小 e 中直接使用 DeepSeek；泛微提供统一的 AI 大模型连接底座支撑；支持私有化部署 DeepSeek 大模型。
应用	法本信息	实现 FarAI GPTCoder 企业级辅助编码平台、FarAI GPTBrain 企业级知识大脑产品、FarAI IGPTRecruit 智能招聘平台等多款产品与 DeepSeek 的深度融合及优化。
应用	汉得信息	在自主 PaaS 平台 H-ZERO 之上，基于 DeepSeek-Coder 模型和 H-ZERO 平台的源码训练代码大模型，为企业加持代码智能辅助能力。
应用	金蝶国际	将 DeepSeek 全面集成至金蝶云全线 SaaS 应用及金蝶云苍穹平台，为客户提供更高效、更安全、更低成本的智能解决方案。
应用	三维天地	SunwayDeepLink 智能体，集 ChatBI 和 ChatDOC 两大功能于一身，与 DeepSeek（深度求索）实现全面深度适配，旨在为企业构建一个“数据—洞察—决策”一体化的管理平台，为企业数据资产管理进一步赋能增效。
应用	彩讯股份	彩讯股份对 Rich M@il 第一时间开展测试，成功完成了与 DeepSeek 的集成，并将借助 DeepSeek 的强大特性，对现有的智能场景展开全方位升级：更精准的邮件分类与归纳；更智能的指令响应；反垃圾邮件机制强化；反钓鱼防护机制强化。
应用	微盟集团	自研大模型应用型产品微盟 WAI 已集成 DeepSeek，开始赋能 WAI Agent 能力升级，通过具备更强的推理、调度能力，从而帮助 WAI SaaS、WAI Pro 以及 WIME 等覆盖不同客户场景的产品提升能力。
应用	云天励飞	云天励飞芯片团队完成 DeepEdge10 “算力积木” 芯片平台与 DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Llama-8B 大模型的适配，可以交付客户使用。
应用	中文在线	公司已在部分内部 AI 网文创作流程中部署 DeepSeek-R1，通过调用 DeepSeek-R1 的能力，增强创作效率。
应用	中软国际	JointPilot(灵析)人工智能应用平台成功接入 DeepSeek-R1 模型，为企业提供更加丰富、高效的大模型应用场景落地服务。
应用	国投智能	“星盾”、“Qiko”等自研平台第一时间成功接入 DeepSeek,进一步满足行业客户的多样化需求，标志着国投智能在落实“All in AI”战略方面迈出了重要一步
应用	新致软件	将 DeepSeek-R1 和 DeepSeek-V3 顺利接入新致新知平台，提升了新知平台的推理和泛化能力，通过 DeepSeek 的深度学习模型，新知平台能够更准确地挖掘和利用数据资源，为企业数智化转型提供了更加精准的支持。
应用	万兴科技	AIGC 软件 A 股上市公司万兴科技（300624.SZ）率先完成深度求索（DeepSeek）最新推理大模型 DeepSeek-R1 的深入适配，涵盖旗下视频创意、绘图创意及文档创意软件业务多款产品。目前，万兴喵影、亿图图示、亿图脑图 MindMaster、万兴 PDF 等均已融合 DeepSeek-R1 大模型相关能力。
应用	梦网科技	梦网科技将 DeepSeek 大模型深度集成至多源 AI 调度引擎“天慧智汇台 2.0”，推动消息



