

无人值守验收：完整演示流程

本文记录 2026-08-08 在 SG 生产环境真实运行的 EPIC-23 演示。

演示数据全部为合成数据；截图来自 Windows 原生 Chrome 中保留的 Liuma 会话与同哈希交付物。

这项能力解决什么问题

把一件需要多轮尝试的工作交给流马后，用户不必守在旁边反复问“做好了吗”。流马会先把目标、验收标准和搜索预算写成可见契约；每轮候选都必须经过买方提供的验收器，并把具名判据、结论和环境证据写进判定卡；交付前还要把结果实际渲染出来看过。

本条演示的业务载体是一个单文件内部系统原型，但被验证的是通用机制：

一句话交活 + 买方自带验收器 → 契约可见 → 候选被机械判据拦截 → 按错误收敛 → 自看 → 具名判定 → 用户追加判据 → 再验收 → 最后交付

0. 演示现场

演示会话为 `e91ce597-2d8d-4ca4-b0d6-cca477550c9c`，PursueRun 为 `e78f5a91-ee4d-4eef-aae5-1c666f8c6658`。任务只引用组织工作区里的 `base/demo-base.json`；公司名、系统名、字段和数据均来自底座，任务话术中无业务值。

交付物 `prototype.html` 为 23,959 字节，sha256 为 `af063c01edc2845bea3156bad39602cdb3d08383fe2aaab2a4608024c6f71ad8`。

1. 用户交活，并交出验收脚本

用户用 `/pursue` 说明目标：读取数据底座，生成一个可双击打开、零外部依赖的单 HTML 内部系统；同时把冻结的 `verify_prototype.py` 作为附件交给流马，并明确要求脚本不得修改、必须跑到 `verdict=pass`，交付前还要实际渲染。



用户交活并附上验收脚本

这里的关键不是原型本身，而是验收口径由买方带入。流马可以执行它，但不能改写它。

2. 流马把验收契约回显出来

流马读取底座和验收脚本后，回显唯一——张契约卡：目标引用为 `single_html_file / prototype.html`；验收标准为 `single_file`、`reference_integrity`、`positive_connectivity`、`data_consistency` 四项结构判据；预算为 10 轮、30 分钟、plateau 3。



目标、验收标准和预算组成可见契约

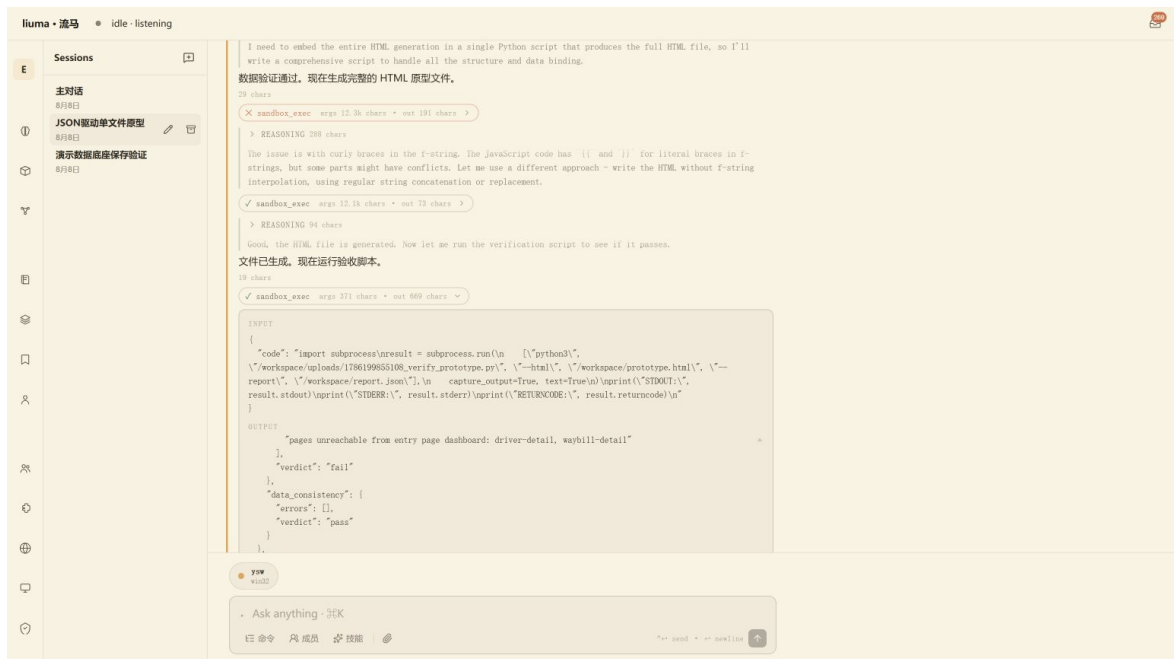
这不是审批流。契约的作用是把流马理解到的判据摊在明面上，用户仍可以继续用自然语言修改要求。

