

PUBLIC CASE STUDY

Agent 长任务实施质量量化评估

SalesTouch Max 两阶段计划 · 评估日期：2026年9月11日

总体判断：实施能力较强，但独立交付可靠性不足。本次有实质性工程产出，也暴露了重大边界缺陷与验收证据缺口。适合承担长任务实施，不宜由同一个 Agent 独立完成最终验收与发布放行。

计划实施难度

8.5 / 10

跨业务域、高约束任务

本次长任务综合表现

60 / 100

有产出，仍需强复核和返工

独立验收结论

Blocked

两个阶段均存在未关闭阻塞

1. 计划难度与分阶段表现

任务范围	难度 / 10	表现 / 100	判断依据
Stage A：日常文字工作台	7	55	涉及文档状态、版本、响应式与真实使用验收；局部 UI 有改进，但状态隔离与计划内闭环漏验。
Stage B：销售专家能力层	9	62	横跨流程事实、权限、B2B/B2C、迁移和评测；实现量大，但权限、溯源与业务入口仍有缺口。

分阶段表现为辅助专家判断，不是需求完成率。综合难度约为 8.5/10；综合表现按下表独立加权，不由两个阶段分数简单平均。

2. 综合得分：60 / 100

评估维度	权重	得分 / 100	加权得分
任务拆解与持续推进	15%	85	12.75
代码实现与集成	25%	75	18.75
安全边界与数据一致性	20%	35	7.00
测试设计与反例防线	15%	65	9.75
真实业务闭环与 UI 验收	10%	40	4.00
证据准确性与结论校准	10%	35	3.50
执行纪律与可追溯性	5%	85	4.25
合计	100%	不适用	60.00

计算方式：各维度得分乘以权重后求和。任务难度不抵消安全缺陷；60 分不等于验收通过，任何未关闭的 SEV1 重大问题均继续阻断发布。

优势与劣势：实施强，闭环弱

3. 主要优势

能力	得分 / 100	本次体现
持续推进与任务拆解	85	能持续完成跨前后端、测试和文档工作。发现真实反馈不足后，按计划将自动演进收缩为“采集先行”，而非机械执行原清单。
工程实现与集成	75	多个模块有真实实现，类型检查及大量专项测试通过；补齐部分国际化、窄容器和移动端问题，不是仅提交说明文档。
执行纪律与可追溯性	85	遵守授权和分支边界，保留迁移与清理记录；没有为了获得通过结论而擅自注册真实组织。

适合承担的工作： 在合同和边界清晰、且有独立复核的条件下，完成大范围实现、迁移准备、测试补充和技术文档整理。

4. 主要劣势

能力	得分 / 100	关键不足及影响
安全边界与数据一致性	35	复核发现三组重大问题：可信身份与权限边界、文档作用域及版本隔离、溯源信息往返一致性。部分来自既有实现，但本轮验收仍未识别。
真实业务闭环与 UI 验收	40	存在组件已完成但业务入口未接线的情况；页面可见或仍在加载，不能证明用户任务完成。写入读回、异常恢复及回滚证据仍不足。
证据准确性与结论校准	35	“唯一阻塞”“结构性不可能”等结论超出证据范围。把局部测试、截图或函数预检等同于完整验收，导致对交付可靠性的判断偏高。
测试设计与反例防线	65	真实运行评测和可判失败的测试是进步；但对重复处理、存储往返、切组织、切文档和检查器自身失败等相邻场景覆盖不足。

最突出的问题不是“没有写代码”，而是从“局部通过”过早推导“全链路完成”。 单点修复后，尚未稳定地验证同类问题是否在相邻输入、状态切换和正常读写往返中继续存在。

不因守住授权边界扣分：未获授权而不执行真实注册或写入，是正确行为。扣分来自未接线、漏验却声称通过，以及遗漏或低估其他阻塞项。

证据边界与使用建议

5. 数字如何解读

复核结果	能证明什么，以及不能证明什么
840 项既有专项测试通过	前端 681 项、后端专项 116 项、行业边界 43 项；前后端类型检查通过。证明这些工程检查通过，不代表全部计划完成或用户闭环可靠。
7 项定向边界检查失败	分属四类问题，说明现有测试遗漏了关键不变量；不是七个独立产品缺陷，也不是随机抽样，不能据此计算整体缺陷率。
三组 SEV1 重大问题	来自本地合成复现与代码复核，属于权限、文档隔离和溯源一致性问题；不表示已经观察到生产攻击或真实客户数据泄露。
真实登录与模型运行证据	本次复核审阅已有材料，未重新执行真实登录、实时模型调用或查询当前生产数据库；不将历史结果冒充本次实时验证。

6. 推荐的使用方式

工作环节	建议安排
长任务实施	可以交由 Agent 承担，但要求拆分里程碑、保留真实结果，并让测试具备判失败能力。
安全与数据复核	由独立评审者检查可信身份、组织与对象作用域、状态往返、异常恢复，以及门禁自身是否可靠。
真实业务验收	独立验证“用户操作、业务写入、读回、刷新、异常和回滚”链路，不能仅看页面存在或单测数量。
发布放行	依据逐项证据关闭阻塞。注册授权仅解决一个前置条件，不代表其他缺陷自动关闭，也不意味着自动 Ready。

最终建议：让该 Agent 做长任务实施者，不让它成为唯一的验收签字者。当前两个阶段均为 **Blocked**；安全、数据和真实业务闭环应由独立验收把关。

7. 评分方法与适用范围

- 这是单次任务的事后专家评分。不代表模型通用排名、长期稳定性、线上成功率或相对人效；也不是需求完成百分比。
- 分数量表：90 分及以上为高质量闭环；80-89 分主要依赖抽查；70-79 分仍有明确且可控的补项；60-69 分有实质产出但需要强复核和返工；低于 60 分为交付可靠性不足。所有档位均不能豁免重大阻塞。
- 难度量表：1-3 为单点修改；4-6 为单模块闭环；7-8 为跨前后端状态与集成；9 为多业务域、迁移、安全及真实验收共同约束；10 为长周期多系统重大迁移与生产连续性。
- 结论仅对应本次复核快照。后续修复须重新验证；本报告不预先认定修复已完成或授权任何真实业务操作。

评估依据：2026年9月8日、9日两阶段实施计划及对应实现、测试和验收材料，结合9月10日至11日的独立复核。本分享版保留评分与判断依据，不附仓库路径、提交记录、访问凭据或客户数据。