一体机行业深度：驱动因素、行业现状、产业链及相关公司深度梳理

随着人工智能技术的迅猛发展，大模型已成为推动产业变革的关键力量。然而，大模型的应用落地面临着算力不足、部署复杂、数据安全等诸多挑战。在此背景下，一体机应运而生，以其软硬件深度融合、开箱即用、私有化部署等优势，为政企客户提供了便捷高效的 AI 解决方案，有望引领 AI 产业进入新的发展阶段。

本报告将深入剖析一体机，首先简述一体机的概念、优势等基本内容，接着分析驱动行业发展的关键因素及当前行业现状。在此基础上，以 DeepSeek 一体机为例，详细分析其部署过程，探讨相关产业链环节及重点公司。最后，我们将展望未来市场趋势。通过这些内容的系统阐述，旨在为读者提供全面的行业洞察和决策参考。

目录

一、行业概述

1、什么是 AI 大模型一体机

AI 大模型一体机是专为人工智能大模型应用和部署而设计的集成计算设备，本质上可以看作是 AI 服务器与大模型私有化部署的融合。其作为“软硬协同、开箱即用”的智能化基础设施，具备高效耦合计算、存储、网络等硬件设备、大模型微调部署软件平台和预置大模型。

高性能硬件：大模型一体机通过适配专用 AI 芯片，针对大模型算法进行深度优化，可以充分释放 AI 芯片性能；通过配置大容量内存和高速存储，支持大模型的加载和运行，提高数据读写速度。

基座大模型：目前的大模型一体机可以提供包括 DeepSeek 系列、LLaMA 系列、Baichuan 系列、Qwen 系列等在内的多种主流开源大模型，企业用户可以根据特定应用场景需求，对预置模型进行微调及增量训练，降低 AI 大模型的落地门槛。

全栈工具链：大模型一体机通过集成 AI 全流程开发工具，实现从数据处理、模型训练到推理部署的全栈式开发，提高模型训练效率。同时，通过外挂用户专属知识库，结合检索增强技术，实现专业领域的知识问答，为企业提供定制化、便捷化、场景化的 AI 服务；通过可视化管理工具，实现硬件组网、资源监控、故障定位清晰可见，降低运维门槛。

图表1：AI 大模型一体机主要构成

资料来源：云计算开源产业联盟，太平洋证券整理

2、云计算和一体机部署大模型的特点对比

相较于云部署模式，采用一体机对大模型进行私有化部署具有以下优点：

高稳定性：相比公有服务器容易受到巨大流量的冲击，私有算力显然更有保障，稳定性更好。

简化部署：传统大模型部署需经历硬件调试、框架适配、算子优化等流程，一体机预装的大模型和配套工具链极大地降低了企业使用门槛，真正实现开箱即用，缩短了项目部署周期，助力企业快速开启 AI 应用实践。

模型定制化：企业可以通过私有数据持续训练模型，或搭载内部知识库，让通用大模型转化为垂直领域的“专家”，适配企业特定业务场景。

经济效应：短期看通过云服务使用大模型无需一次性硬件投入，但长期使用公有云 API 按 token 付费成本较高，通过一体机私有化部署有助于降低总体成本并更好地掌控预算。

数据安全：对于一些具备大量敏感数据的用户，部署本地化大模型可以不联网使用，确保数据在本地处理，避免敏感信息外流，能够满足金融、能源、政务、医疗等对于等数据敏感型行业对于安全和隐私的要求。

图表2：云计算和一体机部署大模型的特点对比

	传统云计算服务	大模型一体机
数据敏感性	需上传云端，存在泄露风险	数据本地处理，安全性高
网络依赖	依赖稳定网络连接	本地运行，低延迟、离线可用
部署成本	按需付费，长期成本可能较高	一次性投入，适合长期高频使用
定制化需求	灵活性高，但需自行优化	预置行业解决方案，开箱即用

资料来源：公开资料，太平洋证券整理

3、央国企及党政机关是一体机的重要客户

政府机构和央国企往往涉及公民信息、政务数据、国家安全等大量敏感数据信息，对本地化、私有化部署要求较高。智能一体机低门槛、低部署成本的私有化部署方案完美契合相关需求，不需要额外部署服务器、雇佣庞大运维团队，仅需支付购买费用，购买后立即就能投入使用。

表1：部分党政机关一体机部署情况

地区	机关	部署情况
北京市	应急管理部大数据中心	部署“工业互联网+安全生产”数据分析决策与应用处置系统项目-多模态训推一体机
深圳市	深圳市龙岗区政府 深圳市南山区政府	云天天书大模型训推一体机已经在深圳市龙岗区、南山区实现双区部署

资料来源：中国政府采购网、云天励飞官网、浙商证券研究所

表2：部分央国企一体机部署情况

企业	服务商	部署情况
中国石化	天翼云、中国电信	全尺寸 DeepSeek-R1(671B 版)大模型国产化部署，推理效率提升近一倍
多家央国企及高校	天翼云、中国电子云等	DeepSeek 智算一体机部署，覆盖国产芯片、推理引擎到模型服务全栈国产化需求

资料来源：中华网、财联社、浙商证券研究所

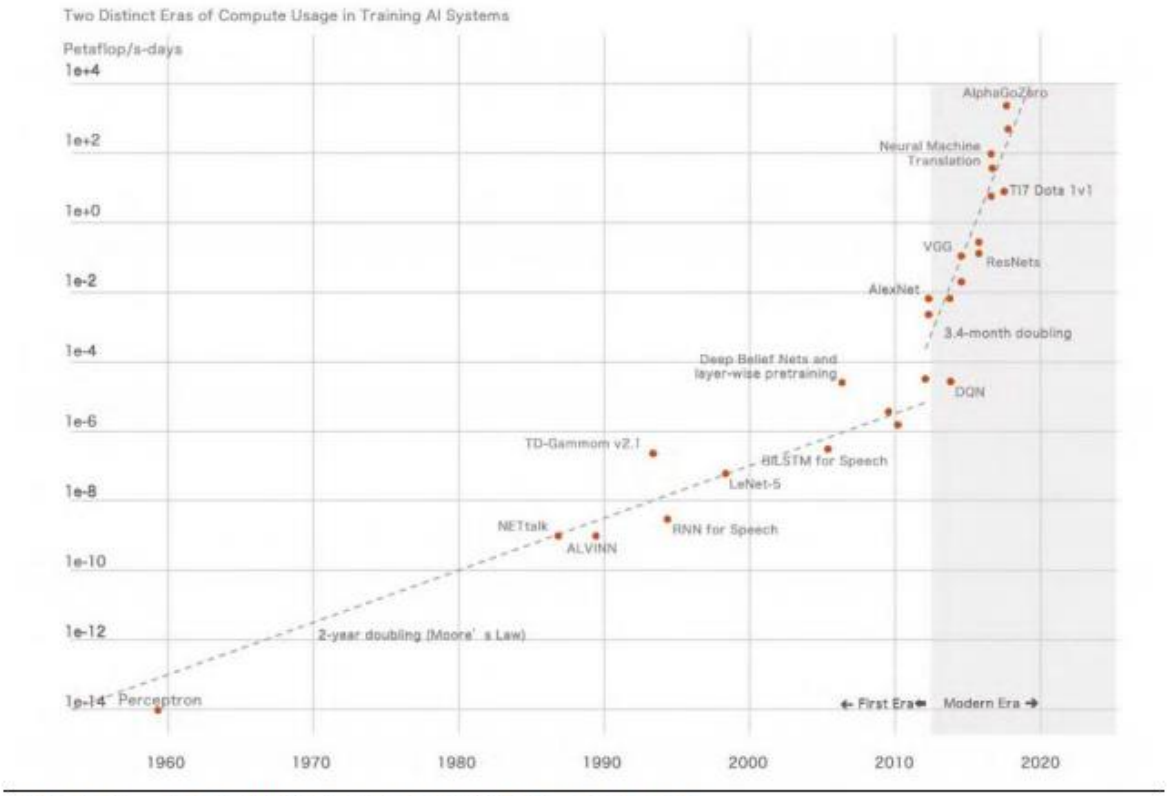
二、驱动因素

AI 计算需求爆发+政策指引+数据安全考虑，推出一体机大势所趋。

1、AI 计算需求呈爆发式增长，预计 2023-2028 年我国智能算力 CAGR 有望达 46.2%

AI 计算的需求爆发式增长。据预测 2025 年全球企业对 AI 的采用率将达 86%，企业数据利用率将剧增至 80%左右。据《昇腾计算产业发展白皮书》统计，AI 模型的规模和需要学习的数据开始爆炸性的增长，从 2012 年开始的 6 年中，AI 计算的需求增加了 30 万倍。

