# 云融智算魔方产品矩阵

| 擎天魔方 | 炫彩魔方 |
| --- | --- |
|  |  |
| **睿捷魔方** | **先锋魔方** |
|  |  |

## 产品概述

云融智算魔方产品，采用多种模型优化技术，大幅降低模型大小适配多种异构 GPU 卡，构建了从数据中心、工作站到办公终端的全系列智算魔方产品矩阵，具备高性价比优势；集合硬软模一键部署的便捷优势，广泛适用于生成式大模型推理场景。在同等推理效率场景下，成本降低 10%以上，便于 AI 赋能各行各业。

该产品选用原生态组件，依托从底层算力、资源调度、AI 框架、镜像管理到模型部署的全栈 AI 能力，为企业高效构建智能化能力提供有力支持，可以在多个垂直行业领域发挥智能辅助效果：

- 在金融风控领域，可实时处理合同文档审计，仅需数分钟即可完成合同审计，生成精准风险评估报告；
- 在智能客服方面，能承载百人并发对话，实现录音、文字、关键信息摘要的同步记录；
- 在科研教育方面，帮助科研工作者完成海量论文的快速学习，并全面且准确的基于材料回答用户问题，帮助用户快速理解材料；
- 在内容生成领域，支持长篇小说、代码项目级别的创作，助力创作效率提升 10 倍。

## 产品规格

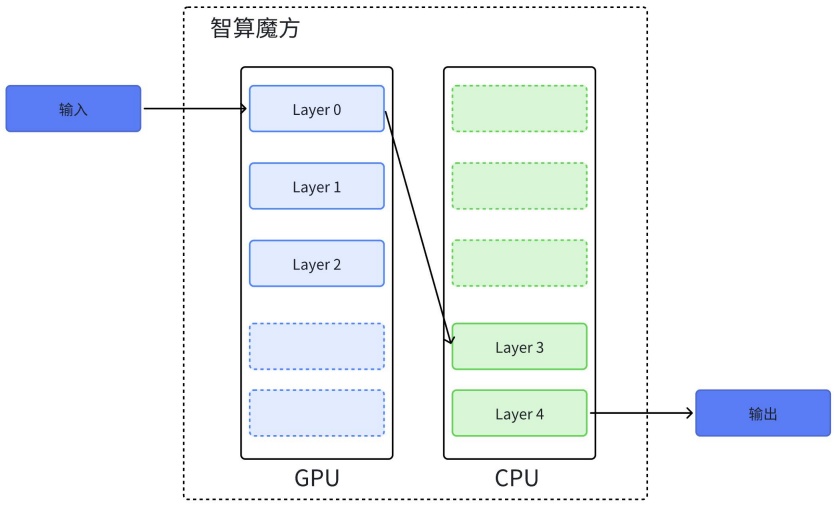
云融基于新型 GPU 高显存、UMA 等技术特点，面向不同的业务场景进行模型优化，整合大模型硬件与软件框架，构建高性价比智算魔方产品矩阵，在不同场景下提供对应的解决方案。

| 产品型号 | 核心技术 | 模型规格 | 形态 | 精度 | 适用场景 |
|---|---|---|---|---|---|
| 擎天魔方 | 高显存，海量记忆能力 | 671B 满血版本 | 服务器 | FP8 | 适合百人团队场景：支持几十人并发的文案写作、合同审计、票据审计、智能客服等场景 |
| 炫彩魔方 | 异构计算，充分利用 GPU 与 CPU 的算力 | 671B 满血版本 | 服务器 | FP8 | 适合 2-3 小团队场景：支持单人并发的文案写作、代码生成等场景 |
| 睿捷魔方 | 动态优化 | 671B 动优版本 | 服务器 | FP8~1.75bit 动态 | 适合小企业场景：支持知识库、代码生成、智能客服等场景 |
| 先锋魔方 | UMA 共享内存，高效数据搬移 | 32B 蒸馏版本 | 工作站 | FP16 | 适合单用户场景：支持知识库、代码生成等场景 |

## 产品特点

### 异构计算

CPU 与 GPU 异构计算，CPU 和 GPU 各自承担部分计算，利用 DeepSeek 模型的混合专家（MoE）架构的稀疏性，将非共享稀疏矩阵卸载至 CPU 内存处理，同时将稠密部分保留在 GPU 上，在系统内存和 GPU 显存各自存放部分模型，利用流水线并行及通信技术，充分利用异构算力，显著降低了显存需求，大幅降低部署成本。



