DeepSeek：
重塑全球AI格局的中国力量

国盛计算机　　2025.5

国盛证券
GUOSHENG SECURITIES

目录

关注技术公众号

DeepSeek震撼科技圈：一月时间差肩最强推理模型o1

2024.5 DeepSeek-V2 发布
提出MLA和DeepSeekMoE架构相比第一代DeepSeek 67B实现了更强的性能，节省了42.5%的训练成本，减少了93.3%的KV缓存

2024.11 推理模型 DeepSeek-R1.Lite 预览版 发布
媲美 01-preview的推理效果并为用户展现了 o1 没有公开的思考过程

2024.12 DeepSeek-V3发布
DeepSeek-V3 671B**在 2048 块 NVIDIA H800 集群上训练2个月，训练成本仅558万美元，**达到GPT-4o和Claude Sonnet 3.5水准

2025.1.20 DeepSeek-R1发布
从数学（AIME/MATH）、编程(Codeforces/SWE)、学科推理(GPQA)的各个高难度benchmark结果来看，DeepSeek-R1推理能力比肩OpenAI-o1-1217版本。同时DeepSeek-R1蒸馏得到的Qwen和llama小模型也与OpenAI-o1-mini相当的效果。

2025.1.28 Janus-Pro、JanusFlow 发布
Janus-Pro是一款统一多模态理解与生成的创新框架，解耦视觉编码。JanusFlow 是一款通过生成流与自回归语言模型融合实现统一的框架，能生成高质量图像。

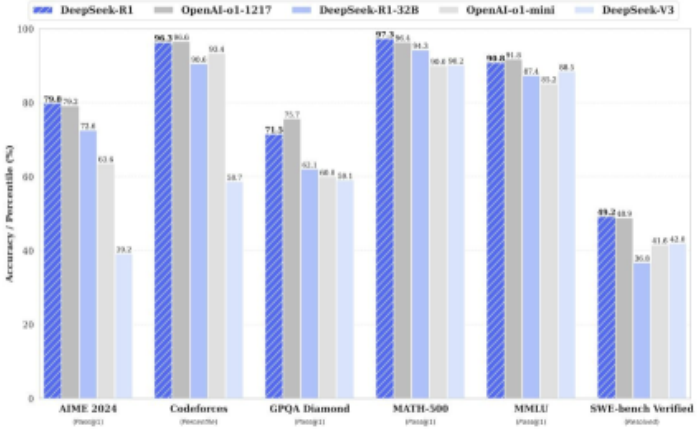


Figure 1 | Benchmark performance of DeepSeek-R1.

数学、代码能力追平o1

	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

蒸馏小模型效果超越o1-mini

DeepSeek震撼科技圈：开源、低成本倒逼OpenAI加速发布

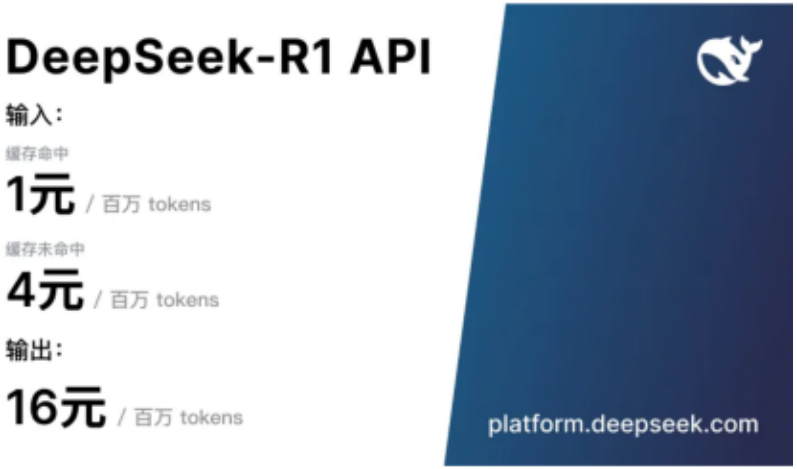
开放的许可证和用户协议：DeepSeek-R1 统一采用标准化、宽松的 MIT License，完全开源，不限制商用，无需申请。产品协议明确可“模型蒸馏”。

OpenAI加速发布：

2月1日，OpenAI推出o3-mini, Plus和Team用户的速率限制从原来o1-mini的每天50条消息增加3倍到o3-mini的每天150条消息。

2月3日，OpenAI发布了基于OpenAI的o3模型之上开发而成的Deep Research。能够像人类分析师一样，对复杂的任务进行逐步分解，并在互联网上进行多轮的信息搜索与验证，直到找到最合适的答案

2月13日，OpenAI宣布将在未来几个月内推出GPT-5, 该模型将整合OpenAI的大量技术，包括o3，在GPT-5推出之前，OpenAI计划在未来几周内先发布GPT-4.5代号“Orion”，这将是OpenAI最后一个“非思维链模型



Model	Pricing	Pricing with Batch API***
o1	$15.00 / 1M input tokens	$7.50 / 1M input tokens
	$7.50 / 1M cached* input tokens	
	$60.00 / 1M output** tokens	$30.00 / 1M output** tokens

Model	Pricing	Pricing with Batch API***
o3-mini	$1.10 / 1M input tokens	$0.55 / 1M input tokens
	$0.55 / 1M cached* input tokens	
	$4.40 / 1M output** tokens	$2.20 / 1M output** tokens

DeepSeek震撼科技圈：算法优化彰显创新能力

美国总统特朗普在佛罗里达表示：“中国公司发布 DeepSeek AI 应该给我们的行业敲响警钟，我们需要集中精力进行竞争。”

DeepSeek提出大量算法创新，中国从AI跟随者走到前沿探索贡献者

DeepSeek-R1：
DeepSeek-R1-Zero 提出不用监督微调直接进行强化学习，也能取得不错的效果。

DeepSeek-R1 加入少量CoT数据进行监督微调作为冷启动，然后再进行多阶段强化学习，可以取得更优的性能，同时回答更符合人类偏好。强化学习不需要进行过程监督和MCTS搜索，直接进行基于规则的奖励

DeepSeek-V3：
注意力层状态压缩(Multi-Head Latent Attention）：对attention层隐向量降维, 减少推理时显存，提高推理效率

细粒度稀疏MoE架构： 671B总参数，平均激活参数37B

多词元前瞻性预测(Multi-Token Prediction）：丰富训练监督信号，加速推理

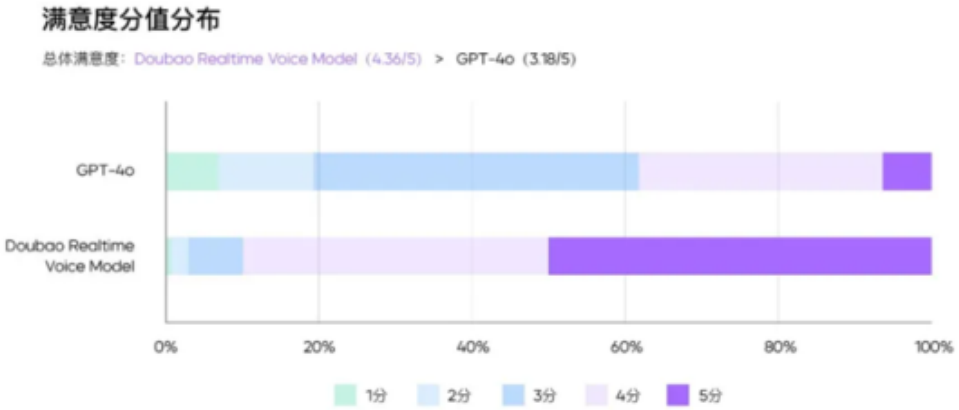
混合浮点精度运算(FP8 Traiming）：对训练算子进行细粒度拆分，降低精度损失，首次在超大模型训练中验证FP8的有效性

PTX层优化：在比Cuda底层的编程语言上优化硬件效率

大厂模型持续进步：字节

字节豆包实时语音大模型情绪理解与表达能力突出。

1月20日豆包实时语音大模型在豆包 APP 全量开放，在情绪理解和情感表达方面与GPT-4O相比优势明显。豆包团队围绕拟人度、有用性、情商、通话稳定性、对话流畅度等多个维度进行考评。整体满意度（以 5 分为满分）方面，豆包实时语音大模型评分为 4.36，GPT-4o 为 3.18。其中，50% 的测试者对豆包实时语音大模型表现打出满分。在模型优点评测中豆包实时语音大模型在情绪理解和情感表达方面与GPT-4O相比优势明显。尤其是“一听就是 AI 与否”评测中，超过 30% 的反馈表示 GPT-4o “过于 AI ”，而豆包实时语音大模型相应比例仅为 2% 以内。



豆包团队评测语音大模型满意度超过GPT-4o

2025年1月，豆包MAU达到7816万，月增速10.47%

大厂模型持续进步：阿里

阿里通义千问登顶非推理国产模型冠军。
Qwen2.5-Max 于1月29日发布，2月4日凌晨，Chatbot Arena 公布了最新的大模型盲测榜单，通义千问Qwen2.5-Max 凭借1332分的成绩，位列全球第七，并成为非推理类中国大模型的冠军。同时，Qwen2.5-Max 在数学和编程等单项能力上排名第一

“通义”比肩美国Meta的LlaMA，影响力稳居全球开源模型的第一阵营。据阿里研究院院长袁媛介绍，Hugging Face社区上，目前全球开发者基于阿里自研“通义”开源模型二次开发的衍生模型已经突破8万个。

4月29日阿里正式发布Qwen3并全部开源8款混合推理模型，亮点包括多种思考模式&多语言&Agent能力强化。
1）思考模式下，模型会逐步推理，经过深思熟虑后给出最终答案；非思考模式下，模型提供快速、近乎即时的响应。2）Qwen3 模型支持 119 种语言和方言。3）优化了 Qwen3 模型的 Agent 和 代码能力，同时也加强了对 MCP 的支持。

5 月 6 日，彭博社记者马克·古尔曼表示，iOS 18.6 预计将在中国大陆启用部分 Apple Intelligence 功能，将由阿里云和百度提供技术支持。苹果公司将与阿里巴巴合作开发审查引擎，另外还将把百度的软件与 Siri 和 Visual Intelligence 实现集成。

大厂模型持续进步:腾讯

腾讯:

2025年3月元宝同时接入DeepSeek和深度思考模型混元T1。混元 T1 采用mamba架构，能秒回、吐字快、擅长超长文处理的强推理模型，推理能力进一步提升，深度理解长文本内容，快速提炼要点，适合论文、报告、策划案等。

2025年4月微信上线元宝好友用户可在微信搜索“元宝”之后将其添加为好友，直接在微信聊天界面与其进行互动。账号介绍显示其搭载混元和DeepSeek双模引擎，无缝衔接微信生态。一键解析公众号文章和任何图片和文档，短评后奉上秒开详解，支持对解读内容做各种智能互动;支持陪伴互动，越聊越懂用户。

推出AI智能工作台ima

2024年11月15日腾讯推出AI智能工作台ima.copilot(简称ima)。ima搜索得出的答案，除开全网信源，还打通微信公众号文章的生态。整个公众号世界里的优质知识，都可为用户所用。能为用户提供好答案和高质量的问题相关信息，有效提升信息获取效率。除了能搜出答案，ima还有一个特点——边问边看，边搜边记，让用户可以轻松弄懂知识点。

混元 3D 生成大模型 2.0：2025 年 1 月 21 日，腾讯宣布开源 3D 生成大模型的 2.0 版本，并推出业界首个一站式 3D 内容 AI 创作平台 —— 混元 3D AI 创作引擎。该版本在生成质量上有显著提升，简化了创作流程，在几何和纹理解耦生成方面取得进步，生成效果更精细，几何结构更准确，纹理色彩更丰富。用户通过一句话、提示词或图片即可生成 3D 模型，还可实现角色塑造、动画形成等功能。

大厂模型持续进步：华为

2024年6月22日，华为云盘古大模型迎来了新的里程碑——盘古大模型5.0。这一版本的发布，不仅标志着盘古大模型在技术上的又一次飞跃，也预示着其在行业应用中的无限可能。盘古大模型5.0在全系列、多模态、强思维三个方面进行了全面升级。

全系列：盘古大模型5.0包含不同参数规格的模型，以适配不同的业务场景。十亿级参数的Pangu E系列可支撑手机、PC等端侧的智能应用；百亿级参数的Pangu P系列，适用于低时延、高效率的推理场景；千亿级参数的Pangu U系列适用于处理复杂任务；万亿级参数的Pangu S系列超级大模型能够帮助企业处理更为复杂的跨领域多任务。

多模态：盘古大模型5.0能够更好更精准地理解物理世界，包括文本、图片、视频、雷达、红外、遥感等更多模态。在图片和视频识别方面，可支持10K超高分辨率；在内容生成方面，采用业界首创的STCG（Spatio Temporal Controllable Generation，可控时空生成技术，聚焦自动驾驶、工业制造、建筑等多个行业场景，可生成更加符合物理规律的多模态内容。

强思维：复杂逻辑推理是大模型成为行业助手的关键。盘古大模型5.0将思维链技术与策略搜索深度结合，极大地提升了数学能力、复杂任务规划能力以及工具调用能力。

国内大模型核心创业公司：智谱AI

智谱AI

CEO：唐杰（曾任清华大学计算机系教授、计算机系副主任、清华-工程院知识智能联合实验室主任、杰青）

主打产品：

GLM系列大模型：包括中英双语千亿级超大规模预训练模型GLM-130B，以及新一代基座大模型GLM-4。

ChatGLM：基于GLM系列，具备多轮对话、创意写作等能力。

智本GPT：专为金融领域设计的智能化应用，特别是在资本管理方面。

最新进展：

2024.8

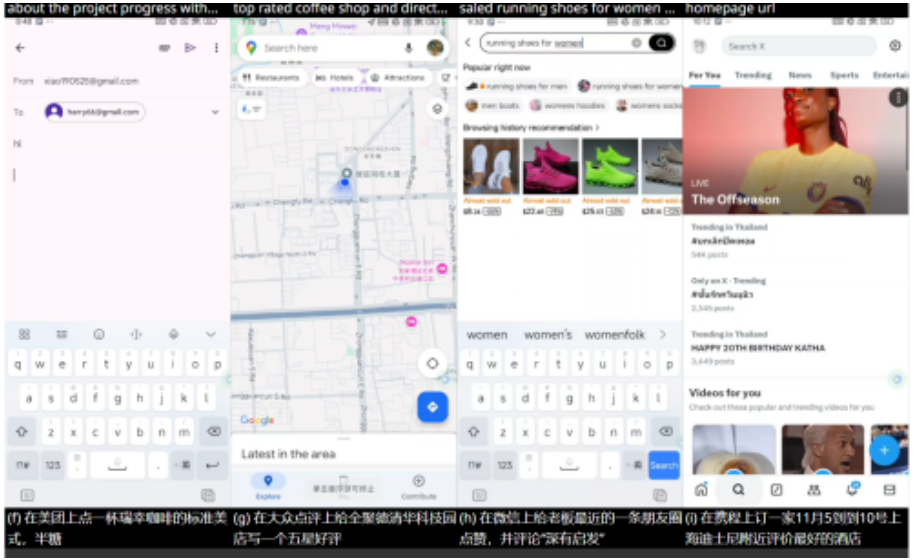
发布新一代基座大模型 GLM-4-Plus，推出国内首个AI视频通话功能。

2024.10

推出最新端到端情感语音模型 GLM-4-Voice和大模型交互智能体（Agent）AutoGLM。

2024.11

Agent 新升级：智谱清言插件上线 AutoGLM 功能；推出 GLM-PC。



智谱AutoGLM可以通过语音操作多种手机APP

国内大模型核心创业公司：MiniMax

MiniMax

创始人：闫俊杰（前商汤科技副总裁、商汤科技研究院副院长、通用智能技术负责人）

文生视频海螺AI。2024年8月31日，成立两年半的AI创业公司MiniMax发布了旗下首个文生视频模型 abab-video-1并在海螺AI中上线。根据谷歌统计的搜索数据，自9月推出文生视频能力以来海螺AI的搜索热度持续飙升。10月10日MiniMax宣布在过去的一个月内，海螺AI网页版访问量增速超800%，荣登AI产品榜（web）9月全球增速榜、国内增速榜双榜单 TOP 1。第三方流量统计网站SimilarWeb的数据也显示海螺AI在网页端的访问量持续飙升，近28天仅海外版海螺AI的日均访问量就达到28万次，平均访问时长16分钟，都超越了视频生成应用runway、pika和可灵，国内版海螺AI的日均访问量也达到15万次。识别准确+一致性强+画面质感佳，海螺文生视频模型综合能力测评排名第一。abab-video-1具有压缩率高、文本响应好、风格多样，支持原生高分辨率、高帧率视频等特点，媲美电影质感。在视频生成模型评测基准Vbench-long中MiniMax的模型综合能力排在第一。

商业化进展迅速，出海表现优秀。 MiniMax在国内大模型中商业化进展迅速，很可能能在比较短的时间内实现自负盈亏及盈利。海外财经媒体Financial Times报道，MiniMax近年的收入将达到7000万美元。MiniMax的出海产品Talkie的下载量为1700万次，紧随类似产品Character AI之后，在美国其下载量已超过Character AI。MiniMax的大模型还赋能了众多行业客户，例如为金山办公WPS提供大模型能力，为小红书提供生成式搜索等。公司的模型每天为全球用户提供超过30 亿次交互，日均处理3万亿文本令牌，生成2000万张图像和7万小时语音，是国内大模型日处理交互量最大的公司。