图 1：训练 AI 系统计算使用的两个不同时代



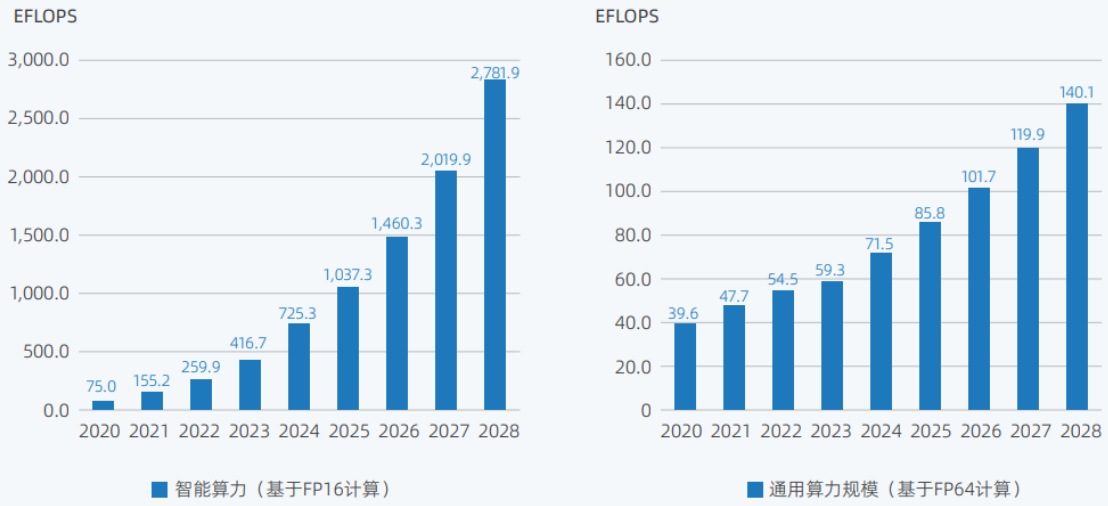
数据来源：《昇腾计算产业发展白皮书》，东北证券

图片、语音、视频等非结构化数据采用 AI 算力处理的效率远远高于通用算力。伴随 5G、智慧城市、物联网等相关领域多样化应用的普及，海量文本、图片、语音、视频等非结构化数据的生成速度不断加快，处理需求呈现指数级增长，而这些非结构化数据采用 AI 算力处理的效率远远高于通用算力。最近，随着算法的持续增强，大规模预训练模型，如 BERT,GPT-3 等，需要的算力从 TFLOPS 级别增加到 PFLOPS 级别，甚至开始进入 EFLOPS 级别。同时，超大规模的批处理、自动模型结构搜索等新方法的涌现，导致 AI 计算需求持续增加。

2025 年中国智能算力规模将达到 1,037.3EFLOPS，预计到 2028 年将达到 2,781.9EFLOPS。2025 年中国通用算力规模将达到 85.8EFLOPS，预计到 2028 年将达到 140.1EFLOPS。预测显示，2023-2028 年

期间，中国智能算力规模的五年年复合增长率预计达到 46.2%，通用算力规模预计达到 18.8%。较上一版本预期值 33.9%和 16.6%，均有显著提升。

图5 中国智能算力和通用算力规模及预测，2020-2028



来源：IDC，2025

2、政策明确 AI 为新质生产力重要引擎，国央企加大算力基础设施建设

国务院国资委中央企业要把发展人工智能放在全局工作中统筹谋划。2024 年 2 月 21 日，国务院国资委召开“AI 赋能产业焕新”中央企业人工智能专题推进会。会议认为，（1）加快推动人工智能发展，是国资央企发挥功能使命，抢抓战略机遇，培育新质生产力，推进高质量发展的必然要求。中央企业要主动拥抱人工智能带来的深刻变革，把加快发展新一代人工智能摆在更加突出的位置，不断强化创新策略、应用示范和人才聚集。加快构建数据驱动、人机协同、跨界融合、共创分享的智能经济形态。会议强调，（2）中央企业要把发展人工智能放在全局工作中统筹谋划，深入推进产业焕新，加快布局和发展人工智能产业。加快建设一批智能算力中心，进一步深化开放合作，更好发挥跨央企协同创新平台作用。（3）开展 AI+专项行动，强化需求牵引，加快重点行业赋能，构建一批产业多模态优质数据集，打造从基础设施、算法工具、智能平台到解决方案的大模型赋能产业生态。

能够认为，国务院国资委召开“AI 赋能产业焕新”中央企业人工智能专题推进会，有望推动中央企业在人工智能领域发挥更大作用，从数据、算法、算力多个维度加速我国人工智能产业建设。

表 1：2021-2023 年我国智能计算促进政策

时间	政策	主要内容
2023 年 12 月	《深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》	到 2025 年底，综合算力基础设施体系初步成型。国家枢纽节点地区各类新增算力占全国新增算力的 60%以上，国家枢纽节点算力资源使用率显著超过全国平均水平；1ms 时延城市算力网、5ms 时延区域算力网、20ms 时延跨国家枢纽节点算力网在示范区域内初步实现
2023 年 2 月	《数字中国建设整体布局规划》	系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局。
2023 年 1 月	《关于推动能源电子产业发展的指导意见》	面向新型电力系统和数据中心、算力中心、电动机械工具、电动交通工具及充换电设施、新型基础设施等重点终端应用，开展能源电子多元化试点示范。
2023 年 1 月	《关于促进数据安全产业发展的指导意见》	推动先进适用数据安全技术产品在电子商务、远程医疗、在线教育、线上办公、直播新媒体等新型应用场景，以及国家数据中心集群、国家算力枢纽节点等重大数据基础设施中的应用。
2022 年 2 月	《关于同意京津冀地区启动建设全国一体化算力网络国家枢纽节点的复函》	同意在京津冀地区启动建设全国一体化算力网络国家枢纽节点，发展高密度、高能效、低碳数据中心集群。积极承接北京等地实时性算力需求，引导温冷业务向西部迁移。
2022 年 1 月	《“十四五”数字经济发展规划》	加快构建算力、算法、数据、应用资源协同的全国一体化大数据中心体系。建设数据中心集群，结合应用、产业等发展需求优化数据中心建设布局。
2021 年 11 月	《“十四五”软件和信息技术服务业发展规划》	前瞻布局新兴平台软件，加快培育云计算、大数据、人工智能、5G、区块链、工业互联网等领域具有国际竞争力的软件技术和产品。
2021 年 11 月	《“十四五”大数据产业发展规划》	加快构建全国一体化大数据中心体系，推进国家工业互联网大数据中心建设，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群。建设高性能计算集群，合理部署超级计算中心。
2021 年 7 月	《新型数据中心发展三年行动计划(2021-2023 年)》	用 3 年时间，基本形成布局合理、技术先进、绿色低碳、算力规模与数字经济增长相适应的新型数据中心发展格局。技术能力明显提升，产业链不断完善，国际竞争力稳步增强。
2021 年 3 月	《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》	加快构建全国一体化大数据中心体系，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群，建设 E 级和 10E 级超级计算中心。
2021 年 1 月	《工业互联网创新发展行动计划（2021-2023 年）》	推动工业互联网大数据中心建设，打造工业互联网大数据中心综合服务能力，到 2023 年基本建成国家工业互联网大数据中心体系，建设 20 个区域级分中心和 10 个行业级分中心。

