DeepSeek：技术颠覆 or 创新共赢

计算机行业首席分析师：吴砚靖；通信行业首席分析师：赵良毕；电子行业首席分析师：高峰；传媒行业分析师：岳铮。

计算机行业分析师：邹文倩、李璐昕，研究助理胡天昊、冯雨淇；通信行业分析师：赵中兴；传媒行业研究助理：祁天睿。

DeepSeek：技术颠覆 or 创新共赢

2025 年 02 月 03 日

核心观点

- **DeepSeek 通过算法及工程创新，显著降低成本，技术变革算力新方向：**DeepSeek 模型通过使用 PTX 编程语言，以及工程能力上的创新，使得其在具有更强的性能的同时，实现更低的训练与推理成本，或将加速推动 AI 应用与硬件的普及和落地。与市场认为的不同，我们认为更低的训练与推理成本对算力需求呈现短期减少，长期高增的趋势，AI 能力边际扩张依然需要依赖更大的模型和强大的算力，DeepSeek 在算法和架构上的创新给 AI 发展增加了一条新的道路，有望开拓 AI 行业的共赢局面。

- **结合我们对芯片、硬件、软件、应用端等的影响分析，我们认为 DeepSeek 的技术颠覆带来的是 AI 行业的多元化，有望加速 AI 行业的普及繁荣，具体细分到行业子板块来看：**

  电子板块-后训练时代看好推理侧算力部署，以及 AI 端侧加速落地：DeepSeek 的创新并没有完全打破 scaling laws，且正从 pre-training 转向 post-training 和推理，通过增加模型规模、扩展训练数据、提高计算资源以及合理的任务设计，可加速模型学习更复杂的推理能力。随着模型规模、数据量和计算资源的增加，模型能够更好地进行推理，通过平衡性能、内存占用和推理速度来提高大语言模型的运行效率，有利于 AI 硬件端的落地与普及。

  通信板块-推理侧算力有望增加利好国产光芯片，看好 AI 时代运营商角色转换，光模块景气度无虞。我们认为运营商作为我国最大的流量管道，具备数据优势及接口优势，AI 应用的普及将持续推进，同时，更强训练模型的未来需求将带动光模块产业链快速发展，在全球经济形势复杂化趋势下，核心器件光芯片等方向自主可控进程进一步加速。

  计算机板块-看好算力向推理，基础设施向应用侧投资变化机遇：当下投资中的结构性机会主要体现在“从训练算力为主到推理算力为主过渡”、“从高端 GPU 到 ASIC 芯片过渡”，以及“从基础设施投资机会向应用侧投资机会过渡”。其开源策略和低成本模型使得更多企业和开发者能够使用先进的 AI 技术，加速了 AI 技术在各行业的应用和发展。

  传媒板块-大模型推陈出新进程加速，AI+赋能进行时：在 C 端，用户渗透率不断提升，主要 AI APP 活跃数据持续环比增长；在 B 端，AI 营销等领域的商业化模式已经逐步得到验证。DeepSeek 有望加速推动在影视、广告、社交陪伴等多个领域 AI+应用落地。

- **投资建议：**建议关注电子板块消费电子相关产业链、AI 终端硬件等方向；通信板块运营商、光模块、光芯片等方向；计算机板块看好边缘算力、AI 应用开发、数据服务与处理、端侧 AI 设备等方向；传媒板块“AI+”等细分子领域方向。

- **风险提示：**国际经济形势复杂度进一步提升的风险；AI 硬件发展速度不及预期的风险；AI 产业链上下游短期波动的风险；AI 应用发展不及预期的风险等。

分析师

赵良毕 首席通信分析师
☎：010-8092-7619
✉：zhaoliangbi_yj@chinastock.com.cn
分析师登记编码：S0130522030003

吴砚靖 首席计算机分析师
☎：010-66568589
✉：wuyanjing@chinastock.com.cn
分析师登记编码：S0130519070001

高峰 首席电子分析师
☎：010-80927671
✉：gaofeng_yj @chinastock.com.cn
分析师登记编码：S0130522040001

岳铮 传媒分析师
☎：010-8092-7630
✉：yuezheng_yj @chinastock.com.cn
分析师登记编码：S0130522030006

计算机行业分析师：
邹文倩，登记编码：S0130519060003、
李璐昕，登记编码：S0130521040001
研究助理胡天昊、冯雨淇；

通信行业分析师：
赵中兴，登记编码：S0130524090002；

传媒行业研究助理：祁天睿

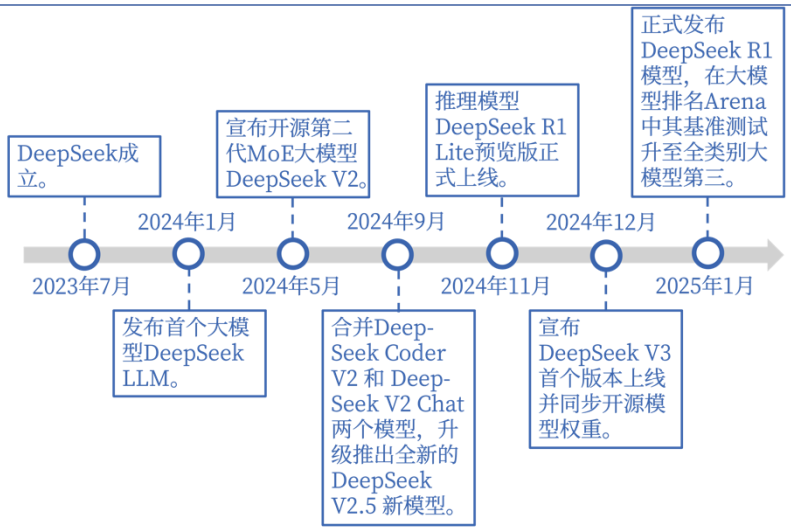
目录
Catalog

一、DeepSeek：算法革命带动 AI 景气进一步上行

（一）DeepSeek 公司成立背景与发展历程

DeepSeek，全称杭州深度求索人工智能基础技术研究有限公司，由幻方量化的联合创始人梁文峰创立。公司自 2023 年 7 月年成立以来，始终专注于大语言模型（LLM）及其相关技术的深度研发。公司坚持技术创新路线，开创性地提出多头潜在注意力机制（MLA）和 DeepSeekMoE 等创新架构。凭借这些创新成果，DeepSeek 的大模型在多项权威测评中展现出顶尖的性能表现。

图1：DeepSeek 发展历程



资料来源：DeepSeekAPI 文档，中国银河证券研究院

DeepSeek 的团队成员大多来自清华大学、北京大学、中山大学、北京邮电大学等国内顶尖高校，整体呈现出“年轻高学历、注重开源、重视创新”的特点。

根据彭博社报道，DeepSeek 的 AI 助手在 140 个市场中成为下载量最多的移动应用。根据 Appfigures 的数据，DeepSeek 的推理人工智能聊天机器人在 1 月 26 日登上苹果公司 AppStore 的榜首并保持全球第一，1 月 28 日起在美国的 AndroidPlayStore 中也位居榜首。根据 SensorTower 的数据，DeepSeek 在发布后的前 18 天内获得了 1600 万次下载，约为 OpenAI 的 ChatGPT 发布时 900 万下载量的两倍，印度贡献了所有平台下载量的 15.6%。

在用户体验方面，DeepSeek 表现不俗。用户普遍认为 DeepSeekR1 的性能出色，特别是在数学推理、编程能力和自然语言理解等领域。其推理速度和准确度在多个测试场景中达到业界领先水平。此外，DeepSeekR1 的“聪明”特性使得用户无需复杂的提示词技巧，即可获得高质量的回答。在实际使用场景中，无论是游戏、视频播放还是日常工作的辅助，其流畅的操作体验都得到了用户的高度评价。用户反馈显示，DeepSeek 界面简洁直观、操作简单，在实时数据推送和内容推荐上十分出色，能够有效提升工作效率，减少用户在信息检索上的时间投入。

DeepSeekR1 的发布引起了硅谷科技领袖、国际媒体及学术界的广泛关注。其性能和开源策略获得了高度评价，被认为是“非美国公司践行 OpenAI 初心”的典范。DeepSeekR1 的发布引发了全球科技市场的连锁反应。其开源策略、低成本、高性能的特性，对科技巨头形成了压力。其训练成本仅为 600 万美元，远低于 OpenAI 和谷歌等公司的同类模型，《MITTechnologyReview》提到，R1 在数学、代码等复杂任务上的表现与 OpenAIo1 相当，而训练成本仅为其 1/70，定价低至 OpenAI 的 3%。这种成本效益优势使得更多企业和开发者能够以较低的成本使用先进的 AI 技术，将大大加速 AI 技术的普及和应用。

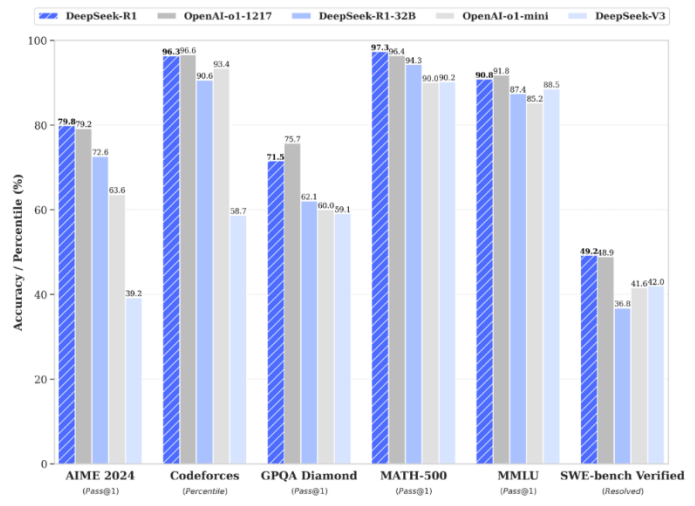
（二）DeepSeek：从硬件竞赛到算法效率革命的技术颠覆

大模型在 AI 行业中占据核心地位，是推动技术创新、拓展应用场景及提升行业效率的关键因素。全球范围内的领军企业持续推动大模型性能的提升，随着模型规模的不断扩张，其性能也实现了显著提升。然而，这种规模的扩大也相应地带来了训练和部署成本的急剧增加，成为制约大模型广泛应用的瓶颈。

在机器学习领域，尤其是大型语言模型（LLMs）的应用场景中，模型性能的提升与模型规模、数据集的大小以及计算资源之间存在着紧密的关联，这一关系通常被描述为“规模定律”（Scaling Law）。根据规模定律，模型的性能会随着模型规模的指数级增加而实现线性提升。目前，国际上主流的大模型，诸如 OpenAI 的 GPT 系列、Anthropic 的 Claude 以及谷歌的 Gemini 等，其最新版本的规模均已突破千亿参数大关。尽管这些模型在性能上展现出了卓越的表现，但对于众多公司和开发者而言，其高昂的硬件资源使用成本、计算时间等依然构成了巨大的挑战。长期以来，大算力训练一直是基座模型厂商用于融资与构建竞争壁垒的重要手段。

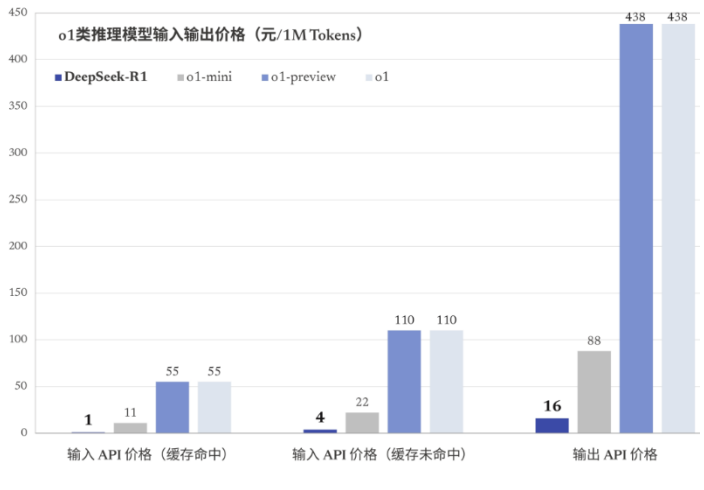
从技术层面来看，GPU 等硬件设施效率的提升以及算法的优化等方式，均有望带动大模型成本的显著下降。在全球 GPU 短缺以及美国限制政策的双重压力下，我国的人工智能公司 DeepSeek 通过算法优化的创新路径，进一步降低了训练成本，为大模型的大规模应用提供了前所未有的可能性。DeepSeek 在 1 月 20 日正式发布了其 R1 模型，并同步开源了模型权重。在第三方的基准测试中，DeepSeek-R1 的表现优于 OpenAI、Meta 和 Anthropic 等美国领先的人工智能公司。在 AIME2024 数学基准测试中，DeepSeek-R1 的成功率高达 79.8%，成功超越了 OpenAI 的 o1 推理模型。在标准化编码测试中，DeepSeek-R1 更是展现出了“专家级”的性能，在 Codeforces 上获得了 2029Elo 的评级，并超越了 96.3%的人类竞争对手。**同时，DeepSeek-R1 真正令人瞩目的地方并不仅仅在于其卓越的性能，而在于其极低的成本**。它打破了硅谷传统的“堆算力、拼资本”的发展路径，仅用 557.6 万美元和 2048 块英伟达 H800 GPU 便完成了性能对标 GPT-4o 的模型训练，成本仅为 OpenAI 同类模型的十分之一，推理成本更是低至每百万 Token 0.14 美元，而 OpenAI 的推理成本则为 7.5 美元每百万 Token。

图2：DeepSeek 性能对齐 OpenAI-o1 正式版



资料来源：DeepSeek 官网，中国银河证券研究院

图3：推理成本低至每百万 Token 0.14 美元



资料来源：DeepSeek 官网，中国银河证券研究院

图4：DeepSeek 蒸馏小模型超越 OpenAI o1-mini

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0

资料来源：DeepSeek 官网，中国银河证券研究院

与专有模型不同，DeepSeek-R1 的代码和训练方法均在 MIT 许可下完全开源，这意味着任何人都可以无限制地获取、使用和修改该模型。全球开发者对 DeepSeek-R1 的贡献代码使其推理效率每小时提升 0.3%，这一开放性的举措极大地激发了业界的创新活力。DeepSeek-R1 在芯片资源利用、算法复杂性和推理速度上实现了重大突破，为 AI 行业的发展树立了新的标杆。

DeepSeek-R1 的崛起和其所展现出的成本优势和开源策略，一度让华尔街对传统的“烧钱”信仰产生了怀疑。

DeepSeek 突破的核心在于算法层次和系统软件层次的创新等：

1）首先是算法层次的创新。他们采用了新的 MoE 架构，使用了共享专家和大量细粒度路由专家的架构。通过将通用知识压缩到共享专家中，可以减轻路由专家的参数冗余，提高参数效率；在保持参数总量不变的前提下，划分更多的细粒度路由专家，通过灵活地组合路由专家，有助于更准确和针对性的进行知识表达。同时，通过负载均衡的算法设计，有效地缓解了传统 MoE 模型因负载不均衡带来训练效率低下的问题。

2）其次在系统软件层次的创新。DeepSeek 采用了大量精细化的系统工程优化。例如，在并行策略方面，采用双向流水的并行机制，通过精细的排布，挖掘了计算和通信的重叠，有效的降低了流水并行带来的气泡影响；在计算方面，采用 FP8 等混合精度进行计算，降低计算复杂度；在通信方面，采用低精度通信策略以及 token 路由控制等机制有效降低通信开销。

DeepSeek-R1 的成功或许证明，未来的 AI 竞赛将不再单纯依赖于芯片的纳米级较量，而是算法效率、生态活力与政策弹性的多维度博弈，AI 行业的发展将呈现出更加多元化和复杂化的变化趋势，有望带来 AI 行业的繁荣。

图5：DeepSeek-R1 与其他代表性模型在各个维度性能上的对比

	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

资料来源：DeepSeek 官网，中国银河证券研究院

（三）DeepSeek 开辟了效率提升新赛道，创新优势明显

更强的性能，更低的训练与推理成本，将加速推动 AI 应用与硬件的普及和落地。虽然更低的训练与推理成本减少了当前的算力需求，但是并不意味着 AI 的未来发展对半导体整体需求的减少，相反由于其模型架构、基础设施数据等方面的优化，以及更低的成本，使得其更加容易布置在端侧，从而加速 AI 的普及。AI 能力边际的扩张依然需要依赖更大的模型和强大的算力，DeepSeek 在算法和架构上的创新给 AI 的发展增加了一条新的道路。

Scaling laws 指出，模型的性能伴随着三个关键因素的增加而提升，即：模型参数量、训练数据量、计算资源，且性能和资源之间存在对数线性关系，DeepSeek 的技术创新表现在很多方面。

