# 深信服AI创新平台

让AI创新更简单、更高效

## 企业 AI 建设新范式

进入 AI 2.0 时代，DeepSeek 等开源模型带来效果提升和成本优势，释放场景化 AI 创新潜能，驱动 AI 应用快速落地。

**DeepSeek 等开源模型的技术优势<br>奠定 AI 爆发基础**

- **高性能突破：**在高难度推理问题、代码能力、数学能力等方面，拔得头筹，位居第一
- **低成本革命：**训练成本断崖式下降，推理成本极致压缩，API 近乎免费
- **强开源基因：**MIT 协议，更为宽松友好，支持商用，支持修改和二次开发

**构建基于 DeepSeek 等开源模型的<br>AI 推理应用是企业智能化转型的首选路线**

- AI 发展重心由训练转向推理，本地模型部署后的推理成本降低一个数量级
- 私有化部署大模型服务组织内部业务成为优先选择
- 大模型应用会向中小企业延伸、场景快速扩展，AI 应用的奇点时刻将加速到来
- DeepSeek 等开源模型生态圈已快速形成并规模扩张

**AI 应用效果为王，Agentic AI system<br>正走向主流**

- Agentic 系统是 1-2 年内构建应用的主流形态，多智能体效果＞智能体效果＞模型效果是确定的趋势
- 端到端的黑盒 Agent 单体短期无法满足企业特定场景的精准控制要求，需根据应用场景来调整可预测性和自主性
- 复杂智能体应用可以通过多次调用大小模型、垂域领域模型调用来提升效果（反思、工具调用、工作流处理等）

**从拥抱 AI 的一开始到大范围使用<br>都要考虑 ROI**

- 基于场景选择模型，基于模型选择显卡，关注平台性能优化效果，用有限的 AI 算力发挥更大的效能

## 用户获取 DeepSeek 等 AI 能力面临挑战

**研发门槛高**

- **上手难：**>30 天搭建一套企业级大模型推理或训练环境
- **踩坑多：**GPU 硬件故障、推理或训练中断 ...
- **效果差：**AI 应用问答回复准确率低

**管理复杂度高**

- 通算到智算，基础设施复杂性大幅提高，使运维成本进一步提高
- 硬件算力差异大，软件技术栈多，管理技术要求高
- 模型类型多、迭代快，维护更新工作量大

**推理性能要求高**

- AI 应用向长输入、高并发、高可靠演进，显存需求指数级增长，“缺卡少算”依然是主要矛盾
- 单纯通过提升硬件性能无法有效提升应用体验，硬件性能不等于模型性能，也不等于应用性能，需要软硬一体化的上下联动调优

**成本代价巨大**

- 专业人才能力要求高、培养成本高，供需失衡、行业竞争激烈
- 基础设施硬件资源成本、存储管理成本高，平台工具成本、开发运维成本高

**应用落地难、效果差**

- 复杂场景应用开发难度大，对接各种业务系统复杂度高，资源隔离和权限管理难；
- 现实场景数据复杂，应用初始效果差；效果优化门槛高、难度大、成本高

**安全风险高**

- 模型泄漏比数据泄漏危害更大
- 企业专属大模型泄漏会丢失知识产权优势，阻碍企业发展



## 深信服 AI 创新平台承载企业级 AI 大模型方案

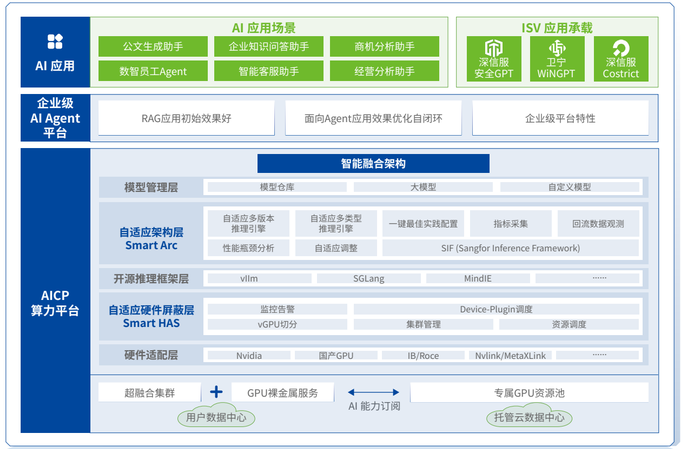
基于线上线下一朵云方案，通过 AI 基础设施和 AI 平台能力，提供企业 AI 大模型承载方案。线下用户在原有平台扩展 GPU 资源即可构建 AI 大模型，线上用户一键订阅即可获取 AI 大模型服务。

