

RoycomKT智算一体机



PART 01

RoycomKT智算一体机简介



痛点

满血大模型 如何落地？

价格高？

671B大模型所需要显存需达到**1500G**以上，部署一个满血版大模型价格需要**百万**以上？

运维难？

模型如何**部署**？如何进行有效部署与管理模型？如何一键替换模型，成功**调用模型**能力为企业效率助力？

落地少？

裸奔数据该如何放心使用？企业如何才能以**低成本**方式促进满血大模型应用的**高性价比**落地？

RoycomKT智算一体机



RoycomKT智算
一体机

唯一**单机**推理
Kimi-K2万亿满血版模型
大幅度**降低部署**门槛，提
高运维效率
超高性价比，为满血版模
型**轻量化**落地成为可能

★ **KLLM**推理引擎
实现异构推理，**以存换算**

非国产配置 intel. + 

2 * CPU + 8 * GPU

信创配置 **HYGON** + 
iluvatar

2 * CPU + 8 * XPU

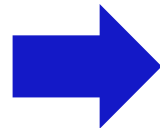
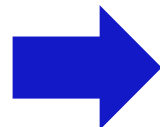
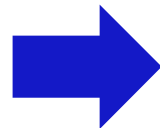
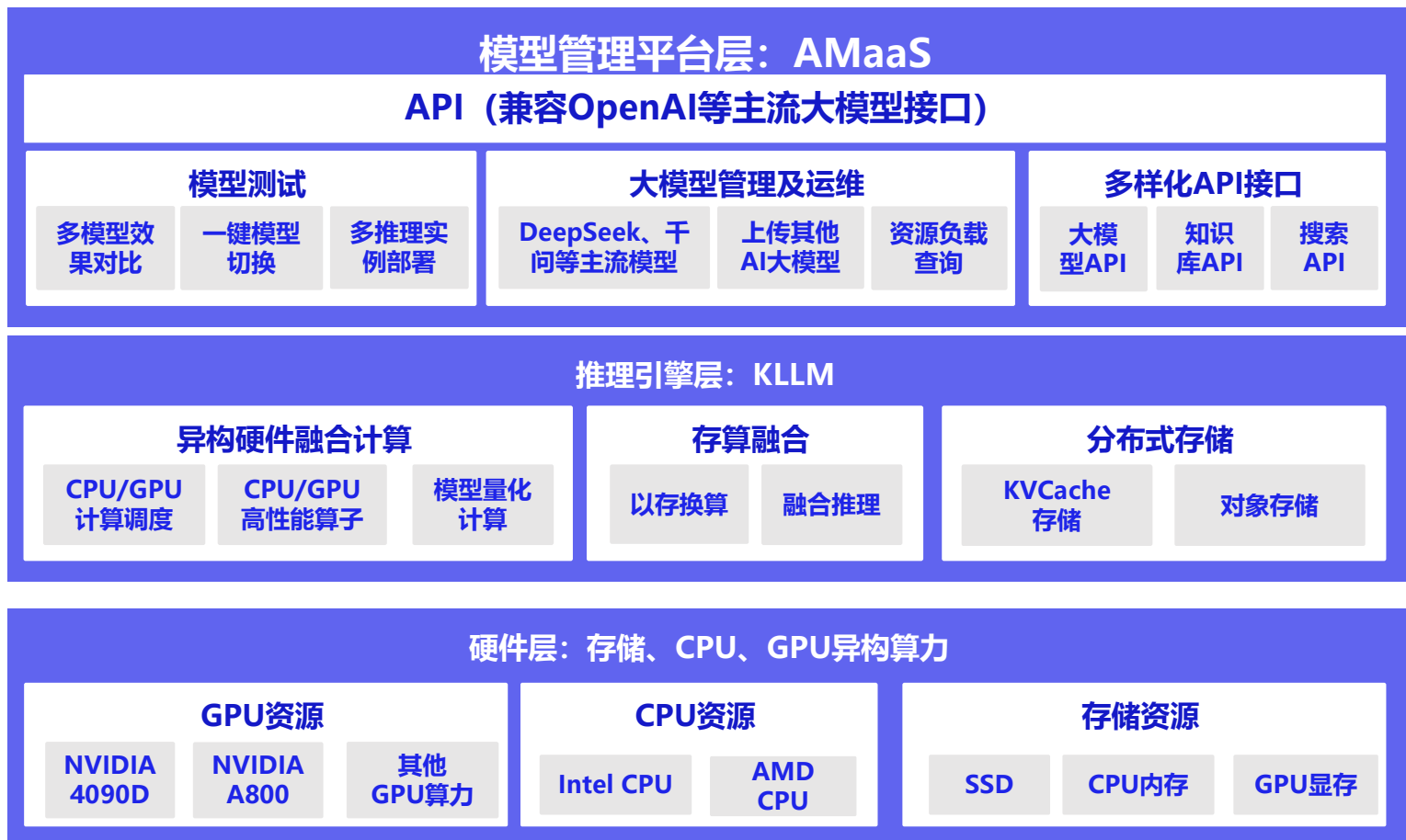