海螺AI视频demo, Prompt：小猫嘴里喷火，它面前的鱼被烤焦

MiniMax大模型客户案例

海外大模型头部玩家梳理

OpenAI

4月17日发布o3和o4-mini模型。s这两款模型实现了对 ChatGPT 全部工具的自主调用能力，包括网页搜索、Python 编程、图像分析与生成、文件解析等。它们能够根据任务需求自动判断何时以及如何使用这些工具，以生成更深入、结构更清晰的回答。模型还具备“图像思考”能力，能够理解和操作图像内容，如放大、旋转等，进一步提升了处理多模态任务的能力。

战略合作“星际之门”项目，探索AI基础设施全球布局，初始股权投资者预计包括软银、OpenAI、甲骨文和中东全球投资集团MGX。

谷歌

Gemini大模型升级：谷歌在2025年1月推出了支持百万级上下文窗口的Gemini 2.0 Flash Thinking，进一步扩展多模态处理能力。同时计划在5月的I/O开发者大会上发布Gemini系列新版本，并可能整合到Pixel 10系列手机及安卓16操作系统中

Anthropic

核心团队成员来自OpenAI，**Claude模型在编程方面能力突出**。Anthropic正探索通过企业级合作（如与软银的合资公司SB OpenAI Japan）推动AI技术的实际应用，尤其是在金融和自动化领域

Meta

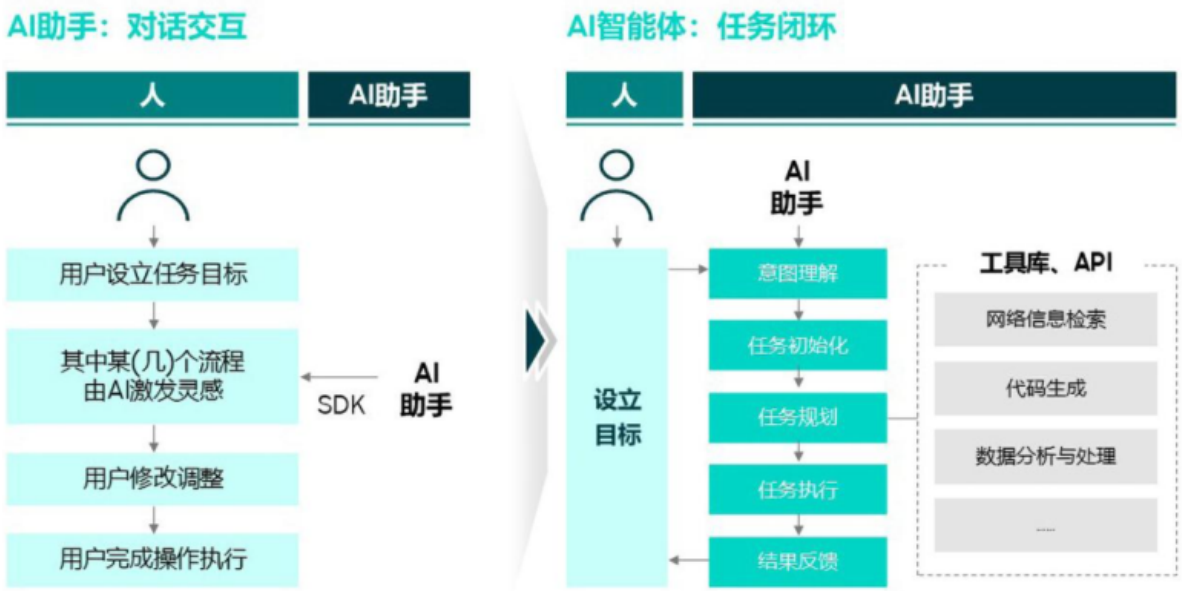
海外开源大模型领军。2024年7月发布了Llama 3.1系列，包括8B、70B和450B参数规模的模型。450B模型在多项基准测试中表现优异，甚至在某些任务上超越了OpenAI的GPT-4o和Anthropic的Claude 3.5 Sonnet。

目录

MCP影响力扩大，Agent生态构建加速

当前以大语言模型为核心驱动的Agent(智能体)技术正在快速发展中，展现出理解复杂指令、自主规划、调用工具并执行多步骤任务的强大能力。从简单的问答、文本生成，到复杂的市场分析、代码编写、差旅预订，Agent正逐渐渗透到我们工作和生活的方方面面。

从对话交互到任务闭环



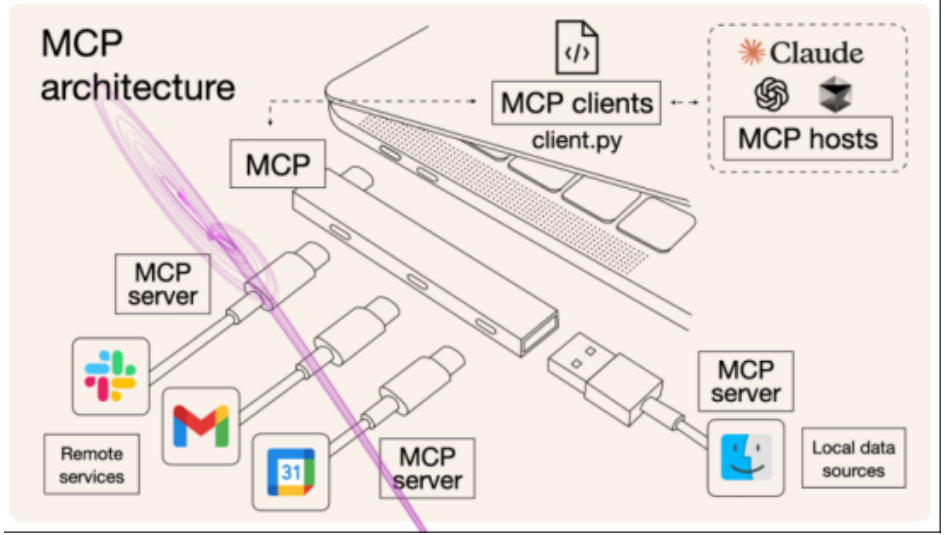
资料来源：罗兰贝格

Roland Berger

MCP影响力扩大，Agent生态构建加速

MCP(模型上下文协议)是一种开源协议，旨在标准化如何为大模型提供上下文。可以将MCP想象成 Agent的 USB-C接口:为大模型提供了一种连接到各种工具和数据源的统一方法。

传统上将AI统连接到外部工具涉及集成多个API。每个API 集成都意味着单独的代码、文档、身份验证方法、错误处理和维护。MCP旨在替换碎片化的Agent代码集成，从而使 AI 系统更可靠，更有效。通过建立通用标准，服务商可以基于协议来推出它们自己服务的 AI 能力，从而支持开发者更快的构建更强大的 AI 应用。开发者也不需要重复造轮子，通过开源项目可以建立强大的Agent生态。



MCP架构示意图

特性	MCP	传统 API
集成难度	单个标准化的集成	每个 API 的单独集成
实时通信	✅ Yes	❌ No
动态发现	✅ Yes	❌ No
可扩展性	插拔即用	需要额外的集成
安全性与控制	一致的工具	每个 API 的单独控制

MCP相比传统API的优点

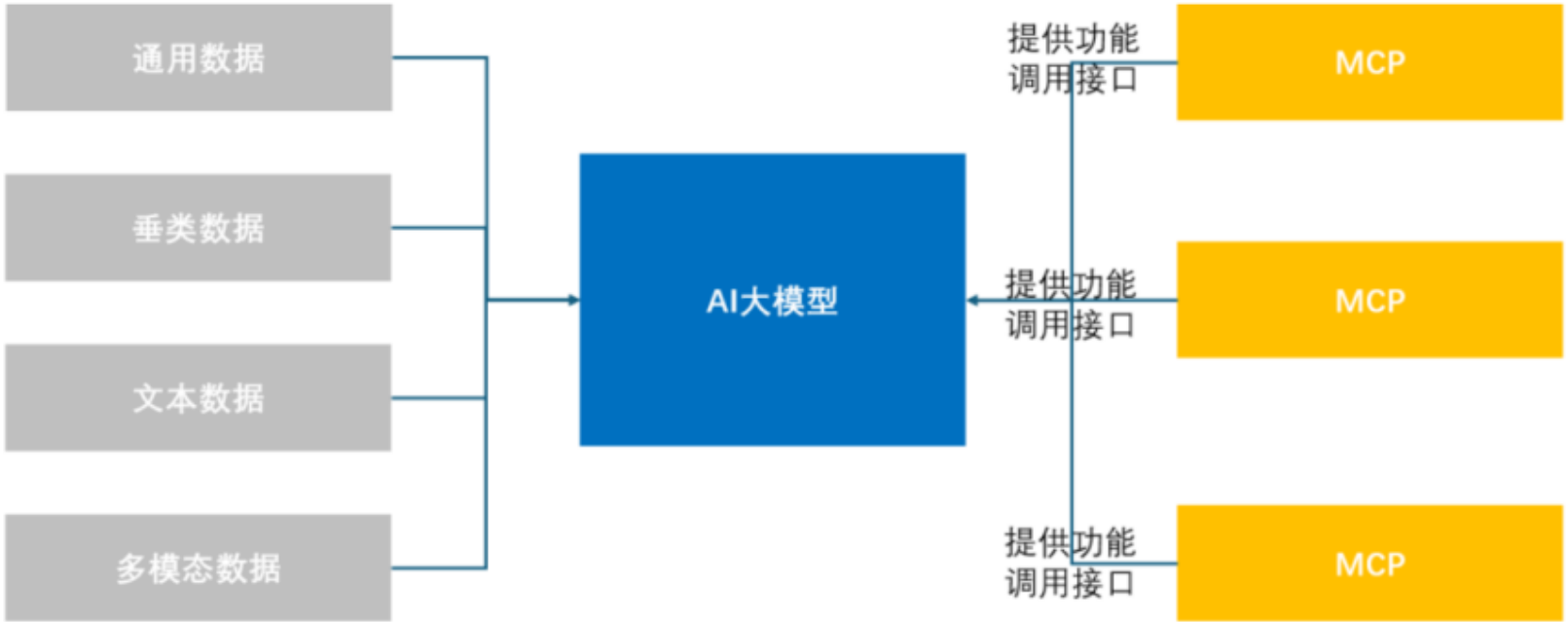
MCP影响力扩大，Agent生态构建加速

MCP最早由Anthropic开源，目前已有越来越多的公司和开发人员正在加入，成为未来 AI 工具交互的新标准

Anthropic	2024 年 11 月 25 日，率先开源 MCP，旨在使 AI 模（如 Claude）更容易与工具和数据源交互。
OpenAI	2025 年 3 月 27 日，OpenAI CEO Altman 在 X 宣布对 Agents SDK 进行了重大更新，支持了 Anthropic 推出的 MCP 服务协议。
谷歌	2025 年 4 月 10 日，Google DeepMind CEO Demis Hassabis 发帖称 MCP 是一个优秀的协议，正在迅速成为 AI 智能体时代的一个开放标准。同时宣布将为 Gemini 模型和 SDK 提供支持。
阿里	2025 年 4 月 9 日，阿里云百炼平台宣布上线业界首个全生命周期 MCP 服务，无需用户管理资源、开发部署、工程运维等工作，5 分钟即可快速搭建一个连接 MCP 服务的 Agent（智能体）。百炼平台首批上线了高德、无影、Fetch、Notion 等 50 多款阿里巴巴集团和三方 MCP 服务，覆盖生活信息、浏览器、信息处理、内容生成等领域，可满足不同场景 Agent 应用开发需求。
腾讯	2025 年 4 月 14 日，腾讯云宣布大模型知识引擎升级支持 MCP 协议，用户在搭建应用时，可以通过大模型知识引擎调用平台精选的 MCP 插件或插入自定义 MCP 插件。目前，知识引擎平台已经精选了多款 MCP Server，包括腾讯位置服务、腾讯云 EdgeOne Pages、Airbnb、Figma、Fetch、微信读书等，涵盖各类专业信息获取、网页部署和预览、网页解析获取等场景。

MCP影响力扩大，Agent生态构建加速

从AI应用底层体系来看，MCP为大模型带来了上游的数据资源，下游接入外部第三方能力的接口，分别增强AI的决策能力与执行能力，更容易孕育出好的Agent产品。



MCP驱动Agent体系示意图

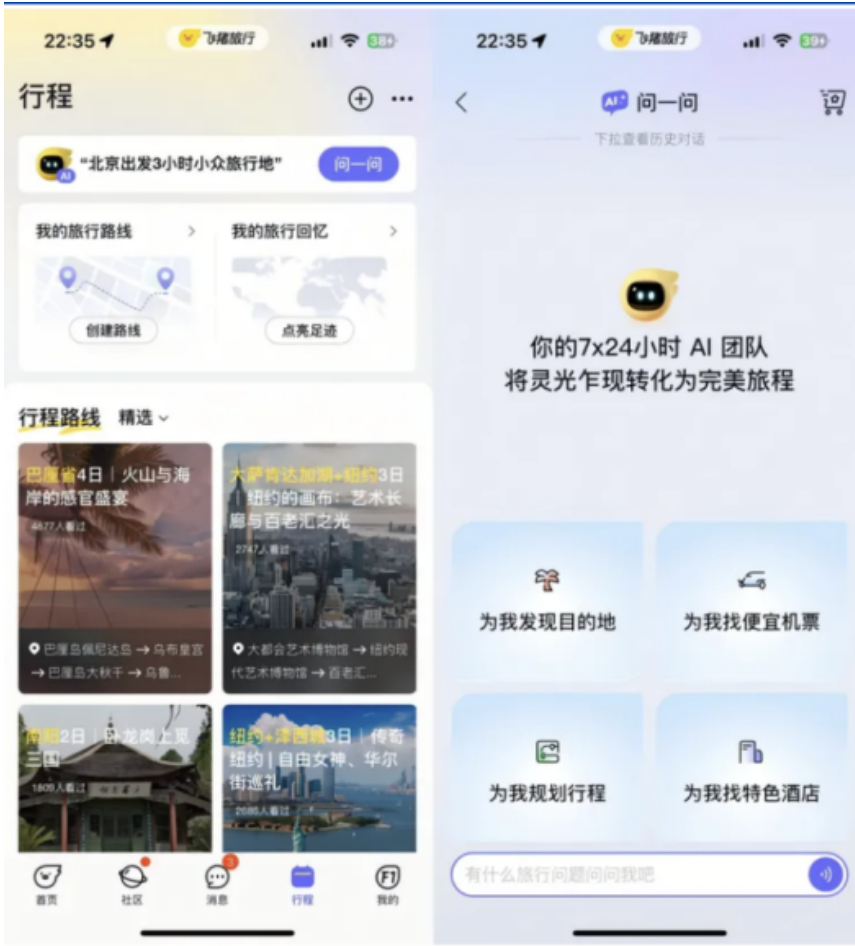
传统功能性APP走向Agent化，AI来全新功能升级

传统功能性APP走向agent化，AI有望带来全新功能升级。飞猪推出Agent“问一问”，旅游攻略智能化体验全新升级，对着手机喊一句“带娃游成都，预算5000”，10秒内「问一问」便召唤出机票比价师、酒店顾问、路线规划师等多位AI助手，从直飞航班、亲子酒店到“熊猫基地早起攻略”一键生成，费用明细精确到元，基于实时库存和真实用户评价推荐方案。而且方案中涉及到的机票、酒店等商品，用户可以直接一键下单预订。

飞猪Agent的关键要点：

1) 打通数据接口，实时连接最新数据。飞猪直接接入了自己的机票报价引擎，Agent在解析完消费需求后，会通过报价引擎从航司和全球机票分销系统获取信息，并打通了酒店、景区品类的供应链管理系统，确保机票、酒店价格和库存等信息秒级更新。

2) 多智能体协作，多维分工提升用户体验：「问一问」采用了多智能体协作机制打造核心决策层。系统内置了行程助手、交通顾问、酒店管家等多个专业AI角色，每个角色负责特定领域的专业判断，方案准确性和可用性得到了大幅度提高。



传统功能性APP走向Agent化，AI来全新功能升级

阿里钉钉AI助理：钉钉是阿里巴巴集团打造的企业级智能移动办公平台，现已推出钉钉AI助手，可以智能化帮助用户完成一系列办公任务，包括请假助手、智能日程、智能创作、消息总结、工作概览等。

伴随着MCP带来的功能聚合，通用Agent不断涌现

3月5日Monica团队发布通用Agent产品Manus，Manus 能通过独立思考和系统规划，在自己的虚拟环境中灵活调用各类工具，编写并执行代码、智能浏览网页、操作各类网页应用，直接交付完整任务成果

