# 私有化部署需求提升带来大模型一体机投资机会

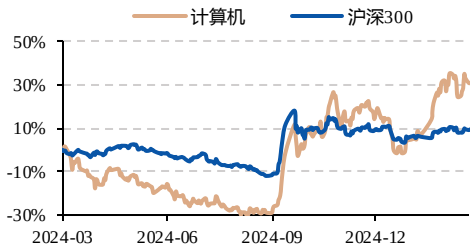
计算机行业深度报告

2025 年 03 月 18 日

**评级** 领先大市

评级变动： 维持

**行业涨跌幅比较**




| % | 1M | 3M | 12M |
|---|---|---|---|
| 计算机 | -0.17 | 15.27 | 34.42 |
| 沪深 300 | 1.53 | 2.18 | 12.26 |

**何晨** 分析师
执业证书编号:S0530513080001
hechen@hnchasing.com
**黄奕景** 研究助理
huangyijing@hnchasing.com

**相关报告**

1 计算机行业点评：OpenAI 发布 Agent 开发组件，Agent 应用开发周期有望明显缩短 2025-03-12

2 计算机行业点评：央国企、政务单位积极推动 DeepSeek 私有化部署 2025-02-18

3 计算机行业双周报：DeepSeek 重大突破，重视 AI 应用与算力的再平衡 2025-02-13

| 重点股票 | 2023A | | 2024E | | 2025E | | 评级 |
|---|---|---|---|---|---|---|---|
| | EPS（元） | PE（倍） | EPS（元） | PE（倍） | EPS（元） | PE（倍） | |
| 紫光股份 | 0.74 | 40.12 | 0.82 | 36.07 | 1.00 | 29.41 | 买入 |
| 浪潮信息 | 1.21 | 49.42 | 1.48 | 40.46 | 1.78 | 33.70 | 增持 |

*资料来源：iFinD，财信证券*

**投资要点：**

- **DeepSeek 技术创新有望推动政企私有化部署需求提升。**私有化部署凭借物理隔离、数据闭环、自主管控、定制服务等特征，正成为政企部署 AI 大模型的主流选择。DeepSeek 的技术创新显著降低了私有化部署的模型和算力门槛，有望解决政企落地 AI 应用的部分痛点难点问题，进一步推动政企私有化部署需求提升。

- **私有化部署方案的选型考虑包括模型参数、运行参数、算力硬件、配套生态及软件栈支持等。**首先需要根据企业实际业务场景需求确定合适的模型参数和运行参数，再基于推理性能、并发需求和投入成本等多维度考虑确定算力硬件，同时也需要重点考量 AI 计算卡的配套生态和软件栈支持。

- **私有化部署方案的选型考虑和部署流程颇为复杂，大模型一体机应运而生。**选型考虑中通常会面临硬件型号多且复杂、与原有 IT 设施的适配性、与实际业务场景需求的匹配度等等问题，部署流程中通常会出现环境配置多且复杂、AI 软件栈多且复杂、依赖版本配套复杂、性能调优复杂等等问题。私有化部署的高复杂性和高门槛催生了大模型一体机的服务器形态，这类服务器通过预集成算力硬件+优化软件栈+预装模型，将 AI 计算卡、配套生态及软件栈、模型算法、数据安全等核心要素深度融合，从而实现软硬结合，通电即用，一键部署 AI 大模型。

- **一般政企不具备较成熟的 AI 基础设施团队，而大模型一体机有望解决私有化部署中的痛点难点，契合政企需求。**央国企、政务机构、学校、医院等泛政府类单位大多不具备较高水平的 AI 基础设施团队，而大模型一体机有望解决私有化部署中硬件选型难、软件适配慢、调优成本高等痛点难点，让政企无需组建专业团队即可实现敏捷部署。

- **私有化部署需求增加有望推高大模型一体机销量。**根据初步测算，现阶段央国企、政务机构、学校、医院等泛政府类单位私有化部署 AI 大模型所需的服务器（一体机）开支空间约在 1000 亿元左右，且随着 AI 应用场景逐渐拓宽，服务器（一体机）需求仍有较大提升空间。

- **建议关注：**紫光股份、浪潮信息、中科曙光、拓维信息等。

- **风险提示：宏观经济波动风险；AI 技术发展不及预期；下游客户 AI**

**算力支出意愿不及预期风险；供应链风险；服务器（一体机）行业竞争加剧风险。**

# 内容目录



# 图表目录

# 1 DeepSeek 技术创新有望推动政企私有化部署需求提升

## 1.1 什么是私有化部署？

本文所指私有化部署（Private Deployment）是指将 AI 大模型及相关基础设施完全部署在客户自主掌控的物理或虚拟环境中，属于客户的内部资产。

与公有云服务相比，私有化部署具有以下核心特征：

**1）物理隔离性：**部署环境与企业内网或专用服务器集群完全隔离，不与其他组织共享计算资源。

**2）数据闭环性：**所有训练数据、推理数据、模型参数均在本地存储和处理，形成完整的数据生命周期闭环。

**3）自主管控性：**用户拥有完整的系统管理权限，可自主进行版本迭代、权限管理、日志审计等操作。

**4）定制化服务：**可根据业务需求深度定制模型架构、算法模块、交互接口等核心组件。

**表 1：公有云对比私有云**

| 对比项 | 公有云 | 私有云 |
|---|---|---|
| 用户类型 | 创业公司、小型公司、个人 | 政府、大企业 |
| 业务类型 | 对外提供交互的业务 | 组织内部业务 |
| 安全 | 主机层面实现安全隔离 | 网络层面实现安全隔离 |
| 成本 | 初期成本低，后期当业务量大时，成本较高 | 初期成本高，随着业务量增加，后期平均成本低 |
| 定制 | 很少定制 | 灵活定制，可与现有系统进行集成 |
| 技术架构 | 自研架构、主要关注分布式、大集群 | 开源架构，主要关注高可用，灵活性 |
| 兼容性 | 根据公有云的要求来修改自身业务 | 主动兼容和适配自身业务 |
| 运维 | 用户无法自主运维，公有云服务商统一运维 | 自主运维，也可托管给第三方运维 |

*资料来源：华为知识百科，财信证券*

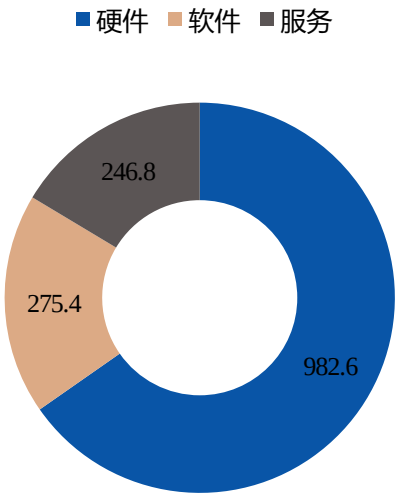
**私有化部署具体可以分为本地化部署和托管私有云两种形式。**1）本地化部署：在企业自有数据中心部署全套系统，完全自主掌控硬件和软件，自主负责日常运维和管理，对数据安全性和系统运行稳定性有极其严格的监管和合规要求的政府部门和企业大多采取本地化部署。2）托管私有云：由第三方 IDC、运营商、云厂商等提供物理隔离的专属硬件资源，企业远程管理，即拥有按需购买，灵活扩容等公有云优势的同时，也能满足用户对硬件资源的专属使用，满足政企上云的安全性要求。

**私有化部署存在软硬一体和裸金属两种服务模式。**1）软硬一体：以大模型一体机为代表，软件平台和硬件设备统一适配，强调开箱即用，用户购买 Deepseek 模型一体机后仅需联网通电即可实现 Deepseek 模型的部署。2）裸金属：软件系统平台与硬件设备解

耦，通过软件平台的兼容适配能力对用户采购的不同品牌、类型的硬件设备实现资源池化和统一管理，可以避免被单一厂商捆绑销售且软硬件深度绑定，便于后续节点的扩建和迁移。

**硬件设备占据私有化部署的约七成价值，软件及服务占据三成。**从销售额来看，服务器、存储、网络设备为代表的硬件设备仍然为私有化部署市场的核心环节，占比达到65.3%，软件及服务占到34.7%。

**图 1：2022 年中国私有云细分市场结构（亿元）**




*资料来源：赛迪顾问《2023 中国企业私有云市场研究报告》，财信证券*

## 1.2 为何要私有化部署？

**对于央国企、政务机构、学校、医院等泛政府单位而言，私有化部署是更主流的 AI 大模型部署方案，主要是出于数据安全、业务需求、部署成本、性能和稳定性等多维度的考虑。**

**出于数据安全考虑，私有化部署相比公有云服务，能更好地满足政企严格的数据安全要求。**央国企、政务机构、学校、医院等泛政府单位通常涉及很多国家经济命脉数据（如能源、金融、交通等）以及公民隐私信息（如户籍、社保、医疗等），采取私有化部署的形式可以避免由于云端存储数据带来的数据泄露风险，并且可以根据实际需求定制专属的安全防护体系，如设置多层次访问权限、加密关键数据、实施实时监控与预警机制等。

