

企业 AI Agent 落地白皮书

企业智能体的架构选择 可靠性评估与项目交付

面向企业业务负责人 技术负责人及项目采购团队

讨论智能体如何进入真实流程 并以结果验证交付价值

广州数娱信息科技有限公司

2026 年 10 月 29 日 版本 1.0

执行摘要

企业部署 AI Agent 的价值，应通过任务结果、执行可靠性和每次成功交付的成本衡量。能进行流畅对话、调用多个接口或展示复杂计划，只能证明系统具备部分技术能力。企业需要进一步确认，它是否在真实数据、业务规则和异常条件下完成任务，以及失败后能否安全恢复。

本白皮书建议以明确业务流程为起点：先建立人工基线，再比较固定流程、单智能体和多智能体方案；把身份、授权、状态、预算和审计交给确定性系统执行；根据真实任务集评估最终业务状态。初次建设优先选择结果可核验、接口可用、异常可交接的场景。知识查询、资料整理、营销生产与内部协作可采用不同自主程度，不能以同一套自动化目标验收。

对采购方，项目应交付业务流程定义、工具与权限清单、评测集、运行记录及运维机制。对技术团队，优先建立可观察、可回退的最小系统，再依据失败证据增加规划能力或并行协作。对业务负责人，先证明一个完整任务的收益，再扩大岗位和系统覆盖范围。

研究方法与应用范围

本文采用公开论文、标准组织资料和技术提供方工程文档的桌面研究方法，资料检索截至 2026 年 10 月 29 日。学术研究用于分析能力与局限；厂商工程资料用于说明设计模式，不视为行业普遍效果；架构、评测门槛和成本算例属于本文的工程建议。

本文未实施自有企业问卷或跨产品统一实测。文中示例不代表客户已上线功能或数娱项目成效。基准中较早模型的结果只说明当时实验条件，不代表当前产品排名。面向财务支付、合同承诺等高影响操作，仍需按企业制度确定审批与责任。

阅读导航

01 企业任务与技术边界

03 架构与工具接口

05 企业场景设计

07 成本与收益

09 采购与交付清单

02 研究证据及其含义

04 权限与执行控制

06 评测与验收

08 实施与持续运营

参考资料

01 企业任务与技术边界

先定义一个完整业务任务

以“整理销售跟进事项”为例，完整任务包括确定客户与时间范围、读取允许访问的记录、提取待办、识别冲突、形成草稿、由负责人确认，再写入指定系统。回答一段建议和真正写入一条正确待办，其验收条件不同。项目立项时需要明确输入、允许动作、最终状态、异常出口和责任人。

任务也要定义停止条件。缺失客户编号时询问用户；没有相应权限时停止；发现重复事项时展示差异；达到预算上限时返回已有结果并解释未完成范围。若只写“帮助销售提高效率”，模型和供应商都无法据此判断完成与否。

按任务不确定性选择系统

方案	适用条件	优点	主要代价
单次模型调用	输入充分且输出可直接检查	易实现 易测量	难处理多步异常
固定 AI 工作流	步骤相对稳定 含分类或生成环节	路径与成本较可预测	流程变化需维护
单 Agent	后续步骤依赖中间结果	能按观察调整行动	需限制循环与工具范围
多 Agent	子任务可独立处理并能合并验收	可隔离上下文与分工	协调 通信及重复成本

本文将工作流定义为由程序预先确定主要路径的系统，将 Agent 定义为模型根据中间结果动态选择部分行动的系统。该区分参考 Anthropic 的工程分类，但企业可以使用不同产品名称；采购验收应看具体控制方式。[A2]

RAG 是为生成或行动提供知识证据的方法；MCP 是交换上下文和工具能力的协议；它们均不单独构成一个完整的业务 Agent。一个系统可以同时使用固定流程、检索增强和有限自主规划，三者之间不存在必须逐级升级的关系。