单机部署
满血模型

大模型一站式解决方案

RoycomKT智算一体机是针对**千亿级参数量**模型推出的本地化部署方案；采用**全系统异构推理架构**和**以存换算**模型推理加速引擎，减少算力开销，为企业客户提供超高性价比的模型部署方案，大幅降低大模型落地门槛，助力为AI大模型行业化私有化应用落地提供强大的支持！

框架图



产品亮点

1



原创推理引擎



门槛断崖式下降

2



硬核底层软硬一体技术创新



性能强劲

3



模型服务与应用部署便捷



开箱即用

4



线性延展，即需即建



延展容易

5



企业级服务为模型性能兜底



运维无忧

满
血
版
大
模
型



deepseek

671B大模型 (R1)

deepseek-R2大模型(即将发布)



KIMI

KIMI-K2大模型(1000+B)

40% 部署门槛

100% 性价比

单机 部署方案

万亿级大模型本地化

以存换算推理引擎

3简化，降低模型部署门槛

简化 大模型算力底座

- 单台可用；
- 计算、存储、网络统一调度和管理；
- 内置容器集群，安全稳定可靠；

简化 大模型推理运行环境

- 多模态大模型支持；
- 模型推理、模型管理和运维监控，并提供MaaS服务；
- 推理的加速库、框架稳定性支持；

简化 AI部署运维难度

- 提前预装平台模型，实现开箱即用；
- 提高可视化界面进行统一管理，减少操作的复杂性
- 部署、测试、调用、运维一站式交付；

3强化，助力模型价值变现

强化 算力资源利用率

- 以最少资源调用满血模型，提高性价比；
- 全新推理架构，助力模型推理应用部署实现
- 以存换算方式，实现单机算力的充分利用

强化 数据安全能力

- 私有环境部署，避免上云的数据隐患
- 数据安全保障，本地数据方式调用

强化 纵向系统工程交付能力

- 硬件、MaaS、数据、AI、推理，全栈能力集和；
- 建设部署服务，数据安全服务，咨询培训服务；
- 多款大模型并发/切换服务

PART 02

RoycomKT智算一体机



RoycomKT智算一体机

图形化界面

- 机器硬件资源、内置大模型、各推理引擎统一纳管，资源使用情况**图形化展现**
- 模型可一键启动和停止运行；根据模型大小，自动化资源调度和分配
- 内置管理所有正在运行的模型的**图形化仪表盘**，方便管理员掌握模型状态



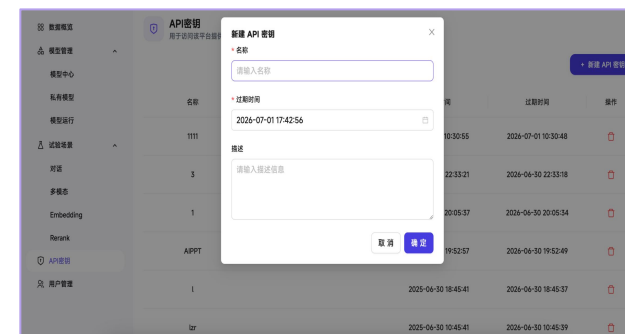
模型管理

- 内置数十个常用大模型，包含大语言、多模态、rerank、embedding等，同时支持私有模型的上传
- 支持**多个模型并行运行调参**和效果对比，以及对不同模型测试效果的全面评估
- 实现**多模型混合推理能力**，快速调用多款模型