**出于实际业务场景考虑，私有化部署方便深度定制开发，并与现有系统无缝对接。**1）各类政企的业务流程、管理模式等各有特点，私有化部署可依据企业特定业务逻辑与操作流程对 AI 大模型进行深度定制开发。2）各类政企通常已有 OA、ERP、内部政务系统等多种 IT 系统，私有化部署的 AI 大模型能与现有系统实现无缝对接，打破数据孤岛，实现数据在各系统间的自由流通与协同应用。

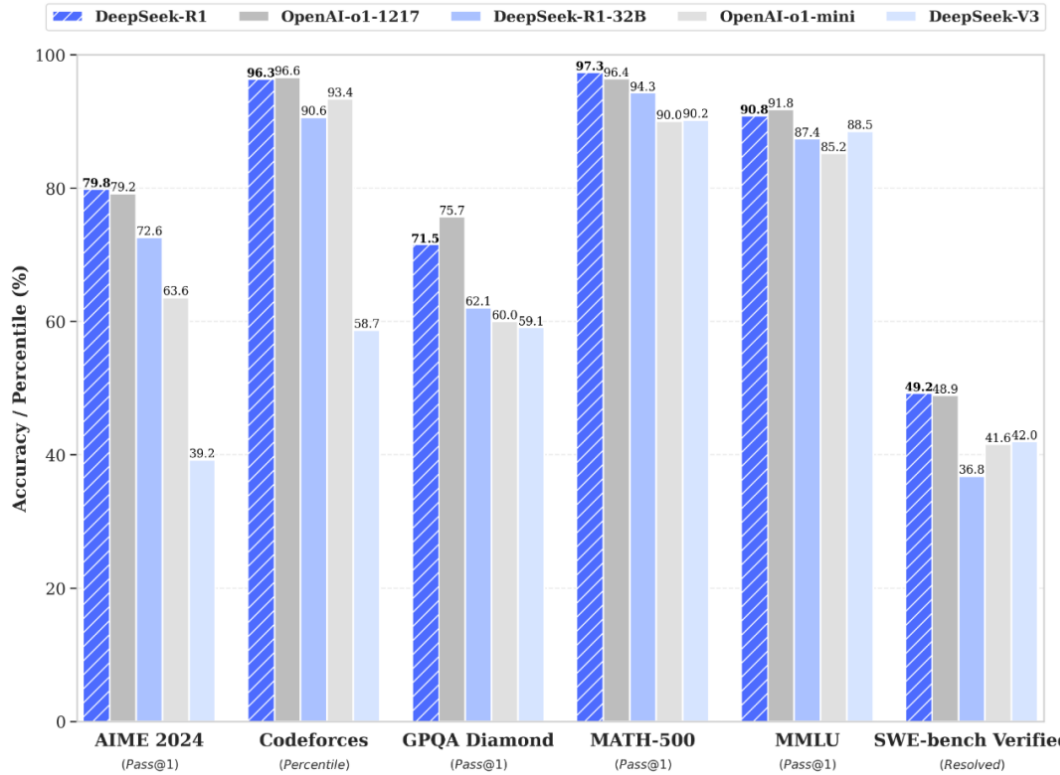
**出于部署成本考虑，私有化部署的长期成本相对可控。**虽然私有化部署初期需投入一定硬件设施、软件授权与维护成本，但从长期看，可避免公有云模式下按使用量或用户数计费带来的成本不确定性。随着企业数据量与业务规模增长，公有云成本可能大幅攀升。而对于央国企、政务机构、学校、医院等泛政府单位而言，其业务需求和应用场景相对稳定，私有化部署的固定成本模式可能更利于企业进行成本规划与控制。

**出于性能和稳定性考虑，私有化部署可以提供低网络延迟与高稳定性。**1）在实时人机对话、智能客服、工业自动化控制、边缘计算等应用场景下，网络延迟是影响用户体验与系统性能的关键瓶颈，私有化部署将 AI 大模型部署在企业内网环境或边缘节点，能够显著降低网络传输延迟，实现毫秒级的快速响应，大幅提升用户交互体验与系统实时性。2）政企自建的 IT 环境具备更高的可控性，有助于保障系统运行的稳定可靠，为关键业务应用保驾护航。

## 1.3 Deepseek 技术创新显著降低大模型落地的模型和算力门槛

**DeepSeek-R1 发布，模型性能对齐 OpenAI-o1 正式版。**DeepSeek-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。

**图 2：DeepSeek-R1 性能对比**




*资料来源：Deepseek 官网*

**DeepSeek-R1低参数蒸馏小模型性能对标国际领先水平。**DeepSeek 利用 DeepSeek-R1

模型的输出进行监督微调，在 Qwen、llama 等开源小模型的基础上蒸馏了 6 个小模型并开源，其中 32B 和 70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果。

**图 3：DeepSeek-R1 蒸馏模型性能对比**

| | 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*

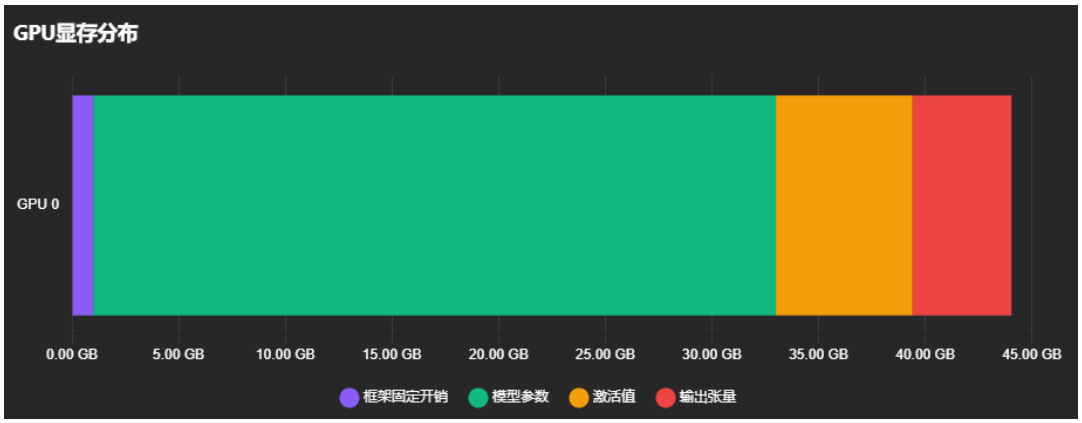
**私有化部署的模型门槛显著降低。**对于企业内部知识问答等中等复杂应用场景，32B 和 70B 等中等规模蒸馏模型已经可以出色地完成任务。同时，DeepSeek-R1 模型完全开源，用户可以无门槛下载模型到本地部署。其模型开源仓库（包括模型权重）统一采用标准化、宽松的 MIT License，完全开源，不限制商用，无需申请，无论是企业用户还是个人用户都可以无门槛地下载模型到本地并部署。并且产品协议明确可“模型蒸馏”，明确允许用户利用模型输出、通过模型蒸馏等方式训练其他模型。

**图 4：DeepSeek-R1 及其蒸馏模型完全开源**


*资料来源：Deepseek 官网*

**蒸馏小模型所需算力硬件门槛也相对较低。**根据 ThinkInAI 测算数据，FP8 精度、序列长度为 2048、批次大小为 16 的情况下，部署 32B 模型需要 44.04GB 显存，仅需一张 NVIDIA L40（48GB）或一张华为 Ascend 910B (64GB)即可；部署 70B 模型需要 92.41GB 显存，仅需两张 NVIDIA L40（48GB）或两张华为 Ascend 910B (64GB)即可。私有化部署的模型和算力门槛显著降低，企业可以灵活地根据实际业务需求和 IT 资源情况选择对应参数规模的蒸馏小模型。