02 研究证据及其含义

学术结果支持机制判断

ReAct 研究将推理与行动交替组织，使模型能够根据外部观察更新后续操作。[A1] 其启示是为系统提供可信观察和纠错机会。企业仍需额外实现身份认证、任务状态、写入校验和故障恢复，不能将一种提示模式等同于生产架构。

r-bench 在模拟用户、业务工具和规则约束下，以任务结束后的数据库状态检验是否成功，并使用多次运行评估一致性。[A3] GAIA 则覆盖浏览、工具使用和多模态等组合任务。[A4] 两者表明，最终结果和跨步骤能力值得独立测试；其数据、接口和模拟用户并不等同于某家企业的真实工作环境。

工程经验需要按场景复验

Anthropic 的多智能体研究系统报告，在其工作负载中，多智能体系统的 token 使用量约为普通对话的 15 倍。[A9] 这是特定架构下的运行观察，不是所有多 Agent 系统的成本系数。它提醒项目组为并行任务计入重复检索、上下文传递和结果合并成本。

证据类型	能支持的判断	不能直接推导的结论
学术基准	某种机制在给定数据中的表现	本企业同样达到某个成功率
厂商工程报告	一个实现模式的收益与代价	所有模型或部署环境的统一效果
企业受控试点	特定岗位与任务的实际效果	所有部门均适合复制
日常运行记录	当前版本的稳定性与使用行为	无对照条件下的因果收益

因此，技术选型不应仅依据榜单、演示或“自主程度”。先用企业自己的代表任务比较简单方案，再决定增加的复杂度是否带来足够收益。

03 架构与工具接口

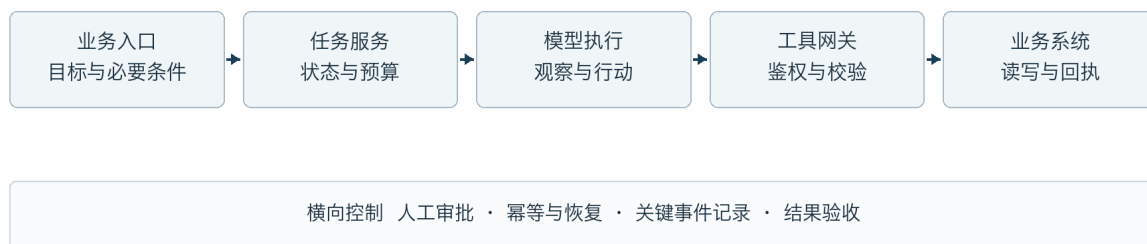


图 1 本文建议的企业 Agent 架构 其中授权和执行约束由系统实施

运行架构需要明确分工

业务入口负责收集目标与必要条件；任务服务建立编号、状态与预算；模型负责分类、生成或提出工具调用；工具网关负责身份绑定、参数校验与访问范围；业务系统提供可确认的读写结果；审计服务记录关键事件并支撑恢复。

执行过程建议维护显式状态，例如“待补信息、准备执行、等待审批、已完成、部分完成、已终止”。模型生成的自然语言不能直接改变状态，实际工具回执和程序校验才是依据。长任务应保存检查点；用户重新进入时先读取当前状态，避免重新执行已完成操作。

工具契约比工具数量更重要

每个工具需要说明输入字段、身份来源、可操作对象、返回结构、超时行为、错误码和重试规则。写入工具还需定义幂等键：同一任务重试时返回已有操作结果，避免重复创建记录。写入成功后若网络超时，先按幂等键查询状态，再决定是否重试。

例如 `create_followup` 不应只接受一段自由文本，而应接收客户标识、负责人、截止日期、事项内容、来源记录标识和幂等键。工具必须验证负责人是否存在、客户是否在授权范围内，不能把模型提供的 `user_id` 当成已认证身份。

MCP 官方文档规定了上下文和工具交换的参与方与基本能力，但协议并不决定业务应用怎样使用模型或管理上下文。[A5] 因此，采用 MCP 可以统一接入方式，权限继承、工具可信度、审计及高影响动作审批仍需由应用架构实现。协议版本和供应商支持状态应在采购时核对。

