abit

升技云计算 · 国产化服务器与算力解决方案提供商

abit
AI APPLIANCE

私有化部署 · 数据不出域 · 开箱即用

AI 本地大模型一体机方案介绍

GPU 算力 × 本地大模型 × 对话界面 × API 服务，打包成一台可交付设备 —— 开机配置网络即可用浏览器访问，并通过 API 对接业务系统。



一台设备，四件事

算力、模型、门户与接口在同一套设备内完成部署，客户拿到的是可直接使用的环境

Web 对话门户

标准 API 服务

本地大模型推理服务

GPU 服务器 • 存储 • 网络

四层由**同一个供应商**统一交付 —— 服务器、推理服务与门户由一份方案定义，出问题只有一条责任线。

01 统一软硬件交付

服务器、推理服务与门户由一个供应商负责，不拼硬件、不凑软件

02 浏览器即可访问

开机完成网络配置后，员工用浏览器登录就能对话，不装客户端

03 标准化 API

业务系统与自动化流程按统一接口调用模型能力，不必各自对接推理层

04 按需扩展业务能力

知识库、Agent 工作流、统一身份认证按业务节奏叠加，非一次性标配

一句话：交付一台能直接用的设备，不是一堆需要集成的零件。

企业应用 AI 面临的五道坎

客户真正在判断的，不是「能不能用大模型」，而是用得放不放心、接不接得上业务。

01 数据边界与合规要求

数据出网范围、存放位置与审计要求需逐项对齐；公有云弹性好，但数据边界由供应商定义

02 软硬件与模型集成

GPU 驱动、推理框架、模型版本与并发调优分散在不同团队，联调周期长、责任界面不清

03 知识与模型能力脱节

通用模型不了解企业制度、产品与技术资料，回答看似流畅却难以直接采用

04 业务接入与自动化落地

从「聊得好」到「嵌进业务」，还需要接口、权限、审批与运维配套，缺一环就停在试用阶段

05 可管理与可扩展的基础设施

资源分配、监控、备份与扩容缺少统一入口，规模一上来就回到人工救火

一体机 · 逐一回应

数据不出域 整机部署在客户机房，数据与访问策略由客户定义

出厂即集成 驱动、框架、模型版本出厂调通，开机即用

接自己的知识 知识库把企业制度与资料喂给本地模型（选配）

接口直达业务 标准 API + 权限审批，能力嵌进流程而非停在试用

统一运维入口 监控、日志、备份随设备交付，扩容按规格换配

这五道坎，在出厂前一并跨过。

五层能力架构

自上而下解耦 —— 默认交付层与可选项分开，不含糊

● 金色标注 = 可选 / 按项目配置，以项目确认为准

应用层	Web 对话门户	企业应用	Agent / 工作流 · 可选
服务层	API 接入	身份认证	访问控制 · 按项目配置
模型层	本地大模型推理	Embedding · 可选	Reranker · 可选
基础设施层	GPU 服务器	存储	网络	操作系统
运维层	监控	日志	备份	维护

默认交付：算力、模型、门户与接口 —— 开机即可对话与调用；**选配能力：**知识库、Agent、身份认证对接，按项目实施。

产品矩阵 · GPU 算力机型选型

全部为在售量产型号 —— 一体机按下面两款主力配置交付，其余型号可替换

型号	平台	处理器	内存上限	GPU 扩展	一体机定位
ABX3G2-404	X86	XEON 3代	1.5TB	4× 双宽 GPU	轻量推理 · 部门级
ABX5G2-408	X86	XEON 4/5代	8TB	8× 双宽 GPU	主流推理 · 训练入门
ABX5G2-410	X86	XEON 4/5代	8TB	10× 双宽 GPU	**X86 主力（本方案推荐）**
ABE4G2-410	X86	EPYC 4/5代	8TB	10× 双宽 GPU	EPYC 平台备选
ABH4G2-408	国产	2× 海光4号 74XX	3TB	8× 双宽 GPU	**国产主力（本方案推荐）**
ABK3G2-410	国产	2× 鲲鹏920	256GB DDR4	10× 双宽 GPU	鲲鹏路线备选