图6：Scaling laws

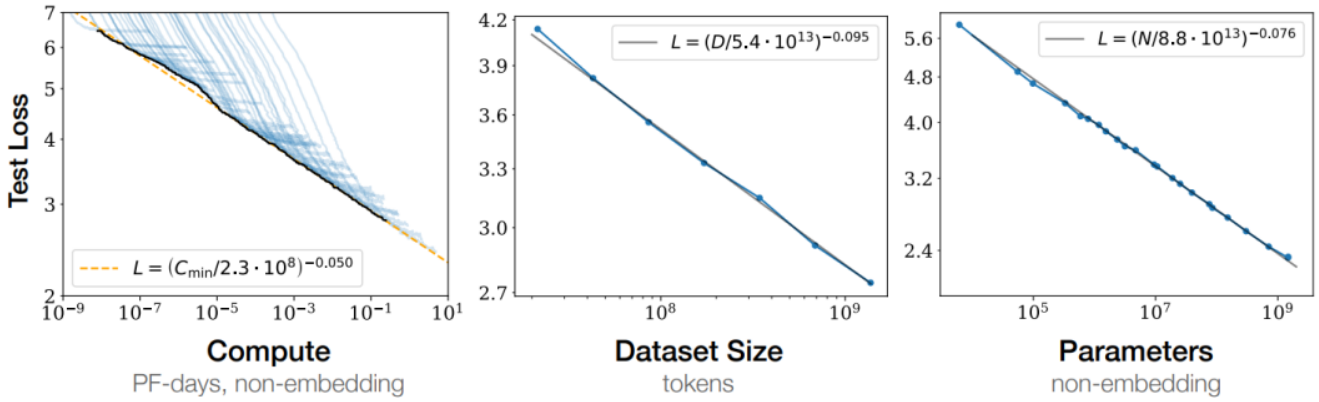


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

资料来源：Cornell University，中国银河证券研究院

模型架构的创新：以 DeepSeek V3 为例，采用了先进的 MoE 架构，具备 6710 亿总参数，但每次仅激活 370 亿参数，从而实现了高效的资源利用。与传统的全参数激活模型相比，MoE 动态激活机制显著降低了计算资源的需求，同时保持了高性能。DeepSeek 提出的多头潜注意力(MLA) 在不牺牲模型质量的前提下，大幅减少了 KV 缓存的大小。MLA 的核心思想是将键和值向量的计算分解成两个步骤，并在推理过程中只缓存中间的“潜向量”，而不是完整的键和值向量，大幅提升效率，降低推理成本。其他架构上的创新还包括：多令牌预测，提升训练效率，推测性解码提高推理速度。使用多 token 预测（MTP）训练目标，提升数据效率。

高效训练：DeepSeekV3 在一个配备 2048 块 NVIDIA H800 GPU 的集群上进行训练，使用 FP8 混合精度加速训练。设计了 DualPipe 算法以实现高效的管道并行性，开发了高效的跨节点全对全通信内核，在训练过程中仔细优化了内存占用。完整训练仅需 278.8 万 H800 GPU 小时，展现高效成本效益。训练成本仅为 557 万美元。

后续 DeepSeek 推出的 R1，在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版。通过 DeepSeek-R1 的输出，蒸馏了 6 个小模型开源给社区，其中 32B 和 70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果。

强化学习：在过去的研究中，大型语言模型往往需要先进行监督微调（SFT），再结合强化学习来提升推理性能。然而，DeepSeek-R1-Zero 直接用强化学习训练基座模型 DeepSeek-V3-Base，不依赖任何监督数据作为起点，证明了大型语言模型只要具备合适的奖励机制，就能纯粹依靠强化学习自主进化，学会复杂且深度的推理。而 DeepSeek-R1 通过冷启动数据和多阶段训练，使模型同时兼具高水平推理与高质量表达。

图7：DeepSeek V3 的架构概览

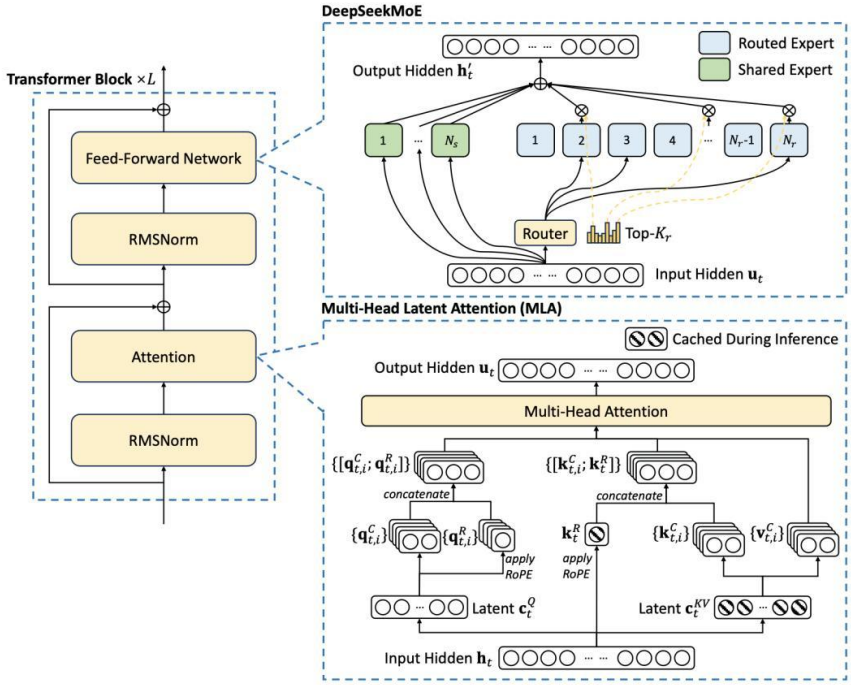


Figure 2 | Illustration of the basic architecture of DeepSeek-V3. Following DeepSeek-V2, we adopt MLA and DeepSeekMoE for efficient inference and economical training.

资料来源：DeepSeek-V3 技术报告，中国银河证券研究院

图8：MTP 的实现环节

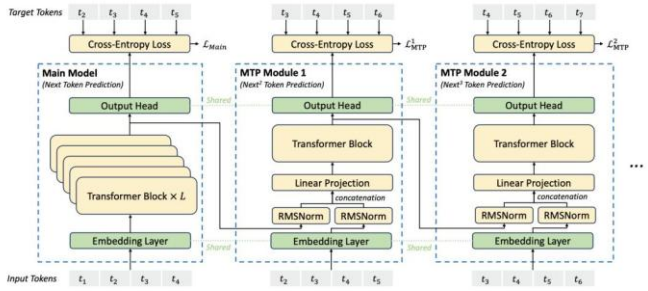


Figure 3 | Illustration of our Multi-Token Prediction (MTP) implementation. We keep the complete causal chain for the prediction of each token at each depth.

资料来源：DeepSeek-V3 技术报告，中国银河证券研究院

图9：FP8 训练混合精度框架

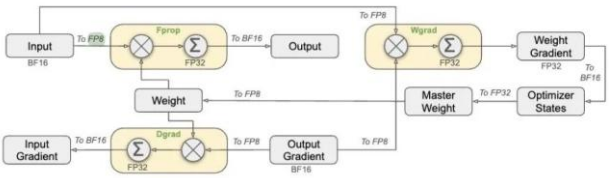
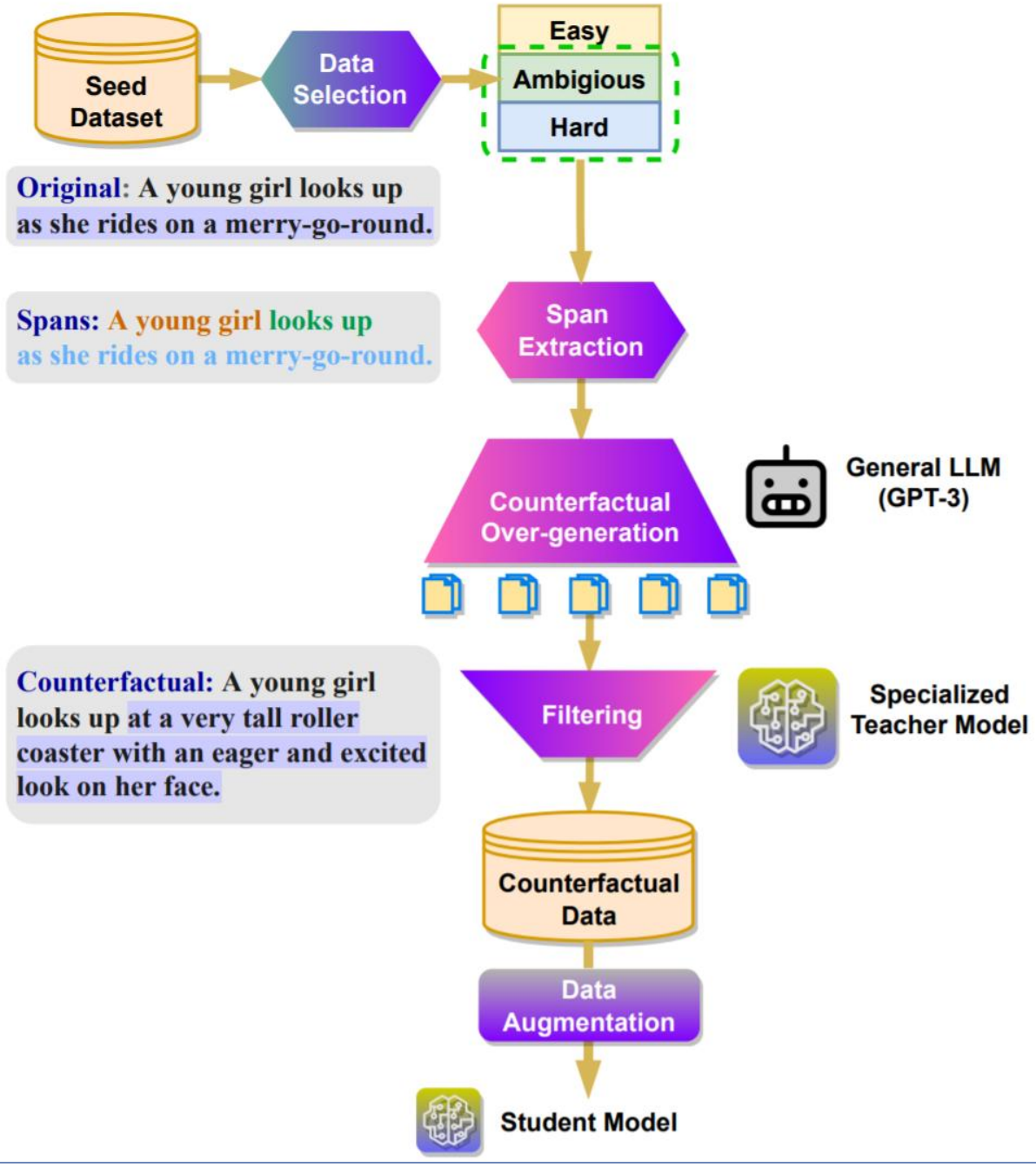


Figure 6 | The overall mixed precision framework with FP8 data format. For clarification, only the Linear operator is illustrated.

资料来源：DeepSeek-V3 技术报告，中国银河证券研究院

蒸馏模型：据 DeepSeek-V3 的技术文档，该模型使用数据蒸馏技术生成的高质量数据提升了训练效率。通过已有的高质量模型来合成少量高质量数据，作为新模型的训练数据，从而达到接近于在原始数据上训练的效果。DeepSeek 发布了从 15 亿到 700 亿参数的 R1 蒸馏版本。这些模型基于 Qwen 和 Llama 等架构，表明复杂的推理能力可以被封装在更小、更高效的模型中。蒸馏过程包括使用由完整 DeepSeek-R1 生成的合成推理数据对这些较小的模型进行微调，从而在降低计算成本的同时保持高性能。让规模更大的模型先学到高水平推理模式，再把这些成果移植给更小的模型。

图10：知识蒸馏小模型



资料来源：Cornell University，中国银河证券研究院

以上的创新主要是利用了更好的技术手段，解决很多实际“问题”，在理论应用和工程上打成平衡，体现了对 transformer 架构的深度理解，成功降低了对高端硬件的依赖，为 AI 的发展打开了一条新的道路。我们认为 DeepSeek 的创新并没有完全打破 scaling laws，对于计算量，模型参数量和数据集大小，当不受其他两个因素制约时，模型性能依然与每个因素都呈现幂律关系。DeepSeek 的创新为大模型的发展提供了新的“基准”，推动大模型发展进入新的阶段，AI 大模型的效率革命已经到来，而算力依然是推动人工智能进步核心因素之一。

（四）DeepSeek 引领 AI 成本革命，算法突破有望促进算力需求正向循环

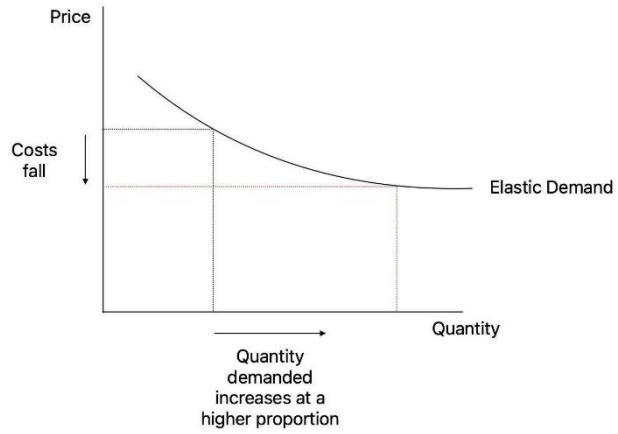
DeepSeek-R1 以超越美国顶尖模型的性能、更低的研发成本和较弱的芯片性能要求，引发了美国对其 AI 全球领先地位的担忧，同时也对科技公司在构建 AI 模型和数据中心上的巨额投入提出了质疑。在 DeepSeek 的冲击下，国内外大模型厂商紧急上线新模型，采取降价、免费等措施以证明自身的领先优势。同时，亚马逊、英伟达、微软等海外 AI 巨头纷纷上线部署支持用户访问 DeepSeek-R1 模型。2 月 1 日，OpenAI 发布全新推理模型 o3-mini 并首次向免费用户开放。这是 OpenAI 推理模型系列中最新、成本效益最高的模型。在定价方面，o3-mini 每百万 token 的输入（缓存未命中）/输出价格分别为 1.10 美元/4.40 美元，比完整版 o1 便宜 93%。不过，o3-mini 的性价比或依然不及 DeepSeek。作为对比，DeepSeek 的 API 提供的 R1 模型，每百万 token 的输入（缓存未命中）/输出价格仅分别为 0.55 美元/2.19 美元。在 o3-mini 推出后，OpenAI CEO 表示，中国竞争对手 DeepSeek 的崛起削弱了 OpenAI 的技术领先优势，并就开源与闭源的问题回应称，OpenAI 过去在开源方面站在“历史错误的一边”，公司曾经开源部分模型，但主要采用闭源的开发模式，未来将重新制定开源战略。与此同时，国产大模型的降价浪潮仍在持续。1 月 30 日，阿里云发布百炼 qwen-max 系列模型调整通知，qwen-max、qwen-max-2025-01-25、qwen-max-latest 三款模型输入输出价格调整，qwen-max batch 和 cache 同步降价，AI 大模型行业竞争加剧。

DeepSeek 的成本突破不仅是大模型训练的从“硬件驱动”向“算法驱动”的范式拓展，更为普惠化应用打开了新空间，反映 AI 技术向实用化、低成本化演进。行业对算力的依赖相较之前发生了“结构性”而非“总量性”变化：DeepSeek 的技术进步短期内或许能够局部缓解算力压力，但由于算法与算力的“螺旋上升”关系、应用场景的爆发式扩展以及数据增长的不可逆等特性，我们认为算力资源需求会从预训练端逐渐转移到推理端，DeepSeek 的兴起不会削弱高端芯片需求，而会促使大模型发展进入“算法进步→模型复杂化→硬件升级”的正向循环。

大模型成本优化与算力需求之间相互成就，高资源使用效率反而可能会增加算力的总消耗量。DeepSeek 通过降低训练成本，提高训练效率，看似减少算力需求，但同时，大模型成本缩减意味着降低了企业的训练与推理门槛，即每单位成本所能提供的训练和推理服务更多了，算力效率提升有望激活更广泛的用户与应用场景，从而引发对更大参数以及更复杂的大模型迭代需求。算法优化（如模型压缩、蒸馏）确实能提升单次任务效率，但 AI 能力的边界扩展（如多模态、复杂推理、通用人工智能）仍依赖更大规模模型和更复杂计算。这可能会对均衡下的算力需求产生正面影响，整体算力需求不会减少而是更加旺盛，从而形成对硬件需求的新一轮推升，即步入“算法进步→模型复杂化→硬件升级”的正循环。