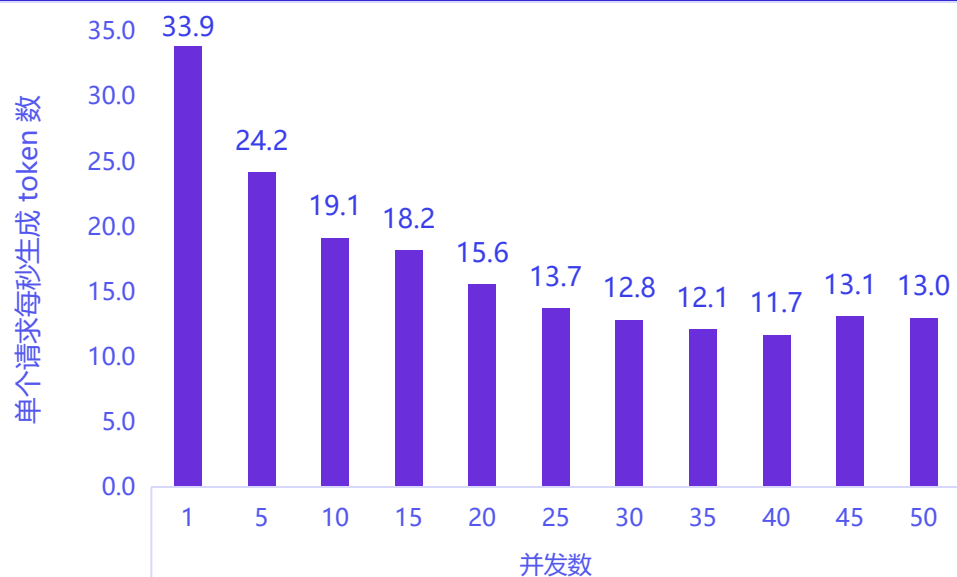
运维管理

- 针对长上下文、文生图及文生视频等场景，进行特定性能优化，打造高可用能力
- 实现用户注册管理，模式实例监管控制，模型调用使用情况
- 内置**API访问接口**，管理员可通过API密钥管理用户使用情况

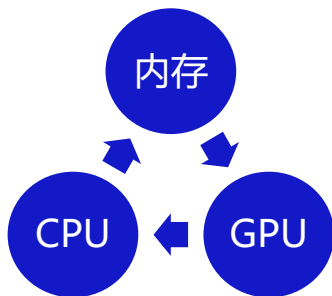
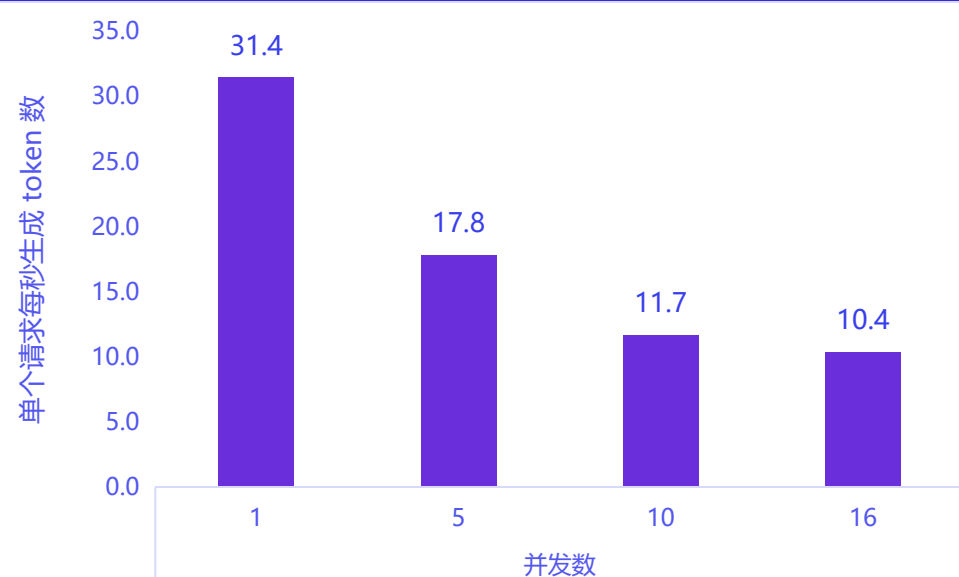


RoycomKT智算一体机

短输入测试 (input=128 tokens)



长输入测试 (input=4096 tokens)



采用**异构协同大模型推理架构**，充分利用CPU、GPU、内存等资源，提供**中并发，高性价比**的满血大模型落地方案

模型	输入	输出	并发	首token (s)	吞吐 (output token)
Deepseek-R1(671B)	128	512	1	0.4	34
			30	2.1	391
			50	15	367
	1024	512	1	0.3	34
			30	11	286
	2K	512	1	1	31
			20	14	194

测试机器为：单机8卡4090（48GB）

结论展示：8卡4090情况下运行671B模型，采用全新KLLM推理引擎运行，并发可达50，提升推理吞吐量

RoycomKT智算一体机

➤ 大语言模型

Deepseek-R1/V3: 671B、685B
Deepseek-R1-Distill: 70B、32B、14B、8B、7B
Qwen2.5: 72B、32B
QwQ: 32B
Qwen3: 235B-A22B、30B-A3B
此外还有: Llama、GLM、KIMI等等

➤ 多模态模型

Qwen-VL、Kimi-VL、Deepseek-VL等等

➤ Embedding&Rerank模型

bge-base/small/large、m3e-base/small/large等
Bge-ranker-base/large等

ROYCOM!
— 融 科 联 创 —

无需担心多模型同时运行
无需担心模型频繁升级

➤ 多模型混推能力

支持多个模型推理任务并发执行, 互不干扰

➤ KLLM引擎

更强的性能
更好的稳定性
更多的硬件适配
更快的模型升级
... ..