3. 冻结验收器拦下第一个候选

首个候选生成了运单详情页和司机详情页，但配置中没有从列表指向详情页的 `navigate` 边。

验收器没有接受“页面已经生成”的自述，而是给出 `verdict=fail`：

`pages unreachable from entry page dashboard: driver-detail, waybill-detail`



[冻结验收器拒绝不可达的详情页](#)

4. 流马按错误清单收敛

流马从错误中确认根因：详情页存在，但入口页无法沿配置图到达。它在两个列表页补齐 13 条下钻边（8 条运单 + 5 名司机），并保留两个详情页的返回边，然后重新生成候选。



流马识别详情页缺少 navigate 边并修正

这是“判据不是橡皮图章”的直接证据：失败项改变了下一轮候选，而不是被文字解释绕过。

5. 交付前先实际渲染

四项脚本检查通过后，流马没有马上发文件，而是调用 `render_file`。渲染结果以图片出现在会话 timeline 中，能看到左侧菜单、看板指标和数据表。



`render_file` 结果以图片进入会话

6. 第一轮判定可查

第一轮判定卡记录 **pass / 100**，并逐项列出四条判据及 `/workspace/report.json` 证据引用。



第一轮具名判定卡

这张卡证明的是：本轮机械检查四项全过，并且看板、菜单和列表已被渲染确认；它不等于证明所有页面的运行时渲染都正确。

7. 用户用一句话追加验收标准

用户追加两条业务标准：看板必须增加覆盖全部运单的“运费总额”；司机详情页必须展示该司机名下的所有运单。验收脚本仍不允许修改。



用户追加判据，流马确认按新要求继续

本次会话没有生成第二张契约卡；变更由用户消息、流马回应和第二轮判定共同追踪。

8. 改完再过一遍

流马加入第 6 个指标，使用没有 `where` 过滤的全表求和：

$4200 + 1800 + 3150 + 900 + 2400 + 1250 + 0 + 3300 = 17000$

同时给司机详情页加入按 `driver_id` 过滤的关联运单子表。第二轮判定卡再次记录四项具名判据全部 通过，best iteration 为 `b45da0fd...`。



第二轮判定卡与运费总额证据

9. 判定之后才发交付卡

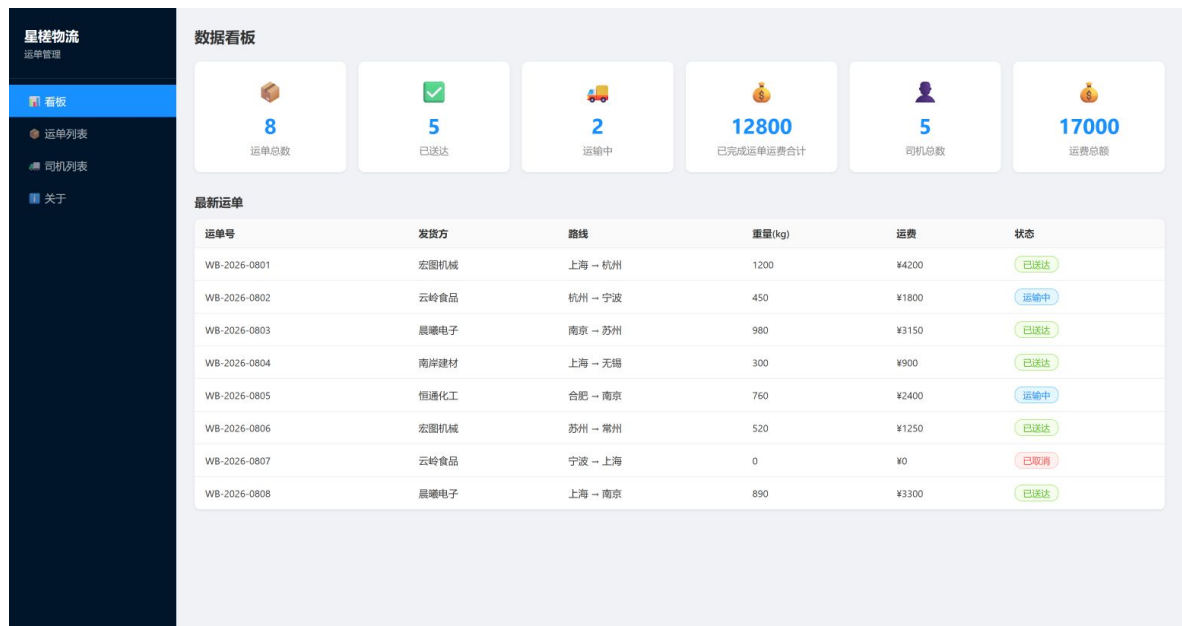
最终文件通过 `preview_file` 进入会话，交付卡指向 `prototype.html`。



