metakina · metakina-agent

元话 AI 私有化平台
产品说明

Powered by metakina-agent · Hermes 运行时 + 企业控制面 + 国产模型与硬件适配

文档版本：v2.0 · 更新日期：2026-06-20

> 面向高敏感企业的 AI FDE 私有化交付平台。在企业内网完成 Agent、知识库、工作流与系统集成的闭环，全程数据不出域，按交付清单验收。

技术代号：metakina-agent

品牌名称：元话 AI 私有化平台

文档版本：v2.0

更新日期：2026-06-20

适用对象：企业决策者、IT 负责人、安全合规团队、AI FDE 交付团队

目录

---

1. 产品概述

1.1 产品定义

metakina-agent 是 metakina 母品牌下的企业桌面 AI Agent 产品；在中国区以 **元话 AI 私有化平台** 对外交付。产品直接基于 `nousresearch/hermes-agent` 的 Desktop / Runtime 能力二开，在客户内网

完成 Agent、知识库、工作流与系统集成的闭环，支持国产大模型与国产/主流硬件，全程可审计、可验收、可运维。

一句话价值：把 Hermes Agent 的执行能力改造成面向中国企业的元话 AI 桌面工作台，把「企业知识 + 业务流程 + AI Agent 执行」装进客户自己的机房，数据不出域，按行业方案包快速落地，按 FDE 标准交付验收。

1.2 产品组成

metakina-agent 由四层构成：

层级	组件	职责
体验层	Hermes Desktop 企业二开、管理控制台、业务门户、人工复核台、企业 IM/API 入口	面向管理员与业务人员
控制面	metachina 企业控制面 + **Connector Hub**	组织权限、知识治理、流程编排、连接器中心、模型网关、审计合规
运行时适配层	**Hermes Agent Runtime / ACP / Workflow / MCP Adapter**	Hermes 负责 Agent 执行与桌面会话；Workflow 负责 RAG/流程；MCP 负责定制系统
基础设施	本地/国产大模型、向量库、文档存储、SSO/ERP/MES/QMS/工单	推理与国内业务系统底座

桌面核心：Hermes Desktop 企业二开。

执行核心：Hermes Agent Runtime 企业加固版（可配合 ACP / Workflow / MCP Adapter）。

企业核心：企业控制面 + Connector Hub + Industry Pack + FDE 交付体系。

1.3 解决的核心问题

企业痛点	metakina-agent 的回应
数据不能上公有云	全组件支持内网离线部署，默认零外网出站
通用 Agent 不够企业化	补齐组织权限、知识审批、人工复核、审计导出
不同行业需求差异大	Industry Pack 配置化适配，不重做平台
国产模型/国产硬件要求	模型网关 + Hardware Profile，适配 GLM、Qwen、昇腾等
项目难验收	AI FDE 交付清单、KPI 口径、压测基线内置
运维风险高	版本灰度、一键回滚、监控告警、备份恢复

1.4 品牌与仓库关系

层级	品牌/仓库	职责
中国区 GTM	**metachina / 元话 AI**	官网、方案页、线索转化、国内交付合同
技术母品牌	**metakina**	开源组织、全球技术品牌、RWA 远期能力预留
企业桌面 Agent	**metakina-agent**	Hermes Desktop 二开、控制面、Connector Hub、Industry Pack、离线部署包
Agent 运行时	**metakina-com/hermes-agent**	Hermes 企业加固版上游跟踪
RWA 产品（未来）	metakina-vault / metakina-rwa	链上发行与资产托管，复用 metakina-agent 链下 AI 与审计能力

命名原则：技术文档与开源仓库统一使用 `metakina-agent`；metachina.ai 官网聚焦企业 AI，不在首屏混宣 RWA 相关内容。

2. 产品架构

2.1 三层架构

2.2 架构设计原则

1. 1. **控制面与运行时解耦**：控制面管策略，Hermes 管执行；升级可独立进行。
2. 2. **策略下发，数据不出域**：配置、Prompt、Skills 由控制面下发；推理与文档留在客户环境。
3. 3. **行业差异配置化**：行业能力 = 基础平台 + Industry Pack。
4. 4. **FDE 可交付**：每个 Pack 自带交付清单、验收用例、KPI 口径。