选型规则：通用业务与性能优先选 **ABX5G2-410**，信创与国产化要求选 **ABH4G2-408**；其余型号同流程可换配，交付与运维体系不变。

X86 平台：ABX5G2-410

通用 X86 生态 · 面向高密度 GPU 加速负载的 4U 机架式算力平台



4U
机架式

10×
双宽 GPU

8TB
DDR5 内存

5.0
PCIe 通道

处理器

第四/五代英特尔® 至强® 可扩展处理器

内存

32× DDR5 插槽 · RDIMM/LRDIMM ECC · 最大 8TB

GPU 扩展

PCIe 5.0 Switch 架构 · 11× PCIe 5.0 x16 · 标配 10 张双宽 GPU

存储

4× 2.5" SSD，选配 12× 3.5"/2.5" 热插拔

电源

2700W 冗余 · 3+1 / 2+2 模式 · 预留 1 槽可插 IB 网卡 / NVMe 扩展卡

典型场景

大模型在线推理 | Llama / 文心等中小模型微调 | 计算机视觉 | 语音识别 | HPC 仿真 | GPU 虚拟化 VGPU | 基因测序 / 金融量化

一体机推荐配置：10 卡版预装本地推理服务与对话门户，开箱承载 7B~70B 级模型。

国产平台：ABH4G2-408

海光 X86 技术路线 · 兼顾国产化适配能力与生态兼容性

处理器

2× 海光 4 号 7400 系列 · 单颗 64 物理核心

内存

24× DIMM · DDR5 4800MT/s · 最大 3TB

GPU 扩展

12× PCIe 扩展插槽 · 最高 8 张国产双宽 AI 加速卡

存储

12× 3.5" SAS/SATA 热插拔 + 4× NVMe U.2 + 2× M.2

电源 / 管理

2000W / 2700W CRPS · 1+1 冗余 · 板载 BMC（IPMI 2.0 / Redfish）



128	8×	3TB	1+1
核心（2×64）	双宽 GPU	DDR5 内存	电源冗余

典型场景：国产化 AI 训练 / 推理 · 信创环境适配 · 兼容通用 X86 软件栈 · 国产双宽 GPU 生态 · 行业信创替代

双技术路线价值：国产平台跑得动，通用软件不用改 —— 适配国产软件栈的同时兼容 X86 生态。

双平台对比与选型

同一套软件栈、同一套交付流程，两种算力底座按需选择

ABX5G2-410 X86 平台

算力底座

XEON 4/5代 · 10× 双宽 GPU

生态

通用 X86 软件栈，CUDA / 主流推理框架原生支持

内存

32× DDR5 · 最大 8TB

适用

性能与生态优先，AI 训练 / 推理 / HPC

ABH4G2-408 国产平台

算力底座

海光 4 号 74XX ×2 · 8× 国产双宽 GPU

生态

国产软件栈 + 通用 X86 兼容，信创评审友好

内存

24× DDR5 · 最大 3TB

适用

国产化与合规优先，政企 / 信创场景

分阶段选型建议

首期试点

单部门先跑起来：8 卡版即可承载对话 + 知识库

主力上线

多部门共用 / 接业务系统：10 卡版 X86 或国产版

规模扩展

并发与模型规模上来后按台扩容，软件栈不变

开箱即用 · 四步上线

全程不装客户端，不需要用户自行配置 GPU 驱动与推理框架

1. **开机并完成网络配置**
   按交付清单上电，配置管理地址与业务网段；DNS、代理按客户网络策略设置
2. **浏览器访问**
   在受控网段内打开浏览器，输入设备地址即可进入登录页，无需安装客户端
3. **登录**
   账号由管理员在本地创建，角色与可用功能按部门与岗位分配
4. **开始对话**
   进入对话界面选择模型与知识范围，会话记录保存在本地环境中



标准 API，连接企业业务

接口只做一件事：让现有系统用统一方式调用模型能力，不必各自对接推理部署。



1 统一调用入口

门户、业务系统、脚本与工作流共用一套模型调用方式

2 鉴权与配额

API 凭据、调用配额与访问日志按项目配置，可与客户账号体系对接

3 兼容范围以交付版本为准

接口形态、字段与兼容范围在项目实施前确认，并写入交付文档

注意：API 是能力入口，不等于 Agent

自动化仍需要工作流编排、工具调用与权限审批配套，才能安全地落到业务流程里（见后页）。

企业知识库，激活内部知识

可选能力 · 默认不随设备交付，按项目选配与实施

01　内部文档

制度文件、产品资料、项目文档与技术手册

02　处理与索引

解析、分块、向量化，为每段内容建立索引

03　语义检索

按问题召回相关片段，带上原文出处

04　生成回答

模型结合片段作答，并标注引用来源

制度与流程查询	员工手册、审批流程、合规要求
产品资料问答	型号参数、配置说明、报价口径
项目文档检索	方案、会议纪要、验收材料的查找
技术手册查询	部署手册、参数说明、故障处理条目