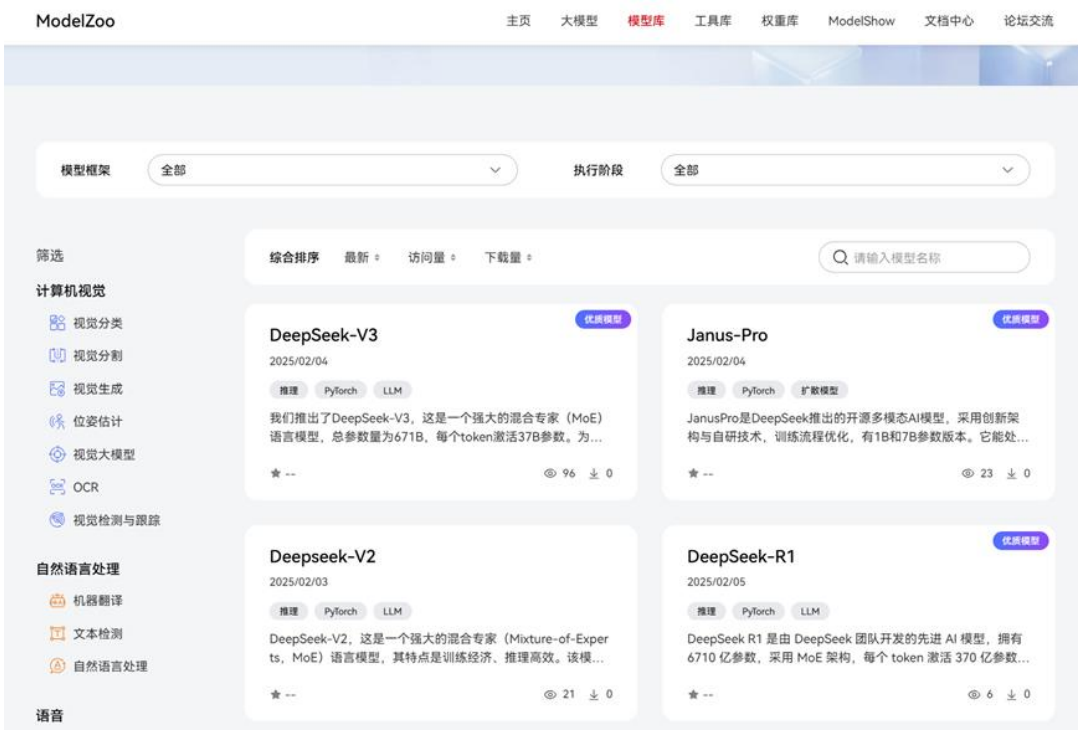
**图 5：部署 DeepSeek-32B 模型所需显存分布**




*资料来源：ThinkInAI*

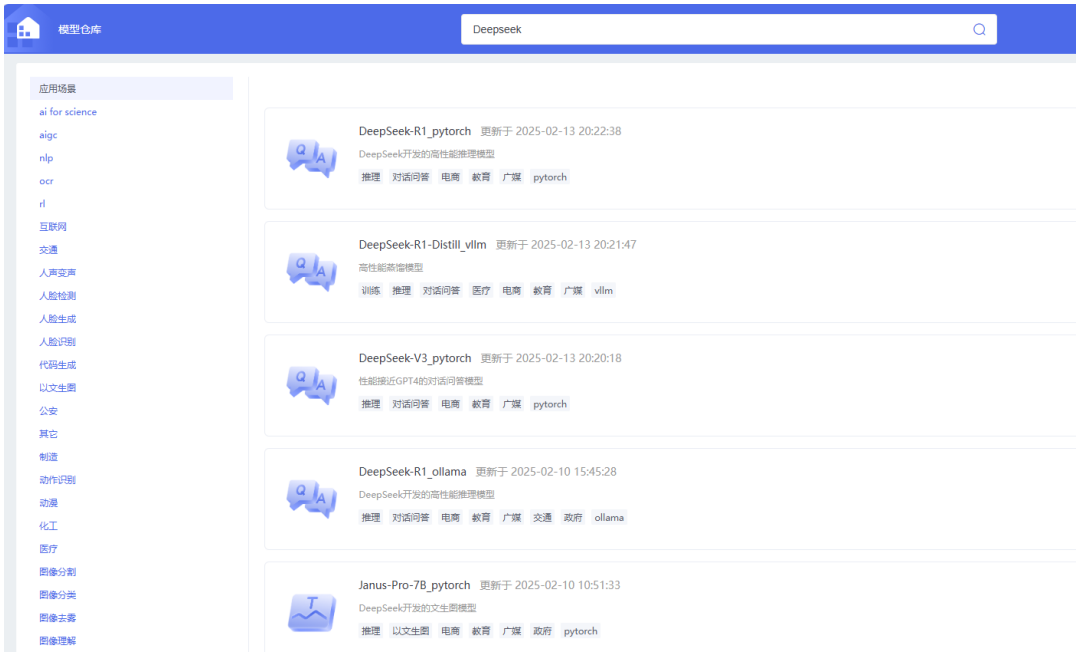
**硬件厂商迅速适配 DeepSeek 全系列模型并推出模型部署文档，私有化部署门槛进一步降低。**2 月 4 日，华为昇腾宣布适配 DeepSeek-R1、DeepSeek-V3、DeepSeek-V2、Janus-Pro 模型，支持一键获取 DeepSeek 系列模型，支持昇腾硬件平台上开箱即用，推理快速部署。同日，海光信息宣布完成 DeepSeek V3 和 R1 模型与海光 DCU（深度计算单元）的适配，支持基于 DCU 平台的快速部署和使用相关模型。

**图 6：华为昇腾宣布适配 DeepSeek 系列模型并发布部署文档**




*资料来源：华为昇腾社区*

**图 7：海光 DCU 宣布适配 DeepSeek 系列模型并发布部署文档**

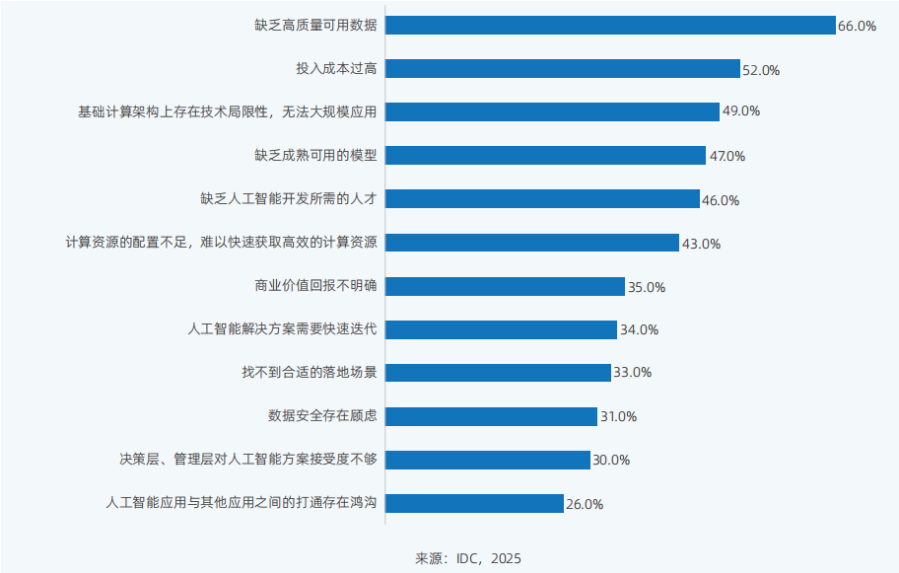



*资料来源：海光开发者社区*

## 1.4 DeepSeek 系列模型有望提升政企部署热情

**DeepSeek 技术创新有望解决政企落地 AI 应用的部分痛点难点问题，提升部署热情。**相比于互联网企业和云厂商对于 AI 应用落地的热情，政企侧的需求相比稍显不足，主要源于缺乏高质量数据、投入成本过高、计算架构局限、模型能力不成熟、人才缺乏、管理层接受度不高等方面的痛点难点问题。而此次 DeepSeek 系列模型的问世，有望解决投入成本、模型能力、管理层接受程度等方面的部分痛点难点，促使政企落地 AI 应用的热情提升。