上下文按任务组织

上下文建议分为当前目标、必要约束、已确认事实、工具观察和历史摘要。外部文档、网页和邮件作为资料处理，不自动获得指令权限。长任务摘要应保留来源和未解决事项；需要精确核实时回读原始记录。上下文工程资料支持按相关性选择信息，而非持续堆叠完整历史。[A8]

04 权限与执行控制

自主范围需要落实到动作

动作等级	示例	建议控制
读取	查询已授权知识与业务记录	使用真实身份 校验对象范围
形成草稿	编写方案 生成待办建议	标明来源与待确认字段
可恢复写入	创建内部草稿 修改测试记录	校验参数 幂等处理 留存前后状态
外部影响操作	对外发布 发出报价 修改生产配置	按企业制度审批 绑定具体内容和对象
高影响或难恢复操作	支付 删除关键数据 合同承诺	明确人工责任与最终执行方式

授权应绑定任务、对象、动作、有效期及预算。用户说“帮我处理这批资料”不能被扩展为允许向任何外部服务发送全部原文。系统应在执行前校验数据流向，在操作后保存结果，避免依赖一句提示词约束所有风险。

提示注入按不可信输入处理

知识库、邮件和工具回执可能包含伪装成系统要求的内容。防护应包含来源标记、工具白名单、参数验证、数据外发限制和动作审批。文本清洗或“忽略恶意指令”提示可以辅助，但不应承担唯一防线。OWASP 的 Agentic Applications 风险资料为识别目标劫持、工具误用和身份滥用提供了框架。[A6]

还需控制无上限循环和恢复动作。每任务设置最大轮次、工具调用数、累计时间与预算；重试应区分网络故障、权限错误和业务冲突。权限错误通常不能通过再次调用解决；已发生的业务写入需要补偿或人工修复，不应笼统执行“回滚”。

NIST AI 600-1 是自愿性风险管理资料，可用于组织角色、测量和事件管理，并非中国法律要求或项目合规认证。[A7] 企业应结合自身适用制度确定控制强度。

05 企业场景设计

内部项目协作助手

输入为项目编号、会议记录和允许读取的任务清单。系统提取事项、匹配负责人、检索前序决策，生成待办草稿；存在相互矛盾的截止日期时列出冲突并请求确认。业务负责人确认后，工具写入项目系统，回读结果并返回记录链接。验收同时检查事项正确性、重复率、权限和后续可追踪性。

该场景适合先以固定流程配合模型抽取完成。只有在信息来源分散、补查步骤变化明显时，才需要增加有限的自主检索。模型提出的项目风险应附证据，不能自动把推测写成已经发生的事实。

艺术文博知识与内容助手

这是本文的设计示例。用户选择文物或艺术主题，系统读取已审定资料，组织鉴赏提纲、生成讲解草稿和多语种版本，并展示可追溯依据。存在归属争议或年代不确定时，保留原文限定；图像识别结果作为候选线索，不直接替代专家鉴定。

内部研究资料、已审定公众内容和版权受限图片应设置独立范围。对外讲解使用审定知识；内部策划可以引用更广资料，但发布前转交审核。该设计的关键是把知识、生成和审定状态连接起来，而非让同一会话无限访问所有资产。

营销生产协作助手

从结构化 brief 开始，调用产品事实库、品牌规范和授权素材，生成脚本与候选物料。系统检查产品属性、禁用词、素材授权和发布规格；人工选择后进入渠道发布草稿。预算、对外承诺和发布对象由明确规则控制。效果回流用于版本比较，不允许把一次点击提升自动解释为长期销售增长。

三类场景都可以复用知识检索与生成能力，但它们的用户、权限、交付物和验收方式不同。平台提供通用底座，定制交付负责把底座映射到具体业务任务。

06 评测与验收

任务集应包含失败路径

建议先建立 120 个任务的启动评测集，作为本文规划示例：正常任务 50 个、信息缺失 20 个、权限限制 15 个、冲突或过期信息 15 个、工具故障与重试 10 个、提示注入及重复请求 10 个。数量可以随业务调整；任务应来自脱敏真实流程，并由业务与技术共同标注。