Manus能力全方位覆盖多领域任务：

旅行规划：不仅整合旅行信息，还为用户创建定制旅行手册。为用户规划日本四月旅行，提供个性化的旅行建议和详细手册。

股票分析：进行深入的股票分析，设计视觉上吸引人的仪表盘展示全面的股票洞察。例如，对特斯拉股票进行深度分析，创建可视化仪表盘。

教育内容创建：为中学教师创建视频演示材料，解释动量定理等复杂概念，帮助教师更有效地教学。

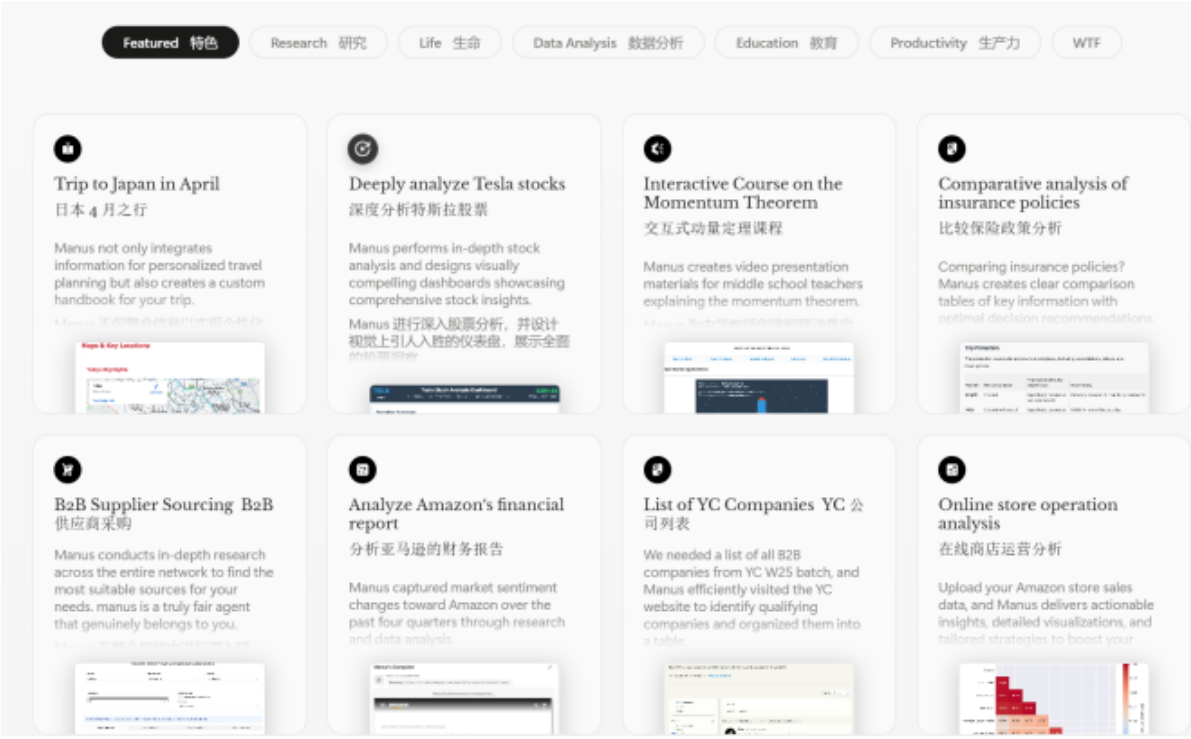
保险政策比较：创建清晰的保险政策比较表，提供最佳决策建议，帮助用户选择最适合的保险产品。

供应商采购：在整个网络中进行深入研究，找到最适合用户需求的供应商，作为真正公平的代理为用户服务。

财务报告分析：通过研究和数据分析捕捉市场对特定公司（如亚马逊）的情绪变化，提供过去四个季度的市场情绪分析。

创业公司列表整理：访问相关网站识别符合条件的公司，并将其整理成表格。

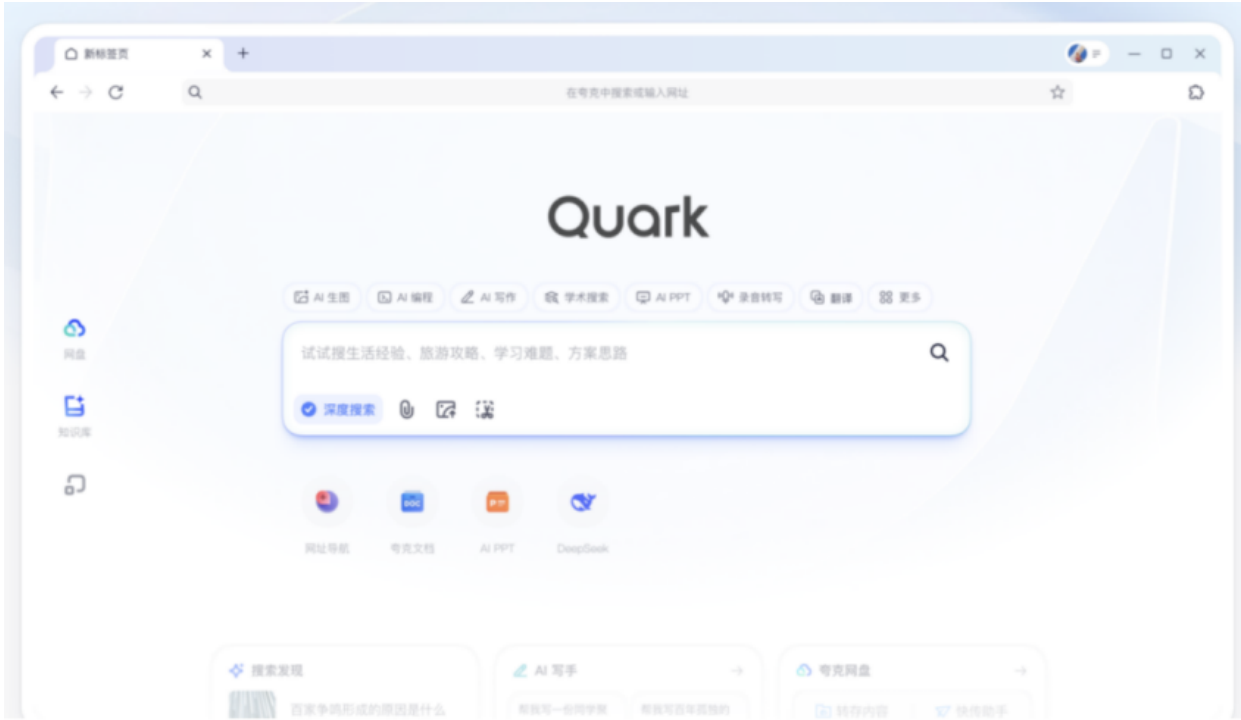
在线商店运营分析：分析亚马逊商店销售数据，提供可操作的洞察、详细可视化和定制策略，帮助提升销售业绩。



伴随着MCP带来的功能聚合，通用Agent不断涌现

夸克：以浏览器起家，现已成为了聚合AI搜索/AI生图/AI写作/AI PPT等多种功能的流量入口。

夸克最早为浏览器软件，然而伴随着AI技术的加入，如今已经成为多功能聚集体，其范围远超一个单纯的浏览器。从夸克首页界面可以清晰看出，从首页就可以直接进入AI生图/AI编程/AI写作/学术搜索/AI PPT/录音转写/翻译等一系列功能，已经成为了一个巨大的AI多agent聚合体，可以通过其功能多样性成为流量入口。



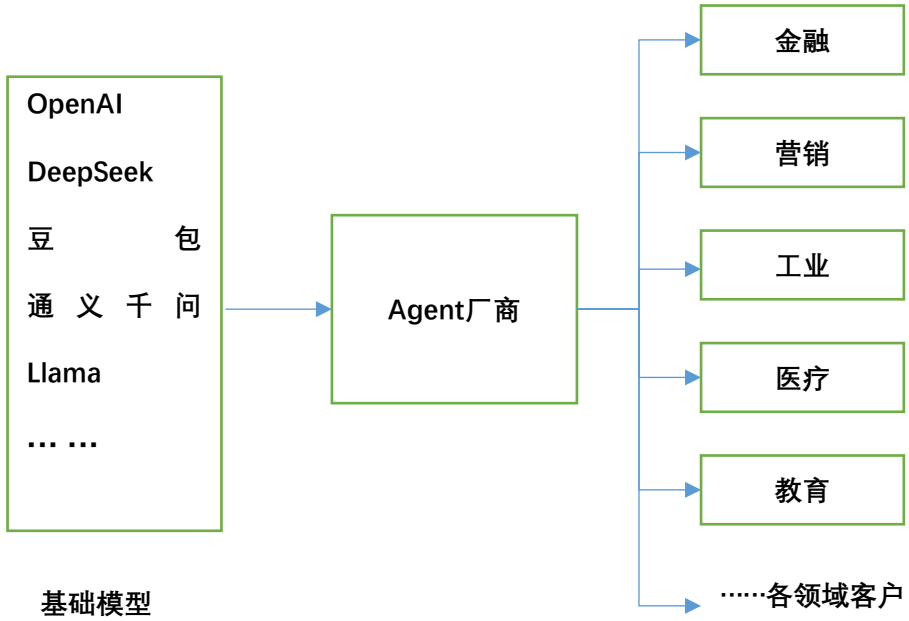
伴随着MCP带来的功能聚合，通用Agent不断涌现

百度心响：心响是百度发布的通用AI智能体，通过自然语言交互帮助用户实现复杂任务拆解、动态执行与可视化结果交付。

除了常见的外部MCP工具调用（Tool Use），在健康、法律等专业场景中，它还实现了「多智能体协作」（Agent Use）机制。比如，面对健康咨询时，系统可自动调度多位“医生AI分身联合会诊”；在法律服务中，则支持由多个律师AI分身组成的“律师智囊团”协同答复与服务

ToB垂直领域Agent落地需要连接大模型与具体行业的软件服务商

在AI Agent落地各ToB行业的过程中，连接大模型与具体行业的软件服务商是必不可少的环节。目前以互联网大厂和独角兽创业公司为主的大模型厂商不一定在各领域具备深耕的行业know-how，另外大模型厂商的研发人员通常在软件行业内属于较高端人才，人力成本相对较高，由其他服务商去对接具体的行业客户是更高的选择。我们认为在垂类深耕的公司以及具备长期软件服务经验的公司有望把握Agent落地机遇，在对接大模型和具体行业的过程中深度受益。



Agent内在技术对算力天然存在高需求

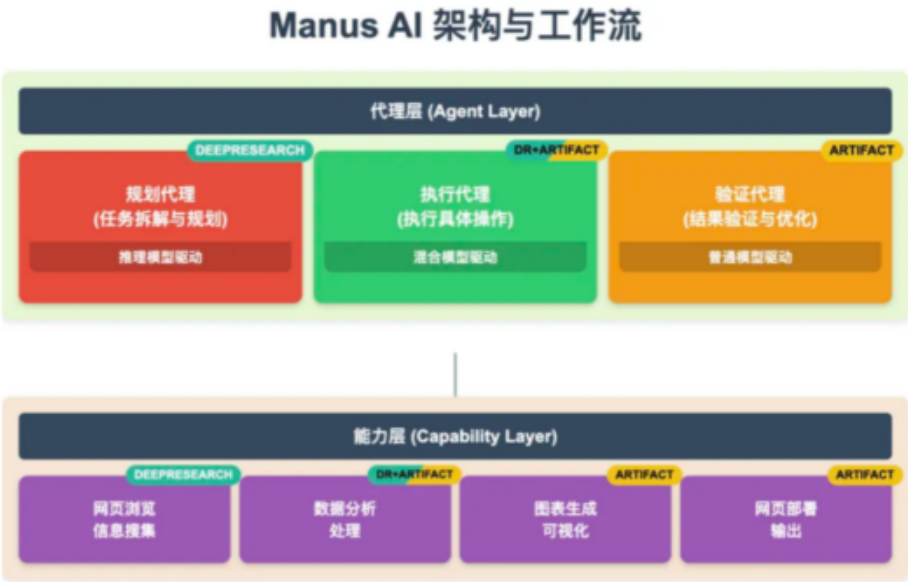
Agent 技术展现出强大的智能交互与任务处理能力，其背后也蕴含庞大的算力需求。

1）Agent接入外部数据以及多次调用模型带来的长上下文。Agent工作时需要维持庞大的上下文信息，包括用户输入、系统提示、通过检索增强生成从外部数据源获取的补充信息。即使是看似简单的单轮对话，背后也可能涉及复杂的上下文处理和推理过程，对于需要多步规划、工具调用或持续交互的任务，模型推理次数会更多。

2) Agent执行任务验证带来算力开销，为确保任务执行的准确性、可靠性与合规性，如Manus技术架构中包含一个专门的验证模块，通过三重校验体系保障输出可靠性：逻辑验证器检测任务链的因果合理性；事实核查器交叉比对多信源数据真实性；合规审查器确保输出符合法律法规

3)多模态发展趋势下，Agent需处理整合文本、图像、音频等多种类型数据，且多模态交互往往要满足实时性交互需求。

4）模型训练阶段对更大规模数据和模型参数量的需求。4月29日阿里开源的Qwen3系列模型在预训练使用的数据量是Qwen2.5 两倍，达到了约 36 万亿个 token。

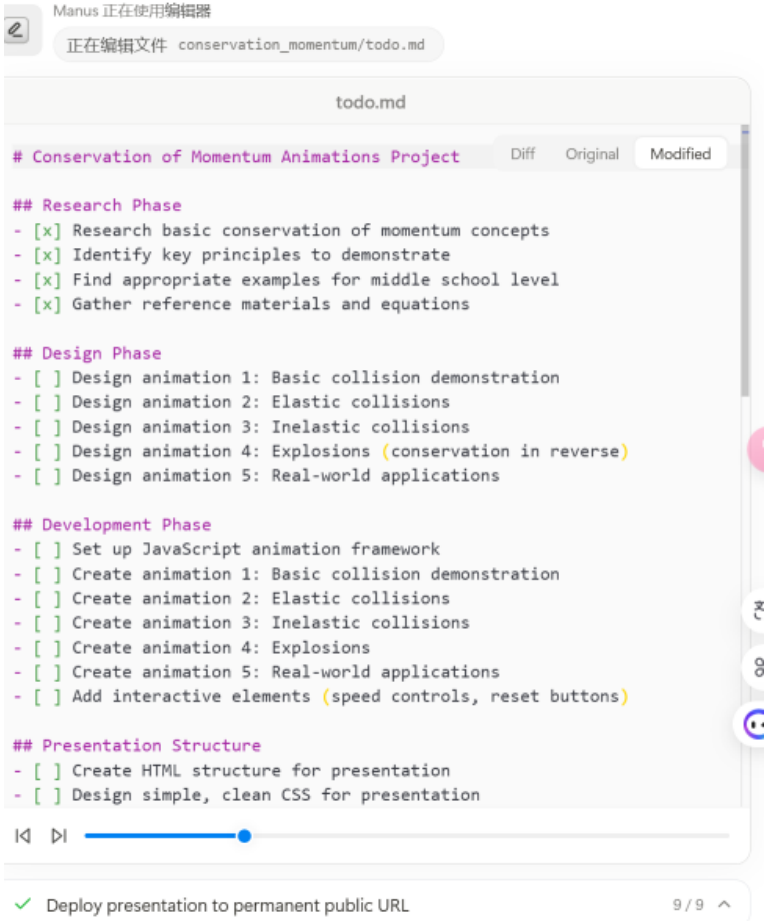


Manus技术架构包括规划、执行和验证三个代理

算力瓶颈与Agent服务的用户体验紧密相关

在Agent用户量激增、模型复杂度提升、应用场景多样化的背景下，算力瓶颈问题日益凸显，具体表现为服务响应延迟、服务不稳定甚至服务中断等情况，导致用户体验受损，虽然可以通过优化API调用方式（如批量请求、异步请求）等方法缓解，但根本原因在于瞬时或持续的算力需求超出了服务提供商的承载能力。

Manus回答问题一般耗时15分钟。据新京报贝壳财经记者测试发现，根据任务难度的不同，Manus执行任务的时间也不同，如对“设计采访提纲与视频采访脚本方案”等几项文字类任务，Manus的执行时间约为15分钟至20分钟，而对于“设计金融科普互动产品”这项涉及网页交互的任务，Manus耗时31分钟。