**图 8：企业在应用人工智能过程中面临的挑战**




*资料来源：IDC*

**石油、金融、能源、建筑等行业央国企加速推进 Deepseek+行业应用落地。**人民邮电网 2 月 17 日讯，在中国移动的助力下，中国石油高效完成了 DeepSeek V3/R1 全栈国产化的训推适配和私有化部署（昆仑大模型），实现了从底层芯片到框架、模型的全栈自主可控。在应用层面，昆仑大模型的问答应用“行业大家”目前已新增 DeepSeek 深度推理能力，用户在使用该应用时，除了可以得到昆仑大模型生成的能源化工领域专业问答结果，还能选择“深度思考”模式，体验知识推理、场景理解等 AI 服务。在模型层面，昆仑大模型的 AI 中台模型广场目前已上线 DeepSeek-V3 与 DeepSeek-R1 模型版本，并实现全尺寸适配，用户可基于 AI 中台调用 DeepSeek 模型 API 服务，并使用 AI 中台组件及工具构建智能体，以满足不同场景的需求。

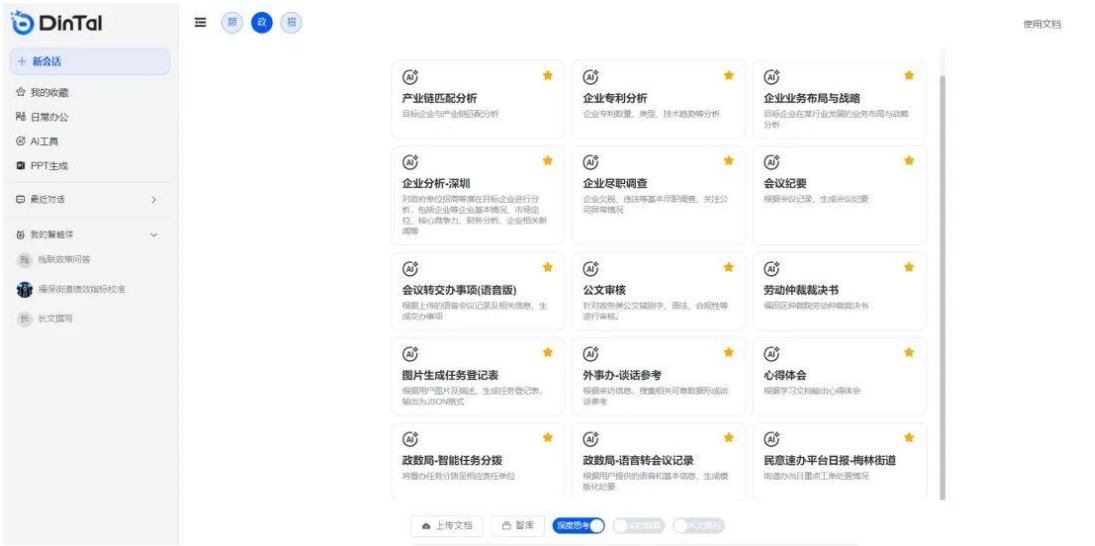
**图 9：中国石油昆仑大模型使用指南**


*资料来源：中国石油报*

**广州、深圳等地政务机构部署上线 DeepSeek，助力政务服务场景应用。**1）广州政府网 2 月 15 日讯，广州市政务服务和数据管理局在政务外网正式部署上线 DeepSeek-R1、V3 671B 大模型，通过融合海量政务数据要素，大模型将丰富政务服务场景应用。本次大模型上线工作依托数字广州创新实验室实施已完成 DeepSeek-R1 等模型深度适配国产硬件，将通过政务

专网算力推动人工智能大模型在民生政策解读系统、12345 热线工单分派等政务领域应用，政务办公、城市治理、民生服务等多个热门政务领域已率先探索应用 DeepSeek 模型优化工作和服务场景。2）财联社 2 月 18 日讯，深圳福田区推出了基于 DeepSeek 开发的 AI 数智员工，上线福田区政务大模型 2.0 版，除了有 DeepSeek 通用能力外，还结合各部门各单位实际业务流程量身定制个性化智能体，首批满足 11 个类别，240 个业务场景使用。通过构建“需求-训练-场景应用-迭代”闭环生态体系，联合 Dintal 数智员工实现“技术穿透业务”的智能化服务升级，覆盖公文处理、民生服务、应急管理、招商引资等多元场景。引入“AI 公务员”后，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。“执法文书生成助手”将执法笔录秒级生成执法文书初稿。民生诉求分拨准确率从 70%提升至 95%，民情周报日报初稿一键生成。“安全生产助手”生成演练脚本效率提升 100 倍。

**图 10：深圳福田基于 DeepSeek 开发的 AI 数智员工**




*资料来源：幸福福田公众号*

# 2 私有化部署方案的选型考虑：模型参数、运行参数、算力硬件、配套生态及软件栈支持等

## 2.1 私有化部署大模型的一般流程

以昇腾 Atlas 800I A2 (8*64G)裸金属服务器为例，企业级部署 Deepseek-R1 模型的流程大致如下：

1、软件栈准备：

1）安装与配置服务器的底层操作系统，如 Ubuntu、Debian、openEuler 等。

2）安装昇腾 NPU（AI 计算卡）固件及驱动。

3）安装与配置昇腾提供的各类配套软件包，包括 Mindle（推理引擎）、CANN（异构计算架构）、MindSpore（AI 框架）等。

**图 11：裸金属服务器所需安装软件栈**


*资料来源：华为昇腾社区*

2、模型获取：下载对应参数大小（671B 满血版或 70B 等蒸馏模型）的模型代码及权重，并转换为相应精度（FP8 或 FP16 等）。

3、推理服务部署：配置环境变量，启动推理服务容器并验证。

4、性能调优：调优推理引擎等软件栈的参数配置，从而达到最优推理效率。

5、安全与监控：进行网络安全设置、管理日志信息、配置监控看板等。

## 2.2 私有化部署方案的选型考虑一：模型参数和运行参数

**企业级私有化部署 LLM 模型，首先需要考虑模型参数和运行参数。**模型参数（满血版 or 蒸馏版）和运行参数（长下文长度、批次大小等）的大小决定了后续需要多少算力硬件，需要综合考虑企业实际业务场景需求。复杂决策场景，如金融研究分析、医疗影像诊断、法律文书分析等，需要较强的模型推理和上下文记忆能力，对于模型参数（70B 以上）和上下文长度（32K 以上）的要求较高。一般复杂场景，如企业内部知识库、线上客服等，对于模型参数和运行参数的要求相对较低。

**图 12：不同参数大小的 DeepSeek-R1 模型对应的特点与适用场景**

| 版本 | 参数量 | 特点 | 适用场景 | 硬件需求 |
|---|---|---|---|---|
| deepseek-r1:1.5b | 1.5B | 轻量级模型，运行速度快，性能有限。 | 低配硬件，简单任务 | 低配硬件 |
| deepseek-r1:7b | 7B | 平衡型模型，性能较好，硬件需求适中。 | 多数常见任务 | 中等硬件 |
| deepseek-r1:8b | 8B | 性能略强于 7B 模型，适合更高精度需求。 | 需要更高精度的任务 | 中等硬件 |
| deepseek-r1:14b | 14B | 高性能模型，擅长复杂任务（如数学推理、代码生成）。 | 复杂任务（数学推理、代码生成等） | 高硬件需求 |
| deepseek-r1:32b | 32B | 专业级模型，性能强大，适合高精度任务。 | 研究、高精度任务 | 高端硬件 |
| deepseek-r1:70b | 70B | 顶级模型，性能最强，适合大规模计算和高复杂度任务。 | 大规模计算、高复杂度任务 | 专业级硬件 |
| deepseek-r1:671b | 671B | 超大规模模型，性能卓越，推理速度快，适合极高精度需求。 | 前沿科学研究、复杂商业决策分析 | 极高硬件需求 |

*资料来源：菜鸟教程*

## 2.3 私有化部署方案的选型考虑二：算力硬件

**AI 计算卡的性能直接决定了模型的推理性能和推理效率，从模型部署的最低算力硬件要求出发，显存容量是 AI 计算卡选型时所考虑的首要因素。**AI 计算卡参数配置包括显存容量、显存带宽、计算能力、互联带宽等。其中，计算能力、显存带宽、互联带宽等直接影响模型推理的性能和效率，而显存容量则直接决定了模型能否正常部署。

