WHITE PAPER

企业智能体落地指南

从四问法判断、选型、数据接入，到评测口径、灰度陪跑与上线巡查

版本	出版	页数	适用读者
v1.0	2026-10	约 6 页	业务牵头人 / 信息技术部门 / 项目与采购负责人

编制单位

写逸网络科技（上海）有限公司

电话咨询

公众号

摘要：智能体落地失败，多数不是模型问题。这份指南给出七个步骤的判断口径与检查表，每一步都有可交付物。

摘要

智能体项目失败，多数不是模型问题，而是五件事没提前定：**该不该做**、**用哪一种**、**数据从哪来**、**怎么算成功**、**上线之后谁维护**。本指南按七个步骤把这五件事拆成可执行的判断口径与检查表，每一步都给出可交付物。

一、先判断该不该做：四问法

我们通常在第一次沟通时就把这四个问题问完。答不上来的，说明场景还没到可以立项
答案是否定的，说明这件事不该用智能体做。

问题	判断要点	否定的结论
一、这件事的输入是什么形态?	文本、表格、扫描件、图片、录音？是否可以被解析成机器可读的字段	输入无法结构化（例如大量无标注的料），先做数据结构化，不要直接上
二、输出的对错谁来判?	有没有明确的判定标准；错误代价是“返工”还是“事故”	判定标准说不清的，先和业务方把口径写成文档，再谈技术方案
三、数据能不能出网?	内网 / 专网 / 互联网；是否可以调用外部模型服务	不能出网就必须按私有化与信创路径设计，模型选型与算力预算都要重算
四、出了错谁接得住?	每个环节有没有复核人；复核成本是否可接受	没有复核环节的场景，智能体只能做“建议”而不该做“执行”

可交付物：《场景与边界说明》——一页纸写清输入、输出、判定标准、部署环境与责任人。这份说明是后面所有工作的前提，也是验收时的对照基准。

电话咨询
公众号

二、选型：单智能体、多智能体与混合方案

选型不是“越复杂越先进”，而是“用刚好够用的复杂度做完这件事”。

方案	适用	成本与维护	建议
单智能体	单一数据源、步骤固定、单步判断	成本低、维护简单	先把这一类做透，它会成为后面所有角色的基础件
多智能体	跨角色、跨系统、步骤多且需中途判断	调用次数与排查路径都增加	只在单智能体结构性做不到时升级，并给出成本估算
规则 + 模型混合	有硬性规则的部分（校验、路由、限额）与需要理解的部分混合	最容易被低估，但最稳	能写成规则的不要交给模型；模型只负责理解与生

实践中，我们交付的项目几乎都是混合方案：规则负责“不许出错”，模型负责“读懂好”，人工负责“拍板”。

电话咨询
公众号

三、数据与系统接入

这一阶段的目标不是“打通所有系统”，而是**用最小的通道拿到够用的数据，并把前提死。**

- **通道按现有条件选**：数据库只读账号优先；没有只读账号走标准接口；接口也没有就通道定时同步。三条通道的实时性、开发量、稳定性不同，要写进方案。
- **权限对齐既有体系**：智能体的可见范围应当来自客户既有的角色与组织架构，而不是另建一套权限。敏感字段按规则脱敏后才进入模型上下文。
- **质量先摸底，再谈治理**：抽一批真实材料统计可解析率、字段缺失率、版本冲突数量。这三个数字决定了智能体效果的**上限。**
- **把前提条件写死**：需要数据责任人按期更新、需要某个接口开通、需要某张表补充字段——逐条写进《数据与集成清单》，并明确责任方。

可交付物：《数据与集成清单》。这份清单里的前提条件没有落实之前，智能体不进入开发阶段；这是我们避免“做完了跑不动”最有效的一条纪律。

四、评测集与验收口径

没有评测集的效果描述等于广告。评测集必须来自真实业务问题，并且用可判定的口径定义“对不对”。

三级判定

- **正确**：结论点全部命中，出处正确，无越权与编造。
- **可用**：主要结论正确，需要人工补细节，出处可用。
- **需返工**：结论方向对但关键要素缺失或出处有误，必须人工重做。

每一级都要附判定说明，并请业务人员复判一次。判定标准不一致，比评测集太小更致命。

六项指标

1. **正确率 / 可用率**：分开统计，不要把“能凑合用”和“必须重做”混在一起。
2. **有出处率**：结论句中带有效出处的比例。
3. **错引率**：引用与结论不匹配的比例，比漏答危险得多。
4. **拒答正确率**：材料中确实没有答案时，系统是否拒答而不是编。
5. **人工返工率**：上线后实际被打回的比例，最诚实的指标。
6. **耗时与成本分布**：看分位数而不是均值。

可交付物：《评测集与基线报告》。从 50 条真实问题起步即可，关键是从此每次改动回归，并明确“不得低于当前基线”。

电话咨询
公众号

五、灰度、陪跑与知识固化

- **灰度**：先在真实但少量的任务上跑，人工双复核，逐条记录错例。错例清单比成功率更有价值——它直接告诉你下一版要改什么。
- **FDE 陪跑**：上线后由交付工程师跟着真实业务走一段时间，把体例与口径的偏差一条条改回来。这一步无法用文档替代。
- **知识固化**：把陪跑期间反复出现的口径问题写进配置或知识库，而不是留在个人经验里。判断标准很简单——换个新同事接手，能不能不改代码就跑对。

可交付物：《灰度报告与错例清单》《口径变更记录》。

六、上线后的运维与巡检

关注点	具体动作	频率
效果回归	按评测集跑一遍，与基线对比，定位退化条目	每次提示词 / 模型 / 工具 / 知识库结构性变更，以及每月一次
稳定性	看成功率、时延分位数、工具超时、降级触发与转人工率	实时监控 + 每周复盘
数据	知识库与源目录对账：多出来的删、缺的补；失效版本清理	每日自动 + 每月人工抽检
成本	token 消耗与单任务成本，按角色与任务类型拆开看	每月
审计	抽查留痕完整性，演示“从结论倒推到原始材料”	每季度

值班时先看三张图：任务成功率（是整体掉还是某类任务掉）、工具失败与超时排行、转人工触发记录。三张图定位不了，再往下看单个任务的回放。

可交付物：《运维手册》《巡检脚本》《培训与交接文档》。判断交接是否成功的标准只有一个：不依赖我们也能日常运行。

七、可复制的落地检查表

1. 《场景与边界说明》是否写完，输入、输出、判定标准、部署环境、责任人是否都在？
2. 方案类型是否说清了理由——为什么不是单智能体，为什么不是纯规则？
3. 数据通道是否选定，只读/脱敏/权限对齐方案是否明确？
4. 《数据与集成清单》里的前提条件是否都已落实，责任方是否签字确认？
5. 评测集是否建成、基线是否跑出、三级判定是否有说明？

电话咨询
公众号

6. 灰度是否有人工双复核，错例清单是否逐条闭环？
7. 上线后是否有人跑回归与巡检，指标是否配了阈值告警？
8. 留痕与回滚是否做过演练——包括一次真实的“从结论倒推到原始材料”？
9. 交接是否完成——新同事能否在不改代码的前提下把系统跑对？

这九条都不涉及算法，但每一条出问题都会让项目在半年后变成“没人敢改的系统”。智能体工程的难点，多半就在这里。

FAQ

关于这份白皮书

电话咨询
公众号