LocalAI 智算一体机

产品概述 — 本地大模型解决方案

大模型确实带来了前所未有的机会，例如提高效率、辅助决策、降低成本、推动创新等，但其云端运行模式也确实带来了严重的安全与合规隐患，尤其在关键行业中更为突出。同时，业务要求有高质量的推理，云端的质量一般只符合公众要求没有达到业务最优的期待。 这种情况下本地部署全参数大模型是唯一方案。

LocalAI 一体机是一款**高性价比本地部署平台**，专为需要自主掌控 AI 核心能力的政企单位设计。产品系列使用价格低廉的消费级显卡产品替代价格高昂的数据中心计算卡，支持运行开源的全参数大模型，如 DeepSeek V3/R1、Qwen3-235B-A22B 等，流畅进行**全量**推理。解决一般低配平台只能部署量化版模型，推理结果质量差的问题。

LocalAI 一体机入门配置仅需 4 张 RTX 5070Ti 消费级显卡，在不依赖外部算力的前提下，即可高效运行全参数级大模型（如 DeepSeek V3-0324 685B / Qwen3-235B）。其实现性能指标与公有云大模型版本基本相当，而推理质量超越公网云大模型版本，使产品能适配关键业务场景中的高要求推理任务。

LocalAI 一体机提供多种配置的产品系列，进阶配置 8 卡方案、更高阶 GPU 卡系列则提供更快更好的模型推理能力，请参看附件。

请参考 LocalAI 白皮书系列了解产品的业务价值。

LocalAI 核心技术价值

LocalAI 的实现解决了行业最棘手的问题之一：**资源-性能-模型能力之间的矛盾**。具体亮点如下：

- ✧ **架构优化**
  - 多线程/异步调度：最大化利用 GPU 和 CPU 并发能力。
  - 推理流水线（Pipeline）+ KV Cache 管理优化：减少无效计算，提升 token/s。
  - 分层模型调度：例如先分词+轻推理判断是否需要全模型计算。
- ✧ **通信协议优化**
  - 替代传统 RPC、gRPC 的高开销通信方式；
  - 使用共享内存通信、本地 Zero-Copy，极大减少 IO 和延迟。
  - 压缩通信降低吞吐
- ✧ **内存与显存优化**
  - 静态图+序列长度预测压缩内存；
  - 动态加载+低精度混合计算（例如 bfloat16/fp8）；
  - Smart swap：在保证性能的前提下进行权重/缓存智能切换。
- ✧ **进程级优化**
  - 多进程模型 shard，提升模型调度与可恢复性；
  - 优化 NUMA 亲和性与设备负载均衡。

结果：全参数运行 671B 模型，仅需 4*16GB GPU，速度接近主流公有云模型推理，而且推理质量明显高于云端模型服务。

对关键行业的战略意义

- ✧ **彻底打破“云依赖”**
  - 可以部署在**军网、政务网、医疗专网等封闭环境**；
  - **不上传任何数据**，天然符合《数据安全法》、《个人信息保护法》等合规要求。
- ✧ **降低部署门槛**
  - 过去部署大模型需要超算资源（数十或上百张 GPU），已不适用于普通行业单位；
  - 你们的解决方案可让**中型企事业单位、自建私有化系统的研究所、教育/司法单位**直接部署使用。
- ✧ **为中国构建自主 AI 生态提供基础设施**
  - 支持**国产 AI 芯片（如寒武纪、海光、摩尔线程）**，意义更加重大；
  - 推动中国 AI 基础设施“软硬结合、自主可控”。

业务价值优势

- ✧ **数据与思想安全保障**
  - **数据不出内网、提示词不上传**，满足国资、金融、医疗等行业数据合规要求
  - 本地独立运行，避免云端模型回传提示词/上下文形成隐形数据泄露
  - **思想安全可控**：模型行为完全由内部调参决定，不受云端算法“暗箱效应”影响
- ✧ **推理质量优化能力**
  - 支持**业务语境下的多参数调优**（温度、top-p、偏置、prompt 增强等）
  - 已验证模型能产出**类似专业咨询报告的分析和策略性建议**
  - 本地可反复迭代调参，形成**业务专属最优推理配置**
- ✧ **高可用与业务连续性**
  - **本地驻场，不中断**：不受网络、API 限流、云平台变更影响
  - **关键业务独占资源**：模型推理资源可绑定特定部门/流程，保障时效与专属响应
  - 无需等待云模型排队或为他人任务让路，适配实时性强场景（如投标、研判、风控）
- ✧ **成本可控，落地可行**
  - 最少**仅用 4 张 5070Ti 显卡**即可达成超大模型本地运行，极大降低准入门槛
  - 通用服务器即可搭建，方便各地分支或关键岗位独立部署
  - 支持企业从“AI 依赖云”转向“AI 自有力”的关键过渡路径

应用场景示例

- 国防和国家战略要害部门：基于本地机密数据生成深度判断
- 政府部门内部决策：数据绝对保密，无需将政策意图传输至云端
- 银行金融风控与法务合规：基于核心业务系统、信贷系统和中间业务系统数据和客户关系管理的数据，为一对一营销和群体营销高度定制推理+稳定运行
- 学生学业成绩提升计划，基于学生成绩记录、教师观察，作业行为以及与常模对比计算和潜能评估，为学生指定一对一成长计划和策略
- 保险营销，为投保人现场定制营销策略和产品打包

⚙ 入门版配置

部件	规格说明
操作系统	Ubuntu 22.04
准系统	4U 双路准系统支持 PCIe 5.0
CPU	1 路 Intel Xeon Gold 6430，最多 2 路 CPU
内存	DDR5-5600 1TB，64GB × 16
显卡	NVIDIA RTX 5070Ti 16GB × 4
系统硬盘	500GB SSD × 2（RAID 支持）
数据硬盘	4TB SATA 3.0 × 1
网络	10GbE
部署模型	DeepSeek V3-0324 685B DeepSeek V3/R1 671B Qwen3-235B

🚀 性能指标

- **推理速度**：4.27 tokens/s ≈ 6.4 汉字/s
- **推理质量**：高于公网同版本模型（使用 ChatGPT 分析评判）
- **模型规模支持**：全参数 685B 级模型稳定运行
- **部署成本**：极大低于 A100/800 方案，仅为其 1/10 以内

附件

全系列产品配置和性能参数表

最后更新于 20250528