**图 13：英伟达 H200 参数配置**

| Technical Specifications | |
|---|---|
| Form Factor | H200 SXM[1] |
| FP64 | 34 TFLOPS |
| FP64 Tensor Core | 67 TFLOPS |
| FP32 | 67 TFLOPS |
| TF32 Tensor Core | 989 TFLOPS[2] |
| BFLOAT16 Tensor Core | 1,979 TFLOPS[2] |
| FP16 Tensor Core | 1,979 TFLOPS[2] |
| FP8 Tensor Core | 3,958 TFLOPS[2] |
| INT8 Tensor Core | 3,958 TFLOPS[2] |
| GPU Memory | 141GB |
| GPU Memory Bandwidth | 4.8TB/s |
| Decoders | 7 NVDEC<br>7 JPEG |
| Max Thermal Design Power (TDP) | Up to 700W (configurable) |
| Multi-Instance GPUs | Up to 7 MIGs @16.5GB each |
| Form Factor | SXM |
| Interconnect | NVIDIA NVLink®:<br>> 900GB/s<br>> PCIe Gen5: 128GB/s |
| Server Options | NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs |
| NVIDIA AI Enterprise | Add-on |

*资料来源：英伟达官网*

### 2.3.1 显存容量

从满足模型部署的最低要求出发，首先需要考虑显存容量是否足够。

**不同的参数和计算精度的模型所需占用的显存容量不同，计算公式为模型参数×计算精度。**以常见企业级生产部署环境为例：DeepSeek-R1-70B 模型，FP8 计算精度，序列长度（模型一次能处理的最大 token 数）8192，批次大小（Batch size，决定了模型一次处理的请求数量）16，一共需要约 70GB 的显存容量=模型参数：70B×模型精度：1 字节（FP8）。

此外还需要考虑一部分其他显存花销：

1）**激活值缓存：**模型运行时产生的中间计算结果，与模型参数和精度相关，计算公式为模型参数*模型精度*动态系数（0.1-0.5，取决于模型参数）。常见企业级生产部署环境下，一共需要约 17.50GB 的激活值缓存=模型参数：70B×模型精度：1 字节（FP8）×动态系数：0.25。

2）**输出张量缓存：**模型生成结果所需的临时存储空间，与批次大小、序列长度和词表大小相关，计算公式为批次大小×序列长度×词表大小×模型精度÷（$1024^3$ ）。常见企业级生产部署环境下，一共需要约 15.66GB 的输出张量缓存=批次大小: 16×序列长度: 8192×词表大小: 128256×模型精度: 1 字节（FP8）÷（$1024^3$ ）。

3）**固定开销：**AI 计算卡和模型初始化时的固定显存开销，包括软件栈缓存、算子编译缓存等，每个 AI 计算卡需要约 1.00GB。

综上，常见企业级生产部署环境下，一共需要约 104.16GB 的显存容量=模型占用:70.00GB+激活值缓存:17.50GB+输出张量缓存:15.66GB+固定开销:1.00GB。

**图 14：常见企业级生产部署环境（DeepSeek-R1-70B，单卡）所需的显存容量分布**


*资料来源：ThinkInAI*

根据上述计算结果，1 张 NVIDIA H200（显存容量：141GB）或 2 张 NVIDIA H20（显存容量:96GB）或 2 张华为 Ascend 910B （显存容量：64GB）均可满足 70B 模型部署最低要求。但是若考虑到生产/开发/测试环境的隔离以及安全性与高可用性冗余等因素，实际业务场景下的模型部署最低要求可能会有所提高。

### 2.3.2 AI 算力大小、显存带宽、互联带宽等

在满足显存容量要求的前提下，AI 计算卡的计算能力、显存带宽、互联带宽等直接决定模型推理的性能和效率。

**计算能力决定算力天花板。**计算能力代表芯片在单位时间内完成矩阵乘法、卷积等核心运算的峰值能力，即每秒浮点运算次数的理论峰值。不同 AI 计算卡的计算架构与配套软件栈的优化情况存在差异，其实际计算效率会存在不同程度的折扣。

**显存带宽决定数据传输效率。**显存带宽代表显存与计算核心间的数据传输峰值速率，当模型参数或激活值的数据量（主要由 batch size 决定）超过带宽供给能力时，则模型推理性能与效率的瓶颈由显存带宽决定。

**互联带宽则决定多卡互联的效率。**在实际企业生产环境中，多为服务器内多卡互联的场景，互联带宽决定了服务器内多张 AI 计算卡之间的数据传输峰值速率。

**硬件选型需要综合考虑推理性能、并发需求和投入成本。**在企业级私有化部署的算力硬件选型中，除了需要满足显存容量的最低要求，还需要综合考虑模型推理的性能和效率（多少 token/s）以及并发需求量（多少并发量），具体包括 AI 计算卡的数量以及计算能力、显存带宽和互联带宽等参数，此外可能还需要考虑生产/开发/测试环境的隔离以及安全性与高可用性冗余等因素。根据拓维信息官方公众号，企业级部署 DeepSeek-R1-70B 模型的推荐配置为 512G 显存容量，相当于 8 张华为 Ascend 910B （显存容量：64GB）的计算性能。

**图 15：拓维信息大模型一体机方案一览**


*资料来源：拓维信息*

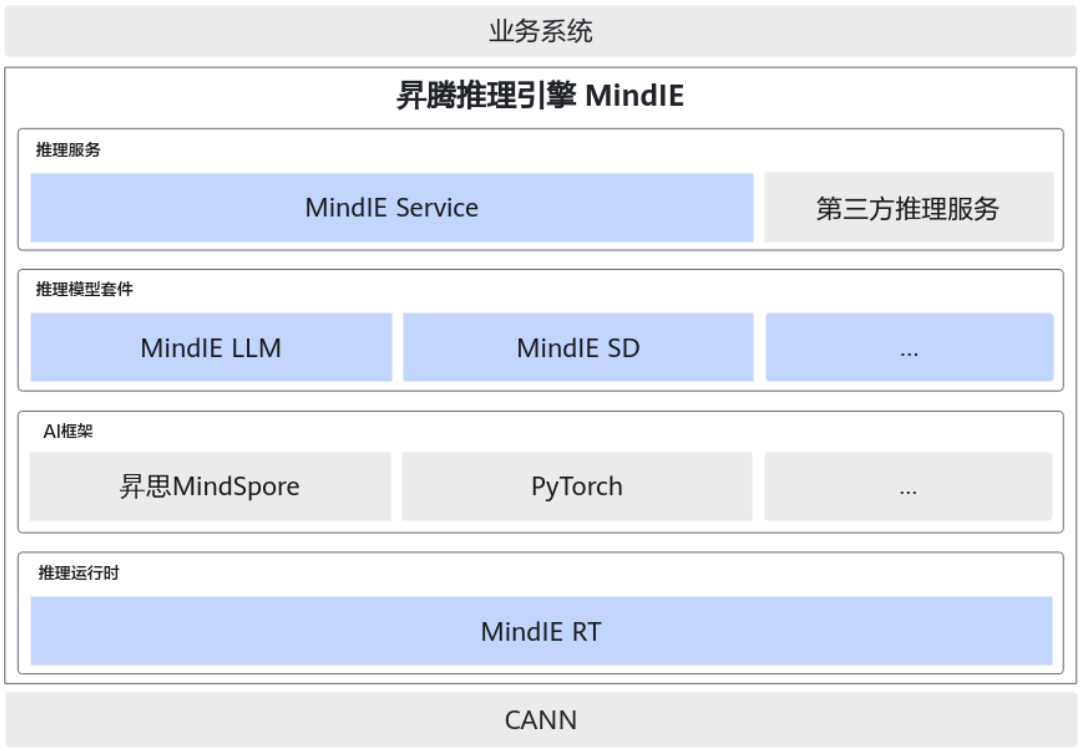
## 2.4 私有化部署方案的选型考虑三：配套生态及软件栈支持

**AI 计算卡的配套生态及软件栈直接影响算力利用效率，同样很大程度上决定 AI 大模型的推理性能和效率。**配套生态及软件栈支持主要包括算力硬件的固件及驱动和面向 AI 大模型部署的各类配套软件包，其决定了算力使用效率、算力兼容性、模型部署及后续维护更新的难易程度，也是 AI 大模型部署解决方案选型时所考虑的重要一环。

**AI 计算卡的固件及驱动决定了其底层计算效率，由芯片厂商提供与维护。**以华为昇腾为例，固件的主要功能包括昇腾计算芯片自带的 OS、电源器件和功耗管理器件控制软件，分别用于后续加载到 AI 处理器的模型计算、处理器启动控制和功耗控制。驱动主要用于管理查询昇腾 AI 处理器，同时为上层 CANN 软件提供处理器控制、资源分配等接口。