微软首席执行官引用了“杰文斯悖论”来解释这一现象：Jevons 在《煤炭问题》中发现，随着蒸汽机效率的提升，煤炭消耗量不降反增。其核心观点为：技术进步提高了资源使用效率，效率提高降低了资源使用成本，成本下降刺激了资源需求的增长，需求增长可能超过效率提升带来的节约，最终导致资源总消耗增加。在算力日益成为数字经济“水电煤”的今天，DeepSeek 的技术方向与开源定位，恰恰是算力普及化革命的关键参与者。我们认为，大模型成本优化与算力需求并不是直接的此长彼消关系，而是互相搭台、相互成就，高资源使用效率反而可能增加算力的总消耗量。定价的持续走低有望带来更快的商业化落地，进而会衍生出更多的微调及推理等需求，将逐步盘活全球 AI 应用及算力发展。

图11：“杰文斯悖论”——高资源使用效率反而可能增加总消耗量



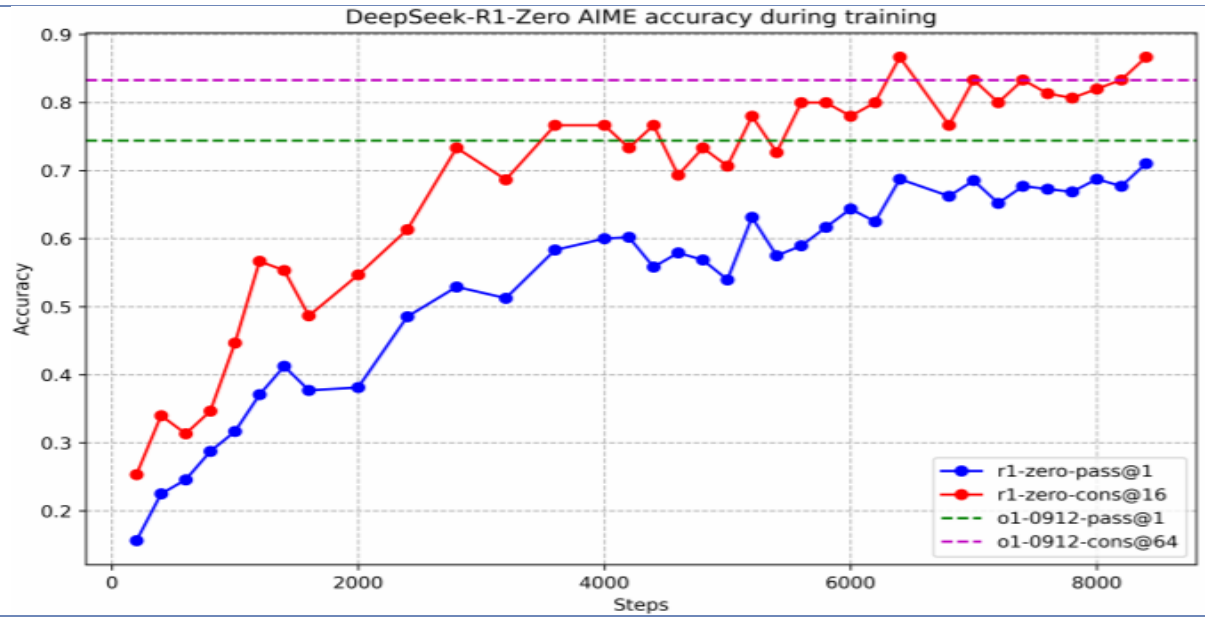
资料来源：维基百科，中国银河证券研究院

二、芯片端：推理+端侧发展星辰大海，高需求无虞

（一）推理算力需求持续增长正向影响芯片需求

在 OpenAI 提供的 O1 模型后训练阶段的缩放定律显示，随着强化学习时间和推理思考时间的增长，O1 模型性能得到显著提升。DeepSeek R1 系列模型推理过程包含大量反思和验证，思维链长度可达数万字。随着思考长度的增加，模型性能在稳步提升。Scaling Law 已经从预训练向推理层转向。

图12：DeepSeek-R1 展现出的推理 scaling law



资料来源：DeepSeek-R1 技术报告，中国银河证券研究院

通过增加模型规模、扩展训练数据、提高计算资源以及合理的任务设计，可以加速模型学习更复杂的推理能力，这一过程遵循 scaling law。随着模型规模、数据量和计算资源的增加，模型能够更好地进行推理。OpenAI 的 O1 模型，以及其背后所强调的后训练(Post-training)和推理阶段(Inference-time)的计算投入，正在重新定义我们理解 AI 模型性能增长的方式。通过模仿人类思考过程，进行多步骤、多路径的推理，最终选择最优的答案。这种“隐式思维链”(Implicit Chain of Thought) 的方法，需要在推理阶段投入更多的计算资源进行探索和评估。

谷歌研究发现，当合理分配推理计算资源时，检索增强生成（RAG）的性能能够呈现近乎线性的增长，RAG 在长上下文的大语言模型上的性能最高可提升 58.9%。这意味着，模型的表现提升和投入的计算量几乎是成正比的，这种现象被称为推理扩展定律。也进一步印证了，在提升推理能力上，算力依然是最为重要的需求之一，无论其模型是开源或者闭源。

未来的 AI 系统计算开销将更多地集中在推理服务上，而非单纯的预训练计算。这意味着算力基础设施的建设和优化需要重新思考，以满足日益增长的推理需求。在 AI 技术的实际应用落地过程中，用户感受最直观、最强烈的往往是推理环节的性能表现。虽然过去我们一直在强调大模型训练的重要性，但真正到了企业应用层面，推理的需求规模往往是训练需求的 5-10 倍。以字节为例，根据 QuestMobile 的数据，目前抖音集团旗下豆包 app 的月活跃用户达到了 7522 万，AI 原生 APP 在 24 年 12 月的月活跃用户规模达到了 1.2 亿，同比增速达到 232%。

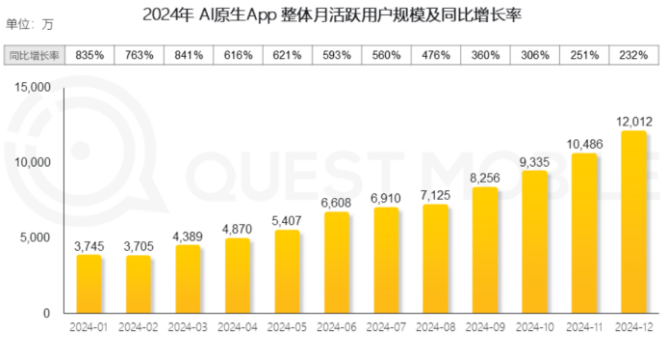
图13：互联网企业 AI 原生 APP 产品矩阵

典型互联网企业 AI原生App 产品矩阵

	原生App	月活跃用户规模（万）	AI场景分类		
抖音集团	豆包	7,522.6	综合场景		
	猫箱	537.3	垂直场景		
	即梦AI	196.0	垂直场景	图像生成	视频生成
	星绘	96.5	垂直场景	图像生成	
	小悟空	4.3	垂直场景	办公	
百度集团	文小言	1,224.1	综合场景		
	橙篇	13.7	综合场景		
阿里巴巴	通义	290.9	综合场景		
	妙鸭相机	28.2	垂直场景	图像生成	
360	360AI搜索	12.9	垂直场景	搜索	
	360智脑	4.6	综合场景		

资料来源：QuestMobile，中国银河证券研究院

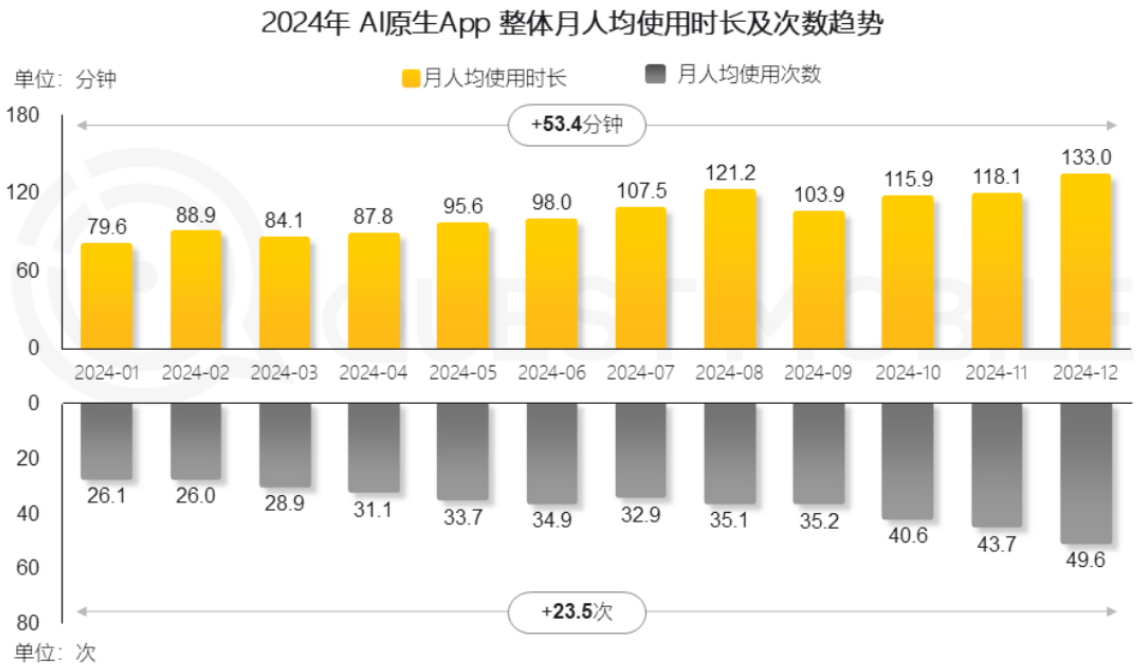
图14：2024 年原生 APP 整体月活跃用户规模及同比增速



资料来源：QuestMobile，中国银河证券研究院

QuestMobile 数据显示，2024 年 12 月，AI 原生 App 整体月人均使用时长达 133.0 分钟，较 1 月增加 53.4 分钟；月人均使用次数从 1 月的 26.1 次增加至 49.6 次。考虑到未来的潜在推理需求，预计将推动国内推理算力需求的持续增长。目前 2024 年豆包大模型的日均 token 调用量在 40000 亿左右，预估 2025 年日均 token 调用量将提升到最高 40 万亿，将大幅提升对推理算力的需求。国产算力寒武纪、海光信息等厂商有望受益。

图15：2024 年 AI 原生 APP 整体月均使用时长和次数趋势



资料来源：QuestMobile，中国银河证券研究院

（二）后训练增长及国产化需求提升有望带动光芯片需求增长

光芯片是光模块核心器件，应用场景较广。光芯片分为激光器芯片及探测器芯片，当前人工智能相关光模块内光芯片主要以 VCSEL 及 EML 芯片为主，高速率光芯片主要以 VCSEL 芯片为主，该款芯片具备线宽窄，功耗低，调制速率高，耦合效率高，传输距离短等特点，主要应用于 500 米内短距离传输，集中在数据中心机柜内布传输，消费电子等领域；EML 芯片则由于其调制频率高，稳定性好，传输距离长等特性，广泛应用于长距离传输，高速率远距离的电信骨干网，城域网和 DCI 等领域，但其具备成本较高的特性，虽然可以部署于短距离传输市场中，但不具备性价比优势。

表1：光芯片按功能分类

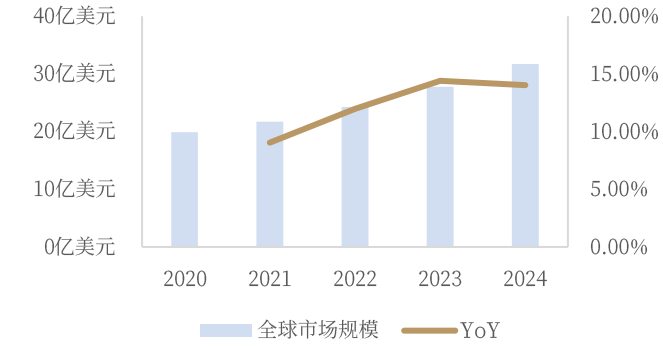
分类	产品类别	工作波长	产品特性	应用场景
激光器芯片	VCSEL	800-900nm	线宽窄，功耗低，调制速率高，耦合效率高，传输距离短，线性度差	500 米内短距离传输，数据中心机柜内布传输，消费电子等领域
	FP	1310-1550nm	调制速率高，成本低，耦合效率低，先行督察	中速度无线接入短距离市场，部分应用场景逐步被 DFB 激光器芯片取代
	DFB	1270-1610nm	谱线窄，调制速率高，波长稳定，耦合效率低	中长距离传输，如 FTTx 接入网、传输网、无线基站、数据中心内部互联等
	EML	1270-1610nm	调制频率高，稳定性好，传输距离长，成本高	长距离传输，高速率远距离的电信骨干网，城域网和 DCI
探测器芯片	PIN	830-860/1100-1600nm	噪声小，工作电压低，成本低，灵敏度低	中长距离传输
	APO	1270-1610nm	灵敏度高，成本高	长距离单模光纤

资料来源：中商情报网，中国银河证券研究院

光芯片市场规模持续提升，推理侧算力部署加速及国产化进程加速有望直接带动行业增长。随

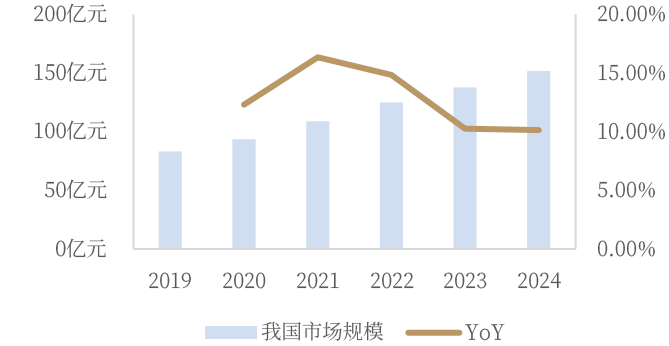
着云计算、大数据、人工智能等技术的快速发展，对高速、高效、低能耗的数据传输需求日益增长，光芯片的市场需求也随之增加，推动全球光芯片市场规模持续扩大。根据中商产业研究院发布的《2024-2029 年全球及中国光芯片行业发展趋势与投资格局研究报告》显示，2023 年全球光芯片市场规模约 27.8 亿美元，较上年增长 14.4%。中商产业研究院预测，2024 年全球光芯片市场规模将达到 31.7 亿美元。随着国产替代的加速推进，中国光芯片市场规模持续增长，并展现出强劲的发展势头。中商产业研究院发布的《2024-2029 年全球及中国光芯片行业发展趋势与投资格局研究报告》显示，2023 年中国光芯片市场规模约为 137.62 亿元，较上年增长 10.24%。中商产业研究院预测，2024 年中国光芯片市场规模将增长至 151.56 亿元。从国产化率来看，国内相关企业仅在 2.5G 和 10G 光芯片领域实现核心技术的掌握，2.5G 及以下速率光芯片国产化率超过 90%；10G 光芯片国产化率约 60%；25Gbs 及以上的光芯片国产化率低，仅有 4%。预计随着推理侧算力部署的逐步增多，以及后训练算力的规模提升，相对较低速率光芯片市场空间将进一步提升，在我国光芯片相关企业有望直接受益。

图16：全球光芯片市场规模



资料来源：中商情报网，中国银河证券研究院

图17：我国光芯片市场规模



资料来源：中商情报网，中国银河证券研究院

三、硬件端：光通信仍然靓丽，智能硬件边际改善

（一）运营商、光模块等细分板块仍旧具备较大投资价值

我们认为 DeepSeek 对通信行业的推动作用主要体现在两方面：

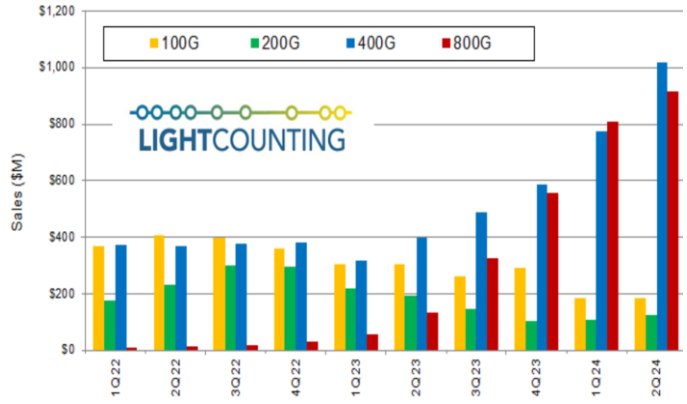
1）强化国产算力产业链：为中国 AI 发展带来新机遇，为中美科技竞争增添变数。DeepSeek-R1/V3 支持华为昇腾平台及 MindIE 引擎，自研推理加速引擎使硅基流动与华为云昇腾服务上的模型效果媲美高端 GPU，同时降低成本。这一突破为 AI 生态提供自主多元化方案，助力我国本土芯片厂商商业化落地，促进高效能 AI 的普及。