Manus完成任务时创建的规划步骤

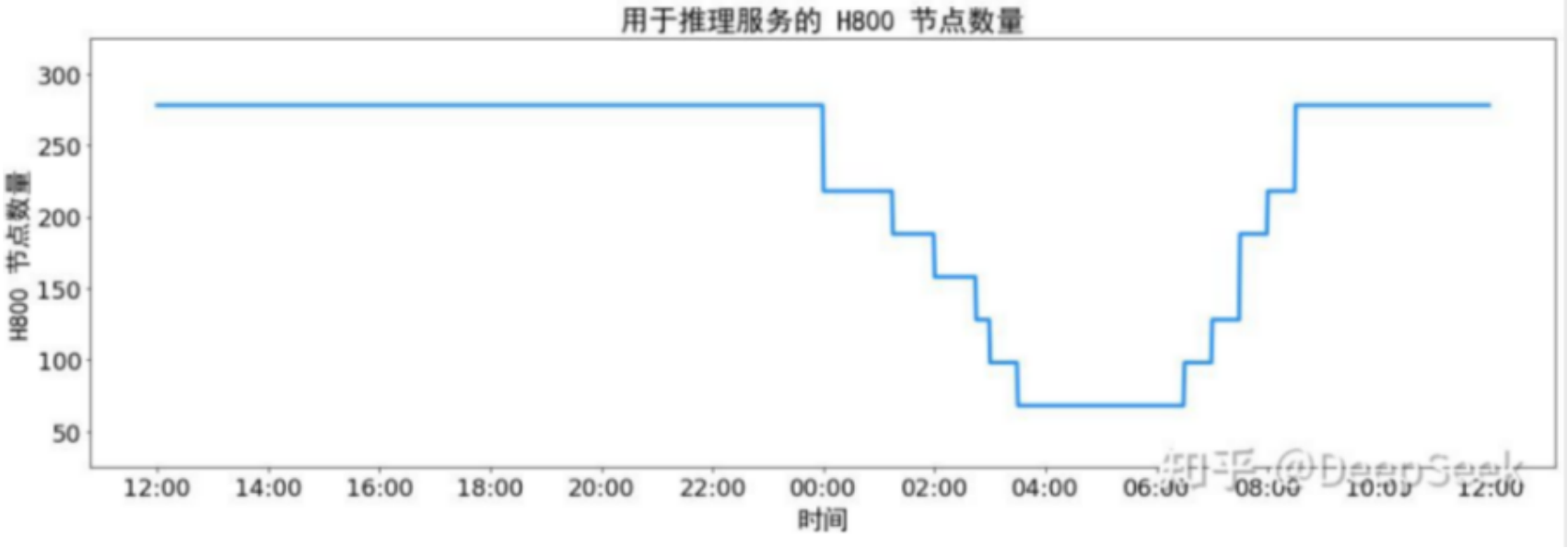
算力瓶颈与Agent服务的用户体验紧密相关

扣子空间是字节推出的智能体协作系统，据量子位4月23日报道，由于放出来的demo效果惊艳，出现了挤爆服务器的场面。扣子空间有探索和规划两种模式，探索模式让AI 自主动态思考，完成速度更快，规划模式由AI帮助规划步骤，分步执行。

据极客公园测试，用扣子的探索模式制定一份旅行攻略，时间在10 分钟以上，可以看到扣子将推理过程的思维链与搜索深度结合，践行“边想边搜”，在已获取到日本关西和熊本的小众景点、海边景点以及适合三十岁生日庆祝的特别地点信息后，扣子保存了景点信息，开始“边想边做”，从景点中筛选出合适的景点并规划出行程安排，在完成行程安排后，开始生成包含地图、景点介绍、必备日语短语及旅行提示的 html 旅行手册。

算力瓶颈与Agent服务的用户体验紧密相关

为了保证用户体验，Agent服务需要留出一定应对用户流量波动的冗余算力。用户对服务的访问量往往具有不确定性，会因各种因素如节假日、特殊事件、营销活动等出现峰值。DeepSeek官方在知乎发布的技术报告指出，由于白天的服务负荷高，晚上的服务负荷低，DeepSeek实现了一套机制，在白天负荷高的时候，用所有节点部署推理服务。晚上负荷低的时候，减少推理节点，以用来做研究和训练。**并非所有Agent服务提供商都有训练模型等需求可以充分利用闲时算力，因此能满足用户峰值的算力必然会存在一定的冗余。**



DeepSeek用于推理服务的H800 节点数量随时间峰值变化

目录

AI应用落地加速：端侧AI

各类智能终端（手机、耳机、眼镜等）将是AI 重要应用场景。荣耀产品线总裁方飞表示：“AI Agent应该是帮助你做你想做不会做，会做不想做的事，以及提供情感陪伴。AI Agent将会成为你智能化的助理，你可以把它理解为‘贾维斯’”。

端侧AI应用场景按人类的基本活动生活、生产划分可大致分为：1）生活：主要覆盖衣食住行教育医疗娱乐场景，如智能家居、智能穿戴设备、智慧交通、智能出行、智慧医疗、智慧娱乐、生活助理等；2）生产：生产主要覆盖物质生产、精神生产场景，如智慧农业、智慧制造、内容生产（游戏、音乐、影视）等。

类别	应用场景	具体场景
生活	智能家居	智能音箱、智能家电（智能台灯、智能冰箱、智能门锁、智能摄像头、智能电视、智能冰箱、扫地机器人、智能窗帘、烟雾警报设备等）
	智能穿戴设备	智能手环、VR 眼镜、多功能耳机等
	智能出行	自动驾驶（车载设备如自动驾驶车辆、路侧摄像头、智能红绿灯、车载娱乐终端设备等）、智慧交通等
	智能健康	通过各类可穿戴的终端设备组合，构成专注于个人健康的智慧医疗系统，如智能健康检测设备（如跌倒监控器、家用血压计、婴儿呼吸监测仪、非接触睡眠监测仪、跌倒传感器、心率监控设备、人体微动监控设备等）、智慧医疗影像分析、增强现实用于医学成像、手术机器人、康复机器人等
	休闲娱乐	AI游戏伙伴、智能音乐推荐、社交（聊天表情包生成）、XR等增强现实设备作为边缘游戏工具等
	智能教育	智能个性化学习系统、AI教师助手、多媒体教室等
	生活助理	信息摘要、邮件梳理与回复、行程规划助手、日程管理、智能搜索、智能问答、AI字幕、购物辅助、法律助手、个性化健身教练、个性化导游、陪伴机器人、聊天机器人等
生产	智慧农业	环境指标监测（通过部署众多不同功能的传感器收集作物的生长数据、土壤数据、环境数据以及气象数据并进行数据分析）、智慧温室大棚（安装加温补光、内外遮阳、风机、滴水灌溉、水肥一体化设备，精准控制并监测设备运行状态）、智慧农业园区（智能摄像头、温控设备、智能网关、采摘机器人等设备）、生长状态监测（作物长势监测、遥感、可见光，制备生态分析）、智慧农业机器人等
	智慧工业	研发规划类（需求识别的产品预测、AI辅助产品研发设计、设计方案生成等）、生产过程管控（预测性维护、设备系统故障诊断、车间调度与规划等）、机器人辅助生产（智能分拣、打标、搬运、上下料、焊接、喷涂、加工、装配、清洁等）、经营管理优化类（客户需求识别与预测、智能营销匹配、物流路径优化、智能销售终端等）、生产安全（智慧矿业、智慧能源等行业安防、侦察、排爆、采掘、专业巡检等机器）
	智慧服务	智慧零售（智能零售终端、智慧讲解导引、餐饮配送和服务）、智慧交通管理（路侧监控终端、道路指示灯终端等）、智能客服、智能销售助理、智能专业服务咨询等
	内容生产	营销文案创作、图片创作、短视频剪辑或创作、影视制作、AI辅助编程、辅助设计、游戏生成等
	智慧办公	文档助理、企业知识管理、代码开发助理等

资料来源：SeedV实验室

AI应用落地加速：端侧AI

端侧AI目前主要参与者有：

1)芯片制造商，如NVIDIA, Qualcomm, Intel, Apple（Neural Engine）, Google（Edge TPU）等；

2）终端设备制造商，如PC/Phone OEM厂商联想、华为、华硕、戴尔、宏碁、三星、荣耀、Apple、惠普、雷神、小米、oppo、vivo等，再如摄像头、麦克风、耳机、眼镜、音箱、传感器、白色家电、智能汽车等供应商；

3）操作系统供应商，如Microsoft（Windows）、Google（Android、Chrome OS）、Apple（macOS、iOS）、华为（HarmonyOS、EulerOS）等；

4）模型/算法供应商，如OpenAI, 字节，阿里等。

	具体进展	现有产品
算力层	CPU+GPU+NPU异构计算	PC处理器： X86架构：intel Meteor Lake（酷睿系列）、Lunar Lake；AMD Hawk Point（Ryzen AI 300系列）、Strix Point Arm架构：高通 Snapdragon X Elite/X Plus;Apple M3/M4 Phone处理器： 高通 Snapdragon 8Gen 3；联发科天玑9300+；三星Exynos 2400 ；谷歌 Tensor G3； A17 Pro
平台层	系统厂商强化软硬件兼容性	PC 系统：微软 Windows 11；谷歌 Chromebook plus OS；Apple MacOS 15 手机系统：安卓15、iOS18、HarmonyOS等
模型层	端侧小模型陆续推出	Phi-3小语言模型、Gemma模型、AI Hub库等
应用层	开发框架	NVIDIA AI Workbench、CUDA、Qualcomm AI Stack、AMD software 、天玑AI开发套件、华为HarmonyOS NEXT、TensorFlow Lite、PyTorch Mobile、高通AIMET等
	应用+系统	Copilot融入Windows 11；Android 15将深度融合谷歌的Gemini大模型；Apple Intelligence系统集成到到面向iPhone、iPad和Mac的新OS中；HarmonyOS NEXT 将 AI 能力融入系统等

AI应用落地加速：端侧AI

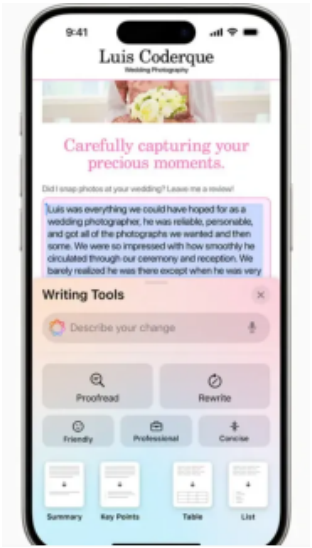
手机、PC芯片已经支持百亿级别参数模型

芯片类型	厂商	芯片	发布时间	算力性能	支持模型规模
手机芯片	高通	骁龙8 Gen3	2023年10月	AI算力超过73TOPS，其中NPU算力34TOPS	支持100亿参数模型，支持端侧多模态，Llama2-7B每秒20tokens
	联发科	天玑 9300	2023年11月	NPU算力33TOPS	支持130亿参数模型，70亿参数模型20 tokens/秒
		天玑 9300+	2024年5月	AI算力68TOPS，NPU算力48 TOPS	
	Apple	A17 Pro	2023年9月	NPU算力35TOPS	
PC芯片	高通	骁龙X ELite	2023年10月	AI算力75TOPS，其中NPU 45TOPS	12核OryonCPU，支持130亿参数模型，Llama2-7B30tokens/秒
	intel	Meteor Lake	2023年12月	AI总算力34TOPS,共中NPU IITOPS	
		Lunar Lake	2024年6月	AI总算力120TOPS，其中NPU 48 TOPS	首次使用封装级内存，采用LPDDRSX,节省40%功耗和250mm主板空间
	AMD	锐龙8040	2023年12月	AI总算力39TOPS，共中NPU 16TOPS	
		Ryzen AI 300	2024年6月	NPU 50 TOPS	
	Apple	M3系列芯片	2023年10月	NPU算力18TOPS	M3/M3 Pro/M3 Max分别拥有10/18/40核GPU
		M4	2024年5月	NPU算力38TOPS	10核GPU，相比M2，CPU性能提升最高1.5倍，GPU渲染性能提升最高4倍
	Navida	RTX40 Super系列	2024年1月	686 AI TOPS	

资料来源：高通、联发科、Apple、AMD、intel、Navida官网

苹果：

2024年6月10日，苹果在全球开发者大会上推出了Apple Intelligence，这是一款深度集成到 iOS 18、iPadOS 18 和 macOS Sequoia 中的个人智能系统。Apple Intelligence 由多个功能强大的生成模型组成，这些模型专门用于用户的日常任务，并可即时适应用户当前的活动。Apple Intelligence 内置的基础模型针对用户体验进行了微调，例如编写和优化文本、确定通知的优先级和摘要、为与家人和朋友的对话创建有趣的图像，以及执行 App 内操作以简化跨 App 的交互。



Apple Intelligence写作工具

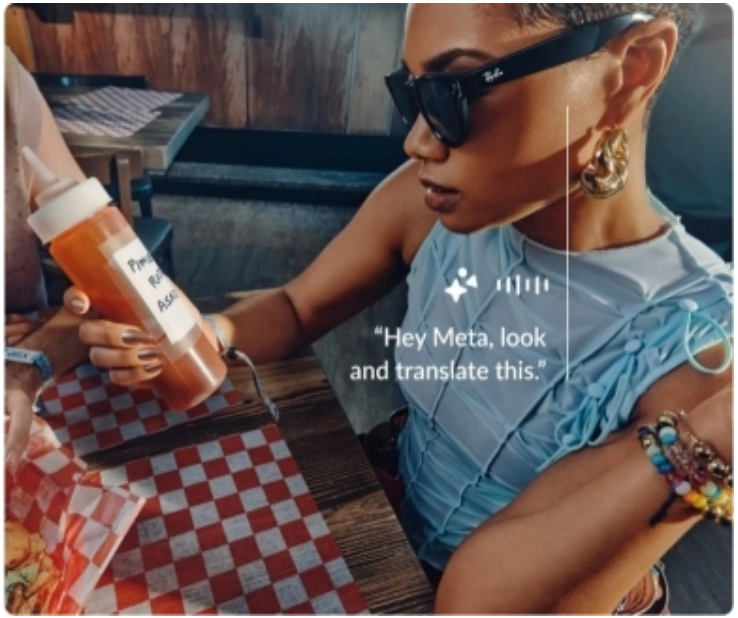
Memories可以根据用户的描述挑选照片和视频，设计故事情节制作电影

Siri可做出数百种跨 app 的新操作，包括为用户找到朋友在信息 app 或邮件 app 里推荐的图书

Meta

2023年9月，Meta与雷朋合作推出第二代联名产品Ray-Ban Meta眼镜，产品迅速赢得了市场的认可。根据IDC的数据，2023年第四季度和2024年第一季度，Ray-Ban Meta的出货量分别达36万台和10万台；截至2024年第二季度末，该产品的出货量已经超过了100万台，预计2024 年全年出货量有望超过150万台。使用 Meta AI，用户可以提出一般问题并接收音频回复，或者拍摄图像并询问有关图像内容的问题。例如，如果用户正在为一群朋友做饭，可以问“Hey Meta。我正在烤扇贝、玉米棒和汉堡。我应该把它们每个煮多长时间？”甚至可以跟进Hey Meta.什么沙拉配得好呢？”

Ray-ban Meta眼镜能力



资料来源：Meta官网，国盛证券研究所

AI应用落地加速：端侧AI

DeepSeek开源的蒸馏小模型超越 OpenAI o1-mini加速AI在端侧落地。DeepSeek开源的32B和70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果，有望促进AI PC、AI手机、AI眼镜、AI玩具等各类智能终端发展。

	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

DeepSeek发布的蒸馏小模型效果超越o1-mini

字节旗下oladance推出的AI耳机

字节AI玩具显眼包

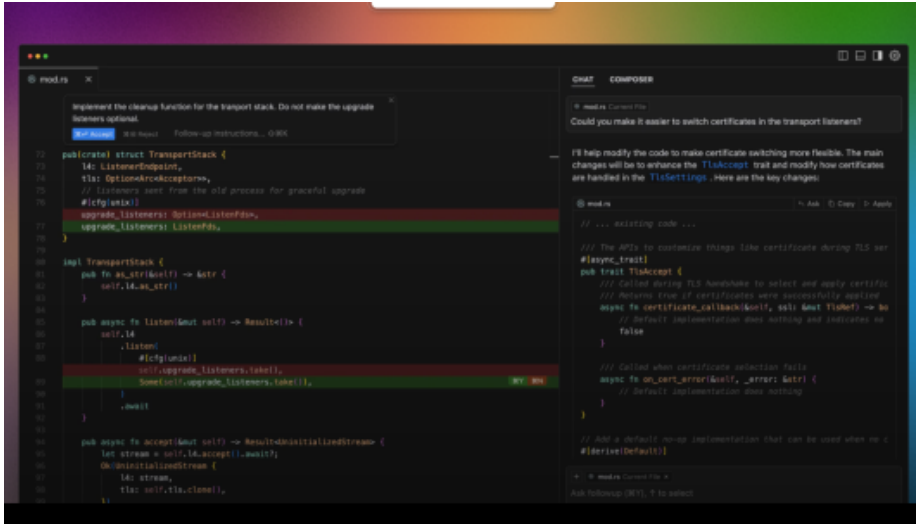
AI应用落地加速：AI编程

编程有比较准确的评判标准，Github等社区拥有海量高质量代码数据，是大模型能较快取得进步的方向。

各类AI代码工具持续涌现，商业化进展迅速。集成大模型的IDE Cursor到2024年8月已有超过40000名企业客户；IDE插件GitHub Copilot自全面推出以来有超过77000个组织采用；国内也有阿里通义灵码AI程序员等应用

2月7日，微软CEO纳德拉宣布GitHub Copilot正all -in到智能体，并推出自主的SWE agent，用于协助软件工程师，可以执行各种开发任务。例如生成和审查代码、重构或优化代码库、自动化测试或管道等工作流程，以及提供架构、错误故障排除和最佳实践方面的指导。