2.3 控制面与 Hermes 对接

控制面通过 **Agent Control API** 管理 Hermes 实例，下发：

- agent_profile（人格、边界、可用工具集）
- skills_bundle（行业 Skills + 企业自定义 Skills）
- knowledge_bindings（知识库与检索策略）
- workflow_policies（转人工阈值、审批节点、敏感词策略）
- model_routing（场景 → 模型映射）
- gateway_channels（IM/门户渠道及白名单）
- audit_sink（审计日志回传端点）

Hermes 回传：session_events、execution_metrics、knowledge_hits、policy_violations。

---

3. 核心能力模块

3.1 企业控制面

模块	职责	关键能力
组织中心	多部门、多项目隔离	组织树、项目空间、数据域标签
身份与权限	企业级访问控制	SSO/LDAP、RBAC、ABAC、审批流
Agent 工作室	智能体生命周期	创建/版本/发布/回滚、灰度、人工接管规则
知识治理中心	企业 RAG 治理	知识源接入、分类、版本、审批发布、来源追溯
流程编排中心	业务工作流	意图路由、置信度分流、人工复核、工单联动
连接器中心	系统集成	CRM、工单、OA、邮件、IM、文档系统适配器
模型网关	多模型统一治理	路由、降级、配额、调用审计、国产模型适配
审计合规中心	取证级留痕	全链路日志、会话归档、导出、保留策略
运维监控	私有化运维	健康检查、告警、备份恢复、版本升级
方案包中心	行业适配入口	Pack 安装、参数化、场景开关、交付向导

3.2 Hermes 运行时（企业加固版）

能力	说明
Agent 执行	多轮对话、工具调用、子 Agent 并行
Skills / Memory	企业 Skills 审核发布、跨会话记忆
Gateway	企业 IM、门户、API 多渠道接入
MCP / 工具集	扩展业务能力与系统集成
Cron 调度	日报、巡检、备份等无人值守任务
安全加固	命令审批、工具白名单、容器隔离

3.3 与 Hermes 的分工

能力	Hermes 负责	metachina 控制面负责
Agent 对话与工具执行	是	策略下发
Skills / Memory / MCP	是	企业审核与发布
多渠道 Gateway	是	渠道权限与企业白名单
Cron / 子 Agent	是	任务审批与审计
组织 / RBAC	否	是
知识审批与版本治理	否	是
行业流程编排	部分（执行）	是（设计）
企业审计报表	事件产生	是（汇聚/导出）
离线安装与升级	运行时包	整体编排与清单

4. 产品版本 SKU

版本	包含内容	适合客户
metakina-agent Core	控制面 + Hermes 运行时 + 基础 Pack	有自建能力、先做平台底座的企业
metakina-agent Industry	Core + 指定行业 Pack + FDE 标准交付 + 验收清单	有明确业务场景、要快速上线的企业
metakina-agent HA	Industry + 高可用 + 灾备 + 7×24 运维选项	金融、法律、政务等高可用要求客户

5. 行业方案包

5.1 Industry Pack 结构

每个 Pack 包含 7 类资产：

资产类型	内容示例
场景目录	业务场景、用户故事、禁用场景
知识模板	文档分类树、元数据字段、审批规则
Agent 模板	角色定义、Prompt、工具白名单
Skills 包	行业专用 Skills
流程蓝图	识别→检索→生成→复核→归档
连接器清单	推荐对接系统
交付与验收	6 周交付节奏、KPI、测试用例、合规检查项

5.2 方案包一览

方案包	对应元话方案	私有化推荐度	核心价值
pack.knowledge-base	企业知识库智能体	★★★★	制度/FAQ/培训知识统一检索，来源可追溯
pack.customer-agent	智能客服	★★★★★	降重复咨询、置信度分流、人工协同
pack.asset-ai	金融文档智能体	★★★★★	尽调材料、合规文档、投资人 FAQ
pack.legal-agent	法律与尽调智能体	★★★★★	合同审查、风险建议、报告辅助
pack.ecommerce-agent	跨境电商运营智能体	★★★★	多语言客服、商品文案、评论分析
pack.sales-agent	销售助手（扩展）	★★★	CRM 联动、线索跟进、话术辅助
pack.manufacturing-ops	制造运营智能体	★★★★★	工艺/质量/设备/计划协同，ERP/MES/QMS 只读联动

5.3 Connector Hub 分级

级别	范围	示例
L1 标准连接器	产品内置	飞书、企微、钉钉、LDAP、文档、Webhook
L2 行业连接器	随 Pack 提供	CRM、客服/工单、用友/金蝶只读、MES/QMS/EAM
L3 定制连接器	FDE 项目交付	自研 ERP、SCADA/时序库、政务/行业专用系统