**承载 AI 大模型，越用 ROI 越高，越用越安全**

基础设施层深度兼容英伟达和国内厂商 GPU，通过统一管理实现大模型在异构算力上的平滑迁移调度，同时支持 GPU 按照算力和显存声明式细粒度切分，提升 GPU 资源利用率。持续兼容 DeepSeek-R1、Qwen 等主流开源大模型，自定义模型功能支持部署垂域 AI 大模型及传统小模型，实现大小 AI 模型混合部署，并通过对并行调度器、推理缓存、负载均衡的性能优化、昇腾 NPU 定向优化，大幅提升大模型推理并发吞吐性能，让大模型越用 ROI 越高。支持模型服务 API 多 Key 管理、token 级别监控、服务限流能力及用量观测看板。支持模型微调训练等能力，降低大模型使用门槛，并通过模型动态加密保障专属模型的知识产权，越用越安全。

**让非 AI 专业人员能简单快速创建初始效果好、可不断评估调优的 Agent 应用**

帮助用户简单快速地构建高质量的 AI Agent 应用，并提供自动化测试数据生成和质量评估、实现自动效果评测，让应用效果持续提升。




深信服 AI 创新平台承载企业级 AI 大模型应用方案



## 核心价值及优势

**AI 实践落地综合 ROI 提升 2-5 倍**

- 自研自适应架构层，通过智能分块调度优化、上下文智能感知缓存优化、参考文本 Cache 等多级 Cache 优化技术提升 3-5 倍推理性能
- 自研的检索投机推理技术、多实例智能负载、推理阶段分离、大小模型混合调度等智能负载调度技术让大并发性能翻倍
- 通过结合业务数据的稀疏化、业务感知量化 BAQ、混合精度推理缓存压缩技术提供翻倍利用率

**智算管理能力大幅提升**

- Sangfor-vGPU 技术可按显存大小快速调整 GPU 资源分配，解决只能整卡调用难题，实现大小 AI 模型混合部署，优化算力资源利用率
- 结合深信服 AI 创新平台、超融合，实现通算和智算统一管理。订阅线上 AICP 一键部署模型服务，完成 GPU 服务器及公网 IP 等资源的快速下发和 DeepSeek-R1 系列模型搭建，快速构建线上线下网络互通方案，提供多重安全防护，全面构建网络、身份、资源、大模型防护体系，保证数据安全

**开放性：向下解耦显卡，向上广泛兼容模型**

- 与 GPU 厂商深度合作调优，持续兼容异构主流 GPU，实现高质量异构显卡管理，同时满足高性能及合规诉求
- 支持 DeepSeek-R1、Qwen 全系列开源模型，持续适配各种开源新模型

**模型安全，有保障**

- **模型动态加密技术：**采用结构加密（增加伪分支），权重加密（替换为伪权重），算子加密（增加伪算子）的方式对模型结构进行混淆加密，保护模型运行态安全，且加密算法不依赖硬件，性能损耗 <5%，正确解密时推理结果不变
- **自适应加密算法：**能够针对不同模型，自适应调整和组合不同强度的结构加密 / 权重加密 / 算子加密技术，保护模型安全的同时，也保障了部署可行性和推理性能
- **保护模型运行态安全：**大模型经过动态加密后，当且仅当密钥设置正确时模型才可以正常运行，当模型被盗用后会因为扰动而输出乱码，使恶意盗窃者无法盗取模型推理内容

**AI 应用开发，效率高、效果好**

- RAG 应用初始效果更好，超过同类开源产品
- 评估调优，助力企业自助实现效果评估，持续提升应用效果，大幅减少对外购调优服务的依赖
- 具备多种企业级平台特性，满足企业个性化的数据隔离、权限管理（知识库和应用）、系统对接要求，有效支撑生产级应用落地