**配套软件包的作用在于帮助开发者优化基于 AI 计算卡训练和推理的效率和流程，更方便快捷地开发 AI 应用。**以华为昇腾硬件平台为例，部署 Deepseek-R1 时可能需要的配套软件包有异构计算架构（CANN）、推理引擎（Mindle）、集合通信库（HCCL）、基础设施管理平台（DCS 套件）等。

**图 16：昇腾配套软件包**

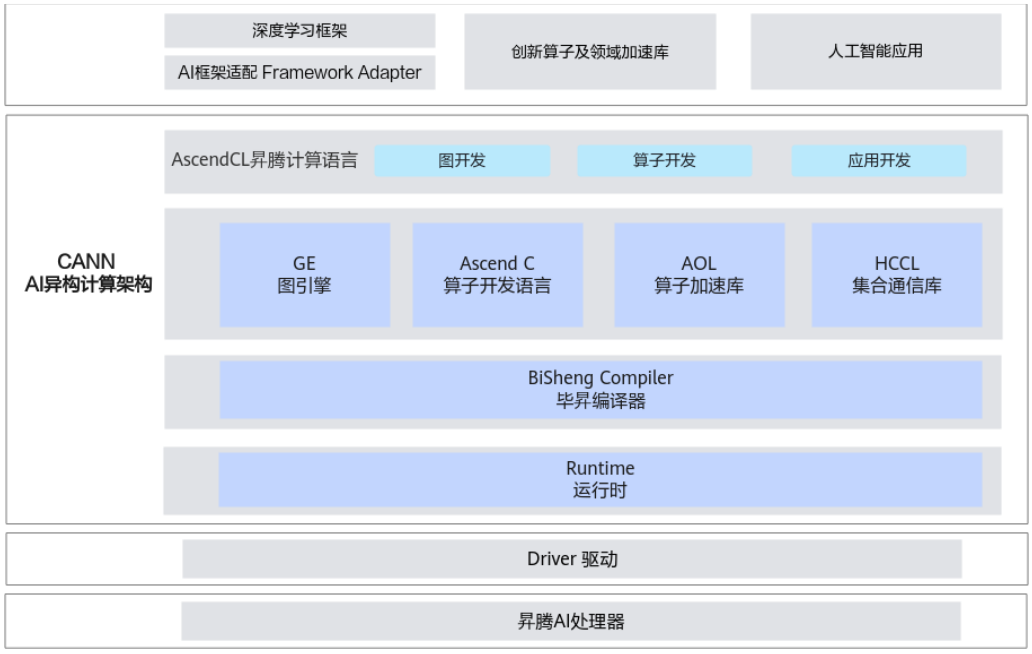



*资料来源：昇腾社区*

**异构计算架构：**整合 CPU、GPU、NPU 等不同处理器协同工作的计算模式，通过分工协作（如 GPU 加速并行计算、CPU 处理逻辑控制）来最大化硬件效能，适配 AI 大模型对海量算力的需求。典型代表包括英伟达 CUDA、华为昇腾 CANN。以 CANN（Compute

Architecture for Neural Networks）为例，其是昇腾针对 AI 场景推出的异构计算架构，向上支持多种 AI 框架，包括 MindSpore、PyTorch、TensorFlow 等，向下服务 AI 处理器与编程，发挥承上启下的关键作用，是提升昇腾 AI 处理器计算效率的关键平台。

**图 17：CANN 软件架构**




*资料来源：昇腾社区*

**推理引擎：**专为模型部署设计的优化工具，将训练模型转换为硬件高效执行的格式，集成量化压缩（FP32→INT8）、算子融合、内存复用等技术，显著降低推理延迟与资源消耗。典型代表包括 vLLM、SG-Lang、英伟达 NIM、华为 MindIE。以 MindIE（Mind Inference Engine，昇腾推理引擎）为例，其是华为昇腾针对 AI 全场景业务的推理加速套件。通过分层开放 AI 能力，支撑用户多样化的 AI 业务需求，使能百模千态，释放昇腾硬件设备算力。向上支持多种主流 AI 框架，向下对接不同类型昇腾 AI 处理器，提供多层次编程接口，帮助用户快速构建基于昇腾平台的推理业务。

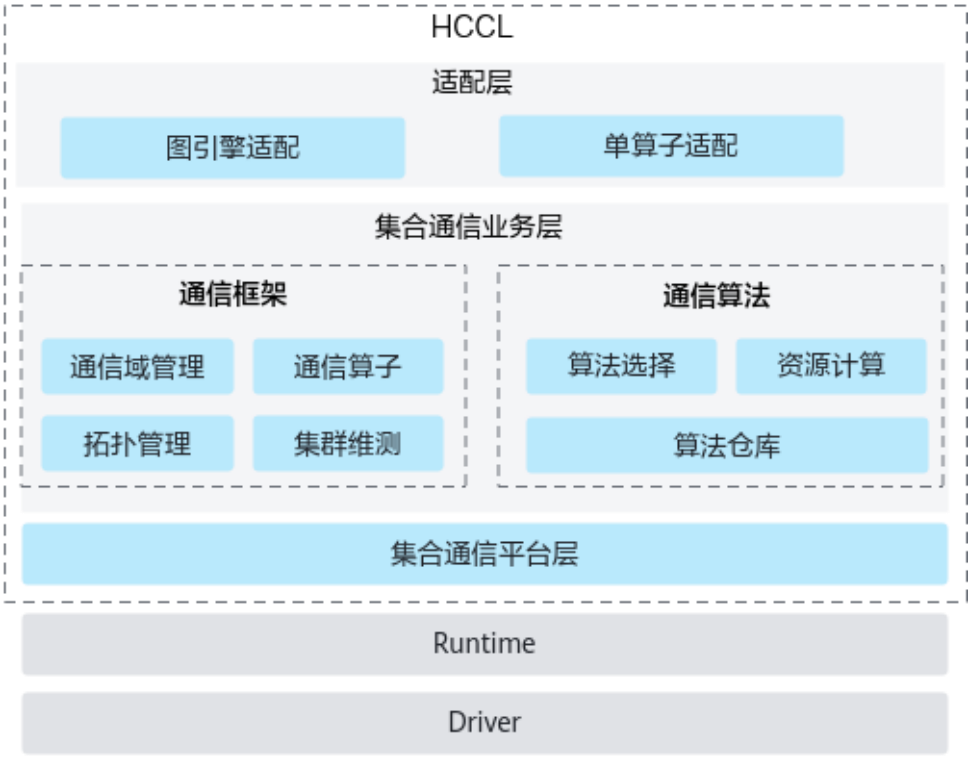
**图 18：MindIE 组件介绍**

| 名称 | 说明 |
| --- | --- |
| MindIE Service | MindIE Service针对通用模型的推理服务化场景，实现开放、可扩展的推理服务化平台架构，支持对接业界主流推理框架接口，满足大语言模型、文生图等多类型模型的高性能推理需求。MindIE Service包含MindIE MS、MindIE Server、MindIE Client和MindIE Benchmark四个子组件。<br>• MindIE MS：提供服务策略管理和运维能力；<br>• MindIE Server：作为推理服务端，提供模型服务化能力；<br>• MindIE Client：提供服务客户端标准API，简化用户服务调用；<br>• MindIE Benchmark：提供测试大语言模型在不同配置参数下推理性能和精度的能力。<br>MindIE Service向下调用了MindIE LLM组件能力。 |
| MindIE LLM | MindIE LLM（Mind Inference Engine Large Language Model，大语言模型）是针对大模型优化推理的高性能SDK，包含深度优化的模型库、大模型推理优化器和运行环境，提升大模型推理易用性和性能。 |
| MindIE SD | MindIE SD（Mind Inference Engine Stable Diffusion，Diffusion系列大模型）是MindIE解决方案下的多模态生成推理框架组件，其目标是为多模态生成系列大模型推理任务提供在昇腾硬件及其软件栈上的端到端解决方案，软件系统内部集成各功能模块，对外呈现统一的编程接口。 |
| MindIE Torch | MindIE Torch对接PyTorch框架，提供PyTorch模型推理加速能力。PyTorch框架上训练的模型利用MindIE Torch提供的简易C++/Python接口，少量代码即可完成模型迁移，实现高性能推理。MindIE Torch向下调用了MindIE RT组件能力。 |
| MindIE RT | MindIE RT（Mind Inference Engine Runtime，推理引擎运行时）基于昇腾异构计算架构CANN提供图引擎能力，当前支持小模型推理场景，更多关于推理运行时能力的构建工作正在进行中，敬请期待。 |