2）提升中国 AI 国际影响力：DeepSeek-R1 的开源实践标志着开源模式对闭源模式的一次重要胜利，这一开源模式对社区的贡献能够迅速转化为整个开源社区的繁荣。中小企业和个人开发者等长尾市场原本因成本限制无法参与的领域（如小规模模型微调、实验性研究）将被激活，形成分散的算力需求增量，产业或将迎来结构性变化，有望为 AI 技术的多元化创新增添更多可能性。同时，AI 加速走向千行百业，智能驾驶、机器人、元宇宙等新兴领域对实时计算和低延迟的高要求，将持续助推高算力需求。DeepSeek 迅速吸引了全球开发者瞩目，曾短时间内即在苹果中国及美国应用商店免费应用下载榜超越 ChatGPT 登顶，彰显了中国 AI 技术以更开放姿态融入全球。未来创新将聚焦于效率、开放性和生产力转化，DeepSeek 或成全球 AI 科技发展转折点。

虽然 DeepSeek 的推出，对降低推理侧成本带来巨大降低，但我们认为推理侧的成本降低，将显著带来训练测迭代的加速，由于推理侧成本的降低，应用场景落地或将进一步加速，推动推理侧模型效率的进一步提升，从而带动通信行业相关方向的持续性繁荣。故而我们维持此前对通信细分子领域运营商、光芯片、光模块的推荐方向，认为 DeepSeek 的推出，运营商作为我国最大的流量管道，具备数据优势及接口优势，AI 应用的普及将持续推进，同时，更强训练模型的未来需求将带动光模块产业链快速发展，在全球经济形势复杂化趋势下，核心器件光芯片等方向自主可控进程进一步加速。

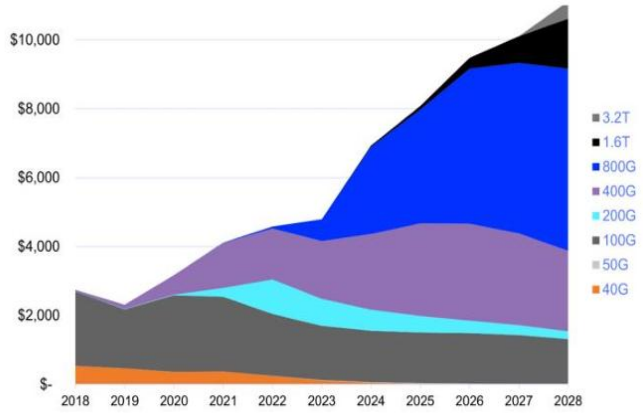
未来 5 年数通市场的增长驱动力主要来自 400G 以上高速率光模块的需求。全球云计算服务提供商对计算能力和带宽需求的持续增长，以及他们在服务器、交换机和光模块等硬件设备上的资本支出的增加，将推动光模块产品向更高速率的 800G、1.6T 甚至更高端产品的迭代升级。根据我们估算，全球光模块 400G 客户主要集中于亚马逊（约 45%）和谷歌（约 25%）、800G 主要集中于英伟达（约 50%）、谷歌(约 30%）和 Meta（约 20%）等，2025 年 1.6T 光模块的主要需求方预计将是英伟达和谷歌。在 GTC 2024 大会上，英伟达发布了其最新产品 GB200，其服务器与交换机端口速率也实现了翻倍提升，更有望引领 AI 光模块从现有的 800G 向更高性能的 1.6T 升级。此外，英伟达明确了 2026 年将使用 1.6T 网卡，对应 3.2T 光模块需求，明确了光模块升级迭代的节奏。LightCounting 预测，到 2029 年，400G+市场预计将以 28%以上的复合年增长率（每年约 16 亿美元以上）扩张，达 125 亿美元。其中 800G 和 1.6T 产品的增长尤为强劲，这两个产品共占 400G+市场的一半以上。与此同时，200G 以下速率光模块产品的市场规模预计将以每年约 10%的速度缩减。光模块头部厂商产品的高度可靠性、领先的研发实力及交付能力等优势将进一步凸显，行业集中度有望进一步提高。因此，那些能够与客户同步研发、快速融入客户供应链，并能提前把握客户需求的光模块厂商，将有机会在产品更新换代时抢先获利。

图18：2Q24 400G 和 800G 光模块需求强劲（百万美元）



资料来源：LightCounting，中国银河证券研究院

图19：2018-2028 年全球数通光模块各速率市场空间（百万美元）



资料来源：LightCounting，中际旭创 24 半年报，中国银河证券研究院

图20：英伟达数据中心芯片产品迭代线路图

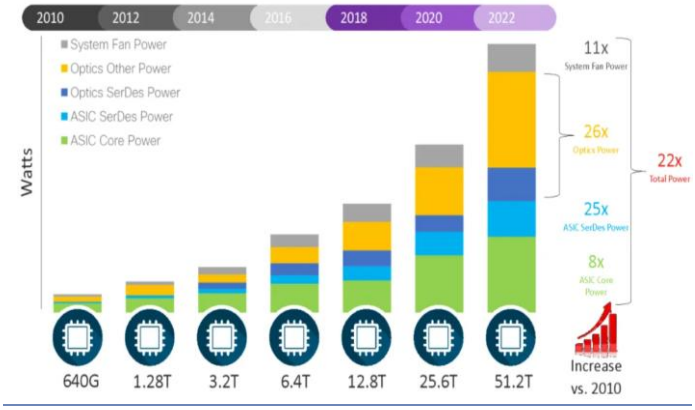
资料来源：英伟达，中国银河证券研究院

图21：互联速率在过去每四年翻一倍，2023 年开始每两年翻一倍

资料来源：Marvell，中国银河证券研究院

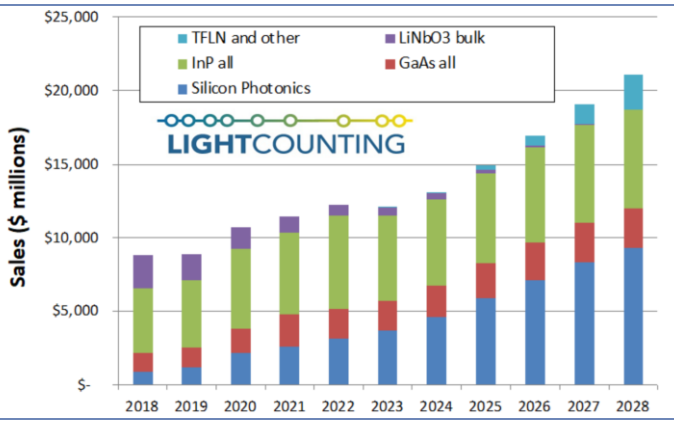
高速光模块的应用导致网络设备功耗大幅增加，硅光等新技术加固护城河。在以 400G 和 800G 光模块为典型配置的 51.2T 和 100T 交换机中，光模块加驱动 SerDes 的功耗占比在 40~45%。预计到 2030 年，在 400G+SerDes 和 6.4T 光模块代际时，OSFP 光模块功耗、SerDes 驱动距离将成为很难突破的瓶颈。据统计 2010-2022 年全球数通光模块的整体功耗提升了 26 倍，2024 年 800G 光模块正式放量后该问题更为突出，这种能耗增长对智算中心的运营成本构成了重大压力，降功耗成为光模块技术发展的核心诉求之一。硅光技术利用现有的 CMOS 工艺将光器件与电器件开发和集成到同一个作为光学介质的硅基衬底上，令光电处理深度融合，较传统分立器件更能发扬“光”（高速率、低功耗）与“电”（大规模、高精度）的各自优势。目前由于良率和损耗问题，硅光模块方案的整体优势尚不明显，在功耗、速率、成本、体积四个方面的突破是未来新技术发展的重点方向，也是未来光模块厂商竞争力的体现。根据 LightCounting 的预测，使用基于 SiP 的光模块市场份额将从 2022 年的 24%增加到 2028 年的 44%，硅光有望凭借硅基产业链的工艺、规模和成本优势迎来产业机遇。

图22：2010-2022 年光模块的整体功耗提升了 26 倍



资料来源：Cisco，Rosenberger，中国银河证券研究院

图23：基于 SiP 光调制器的光模块市场份额在 2028 年占 44%



资料来源：LightCounting，中国银河证券研究院

LPO 和 CPO 技术在功耗及成本上也各具明显优势，或成未来发展方向之一。LPO（线性驱动）技术通过移除 DSP 降低了光模块的成本和功耗，以 400G 光模块为例，其 7nm DSP 的功耗约 4W，占模块总功耗的一半，而 BOM 成本则占 20-40%，无 DSP 的 LPO 在功耗和成本上更具优势。然而，由于 DSP 的功能不能完全由 TIA 和驱动芯片替代，LPO 可能会增加误码率，进而缩短传输距离。因此 LPO 更适合短距离应用，如数据中心内部服务器与交换机的连接，以及机柜间的连接。而在 CPO（光电共封装）技术中，光学组件被直接封装在交换机芯片旁边，进一步缩短了光信号输入和运算单元之间的电学互连长度，在减少信号损耗问题的同时实现了更低的功耗，还有助于缩小设备体积，使得数据中心的布局更加紧凑。LightCounting 统计，CPO 出货预计将从 800G 和 1.6T 端口开始，并于 2024 至 2025 年开始商用，2026 至 2027 年开始规模上量，CPO 端口在 2027 年 800G 和 1.6T 出货总数中占比预计达约 30%。

表2：传统光模块与 LPO 与 CPO 方案技术优缺点对比

特性	传统光模块	LPO	CPO
功耗	高	较低	低
成本	高	较低	低
时延	高	较低	低
产品成熟度	高	较低	较低
可维护性	好	好	较差
链路性能	好	一般	好
互联性	好	较差	较差

资料来源：Rosenberger，中国银河证券研究院

在光电子器件方面，随着算力资源的广泛部署及其网络基础设施建设的加速推进，MTP、MPO 这类密集连接的典型产品，以其独特的高密度设计显著降低了布线成本，同时增强了系统的可靠性和可维护性，为数据中心的长期发展提供了有力支持，需求展现出快速增长的态势。此外，传输速率的显著提升也驱动了光有源器件光口向多通道方向的快速发展，进而带动了市场对多通道密集连接器件产品的需求增长。在此背景下，研发、制造 MTP、MPO 等高密度光网络关键无源器件的企业将显著受益。太辰光是全球最大的密集连接产品制造商之一，其中 MT 插芯及部分无源光器件产品的技术水平在细分行业处于领先地位，公司凭借产品的高性价比优势，有望进一步提升在产业链的市场份额。

目前在数据中心和算力点内部，美国已经完成 400G 光口向 800G 光口的演进，正在向 1.2T、1.6T 推进。我国目前仍然以 400G 光口为主，预计明年 800G 光口成为主流。因此在数据中心、算

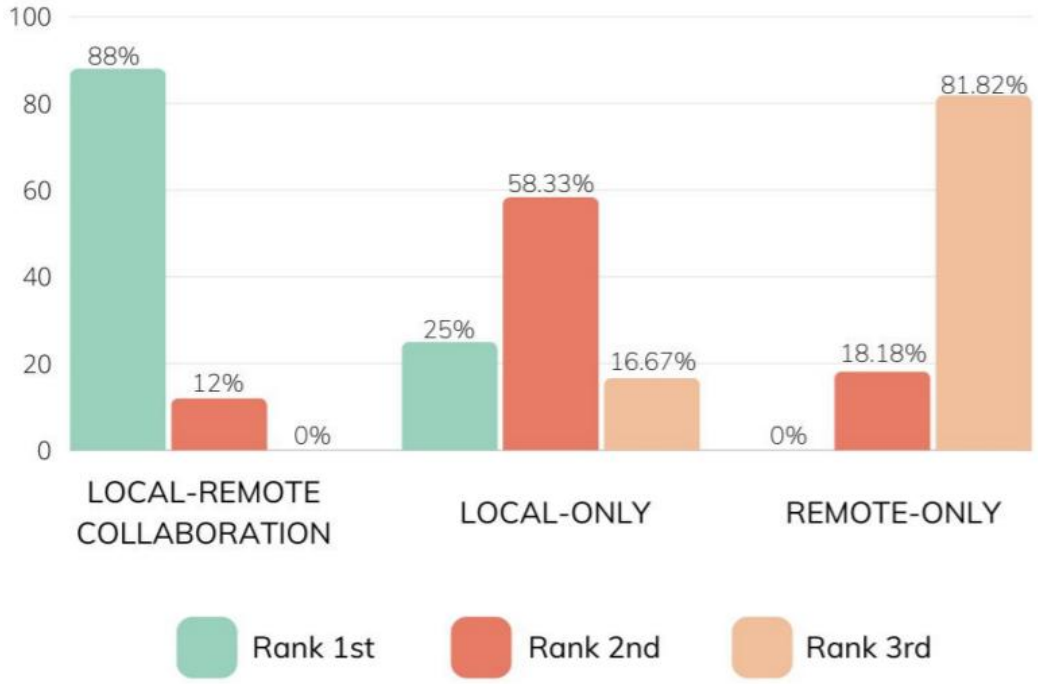
力点和算力集群之间迫切需要 400G/800G 光传送设备进行承载和传输。德科立在长距离光电子器件产品上不断推陈出新，在宽谱放大器、小型化可插拔放大器、高速率长距离相干和非相干光收发模块等领域保持较强的技术优势，有望随数据中心互联互通的建设升级而迎来更加广阔的发展空间。

TSV（硅通孔）技术是硅光芯片封装中的关键技术，其通过在硅片中创建垂直通道实现光芯片与电芯片间的高效电互连，促进了高密度集成和 3D 堆叠，增强了光电混合集成的性能和可靠性，对提升硅光芯片封装技术至关重要。晶方科技作为全球晶圆级芯片尺寸封装服务的主要技术引领者，拥有包括 TSV 在内的多样化先进封装技术，具备 8 英寸、12 英寸晶圆级芯片尺寸封装技术规模量产封装线，有望在提升高端光模块性能方面发挥关键推动作用。

（二）端侧大模型落地，智能硬件迎来星辰大海

LLM 单纯云端部署（例如 ChatGPT）并不广泛接受。如下图统计所示，88%的参与者倾向于边缘-云协作架构，其中 58.33%支持本地部署，81.82%对现有的仅云端解决方案不满意。他们的主要担忧是:1)远程大型语言模型服务的高延迟，2)将个人数据传输到云端的风险，3)云端大型语言模型服务的成本。

图24：个人对不同 LLM 部署策略的投票分布

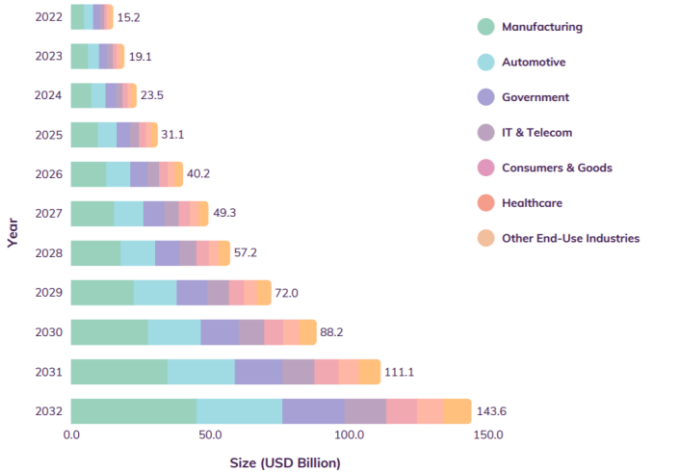


资料来源：Cornell university，中国银河证券研究院

2023 年边缘大型语言模型开始陆续爆发，当时出现了几个参数量低于 10B 的模型，使其能在边缘设备上运行，包括 meta 的 LLaMA 系列，微软的 Phi 系列，智谱的 ChatGLM，阿里巴巴的 Qwen 等。进入 2024 年创新步伐加快，边缘端部署的优势是能够缩短响应时间，并直接应用在如手机、汽车、可穿戴设备上。2022 年至 2032 年，按终端用户划分的全球设备边缘人工智能市场规模。市场将以 25.9%的复合年增长率增长，预计 2032 年的市场规模为 1436 亿美元。