纳德拉宣布GitHub Copilot正all -in到智能体



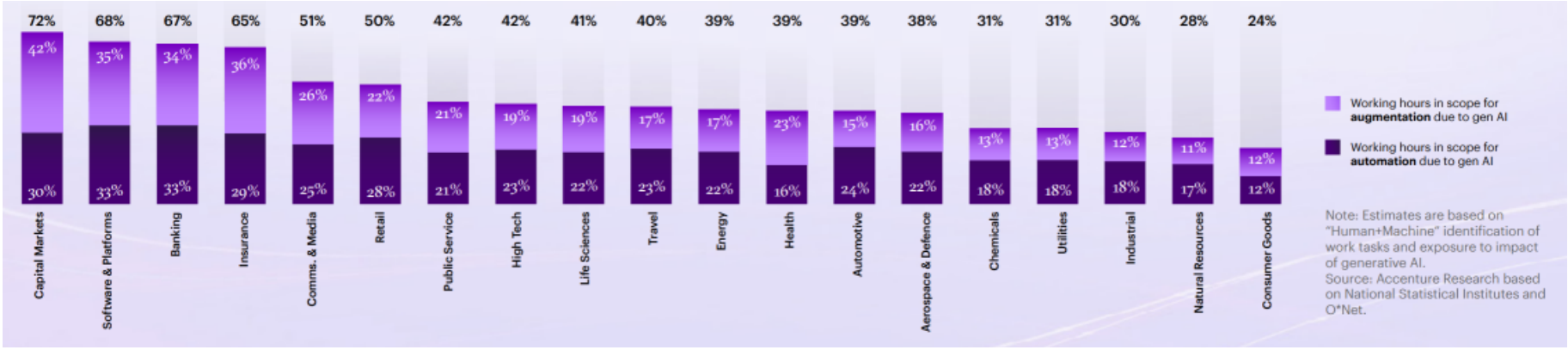
AI编程工具 Cursor到2024年8月已有超过40000名企业客户

AI应用落地加速：AI编程

我们预计2025年下半年起，AI编程会产生显著商业化效果，由于计算机行业人员成本是重要的成本组成部分，有望给计算机行业公司整体利润率带来极大的提升空间。

2024年1月，埃森哲发布了一份标题为《Work, workforce, workers: Reinvented in the age of generative AI》（在生成式人工智能时代重新定义工作、劳动力和工作者）的咨询报告，深入探讨了生成式AI对工作、劳动力和工作者所带来的影响。埃森哲基于对工作任务中人类与机器交互程度的分析，考虑了哪些工作任务目前或将来可以由人工智能系统自动完成或辅助完成，分析了生成式AI对工作时间的潜在影响。从各行业看，软件平台的工作在生成式AI的影响范围内的工作时间占比位居第二位，达到68%。

谷歌首席执行官Sundar Pichai在2024三季度财报电话会议上介绍了AI在软件开发中日益增长的影响力。根据Pichai的说法，人工智能系统现在负责为谷歌产品生成超过 25% 的新代码，而人类程序员则监督和管理这些人工智能生成的贡献。



埃森哲：生成式AI能影响（自动化或增强）各行业的工作时间

AI应用落地加速：AI编程

人员薪酬是计算机企业最重要的成本及费用，占申万计算机板块收入比例可达约44%。由于最终产品以软件项目或者软硬一体化项目的形式向客户交付，计算机板块是典型的轻资产行业，实体硬件成本及费用并不显著。计算机行业最重要的资产是人力资源，最主要的成本和费用则是人员薪酬，即软件工程师的工资。通过wind数据计算，一分为二，从成本和收入两端来看：

成本端：根据wind数据统计，申万计算机板块2018-2023年平均销售毛利率为26.6%，即营业成本占营业收入的比重约为73.4%。而根据我们对计算机板块一些典型领军企业财报数据的统计可知，人员成本是公司营业成本的重要组成部分，平均来看，2023年人员成本占总体营业成本的比重可达40%左右。（注：营业成本结构这一数据，并非所有公司都在财报中披露，因此无法直接提取计算全行业情况，只能选取典型公司获得平均值）因此，若按以上数据计算，我们可以大致推算出，人员成本大约占计算机板块总收入的29%。

费用端：根据wind数据计算可知，2023年申万计算机板块，人员费用占总收入比例为15%。

将成本端与费用端加总可知，人员薪酬占计算机板块总收入的比例约为44%

图表1：*2018-2023年计算机板块营业成本占营业收入比例情况*

	销售毛利率	营业成本占营业收入比例
2018/12/31	*26.9%*	*73.1%*
2019/12/31	*27.8%*	*72.2%*
2020/12/31	*26.8%*	*73.2%*
2021/12/31	*25.8%*	*74.2%*
2022/12/31	*25.3%*	*74.7%*
2023/12/31	*27.1%*	*72.9%*
平均数	***26.6%***	***73.4%***

资料来源：wind，国盛证券研究所

图表1：*申万计算机板块整体人员成本及人员费用情况*

费用端	
2023年销售费用工资薪酬总和（亿元）	597.22
2023年管理费用工资薪酬总和（亿元）	365.74
2023年研发费用工资薪酬总和（亿元）	851.92
2023年申万计算机板块总收入	12,114.38
人员费用占总收入比例	**15%**
成本端	
2018年至2023年，计算机板块营业成本占收入比重平均数	73%
计算机板块公司营业成本中，人工成本占比	39.88%
人员成本占总收入比例	**29%**
人员薪酬（成本+费用）占计算机企业营业收入的比重	**44%**

资料来源：wind，国盛证券研究所

AI应用落地加速：AI编程

进行AI编程对利润端弹性的敏感性测算可知，AI编程带来的效率提升可以为计算机板块带来接近翻倍的净利率空间：

1）2018-2023年，计算机板块平均销售净利率为3.4%。人员薪酬占计算机板块收入比例可达约44%。

2）假设人员效率分别提升5%/10%/20%/25%/30%，即能够减少原有44%人员成本中的5%/10%/20%/25%/30%，也就是说，按这两者相乘计算，可以分别带来2%/4%/9%/11%/13%的增量利润空间。

3）因此，在人员效率分别提升5%/10%/20%/25%/30%的假设下，AI编程带来的全新净利率水平分别约为6%/8%/12%/14%/17%（按原本的净利率+增量利润空间计算），相比原本的计算机行业净利率水平有极大的提升。**即使是人员效率仅提升5%的偏低假设下，净利率也可提升至6%，相比原本3.4%的水平，有接近翻倍的提升。**

我们认为，北美AI工具多用于B端降本，计算机厂商对内对外都可输出能力， AI代码工具出现后，预计作为第一个高准确度AI产品会快速爆发。由于计算机行业人员成本是最主要的成本组成部分，B端提效+AI编程有望带来行业利润极大的潜在提升空间。

图表1：*2018-2023年计算机板块销售净利率情况*

	销售净利率
2018/12/31	3.2%
2019/12/31	2.9%
2020/12/31	3.7%
2021/12/31	4.8%
2022/12/31	2.9%
2023/12/31	2.7%
平均数	3.4%

资料来源：wind，国盛证券研究所

图表1：*AI编程效率提升给计算机板块带来的利润空间敏感度测算*

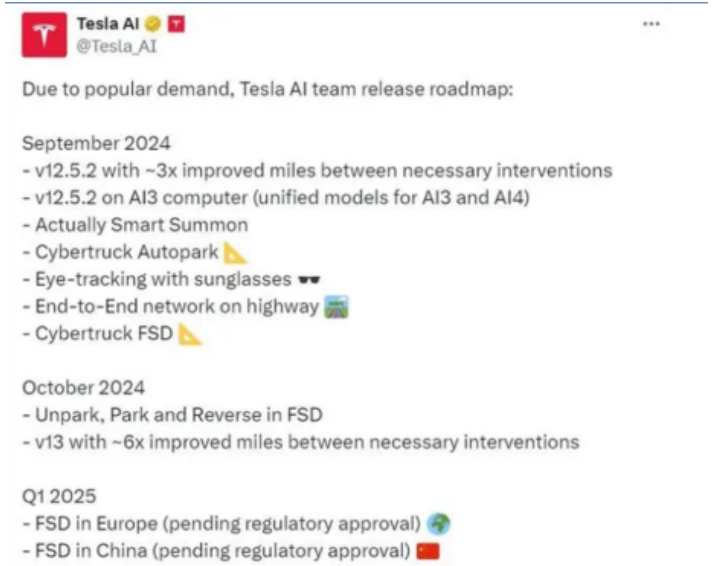
AI编程效率提升带来的利润空间敏感度测算					
人员薪酬占计算机企业营业收入的比重			44%		
假设人员效率提升	5%	10%	20%	25%	30%
可节省出的利润空间	2%	4%	9%	11%	13%
2018至2023年计算机板块平均净利率水平			3%		
AI编程带来的全新净利率水平	6%	8%	12%	14%	17%

资料来源：wind，国盛证券研究所

AI应用落地加速：智驾/机器人

特斯拉：FSD V12为首个端到端自动驾驶系统。传统的自动驾驶按照感知、决策和控制划分为不同的模块，系统先对周围的动静态交通参与者和路网结构进行准确感知，再规划车辆的行车轨迹，最后通过执行机构对进行闭环控制。2023年8月，特斯拉CEO马斯克在做FSD Beta V12试驾直播时，重点介绍说“这是世界上第一个端到端AI自动驾驶系统”，首次将端到端大模型的概念引入自动驾驶。从特斯拉的端到端方案来看，它将自动驾驶系统的感知和定位、决策和规划、控制和执行之间的断面整合在了一起，形成一个大的神经网络，即通过传感器采集到原始数据，将原始数据输入神经网络系统，直接给车辆底层控制器输出加速、制动、转向等驾驶指令。本质上，特斯拉的端到端FSD是将上千万个视频片段包含的人类驾驶知识压缩到了端到端神经网络参数中。

CES 2025一月大会的采访上，马斯克表示，“（现阶段）自动驾驶汽车的性能提升速度呈指数级增长。有信心在三个月内，也就是今年第二季度，实现自动驾驶汽车的性能超越人类驾驶。”，展现出马斯克对FSD迭代速度的信心。2）根据特斯拉自动驾驶工作人员Ashok Elluswamy的推特信息披露，特斯拉FSD v13.2 已开始向有限的外部客户推出。而根据马斯克在推特上的表述，V13 在每次必要干预之间的英里数将比现在好5到10倍。特斯拉FSDv13版本的发布显示了公司端到端算法迭代速度，我们预计伴随着特斯拉数据和算力的持续积累，其算法将继续加速迭代，为自动驾驶带来更好的乘驾体验。



特斯拉在社交平台X宣布FSD将在25年一季度进入中国和欧洲

AI应用落地加速：智驾/机器人

华为：

2019年可视作华为正式进入汽车赛道，早期定位为2B零部件提供商。2013年，华为成立“车联网业务部”，且在同年推出车载通信模块ME909T。此后，2014到2018年，华为先后成立车联网实验室，与东风、长安、一汽签署合作协议，合作技术与产品研发。2019年，华为首次在上海车展以零部件供应商身份出现，并于5月正式成立智能汽车解决方案事业部（“车BU”），可视作正式进入汽车赛道。

发布Huawei Inside品牌，转型解决方案提供商。2020年10月，全栈智能汽车解决方案品牌Huawei Inside（HI）发布，明确自身作为车厂智能软硬件供应商身份，确认了与车厂合作模式。

- HI全栈智能汽车解决方案包括：1个全新计算与通信架构和5大智能系统，智能驾驶，智能座舱、智能电动、智能网联和智能车云，及激光雷达、AR-HUD等全套的智能化部件。HI 技术帮助汽车产业实现技术升级，快速开发领先的智能电动汽车。
- HI提供强大算力和OS，包括三大计算平台，智能驾驶计算平台、智能座舱计算平台和智能车控计算平台，以及三大操作系统AOS（智能驾驶操作系统）、HOS（智能座舱操作系统）和VOS（智能车控操作系统）。依托强大算力和OS支持，汽车可实现软件定义，持续开发新功能，提升和优化用车体验。

“华为智选”合作模式，兼顾2C优势整合。2020年11月，华为将车BU业务管辖关系调整至消费者管理委员会，由消费者业务CEO负责，一方面有利于研发、品牌、零售资源协同，另一方面也可发挥手机端与车端业务的融合。

2024年1月16日，华为成立深圳引望智能技术有限公司，注册资本为10亿元人民币，由华为技术有限公司全资持股。2024年3月，引望在上海、东莞、苏州、杭州、南京5地注册全资子公司，5月在武汉成立分公司。2024年8月，赛力斯集团和长安汽车的子公司阿维塔科技均宣布将斥资115亿元人民币收购引望公司10%的股份，这使得引望的估值飙升至1150亿元人民币。交易完成后，华为将持有引望80%的股份，而赛力斯和阿维塔则各自持有10%。此次阿维塔和赛力斯的入股标志着引望在成立不足8个月的时间里，其估值已经达到了1150亿元。引望整合华为车BU的现有技术和资源，并通过股权合作的方式吸引更多汽车制造商参与到智能汽车解决方案的研发与应用中，以此促进汽车行业的智能化进程。

华为智能汽车解决方案微博

AI应用落地加速：智驾/机器人

百度：萝卜快跑是百度旗下自动驾驶出行服务平台，已于全国11个城市开放载人测试运营服务，实现超一线城市全覆盖。此外，萝卜快跑已经开始在北京、武汉、重庆、深圳、上海开展全无人自动驾驶出行服务与测试。截至2024年4月19日，百度萝卜快跑在开放道路提供的累计单量超过600万，稳居全球最大的自动驾驶出行服务商。

近两年来，武汉市加快开放自动驾驶测试道路，在全国率先发布全无人驾驶商业化试点政策，实现跨区通行、跨江通行、机场高速通行等多个自动驾驶商业应用场景的全国创新突破。“从常态化运行的自动驾驶出行服务车辆数量、订单量，以及开放道路里程、面积等核心数据来看，武汉已成为全球最大的自动驾驶出行服务区。”国家智能网联汽车（武汉）测试示范区相关负责人介绍。2022年8月，萝卜快跑在武汉经开区启动车内无人商业化示范，越来越多的武汉市民已经接受了自动驾驶出行服务。截至2024年5月15日，百度萝卜快跑已攻克武汉的复杂道路场景，实现了武汉城市全域、全时空场景覆盖，为近半数的武汉市民提供便捷的无人化出行服务。

自用无人车成本已降至20万以下，有望带来运营服务成本的显著降低。搭载百度Apollo第六代智能化系统解决方案的萝卜快跑第六代无人车，整车成本相较于5代车直接下降60%，价格只需要20万，再次刷新了行业纪录。首批交付的第六代无人车，将在武汉投入使用，年内在武汉完成千台无人车的部署，让更多用户享受到绿色、安全的美好出行。与此同时，随着萝卜快跑无人车自动运营网络完成建设，营运成本将降低30%，通过自动驾驶技术和人车舱效率的持续优化，服务成本将降低80%。精细化的成本管控，使得萝卜快跑成本持续降低

百度第六代无人车Apollo RT6成本降至20万

AI应用落地加速：智驾/机器人

比亚迪

2月10日，比亚迪董事长兼总裁王传福在比亚迪智能化战略发布会上表示，高阶智驾系统“天神之眼”正式发布，分为三个版本，其中三激光版本主要搭载于仰望车型。王传福表示，“天神之眼”高阶智驾可实现全程高速0接管

王传福表示，比亚迪将通过车辆卓越的性能与智能控制能力真正做到赛道无人驾驶，该系统可以全方位无死角采集赛道上的各种信息。同时，赛道无人驾驶通过扭矩矢量自动控制、分配，达成整车智能战略下性能与智能的极致融合。比亚迪集团高级副总裁、汽车新技术研究院院长杨冬生在比亚迪智能化战略发布会上介绍，比亚迪“璇玑架构”全面接入DeepSeek。比亚迪旗下共计20款车系完成上市，这20款车系均搭载天神之眼智驾。其整体售价区间为6.98-24.98万元。

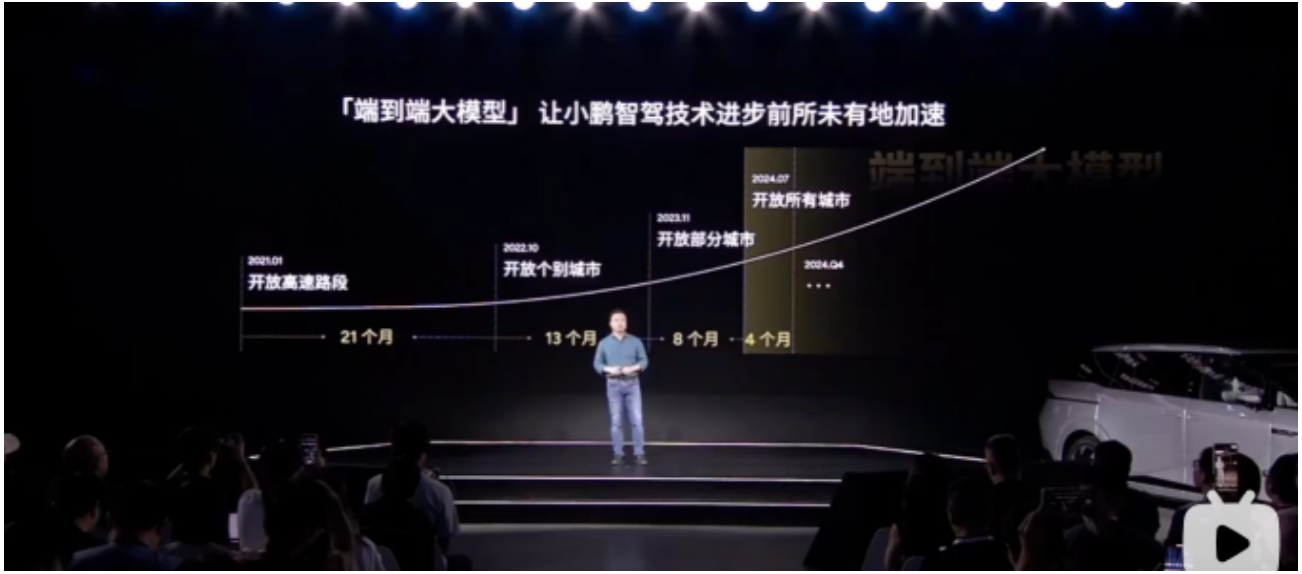
比亚迪“天神之眼”智驾系统发布

AI应用落地加速：智驾/机器人

小鹏：

率先在国内实现端到端量产上车。2024年5月20日，小鹏汽车全量推送AI天玑系统，在国内率先实现了城区智驾100%无图覆盖、端到端自动驾驶大模型的量产应用。据悉，小鹏的端到端大模型由三部分组成：神经网络XNet+规控大模型XPlanner+大语言模型XBrain。上述三网融合的大模型能2天迭代一次，XNGP的能力在18个月内提升30倍，而用户能明显感知到的结果是前后顿挫减少50%、违停卡死减少40%、安全接管减少60%。