		通信服务从基础连接向智能决策升级，为行业开辟降本增效新路径。
应用	用友网络	· 用友 BIP 全面上线以 DeepSeek-V3 和 DeepSeek-R1 作为基座大模型的智能服务，企业客户可以通过用友 BIP3 R6 智能平台的公有云、专属云模式使用相关服务。
应用	赛意信息	· 赛意善谋 GPT、知识库等产品也已全面接入 DeepSeek-R1 模型，为客户带来更加智能、高效的服务体验。
网安	中新赛克	· 借助 DeepSeek 的深度理解和分析技术，调用其他安全工具的数据进行交叉验证，有效减少无效告警，同时提供针对性的处置建议和策略，让安全运维人员能够迅速锁定问题，高效解决问题。
网安	启明星辰	· 完成了“安星”智能体与 DeepSeek 大模型的全面对接，并应用到安全运营、威胁检测、威胁情报、数据安全等全业务场景中，实现了安全智能化能力的再升级。
网安	奇安信	· 自研 QAX 安全大模型通过 DeepSeek R1 进行了一系列的优化和蒸馏后，不仅运营成本实现了大幅降低，同时在威胁研判等多个场景下的模型性能方面获得了显著提升，这势必将进一步扩大奇安信在人工智能与网络安全融合创新的领先优势。
云	开普云	· 在开悟大模型智能体平台接入 DeepSeek V3、R1 在线模型 API，在运营管理平台部署量化版 R1 模型服务，在端侧一体机中部署蒸馏量化版 R132B 模型。
云	神州数码	· 旗下神州鲲泰推理服务器产品搭载昇腾硬件，可全面支持 DeepSeek 系列模型的快速部署，为用户带来更快、更高效、更便捷的 AI 开发和应用体验；将 DeepSeek 集成到其自主研发的神州问学平台中，仅需 3 分钟部署 DeepSeek 模型。
云	协创数据	· 平台对 DeepSeek 系列模型的适配与应用已全面完成并正式上线，为用户带来了更强大、更高效的智能服务体验。
云	优刻得	· 基于壁仞科技国产芯片的先进内存架构、多模型适配能力、广泛的数据精度支持以及解码能力，全面开展包括 R1 在内的 DeepSeek 全系列模型适配工作，以满足不同规模参数量模型的个性化部署需求。
云	每日互动	· 温州数安港携手 DeepSeek 的可信合作伙伴——浙江省大数据联合计算中心、每日互动及温州市数据集团，依托数安港可信数据空间，部署 DeepSeek R1、v3 等一系列模型，为相关产业提供基于可信数据空间的可控大模型租用服务、私有化部署和精调服务，为各行业提供更为安全、高效的大模型应用解决方案。
云	青云科技	· 基石智算 CoresHub 正式上线 DeepSeek Janus-Pro-7B WebUI 版文生图模型，支持 Multimodal Understanding 和 Text-to-Image Generation 两种服务，可通过基石智算 AI 算力云服务的容器实例进行直接使用。由于模型文件存放在系统盘，也可通过 Jupyter 进行算法二次开发使用。

来源：iFind，各公司公众号，国家数据局公众号，36 氪，科创板日报公众号，国金证券研究所

4. 风险提示

■ 底层大模型迭代发展不及预期

若底层大模型迭代发展不及预期，可能会影响 AI 应用落地的深度，使其难以在金融、教育、游戏等领域进行更加深入的应用。若底层大模型的蒸馏剪枝发展不及预期，可能会使其难以在边缘硬件上充分发挥性能。

■ 国际关系风险

若出于国际关系原因，OpenAI 等海外大模型的调用或其他软硬件的进口受到影响，有可能使得国内 AI 应用的发展不及预期。

■ 应用落地不及预期

若相关应用公司不能找到人工智能算法较好的商业应用落地场景，或相关场景客户没有较强的付费意愿，可能算法应用落地会不及预期。



■ 行业竞争加剧风险

若相关企业加快技术迭代和应用布局，整体行业竞争程度加剧，将会对行业内已有企业的业绩增长产生威胁。



行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告版权归“国金证券股份有限公司”（以下简称“国金证券”）所有，未经事先书面授权，任何机构和个人均不得以任何方式对本报告的任何部分制作任何形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海

电话：021-80234211

邮箱：researchsh@gjzq.com.cn

邮编：201204

地址：上海浦东新区芳甸路 1088 号

紫竹国际大厦 5 楼

北京

电话：010-85950438

邮箱：researchbj@gjzq.com.cn

邮编：100005

地址：北京市东城区建国门内大街 26 号

新闻大厦 8 层南侧

深圳

电话：0755-86695353

邮箱：researchsz@gjzq.com.cn

邮编：518000

地址：深圳市福田区金田路 2028 号皇岗商务中心

18 楼 1806



【小程序】
国金证券研究服务



【公众号】
国金证券研究