危机预演平台

商业化改造方案

从“一次性买断的证据包”到“可复算的决策沙盘”

文档版本	v1.0
日期	2026-11-23
对应代码	crisis-rehearsal-pro 0.4.0（113 项测试通过）
适用范围	产品改造、定价、开源发布与对外尽调

阅读指引

本文回答四个问题：**现状为什么卖不掉、改什么才能卖得掉、每一项怎么落地、改完值多少钱**。所有引用的数字都来自当前仓库的实跑结果，不引用记忆中的估计。

章节	标题	回答的问题
一	执行摘要	三句话结论与关键数据
二	现状诊断	客观数据、四个结构性折价、真正的瓶颈
三	改造总纲	一句话方向与三个替换
四	引擎重构	从多智能体聊天到可审计响应模型
五	产物重构	从报告到决策卡与会议包
六	数值层	已交付：区间、覆盖度与发布门禁
七	物理与实验数值	Oracle 适配器协议与前置条件
八	证据闭环	三臂盲评、样本外、人工 κ、敏感性
九	语料改造	三个新字段与英文语料
十	交付与运营	工作坊模式、导出、离线运维
十一	路线图与验收	P0–P4 的范围、产出与验收线
十二	商业模式	三种卖法、定价阶梯与估值影响
十三	风险与边界	合规、方法论风险、明确不做的事

表 0 文档结构

一、执行摘要

结论一：现在的资产卖不出价，不是因为技术差，而是因为“证明”依赖一个你已经不发布的引擎。 仓库能重算 82.2%，但重跑不了这 82.2%——买家问一句“能重现吗”，答案是不能。这是尽调桌上的硬伤。

结论二：商业价值的瓶颈是卖错了东西。 一份风险报告值 2000 元，一场带决策记录的演练值 3–5 万元。买家付钱买的是“在时间压力下做对决定并且能证明当时为什么这么决定”，不是一份 PDF。

结论三：精确点位不可交付，但校准区间可以。 实测 NVDA 次日收盘的朴素误差是 3.50 美元（2.02%），BTC 是 1387 美元（1.65%），动量模型比随机游走更差。任何“明天精确到某价位”的承诺，误差由波动率决定，不由模型决定。可交付的是分布、覆盖度，以及“什么时候不可知”的判断。

指标	数值	口径
事件均值命中率	**82.2%**	LLM 判分，30 事件 × 3 跑，t 区间 78.5–85.9%

指标	数值	口径
跨跑一致率	5 / 30（17%）	三次运行得分完全一致的事件数
离线基线	24.4% / 45.6%	关键词判分，严格 / 宽松两档
NVDA 次日误差	3.50 美元（2.02%）	朴素预测，501 个交易日，walk-forward
BTC 次日误差	1387 美元（1.65%）	朴素预测，900 个交易日，walk-forward
工程状态	113 项测试通过	46 个语料文件哈希校验，wheel 62.8 KB，零依赖

表 1 当前基线（全部实跑可得）

改造分四步走，每步都有可验收的产物，而不是“先重构再说”：

阶段	周期	核心产出	验收线
P0（已完成）	—	数值层 crh forecast、发布门禁	打不过随机游走则不输出点位
P1	1 周	条件重估：事件 → 分布怎么平移	同一事件两次运行结果逐字节一致
P2	2 周	三臂盲评、样本外、人工 κ	留出集不低于人类基线，κ ≥ 0.6
P3	2 周	工作坊模式、决策日志、DOCX/PPTX 导出	一场 90 分钟演练只用一个工具
P4	1 周	英文语料 30 事件	海外演示可用英文端到端跑通

表 2 四步路线

二、现状诊断

2.1 客观数据

仓库公开后四周的状态（GitHub API 与作者自建的流量快照）：

项目	数值	含义
Star / Fork / Watcher	2 / 0 / 0	其中一个是作者本人
外部 Issue	0	4 个 issue 全部由作者自己创建
提交	15 个，集中在 5 天内	之后近四周无提交
14 天流量	40 浏览 / 7 独立访客	114 次 clone / 55 个独立 cloner
品类规模	12 个仓库，最高 11 star	crisis-simulation 全 GitHub
品类第一	attackgen 1258 star / 3 年	约 0.4 star/日

表 3 需求侧的客观规模

注意第 4 行：**55 个独立 cloner，0 个外部 star。** 这说明拿到代码的人读得懂，但没被说服。问题不在可见度，在“读完之后没有得到一个可验证的承诺”。