## AI 大模型推荐配置

**根据应用场景特点使用不同量级 AI 大模型：**

**严谨 AI 应用场景、重要生产应用场景，大参数量级开源 MOE 模型是最优选。**

- 如 DeepSeek-R1、Kimi2、GLM4.5 等，在 AI 编码、销售助手、toC 端问答助手等通用场景的理解和推理过程，表现出更高准确率。

**通用 AI 应用场景下，32B 稠密模型效果已经被广泛验证认可。**

- 如行政问答、合同审查、OA 系统智能助手等场景，都基于 32B 稠密模型完成大量交付应用。

**以下是 AI 大模型部署推荐配置，请根据大模型业务需求酌情参考**

| 蒸馏模型名称 | 承载最小配置 | 并发 |
|---|---|---|
| DeepSeek-R1-671B（FP8） | 1 台 H20*8（总显存 1152GB） | 256 |
| DeepSeek-R1-671B（混合精度） | 4090D*8（总显存 192GB）（CPU 架构为 AMD） | 8 |
| Qwen3-235B-A22B（FP8） | 1 台 H20*4（总显存 576GB） | 256 |
| **蒸馏模型名称** | **常规最小配置** | **并发** |
| DeepSeek-R1-Distill-LLama-70B（BF16） | 4090D*8（总显存 192GB） | 128 |
| DeepSeek-R1-Distill-Qwen-32B（BF16） | 4090D*4（总显存 96GB） | 256 |
| Qwen3-32B（BF16） | 4090D*4（总显存 96GB） | 256 |
| DeepSeek-R1-Distill-Qwen-14B（BF16） | 4090D*2（总显存 48GB） | 256 |
| DeepSeek-R1-Distill-Qwen-7B（BF16） | 4090D*1（总显存 24GB） | 256 |

## 大模型应用场景

| ISV | 金融 | 医疗 | 企业 | 交通 |
| --- | --- | --- | --- | --- |
| 应用 Copilot 化 | 知识问答（投研领域、产品投教） | 病历生成 | AI 物料选型 | TFDS |
| 代码自动生成 | 保险销售机器人 | 辅助导诊 | 工程设计知识库 | AI 客运服务 |
| 专业知识库 | 智能风控 | 内涵质控 | 设备维修 AI 助手 | AI 运维 |

## 客户案例

### 某大型新能源科技股份有限公司案例

- 该企业专注于锂电池材料研发与生产，在推进 AI 建设中面临算力不足、知识沉淀有限及流程复杂工作效率低等挑战。
- 基于 3 台共 24 卡 L20 GPU 裸金属服务器，建设深信服 AI 创新平台，系统性规划 50 个 AI 应用场景，陆续开发膜材料知识问答助手、订单设计生成系统、采购合同 AI 审计、OA 轻工作流等应用。
- 目前，人事行政智能问答助手已在企业微信工作台上线，单项任务处理时间由小时降至分钟，显著提升工作效率。项目实现了从算力建设到智能化落地的闭环，为制造业 AI 转型提供了可复制路径。

### 某医疗健康集团案例

- 该集团依托深信服 AI 创新平台（AICP）统一管理 4 节点 32 卡 H100 算力资源，部署 DeepSeek -V3-671B 大模型，上线 AI 智能移动端 APP，为数十万会员提供报告解读、在线问诊、用药咨询等智能健康服务。
- 随着用户增长，系统出现性能瓶颈，限制了会员接入规模。深信服基于 AICP 算力平台优化能力，针对不同业务对大模型调用特征进行对 DeepSeek -V3-671B 大模型服务定向调优，使 APP 并发能力翻倍、系统稳定性显著提升。
- 项目显著提高了算力利用率与用户承载规模，AI 建设 ROI 翻倍，为医疗行业 AI 应用落地提供了标杆经验。

### 某省港口集团案例

- 项目一期以建设高性能 H20（141G）GPU 资源池为目标，采用 8 卡整机方案，通过 AICP 实现 GPU 算力池化与显存级切分，提升大小模型混合部署的资源利用率。
- 集团同步部署多模态模型，测试知识库问答、SQL 分析、多模态生成等场景；结合 HCI 平台，实现通算与智算统一管理，并以“单节点起步、按需扩展”模式降低建设成本。
- 二期将规划国产化异构智算资源池，AICP 提供跨架构统一调度与管理，降低运维成本，推动大模型深度融入港口生产与经营，助力港口数字化转型升级。



SANGFOR 深信服科技

让每个用户的数智化更简单、更安全

深信服官方微信　深信服移动官网

深圳市南山区西丽街道西丽社区仙洞路16号深信服科技大厦
售前咨询：400-806-6868　售后服务：400-630-6430
邮编：518055　邮箱：market@sangfor.com.cn

SANGFOR 深信服科技


让每个用户的数智化更简单、更安全

# 深信服AI创新平台解决方案