5.4 跨行业复用内核

无论行业，以下能力共用同一套实现：组织/权限/审计、知识接入管道、Agent 发布与版本管理、模型网关、人工复核工作台、运维备份与灾备。行业 Pack 仅覆盖 20%–40% 的差异层。

---

6. 国产大模型与硬件适配

6.1 模型分层策略

档位	代表模型	适用场景
L1 旗舰档	GLM-5 系列、DeepSeek-V3 级 MoE	复杂推理、长文档、多步 Agent
L2 生产档	GLM-4.7、Qwen2.5-72B	客服、知识库、流程编排主模型
L3 轻量档	GLM-4-9B、Qwen2.5-7B、Embedding/Rerank	意图分类、检索重排、快响

原则：GLM-5 系列适合旗舰推理引擎，但不是唯一引擎；按场景分级路由可控制成本与硬件门槛。

6.2 模型角色分工

角色	推荐国产模型
主推理	GLM-5 系列、GLM-4.7、Qwen2.5-72B
快速响应	GLM-4-9B、Qwen2.5-7B/14B
Embedding	BGE-M3、Qwen3-Embedding
Rerank	BGE-Reranker、Qwen-Rerank
代码/工具	GLM-4.7、DeepSeek-Coder

6.3 硬件 Profile

Profile	目标模型	典型硬件	适用客户
HP-Entry	L3（7B—14B）	1× RTX 4090 / 1× 910B	部门试点
HP-Standard	L2（32B—72B）	2—4× A800 / 910B	知识库、客服
HP-Enterprise	L2 集群 + L3	4—8× A800/H800	多部门生产
HP-Flagship	L1（GLM-5 级）	8× H200 或 8× 910B	金融/法律旗舰
HP-Ascend	L2 + 部分 L1 量化	Atlas 800I A2/A3	信创合规优先
HP-Edge	L3 only	CPU + 单卡低功耗	分支机构

6.4 GLM-5 系列硬件参考

NVIDIA 路线

精度	显存需求	推荐配置
FP8	~800GB+	8× H200（141GB）
INT4/W4A8	~400—500GB	8× H100（80GB）

配套：CPU ≥64 核，内存 ≥512GB，存储 ≥2TB NVMe。

昇腾路线（国产化优先）

- 硬件：Atlas 800I A3（8× 910B）或 A2 系列
- 框架：vLLM + vLLM-Ascend
- 量化：W8A8 / W4A8
- 软件栈：CANN 8.3+、torch-npu

6.5 模型网关路由示例

- 客服快响场景 → L3（Qwen-7B）
- 知识库长文档（>8000 tokens）→ L1（GLM-5 系列）
- 法律/金融需引用场景 → L1（GLM-5 系列）
- 默认生产 → L2（GLM-4.7）
- 超时/OOM → 降级至 L2；硬失败 → 人工接管

6.6 分行业模型 + 硬件推荐

行业 Pack	主模型	辅模型	硬件 Profile
知识库	GLM-4.7 / Qwen-72B	BGE + Rerank	HP-Standard
智能客服	Qwen-14B + GLM-4.7	意图 7B	HP-Standard
金融文档	GLM-5 系列	GLM-4.7 备用	HP-Flagship / HP-Ascend
法律尽调	GLM-5 系列	Rerank + 规则引擎	HP-Flagship
跨境电商	GLM-4.7 + 多语言 7B	Embedding	HP-Standard
政务/国企	Qwen-72B + 昇腾	国产 Embedding	HP-Ascend

7. 部署架构

7.1 标准离线部署

```
客户内网
├── 接入区：反向代理 / WAF / SSO
├── 应用区
│   ├── metakina-agent-console
│   ├── metakina-agent-api
│   ├── metakina-agent-model-gateway
│   ├── hermes-gateway
│   └── hermes-agent-workers
├── 数据区
│   ├── PostgreSQL（元数据、审计、流程）
│   ├── 向量库（Milvus / Qdrant / pgvector）
│   ├── MinIO（文档与附件）
│   └── Redis（队列、缓存）
└── 模型区
    ├── 推理池 A：GLM-4.7 / Qwen-72B（生产）
    ├── 推理池 B：GLM-5 系列（旗舰，可选）
    └── 推理池 C：7B 快响 + Embedding + Rerank
```

