AI 员工算力与部署白皮书：从模型参数到任务级服务目标

以真实并发、长上下文、检索与工具执行选择云、边、端和数据中心组合

在线技术白皮书

6 分钟阅读

发布日期 / 更新日期
2026-11-22

适合读者

基础设施负责人、架构师、运维与安全团队、AI 项目采购人员

核心结论 / DIRECT ANSWER

AI 硬件选型应从任务级服务目标开始：在什么模型、上下文、并发和工具链下，首 token 等待、生成速度、成功率与单任务成本是否达到业务要求。加速器峰值算力或显存容量只是约束条件，不能单独证明 AI 员工的实际体验、稳定性或项目成效。

01 / KEY QUESTIONS

本白皮书回答的关键问题

Q1

为什么不能只按 GPU 峰值算力采购?

Agent 工作负载还受模型、量化、上下文、显存带宽、批处理、检索、网络、工具执行和调度影响，必须端到端测量。

Q2

本地部署是否天然更安全、更便宜?

不是。安全性取决于完整架构和运营，成本还包括采购、供电、散热、运维、升级与闲置；需要按数据、风险和利用率判断。

Q3

PoC 最重要的输出是什么?

在固定版本和数据条件下得到可复现的任务级延迟、质量、稳定性、资源消耗与成本记录，而不是一次演示。

02 / 正文

01

先描述工作负载，再讨论设备

AI 员工的负载不是单次模型推理，而是一条包含检索、规划、工具与复核的任务链。

知识问答、长文研究生成、代码任务和多角色协同对计算的要求不同。输入长度影响预填充，输出长度影响解码，工具调用引入等待与外部依赖，知识检索增加存储和向量计算，多角色流程则可能产生重复上下文和并发峰值。选型前应记录每类任务的典型与高负载形态。

中国信通院人工智能产业研究资料持续关注模型、算力、数据与应用协同。对企业采购而言，可落地的含义是建立“模型—系统—任务”联合视图：明确模型版本与精度、服务框架、知识库、VROS Agent 运行时、Worker、网络和存储，而不是把硬件视为孤立盒子。

负载变量	需要采集的条件	影响
模型与精度	参数规模、量化、上下文上限	显存、计算与质量
请求形态	输入／输出长度、并发、峰值	排队、吞吐与尾延迟
Agent 流程	轮次、检索、工具与重试	总时长与单任务成本
服务目标	交互等待、批处理窗口、可用性	部署拓扑与冗余

本章来源 [1] [2] [4]

02

云、边、端与数据中心是组合，不是阵营

部署位置应由数据边界、时延、弹性、运维能力和业务连续性共同决定。

桌面或工作站适合开发验证和小规模交互，办公或边缘节点可承接局部知识与低时延任务，数据中心适合集中服务和统一资源治理，云端则提供弹性与托管能力。实际项目可以采用混合路径，但跨环境传输的数据、模型、日志和身份必须一致纳管。

AIBotStore 公开方案关注 VROS Agent 与 FusionXpark、TokenBox、FusionOne AI、FusionServer 等多形态算力的融合适配。这些名称代表候选部署形态，不构成对任意模型、规模或性能的默认承诺；具体设备、兼容矩阵和可用能力须按项目验证。

- ✓ 数据敏感且边界稳定：优先评估本地或专有云路径。
- ✓ 需求波动明显：评估弹性资源与峰值溢出策略。
- ✓ 边缘场景：额外检查断网续作、远程运维与物理安全。
- ✓ 混合部署：明确模型、知识、事件和日志在各环境间的流向。

本章来源 [1] [5] [3]

03

模型服务层决定硬件能否转化为体验

相同设备在不同运行时、批处理与量化策略下，可能呈现完全不同的质量与吞吐。

服务栈通常包括模型权重与精度、推理引擎、显存管理、连续批处理、KV Cache、请求队列、路由、可观测性和模型更新机制。上下文缓存、 speculative decoding 等技术是否适用，需要按模型与服务质量验证，不能作为无条件收益写进方案。