据小鹏汽车董事长、CEO何小鹏透露，基于折算10亿+里程的视频训练、超646万累计公里数的实车测试、超2.16亿累计公里数的仿真测试，小鹏汽车端到端大模型能够做到“两天迭代一次”，在未来18个月内智驾能力提升30倍。按照规划，2024年第三季度，小鹏汽车的智驾即可实现“全国都能开，每条路都能开”，2025年实现城区智驾比肩高速智驾体验。同时小鹏汽车也正在全球范围对XNGP端到端的能力进行测试，智驾技术开始走向全球。



小鹏汽车AI智驾技术发布会

AI应用落地加速：智驾/机器人

理想：

2024智能驾驶夏季发布会发布了基于端到端模型、VLM视觉语言模型和世界模型的全新自动驾驶技术架构，并开启新架构的早鸟计划。新架构由端到端模型、VLM视觉语言模型和世界模型共同构成。VLM视觉语言模型具备强大的逻辑思考能力，可以理解复杂路况、导航地图和交通规则，应对高难度的未知场景。世界模型结合重建和生成两种路径，构建的测试场景既符合真实规律，也兼具优秀的泛化能力。 理想汽车的自动驾驶全新技术架构受诺贝尔奖得主丹尼尔·卡尼曼的快慢系统理论启发，模拟人类的思考和决策过程，形成更智能、更拟人的驾驶解决方案。快系统，即系统1，善于处理简单任务，是人类基于经验和习惯形成的直觉，可以应对驾驶车辆时的常规场景。慢系统，即系统2，是人类通过更深入的理解与学习，形成的逻辑推理、复杂分析和计算能力，在驾驶车辆时用于解决复杂甚至未知的交通场景。系统1和系统2相互配合，分别确保大部分场景下的高效率和少数场景下的高上限，做出决策的基础。

理想汽车基于快慢系统系统理论形成了自动驾驶算法架构的原型。系统1由端到端模型实现，具备高效、快速响应的能力。端到端模型接收传感器输入，并直接输出行驶轨迹用于控制车辆。系统2由VLM视觉语言模型实现，其接收传感器输入后，经过逻辑思考，输出决策信息给到系统1。双系统构成的自动驾驶能力还将在云端利用世界模型进行训练和验证。

理想汽车新自动驾驶新架构

机器人：宇树等国产头部科技公司的机器人进展不断，伴随人工智能的技术突破，政府政策的大力支持，各个大厂的参与布局，人形机器人行业发展踏上了快车道，特斯拉、1-X、Figure、智元、傅里叶等国内外知名整机厂均已实现小批量生产，量产时间越来越近。央视春晚16个宇树科技H1机器人手持手绢上演了创意融合舞蹈《秧Bot》全网。国内机器人进展与Optimus形成中美共振，赛道前景广阔.

春晚《秧Bot》

AI应用落地加速：军工AI

OpenAI与武器厂商合作，AI+军事应用场景有望持续扩大。根据财联社报道，12月4日OpenAI与国防科技初创公司Anduril Industries共同宣布，双方将建立战略合作伙伴关系，以开发和负责任地部署用于国家安全任务的先进人工智能 (AI) 解决方案。这标志着OpenAI首次与一家商业武器制造商合作，这也是迄今为止该公司与美国国防部最深入的合作。Anduril在去年11月宣布了一项2亿美元的合同，向美国海军陆战队提供该公司的反无人机系统。两家公司表示，合作将专注于提高美国的反无人机系统(CUAS)及其实时检测、评估和应对潜在致命空中威胁的能力。Anduril和OpenAI将探索如何利用前沿人工智能模型快速合成时间敏感数据、减轻人类操作员的负担并提高态势感知能力。这些模型将在Anduril业界领先的CUAS威胁和行动数据库中进行训练，以帮助保护美国和盟军军事人员并确保任务成功。

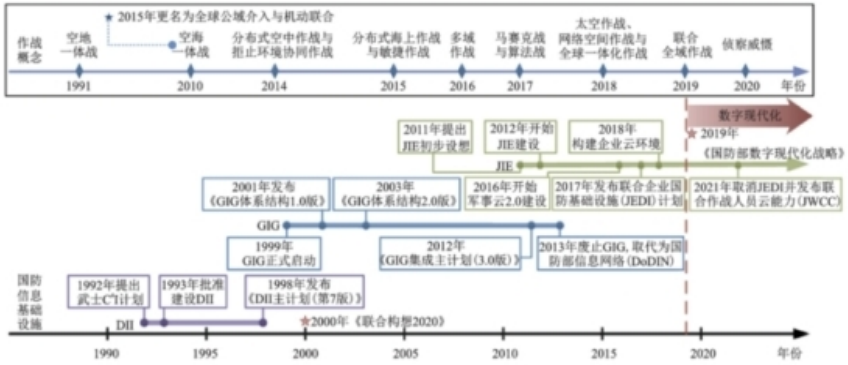
随着AI驱动致命性自主武器系统（LAWS）投入实战，继火药和核武器之后的“第三次军事革命”将可能发生。根据北京日报2024年5月报道，美国前军官保罗·沙尔，入选《时代》杂志的人工智能领域最具影响力100人。今年2月，他在《外交事务》杂志上发出警告称：“如不对自主武器系统加以严格管控，未来人类的角色将沦为打开机器并坐在场边，且几乎没有能力控制甚至结束战争。”据彭博社报道，在美军于2024年初在中东进行的空袭轰炸中，梅文智能系统至少帮助美军锁定了85个打击目标，地点涉及伊拉克、叙利亚、也门和红海地区。据《时代》周刊报道，2024财年，美国军方对人工智能的投资迅猛增加。根据美国公布的2024财年政府预算，包括国防部、能源部、国土安全部等多个机构，累计向AI领域计划投入超过2511亿美元。如果将政府外部筹资、资本市场的投入计算在内，2024年，美国在AI领域投资预计超过数万亿美元。

AI应用落地加速：军工AI

美国AI大幅应用于军事的核心前提是军队云充分建设。美军云基础设施建设分为四个阶段。根据防务快讯《指挥信息系统与技术》，在持续创新的作战概念牵引下，美军不断推进国防信息基础设施体系的建设进程，主要分为以下4个阶段：

- **以国防信息基础设施（DII）为代表的第1个阶段：**美军针对海湾战争暴露的“烟囱”问题提出了通过构建军事信息基础设施，助力跨军兵种信息系统综合集成的理念。1992年，美军提出武士C4I计划，并于次年批准DII计划作为其基础支撑。1992—1998年，美军相继发布7个版本的《DII总计划》，国防基础设施建设得以持续深化。
- **以全球信息栅格（GIG）为代表的第2个阶段：**波黑战争和科索沃战争后，美军进一步总结经验，于1999年提出GIG概念，并在《2020联合构想》中将GIG作为实现网络中心战的重要基础，支撑网络中心和面向服务的技术理念落地。截至2012年，GIG经历了3个发展阶段。
- **以联合信息环境（JIE）为代表的第3个阶段：**2011年10月，美国防部发布了信息技术企业战略和路线图，针对GIG建设过程中暴露的互操作性差、规模过于复杂庞大及成本高昂等问题提出了JIE的初步设想，并自2012年起分阶段推进JIE建设，旨在提供一个安全的联合信息环境，包括单一安全架构、共享IT基础设施和企业服务，以满足美军全球一体化联合作战的需求。
- **4）以数字现代化战略（DMS）为代表的第4个阶段：**2019年7月，美军发布《国防部数字现代化战略》，并将其作为IT领域的顶层战略。2020年，国防部首席信息官批准数字现代化基础设施执行委员会章程，将DMS视为具体计划，将JIE纳入其中并对工作内容进行了延展。2021年1月，美国防部作战试验和评估办公室（DOT&E）发布的2020年度报告将JIE计划更名为DMS相关企业IT倡议，意味着DMS取代了JIE，并成为美军国防信息基础设施体系未来的发展方向。

美军国防信息基础设施发展历程



资料来源：安全内参，国盛证券研究所

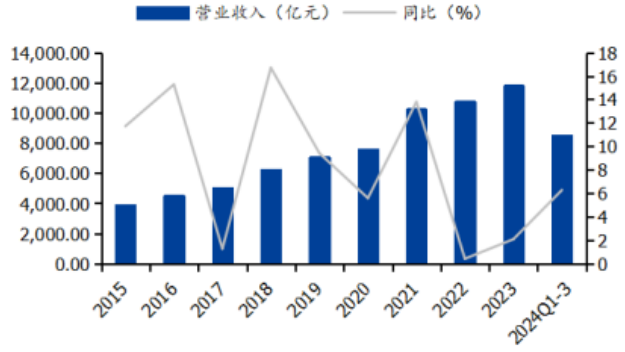
我国军队云建设有望加速。AI持续应用于军事，依赖基础设施建设以及大数据分析软件，我国特种云建设有望加速。2024年8月，解放军报发表《新时代推动军队高质量发展的科学指南》，指出当前，世界之变、时代之变、历史之变正以前所未有的方式展开，要求我们必须增强本领恐慌意识，找准短板弱项，优化素质结构，在实践中占据主动、赢得未来。要靠融合共享赋能，坚持用开放的胸襟接受新事物，用改革的精神创造新办法，紧跟信息化智能化发展趋势，把大数据、物联网、云计算等先进技术运用到军队建设中，以发展模式改革带动本领全面增强。

图表14：AI 应用于军事依赖上游云建设和大数据分析帮助

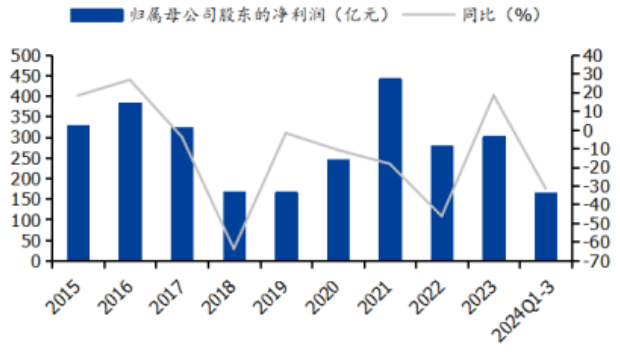
目录

2024年外部环境承压，国内IT企业估值仍处于历史低位。近年外部环境有所波动，以党政军、央国企为代表的群体IT开支景气度不高，传导至计算机公司层面，其利润、现金流、收入增速即相对较低，估值亦处于历史低位。2024Q1-3，计算机（申万）行业营业收入约为8498.48亿元，同比增长6.27%；归母净利润约为163.79亿元，同比下降31.55%。

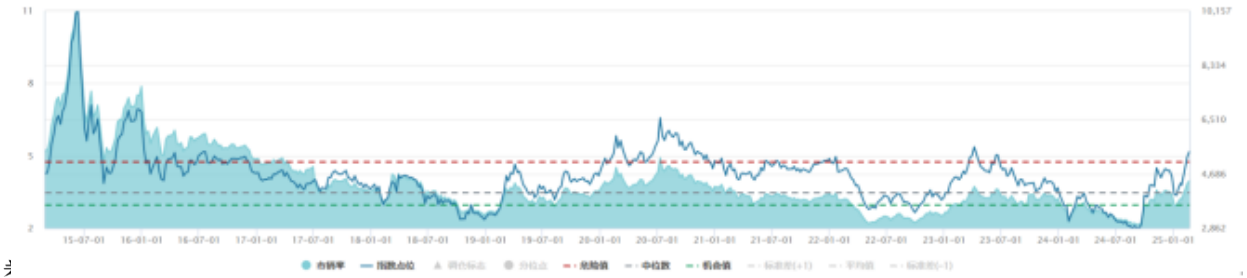
图表 1：2015-2024Q1-3 计算机（申万）行业收入及同比



图表 2：2015-2024Q1-3 计算机（申万）行业利润及同比



图表 3：计算机（申万）近十年 PS



DeepSeek驱动国内IT基本面质变，估值修复空间显著

1）DeepSeek 提出多项算法创新，打破了海外算力堆砌的传统大模型提升路径，展示出中国在AI领域已经具备与全球顶尖水平竞争的实力，我们认为这将使全球投资者对中国科技企业的技术创新能力塑造新的认知，进而对中国科技资产重新进行价值评估

2）Deepseek开源+高性能+低成本显著降低应用开发门槛，应用落地方面中国优势显著。中国拥有庞大的工程师队伍，自移动互联网时代以来已经积累了丰富的数据资源以及社交网络和电商领域完善的生态系统，DeepSeek降低了AI技术门槛，让这些资源与 AI 技术更好地结合，完善了中国 AI 产业生态，使中国在全球 AI 产业竞争中占据更有利的位置推动科技资产价值上升。

3）应用爆发带来广泛推理需求，为国内算力产业带来发展机遇。相比于模型训练，推理需求对硬件要求相对较低，集群互联的要求也更为宽松，为国产芯片、英伟达中低端卡以及ASIC芯片等国内算力提供了有利的发展机遇。寒武纪、昇腾、沐曦、天数智芯、摩尔线程、海光信息、壁仞科技、太初元碁、云天励飞、燧原科技、昆仑芯、灵汐科技、鲲云科技、希姆计算、算能科技、清微智能、芯动力、龙芯中科、瀚博半导体等。

4）财政政策或更积极，IT企业增速有望进一步修复。2025年1月10日，财政部副部长廖岷表示，加积极的财政政策未来可期。财政将围绕加快发展新质生产力，要加大对教育人才、科技攻关、乡村振兴、绿色低碳等领域的支持。在应对国际国内形势的变化上，将保持密切跟踪，适时进行科学设计和动态调整，梯次拿出政策“后手”，为经济社会发展提供强有力支持。我们认为，自主可控、人工智能、国产算力等作为新质生产力重要方向，未来有望得到更多支持。

腾讯全面拥抱DeepSeek。

腾讯元宝、微信、ima、腾讯文档、QQ浏览器、QQ音乐等多款腾讯产品接入DeepSeek-R1模型。腾讯及其投资版图具备丰富的应用场景和海量用户，有望加速AI技术普惠全民。同时腾讯作为自身有混元大模型的互联网大厂率先拥抱开源模型，也为更多企业整合开源技术+场景深耕提供良好的示范效应。

微信搜一搜接入DeepSeek，上线“AI搜索”入口

腾讯文档的AI文档助手已与DeepSeek-R1结合，可以直接生成文档、表格、幻灯片、思维导图、智能文档

图表 5：微信搜一搜 AI 搜索入口。

图表 6：腾讯文档的 AI 文档助手。

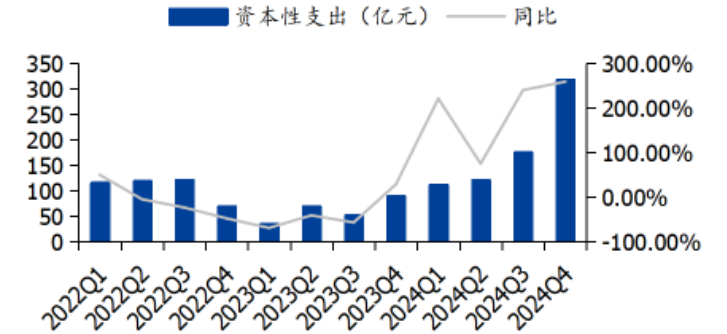
阿里云收入及资本开支大超预期

2月20日阿里发布2024年12月底截止季度报告，阿里云实现收入317.42亿元，同比增长13.10%。且DeepSeek R1在2025年1月20发布，我们认为DeepSeek对国内云计算需求将进一步带来巨大增量。