数据来源：中商产业研究院、中国经济网，东北证券

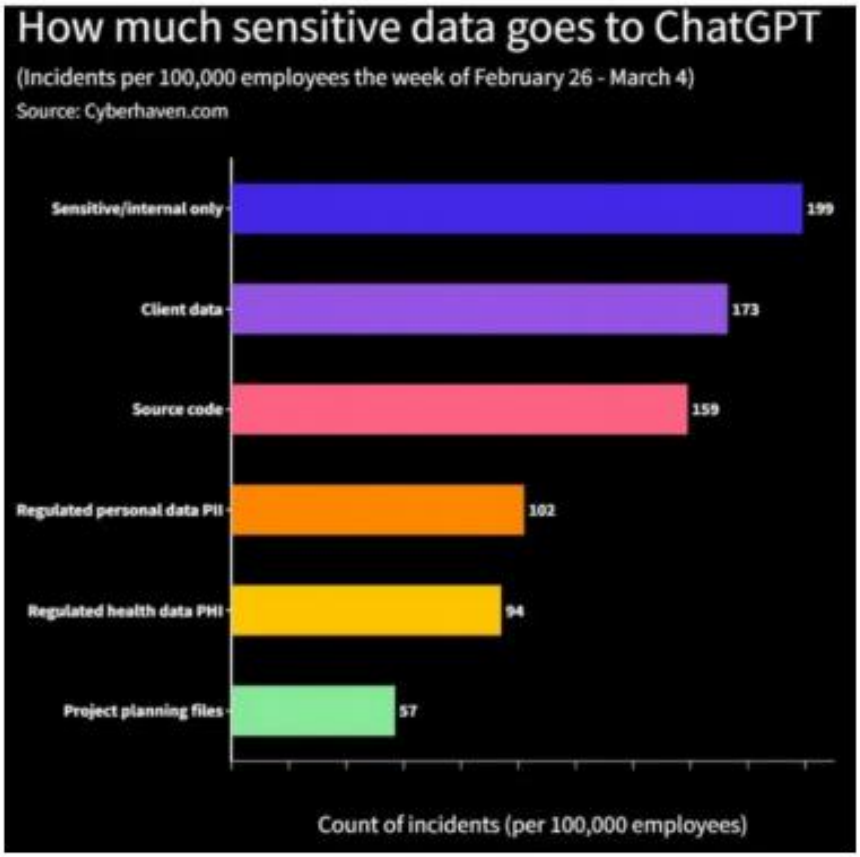
3、使用外部模型或造成数据泄露，AIGC 时代下数据安全引担忧

AI 浪潮下数据安全成为全球焦点之一，诸多公司和国家对外部人工智能应用的使用采取限制措施。例如，JPMorganChase 出于对第三方软件控制及财务信息安全性的考虑，限制了其员工使用 ChatGPT。此外，全球领先的半导体制造商台积电近期也指示其员工在使用 ChatGPT 时，禁止泄露公司敏感信息

以保护个人隐私。同样，意大利的个人数据保护机构也对 ChatGPT 采取了禁令，并限制其开发商 OpenAI 处理意大利用户的数据。

三星引进 ChatGPT 后已遭遇了三次数据泄露事件，问题的根源在于内部员工将公司机密通过提问形式输入 ChatGPT。导致自从三星引进 ChatGPT 以来，公司遭遇了三次数据泄露事件，其中两起涉及半导体设备，一起与内部会议相关。报道称，这些泄露事件导致重要的半导体设备测量数据和产品质量率信息直接被传送给了一家美国公司。韩国媒体指出，问题的根源在于三星的员工将公司机密通过提问形式输入 ChatGPT，使得这些机密信息有可能被集成进学习数据库并被广泛传播。为了防止此类事件再次发生，三星已经指示其员工在使用 ChatGPT 时必须格外小心。三星还表示，如果类似的情况继续出现，公司将考虑禁止在其内部网络上使用 ChatGPT。

图 4：敏感数据泄露给 ChatGPT 的数量



数据来源：Cyberheaven、东北证券

三、行业现状

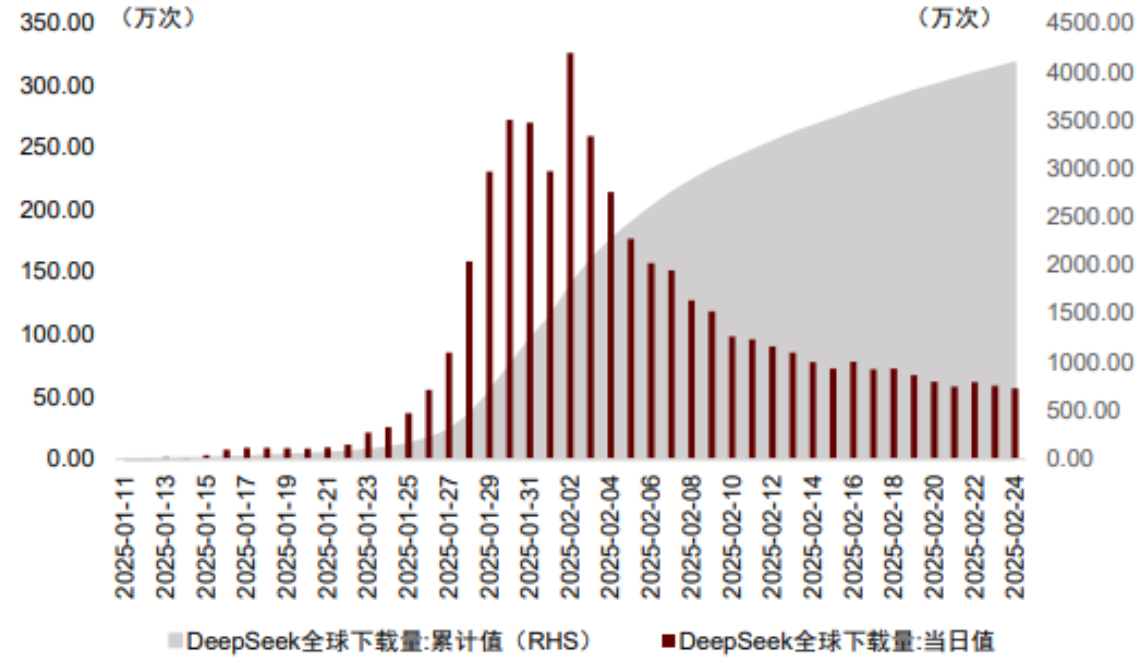
1、DeepSeek 开源大模型推动私有化部署新趋势

中国企业 DeepSeek（简称 DS）全面开源的创新成果引发了市场对生成式 AI 技术发展与算力硬件需求的热烈讨论，其 V3 版本模型以仅 1/10 训练成本消耗便获得了与海外领先模型 GPT-4o/Llama3.3 对标的能力，并通过对 V3 同一基础模型的后训练获得 R1 模型，R1 在后训练阶段大规模使用了强化学习技

术，在仅有少量标注数据的情况下，提升了模型推理能力，在数据、代码、自然语言推理等任务上，性能比肩 OpenAIo1 正式版。此外，DS 于 2025 年 2 月 24 日正式启动“开源周”，连续 5 天每天开源一个项目，丰富 AGI 领域的开源生态。

高质量的开源模型有望推动 AI 大模型的能力边界探索，并加速 AI 应用落地，利好作为底层支撑的算力硬件需求。DeepSeek 的创新是在命题作文下（中美贸易摩擦背景下 AI 硬件采购受限）的较优解，并未提出任何反“Scaling Law”的趋势，杰文斯悖论（Jevons Paradox）为 DeepSeek 带来的“大模型平权”创新行为影响指明了方向——全行业算力资源使用效率的提升，可能会创造更大的需求。在应用推理方面，能够看到 DeepSeek 在 C 端表现亮眼，根据 Data.ai 数据，DS APP 自 2025 年 1 月 11 日发布以来，下载量呈指数级增长，1 月 20 日至 26 日单周下载量达 170 万次，次周（1 月 27 日至 2 月 2 日）达到 1576 万次，环比增长超 800%，截至 2025 年 2 月 24 日，累计下载量已突破 4000 万次；在 B 端，DS 的开源属性与模块化设计加速了其在垂直领域的渗透，包括政务、医疗、汽车、工业、金融等领域，根据爱分析的统计数据，截至 2025 年 2 月 21 日，已有 45%的央企完成了 DeepSeek 模型的部署。

图表 1：DeepSeek APP 下载量



资料来源：Data.ai，中金公司研究部

大模型云端部署带动的云端算力需求提升，头部云厂商进入资本开支上行周期。大模型的云端部署以弹性算力和快速迭代见长，能够认为，R1 模型作为高质量、低成本的模型代表，开发者通过云厂商调用 API、部署模型并开发应用，有望推动云资源消耗量提升。根据阿里巴巴财报，AI 推动阿里云收入增速持续提升，4Q24 季度收入重回 13%的同比双位数增长，AI 相关产品收入连续六个季度实现三位数同比增长，单季度资本开支 44.1 亿美元，环比增长 81%；业绩电话会上集团 CEO 表示，未来三年云和 AI 基础设施投入预计超过过去 10 年总和，AI capex 指引积极。

DS 部署不止于云端，本地化私有部署同样具备广阔的应用场景，私有化部署方式使得企业及个人能够完全掌控自身的数据环境，有力保障数据安全，降低数据泄露和遭受外部干扰的风险：