2.2 四个结构性折价

- **引擎不在手里。** 仓库只有对上游引擎的 5 个补丁（555 行），核心模拟能力不属于这个资产。买家两天就能复现，任何高于 5 万的报价都会被“我自己做”击穿。
- **证明依赖旧引擎。** 82.2% 的 90 条判分记录由旧管线产生，而旧管线已不在交付范围内。能重算、不能重跑，等于“只能自证、不能被证”。
- **许可与派生。** 旧版本以 AGPL 派生发布，最主流的买断动机（改成闭源商业产品）在法律上被堵死。这一条已在本版解决：代码全自研、许可改为 MIT。
- **需求侧接近零。** 整个品类 12 个仓库、最高 11 star。这不是营销能解决的，是赛道本身没有开发者需求。

2.3 真正的瓶颈

把三类买家的付费理由排一下，结论很清楚：

买家	为什么付钱	愿意付	当时能交付的
公关 / 危机公司	向客户交付一场演练	2–5 万元 / 场	一份 8000 字档案
企业 CRO / 法务	决策有据、事后可免责	5–20 万元 / 年	一个风险清单
政府 / 国企	台账、流程、演练记录	10–15 万元 / 项目	离线 CLI，没有流程

表 4 买家付费理由与当时的交付物不匹配

三个买家要的都是**“决定”**，而产品交付的是**“判断”**。这就是瓶颈。

三、改造总纲

一句话方向：把“预测风险的模拟器”改成“可复算的决策沙盘”。

具体是三个替换，每个替换都对应一个可验收的产物：

维度	现在	改造后	为什么更值钱
引擎	多智能体两轮对话，依赖外部引擎与 key	显式响应模型，8 类主体 × 72 时间步，< 1 秒，确定性	会议室里可复算、可改参、可辩护
产物	8000 字风险报告	决策卡 + 会议包（议程、注入、决策日志、异议记录）	买家为“会”付钱，不为“文档”付钱
证明	单臂 LLM 判分，17% 复现	三臂盲评 + 样本外 + 人工 κ + 敏感性	能回答“凭什么信你”

表 5 三个替换

3.1 保持不变的三件事

- **T−72h 干净语料纪律。** 只用事发前 72 小时公开信息，这是产品可信度的地基，不能让步。

- **哈希校验与可复算。** 每个数字都能从语料重算，这是与“AI 黑箱”区分的唯一方式。
- **公开局限。** 敢写“这个数不可比、这个模型打不过随机游走”，在受监管行业是加分项。

四、引擎重构：从多智能体聊天到可审计响应模型

4.1 为什么旧引擎不可辩护

旧管线的问题不是“不够大”，是不可审计：7–12 个 agent 聊两轮，采访接口返回 0/19，三次跑只有 17% 结果一致。这种系统在会议室里无法回答三个问题——为什么是这三个风险、改一个输入会不会变、结论对参数有多敏感。

4.2 状态与动力学

新引擎是一个**可以画在白板上**的离散时间模型：8 类行为主体（公众、媒体、监管、投资者、内部人、渠道、竞品、法律原告），8 个渠道，72 个时间步（每小时一格），每类主体三个状态：注意力 A、立场 S、行动准备度 R。

```
A_g(t+1) = A_g(t)·(1-λ_g) + Σ_c w_gc·V_c(t) + κ·Σ_h C_gh·A_h(t)
V_c(t+1) = V_c(t)·(1-μ_c) + Σ_g α_cg·A_g(t)·max(0, S_g(t)) + shock(t)
R_g(t+1) = R_g(t) + β_g·(A_g(t)-θ_g) + γ_g·[S_g(t) < 0]

监管立案 / 下架 / 刑事追责：当 R_g ≥ 1 且证据存在时触发，按类型带延迟
```

式 1 响应模型的三条状态方程（参数约 40 个，全部版本化）

4.3 参数治理

- 参数存放在 model/params.v1.json，每个参数带一行理由与出处。
- 任何参数改动必须跑一次敏感性分析，输出影响排序。
- 调参只允许在训练集（20 个事件）上做，测试集（10 个事件）只测不调。

4.4 输出指标

指标	定义	用途
峰值 V_max	72 小时内声量指数最大值	判断事件量级
长尾时长	回落到峰值 20% 所需小时数	判断要占用多久的注意力
不可逆标志	监管处罚 / 下架 / 刑责是否触发	判断是否还有退路
路径残差	采取该应对后剩余的复合风险	三条路径的横向比较