➤ 实现架构

GPU + CPU异构推理
以存换算推理, 实现以极少显存运行大参数模型

非国产方案配置（固定）

2 * Intel Xeon Platinum 6530
8 * NV RTX 4090 48G (384G)
16 * 64GB DDR5 4800 RECC (1TB)
1 * 系统盘：480G SSD SATA
1 * 数据盘：3,84T U.2/M.2 NVME
2 * 网卡：千兆电口网卡（板载）

约400G显存

基础模型精度：FP8

- 一键部署DeepSeek满血版模型，开机即用
- 全新kLLM推理引擎实现异构推理，以存换算；最少400G显存运行671模型，满足50并发
- 单个消费级GPU+CPU异构推理千亿大模型
- 实现软硬协同调优，保障性能要求
- 模型升级灵活适配，多模型设配支持

国产方案配置（固定）

CPU: 海光7470 * 2

GPU: 天数天垓150 * 8 (或天数智凯100duo卡)

内存: 24 * 64GB DDR5 4800 RECC

1 * 系统盘: 480G SSD SATA

1 * 数据盘: 3,84T U.2/M.2 NVME

2 * 网卡: 千兆电口网卡 (板载)

基础模型精度: INT8

- 打造国产满血模型部署性价比最优比
- 一键部署DeepSeek满血版模型, 开机即用
- 全新kLLM推理引擎实现异构推理, 以存换算;
- 单个消费级GPU+CPU异构推理千亿大模型
- 实现软硬协同调优, 保障性能要求
- 模型升级灵活适配, 多模型设配支持

PART 3

案例介绍



某中型律师事务所

客户背景

- 专注民商事诉讼、企业合规等领域，每年服务本地中小企业及个人客户超百余家
- 年处理案件量较多，涵盖合同纠纷、劳动仲裁、知识产权等类型，需高频处理法律文书撰写、类案检索及证据分析等工作
- 传统办公依赖人工协作，缺乏智能化工具支持

客户痛点

- **取证时间长**：普通案件平均1-2小时，占据或超办案时间30%
- **案件转化率低**：近80%独立律师年均案源转化率不足15%
- **关键事实证据影响成案**：关键事实证据作用明显，经验不足影响客户信任
- **数据安全与智能化瓶颈**：律所数据涵委托人相关信息，非常敏感，无法使用公有云大模型能力

解决方案

- **本地化智能硬件部署**：通过整合法律智能体并部署律所一体机，搭载671B满血大模型，实现私有化本地运行，无需依赖云端，既保障案件数据安全，又为律所提供本地化高速算力支持
- **智能法律服务平台开发**：开发“AI律师助手”，深度融入案件委托全流程，提供智能案情分析、法律条款检索、文书预生成等服务，助力律师高效处理案件委托事务，显著提升法律服务的精准度与质效，推动律所向智能化、专业化转型

某水利研究院所

客户背景

- 水利行业涉及防洪抗旱、水资源管理、水生态保护、智慧水务等多个专业领域
- 院所所中存在大量的数据文献需要进行整合，在进行处理时效率较低，耗时间又长
- 传统办公依赖人工协作，缺乏智能化工具支持

客户需求

核心需求是构建一个智能科研应用助手，具备以下能力：

- ① **智能问答**：支持水务专业领域知识问答、水务行业政策解读、现有公共知识库实时查询等功能。
- ② **严格依据参考资料回答**：作为智能科研助手，具备写文案、读文献、代码编程、答疑解惑的能力，在涉及到专业领域知识时要求数据能够严格按照参考资料进行回答。
- ③ **文档解析与校验**：支持研究人员上传技术报告、项目案例等附件(支持 PDFNord/Excel 格式)，

解决方案

- **本地化智能硬件部署**：提供搭载满血版671B大模型的私有化部署方案，深度适配检测数据敏感特性，无需接入公有云即可实现报告智能审核，精准识别数据异常、逻辑漏洞等问题，助力机构突破技术限制，提升审核效率，同时保障数据安全合规。
- **智能科研应用助手**：构建本地知识库，从本地数据文档中获取相关政策报告，行业标准、研究报告等进行整理，深度适配检测数据敏感特性，无需接入公有云即可实现报告智能审核，本地知识库问答

融科联创（天津）信息技术有限公司
Roycom (Tianjin) Information Technology Co., Ltd.

欢迎

感谢聆听！



融科联创（天津）信息技术有限公司
公司官网：www.roycom.com.cn
公司邮箱：sales@roycom.com.cn
服务热线：400-018-8995
天津-北京-深圳-成都-西安-杭州-武汉-南京