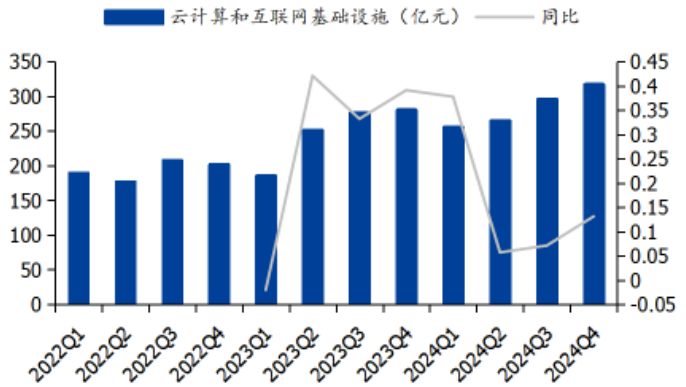
阿里单季度资本开支达到317.75亿元，同比增长258.76%。

阿里业绩会上管理层指出：追求AGI是阿里努力的关键目标。对AGI的追求可以带来巨大的商业价值。AGI的标准可能是人工智能它可以取代或实现80%的人类能力。全球GDP的50%左右是人力工资，包括智力或脑力劳动和体力劳动。因此，如果能够实现AGI，那么这可能会对全球重组行业产生巨大影响。它可能会对全球GDP产生重大影响，甚至取代50%的GDP。阿里管理层认为，在未来95%的输出token将在云端生成并由云端分发，因为只有云计算网络才能以最高的效率生成和分发token。阿里对将人工智能深度整合到自己的场景中保持开放的态度，以在所有的业务中创造价值。**未来三年阿里在云和人工智能方面的投资将超过过去10年的总和**

图表 7：2022-2024 年阿里巴巴资本开支及同比



图表 8：2022-2024 年阿里云收入及同比



对标美国云计算与SAAS，中国企业尚处低估状态

通用软件为企业管理的核心IT系统，将成为AI应用的主要载体。1）成体量的通用软件一般为企业内部的核心IT系统，能为企业的经营流程带来明显的管理效益，才得以收获高粘性的客户、以及较强的付费意愿。2）这类通用软件，一般串联了企业内外部事项的核心流程，覆盖行政、财务、采购、销售、库存、人事、客户关系、市场销售、文档资产等多方面，将成为AI应用升级的主要载体。

CRM：客户关系管理领域CRM 与 AI 结合，能智能处理客户数据、优化销售流程、提升客户服务效率、实现营销自动化等。2024年12月17日Salesforce正式推出其全新平台Agentforce 2.0。

ERP：HR SaaS、AI+财务管理等方向逐步落地。1）HR SaaS可用于多类场景，具体包括假勤月报异常分析、智能人才发现推理搜索类场景等。2）AI+财务管理空间较大、任务也相对复杂，包括全面预算、风险控制、成本控制、财务流程自动化等各方面。

OA：OA为日常工作的高频入口，市场、合同、客服、项目等数智化管理场景众多。1）OA为一般员工使用公司系统的高频入口，伴随移动化、云化发展，OA逐步向广义协同办公平台演进。包括流程审批、文档管理、沟通协作、日程管理、移动办公等功能。2）以泛微数智大脑Xiaoe.AI为例，其具备智能问答及AI搜索、智能摘要及内容解析、智能内容撰写及校对、AI图像识别及数据生成转化、AI智能数据分析及意图识别、智能流程引擎及自动化处理助手、业务与数据处理的RPA自动化引擎等多种功能。

文档办公平台：文档为工作知识的核心载体，AIGC创作从可用向好用演进。以WPS AI为例，其在2024年7月发布全新2.0版本，在个人办公方面发布AI写作/阅读/数据/设计助手；在企业端正式发布面向企业的智能文档库，支持智能问答与创作；在政务版方面针对垂类场景继续优化，包括政务AI写作/问答等。

数据终端：大数据已成为工作决策的基础支撑，针对金融、大宗等高频数据场景，AI有望深度赋能。1）同花顺“问财”为金融助手的代表产品，可提供智能体、选股票、诊股票、看大势、学投资等等一系列能力。2）对于上海钢联，其“小钢”数字智能助手1.6版本在“我的钢铁网”网站、APP、钢联数据终端等多渠道上线，通过与智能助手交互对话的方式，“小钢”数字智能助手能为客户提供查价格、读资讯、写报告、问百科、找商机等内容服务，此外还具备了市场分析、多模态生成和智能客服等功能，并辅助文章阅读，为用户提供AI摘要，生成的内容每天都服务在大宗商品行业各个领域

对标美国云计算与SAAS，中国企业尚处低估状态

在DeepSeek R1开源前，国内大模型进展相对较慢，美国企业更早受益于AI技术，我们认为后续国内企业进一步深入应用AI，营收与估值均有较大提升空间。

根据wind数据，以2025/2/21日股价为准：

港股SW计算机分平均PS（TTM）为5.61；

A股SW计算机分类平均PS（TTM）为11.04；

美股纳斯达克计算机指数成分平均PS（TTM）为77.43。

具体从部分代表性公司PS估值来看，部分业务有相似性的公司对比也可以发现中国企业相对低估，如社交媒体巨头腾讯（PS 6.66）与脸书（PS 10.53）、电商与云计算巨头阿里巴巴（PS 2.48）与亚马逊（PS 3.6）

H股部分云计算公司PS（TTM） 以2025/2/21日股价为准

股票代码	公司	PS	股票代码	公司	PS
9988.HK	阿里巴巴	2.48	300378.SZ	鼎捷数智	4.68
3896.HK	金山云	5.29	300170.SZ	汉得信息	7.15
0700.HK	腾讯控股	6.66	603039.SH	泛微网络	7.94
1357.HK	美图公司	8.80	688369.SH	致远互联	3.74
0268.HK	金蝶国际	7.86	688111.SH	金山办公	34.99
2556.HK	迈富时	11.19	002153.SZ	石基信息	8.20
600588.SH	用友网络	6.11	002410.SZ	广联达	3.83

资料来源：Wind，国盛证券研究所

美股部分云计算和SAAS公司PS（TTM） 以2025/2/21日股价为准

股票代码	公司	PS	股票代码	公司	PS
MSFT.O	微软	11.59	NOW.N	SERVICENOW	17.60
AMZN.O	亚马逊	3.60	INTU.O	INTUIT	9.55
GOOGL.O	谷歌	6.26	APP.O	APPLOVIN	30.54
META.O	脸书	10.53	TEAM.O	ATLASSIAN	15.63
ORCL.N	甲骨文	8.54	WDAY.O	WORKDAY	8.36
SAP.N	思爱普	9.23	FICO.N	FAIR ISAAC	23.35
CRM.N	赛富时	7.97	HUBS.N	HUBSPOT	14.42
PLTR.O	PALANTIR	82.95	RDDT.N	REDDIT	23.15
ADBE.O	奥多比	8.99	MNDY.O	MONDAY.COM	15.44

资料来源：Wind，国盛证券研究所

目录

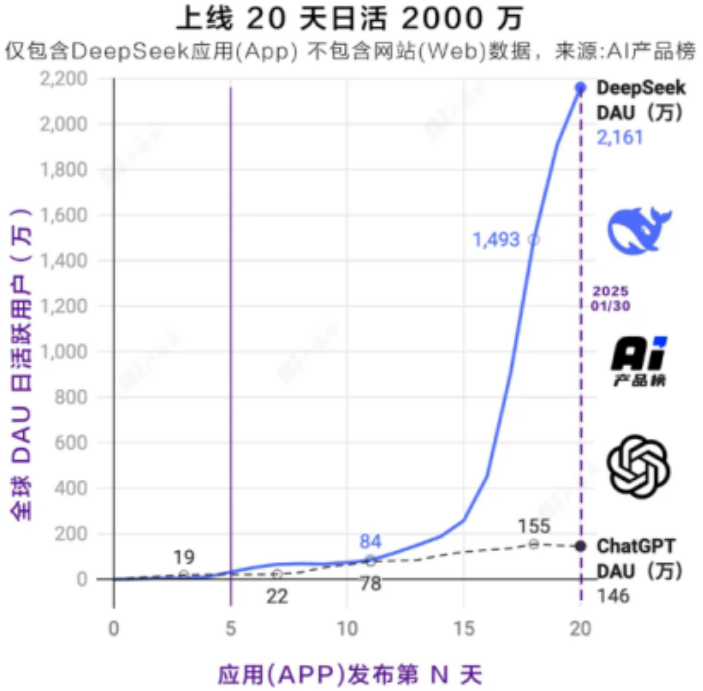
DeepSeek日活飙升，拉动巨大算力需求

DeepSeek凭借“开源、成本低、性能高”迅速出圈火爆全球，正在激活市场热情。

1月27日，Deepseek应用登顶苹果中国地区和美国地区应用商店免费APP下载排行榜，在美区下载榜上超越了ChatGPT。

据AI产品榜，DeepSeek APP上线20天日活突破两千万，成为全球增速最快AI应用。

1月DeepSeek官网多次显示，服务器繁忙/API不可用。DeepSeek回应称可能是由于新模型发布后，用户访问量激增，服务器一时无法满足大量用户的并发需求。



算力需求开始向推理倾斜

1月27日微软CEO在X上发帖引用“杰文斯悖论”，表示随着 AI 的效率和可访问性越来越高，我们将看到它的使用量猛增，将其变成我们无法满足的商品。

训练侧：业界预训练阶段Scaling边际提升放缓，目前主要精力放在后训练。整体训练算力需求短期是否会继续扩张有一定不确定性。

DeepSeek-V3总参数为 6710 亿参数，训练数据量14.8 万亿 token，Llama 405B训练数据15万亿token，随着互联网文本数据的耗尽，预训练阶段的Scaling law 出现放缓趋势。DeepSeek-V3提出预训练阶段多项优化措施。

后训练阶段，DeepSeek-R1 Zero提出只强化学习（RL)不需要监督微调（SFT）也可以取得较好效果，但是可读性较差。R1增加了部分高质量思维链数据做监督微调。

推理侧：
1.模型进步有助于应用落地，调用量提升
2月3日DeepSeek在全球140个市场移动下载量排行榜中位列榜首，自1月28日以来在谷歌Play Store美国区下载量达1600万次，超过ChatGPT同期表现。

2.DeepSeek R1这类推理模型回答时要输出大量token作为思考过程，需要较多算力。

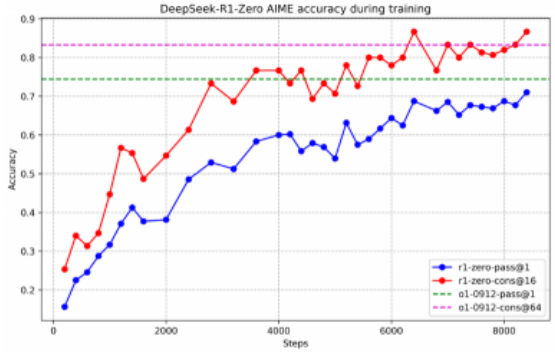


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

强化学习增加训练时间依然可以看到效果显著提升

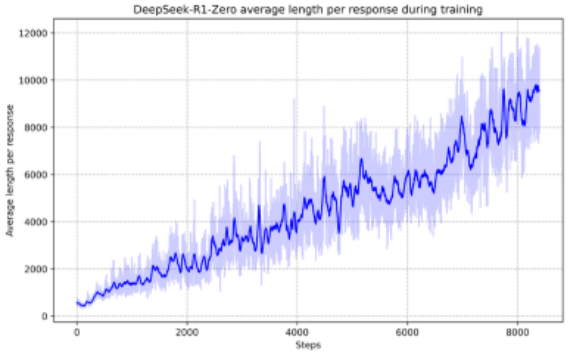


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

强化学习过程中模型的平均响应长度逐渐变长

算力需求开始向推理倾斜

算力需求开始向推理倾斜，豆包模型日均Token已达数万亿。1）技术层面来看，推理阶段优化是下一个发力点。OpenAI联合创始人Ilya Sutskever指出，通过增加数据和计算能力来提升预训练模型的效果已达到瓶颈。当前，OpenAI在o1模型中采用了“测试时计算”（test-time compute）技术，允许模型在推理阶段进行多步推理，类似于人类的思考过程。其他AI厂商，如Anthropic、xAI和Google DeepMind也在开发类似的技术，通过优化推理阶段来提升模型性能。2）AI应用端来看，随着大模型在应用侧规模化部署，算力需求已向推理端倾斜。2024年12月18日举办的火山引擎FORCE大会上，火山引擎CEO谭待表示，截至目前，豆包大模型日均tokens使用量超过4万亿，较5月发布时期增长超过33倍，tokens使用量直接反映了模型的广泛应用和市面需求，大模型应用正在向各行各业加速渗透。自2024年9月至12月，豆包大模型在信息处理场景的调用量增长了39倍，客服与销售场景增长16倍，硬件终端场景增长13倍，AI工具场景增长9倍，学习教育等场景也有大幅增长。

图表1：*豆包日均tokens数量*

资料来源：火山引擎微信视频号，国盛证券研究所

图表2：*豆包大模型应用加速渗透*

资料来源：火山引擎微信视频号，国盛证券研究所

算力产业链各环节均迎增长机遇

算力产业链上游为算力基础硬件设施，主要包括元器件、ICT基础设施、其他硬件设备等。中游为算力网络与平台，上游硬件设备及基础设施共同组成数据中心、算力网络等，提供IDC服务、云计算服务、以及各类算力网络服务等。产业链下游则为应用场景与用户。

中国AI算力产业链结构，资料来源：前瞻产业研究院

美国星际之门计划打造AI基建

2025年1月美国总统特朗普在白宫宣布成立了一家新公司，名为“星际之门计划（Stargate Project）”，计划在未来四年内投资5000亿美元，在美国为OpenAI建设新的人工智能基础设施。他称其为”迄今为止历史上最大的人工智能基础设施项目”，并表示这将有助于将“技术的未来”留在美国。OpenAI将立即开始部署1000亿美元。星际之门的初始股权出资者包括软银、OpenAI、甲骨文和中东全球投资集团MGX。软银和OpenAI是星际之门的主要合作伙伴，软银负责财务，OpenAI负责运营。孙正义将担任董事长。

Arm、微软、NVIDIA、Oracle和OpenAI是主要的初始技术合作伙伴。扩建工作目前正在进行中，从德克萨斯州开始，OpenAI正在评估全国各地的潜在地点，以建立更多园区，并最终确定最终协议。作为星际之门的一部分，Oracle、NVIDIA和OpenAI将密切合作，共同构建和运营该计算系统。这建立在OpenAI和NVIDIA自2016年以来的深度合作以及OpenAI和Oracle之间较新的合作伙伴关系的基础之上。这也建立在OpenAI与微软现有的合作关系之上。随着OpenAI继续与微软合作，利用额外的计算能力来训练领先模型并提供出色的产品和服务，OpenAI将继续增加对Azure的使用。

微软计划今年投资800亿美元建设以人工智能为重点的数据中心。它还参与了包括贝莱德和MGX在内的1000亿美元的合资企业，专注于进行AI数据中心投资。亚马逊也一直在向这些中心投入类似规模的资金，仅在过去两个月就宣布了两个价值约100亿美元的项目。

特朗普与 OpenAI、甲骨文和软银的老板一起宣布了星际之门项目

中国智算规模两年有望翻倍

DeepSeek凭借“开源、成本低、性能高”迅速出圈火爆全球，正在激活市场热情。

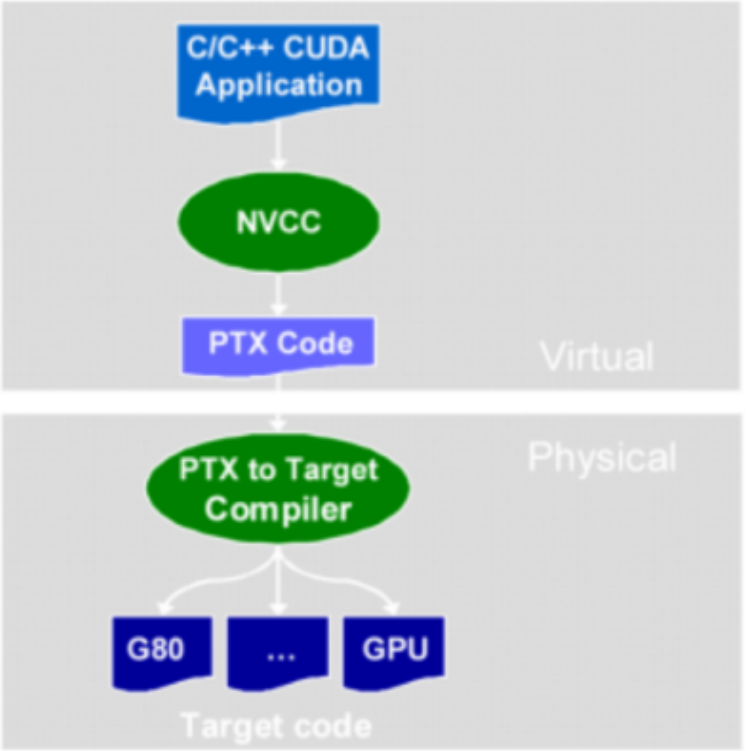
2月13日，国际数据公司IDC与浪潮信息联合发布《2025年中国人工智能计算力发展评估报告》指出，大模型和生成式人工智能推高算力需求，中国智能算力增速高于预期。