7.2 网络策略

- 默认零外网出站
- 模型、补丁、方案包通过离线介质/内网镜像仓更新
- 可选混合 AI：仅模型 API 走专线，数据面仍留本地

7.3 部署规格建议

规模	适用	最小配置
试点版	单部门 POC	8C32G × 2 + GPU 可选
标准版	500—2000 用户	16C64G × 3 + GPU × 1—2
企业版	多部门/高并发客服	控制面 HA + Worker 集群 + 独立向量库

7.4 高可用与灾备

- 控制面 API/DB 主备
- Hermes Worker 无状态，可水平扩展
- 每日增量备份 + 每周全量
- Agent 版本蓝绿发布，支持一键回滚

8. 适用场景

8.1 强烈推荐

（1）数据安全与合规优先

- 客户资料、合同、财报、尽调材料不能出内网
- 需要完整审计日志与合规导出
- 典型客户：银行、证券、保险、律所、会计师事务所、大型国企

（2）智能客服 / 售后（私有化版）

- 售后咨询量大，重复问题多
- 需要 AI 先答 + 低置信转人工 + 服务 KPI
- 场景：FAQ 自动应答、工单预分类、坐席辅助、服务指标看板

（3）企业知识库 / 内部问答

- 制度、流程、培训、项目文档分散
- 要求答案带来源、版本、权限边界
- 场景：内控制度问答、培训答疑、售前检索、项目交接

（4）金融文档 / 尽调辅助

- 尽调材料体量大、更新快
- AI 仅做辅助，关键结论人工确认
- 场景：文档检索摘要、资料核对、合规版本对比、投资人 FAQ

（5）法律 / 合同 / 尽调

- 合同审查重复性高，报告依赖手工汇总
- 场景：条款抽取、风险提示、法规问答、报告草稿、法务审核链

（6）跨境电商 / 运营辅助

- 商品咨询高峰、多语言成本高
- 场景：商品咨询回复、多语言模板、评论分析、运营日报

（7）研发 / IT / 运维内部赋能

- 内部文档难检索，需 Agent 做巡检与日报
- 场景：运维知识问答、SOP 辅助、定时巡检、集成脚本辅助

（8）信创 / 国产算力环境

- 必须使用国产芯片（如昇腾 910B）与国产大模型
- 推荐：HP–Ascend + vLLM–Ascend + 国产模型矩阵

8.2 可使用但需评估

场景	说明	建议
小规模试点（<50 人）	私有化成本偏高	HP–Entry 或评估混合 AI
强实时语音客服	需 ASR/TTS 低延迟链路	需额外语音模块
复杂 ERP 深度改写	平台不是 ERP 替代品	以辅助+集成为主
多分支机构广域部署	需 Edge + 中心同步	中心推理池 + 边缘轻量节点

9. 不适用场景

场景	原因
个人开发者尝鲜	私有化部署与运维成本高
纯营销官网	使用 metachina 静态站即可
公有云多租户 SaaS	应走公有 AI 产品线
无合规压力、追求最快上线	混合 AI 或公有 AI 更合适
期望零人工自动裁决（高敏金融/法律）	产品定位为 AI 辅助 + 人工确认

10. 与元话其他部署模式的关系

模式	定位	何时选择
公有 AI	快速验证、标准化模块	试点、非敏感数据、最快上线
混合 AI	控制面托管 + 数据面本地	有数据边界但不想全自建
metakina–agent	全量离线、数据主权、合规优先	高敏感、信创、金融法律、核心客服数据

选型路径：业务诊断 → 评估数据敏感度 → 低选公有 AI / 中选混合 AI / 高选 metakina–agent。

11. AI FDE 交付方法

11.1 六步交付流程

步骤	内容	私有化附加项
01 业务诊断	目标、边界、指标	数据分级、网络边界、合规约束
02 试点验证	核心场景 POC	模型命中率、权限校验、压测基线
03 系统集成	知识源、CRM、工单	SSO、内网文档源、专线对接
04 生产部署	环境上线	离线镜像导入、密钥托管、安全扫描
05 培训上线	角色培训	管理员/业务/审计分轨培训
06 持续运维	指标复盘	模型升级、知识巡检、KPI 复盘