C 端呼唤云端协同范式：面向消费级 AIPC 等场景，"云端协同"成为优化体验与隐私保护的必然选择。通过将非敏感任务卸载至云端，可突破终端算力限制，支撑复杂模型推理；同时端侧部署轻量化模型处理隐私数据，既满足 GDPR 等法规要求，又减少网络依赖带来的延迟抖动及可能的体验中断。

B 端部分行业刚性需求驱动本地化部署：部分企业级市场对私有化部署呈现强依赖性，1）尤其是金融、医疗等行业公司，处理较多高度敏感的数据，本地化部署能够防止数据离开企业内部网络，降低数据被外部恶意行为者窃取或滥用的风险；2）定制化需求旺盛，需针对行业知识库进行微调训练，从而推动 DS 大模型形成容器化交付、私有化调优的完整解决方案体系，满足企业对模型所有权与控制权的双重诉求。

DeepSeek R1 具有技术开源和成本控制的核心特点，降低了企业及个人部署高水平 AI 大模型的门槛，有望推动包括 DeepSeek 一体机在内的本地私有化部署需求快速提升：

模型性能领先：DeepSeek R1 在性能上对标国际前沿模型，在数据、代码、自然语言推理等任务上，性能比肩 OpenAIo1 正式版。

开源策略：在闭源模式下，企业依赖大模型厂商的服务订阅，成本与技术的不透明度高。对比之下，DeepSeek 采用 MIT 开源协议，允许企业免费商用和二次开发，同时，在最新“开源周”，DeepSeek 陆续开源了 FlashMLA、DeepEP、DeepGEMM、并行优化策略等项目，涉及大模型推理框架、MoE 模型、FP8 计算性能等方面提升，降低了前沿 AI 技术的获取门槛。

硬件成本：DS 团队通过 MLA（多头潜在注意力机制）、NSA（原生稀疏注意力机制）、Prefill/Decode 分离、高度 EP 等技术创新，实现推理成本下降；同时，通过知识蒸馏等技术，实现参数量分别为 1.5B、7B、8B、14B、32B 和 70B 的蒸馏版模型，在保持大模型性能的同时，减少了对显存、内存和存储的需求，进一步降低本地私有化部署的硬件成本。

硬件适配优化：DeepSeek 早期的模型训练基于 NV 硬件实现，例如 DS 团队开源了 MLA 相关内核（Kernel），解密 MLA 结构在 NV 硬件上的具体实现。国产主流 GPU 厂商已宣布适配 DeepSeek，为基于国产卡的 DeepSeek 一体机的快速落地奠定了基础，而 DS 团队基于 NV 硬件的优化方式开源也为开发者优化适配其他硬件提供了思路。

2、DeepSeek 一体机，密集上新

目前 DeepSeek 大模型一体机分为推理一体机和训推一体机。DeepSeek 推理一体机内置 DeepSeek-R132B、70B、满血版 671B 等不同尺寸模型，价格在几十万到数百万不等，主要面向对数据安全、数据隐私较为敏感的企业用户。而训推一体机的售价更高，用于 DeepSeek-R1 32B 模型的预训练和微调的一体机价格就达到数百万。

华为计算推出昇腾 DeepSeek 一体机，适配从 DeepseekR 11.5B 蒸馏板到满血版的 V3/R1。深度融合昇腾高性能算力底座与 DeepSeek 全系列大模型能力，覆盖语言理解、图像分析、知识推理等全场景需求，为企业提供了一站式的 AI 解决方案，助力千行万业迈向高质量发展的新征程。

昇腾 DeepSeek 一体机能力出众。在官方推荐配置下，新一体机 70B 蒸馏版 R1 系统每秒吞吐可达 3300Tokens，确保了大规模数据的高效处理，能够轻松应对海量文本分析和复杂业务场景模拟。新一体机拥有更好的多用户并发和并发时延保障，满血版 DeepSeek V3/R1 支持 2 倍于业界的多用户并发数，最多可以满足 192 个用户同时在线，每用户时延仅 50ms，满足了大型企业多部门协同工作的需求，极大地提升了企业整体和个人的工作效率。

图表3：昇腾一体机 DeepSeekV3/R 及蒸馏模型推理服务部署推荐配置

模型名称	参数	产品	配置	系统吞吐（Token/s）	多用户并发数（路）
DeepSeek V3	671B	Atlas 800I A2	1024GB	1911	192
DeepSeek R1	671B	Atlas 800I A2	1024GB	1911	192
DeepSeek-R1 Distill-Llama-70B	70B	Atlas 800I A2	512GB	3300	165
DeepSeek-R1 Distill-Qwen-32B	32B	Atlas 800I A2	256GB	4940	247
DeepSeek-R1 Distill-Qwen-14B	14B	Atlas 800I A2	256GB	7500	300
DeepSeek-R1 Distill-Qwen-14B	14B	Atlas 300I Duo	1*Duo 96GB PCIE	730	80
DeepSeek-R1 Distill-Llama-8B	8B	Atlas 300I Duo	1*Duo 96GB PCIE	956	115
DeepSeek-R1 Distill-Qwen-7B	7B	Atlas 300I Duo	1*Duo 96GB PCIE	956	115
DeepSeek-R1 Distill-Qwen-1.5B	1.5B	Atlas 300V	1*300V 24GB PCIE	432	16

资料来源：华为计算，太平洋证券整理

昇腾合作伙伴积极推出基于昇腾芯片的 DeepSeek 一体机，助力各行业应用落地。截止 2 月 12 日，已有 13 家伙伴基于昇腾产品打造自有 DeepSeek 一体机产品，满足细分市场的多样化需求，进一步拓展 AI 技术在各行业的应用边界。例如，神州数码推出基于神州鲲泰（搭载昇腾硬件）AI 算力的神州问学企业级 GenAI 私有化部署解决方案，在神州问学平台上，用户仅需 3 分钟便可完成 DeepSeek 模型的部署。在教育领域，可实现智能教学辅助、个性化学习方案制定；在医疗行业，可辅助疾病诊断、药物研发等；金融行业，可用于风险评估、智能投顾等。

图表4：昇腾合作伙伴推出 DeepSeek 一体机

资料来源：华为计算，太平洋证券整理

此外，多家厂商还推出了基于多款国产算力芯片的 DeepSeek 一体机。例如，联想集团与沐曦联合发布基于 DeepSeek 大模型的一体机解决方案，以“联想服务器/工作站+沐曦训推一体 GPU+自主算法”为核心架构。京东云的 DeepSeek 大模型一体机，支持华为昇腾、海光、寒武纪、摩尔线程、天数智芯等国产 AI 加速芯片。

图表5：推出 DeepSeek 一体机的部分厂商及相关产品

公司名称	产品简介
华为	推出昇腾 DeepSeek 一体机，深度融合昇腾高性能算力底座与 DeepSeek 全系列大模型能力
浪潮信息	提供 AI 服务器集群，支撑模型训练，发布 671B DeepSeek 大模型一体机解决方案
联想集团	与沐曦股份联合发布基于 DeepSeek 大模型的一体机解决方案，推出智能体一体机与训推一体服务器
京东云	发布 DeepSeek 大模型一体机，支持多种国产 AI 加速芯片
中国移动	推出智算一体机-DeepSeek 版，基于国产 CPU/GPU 和操作系统
中国电信	推出息壤智算一体机-DeepSeek 版，以华为昇腾芯片为基础，深度融合了 DeepSeek-R1/V3 系列大模型

中国联通	推出 DeepSeek 一体机+GPU 云服务器解决方案，基于联通云“星罗”算力调度平台
拓维信息	与整数智能共同推出业内首款搭载 DeepSeek 全系列模型的智能数据标注一体机，率先打造“数据标注平台+大模型+算力”全栈国产化的软硬一体解决方案
亚康股份	控股子公司北京亚康华创联合昇腾共同推出了国内首款 DeepSeek 桌面级智能一体机 D-BOX Pro，将 DeepSeek 系列模型（1.5B/7B/8B/14B）集成至桌面设备
优刻得	提供多款 GPU 机型配置，已率先完成沐曦、壁仞、昇腾、天数智芯等主流国产芯片的全适配
紫光股份	旗下新华三集团发布基于 DeepSeek 大模型的一体机 UniCube，全面搭载 DeepSeek V3、R1 模型，并实现 671B DeepSeek 大模型单机推理及单机训推一体服务
天融信	发布 DeepSeek 安全智算一体机产品以“算力硬件平台+智算平台”为基座，集成 DeepSeek 大模型
神州数码	基于神州鲲泰（搭载昇腾硬件）AI 算力的神州问学企业级 GenAI 私有化部署解决方案，与 DeepSeek 系列模型的深度结合
云从科技	从容大模型训推一体机成功适配国产开源大模型 DeepSeek，可实现“开箱即用”的私有化部署
中国长城	长城擎天训推一体机适配 DeepSeek-R1 系列模型，支持私有化部署
润建股份	联合希姆计算打造完全自主可控的“算力+算法+应用”全国产闭环生态，发布了业内领先的 DeepSeek 加持全国产算力政务智能一体机
智微智能	搭载第四/五代 Intel 至强®可扩展处理器与 8 张双宽全高 GPU，结合 DeepSeek R1 推理模型，完成私有化部署
中科曙光	曙光云推出的全国产 DeepSeek 大模型超融合一体机，选用国产 X86 CPU 和国产 GPGPU 加速卡，支持全精度、半精度混合训练与推理
新致软件	新致信创一体机，以海光 K100 GPU 服务器为算力基石，深度融合新致新知人工智能平台与 DeepSeek 系列大模型
云天励飞	云天天书大模型训推一体机成功适配 DeepSeek，该一体机由云天励飞与华为联合推出，可以部署在华为昇腾服务器上，支持私有化部署，实现开箱即用，满足本地化、专属化的业务需求