表 6 模型输出指标

4.5 与旧引擎的对照

	旧：多智能体	新：响应模型
复现性	17%	100%（确定性）
单次耗时	约 4 分钟 + 0.03 美元	< 1 秒，0 成本
可解释	“agent 说的”	每一步有规则与参数
现场可改	不能	改一个参数即时看差异
断网可用	不能	可以

表 7 引擎对照（第 4、5 行决定工作坊能不能开）

五、产物重构：从报告到决策卡

新引擎的输出不是“Top-3 风险”，而是**必须在几小时内做的那个决定**：

```
决策点：是否在 6 小时内公开道歉并启动召回?
├ 选项 A 沉默          峰值 0.72 / 长尾 51h / 不可逆风险 低   / 可撤回
├ 选项 B 道歉+召回     峰值 0.55 / 长尾 28h / 不可逆风险 中   / 不可撤回
└ 选项 C 对抗          峰值 0.88 / 长尾 96h / 不可逆风险 高   / 不可撤回
触发条件：24h 内出现监管立案信号 → 立即转 B
反悔条件：道歉后 48h 内二次爆料 → 升级为 B+
责任人 / 截止时间 / 依据（模型轨迹 + 同类事件频次）
```

式 2 决策卡结构

配套产出**会议包**：90 分钟议程、4 个按时投放的注入、角色卡、决策日志、异议记录、会后 readout。这才是能收 3–5 万元的东西——报告只是会议包的附件。

六、数值层（已交付）

针对“精确点位”的要求，本版交付的是**唯一能通过尽调的形态**：校准区间 + 方向概率 + 发布门禁。实测（walk-forward，60 根预热，留出 30% 打分）：

序列	样本	日波动 σ	朴素误差	最好模型	相对朴素	90% 区间
NVDA	501 日	2.9%	3.50 美元（2.02%）	naive	0.0%	120.30 – 130.54
BTC/USDT	900 日	2.4%	1387 美元（1.65%）	naive	0.0%	58494 – 63131

表 8 次日点位的误差由波动率决定，不由模型决定

趋势外推与均值回归均未跑赢随机游走；5 日动量模型更差（−3.1% / −3.8%）。因此门禁判定失败，**拒绝输出点位，只输出区间**，退出码 3。

```
NVDA · last 2026-11-20 = 125.53
  80% 区间   121.79 .. 129.43
  90% 区间   120.30 .. 130.54
  P(上涨)    54%      样本外覆盖 79% / 89%
GATE: FAIL — point forecasts are refused.
```

式 3 实际输出（数值来自仓库内 data/market/nvda.csv）

为什么这是可卖的： 卖点位是卖一个买家 24 小时就能证伪的东西；卖带覆盖度的区间，是卖一个能过审计的东西。后者才有续费。

6.1 商业含义

数值层真正的用途不是"猜收盘"，而是回答条件问题：**如果这个事件落地，分布怎么移动、谁先重估、谁先恐慌。**这一步把行情只当作噪声尺度，把 30 个中文事件语料当作真正的资产——这也正是别人复制不了的部分。

七、物理与实验数值：Oracle 适配器协议

芯片良率、降水毫米、药物毒性不可能由语言模型算出来。它们来自工艺数据与仿真器。平台能提供的价值是**包在真仿真器外面的一层**：分布化、不确定性传播、校准与审计。

```
class Oracle(Protocol):
    domain: str      # weather | process | pharmacology
    unit: str
    def query(inputs) -> Distribution: ...    # 返回分位数；返回点直接拒绝
    def validate() -> ValidationReport: ...   # 留出误差 + 区间覆盖 + 数据溯源
```

式 4 适配器协议（类型层面拒绝点预测）

域	后端	前置条件	交付形态
气象	NWP 集合预报	公开集合数据或商业接口	降水区间与超阈值概率
工艺良率	SPC / TCAD / 工艺数据回归	产线数据协议与漂移假设	良率分布与敏感性
药物毒性	PBPK / QSAR	毒理数据与模型验证	剂量-反应区间

表 9 三类适配器：每一个都需要数据协议与领域验证，不能靠语言模型

四条铁律：只收分布不收点；每个输入带来源与时间戳；没有验证报告一律标记 unvalidated 且不得过门禁；领域模型必须可解释、可替换。**当前未接任何适配器**——协议先落地，第一个适配器是一周的活，不是重写系统。