开发集用于改进，冻结测试集用于版本验收，另留一组上线后新增的困难任务。相同任务重复运行三至五次，观察成功率与成本波动；该次数属于初期诊断建议，不足以证明高可靠场景的统计要求。

指标	定义与测量方式	验收关注点
端到端成功率	最终业务状态正确的任务数除以全部有效任务数	同时满足业务规则
重复运行一致性	同一任务多次均成功的任务占比	单次演示不能替代
未授权动作	越权读取 写入或外发的观察次数	所有发现项必须修复
人工介入率	需要额外介入的任务占比	区分计划内审批与意外救援
成功任务成本	总模型 工具及相关运行费用除以成功任务数	计入失败与重试
时延分布	从收到完整输入至可用结果的 P50 和 P95	分开记录排队与审批时间

阈值应根据人工基线和错误损失设定。例如内部草稿可以接受可识别的待补字段；生产系统写入必须满足对象、数量及状态校验。高影响错误不应被大量普通任务的平均分抵消。

评分需要可复核

对于结构化结果，优先使用数据库状态或规则检查。对于方案和文本，由领域人员按准确性、完整性、可执行性和来源质量评分。模型裁判可协助筛查，但应抽样人工复核并记录分歧。测试失败要定位到知识缺失、规划失误、工具错误、权限问题或验收定义不清。

如果 120 次测试未发现某类错误，不能推导错误率为零。在独立同分布等近似条件成立时，零事件的 95%上界可用 $3/n$ 估算， n 为测试次数； $n=120$ 时约为 2.5%。真实攻击与长尾任务通常不满足这些假设，这一结果只能解释有限样本的边界。

07 成本与收益

用成功任务口径核算

总成本应包含模型输入输出、检索重排、工具调用、长任务状态、日志与存储、人工复核、知识维护及故障处理。不能只比较每百万 token 单价。相同任务在一个方案中一次完成，在另一个方案中需要重试三次，实际单位成本可能发生反转。

成功任务运行成本 = 统计期内模型与工具等运行费用 ÷ 验收成功任务数。项目经济性则需进一步扣除人工审核、持续运维和初始建设投入。

以下均为假设算例，不是市场报价或项目承诺：某团队每月处理 2,000 项任务，人工平均 20 分钟；上线后每项平均复核 8 分钟，实际使用覆盖率 70%，综合人工成本 100 元每小时，则月度可释放工时为 280 小时，对应 28,000 元工时价值。若每月运行和维护成本共 12,000 元，初始投入摊销前净值约 16,000 元。

实际覆盖率	可释放工时每月	对应工时价值	扣除月度成本后的净值
40%	160 小时	16,000 元	4,000 元
70%	280 小时	28,000 元	16,000 元
90%	360 小时	36,000 元	24,000 元

工时释放不等于现金节省。只有企业确实减少加班、外包或增加有效产出时，才能转化为相应财务价值。该算例还假定质量相当、失败返工已包含在平均复核时间中；真实项目应单独检查这些条件。

复杂度增加需要证据

多智能体只有在子任务可拆分、上下文隔离有收益、合并结果能校验时才有明确理由。否则增加的通信可能重复消耗信息和预算。升级更强模型之前，先检查资料完整性、工具定义、任务边界和失败处理；模型升级也要重新跑固定测试集，避免把局部提升当作整体收益。

08 实施与持续运营

以阶段结果决定下一步

阶段	主要工作	进入下一阶段的条件
业务诊断	梳理任务 数据 接口 责任及人工基线	有可验收任务与业务负责人
原型验证	比较简单流程与 Agent 方案	关键任务可以复现并识别失败
有限试点	接入身份 日志 审批与监控	权限及异常路径通过测试
生产上线	灰度发布 备份 恢复与培训	有值守方式和回退方案
持续运营	分析任务失败 成本与使用反馈	依据版本证据扩大范围