资料来源：iFinD，各公司微信公众号，太平洋证券整理

3、DeepSeek 一体机优势明显

DeepSeek 一体机降低 AI 大模型的部署门槛。政企客户行业分布广泛，部分企业客户开发经验有限，通常需要全面的售前售后服务支持。目前，DeepSeek 一体机厂商提供两项解决方案：“开箱即用”的部署模式、通过集成工具降低 AI 开发和应用门槛。DeepSeek 一体机采用单次买断制，有利于企业用户控制资本开支及 AI 部署成本。

“开箱即用”部署模式：实现 DeepSeek 一体机小时级一站式交付、即插即用，配备一个大模型开发平台，综合覆盖多元多模数据处理、RAG（检索-生成）以及数据安全等关键环节。厂商将结合客户的具体需求和数据，对大模型进行开发优化，并配备 ISV 在现场进行数据治理、模型微调等复杂流程的整合和部署，以加速企业 AI 应用的落地。对比传统的数据准备、清洗、治理和跨平台的训练、微调生态流程，DeepSeek 一体机将帮助企业节约大量迭代时间。

自主微调模式：针对企业自主微调大模型的需求，DeepSeek 一体机集成主流有效的微调方法，内置多种大模型计算框架和基础模型，微调采用低代码可视化界面，内置了如 Lora、SFT 等多种微调框架和优化参数，有效降低复杂性和技术门槛，企业用户能根据具体需求和数据特性选择合适的技术、快速开发和部署模型应用。

DeepSeek 一体机核心优势在于，实现硬件与软件的深度耦合。通过软硬件协同优化，DeepSeek 一体机可实现算法与国产芯片的高度适配，通过国产 AI 模型+国产 AI 芯片的组合，国内 AI 生态开始打破英伟达的 CUDA 生态限制，推动“国产算力+国产大模型”生态系统的建设。例如，华为 DS 版训/推超融合一体机使用大模型训练、推理和应用开发的华为 ModelEngine AI 平台，该平台基于昇腾 AI 芯片，通过推理框架优化和 MoE 存算协同，提供动态换入换出和全局统一缓存，实现推理高并发和低时延，全面支持 DeepSeek 大模型 R1&V3 和蒸馏系列模型的本地部署与优化，开发者可通过该平台实现上述 DeepSeek 大模型的“一键部署”。联想配合一体机方案的 AI Force 智能体开发平台，通过"一体集成、私有化部署、简易运维"三大技术创新，重构企业级 AI 开发范式，具备高度定制化与再开发能力。

国产芯片有望成为 DeepSeek 一体机的算力基石。DeepSeek 的异军突起，给昇腾、沐曦、昆仑芯等国产芯片创造了更多可能，本土硬件和软件的紧密结合，将为本土大模型的发展提供一条更可控的成长路线。国产芯片纷纷适配 DeepSeek 模型，昇腾拥有芯片+框架+工具链等全栈 AI 能力，与 DeepSeek 的技术栈适配潜力大；海光 DCU 兼容通用的“类 CUDA”环境，擅长高性能计算；沐曦优势在于 GPU 通用性与 CUDA 兼容性。

四、DeepSeek 一体机的部署

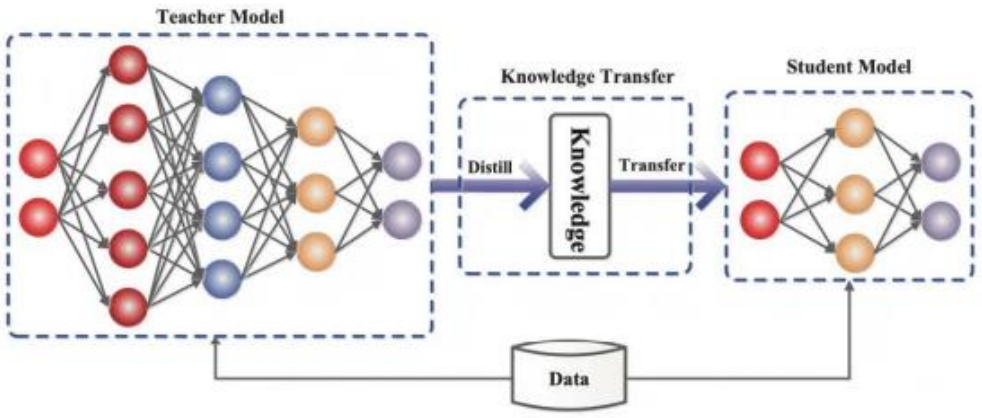
蒸馏模型利好 C 端部署，一体机方案收获 B 端青睐。

1、C 端：DeepSeek-R1+蒸馏技术，轻量化模型推动 AI 端侧部署

（1）DeepSeek-R1 蒸馏：“小模型”蕴含“大智慧”

知识蒸馏（Knowledge Distillation）的本质是知识迁移和压缩，其核心在于将复杂“教师模型”的决策逻辑与特征表征能力迁移至轻量“学生模型”。根据 DeepSeek-R1 技术论文，使用 671B 参数量的 DeepSeek-R1（教师模型）生成 80 万条高质量训练数据，涵盖数学推理、编程、科学问答等场景任务，并通过规则过滤混合预研、冗余段落和代码块，最终数据样本中包括最终答案和多专家协作的决策逻辑；通过知识蒸馏技术，将 671B 参数大模型的复杂推理模式（如长链思考、自我验证等）迁移至轻量模型（学生模型），从而形成参数量为 1.5B、7B、8B、14B、32B、70B 的 6 个不同版本蒸馏模型。

图表 2：“教师模型”通过知识蒸馏后得到“学生模型”



资料来源：Gou, J., Yu, B., Maybank, S.J., & Tao, D. (2020). Knowledge Distillation: A Survey. International Journal of Computer Vision, 中金公司研究部

DeepSeek-R1 蒸馏版模型的推理性能超越同规模传统模型。DeepSeek 蒸馏技术融合了数据蒸馏与模型蒸馏，采用监督微调（SFT）方式，利用 DeepSeek-R1 生成的 80 万个数据样本对基础模型（如 Qwen 和 Llama 系列）进行微调，并且在架构优化中增加层次化特征提取、多任务适应性、参数共享与压缩等设计，实现了高效的知识迁移。得益于模型结构优化和蒸馏技术的应用，蒸馏版本模型在多个推理基准测试中表现优异，根据 DeepSeek-R1 技术论文，DeepSeek-R1-Distill-Qwen-7B 在 AIME 2024 基准测试中实现了 55.5%的 Pass@1（模型首次生成答案即正确的概率），超越了 QwQ-32B-Preview，DeepSeek-R1-Distill-Qwen-32B 在 AIME 2024 上则实现了 72.6%的 Pass@1，在 MATH-500 上实现了 94.3%的 Pass@1，超过了 OpenAI-o1-mini。论文还进一步对比了蒸馏模型和基于 Qwen-32B-Base 模型、使用数学、代码和 STEM 领域数据进行了超过 1 万步大规模强化学习（RL）训练而来的小模型，结果显示，在所有推理基准测试中，蒸馏模型 DeepSeek-R1-Distill-Qwen-32B 均优于后者（DeepSeek-R1-Zero-Qwen-32B），且耗费更少的计算资源，兼具经济性与有效性。能够判断主要得益于蒸馏的知识迁移优势、更高效的学习过程、以及继承了教师模型一部分泛化能力。