*资料来源：昇腾社区*

**集合通信库：**面向分布式训练的底层通信优化库，提供 AllReduce（梯度聚合）、Broadcast（参数同步）等高性能接口，利用 RDMA/NVLink 高速互联技术降低多节点通信延迟。典型代表如英伟达 NCCL、华为 HCCL。以 HCCL（Huawei Collective Communication Library）为例，其是基于昇腾 AI 计算卡的高性能集合通信库，提供单机多卡以及多机多卡间的数据并行、模型并行集合通信方案。

**图 19：集合通信库软件架构图**




*资料来源：昇腾社区*

**基础设施管理平台：**集成了算力硬件虚拟化、异构算力管理、资源分配、弹性扩缩容、运维管理等一系列功能的 AI 大模型工具箱，支持 AI 大模型的全生命周期管理。市场参与者包括芯片厂商、云厂商、ICT 厂商等，典型代表包括英伟达 DGX SuperPOD、华为 DCS 套件、京东云 vGPU 算力池化平台、新华三灵犀平台。以华为 DCS 套件为例，其通过整合 ICT 硬件及进行系统级优化，提供统一运维管理、硬件资源虚拟化、异构算力资源管理和调度、灾备和安全等功能及服务。

**图 20：DCS 组件架构**


*资料来源：《华为 DCS 数据中心虚拟化解决方案技术白皮书》*

# 3 私有化部署的选型考虑和部署流程颇为复杂，大模型一体机契合政企私有化部署需求

## 3.1 大模型一体机：软硬结合，通电即用的服务器

**私有化部署方案的选型考虑和部署流程颇为复杂，大模型一体机应运而生。**选型考虑中通常会面临硬件型号多且复杂、与原有 IT 设施的适配性、与实际业务场景需求的匹配度等等问题，部署流程中通常会出现环境配置多且复杂、AI 软件栈多且复杂、依赖版本配套复杂、性能调优复杂等等问题。私有化部署的高复杂性和高门槛催生了大模型一体机的服务器形态，这类服务器通过预集成算力硬件+优化软件栈+预装模型，将 AI 计算卡、配套生态及软件栈、模型算法、数据安全等核心要素深度融合，从而实现软硬结合，通电即用，一键部署 AI 大模型。

**一般政企不具备较成熟的 AI 基础设施团队，而大模型一体机有望解决私有化部署中的痛点难点，契合政企需求。**央国企、政务机构、学校、医院等泛政府类单位大多不具备较高水平的 AI 基础设施团队，而大模型一体机有望解决私有化部署中硬件选型难、软件适配慢、调优成本高等痛点难点，构建从硬件到软件、从开发到运维的全生命周期技术闭环，让政企无需组建专业团队即可实现敏捷部署。

**图 21：拓维信息一体机三大优势**


*资料来源：拓维信息公众号*

**图 22：拓维信息一体机使能体系**


*资料来源：拓维信息公众号*

## 3.2 私有化部署所需服务器（一体机）开支测算

**私有化部署需求增加有望推高大模型一体机销量。**根据初步测算，现阶段央国企、政务机构、学校、医院私有化部署 AI 大模型所需的服务器（一体机）开支空间约在 1000 亿元左右，且随着 AI 应用场景逐渐拓宽，服务器（一体机）需求仍有较大提升空间。

**核心假设：**

1）私有化部署的下游客户主要是央国企、政务机构、学校、医院等泛政府类单位。

2）参考拓维信息的推荐配置表和草根调研信息，DeepSeek-70B 模型为企业级私有化部署的入门款，能够应付大部分业务场景需求，推荐配置约等效为 1 台配有 8 张华为 Ascend 910B （显存容量：64GB）的服务器（一体机），多用户并发数为 64 路左右，价格假设为 140-180 万元不等；DeepSeek-671B 模型的推荐配置则为 2-4 台服务器（一体机）组成集群。

3）央国企：根据浙江国资数据，截至 2022 年，我国大约共有 29.1 万家央国企，假设 10%左右的央国企有私有化部署需求，每个配备 1 台服务器（一体机）。

4）政务机构：目前主要应用于政务服务场景，假设全国 14 亿左右人口中，每 10 万人配备一台。

5）三甲医院：应用场景较为复杂，包括医疗辅助诊断、医院智能运营、药物研发赋能等等，假设每个三甲医院配备 4 台。

6）高等院校：应用场景较为复杂，包括智慧校园、智慧教学、科研赋能等等，假设每个高等院校配备 4 台。

**表 2：私有化部署所需服务器开支测算**

| | 假设 | 私有化部署所需服务器数量（台） | 裸金属占比 | 价格（万） | 一体机占比 | 一体机价格（万） | 市场空间（亿元） |
|---|---|---|---|---|---|---|---|
| 央国企 | 29.1 万家央国企×10%渗透率 | 29100 | 10% | 140 | 90% | 180 | 512 |
| 政务机构 | 14 亿左右全国人口÷10 万人配备 1 台 | 14000 | 10% | 140 | 90% | 180 | 246 |
| 三甲医院 | 3855 家三级医院×每个配备 4 台 | 15420 | 10% | 140 | 90% | 180 | 271 |
| 高等院校 | 3117 所高等院校×每个配备 4 台 | 12468 | 10% | 140 | 90% | 180 | 219 |
| 总计 | | 70988 | 10% | 140 | 90% | 180 | **1249** |

*资料来源：浙江国资，国务院，国家卫健委，中国政府网，财信证券*

*注：央国企指的是国资委、财政部履行出资人职责的中央企业、中央部门和单位所属企业以及 36 个省（自治区、直辖市、计划单列市）的地方国有及国有控股企业、新疆生产建设兵团所属国有及国有控股企业，不含国有一级金融企业*

# 4 相关公司

## 4.1 紫光股份

**紫光股份旗下新华三集团于近期发布集成"智算-算法-治理"全要素的 DeepSeek 大模型一体机——灵犀 Cube。**依托自主研发的异构算力架构，可实现多品牌 GPU 兼容适配（包括 NVIDIA、AMD 和国产芯片），支持从 14B 到 671B 参数大模型的单机推理及单机训推一体服务。

新华三一体机提供基于“硬件+算法+平台”三位一体的全栈 AI 解决方案。

**1）基础设施层：**提供预集成的智能算力集群，实现网络、存储、安全的统一编排。

**2）算法服务层：**预置 DeepSeek 基础大模型及行业优化版本，支持后续模型的蒸馏和微调。

**3)应用使能层：**内置 H3C AlStore 智能资产平台，免费开放超过 1000 款开源数据集、模型及镜像，商用资源一键付费获取，使用户获取上百 GB 镜像和模型等 AI 资产的时间从几天缩短到几小时。内置 Web 前端可视化操作界面，提供标准化 API，减少重复开发工作量，大大降低操作门槛。

**图 23：新华三灵犀 Cube 一体机**


*资料来源：新华三公众号*

**新华三已成功推动 DeepSeek 大模型在政府、教育、医疗、金融、制造等多领域实现本地私有化部署与深度场景落地，形成显著的行业示范效应。**在教育领域，助力华东师范大学、南昌大学等双一流高校应用于科研文献分析、智能教学评估及校园管理，提升学术效率并保障数据安全；在医疗领域，为无锡中医医院等三甲医院部署复杂医学模型训练与辅助诊断系统，优化临床决策效率与科研能力；金融行业方面，通过私有化部署实现智能客服、信贷风控等场景的精准适配，多家商业银行完成垂直业务智能化升级；政府领域助力杭州、郑州等地构建 AI 政务处理平台，覆盖公文自动化、民生服务及决策支持，形成可复制的智慧城市治理范式；在央企及制造业中，依托大模型的自然语言处理与任务拆解能力，推动研发设计、供应链管理、运营决策等全流程智能化转型。其技术优势体现在数据隐私保护、场景适配性与规模化落地能力，有效助推各行业数字化与智能化升级。

我们预计公司 2024-2026 年实现营业收入 824.17/918.01/1018.94 亿元，同比增长 6.61%/11.39%/10.99%，实现归母净利润 23.39/28.69/35.68 亿元，同比增长 11.24%/22.63%/24.38%，对应 EPS 为 0.82/1.00/1.25 元，对应当前价格的 PE 为 36/29/24 倍。维持“买入”评级。

## 4.2 浪潮信息

**浪潮信息推出专为大模型推理优化设计的元脑 R1 推理服务器。**搭载 EPAI 企业大模型开发平台，具备模型管理、知识检索、提示词工程、智能体等全流程开发技术栈，集成伙伴和客户的企业、医疗、教育、公共事业等行业高质量数据，形成多行业场景示范模板。具体特性如下：