阶段周期由接口接入、数据整理和组织配合决定，不应在需求未明确时承诺统一上线天数。供应商应在早期提供依赖清单：谁批准接口、谁维护知识、谁定义成功、谁处理异常。

上线后跟踪任务量、失败类型、审批等待、单成功任务费用及用户继续使用情况。新增工具、修改权限、变更模型或检索策略都属于需要回归验证的配置变化。先在影子模式中生成建议，再以小范围真实任务试运行；影子结果不能替代写入和恢复测试。

训练用户同样重要。用户需要知道怎样给出任务范围、何时核对证据、怎样报告错误，以及哪些动作仍由人决定。业务负责人应定期检查系统是否把原流程中的低效步骤固化，而非默认所有旧流程都值得自动化。

09 采购与交付清单

采购方宜要求供应商用企业脱敏任务现场展示，并解释失败与人工接管方式。比较时使用相同任务、资料权限、预算和验收口径。演示过程中的人工准备工作、额外脚本和后台干预也应计入交付说明。

交付物	最低应包含的内容
业务与范围说明	用户 任务 输入输出 排除范围 成功定义
数据与工具清单	来源 更新责任 接口契约 访问范围
架构与运行配置	模型路由 状态管理 预算 降级及恢复

交付物	最低应包含的内容
权限与审计说明	身份链路 审批规则 关键事件与留存策略
评测资料	任务集 评分标准 版本结果 失败分析
运维与移交	培训 升级方式 监控告警 数据导出与退出机制

服务商的交付价值在于把行业知识、系统接口和岗位流程连接起来，并持续验证结果。自研平台可以复用身份、知识、生成和协作能力；具体项目仍需明确使用了哪些模块、进行了哪些定制，以及谁负责上线后的维护。

数娱科技的 AI 服务以企业应用定制与落地为总入口，EVIPX 作为产品平台能力之一。企业可从一个具体任务发起需求诊断，再决定采用标准产品、定制应用或组合方案。项目范围和可用能力以双方确认的方案及实际交付为准。

参考资料

[A1] Yao S 等 ReAct Synergizing Reasoning and Acting in Language Models 2022 ICLR 2023

<https://arxiv.org/abs/2210.03629>

[A2] Anthropic Building effective agents 2024 工程方法 <https://www.anthropic.com/research/building-effective-agents>

[A3] Yao S 等 τ bench A Benchmark for Tool Agent User Interaction in Real World Domains 2024

<https://arxiv.org/abs/2406.12045>

[A4] Mialon G 等 GAIA a benchmark for General AI Assistants 2023 <https://arxiv.org/abs/2311.12983>

[A5] Model Context Protocol Architecture overview 官方协议文档

<https://modelcontextprotocol.io/docs/learn/architecture>

[A6] OWASP Top 10 for Agentic Applications for 2026 2025 年 12 月发布 <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>

[A7] NIST Artificial Intelligence Risk Management Framework Generative Artificial Intelligence Profile AI 600 1

2024 自愿性框架 <https://doi.org/10.6028/NIST.AI.600-1>

[A8] Anthropic Effective context engineering for AI agents 2025 工程方法

<https://www.anthropic.com/research/effective-context-engineering-for-ai-agents>

[A9] Anthropic How we built our multi agent research system 2025 厂商系统运行观察

<https://www.anthropic.com/research/multi-agent-research-system>

以上网页访问日期均为 2026 年 10 月 29 日。较早研究的模型版本与任务条件需结合原文理解；本文未将历史分数用于当前产品排名。

关于数娱科技

广州数娱信息科技有限公司围绕企业 AI 应用与文旅数字化开展业务。企业 AI 服务包括知识应用、生产力工具、工作流及多模态应用等方向，具体范围按项目确定。EVIPX 是 AI 板块中的产品平台能力。

官网 <https://shuyukeji.cn/> 产品平台 <https://evipx.com/>

业务联系 market@shuyukeji.cn 020 38080881