图表 3：DeepSeek-R1 蒸馏模型与其他模型推理性能对比

DeepSeek-R1蒸馏模型与其他可比模型对比

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

蒸馏模型与强化学习模型的对比

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
RL结果 ← DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
蒸馏结果 ← DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

注：Pass@1 指模型首次生成答案即正确的概率，主要评估模型的及时响应能力；Cons@64 是通过 64 次独立生成答案后取多数投票结果作为最终答案的评估指标，主要评估多次采用后的稳定性和一致性。
资料来源：DeepSeek-R1 技术报告，中金公司研究部

（2）DeepSeek-R1 蒸馏模型本地部署的硬件要求

传统大模型在推理时需要大量计算资源、以及足够大的内存和存储空间，如满血版 671B 参数量的 DeepSeek-R1 在采用 FP8 训练（精度系数为 1）时，显存需求约 850GB，若采用 INT4 量化，只考虑加载模型参数仍需占用 313GB 的显存，对内存和硬盘空间的要求也较高，超出 PC、手机等终端设备的硬件承载阈值。**蒸馏模型在尽量保持大模型性能的基础上，减少了对显存、内存和存储的需求，更加适合搭载于资源受限的终端设备，适用于 C 端场景。**

DeepSeek-R1 蒸馏模型的本地部署需要根据模型大小和计算需求，选择合适的终端硬件配置。蒸馏后的 DeepSeek-R1 模型可以通过 Ollama 和 AnythingLLM 实现 PC 本地部署。根据联想官网信息以及 ollama，梳理运行不同版本参数蒸馏模型所需的硬件配置：若只运行 1.5B 的超轻量模型，具备实时基础

问答、文本情感分析等功能，集成显卡的配置基本足以支持；若需要执行中等复杂度任务如文本摘要、翻译、图像描述生成等，需部署 7B 或者 8B 端侧模型，INT4 量化假设下的最低显存要求需达到 4-5GB，普通的消费级硬件（如 RTX3060/3070/4060 等）可支持运行，推荐内存配置为 16GB+，硬盘容量大于 10GB；若要部署 14B 中型模型，用于跨模态理解、复杂代码生成、本地知识库检索等任务，需升级硬件配置，采用 RTX4090/A5000 或更高显存的显卡、以及 32GB+的内存和 15GB+的硬盘存储；而对于 32B 或 70B 较大参数量模型的本地部署，以实现多模态任务处理、科研数据分析、复杂语义理解等任务，对 PC 硬件提出了更高要求，通常需要配置专业级 GPU（NVIDIA A100/H100，或采用多卡并行，INT4 量化假设下最低显存要求接近 40GB，推荐显存大小为 80GB，并提出更大的内存和存储、更高的散热和电磁屏蔽等要求。

PC 是承载本地模型的重要终端，更高规格、性能端侧模型的部署正在成为 AI PC 升级的有力推手。2 月 25 日，联想推出全球首批端侧部署 DeepSeek 的 AIPC 产品——YOGA AI PC 元启系列，在消费级设备上实现 70 亿参数端侧模型的流畅运行，用户文档的总结、翻译、撰写等操作无需调用云端大模型即可完成，提升了推理效率，充分保障了数据隐私与离线可用性，还可以根据用户个人需求进行定制化训练。能够认为，AI PC 的换机动力仍有提升空间，此前主要受制于端侧模型能力有限、国外厂商 API 调用限制、以及价格高昂；DeepSeek-R1 基于知识蒸馏的轻量化模型在本地推理性能上表现优异，以更小参数量实现接近原模型的精度，降低了端侧 AI 任务的门槛。能够认为，PC 作为生产力工具，其用户追求性能体验，随着应用场景逐步扩展到多模态任务处理、复杂推理等领域，用户对更优性能、更高规格本地模型部署的需求攀升，传统 PC 的算力与内存配置逐渐成为瓶颈，硬件升级趋势明确，端侧模型进化与硬件迭代形成飞轮效应、有望加速 AIPC 渗透。

图表 4：不同参数量裁剪版 DeepSeek-R1 模型本地部署的硬件要求

模型版本	模型参数量（B）	模型文件大小（GB）	最低显存要求（GB）	推荐GPU配置	推荐CPU配置	推荐内存	推荐存储	适用场景
DeepSeek-R1-Distill-Qwen-1.5B	1.5	1.1	0.8	集成显卡或4GB显存	4核以上	8GB+	5GB+	短文本生成、基础问答等轻量级任务
DeepSeek-R1-Distill-Qwen-7B	7	4.7	3.9	RTX 3060 8GB显存	8核以上	16GB+	8GB+	文案、表格、统计等中等复杂度任务
DeepSeek-R1-Distill-Llama-8B	8	4.9	4.5	RTX 4070 12GB显存	8核以上	16GB+	8GB+	文案、表格、统计、多轮对话中等复杂度任务
DeepSeek-R1-Distill-Qwen-14B	14	9	7.8	RTX 4090 24GB显存	16核以上	32GB+	15GB+	长文本生成与理解、数据分析等企业级复杂任务
DeepSeek-R1-Distill-Qwen-32B	32	20	17.9	RTX A100 40GB显存或多卡并行	32核以上	64GB+	30GB+	科研数据分析、多模态大模型推理等高精度专业领域任务
DeepSeek-R1-Distill-Llama-70B	70	43	39.1	RTX H100 80GB显存或多卡并行	64核以上	128GB+	70GB+	大规模数据分析、创意写作、算法设计等高复杂度生成任务

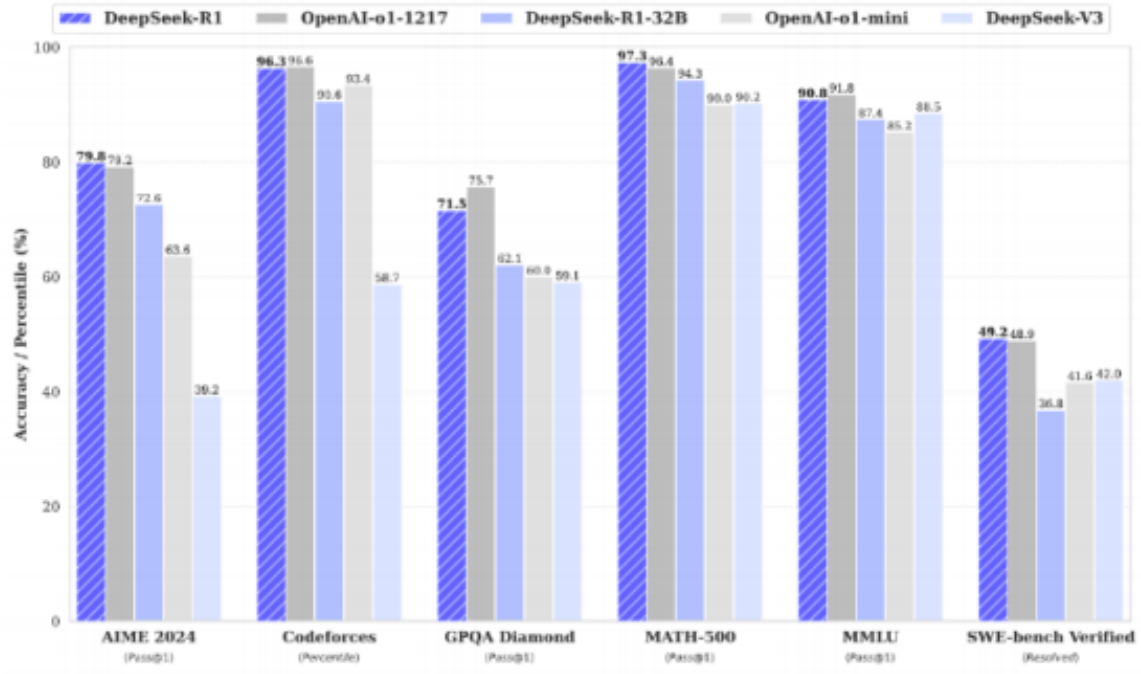
注：1）模型文件大小来源于 Ollama 官网模型下载文件的大小；2）最低显存要求的计算方式：假设均采用 INT4（4 比特）量化，每个参数占用 0.5 字节，且考虑到实际部署时需预留额外显存用于存放中间计算和框架开销，这部分额外开销一般占模型本身大小的 20-50%，我们采用 20%保守计算，显存需求（GB）≈参数规模（B）*每个参数的字节数/ (1.024)^3)*(1+20%)
资料来源：Ollama 官网，CSDN，联想集团官网，中金公司研究部