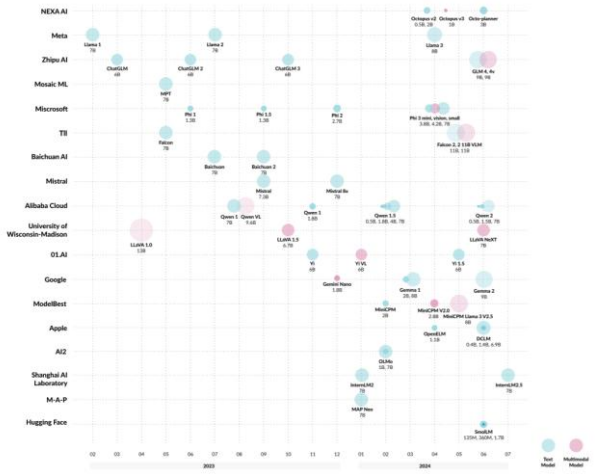
尽管在边缘端部署大模型有诸多优势，但考虑到端侧有限的计算能力、存储能力和能源限制等，使得直接部署基于云端的 LLM 困难重重。再评估设备端大型语言模型的性能时，有几个关键指标需要考虑：延迟、推理速度、内存使用、存储和能耗。通过优化这些性能指标，设备端大型语言模型能够在更广泛的场景中高效运行，提供更好的用户体验。同时针对边缘设备的部署，在保持性能的同时提高计算效率至关重要，通过量化、剪枝、知识蒸馏和低秩分解，这些方法通过平衡性能、内存占用和推理速度来提高大语言模型的运行效率，确保其在设备端应用中的可行性。

图25：边缘 AI 的市场规模（十亿美金）



资料来源：Cornell university，中国银河证券研究院

图26：端侧大语言模型的演变



资料来源：Cornell university，中国银河证券研究院

近年来，人工智能技术的迅猛发展和移动设备硬件的不断升级，使得在边缘设备上部署大型语言模型成为可能。作为人们日常生活中最常用的设备，智能手机上的语言模型引人注目。目前，全球主要手机品牌已开发并发布了多款先进的模型，这些模型采用设备端部署或设备-云协同策略。

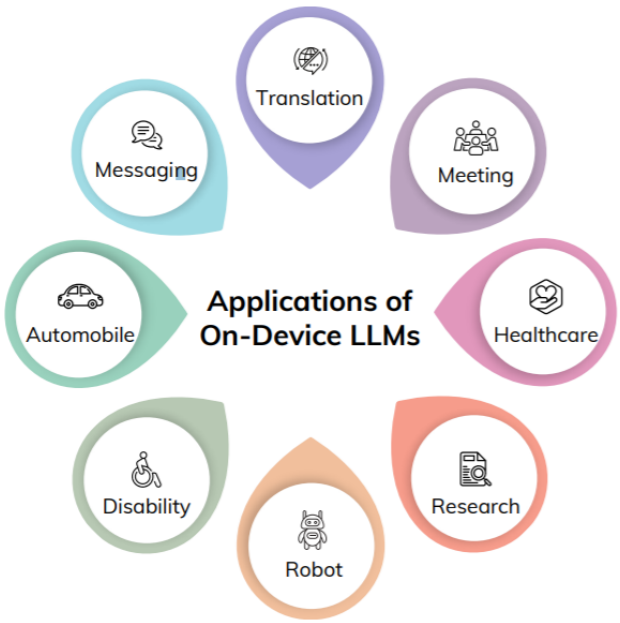
图27：手机厂商发布的设备端 LLM

Year	MODEL NAME	Model Size	Edge	Cloud
2023	Google Gemini Nano	7B	√	
2023	OPPO AndesGPT	7B	√	√
2024	Honor MagicLM	7B	√	
2024	VIVO BlueLM	7B	√	√
2024	XiaoMi MiLM	6B	√	
2024	Apple OpenELM	1.1B	√	√

资料来源：Cornell university，中国银河证券研究院

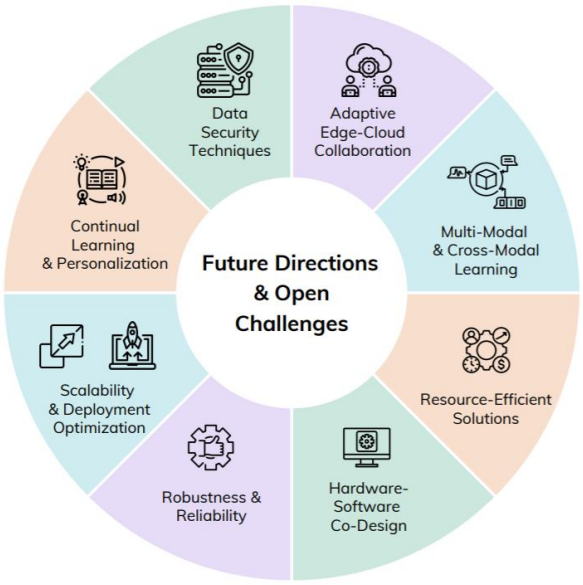
设备端语言模型正开启一个智能、响应迅速、个性化应用的新时代。通过将先进的自然语言处理能力直接引入用户设备，这些模型正在改变人们与技术互动的方式。从即时消息建议到实时语言翻译，从保密医疗咨询到尖端自动驾驶汽车。在资源受限设备上部署 LLM 面临独特挑战，这些挑战与传统的基于云的实施有显著不同。这些挑战涉及多个领域，包括模型压缩、高效推理、安全性、能源效率，以及与多样化硬件平台的无缝集成等。

图28：端侧 LLM 的应用



资料来源：Cornell university，中国银河证券研究院

图29：端侧 LLM 的挑战与未来方向



资料来源：Cornell university，中国银河证券研究院

人工智能技术的快速发展，“AI+”已经成为推动全球创新和经济增长的重要力量。相比 24 年 AI 基础设施相关个股业绩和股价的一骑绝尘，2025 年则可能是“AI+”百花齐放的开始。根据 QuestMobile 的数据，当下 LLM 的落地应用在网页端、移动端都已比较成熟，正逐步拓展至智能硬件端，不断深入用户日常生活并提供更自然便捷的交互体验。

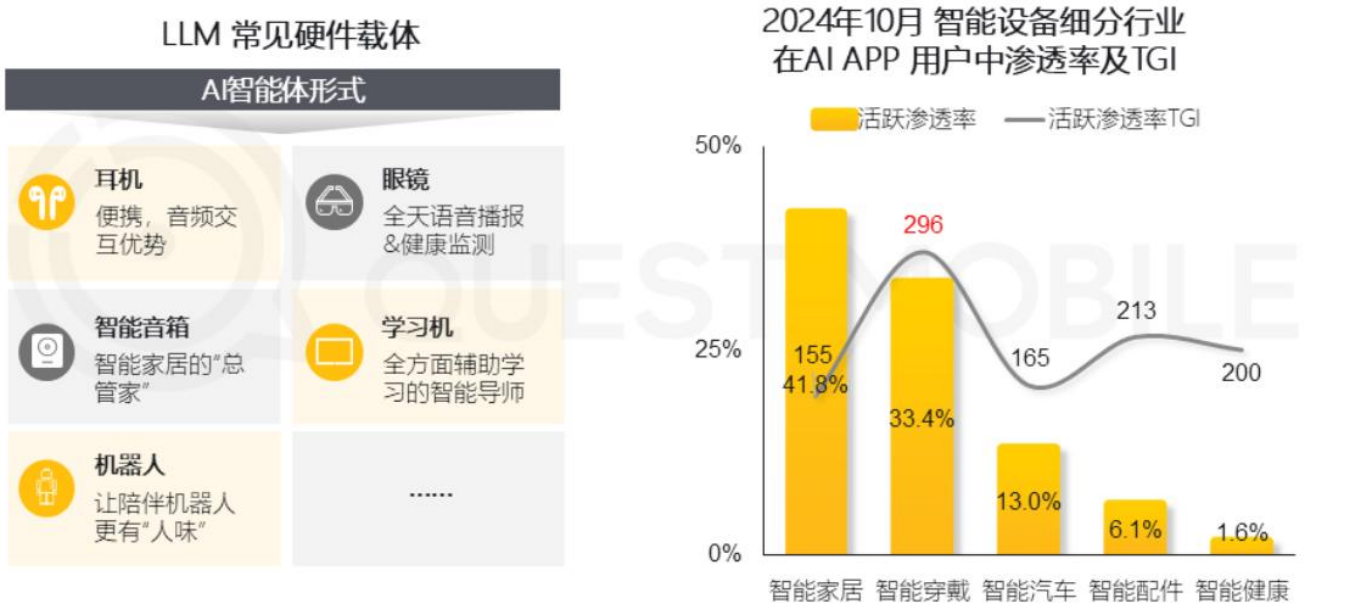
图30：LLM 落地三阶段

资料来源：QuestMobile，中国银河证券研究院

从硬件产品来看，手机是目前 LLM 最成熟的落地硬件载体之一，除手机外，AI 硬件首先以市

场成熟品类为切入点，如耳机、眼镜、智能音箱等。QuestMobile 数据显示，2024 年 10 月，智能穿戴行业在 AI APP 用户中渗透率 33.4%，TGI 达 296。

图31：智能硬件渗透率不断提升



资料来源：QuestMobile，中国银河证券研究院

AI 正在内容、应用、硬件、生态上影响世界，AI 智能体已从“数字”走向“具身”；随着市场发展，大模型更广泛地接入硬件产品，做好软硬件协同发展是未来竞争的关键。

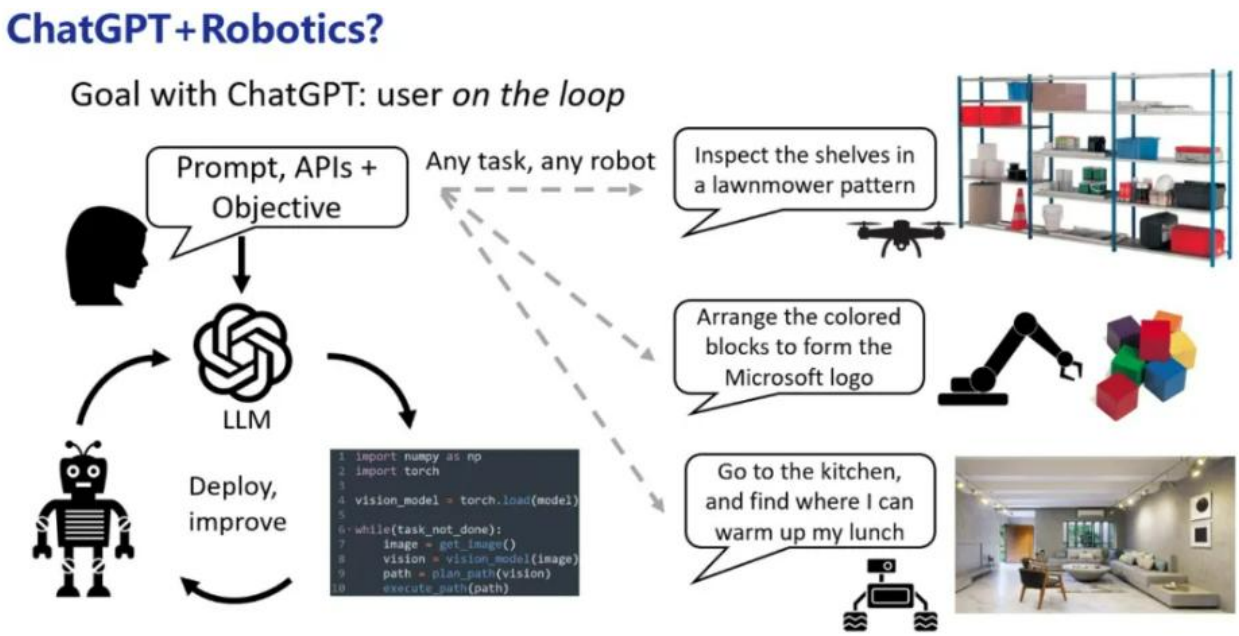
图32：AI 智能体“数字化”走向“具身化”

资料来源：QuestMobile，中国银河证券研究院

AI 技术可以赋予 IoT“人工智能大脑”。物联网可以将人与物、物与物连接成为一个整体，通

过 IoT 智能设备生成海量数据；AI 技术可以对海量数据进行深度学习、判断用户的习惯，提升用户体验，两者相辅相成，推动“万物互联”向“万物智联”进化。ChatGPT 的出现使得人工智能技术在语言交互方面的应用更为广泛，近日推出的插件功能，将进一步促进 AI 技术和其他产业的融合，AIoT 产业也将在 AI 技术升级的推动下不断发展。具身智能将是 AI 终端的最终形态，具身智能的核心在于如何理解世界、对世界进行建模，并基于此进行行为的决策以及与环境进行交互。大语言模型从本质上，只有数据和算法的迭代，而具身智能则需要把本体也一起囊括进来，需要本体、算法和数据一起联合迭代、优化和进化。

图33：ChatGPT+机器人组成具身智能体



资料来源：机器人大讲堂，人形机器人联盟，中国银河证券研究院

随着多模态大模型和世界模型（WMs）的出现，这些架构因其出色的感知、交互和推理能力而被视作具身代理的“大脑”。机器人可以通过接入大模型直接理解人类的自然语言指令，并将其转化为具体的行动。而当前我们依然处在“具身智能”的初级阶段，即智能硬件。AI+硬件也是未来 3-5 年消费电子的主要发展方向，值得关注。

图34：基于 MLMs 和 WMs 的具身智能体框架

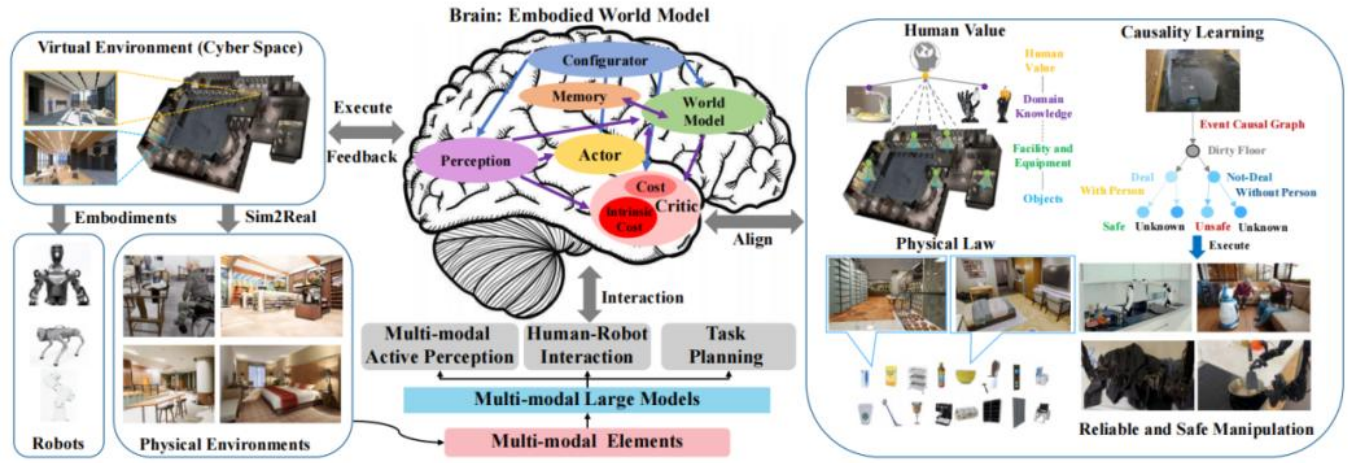


Fig. 2. The overall framework of the embodied agent based on MLMs and WMs. The embodied agent has a embodied world model as its "brain". It has the capability to understand the virtual-physical environment and actively perceive multi-modal elements. It can fully understand human intention, align with human value and event causality, decompose complex tasks, and execute reliable actions, as well as interact with humans and utilize knowledge and tools.