VROS Agent 与模型服务之间应建立清晰接口和超时、重试、熔断策略。检索与 Worker 沙箱也需要独立资源预算，避免工具任务耗尽模型服务资源。升级推理引擎或量化版本时，除速度外还要复测答案质量、工具调用格式与长上下文稳定性。

- ✓ 固定模型、运行时、驱动与依赖版本，形成可复现测试清单。
- ✓ 将交互请求与离线批任务分开排队，避免相互挤占。
- ✓ 为模型、检索、Worker 和存储分别设置资源与健康指标。
- ✓ 所有优化都要同时评估质量、延迟、吞吐和失败行为。

本章来源 [4] [6] [2]

04

用 TTFT、TPOT、吞吐与尾延迟描述真实服务

单看 tokens/s 会隐藏首字节等待、排队和最慢请求，指标必须附带测试条件。

TTFT 是从请求到首个输出 token 的等待，TPOT 表示后续输出 token 的间隔，吞吐描述单位时间处理的 token 或请求，尾延迟则观察高分位请求的等待。对 Agent 任务，还应测量端到端任务时长、工具等待、重试和最终交付合格率。不同指标不可脱离输入输出长度和并发进行比较。

MLPerf Inference 提供数据中心等场景的标准化推理 Benchmark 方法，可用于理解性能测试的场景与指标设计。但第三方设备成绩不能直接转化为 AIBotStore 的性能承诺；项目验收应在指定软硬件版本、模型、数据和任务集上独立完成。

指标	回答的问题	报告时必须附带
TTFT	用户多久看到首个输出?	输入长度、并发、排队与分位数
TPOT	持续生成是否流畅?	输出长度、解码与批处理条件
吞吐	系统总处理能力多大?	质量约束、请求结构与统计窗口
尾延迟	最慢一批请求体验如何?	P95／P99、峰值与超时策略
任务级结果	完整工作是否按时、合格完成?	检索、工具、重试与人工复核

本章来源 [4] [7]

05

本地推理仍需完整安全与运维体系

设备放在机房内并不自动消除模型、接口、供应链和人员风险。

需要保护模型权重、知识索引、事件日志与缓存，限制管理面和推理面访问，并对外部文档与工具返回值进行不可信输入处理。Worker 沙箱、网络出口、密钥管理、镜像与依赖审查、备份恢复和日志保留应纳入同一安全设计。

运维侧还应检查温度、功耗、显存、磁盘、队列、模型加载、节点漂移与服务降级。NIST AI RMF 与部署后监测方法可用于组织风险和运行问题，但不等同于具体设备认证或合规证明。高风险动作继续保留人工确认和可停止路径。

- ✓ 身份与密钥不写入模型上下文、镜像或公开日志。
- ✓ 管理接口与用户推理接口分权，远程运维有审批与审计。
- ✓ 模型、插件、驱动和系统更新先在验证环境复测。
- ✓ 故障时明确切换、降级、数据恢复和人工接管流程。

本章来源 [3] [8] [5] [7]

06

以 TCO 和 PoC 收口采购

真正的成本单位应是合格任务，而不是设备标价或理论 token 单价。

TCO 应包含设备或云资源、网络与存储、供配电与散热、软件适配、模型与知识维护、监控、安全、备件、升级和人员。收益侧则按实际工作量、质量与人工基线评估，不把目标人效或未经验证的节省写成事实。

PoC 需冻结测试条件，覆盖典型与峰值任务、长稳运行、异常恢复和变更复测。输出至少包括兼容矩阵、任务质量、延迟分布、资源曲线、失败样本、成本模型与运维责任。未达到阈值的场景应缩小范围、调整架构或停止，而不是用平均数掩盖。

- ✓ 先定义服务目标与验收数据集，再选择候选配置。
- ✓ 在相同模型与任务条件下比较不同部署方案。
- ✓ 报告均值与高分位，同时保留失败和人工返工。
- ✓ 将扩容、升级、退出和资产移交写入采购条款。