**1）单机高性能支持原版 671B 模型：**元脑企智 DeepSeek 一体机基于元脑 R1 推理服务器，提供 1128GBHBM3e 高速显存和 4.8TB/s 的显存带宽，单机即可支持

Deepseek-R1671B 模型在 FP8 精度下的全量模型推理，实现单用户解码最高 33tokens/s 及最大用户并发超 1000 的优异性能表现。

**2）完备工具链协助快速上线：**元脑企智 DeepSeek 一体机搭载 EPAI 企业大模型开发平台，集成了完备的大模型应用开发工具链，涵盖资源管理、知识检索、提示词工程等基础功能，内置智能助手、工作流、商业智能、RAG 应用等应用开发组件，为企业构建 AI 应用超级工作台。预置 DeepSeek 大模型和全栈工具链，使得元脑企智 DeepSeek 一体机最快仅需 3 小时，即可完成从硬件上电到千亿级模型服务上线，实现真正的开箱即用。

**3）客户业务数据专业化集成：**元脑企智 DeepSeek 一体机具备高效完备的数据处理工具，能够帮助企业基于海量数据快速构建专业知识库，数据处理效率提升 40%，利用高精度知识库检索能力，可整合多源场景数据，实现精准知识检索与生成。一体机通过大模型和 RAG 知识库的协同有效解决了通用模型的知识盲区问题，通过实时检索最新数据抑制模型幻觉，使查询准确率提升 95%以上，可适用于金融、医疗等知识高精度需求场景，同时支持数据加密和安全隔离，保障知识库的安全合规。

**4）多行业场景示范模板：**元脑企智 DeepSeek 一体机接入了浪潮信息与合作伙伴联合开发的多种行业大模型应用模板，覆盖金融风控、工业质检、医疗影像等场景，无需进行复杂编程，而是以可视化引导的方式实现低代码快速微调，快速构建智能应用，大幅缩短开发周期，开发效率提升 5 倍以上。

**图 24：浪潮信息元脑 R1 推理服务器**


*资料来源：元脑服务器公众号*

预计公司 2024-2026 年实现营业收入 1051.54/1209.93/1392.30 亿元，同比增长 59.65%/15.06%/15.07%，归母净利润 21.78/26.14/31.89 亿元，同比增长 22.16%/20.03%/21.99%，EPS 为 1.48/1.78/2.17 元，对应当前价格的 PE 为 40/34/28 倍。维持“增持”评级。

## 4.3 中科曙光

中科曙光的 DeepSeek 人工智能一体机集多形态曙光高端计算服务器、高效能基础模型、全流程 AI 工具链于一体，并内置曙光自研 AI 管理平台 SothisAI3.0，支持从 10 亿级参数模型推理到 1000 亿级参数模型训练的 AI 全场景需求，支持多种大模型适配、智能异构算力调度、数据加密计算等功能，为科研机构、企业、政府等用户提供一站式、开箱即用的 AI 解决方案。

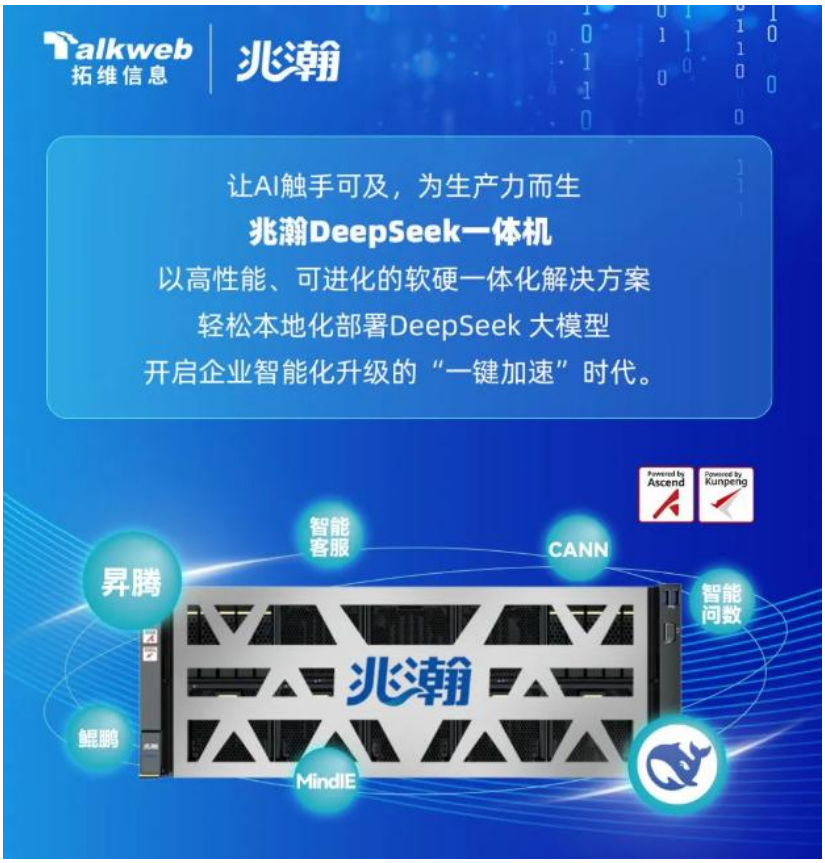
**图 25：中科曙光 DeepSeek 大模型超融合一体机**


*资料来源：天津市软件行业协会*

## 4.4 拓维信息

拓维信息“兆瀚”系列 AI 服务器及相关产品已全面完成与 DeepSeek-R1/V3 系列大模型的深度适配，基于华为鲲鹏+昇腾处理器，全面支持 DeepSeek 实现本地化快速部署，为行业用户提供从训练到推理、从云端到边缘的完整部署能力。同时，公司为客户全面接入 DeepSeek 提供软硬一体化的规划咨询和部署服务，助力客户加速 AI 应用落地。

**图 26：拓维信息兆瀚 DeepSeek 一体机**




*资料来源：拓维信息公众号*

# 5 风险提示

宏观经济波动风险；AI 技术发展不及预期；下游客户 AI 算力支出意愿不及预期风险；供应链风险；服务器（一体机）行业竞争加剧风险。

## 投资评级系统说明

以报告发布日后的 6—12 个月内，所评股票/行业涨跌幅相对于同期市场指数的涨跌幅度为基准。

| 类别 | 投资评级 | 评级说明 |
|---|---|---|
| 股票投资评级 | 买入 | 投资收益率超越沪深 300 指数 15%以上 |
| | 增持 | 投资收益率相对沪深 300 指数变动幅度为 5%—15% |
| | 持有 | 投资收益率相对沪深 300 指数变动幅度为-10%—5% |
| | 卖出 | 投资收益率落后沪深 300 指数 10%以上 |
| 行业投资评级 | 领先大市 | 行业指数涨跌幅超越沪深 300 指数 5%以上 |
| | 同步大市 | 行业指数涨跌幅相对沪深 300 指数变动幅度为-5%—5% |
| | 落后大市 | 行业指数涨跌幅落后沪深 300 指数 5%以上 |

## 免责声明

本公司具有中国证监会核准的证券投资咨询业务资格，作者具有中国证券业协会注册分析师执业资格或相当的专业胜任能力。

本报告仅供财信证券股份有限公司客户及员工使用。本公司不会因接收人收到本报告而视其为本公司当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发送，概不构成任何广告。

本报告信息来源于公开资料，本公司对该信息的准确性、完整性或可靠性不作任何保证。本公司对已发报告无更新义务，若报告中所含信息发生变化，本公司可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指投资及服务可能不适合个别客户，不构成客户私人咨询建议。任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司及本公司员工或者关联机构不承诺投资者一定获利，不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此作出的任何投资决策与本公司及本公司员工或者关联机构无关。

市场有风险，投资需谨慎。投资者不应将本报告作为投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。



本报告由财信证券研究发展中心对许可范围内人员统一发送，任何人不得在公众媒体或其它渠道对外公开发布。任何机构和个人（包括本公司内部客户及员工）对外散发本报告的，则该机构和个人独自为此发送行为负责，本公司保留对该机构和个人追究相应法律责任的权利。

## 财信证券研究发展中心

网址：stock.hnchasing.com
地址：湖南省长沙市芙蓉中路二段 80 号顺天国际财富中心 28 层
邮编：410005
电话：0731-84403360
传真：0731-84403438