2、B 端：AI 私有化部署新趋势，DeepSeek 一体机的全栈式解决方案

（1）DeepSeek 一体机重构本地私有化 AI 部署模式

DeepSeek-R1 全参数模型拥有 671B 参数，相较于 32B 参数的蒸馏版，展现出更强的数学、代码及逻辑推理能力，为 B 端用户所需要；但也对系统显存容量、显存带宽、互连带宽、延迟等提出了更高的要求。根据安擎，MoE 模型运行所需的显存可以由公式——模型参数量×精度系数+激活参数量×精度系数+10%~20%其他消耗——计算得到，对于 DeepSeek R1 而言，模型参数为 671B，单次激活专家参数量为 37B，模型主要采用 FP8 训练（精度系数为 1），则所需的显存约为 850GB。

图表 5：DeepSeek-R1 模型性能表现



资料来源：DeepSeek，中金公司研究部

（2）DeepSeek 一体机软硬件协同难点

当下主流国产 AI 训练芯片缺少对 FP8 精度的支持是运行 DS 模型的一大问题，采用 16 位精度单元计算会大幅降低效率。DeepSeek 采用的是 FP8 混合精度，但当前主流的国产 AI 训练芯片缺少对 FP8 精度的支持，如果采用 BF16 或者 FP16 来计算，理论上对精度影响较小，但是对计算和显存的硬件需求几乎增加一倍。采用上节相同计算方法，采用 FP8 精度部署 671B 的 DS 大模型，显存需求约为 850GB；如果采用 FP16 或者 BF16 部署 DS 大模型，显存需求约在 1.5T 以上，以阿里云百炼专属版 AI 训推一体机为例，单机部署全精度 16/8/4bit 下高并发满血版 DeepSeek-R1/V3，部署 16 卡，显存达到 1,536GB。

图表 7：中国 AI 芯片显存大小及支持的数据精度

公司名称	芯片型号	显存大小	支持的数据精度
摩尔线程	基于新一代计算架构MUSA Compute Capability 3.1的全功能GPU	未公开	FP8等
摩尔线程	MTT S4000	48GB	FP64、FP32、TF32、FP16、BF16、INT8 等
沐曦	曦云C500	64GB HBM2E	FP32、FP16
华为	昇腾910	64GB HBM	FP32、FP16、INT8等
海光信息	海光8100	32GB HBM2	FP64, FP32, FP16, INT8, INT4
昆仑芯	R480-X8	32GB	FP32, FP16, INT16, INT8
寒武纪	MLU370-X8	48GB	FP32, FP16, BF16, INT16, INT8, INT4
天数智芯	天垓100	32GB HBM2	FP32, FP16, INT8
壁仞科技	BR100	64GB HBM2E	FP32、TF32+、FP16、BF16、INT32、INT16、INT8等
燧原科技	燧原S60	48GB	FP32、FP16、BF16、INT8

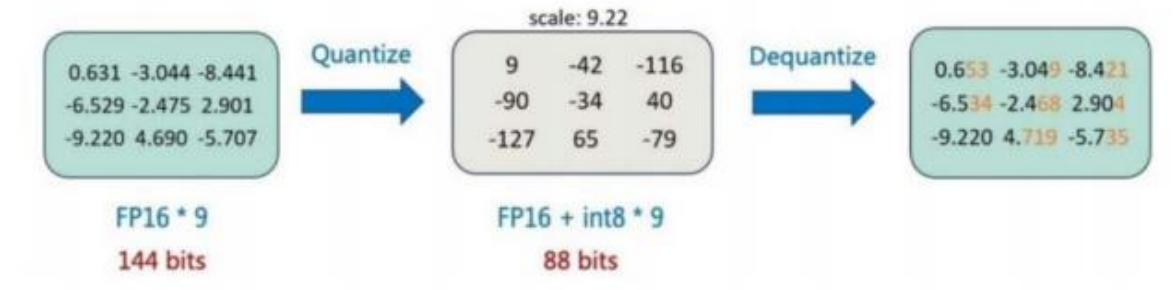
资料来源：摩尔线程官网及公众号，半导体产业纵横，京东，华为，中国算力大会，海光信息，昆仑芯，寒武纪，云轩 Cloud Hin，Hot Chips，燧原科技，中金公司研究部

图表 8：百炼专属版 AI 训推一体机

资料来源：阿里云政企公众号，中金公司研究部

DeepSeek R1 模型全参数配置的显存要求较高，通过定点量化压缩显存占用。以 FP8 精度部署 DS R1 全参数模型所需的显存为 850GB，同时，由于大部分国产 AI 芯片只支持 INT8、FP16、FP32 等格式，如果采用 FP16 精度，单机显存要求将进一步提升至 1.5T 显存以上，超出单台 8 卡 AI 服务器的显存范围。为了在单台 8 卡服务器上实现 671B 全参数 DeepSeek R1 模型，厂商需要进行定点量化，即通过降低模型中的参数精度（如从 16 位浮点数转换为 8 位）来减少模型的大小与计算复杂度，从而降低显存占用并提高吞吐效率，并保持可接受的精度损失范围。以 INT8 量化——即将模型从浮点数转换为 8 位整数——为例，模型的权重和激活值会经过缩放、偏移等量化过程，以尽量多地保留原始浮点数的信息，在推理过程中，这些定点量化值会被反量化回浮点数进行计算，然后再量化回 INT8 进行下一步。

图表 9：INT8 量化示意图



资料来源：53AI，中金公司研究部

DeepSeek 一体机并非 AI 服务器硬件与大模型的简单叠加，可能会遇到无法部署或资源浪费的问题，一体机厂商需要围绕 AI 芯片与大模型进行深度适配，并在优化算力效率与保障模型效果之间寻求平衡点。

五、一体机产业链

以华为昇腾一体机为例，其产业链构成丰富，涵盖了从上游的基础硬件与软件供应，到中游的华为及合作伙伴的研发生产，再到下游的应用场景及相关服务等多个环节。

1、上游：基础硬件与软件供应

（1）芯片及组件供应商

华为海思：提供昇腾系列 AI 芯片，是一体机的核心计算部件，为一体机提供强大的 AI 算力。

其他芯片厂商：为一体机提供 CPU、存储芯片、网络芯片等其他必要芯片。如提供通用 CPU 的英特尔、AMD 等，以及提供存储芯片的三星、海力士等。

基础组件供应商：提供电源、散热器、PCB 板等基础硬件组件。如台达电子为一体机提供电源供应器，猫头鹰等厂商提供散热风扇等散热设备。

（2）软件及算法供应商

操作系统厂商：如麒麟软件、统信软件等，为一体机提供国产操作系统，保障系统的稳定运行和安全。

AI 框架及算法提供商：开源的 AI 框架如 TensorFlow、PyTorch 等，为一体机的 AI 开发提供基础框架和算法支持。

2、中游：华为及合作伙伴

（1）华为

研发与设计：负责昇腾一体机的整体架构设计、硬件研发、软件系统开发以及 AI 算法的优化等工作，将昇腾芯片与其他硬件、软件进行深度整合。

品牌与技术支持：凭借自身的品牌影响力和技术实力，为一体机提供技术支持和售后服务，推动一体机在市场上的推广和应用。

（2）合作伙伴

硬件组装厂商：华鲲振宇、神州数码、拓维信息等合作伙伴，根据华为的设计方案，进行昇腾一体机的硬件组装和生产，确保产品的质量和稳定性。

解决方案提供商：与华为合作，基于昇腾一体机开发行业特定的解决方案，如烽火通信针对数据中心场景，宝德针对高性能计算场景等，为不同行业客户提供定制化服务。

3、下游：应用场景/系统集成商与服务商

系统集成商：将昇腾一体机与其他系统和设备进行集成，为客户提供完整的信息化解决方案，满足客户的多样化需求。

运维服务商：为昇腾一体机提供运维服务，包括设备维护、故障排除、性能优化等，确保一体机的稳定运行。

六、相关公司

1、智微智能

国内领先智联网硬件提供商，传统业务深耕客户。公司产品覆盖行业终端、ICT 基础设施和工业物联网 IIoT 三大业务板块。公司早期与 Intel 共同发布了 OPS-C 的标准，作为标准的制定者将该产品迅速推广，市场占有率居行业前三。公司客户涵盖智能交互平板、PC、云终端市场、网络安全设备、交换机等领域的龙头及核心厂商，是鸿合科技、同方计算机、紫光计算机、宏碁股份、深信服、锐捷网络、深信服等厂商相关业务的主要供应商之一。工业物联网方面，公司以“智微工业”品牌推出全系列工业产品家族，开拓奥普特、比亚迪、理想汽车、汇川技术、先导智能、盛视科技等新客户。