最终 prototype.html 交付卡

10. 产物本身可看、可点、可复核

看板有 6 个指标：运单总数 8、已送达 5、运输中 2、已完成运单运费合计 12800、司机总数 5、全部运单运费总额 17000。下方 8 行运单与底座逐行一致。



交付原型的看板与六个指标

运单列表显示 8 行数据，状态分布为已送达 5、运输中 2、已取消 1。

运单号	发货方	出发地	目的地	重量(kg)	运费	状态	出发日期
WB-2026-0801	宏图机械	上海	杭州	1200	¥4200	已送达	2026-08-01
WB-2026-0802	云岭食品	杭州	宁波	450	¥1800	运输中	2026-08-02
WB-2026-0803	晨曦电子	南京	苏州	980	¥3150	已送达	2026-08-03
WB-2026-0804	南岸建材	上海	无锡	300	¥900	已送达	2026-08-04
WB-2026-0805	恒通化工	合肥	南京	760	¥2400	运输中	2026-08-05
WB-2026-0806	宏图机械	苏州	常州	520	¥1250	已送达	2026-08-06
WB-2026-0807	云岭食品	宁波	上海	0	¥0	已取消	2026-08-07
WB-2026-0808	晨曦电子	上海	南京	890	¥3300	已送达	2026-08-08

交付原型的八行运单列表

在 Windows Chrome 中点击 WB-2026-0801 可进入详情，页面显示发货方、路线、司机、重量、运费、状态和日期；点击 “← 返回” 后 hash 从 #waybill-detail 回到 #waybills。



运单详情页字段与返回按钮

证据索引

对象	证据
演示会话	e91ce597-2d8d-4ca4-b0d6-cca477550c9c
PursueRun	e78f5a91-ee4d-4eef-aae5-1c666f8c6658 , 终态 done
第一轮 iteration	3f5fecbf... , pass / 100
第二轮 iteration	b45da0fd... , pass / 100, best
冻结验收器	sha256 527fd4551e71007ecafae5b4f428d01f2c9393dc0d8b943bd455a97ef9842fe7
最终交付物	sha256 af063c01edc2845bea3156bad39602cdb3d08383fe2aab2a4608024c6f71ad8
数据真值	base/demo-base.json
同哈希交付物副本	evidence/prototype.html

已知边界

以下事实按现场保留，不把缺陷包装成能力：

1. 司机详情页的关联运单子表正确，但司机编号、姓名、电话、车牌、基地五个字段恒空。冻结验收器只查结构，`render_file` 又没有覆盖 hash 路由页，因此两道关都漏掉了这个运行时渲染缺陷。
2. run 进入 `done` 后，第二次 `mark_done` 被拒绝，但 `log_iteration` 和 `select_best` 仍可写入，导致 `completed_at` 早于 best iteration。
3. `read_file` 读取 HTML 一次返回 Markdown 提取、一次只返回 1 字节，流马改用沙箱读取源码。
4. `write_file` 保存底座时丢失文件末尾换行；补回换行后内容哈希一致。
5. 契约预算为 10 轮 / 30 分钟 / plateau 3，本次只用 2 轮 / 11 分 40 秒，但运行中没有机制读取或强制该预算。
6. 非 Recipe 目标仍会在 tick 卡片中看到一大段 Recipe 专属指令，本次由流马自行判定为不适用。