11.2 验收标准示例

知识库 Pack

- 权限内检索命中率 ⩾ 85%
- 高频问题首答一致率提升 ⩾ 35%
- 100% 答复可追溯到来源版本
- 审计日志可按用户/时间/案件导出

智能客服 Pack

- 重复咨询自动化处理比例 ⩾ 45%
- 首次响应时间下降 ⩾ 35%
- 人工转接准确率提升 ⩾ 30%

11.3 性能压测基线

指标	HP-Standard（72B）	HP-Flagship（GLM-5）
首 Token 延迟 P95	⩽ 2s	⩽ 3s
持续吞吐	⩾ 30 tok/s	⩾ 20 tok/s
工具调用成功率	⩾ 95%	⩾ 93%
故障切换	30s 内降级	30s 内降级

12. 安全与合规

维度	措施
身份	SSO、MFA、会话超时、IP/设备白名单
权限	RBAC + 数据域 ABAC、最小权限
数据	静态加密、TLS、字段脱敏
模型	私有化推理、Prompt 防护、输出审查
工具	Hermes 命令审批 + 企业工具白名单
审计	不可篡改日志、保留策略、导出水印

合规交付物：数据流向图、权限矩阵、审计抽样报告、模型与知识版本清单、应急预案与备份演练记录。

13. 运维与版本治理

项目	策略
模型版本	不可变版本号登记，如 glm-5.2-ascend-w4a8-202603
灰度发布	新模型先 5% 流量，对比 KPI 后全量
回滚	保留上一版权重与配置快照
补丁	驱动/CANN/推理框架与模型分包升级
监控	GPU/NPU 利用率、显存、KV Cache、队列、超时率
合规	权重 SHA256 校验、推理日志留痕、禁止私自换模

14. 客户选型速查

如果你...	推荐
是金融/法律/国企，数据绝不能出域	metakina-agent Industry/HA
要做私有化智能客服	pack.customer-agent + HP-Standard
要先做内部知识库	pack.knowledge-base + HP-Standard
要尽调/合同 AI 辅助	pack.legal-agent / pack.asset-ai + HP-Flagship + GLM-5
要信创国产化	HP-Ascend + 国产模型矩阵
只有单部门、预算有限	HP-Entry 试点，或先评估混合 AI
只要官网获客	使用 metachina 官网，不上 metakina-agent

15. 推荐落地路径

```
1. 业务诊断 → 确认是否必须私有化
2. 选择 Industry Pack（建议先从知识库或客服开始）
3. 硬件勘测 → 选定 Hardware Profile
4. 部署 L3 + L2 生产模型池
5. 联调 Hermes + 控制面 + 企业系统
6. 灰度上线 → KPI 验收
7. （可选）增配 GLM-5 旗舰池，覆盖高难场景
```

默认推荐组合

- 合规型：昇腾 4—8 卡 + Qwen-72B + GLM-4.7 + 知识库/客服 Pack
- 旗舰型：8 卡昇腾/H200 + GLM-5 系列 + 金融/法律 Pack
- 性价比型：A800 2—4 卡 + GLM-4.7 + 知识库 Pack

16. 产品边界

metakina-agent 是什么

- 企业级私有化 AI FDE 平台
- Agent + 知识库 + 工作流 + 审计的一体化底座
- 可按行业 Pack 快速落地的交付型产品

metakina-agent 不是什么

- 不是公有云 SaaS
- 不是 ERP/CRM 替代品
- 不是零人工的自动裁决系统
- 不是个人 Agent 工具

---

17. 附录：命名规范与离线部署包

17.1 服务命名

组件	服务名
控制面 API	metakina-agent-api
管理控制台	metakina-agent-console
模型网关	metakina-agent-model-gateway
Hermes Worker	metakina-agent-runtime-worker

17.2 离线部署包结构

```
metakina-agent-offline-bundle/
├── manifests/          # 硬件与模型清单
├── images/             # 内网镜像
├── models/             # 离线权重
├── drivers/            # 驱动与 CANN
├── compose/            # 部署编排
└── acceptance/         # 压测与验收脚本
```

17.3 对外标准介绍

> 元话 AI 私有化平台（metakina-agent）是面向高敏感企业的 AI FDE 私有化交付平台。平台在企业内网部署 Agent 运行时、知识治理、流程编排与审计合规能力，支持国产大模型（如 GLM、Qwen）与国产算力（如昇腾）适配，通过行业方案包覆盖知识库、智能客服、金融文档、法律尽调等场景。全程数据不出域，按交付清单验收，适合对数据主权、审计留痕与信创合规有强要求的企业客户。

元话 AI · metakina-agent 产品说明 v2.0