QUANXING 铨兴

打破显存墙

Llama-2 (70B) 模型

现行AI运算体系结构

Data Data Data Data
Data Data Data Data
Data Data Data Data
Data Data Data Data
Data Data Data Data
Data Data Data Data

- 24张80GB显存GPU
- 3台服务器集群
- 运算能耗超30kW
- 部署成本大于750万

QX AI大模型训练解决方案的运算体系结构

Data Data
Data Data

+

GPU可用显存空间提升20倍
搭载铨兴添翼 AI 扩容卡 ×2

- 4张20GB显存GPU+2张扩容卡
- 1台服务器工作站
- 运算能耗不到2kW
- 成本不到100万，节省超85%

AI训推一体机
LLM本地化部署解决方案

70B-180B 超大模型、高精度、全参微调
整机低于100万

推动 AI 完成最后一公里

70-180B 超大模型全参微调、训推一体、超低成本、私有化部署、极低功耗

政务+AI　科研+AI　教学+AI　法律+微调　金融+微调　代码+微调

智能政务、数字
政府、智慧城市

细分科研
领域专家

大幅降低高校
开课成本

企业专属合同专家
随时待命法律顾问

深度市场研究
权威风险分析

专属架构师
深度代码分析重构

显著降低配置成本、功率低

铨兴添翼AI解决方案所推荐的整机配置，在功率控制方面表现出色，其整体运算功率控制不到2kW，可以使用家用配置供电。在配置成本方面，4张中阶易采购的GPU显卡的整机配置（以70B大模型训练为例），总成本不到100万元，相较于传统方案节省了超过85%。

适配场景灵活多样

除了在低成本和低功率方面具有显著优势外，铨兴添翼AI解决方案更以其卓越的不受大模型尺寸限制的特性（70B-180B）脱颖而出。无论面对何种规模的模型，该方案均能够展现出强大的适应性和应对能力，为用户带来灵活多样的AI训练服务，广泛满足不同行业和场景的需求，展现其强大的应用潜力和价值。

保护数据隐私与安全

铨兴添翼AI解决方案作为一款软硬件结合的先进大语言模型(LLM)微调训练方案，致力于助力中小企业处理敏感机器学习(ML)数据，同时确保本地控制权的稳固保持。该方案能够有效杜绝数据外泄的可能性，从而显著提升数据安全保障水平，有力防范外部入侵与资料泄露等潜在风险。

ai@quanxingtech.cn

QUANXING 铨兴

联系我们

铨兴公众号

AI 训推一体机推荐配置|70B版

内存
铨兴 DDR5 R-DIMM
64GB ×8

固态硬盘
铨兴 SATA / NVMe eSSD
7.68TB ×1

AI扩容卡
铨兴添翼AI扩容卡 ×2

CPU
INTEL W5-3425 ×1

GPU
NVIDIA RTX 4000 Ada
20GB ×4

软件
训推一体软件中台
预置主流大模型

低于60分贝

AI训推一体产品对比

	铨兴添翼AI训推一体机 50-100万	方案A >750万	方案B >100万
软件生态	全参微调 添翼AI世界 大模型训推一体软件平台 超大模型110B-180B 大模型70B-72B 小模型6B-13B	全参微调 超大模型 ⊗ 大模型 小模型	全参微调 超大模型 ⊗ 大模型 ⊗ 小模型
硬件	 可支持 Nvidia主流计算显卡仅需80GB-192GB显存	24张顶级显卡 1920GB 显存	8张顶级显卡 640GB 显存
功耗	2 kW	30 kW	10 kW

铨兴添翼AI

核心产品能力

核心硬件

铨兴添翼AI扩容卡

铨兴 SATA/NVMe eSSD

铨兴 DDR5 R-DIMM

核心软件

添翼AI One
保姆式AI服务

专业AI解决方案
法务 | 财税 | AI教学 | 医疗...

专业AI服务
数据治理 |训练 | 部署 | 应用

添翼AI Pro
一站式AI应用解决方案

超自由度 AI 应用平台

添翼AI Base
面向具备 AI IT 的客户
FREE

添翼AI训练平台