2024年，中国智能算力规模达725.3EFLOPS，同比增长74.1%，增幅是同期通用算力增幅（20.6%）的3倍以上；市场规模为190亿美元，同比增长86.9%。未来两年，中国智能算力仍将保持高速增长。**2025年，中国智能算力规模将达到1,037.3 EFLOPS，较2024年增长43%；2026年，中国智能算力规模将达到1,460.3 EFLOPS，为2024年的两倍**。2025年中国人工智能算力市场规模将达到259亿美元，较2024年增长36.2%；2026年市场规模将达到337亿美元，为2024年的1.77倍。。

国内大模型团队具备针对硬件从较底层架构上调优的能力

DeepSeek使用PTX从更底层调用硬件:
DeepSeek-V3定制了高效的跨节点全对全通信内核（包括分发与合并），以减少专门用于通信的流多处理器（SM）数量。DeepSeek采用定制的并行线程执行（PTX）指令，并自动调整通信块大小，这显著减少了二级缓存的使用以及对其他 SM 的干扰。

豆包的Doubao-1.5-pro也在工程架构上进行了细致优化，据豆包公众号，豆包凭借自研服务器集群方案，灵活支持低成本芯片，硬件成本比行业方案大幅度降低。通过定制化网卡和自主研发的网络协议，显著优化了小包通信的效率。



CUDA（类似C++的高级语言）
编译为PTX(类似汇编的低级语言）
最后编译为机器码(0和1)

算力国产化进程加快

DeepSeek与开源社区和硬件供应商合作，提供多种方式在本地运行模型，大模型厂商和硬件厂商适配利好推理算力百花齐放

6. 如何在本地运行

DeepSeek-V3 can be deployed locally using the following hardware and open-source community software:

1. **DeepSeek-Infer Demo**: We provide a simple and lightweight demo for FP8 and BF16 inference.
   DeepSeek-Infer演示：我们为FP8和BF16推理提供了一个简单而轻量级的演示。
2. **SGLang**: Fully support the DeepSeek-V3 model in both BF16 and FP8 inference modes, with Multi-Token Prediction coming soon.
   SGLang：在BF16和FP8推理模式下完全支持DeepSeek-V3模型，即将推出多令牌预测。
3. **LMDeploy**: Enables efficient FP8 and BF16 inference for local and cloud deployment.
   LMDeploy：为本地和云部署提供高效的FP8和BF16推理。
4. **TensorRT-LLM**: Currently supports BF16 inference and INT4/8 quantization, with FP8 support coming soon.
   TensorRT-LLM：目前支持BF16推理和INT4/8量化，很快将支持FP8。
5. **vLLM**: Support DeepSeek-V3 model with FP8 and BF16 modes for tensor parallelism and pipeline parallelism.
   vLLM：支持DeepSeek-V3模型，FP8和BF16模式，用于张量并行和管道并行。
6. **AMD GPU**: Enables running the DeepSeek-V3 model on AMD GPUs via SGLang in both BF16 and FP8 modes.
   AMD GPU：允许在AMD GPU上通过SGLang在BF16和FP8模式下运行DeepSeek-V3模型。
7. **Huawei Ascend NPU**: Supports running DeepSeek-V3 on Huawei Ascend devices.
   华为Ascend NPU：支持在华为Ascend设备上运行DeepSeek-V3。

Since FP8 training is natively adopted in our framework, we only provide FP8 weights. If you require BF16 weights for experimentation, you can use the provided conversion script to perform the transformation.

相比训练端，推理端对芯片性能及集群互联的要求显著降低，有望盘活国产算力资产。为国产芯片、英伟达中低端卡以及ASIC芯片等国内算力提供了有利的发展机遇。

国产芯片：Deepseek模型框架的优化与需求的推理端转移显著降低了硬件算力需求，利好国产AI芯片在推理端快速实现商业化落地。目前国内芯片厂商已与DeepSeek合作，加速深度学习框架优化和分布式训练适配，推动“国产算力+国产大模型”闭环生态的构建。包括华为昇腾、海光、沐曦、天数智芯、摩尔线程、壁仞、燧原、昆仑芯、云天励飞、灵汐科技、鲲云在内的多家国产芯片厂商纷纷宣布完成对DeepSeek系列模型的适配。

英伟达中低端卡：2025年1月13日美国对华AI芯片出口管制再次加码，对包括中国在内的120个国家中数据中心、人工智能产品所使用的芯片进行限制。国内市场需求量将仍以英伟达中低端显卡为主，且保有较多存量，Deepseek的出现有望盘活中低端显卡市场。

ASIC芯片：ASIC芯片相比通用GPU定制化程度更高，规模量产后，在特定场景下ASIC的单位成本或可更低。以谷歌为代表的海外大厂在ASIC方面持续迭代。我国ASIC行业研发基础扎实且发展势头强劲，如寒武纪、澜起科技、黑芝麻、地平线、华为海思、百度以及阿里巴巴等均已布局ASIC产品，且部分国产ASIC技术已经达到国际领先水平。ASIC特定场景下（如推理端）具有高性价比，利好国内芯片厂商。

国产芯片核心厂商：寒武纪

■ **寒武纪：国产AI芯片龙头，云边端一体协同**

➢ 寒武纪成立于2016年，专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片，让机器更好地理解和服务人类。寒武纪能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。寒武纪秉承开放、合作、共赢的态度与客户紧密协作，为互联网、能源、金融等行业的智能化升级提供充裕的底层算力支撑。兼容“CUDA”生态优势明显，与主流大模型厂商适配良好。海光CPU兼容x86指令集，支持国内外主流操作系统、数据库、虚拟化平台或云计算平台，能够有效兼容目前存在的数百万款基于x86指令集的系统软件和应用软件，具有优异的生态系统优势；海光DCU兼容“CUDA”生态，对文心一言等大多数国内外主流大模型适配良好。依托DCU可以实现LLaMa、GPT、Bloom、ChatGLM、悟道、紫东太初等为代表的大模型的全面应用，从大模型生态上DCU已经达到国内领先水平。同时，公司与多家国内知名服务器厂商达成战略合作，并加速以海光为核心的“光合组织”生态建设。

国产芯片核心厂商：华为

■ **华为：打造全栈自主 AI 基础软硬件**

昇腾计算：全栈AI基础设施及应用服务

- 昇腾计算产业基于昇腾系列处理器和基础软件，构建全栈AI计算基础设施、行业应用及服务，包括系列处理器、系列硬件、CANN、AI计算框架、应用使能、开发工具链、管理运维工具、行业应用及服务等。
- 昇腾310和910处理器为华为AI算力领域核心产品，基于达芬奇架构，覆盖端边云全场景，可满足不同部署环境差异性的算力需求，从算力和功耗来看，目前昇腾910单卡Int8算力大致可达0.6 P，最大功耗为300W。
- 基于昇腾910和310 AI处理器，昇腾计算产业在硬件方面坚持“硬件开放”策略，通过自有硬件和合作硬件相结合的方式为客户提供多样化的算力选择。其中，自有硬件主要为Atlas系列硬件产品，包含模组、板卡、小站、服务器、集群等多个产品形态。

图表：主流 GPGPU/AISC 产品性能对比

厂商	产品型号	制程	FP64 (TFLOPS)	FP32 (TFLOPS)	INT8 (TOPS)	互联带宽	显存	接口	功耗
英伟达	**A100 PCIe**	**7nm**	**9.7**	**19.5**	**624**	**600GB/s**	**80GB**	**PCIe**	**300W**
英伟达	A800 PCIe	7nm	9.7	19.5	624	400GB/s	80GB	PCIe	300W
英伟达	Tesla V100 PCIe	12nm	7.8	15.7	-	300GB/s	80GB	PCIe	300W
AMD	Instinct MI250X	6nm	47.9	47.9	383	100GB/s	128GB	PCIe	560W
AMD	Instinct MI250	6nm	45.3	45.3	362.1	100GB/s	128GB	PCIe	560W
AMD	Instinct MI100	7nm	11.5	23.1	184.6	92GB/s	32GB	PCIe	300W
天数智芯	天垓 100	7nm	-	37	295	64GB/s	32GB	PCIe	250W
壁仞科技	壁砺 100P	7nm	-	240	1920	448GB/s	64GB	PCIe	550W
壁仞科技	壁砺 104P	7nm	-	256	1024	192GB/s	32GB	PCIe	300W
Google	TPUv4i	7nm	-	-	138	300GB/s	8GB	-	-
Google	TPUv4	7nm	-	-	275	1200GB/s	32GB	-	-
华为海思	昇腾 310	12nm	-	-	16	-	-	-	8W
华为海思	**昇腾 910**	**7nm**	**-**	**-**	**640**	**-**	**-**	**-**	**310W**
燧原科技	T20	-	-	32	256	300GB/s	32GB	PCIe	300w
寒武纪	MLU370-X4	7nm	-	24	256	307.2GB/s	24GB	PCIe	150W
寒武纪	MLU3370-S4	7nm	-	18	192	307.2GB/s	24GB	PCIe	75W

资料来源：英伟达官网、TechPowerUp、Unixcloud、AMD 官网、天数智芯官网、云和官网、壁仞科技官网、NextPlatform、海思官网、燧原科技官网、寒武纪官网、国盛证券研究所

芯片	制程	单价	当前产能	备注
鲲鹏920	7nm	平均1.5	台积电，已停产	
鲲鹏920B	28nm堆叠		无限量供应	性能持平，功耗高30%
昇腾310	12nm	0.7	逐步停产	
昇腾310B	28nm堆叠		无限量供应	
昇腾710	14nm	2	无限量供应	
昇腾910	7nm	8-9	台积电，逐步停产	
昇腾910B	14nm堆叠	8-9	中芯南方，月产1.7万	FP32：75 TFLOP
昇腾610	7nm	-	-	

国产芯片核心厂商：海光信息

■ **海光信息：DCU 开放生态加速推广**

海光信息为我国芯片产业领军企业之一，长期以来深入布局国产芯片，产品主要包括 CPU 和 DCU 两大类。

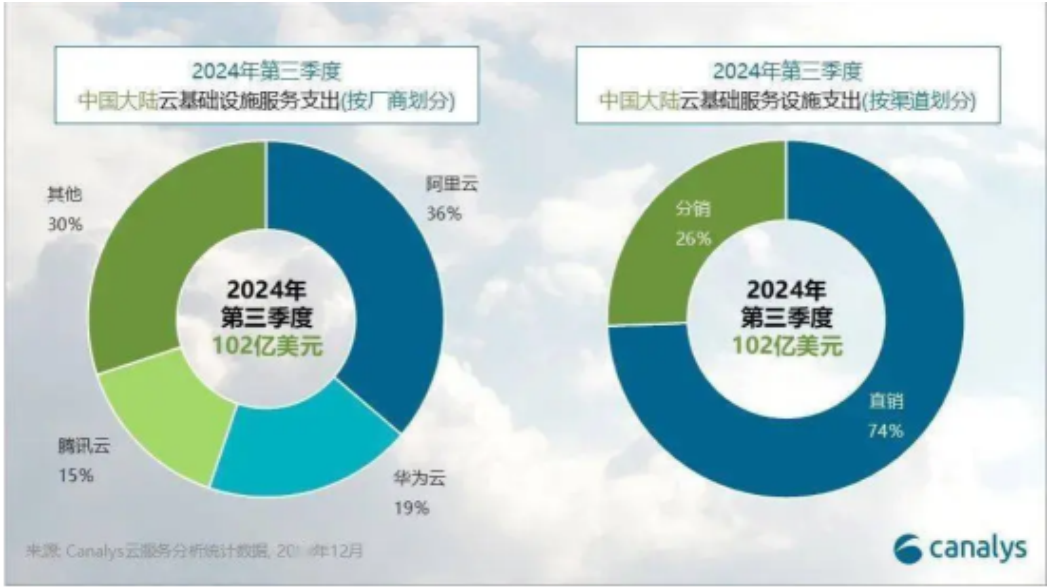
- 兼容“CUDA”生态优势明显，与主流大模型厂商适配良好。海光CPU兼容x86指令集，支持国内外主流操作系统、数据库、虚拟化平台或云计算平台，能够有效兼容目前存在的数百万款基于x86指令集的系统软件和应用软件，具有优异的生态系统优势；海光DCU兼容“CUDA”生态，对文心一言等大多数国内外主流大模型适配良好。依托DCU可以实现LLaMa、GPT、Bloom、ChatGLM、悟道、紫东太初等为代表的大模型的全面应用，从大模型生态上DCU已经达到国内领先水平。同时，公司与多家国内知名服务器厂商达成战略合作，并加速以海光为核心的“光合组织”生态建设。
- 产品迭代或加速，新款CPU、GPU核心性能大幅提升。海光信息秉持 “销售一代、验证一代、研发一代”的研发策略。CPU：海光三号目前为主力销售产品，实测性能较上一代产品提升约45%，产品对标海外一线厂商能力，国内厂商中领先；DCU：深算一号为主力销售产品，深算二号为2023Q3发布并商用，性能较前代翻倍。深算三号研发进展顺利。

产品类型	处理器种类	指令集	主要产品	产品特征	典型应用场景
海光 CPU	通用处理器	兼容 x86 指令集	海光 3000 系列 海光 5000 系列 海光 7000 系列	内置多个处理器核心，集成通用的高性能外设接口，拥有完善的软硬件生态环境和完备的系统安全机制，适用于数据计算和事务处理等通用型应用	云计算、物联网、信息服务等
海光 DCU	协处理器	兼容“类 CUDA”环境	海光 8000 系列	内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用	大数据处理、人工智能、商业计算等

云厂商深度受益，价值有望重塑

据财学堂，随着大模型DeepSeek的热度持续飙升，其生态合作版图加速扩张。华为云、腾讯云、阿里云、百度云等国内头部云厂商近期密集宣布支持部署DeepSeek模型，标志着这场AI技术革命正在深刻重构云计算产业格局。

云厂商通过大模型生态，不仅能够吸引更多企业客户使用其算力服务，更可能颠覆传统以资源租赁为主的盈利模式，推动行业进入“AI+云”双轮驱动的高质量发展阶段。国内云厂有望迎来利润率和规模双击的趋势，价值有望迎来重塑



据Canalys数据显示，2024年第三季度，中国大陆的云基础设施服务支出总额跃升至102亿美元，与前一年同期相比增长了11%，再次迈入两位数的增长轨道。

尽管市场竞争激烈，但中国的前三大云服务提供商依然稳固地占据着头把交椅。阿里云、华为云以及腾讯云三者合计占据了70%的市场份额，继续引领行业发展。中国电信等运营商也在积极探索新的路径，以期在云服务市场中占据一席之地

供给端管控趋向严格，或利好云上游议价能力提升。

1）政策或导向数据中心能耗指标缩紧，具有能耗指标或已建成机房储备的IDC厂商稀缺性上升，有望进一步带来议价能力上升，由于推理需求需要考虑延迟问题，尤其京津冀区域的IDC或将深度受益。3月18日，国家发展改革委等五部门发布关于促进可再生能源绿色电力证书市场高质量发展的意见，提出依法稳步推进绿证强制消费，逐步提高绿色电力消费比例并使用绿证核算。加快提升钢铁、有色、建材、石化、化工等行业企业和数据中心，以及其他重点用能单位和行业的绿色电力消费比例，到2030年原则上不低于全国可再生能源电力总量消纳责任权重平均水平，国家枢纽节点新建数据中心绿色电力消费比例在80%基础上进一步提升

2）美国商务部近期将50余个中国科技企业和机构纳入所谓的“实体清单”，包括一系列与中国AI大模型开发、服务器以及超级计算机产业的12家公司，美国打压也可能将令存量算力设备以及国产算力重要性进一步提升。

3）国内算力建设政策重视程度也逐渐加大：
3月26日上海市经济和信息化委员会印发 《上海市关于促进智算云产业创新发展的实施意见（2025-2027年）》提出到2027年，本市智算云产业规模力争突破2000亿元，智算规模要力争达到200EFLOPS，其中自主可控算力占比超70%。

海外云大厂收入与资本开支持续增长

海外科技巨头业绩超预期，持续加大AI基建支出。

1）谷歌：2025年第一季度营收902.3亿美元，净利润345亿美元，均超预期。一季度谷歌云计算部门的收入同比增长28%达123亿美元。谷歌将维持今年2月公布的资本支出计划，即谷歌2025年全年资本支出达到750亿美元，用于建设数据中心等项目，较2024年的530亿美元显著增加。

2）微软：截至3月31日的2025财年第三财季财报营收为700.66亿美元，同比增长13%；净利润为258.24亿美元，同比增长18%，在云计算业务Azure强劲增长加持下业绩超过分析师预期。其中智能云业务事业部营收为267.51亿美元，较上年同期的221.41亿美元增长21%。剔除财务租赁的资本支出达167.5亿美元，同比增长近53%。2026财年微软预计资本支出将继续增长，但增速将低于2025财年。

3）Meta：2025年第一季度营收为423.14亿美元，同比增长16%；净利润为166.44亿美元，同比增长35%。首席执行官扎克伯格表示目前Meta AI已拥有近10亿月度活跃用户。上调资本开支指引：Meta预计2025年全年资本支出将达到640亿至720亿美元，较此前预期的600亿至650亿美元有所增加。

4）亚马逊：2025年第一季度财报实现了1556.67亿美元收入额，同比增长9%；净利润为171.27亿美元，同比增长64%，一季度资本支出250.2亿美元，同比增长约67.6%