进军 AIGC，子公司腾云智算专注 AI 算力服务。2024 年初，公司投资设立控股子公司腾云智算，持有其 51%股权，进军 AI 算力服务市场。腾云智算成立仅半年就为上市公司股东贡献了 4827.67 万元的净利润，占公司 2024 上半年归母净利润的 85.5%。智微智能在腾云智算加持下，推出 AI 超算系列服务器 SYS-8043，采用 Intel 和 AMD 两大主流平台，支持 8 个扩展槽，适应多种型号的 GPU 加速卡，采用 CPU-GPU 直通架构，CPU 和 GPU 挂载比 1：4，较传统架构数据传输效率提升 20%，已成为多个应用场景的理想选择。据 IDC 预测，预计 2025 年 AI 服务器市场达 317.9 亿美元，2023-2025 年 CAGR 为 22.7%，AI 智算业务将为公司提供增长新动力。

全面拥抱 DeepSeek，开启 AI 新篇章。2 月 11 日，公司发布了搭载 DeepSeek-R1 的高性能算力一体机，采用 SYS-80415R，具备多项优势：1）搭载 Intel C741 芯片组，性能较前代提升 30%，能够轻松应对复杂计算任务；2）最大支持 8 张双宽全高 GPU，能够满足 AI 训练推理的高密度并行计算需求；3）能够进行提供完善及时的定制服务，满足细分场景需求。2 月 13 日，公司端侧产品已经实现了 DeepSeek 大模型的本地化部署，包括一体机、MINI PC、AI 边缘、工作站和信创等最全面产品线，并能够兼容英特尔、AMD、海光等多个平台，可以为用户提供一站式 AI 解决方案。在信创、智能制造等领域，在国产替代与政策支持逻辑加持下，公司将进一步打开 AI 增长空间。

打造自主品牌开源产品，深入鸿蒙新生态。公司积极拥抱开源鸿蒙生态，深度参与构建中国自主可控的操作系统底座。2023 年 12 月，公司已与中软国际签署开源鸿蒙合作协议，实现全面战略合作。25 年 1 月公司产品 T468 获得 OpenHarmony 生态产品兼容性证书，标志着智微智能在开源鸿蒙生态领域的持续深入。目前适配 OpenHarmony 产品线共计 20 余款，包括 OPS 模块、物联网终端、边缘融合终端和平板等产品线，满足不同领域需求。

2、软通动力

国内 IT 服务领军企业，“软硬一体”全栈赋能。软通动力是国内领先的数字信息技术服务企业，长期服务华为、阿里巴巴、百度、腾讯、中国银行等行业头部企业。2024 年，公司收购同方计算机和同方国际，强力构建硬件新增长板块。

“纯血”鸿蒙落地，软通动力是深度受益合作伙伴。2023 年 8 月，HarmonyOS NEXT 发布，不再兼容安卓应用，“纯血”鸿蒙正式落地。华为 Mate60 上市带动终端业务强势反弹，跨过市占率 16%的“生死线”，鸿蒙生态圈逐步成熟。在美国制裁升级和我国信创政策出台的背景下，搭载鸿蒙系统的 PC 产品需求有望增加，成为发展重点。软通动力与华为深度协同，是开源鸿蒙 A 类捐赠人，旗下鸿湖万联基于开源鸿蒙技术自研 SwanLink OS 操作系统，并将该系统落地应用在商显、交通、媒体、金融等多个行业；

公司旗下软通计算（同方计算机）从事信创、商用领域 PC、服务器等的生产销售，有望受益于信创政策落地和鸿蒙 PC 发展。

公司发力互联网，有望受益于 AI 发展。软通动力是阿里巴巴、腾讯、百度、字节跳动等头部互联网厂商的核心信息技术服务和计算产品供应商之一。公司与阿里云构建了全面战略合作，是腾讯、百度的 A 级供应商，并与字节火山引擎保持战略协同，在 AI 大模型相关产品与应用研发方面开展合作。2025 年字节跳动将大力发展人工智能领域，在山西太行的算力中心二期项目建设用地也于 2025 年 1 月获批。字节跳动看好并大力发展人工智能领域，软通动力有望从中受益。

3、神州数码

公司持续获得运营商算力订单，联合华为发布一体机。公司作为华为昇腾和鲲鹏双领先级合作伙伴，持续获得运营商算力订单，2024 年中标中国移动智算中心采购金额约 20 亿元，中标中国电信集中采购项目金额近 10 亿元。公司联手华为打造一体机，2025 年 3 月 21 日，神州鲲泰携手华为发布“昇腾+伙伴大模型应用一体机”。

4、中国长城

深耕国产自主可控计算机硬件制造、人工智能计算、信创产业。旗下长城科技推出长城 AIGC 一体机。基于国产服务器和 GPU 推出产品，在信创领域积累深厚，产品安全自主，且作为重要计算机硬件制造企业，有完整研发生产体系，保障产品质量和供应稳定。

5、麒麟信安

业务涉及信息安全、云计算、智算解决方案。推出麒麟信安全国产化智算一体机。在关键领域信创解决方案上，以高性能、高安全为优势，在信息安全领域技术和产品积累丰富，可提供安全可靠智算解决方案。

6、天玑科技

提供数据中心 IT 基础设施解决方案、云计算、数据库管理服务。自主研发数据库一体机和超融合一体机 PriData，具备一站式解决方案能力，可满足不同客户多样化需求。

七、未来展望

1、一体机是当前 AI Agent 的理想实现载体

从此意义上来看，一体机作为集成化方案同时将 AI Agent 产业链上游的模型、算法、硬件和中游的平台及下游的应用集合在一起，可以给客户提供针对垂类领域便捷高效的解决方案，符合当前主要需求方——政企、金融客户众多细分场景的部署要求。**是当前 AI Agent 的理想实现载体。**

2、DeepSeek 一体机市场空间测算

AI 大模型能够有效提升政府工作效率，深圳市福田区已上线 11 大类 70 名“数智员工”，满足 240 个政务场景的需求，其中，个性化定制生成时间从 5 天压缩至分钟级，公文格式修正准确率超 95%，审核时间

缩短 90%，错误率控制在 5%以内。此外，医疗、金融等行业及央国企由于涉及敏感信息，对于数据安全的要求较高。**能够认为，DeepSeek 大模型一体机作为开箱即用的私有化部署方案，在实现快速部署 AI 大模型的同时，能够满足对于公民信息、关键业务数据等数据安全保障的需求，有望受益于政府及企业的 AI 转型趋势。**

DeepSeek 一体机有望达到 500 亿元级别市场空间。根据 IDC，2025 年中国服务器市场出货量有望达到 488 万台，政府、金融、公共事业、医疗等 6 大政企行业由于涉及隐私数据，存在本地私有化部署需求，2021 年上述行业占中国服务器市场需求约 28%。预计，乐观情景下 2025 年上述行业约 5%的需求转向 DeekSeek 一体机，则需求达到 7 万台，市场规模有望达到 540 亿元。

图表 11：2025 年国内 DeepSeek 一体机市场空间测算

	2025E
中国服务器数量(万台)	488
主要行业服务器需求占比	**28%**
—政府	9%
—金融	10%
—公共事业	2%
—交通	2%
—医疗	2%
—教育	3%
主要行业服务器总需求（万台）	135

	悲观情景	中性情景	乐观情景
DeepSeek一体机渗透率	2.0%	3.5%	5.0%
主要行业DeekSeek一体机需求（万台）	**3**	**5**	**7**
DeepSeek一体机单价（万元）			80
主要行业DeekSeek一体机市场规模（亿元）	**216**	**378**	**540**

注：DeepSeek 一体机单价因配置不同而存有较大差异，测算中采用中位数水平单价
资料来源：IDC，中金公司研究部

八、参考研报

1. 中金公司-科技硬件行业 AI 进化论（3）：DeepSeek 本地部署需求盛行，一体机硬件乘风而上
2. 太平洋证券-计算机行业周报：Deepseek 一体机大潮开启
3. 浙商证券-计算机行业大模型点评：国产化智算一体机助力政企 DeepSeek 部署
4. 东北证券-计算机行业：昇腾一体机深度报告，打造 AI 大模型“最后一公里”
5. 东方证券-电子行业动态跟踪：国产芯驱动 Deepseek 一体机，AI 加速赋能各行各业
6. 北京大学-DeepSeek 私有化部署和一体机


慧博公众号


慧博 PC 版


慧博 APP

免责声明：以上内容仅供学习交流，不构成投资建议。