资料来源：Cornell University，中国银河证券研究院

国内互联网厂商包括科技公司等持续深度布局 AI 硬件赛道，通过 AI 软件+硬件的结合来推动 AI 的落地。

图35：互联网和科技企业在智能硬件的布局

互联网企业 &AI企业 智能硬件布局概览

	企业	硬件类型	具体硬件	场景类型
互联网企业	百度	智能穿戴	小度耳机、小度眼镜	泛场景
	百度	教育终端	Z30学习机、青禾学习手机	垂直场景
	阿里巴巴	智能穿戴	爱富耳机	泛场景
	阿里巴巴	智能家居	天猫精灵、天猫智慧屏	泛场景
	有道	教育终端	有道词典笔、有道听力宝、有道AI学习机	垂直场景
	抖音	智能穿戴	Ola Friend智能体耳机	泛场景
科技公司	科大讯飞	智能穿戴	讯飞会议耳机、智能助听器	垂直场景
	科大讯飞	教育终端	T30学习机	垂直场景
	科大讯飞	办公终端	办公本、智能录音笔	垂直场景
	星海图	机器人	R1系列仿人形机器人	泛场景

资料来源：QuestMobile，中国银河证券研究院

图36：智能可穿戴设备工作流演示

资料来源：QuestMobile，中国银河证券研究院

表3：“AI+”硬件是未来消费电子发展方向

具体产品	品牌厂商	相关 AI 产品
智能手机	三星	三星 Galaxy S24 系列通过融合本地和云端 AI 体验的 Galaxy AI，充分释放了移动设备的生产力潜能，让智能手机可以处理更多、更重要的工作事务，成为真正意义上的智能端。 Galaxy AI 依托三星 Galaxy 设备强大的端侧芯片算力和持续优化的模型压缩算法，通过端侧 AI 即可为用户提供强大的翻译功能。针对交谈场景，三星 Galaxy S24 系列可以进行通话实时翻译和同传。 借助由 Galaxy AI 支持的高精度图像分割、OCR 识别以及大模型理解能力，三星 Galaxy S24 系列还为消费者带来了创新的即圈即搜功能，让用户只需长按 Home 按钮，然后通过简单的圈选手势，即可搜索屏幕上感兴趣的内容。
	苹果	Apple Intelligence 带来了多项新功能，包括写作助手、重新设计的 Siri、在语音命令和输入之间切换的选项、摘要功能、新邮件分类和智能回复等。
智能眼镜	Meta	Orion 包含内置的情境人工智能，能够“感知和理解”佩戴者周围的世界，从而“预测并主动满足”佩戴者的需求。
智能耳机	字节跳动	Ola Friend 为开放式耳机，单耳重量为 6.6 克，接入了豆包大模型，与豆包 App 深度结合，用户戴上耳机后，无需打开手机即可通过语音唤起豆包进行对话。该耳机接入了字节豆包大模型的 Seed-ASR（语音识别）技术模型。该模型可以高精度识别中英文、口音，甚至能通过上下文，“聪明”地识别各类信息。

资料来源：CNMO，IT 之家，中国银河证券研究院

四、软件端：大模型演进加速，看好 AI Agent 发展

总体来说，DeepSeek 通过在算法与工程侧的深度耦合，把相同的算力资源利用率最大化，所以我们看到单次训练成本有显著下降。

表4：DeepSeek 大模型与 GPT-4、Llama3.1405B 对比

	GPT-4	DeepSeekV3	DeepSeekR1	Llama3.1405B
发布时间	2023 年 3 月	2024 年 12 月	2025 年 1 月	2024 年 7 月
参数	1.8 万亿	6710 亿，每次激活 370 亿	15 亿-70 亿（不同规模蒸馏版本）	4050 亿
训练数据 token	13 万亿	14.8 万亿	-	15 万亿
训练成本	6300 万美元	557.6 万美元	-	-
训练时长	-	278.8 万 GPU 小时（2048 张 H800）	-	3080 万 GPU 小时（1.6 万张 H100）
是否开源	否	是	是	是
API 价格	输入端：0.03 美元/千 token 输出端：0.06 美元/千 token（8K 版本）	输入端：0.14 美元/百万 token 输出端：0.28 美元/百万 token	输入端：0.55 美元/百万 token 输出端：3.29 美元/百万 token	-

资料来源：DeepSeek、CSDN、OpenAI，中国银河证券研究院

图37：DeepSeek 发布文生图 Janus-Pro，基础测试中超越 OpenAI

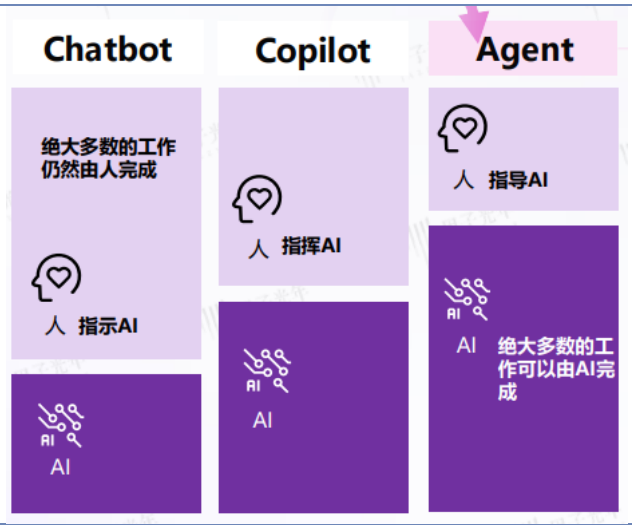
资料来源：DeepSeek 官网，中国银河证券研究院

（一）DeepSeek 加速 AGI 到来，大模型从“训练”向“推理”演进

DeepSeek 将加速 AGI 时代到来，AIAgent 将成为通往 AGI 的基石。DeepSeekR1 是人工智能革命下里程碑式的产品，对标 OpenAI 的 o1 模型，并且在强化学习的推动下，展现出了此前未曾预见的推理能力，同时通过工程与算法等深度耦合，大幅降低成本，让大模型更易触达下游厂商。

并且 DeepSeek 开源其模型预示着开源社区正以全新的方式推进人工智能技术的发展，加速通用人工智能时代到来并推动 AI Agent 技术更加成熟。

图38：AIAgent 处于早期阶段，逐渐由 Copilot 进入到 AIAgent 探索阶段



资料来源：甲子光年，中国银河证券研究院

AIAgent 是一种能自主感知周遭环境，通过内在的智能处理进行决策，并执行相应行动以达成特定目的的智能体。它基于大型语言模型（LLM），集成了规划、记忆、工具和行动能力。从智能助手、个性化推荐系统到自动化客户服务，AIAgent 的应用案例层出不穷，它们在各行各业中展现出巨大的潜力和价值。

AI 代理的工作流程比传统的 LLM 交互方式更高效:通过迭代式的 AI 代理工作流程（例如：先写提纲，再进行网络搜索，再写初稿，再修改），可以显著提高 AI 模型的输出质量，其提升程度甚至超过了模型本身的迭代升级。

表5：AIAgent 可以通过设定目标完成自动化

名称	自动化的实现方式	含义
Chatbot	-	人类完成绝大部分工作，类似于向 AI 询问意见，了解信息：AI 提供信息和建议，但不直接处理工作。
Copilot	借助复杂的提示词完成自动化	人类和 AI 进行写作，工作量相当：AI 根据人类的指令完成工作的初稿。人类负责目标设定、修改调整，并最终确认工作成果。
Agent	通过设定目标完成自动化	AI 完成绝大部分工作，人类负责设定目标、提供资源和监督结果：AI 负责任务拆分、工具选择、进度控制，在达到目标后，AI 能够自主结束工作。

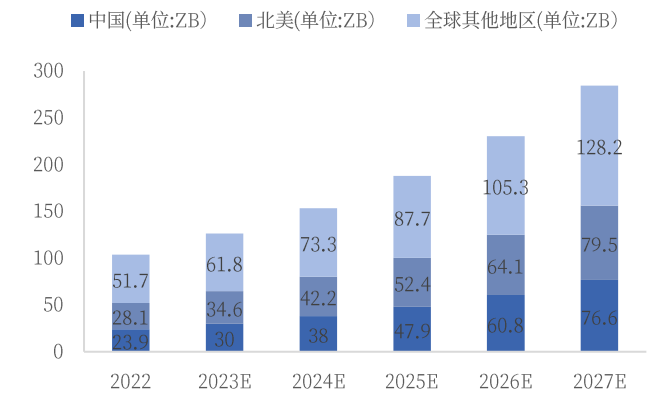
资料来源：甲子光年，中国银河证券研究院

未来人工智能发展关键在于推理，Capex 逐渐转向经营性性质。传统训练任务的 Capex 是研发型投入，而近期 OpenAI 发布的 GPT-o1 所采用的推理模型，是被设计用来处理长时间的思考和多步骤的复杂任务，为用户的决策提供支持，GPT-o1 在推理过程中的成本被描述为数百万甚至数十亿级别的提升，这预示着未来 AIAgent 广泛渗透时，对于推理算力的需求将是指数级爆发增长，推理的 Capex 的日常经营性质越发明显。AIAgent 推理和规划的能力由 LLM 来实现，推理和规划赋能 Agent 学习能力，可以积累知识和经验，并且 Agent 可以对过往的数据和动作进行反思总结，从错误中吸取经验，并为接下来的行动进行纠正，从而适应环境、更有效地执行任务并成功达成目标。

（二）AI Agent 崛起，B 端+C 端应用开启新篇章

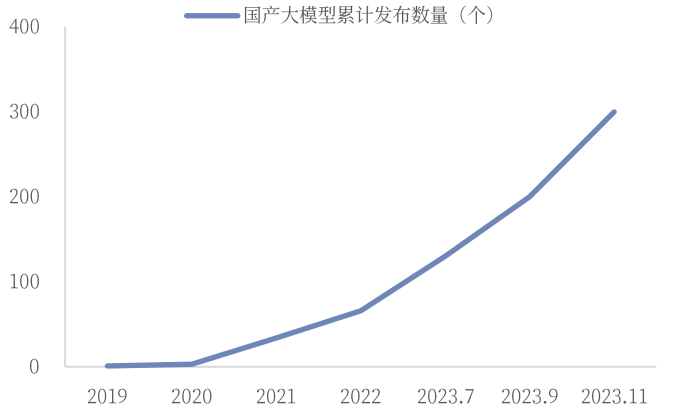
DeepSeek 推动 AIAgent 快速进化，AI 应用进入新时代。当前，伴随全球数据量维持高速增长，为 AIAgent 发展提供数据资源。未来五年，全球数据量将维持增长，2022 年，全球数据规模已达到 103ZB，中国数据规模达到 23.9ZB；预计 2027 年，全球数据规模可达到 284.3ZB，近五年的 CAGR 可达到 22%，中国数据量规模则可达到 76.6ZB，近五年的 CAGR 为 26%，超过全球增长速度。国产大模型自 2023 年 7 月开始进行密集发布，截至 2023 年 7 月，国产大模型累计数量达到 300 个，并且涉及金融、法律、教育、医疗、娱乐等多个垂直细分领域。

图39：全球数据量持续增长，为 AIAgent 发展提供数据资源



资料来源：甲子光年，中国银河证券研究院

图40：国产大模型数量以指数级增长



资料来源：甲子光年，中国银河证券研究院

中国 AIAgent 市场空间广阔，B 端、C 端大有可为。2023 年中国 AIAgent 市场规模为 554 亿元，预计至 2028 年将达 8520 亿元，其年均复合增长率为 72.7%。AIAgent 于 2023 年被业内正式引入并重新定义，随着人工智能技术的迅速发展，垂直领域的 AIAgent 正逐渐成为科技行业的新宠，垂直领域的 AI 代理市场规模可能达到 SaaS 的十倍，创造超过 3000 亿美元的独角兽企业。**AIAgent 市场规模包括 ToC 端和 ToB 端的应用价值：**1）在 B 端场景下，AIAgent 将对 SaaS 应用进行全面重构，与传统知识库结构化管理模式相比，AIAgent 的向量数据库能自动学习和理解文档，实现更加高效知识管理；2）在 C 端场景下，AIAgent 作为生成式 AI 的商业化应用，可以广泛应用于电商、教育、旅游、酒店以及客服等行业，带来传统行业的升级转型。

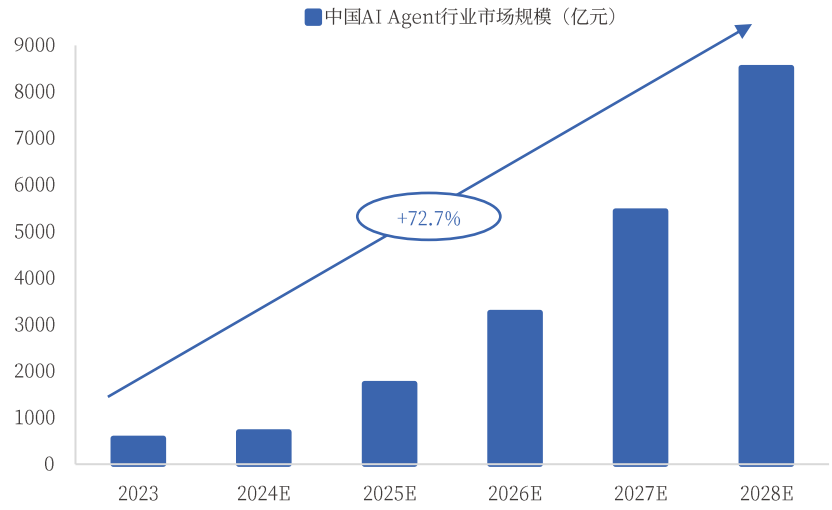
表6：AIAgent 可以通过设定目标完成自动化

	2023	2024E	2025E	2026E	2027E	2028E
中国 SaaS 市场规模（亿元）	550	688	859	1074	1343	1678
yoy	-	25.09%	24.85%	25.03%	25.05%	24.94%
SaaS 重构应用价值倍数	1	1	2	3	4	5
yoy	-	0%	100%	50%	33.33%	25%
B 端 AIAgent 市场规模=中国 SaaS 市场规模×SaaS 重构应用价值倍数	550	688	1718	3222	5372	8390
中国生成式 AI 市场规模（亿元）	80	104	156	234	304	395
yoy	-	30%	50%	50%	29.91%	29.93%
AIAgent 渗透率	5%	7%	11%	16%	23%	33%
yoy	-	40%	57.14%	45.45%	43.75%	43.48%
C 端 AIAgent 市场规模=中国生成式 AI 市场规模×AIAgent 渗透率	4	7.28	17.16	37.44	69.92	130.35

中国 AIAgent 市场规模=B 端 AIAgent 市场规模+C 端 AIAgent 市场规模	554	695.28	1735.16	3259.44	5441.92	8520.35

资料来源：头豹研究院，中国银河证券研究院

图41：2023-2028 中国 AIAgent 行业市场规模及预测



资料来源：头豹研究院，中国银河证券研究院

五、应用端：AI+赋能进行时，行业加速繁荣可期

在 AI 技术的全球竞争中，以 DeepSeek、Kimi、豆包等为代表的国产大模型凭借其前沿的技术创新与精准的场景化应用，正强势突围。在 C 端，用户渗透率不断提升，主要 AI APP 活跃数据持续环比增长；在 B 端，AI 营销等领域的商业化模式已经逐步得到验证。**我们认为，DeepSeek-R1 通过强化学习实现了低成本与高性能的结合，其发布并开源为行业生态发展带来了新的可能性，并且有望加速推动在影视、广告、社交陪伴等多个领域应用落地。**

（一）开源的生态推动 AI 行业高速发展

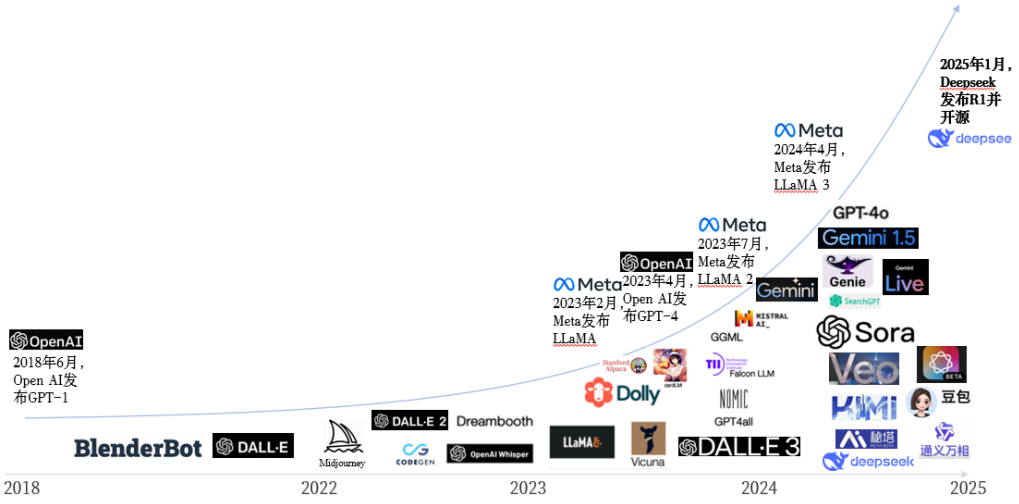
回顾人工智能的发展历程，目前全球人工智能的发展大致经历了两个阶段：

1）1.0 阶段——被动分析与判断：早期的人工智能的主要能力是被动地接受信息来进行分析和判断。比较典型的例子如：生物识别技术（根据人脸、虹膜等特征信息输入判断是否匹配）。这些技术没有主动创造内容的能力，更多地是对人类所输入信息的判断和匹配。

2）2.0 阶段——生成式 AI 出现：随着人工智能的不断发展，AI 的能力不再仅限于对被动输入信息的接受和分析，而是具备了一定的主动输出内容的能力。在这一发展进程中，Transformer 和 Diffusion Model 这两个算法模型对推动生成式 AI 的发展起到了重要的作用。目前的生成式 AI 已经可以自主生成文本、图片、视频等多种模态的信息。