实际效果取决于

- 文档质量与更新频率
- 权限设计：不可见的内容不进入检索
- 模型与检索方案选型
- 文档规模与索引维护投入

回答带原文出处，便于核对 —— 这比「回答得流畅」更重要。

Agent 与自动化办公

可选能力 · 编排与业务系统集成按项目选配

01
用户提出需求
一句话描述目标，例如「汇总本周项目进展」

02
调用本地模型
由工作流拆解任务，模型在本地完成生成

03
检索或调用接口
按需要查资料，或读取业务系统中的数据

04
生成结果
输出草稿、摘要或待提交的表单内容

05
人工确认后执行
写操作由责任人确认，审批通过才落地

案例示意

生成项目周报并提交确认

汇总本周项目数据 → 生成周报草稿 → 项目负责人确认 → 由负责人发出

草稿与引用来源一并留档，便于回溯是谁在什么依据下确认的。

工作流是可选能力

Agent 编排与业务系统集成按项目选配，不属于设备默认交付范围

写操作必须有权限与审批

发邮件、改数据等动作，按客户要求设置角色权限与审批节点

能力顺序：先 API 打通业务，再按场景叠加知识库与 Agent —— 每一步都可单独验收。

典型应用场景

先在一个部门、一类场景跑通，再按同样方式扩展

企业办公

公文与邮件起草、会议记录整理成纪要、长文档摘要与格式改写 —— 高频文字工作先受益

企业知识管理

制度、产品资料与技术文档的统一问答入口，回答带原文出处，便于核对

03 科研与工程

文献整理与摘要、代码辅助、实验数据处理脚本编写

04 客户服务

基于授权资料辅助答疑，不引用未授权内容

05 研发与 IT 运维

脚本生成、日志分析与故障排查思路辅助

以上场景均为辅助性质：涉及对外发布、合同、财务与客户承诺的内容，须由相关责任人审核后使用。

安全与运维设计

以下均为可配置项，具体措施与服务水平以项目方案和合同为准

客户环境本地部署

设备在客户机房或私有环境内运行，数据流向由客户网络边界与策略决定

网络访问控制

按网段、地址与端口限定访问范围，可只开内网入口或经反向代理转发

身份与权限管理

本地账号或对接客户已有账号体系，角色决定可见知识与可用功能

API 身份认证

调用凭据按系统发放，可配置调用配额、有效期与吊销策略

日志与审计

登录、对话与接口调用记录本地留存，可按需导出用于审计

运行监控与备份

GPU 利用率、显存与服务状态可视；数据与配置的备份周期、恢复方式按服务约定执行

关于认证：本资料不声明任何尚未取得的认证或合规资质；如需特定行业合规要求（等保、行业规范等），在项目方案中逐项确认。

部署与交付流程

验证不通过则回到配置环节调整，不带着问题上线

01 需求调研
02 方案设计
03 设备与软件配置
04 安装部署
05 功能与性能验证
06 培训交付
07 运维支持

调研与设计

明确场景、用户规模与数据边界，输出配置与集成方案

实施与验证

到货上架、软件配置、按约定场景做功能与性能验证

交付与运维

培训移交、纳入日常运维，按服务约定提供支持

方案与配置清单

硬件配置、软件清单与网络接入方式

验证记录

功能验收项与实测结果留档

培训与运维文档

用户手册、管理员手册与维护约定

各阶段的具体交付范围、时间与服务内容，以合同和项目方案为准；上图描述的是建议流程，不代表固定工期。

ABIT CLOUD · AI 本地大模型一体机

让大模型跑在自己的机房里

从 GPU 算力到业务应用，构建本地可控的 AI 基础设施。

开机即用 · 数据不出域 · 标准 API 对接业务系统

升技云计算 ABIT CLOUD

国产化服务器与算力解决方案提供商 · **www.abitcloud.cn**

AI 本地大模型一体机 · 方案介绍 · 2026.11

评估测试建议

可先提供一个部门或一类场景，我们按场景给出配置与验证方案。

- 单部门试点 8 卡版起步
- 按场景确认配置与集成
- 先验证再扩容，软件栈不变