本章来源 [4] [5] [7] [1]

---

03 / CHECKLIST

项目启动前检查清单

可将以下项目转化为 PoC 计划、采购问题或验收清单。

01 建立模型、上下文、并发、检索、工具和重试的负载画像

02 按数据、时延、弹性与运维能力选择云边端组合

03 冻结模型服务栈版本并验证质量与性能的共同变化

04 同时报告 TTFT、TPOT、吞吐、尾延迟和任务级合格率

05 将权重、知识、缓存、日志、Worker 与管理接口纳入安全设计

06 以合格任务成本完成 TCO，并在真实条件下执行长稳 PoC

04 / FAQ

常见问题与直接回答

01 **TTFT 和 TPOT 有什么区别?**

TTFT 衡量首个输出 token 的等待，TPOT 衡量后续 token 的生成间隔。前者影响“响应是否及时”，后者影响“阅读是否流畅”，两者都需附带并发与文本长度条件。

02 **第三方 MLPerf 成绩可以作为项目验收吗?**

不能直接替代。MLPerf 可帮助理解标准化推理测试，但项目仍需在自己的模型、数据、版本、任务和服务目标下测量。

03 **TokenBox 或其他硬件是否保证固定性能？**

不保证。适配与性能取决于具体设备、模型、精度、服务栈和任务条件，必须以兼容测试和项目验收结果为准。

04 **本地部署一定优于云端吗？**

没有绝对答案。本地、云和边缘各有数据、弹性、时延与运维权衡，企业应通过任务级服务目标和 TCO 比较。

05 / REFERENCES

参考文献与公开来源

中文按 GB/T 7714 风格编排，英文按 APA 风格编排。访问日期：2026-11-22。外部资料用于行业与技术方法参考，不代表 AIBotStore 获得其背书或认证。

[1] 中国信息通信研究院. 人工智能产业发展研究报告（2026年）[R/OL]. (2026-02)[2026-11-22].
https://www.caict.ac.cn/kxyj/qwfb/bps/202602/P020260202487301304903.pdf →

[2] 中国信息通信研究院人工智能研究所. 智能原生（AI Native）研究报告（1.0）：企业智能原生转型框架与实践[R/OL]. (2026-10)[2026-11-22].
https://www.caict.ac.cn/kxyj/qwfb/ztbg/202610/P020261022558735347833.pdf →

[3] 中国信息通信研究院. 人工智能安全治理蓝皮书（2025年）[R/OL]. (2026-01)[2026-11-22].
https://www.caict.ac.cn/kxyj/qwfb/bps/202601/P020260109784447548497.pdf →

[4] MLCommons. MLPerf Inference：Datacenter 场景与指标[EB/OL]. [2026-11-22].
https://mlcommons.org/benchmarks/inference-datacenter/ →

[5] National Institute of Standards and Technology. AI Risk Management Framework（AI RMF 1.0）[EB/OL]. (2023-01-26)[2026-11-22].
https://www.nist.gov/itl/ai-risk-management-framework →

[6] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile（NIST AI 600-1）[R/OL]. (2024-07-26)[2026-11-22].
https://doi.org/10.6028/NIST.AI.600-1 →

[7] Rao A, Keller A, Kalra N, 等. Challenges to the Monitoring of Deployed AI Systems（NIST AI 800-4）[R/OL]. (2026-03)[2026-11-22].
https://doi.org/10.6028/NIST.AI.800-4 →

[8] OWASP GenAI Security Project. Agentic AI—Threats and Mitigations[EB/OL]. (2025-04-28)[2026-11-22].
https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/ →

范围说明

本白皮书用于技术选型与项目规划参考，不构成法律、合规、安全或投资专业意见。VROS Agent、AIBotStore、硬件适配和受控自进化的实际范围以交付版本、项目合同与验收为准。行业任务示例不代表已交付客户或已实现成效。