在人工智能的发展过程中，开源的生态起到了重要的作用：头部 AI 公司引领着各项技术向前，并使得后来者能够了解到最新的技术进展并发展相关的技术应用，而技术应用又进一步促进 AI 技术的发展。**我们认为，DeepSeek 推出的开源推理模型具备显著的成本优势，大幅降低了企业接入门槛，有利于应用端的开发创新，将极大地推动 AI 应用生态的蓬勃发展。**

图42：AI 开源加速技术发展进程



资料来源：Open AI，Meta，Google Deep Mind，中国银河证券研究院整理

在开源生态的大背景下，随着相关技术的不断迭代，我们认为 AI Agent 有望成为 AI 浪潮的下一个发展方向。AI Agent 将进一步改变人们的日常生活：AI Agent 不仅能够提高工作效率，优化资源配置，还将在个性化服务、智能决策支持等方面发挥重要作用，有望成为推动社会进步和创新的关键。目前，头部互联网大厂相继在 AI Agent 领域积极布局，我们认为凭借其强大的技术资金实力、丰富的数据资源和庞大的用户基础正开启追赶模式，潜力巨大。

表7：2024 年至今国内互联网大厂 AI 重要进展梳理

公司	时间	具体技术/模型
百度	2024 年 4 月	发布文心大模型 4.0 的工具版。可以体验代码解释器功能，通过自然语言交互，就能实现对复杂数据和文件的处理与分析，还可以生成图表或文件，能够快速洞察数据中的特点、分析变化趋势、为后续的决策提供支撑。
	2024 年 6 月	发布了文心大模型 4.0 Turbo，同时发布了飞桨新一代框架——飞桨框架 3.0，具备动静统一自动并行、编译器自动优化、大模型多硬件适配、大模型训推一体等核心技术，支撑大模型效果更好，性能更优。
	2024 年 11 月	发布两大赋能应用的 AI 技术：检索增强的文生图技术（iRAG）和无代码工具“秒哒”。文心 iRAG 用于解决大模型在图片生成上的幻觉问题，极大提升实用性；无代码技术“秒哒”让每个人都拥有程序员的能力，将打造数百万“超级有用”的应用。
阿里巴巴	2024 年 4 月	推出最新语言大模型“通义千问”,并陆续接入阿里巴巴生态的所有商业应用中，如企业通讯、智能语音助手、电子商务、搜索、导航、娱乐等，从而进一步提升用户体验。
	2024 年 10 月	发布 AI 生意助手 2.0，在外贸经营的四大难点：发品、接待、营销、合规领域分别为中小企业配备了 4 个专业的 AI Agent，在各自领域帮外贸人找到更高效的经营方式，实现生意增长。
	2024 年 11 月	阿里发布全新 AI 推理模型 QwQ-32B-Preview 并同步开源，整体推理水平比肩 OpenAI o1。
	2025 年 1 月	发布了 Qwen2.5-Max 旗舰版模型，其预训练数据量超过了 20 万亿 tokens，且综合性能在多项主流模型的评测中均展现出卓越表现。
腾讯	2024 年 9 月	推出 AI Infra 品牌“腾讯云智算”，集计算、储存和网络解决方案于一体，通过优化基础设施，助力企业更高效地开发和训练大模型。 基于 MoE 架构，首次发布了“腾讯混元 Turbo”模型，与上一代混元模型相比，其推理效率提升了一倍，而推理成本则降低了 50%。
	2025 年 1 月	腾讯混元宣布开源 3D 生成大模型 2.0 版本，并上线业界首个一站式 3D 内容 AI 创作平台——混元 3D AI 创作引擎。
字节跳动	2024 年 5 月	发布字节跳动豆包大模型家族、火山方舟 2.0、AI 应用及 AI 云基础设施等一系列最新产品。

	2025 年 1 月	豆包大模型 1.5 正式发布，目前通用模型 pro 已在豆包 APP 灰度上线，同时开发者可在火山引擎直接调用 API。

资料来源：新智元，中国银河证券研究院整理

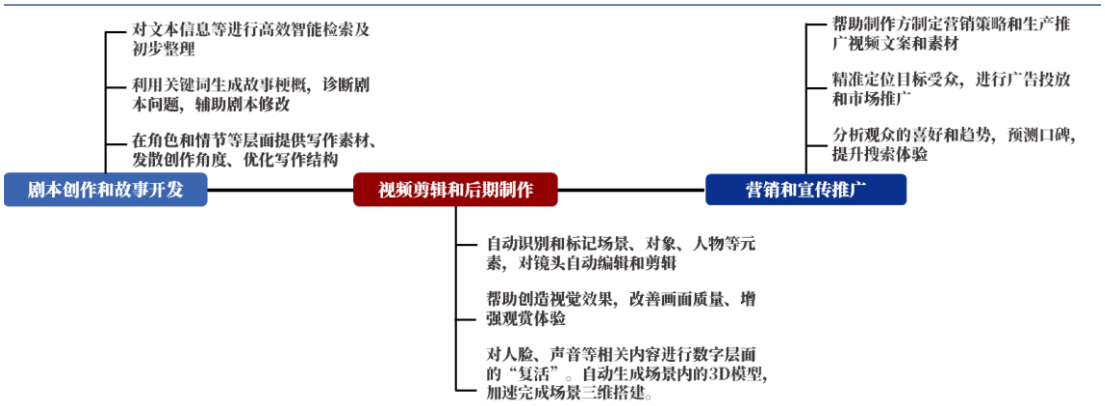
（二）AI 应用：“AI+”行业应用百花齐放

AI+影视：

AIGC 有望在影视生产全环节得到广泛应用。前期策划阶段，AIGC 可根据电影主题、风格、人物等要素自动生成剧本草稿。此外，AI 还可以辅助分镜制作，提供视觉参考，从而加快电影前期准备工作。电影制作阶段，多模态 AI 可以低成本地生产图片、音频、视频等素材，从而提供更多元的内容供给。AI 技术还可用以辅助场景生成，特效制作等环节，从而为影片带来更逼真的视觉效果和更丰富的细节。AI 工具还为视频处理提供了有力的工具，大幅降低了视频去除噪点、模糊、抖动，提升画质，提取关键镜头的剪辑难度。

此外，AIGC 还可用以进行风格迁移，从而加速电影 IP 向周边商品、漫画、游戏等媒介的落地。宣发阶段，AI 模型可根据用户画像定制优化预告片、海报及展示的评论等。同时结合虚拟数字人技术，发行商有望以极低的成本实现映前观众与电影人物的“面对面”交流，从而持续地在维持电影话题热度。

图43：AI 赋能影视生产全环节



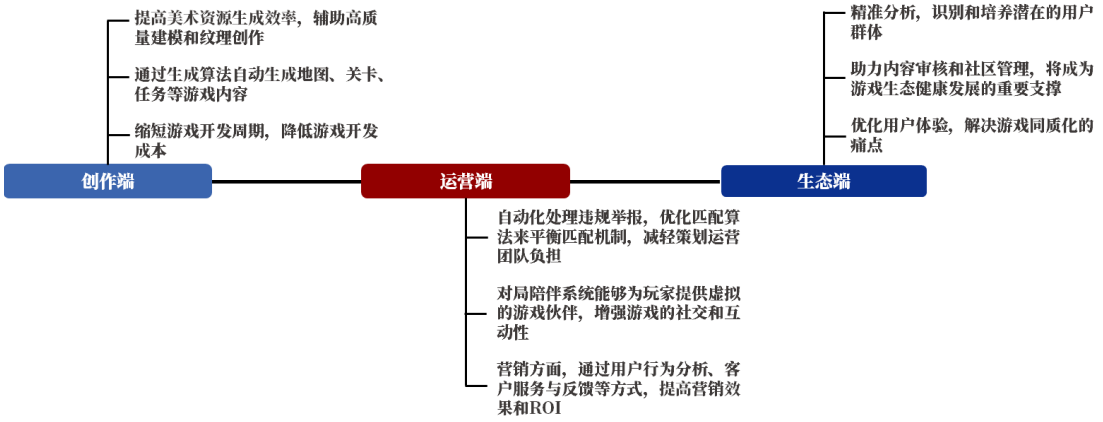
资料来源：Wind，中国银河证券研究院整理

当前，AIGC 蓄势待发，预备在影视行业掀起新一轮的技术变革和竞争力洗牌，国内各大平台与影视公司也在 AIGC 的应用研究与业务协同上开启了应对未来的准备。从《斗罗大陆》《吞噬星空》到《三体》《遮天》，腾讯视频在视效技术规模化运用和长期投入方面展现出了灵敏的嗅觉。自 AIGC 出现后，腾讯视频在动画行业中也开始探索二维和三维界限的打破。**我们认为，未来通过借助三维工业化流程和 AIGC 的辅助，如文生图、图生图、文生视频的能力，行业难题有望被突破，二维动画的效率和产能将有进一步的提升。**

AI+游戏：

游戏作为集合了文字、图像、声音、视频等内容形式的商业化应用，有望更好地在研发端利用多模态大模型的能力，AI 对游戏行业的长期催化作用值得看好：经过技术探索和商业化落地，AIGC 技术已被广泛应用在游戏资产生成，仿真场景渲染等多个环节。遵循供给端降本增效，需求端革新交互体验，挖掘用户付费意愿的逻辑，我们认为，AI 技术将通过 AIGC 工具（绘画工具、文本创作、语音合成等）和 AI 工具来对游戏行业全产业链条进行重塑：

图44：AI 赋能游戏产业链



资料来源：Wind，中国银河证券研究院整理

1）创作端：传统的游戏创作端存在着资源生成效率低、成本高昂的痛点。特别是在美术资源的制作上，高质量的 3D 模型和纹理的创作往往需要大量的手工艺术家工作时间，这不仅使得游戏开发周期延长，而且大幅度增加了开发成本。在 AI 技术的加持下，游戏创作将在自动化内容生成（如 AIGC 绘画工具和 3D 模型生成）方面发生根本性的变革。具体而言，**AIGC 可以通过生成算法自动生成地图、关卡设计、任务等游戏内容，分析玩家数据并进行游戏平衡性调整等，提高开发效率和游戏多样性**。

2）运营端：在游戏运营方面，AI 技术可以通过智能 NPC、智能 BOT 和掉线托管等应用，解决现有运营工作杂、营销转化弱的难题。例如，AI 可以帮助处理违规审判，通过学习判断何为游戏内的违规行为，自动化处理大量的审判工作，减轻运营团队的负担；AI 也能实现平衡匹配，通过分析玩家的技能水平和游戏习惯来优化匹配算法，从而提供更公平、更有趣的游戏体验；同时，对局陪伴系统能够为玩家提供虚拟的游戏伙伴，增强游戏的社交和互动性。在游戏营销方面，AIGC 可以通过用户行为分析、客户服务与反馈、营销预测和广告优化等方式，提高营销效果和 ROI。

3）生态端：在游戏生态构建上，AI 技术的应用将有助于优化用户体验和增强运营工作的自动化，解决游戏体验同质化的痛点。AI 可以通过精准的数据分析来提升营销转化率，为运营团队提供决策支持，并帮助他们识别和培养潜在的用户群体。此外，AI 在内容审核和社区管理方面的应用，如自动化过滤不良信息，也将成为游戏生态健康发展的重要支撑。最终，这些进步将推动游戏行业朝着更加智能化和个性化的方向发展。

AI+社交陪伴：

AI 驱动的虚拟助手，如 Siri、GoogleAssistant，能够通过语音识别和自然语言处理（NLP）提供陪伴服务，如帮助日常任务、提醒事项、甚至进行简单的闲聊。从而增强用户体验，特别是在老年人、孤独人群中的应用，可以提供情感支持和便利。另外,AI 用于生成虚拟人类角色，能够进行情感交流、陪伴聊天，甚至根据用户需求提供心理疏导,能给用户提供个性化、情感化的陪伴服务，尤其在远程工作或社交限制的环境下，满足用户的社交需求。

AI+社交陪伴领域正迎来一场革命性的变革。在这新兴领域，AI 可以通过分析用户的历史对话和行为模式，自动生成符合用户个性的对话脚本和互动建议。这种个性化的服务能够让用户感受到更加贴心的社交体验。借助于先进的情感识别技术，AI 能够识别用户的情绪状态，并据此调整其回应策略，提供更加贴合用户情感需求的陪伴。同时 AI 技术可以创建虚拟角色，这些角色不仅能够进行自然语言对话，还能够模拟真实人类的表情和肢体语言，为用户提供一种全新的社交体验。对于那些社交技能较弱的用户，AI 可以通过模拟社交场景，提供社交技能训练，帮助用户提高社交能力。目前像 ChatGPT、豆包以及 Minimax 等能提供自然生动的语音合成能力，善于表达多种情绪，演

绎多种场景,备个性化的角色创作能力，更强的上下文感知和剧情推动能力，满足灵活的角色扮演需求。

图45：AI 赋能社交陪伴领域

资料来源：Wind，中国银河证券研究院整理

我们认为：AIGC 技术目前已经能从语言、语气等多方面深入洞悉人类的多种情感，同时作出判断给予不同的情感价值和需求。无论是在增强用户体验方面,还是拓展更多个性化的服务等方面，都具有丰富的想象空间,AI陪伴未来可能是成为工作中最得力的助理,也是最了解用户习惯的销售员，进一步的创新有望开发商业化应用落地。

AI+电商：

AI 分析用户的购物历史、浏览记录和偏好，向用户推荐可能感兴趣的商品，提升转化率和购买欲望。平台如 Amazon、淘宝等都在使用此技术,可以提高销售额和客户满意度，同时减少用户的决策疲劳，提升购物体验。此外,AI 驱动的聊天机器人和语音识别技术被广泛应用于电商平台，能够实时响应客户咨询、处理订单问题、解决售后问题,有效提高客户服务效率，降低人力成本，同时提升客户体验。

人工智能（AI）技术的应用领域不断拓展，电商行业也在积极探索新的 AI 应用，以保持竞争力并满足消费者日益增长的期望。最初，电商平台采用聊天机器人提供 24/7 客户服务，解答用户问题并处理订单查询。通过自然语言处理技术，系统能够理解用户意图，提供更准确的服务。此外，预测分析工具被用于预测产品需求，优化库存水平，减少积压和缺货现象。AI 还被应用于供应链管理，提高物流效率等 B 端场景。在 C 端方面，Google、OpenAI 等公司正积极布局 AI 与搜索在电商中的应用。通过对大量数据的深度学习，分析客户的潜在消费需求，实现精准推荐。

我们认为，生成式 AI（AIGC）正在重塑电商格局。在生产端，AIGC 辅助商家拓展业务，降低销售和运营成本。在消费端，购物模式将从“人找货”过渡到“货匹配人”，最终实现“货找人”，这将带来巨大的消费增量。

图46：AI 赋能电商全流程

资料来源：Wind，中国银河证券研究院整理

AI+营销：

AI 技术被用于分析消费者的行为数据，生成用户画像，并根据这些数据进行精准广告定向投放（如 Facebook、Google 广告），更有效地提高广告的点击率和转化率，减少广告浪费，提升广告主的投资回报率（ROI）。同时，AI 可以自动生成与用户相关的个性化内容，如广告文案、电子邮件和社交媒体帖子。基于用户的兴趣、行为模式生成个性化营销信息，增强与消费者的互动，提高品牌忠诚度和参与度。

人工智能（AI）的应用正深刻变革营销服务商的商业模式，推动行业降本增效。通过结合底层大型语言模型（LLM），并利用长期积累的广告投放案例、用户数据和行业数据，企业能够为不同行业定制专业化的广告投放模型。这使广告主将更多预算转向融合 AI 技术的增值服务，从而提升综合毛利率。例如，AppLovin 开发了 AI 广告引擎 Axon 2.0，显著提高了广告投放的精准度，推动公司业绩增长。通过深度学习和机器学习算法，分析用户行为数据，精准识别目标受众，并在适当时机投放最具吸引力的广告。此外，利用大数据分析，预测用户行为，更有效地进行个性化推荐，使广告商能够与更可能下载其应用的用户匹配，以获得更高的留存率。

我们认为，AI 已经重塑了营销行业的生态。它不仅为营销人员提供精准的数据支持，助力营销策略的制定和优化，还推动了“一人多面”的个性化营销，使针对每个用户生成定制化的内容和服务成为可能。同时，AI 保持了大规模营销活动的高效执行，实现了个性化与规模化生产的平衡。

图47：AI 赋能营销环节

资料来源：Wind，中国银河证券研究院整理

六、投资建议：硬件产业链加速发展，应用端方兴未艾

（一）电子板块：Scaling Laws 转向后训练，计算效率提升至关重要

我们乐观看待 DeepSeek 创新对电子行业带来的改变。我们总结以下几条结论：

1，**我们认为 DeepSeek 的创新并没有完全打破 scaling laws** 。DeepSeek 模型具有更强的性能，更低的训练与推理成本，将加速推动 AI 应用与硬件的普及和落地。虽然更低的训练与推理成本减少了当前的算力需求，但是并不意味着 AI 的未来发展对半导体整体需求的减少，相反由于其模型架构、基础设施数据等方面的优化，以及更低的成本，使得其更加容易布置在端侧，从而加速 AI 的普及。AI 能力边际的扩张依然需要依赖更大的模型和强大的算力，DeepSeek 在算法和架构上的创新给 AI 的发展增加了一条新的道路。