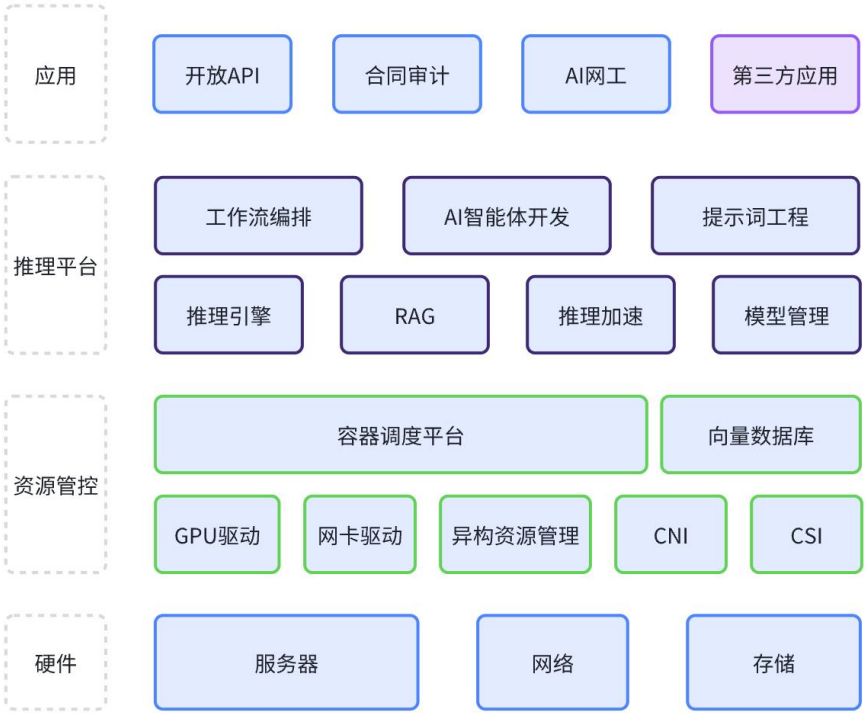
## 动态优化

动态优化是一种模型优化技术，与静态量化不同，在保证精度损失较小的情况下对模型进行分层优化压缩，适合对精度要求较高且输入数据分布差异较大的场景，能更好地保留模型原始精度。采用动态优化技术处理，实现了 Deepseek 满血版从 720GB 缩小到最低 131GB 。

- **关键层保持高精度：**初始的全连接层、下投影矩阵（down_proj）以及注意力模块等对模型稳定性至关重要的部分，保持较高精度（如 4 位或 6 位）。
- **MoE 层激进优化：**模型中约 88% 的权重位于混合专家（MoE）层，这些层可以容忍较低的优化精度，因此被优化到 1.5 至 2 位。
- **重要性矩阵校准：**根据每一层的特性动态调整优化精度，避免了无限循环或输出无意义结果等均匀优化常见的问题。

## 软硬件一体化交付

将高性能硬件、预训练的大模型以及全栈开发工具深度集成，封装为一体化设备。极大地简化了部署流程，用户无需自行安装繁琐的硬件驱动、底层框架和复杂的模型推理环境配置，实现大模型开箱即用。

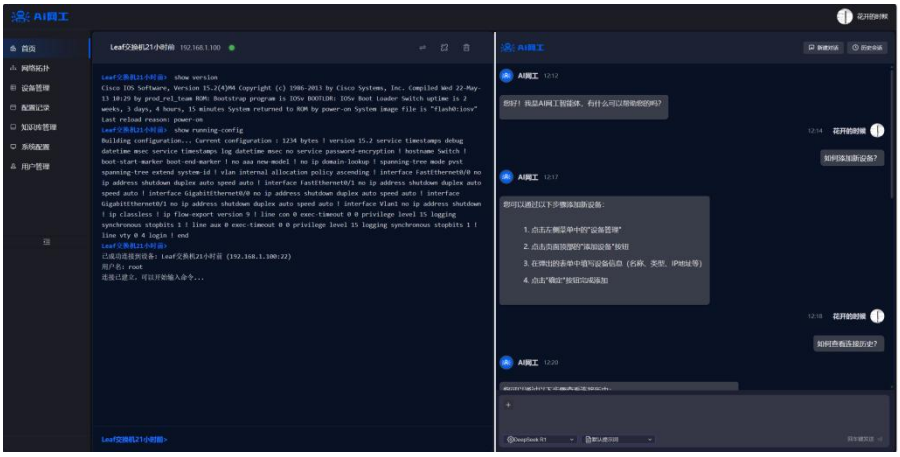



## 典型场景应用

### 港口物流大模型


湖北省港口物流大模型提供智能货运订舱系统、智能客服、智能竞价等港口行业核心业务。该模型已成功支撑湖北港口集团业务运营，显著提升了长江中游航运枢纽的数字化服务能力。

### AI 网工



网络运维智能体基于专业设备手册和海量运维案例，通过 AI 实现自动化配置、智能故障排查、文档生成和安全增强等功能，显著提升网络运维效率，减少人为错误。


## 苏州云融信息技术有限公司

地址：苏州工业园区科营路 2 号中新生态大厦
电话：400-998-7338
官网：www.sdnware.com