2，**Scaling laws 正在从 pre-training 转向 post-training 和推理，**通过增加模型规模、扩展训练数据、提高计算资源以及合理的任务设计，可以加速模型学习更复杂的推理能力，这一过程遵循 scaling law。随着模型规模、数据量和计算资源的增加，模型能够更好地进行推理。

3，**针对边缘设备的 LLM 部署，在保持性能的同时提高计算效率至关重要，**通过量化、剪枝、知识蒸馏和低秩分解，这些方法通过平衡性能、内存占用和推理速度来提高大语言模型的运行效率，有利于 AI 硬件端的落地与普及。我们看好 AI 应用持续落地带来的传统消费电子的换机周期，苹果产业链值得关注，同时看好 AI 终端硬件如耳机、眼镜、桌面机器人、小家电、周边硬件等。建议关注：寒武纪、海光信息、蓝思科技、鹏鼎控股、领益智造、水晶光电、蓝特光学、恒玄科技、中科蓝讯、乐鑫科技、瑞芯微、全志科技、翱捷科技、敏芯股份、兆易创新、普冉股份、艾为电子。

表8：建议关注相关标的盈利预测情况-电子（截至 2025 年 1 月 31 日）

代码	标的名称	总市值（亿元）	EPS（元）			P/E		
			2024E	2025E	2026E	2024E	2025E	2026E
300433.SZ	蓝思科技	1296.05	0.80	1.10	1.37	32.61	23.61	18.94
002938.SZ	鹏鼎控股	938.79	1.55	1.94	2.21	26.04	20.82	18.35
002600.SZ	领益智造	596.40	0.29	0.42	0.55	29.37	20.05	15.37
002273.SZ	水晶光电	304.41	0.74	0.92	1.11	29.45	23.73	19.76
688127.SH	蓝特光学	109.63	0.67	0.92	1.13	40.67	29.47	24.00
688608.SH	恒玄科技	476.59	3.24	4.86	6.58	122.44	81.74	60.31
688332.SH	中科蓝讯	172.04	2.52	3.32	4.14	56.80	43.13	34.50
688018.SH	乐鑫科技	303.21	3.09	4.07	5.32	87.43	66.33	50.81
603893.SH	瑞芯微	689.09	1.22	1.80	2.48	134.37	91.57	66.25
300458.SZ	全志科技	298.58	0.38	0.57	0.78	124.83	82.90	60.81
688220.SH	翱捷科技	290.80	-1.34	-0.70	0.37	-51.85	-98.92	190.05
688286.SH	敏芯股份	39.40	-0.54	0.53	1.37	-130.55	133.09	51.24
603986.SH	兆易创新	846.01	1.68	2.47	3.17	75.80	51.48	40.20
688766.SH	普冉股份	114.36	2.66	3.40	4.16	40.76	31.89	26.02
688256.SH	寒武纪	2388	-1.10	-0.06	1.15	-520.7	8893.3	499.6
688041.SH	海光信息	2975	0.82	1.21	1.64	155.3	106.1	78.2

688798.SH	艾为电子	168.01	0.92	1.65	2.42	78.42	43.68	29.81

资料来源：Wind 一致预期、中国银河证券研究院

（二）通信板块：运营商、光模块及光芯片子板块动能强劲

运营商：低估值高成长，国家政策支持下新业务发展有望超预期。

运营商盈利能力、现金流资产不断改善、资产价值优势凸显，持续增加分红回馈股东，相对历史估值和国外水平，通信运营商均处于估值低位。总体来说，运营商业绩持续增长或超预期，5G“收获期”大有可为。当前运营商云业务发展如火如荼，DeepSeek 对于成本端的降低有望协同运营商云业务部署以及运营商的海量数据资产，推动运营商第二曲线的快速增长。

建议关注：中国移动、中国联通、中国电信等。

光通信：技术迭代产品量价齐升，技术壁垒增强竞争格局有望边际改善。

AIGC 引领新一轮科技革命，DeepSeek 对于成本端的降低或将推动应用端的繁荣，继而反哺推理侧模型的快速迭代，推动应用端的进一步发展。光模块 100G/200G→400G→800G→1/6T 迭代速率持续提升，带来产品量价齐升有望延续，带来业绩高增持续可期。同时国内的算力部署有望推动国内光模块产业链景气度提升，带来较强的规模效应。

建议关注：中际旭创、新易盛、天孚通信、光迅科技、华工科技等。

光芯片产业链国产化进程加速，复杂国际经济形势下渗透率有望进一步提升。

DeepSeek 对于成本端及训练精度的降低或将使得推理侧对光芯片的技术需求产生一定放松，国产光芯片在推理侧算力部署中具备成本优势且技术可靠性较强，产业链渗透率有望跟随推理侧算力部署的增加而有所上升，同时在复杂国际形势下，海外芯片采购难度预计将提升，国产光芯片在推理侧部署的可靠性及成本优势将进一步提升自身在采购链条中的话语权。

建议关注：源杰科技、仕佳光子等。

表9：建议关注相关标的盈利预测情况-通信（截至 2025 年 1 月 31 日）

代码	标的名称	总市值（亿元）	EPS（元）			P/E		
			2024E	2025E	2026E	2024E	2025E	2026E
600941.SH	中国移动	23,877.77	6.48	6.86	7.25	17.13	16.18	15.31
600050.SH	中国联通	1,555.05	0.29	0.32	0.35	17.09	15.45	14.03
601728.SH	中国电信	6,460.40	0.36	0.39	0.41	19.66	18.25	17.03
300308.SZ	中际旭创	1,287.32	4.95	8.11	9.34	23.20	14.16	12.29
300502.SZ	新易盛	891.47	3.50	6.45	8.55	35.93	19.50	14.71
300394.SZ	天孚通信	558.40	2.58	4.16	5.46	39.08	24.22	18.48
002281.SZ	光迅科技	384.18	0.98	1.37	1.72	49.43	35.30	28.14
000988.SZ	华工科技	395.16	1.31	1.68	2.12	29.91	23.37	18.51
688498.SH	源杰科技	124.44	0.31	1.27	2.35	469.71	114.65	61.96
688313.SH	仕佳光子	88.92	0.12	0.24	0.37	167.79	80.62	52.25

资料来源：Wind 一致预期、中国银河证券研究院

（三）计算机板块：看好算力向推理，基础设施向应用侧投资变化机遇

DeepSeek 的爆火和推广，有望加速全球 AI 产业链的发展。DeepSeek 的技术突破将显著降低高质量 AI 模型的训练成本，有望加速 AI 技术的普及和应用。成本的降低使得企业可以将更多资源投入到 AI 技术研发和应用开发中，加速 AI 技术的创新和迭代以及 AI 市场的繁荣。短期而言，DeepSeek 的技术创新可能对以 GPU 为代表的高端算力的芯片企业产生压力，市场预计会从单纯追求高端算力马太效应企业，转向更加注重技术创新和成本控制的企业和项目，但中长期来看，“杰文斯”悖论再次到来，技术进步反而推动资源使用总量上升，DeepSeek 将推动算力需求总量提升。当模型的成本越低，开源模型发展越好，模型的部署、使用就会更高频率、更多数量，对算力的需求将越来越大，我们认为，当下投资中的结构性机会主要体现在“从训练算力为主到推理算力为主过渡”、“从高端 GPU 到 ASIC 芯片过渡”，以及“从基础设施投资机会向应用侧投资机会过渡”。

1、从以训练算力为主到以推理算力为主过渡

DeepSeek 的技术创新显著降低了模型训练成本，同时提升了推理效率。这种变化推动了 AI 产业从以训练算力为主向以推理算力为主过渡。随着推理需求的增长，ASIC（应用特定集成电路）和 LPU（语言处理单元）等专用芯片将逐渐取代部分 GPU 市场份额。此外，推理算力的增长将推动边缘计算设备的需求，边缘侧设备能够直接在本地运行轻量化大模型，减少对云端的依赖，降低延迟和带宽成本，边缘算力机会逐渐凸显。建议关注：宝信软件、润泽科技、海光信息、中科曙光、网宿科技等。

2、从高端 GPU 到 ASIC 芯片过渡

DeepSeek 的低成本、高性能模型展示了 ASIC 芯片在特定任务中的优势。ASIC 芯片通过定制化设计，能够实现更高的能效比和更低的推理延迟。例如，DeepSeek 的 LPU+R1 模型在运行 7B 蒸馏模型时，推理延迟仅为 50ms，功耗约 30W，而英伟达 A100GPU 运行 175BGPT-3 模型时，推理延迟约 350ms，功耗约 300W，ASIC 芯片的崛起将为相关 ASIC 制造商以及 AIOT 端侧只能硬件带来新的增长机遇。

3、从基础设施投资机会向应用侧投资机会过渡

DeepSeek 的技术创新不仅降低了模型训练成本，还推动了 AI 技术在更多领域的应用。其开源策略和低成本模型使得更多企业和开发者能够使用先进的 AI 技术，加速了 AI 技术在各行业的应用和发展，相应的投资机会：

1）AI 应用开发：随着 AI 技术的普及，应用开发将成为新的投资热点。投资者可关注在教育、医疗、金融、办公等领域有技术积累和市场优势的 AI 应用开发公司，建议关注科大讯飞、大华股份、海康威视、同花顺、恒生电子、金山办公、财富趋势、嘉和美康、彩讯股份等。

2）数据服务与处理：高质量的数据集是训练高效 AI 模型的基础，数据的获取、处理和应用将成为 AI 应用的关键。投资者可关注数据采集、存储、处理和分析等环节的技术和服务提供商，如深桑达、拓尔思、达梦数据、上海钢联、英方软件等。

3）端侧 AI 设备：端侧 AI 技术的快速发展将推动终端设备的智能化升级，如智能家居、智能眼镜、车载系统等，建议关注虹软科技、萤石网络、道通科技、东软集团、中科创达等。

表10：建议关注相关标的盈利预测情况-计算机（截至 2025 年 1 月 31 日）

代码	标的名称	总市值（亿元）	EPS（元）			P/E		
			2024E	2025E	2026E	2024E	2025E	2026E
000032.SZ	深桑达 A	185.49	0.4	0.51	0.63	41.83	32.80	26.56
002230.SZ	科大讯飞	1171.59	0.26	0.42	0.58	191.62	118.62	85.90
002236.SZ	大华股份	499.29	1.04	1.23	1.45	14.89	12.59	10.68
002415.SZ	海康威视	2680.40	1.54	1.78	2.05	18.84	16.30	14.16
300033.SZ	同花顺	1500.60	2.63	3.2	3.6	115.21	94.69	84.17
300226.SZ	上海钢联	77.39	0.69	0.9	1.1	34.28	26.28	21.50
300229.SZ	拓尔思	193.16	0.22	0.29	0.36	85.32	64.72	52.14
300496.SZ	中科创达	285.85	0.8	1.18	1.62	81.15	55.02	40.07
300634.SZ	彩讯股份	133.29	0.67	0.81	0.98	38.12	31.53	26.06
600570.SH	恒生电子	491.34	0.79	0.92	1.09	33.53	28.79	24.30
600718.SH	东软集团	114.11	0.25	0.36	0.48	38.36	26.64	19.98
600845.SH	宝信软件	832.84	1.03	1.28	1.6	28.63	23.04	18.43
603019.SH	中科曙光	980.20	1.48	1.82	2.19	44.93	36.54	30.37
688041.SH	海光信息	2975.15	0.82	1.21	1.64	159.76	108.26	79.88
688088.SH	虹软科技	197.78	0.39	0.56	0.74	129.49	90.18	68.24
688111.SH	金山办公	1452.33	3.3	4.13	5.22	92.43	73.86	58.43
688208.SH	道通科技	176.23	1.35	1.63	2.02	30.52	25.28	20.40
688246.SH	嘉和美康	32.92	0.58	0.89	1.26	42.60	27.76	19.61
688318.SH	财富趋势	278.27	1.67	1.86	2.1	96.99	87.09	77.13
300017.SZ	网宿科技	254.44	0.26	0.3	0.34	43.73	37.90	33.44
300442.SZ	润泽科技	1026.35	1.29	1.87	2.36	46.68	32.20	25.52
688435.SH	英方软件	24.28	0.74	1.04	1.53	40.54	28.85	19.61
688475.SH	萤石网络	273.89	0.75	0.94	1.15	46.00	36.70	30.00
872808.BJ	曙光数创	124.54	0.51	0.68	0.89	125.71	94.28	72.03

资料来源：Wind 一致预期、中国银河证券研究院

（四）传媒板块：DeepSeek 加速 AI 向低成本发展，看好 AI+应用落地

DeepSeek-R1 通过强化学习实现了低成本与高性能的结合，其发布并开源为行业生态发展带来了新的可能性，并且有望加速推动在影视、广告、社交陪伴等多个领域应用落地。建议关注与 C 端用户体验密切相关的行业：1）AI+游戏：游戏内 NPC 互动、互动影游；2）AI+影视：赋能生产全环节；3）AI+营销：精准推送，重塑生态；4）AI+专业咨询：情感陪护，应用场景专业解答等。公司层面，建议关注头部互联网大厂：腾讯控股、阿里巴巴-W 等。

表11：建议关注相关标的盈利预测情况-传媒（截至 2025 年 1 月 31 日）

代码	标的名称	总市值（亿元）	EPS（元）			P/E		
			2024E	2025E	2026E	2024E	2025E	2026E
0700.HK	腾讯控股	34,070.69	19.92	22.39	24.99	18.46	16.43	14.72
9898.HK	阿里巴巴-W	15,468.28	3.91	6.19	6.88	16.27	13.13	11.82

资料来源：Wind 一致预期、中国银河证券研究院

七、风险提示

1. 国际经济形势复杂度进一步提升的风险：国内外政策和技术摩擦不确定性，经贸关系面临诸多挑战，直接影响各国贸易往来以及全球经济稳定发展。

2. AI 产业链上下游短期波动的风险：若未来上游原材料的价格受到宏观经济、贸易摩擦等因素的影响而产生大幅波动，将会对科技行业上市公司的经营业绩造成不利影响。

3. AI 硬件发展速度不及预期的风险：据技术发展的趋势和下游客户的需求，不断升级更新现有产品，并研发新技术和新产品，从而保持技术的先进性和产品的竞争力。如果产品研发进度未达预期或无法在市场竞争中占据优势，AI 硬件或将面临新产品研发失败的风险，前期的研发投入也将无法收回。

4. AI 应用发展不及预期的风险：目前 AI 仍处于早期快速发展迭代阶段，技术研发进展存在不确定性，在应用端落地速度存在不及预期的风险。

图表目录

分析师承诺及简介

本人承诺以勤勉的执业态度，独立、客观地出具本报告，本报告清晰准确地反映本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告的具体推荐或观点直接或间接相关。

首席计算机分析师吴砚靖；首席通信分析师赵良毕；首席电子分析师高峰；传媒分析师岳铮。

计算机行业分析师：邹文倩、李璐昕；通信行业分析师：赵中兴。

免责声明

评级标准

评级标准		评级	说明
评级标准为报告发布日后的 6 到 12 个月行业指数（或公司股价）相对市场表现，其中：A 股市场以沪深 300 指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准，北交所市场以北证 50 指数为基准，香港市场以恒生指数为基准。	行业评级	推荐：	相对基准指数涨幅 10%以上
		中性：	相对基准指数涨幅在-5%～10%之间
		回避：	相对基准指数跌幅 5%以上
	公司评级	推荐：	相对基准指数涨幅 20%以上
		谨慎推荐：	相对基准指数涨幅在 5%～20%之间
		中性：	相对基准指数涨幅在-5%～5%之间
		回避：	相对基准指数跌幅 5%以上

联系

中国银河证券股份有限公司 研究院

深圳市福田区金田路 3088 号中洲大厦 20 层

上海浦东新区富城路 99 号震旦大厦 31 层

北京市丰台区西营街 8 号院 1 号楼青海金融大厦

公司网址：www.chinastock.com.cn

机构请致电：

深广地区：程　曦　0755-83471683　chengxi_yj@chinastock.com.cn

苏一耘　0755-83479312　suyiyun_yj@chinastock.com.cn

上海地区：陆韵如　021-60387901　luyunru_yj@chinastock.com.cn

李洋洋　021-20252671　liyangyang_yj@chinastock.com.cn

北京地区：田　薇　010-80927721　tianwei@chinastock.com.cn

褚　颖　010-80927755　chuying_yj@chinastock.com.cn