八、证据闭环

"更好"必须是可测的，否则只是形容词。三件事构成闭环：

手段	做法	回答的问题
三臂盲评	人类分析师 / 旧 LLM 管线 / 新引擎，同一 rubric 盲评	新引擎是否真的更好
样本外	20 事件调参、10 事件只测	数字是否过拟合
人工 κ	两名评分者独立打分，算一致性	判分是否可信
敏感性	逐参数扰动，输出影响排序	结论是否靠某个参数撑着

表 10 证据闭环的四件事

验收线：留出集不低于人类分析师基线；人与 LLM 判分一致性 $\kappa \geqslant 0.6$；敏感性报告中前三个参数合计解释力不超过 60%。达不到就不对外报价，只报区间。

九、语料改造

现有 30 个事件只有“触发 / 结局 / 处罚”。要支撑决策卡，必须补三个字段：

新字段	内容	用来回答
decision_taken	当时实际选择了哪条路（道歉 / 沉默 / 对抗）	同类事件里哪种应对更常见
decision_delay_h	拖了多久才决定	拖延的代价有多大
cost	复盘代价（罚没 + 市值损失 + 渠道损失）	不同应对的成本差

表 11 语料新增字段

有了这三个字段，才能说出“30 个同类事件里，6 小时内道歉的一组长尾中位数短 23 小时”这类话。这句话是定价依据；现在的语料只能说“可能的风险”。

同时补 30 个英文事件（英美欧），否则海外只能演示、不能报价。英文语料必须先定输出结构再采集，避免返工。

十、交付与运营

10.1 工作坊模式

工具要能“主持一场会”，而不只是“生成一份文件”：内置 90 分钟计时、4 个按时投放的注入、角色卡、决策日志录入、异议记录、结束后一键生成 readout。

10.2 导出与集成

- 导出 DOCX / PPTX / PDF 的会后材料，直接进入客户流程。
- 事件包与结论以 JSON 输出，便于接入工单、飞书、Jira 等系统。
- 保留 webhook：客户自有系统可在演练结束触发后续动作。

10.3 运维

项	当前状态
离线运行	verify / stats / rehearse / demo / forecast 全部无网络调用
可观测	Prometheus 指标 5 个，稳定路由标签，含 rehearsal 与 evidence 计数
部署	wheel 62.8 KB，非 root 容器，健康检查，systemd 单元模板
数据完整性	46 个文件 sha256 清单，sha256sum 兼容，篡改即退出 4
性能	单次推演 0.2 ms，stats 20 ms，CLI 冷启 66 ms

表 12 运维现状

十一、路线图与验收

每一步都必须能独立验收，避免“重构三个月、无法演示”。

阶段	范围	产出	验收线	工作量
P0	数值层与门禁	crh forecast、两档判分、样例行情	打不过随机游走则不输出点位（已达成）	已完成
P1	条件重估	事件 → 分布平移、影响链	两次运行逐字节一致；参数改动 1 秒内出对照	1 周
P2	证据闭环	三臂盲评、留出集、κ、敏感性	留出集 ≥ 人类基线；κ ≥ 0.6	2 周
P3	交付层	工作坊、决策日志、DOCX/PPTX 导出	一场 90 分钟演练全程单工具	2 周
P4	英文语料	30 个英文事件与英文输出	海外演示端到端可用英文跑通	1 周

表 13 路线图（P2 是估值开关，不是锦上添花）

十二、商业模式

12.1 三种卖法

形态	买家	单价	交付
演练工作坊	公关 / 危机公司、企业 CRO	3–5 万元 / 场	决策卡 + 会议包 + 会后 readout
年度授权	企业、集团风控	5–10 万元 / 年	离线部署 + 语料更新 + 支持
OEM 组件	集成商、舆情厂商	3–8 万元 / 年 或项目分成	演示件 + 交付件 + 白标

表 14 三种卖法与定价

12.2 为什么不能按份卖报告

报告是一次性交付，买家验证一次就结束，没有续费理由；工作坊与年度授权绑定的是流程与台账，每次事件都要重估，天然复购。**把 2000 元的报告改成 3 万元的会，是这次改造最大的单价跃迁。**

12.3 估值影响

口径	改造前	改造后（P2 完成）
作为代码资产买断	1.5–2 万元	8–15 万元
作为服务交付物	2–5 万元 / 单	3–5 万元 / 场，可复制
作为可授权产品	难以授权	5–10 万元 / 年

表 15 估值口径变化（P2 未完成前，代码资产仍按 demo 计价）

十三、风险与边界

13.1 合规

向第三方分发价格预测可能触及证券投资咨询牌照要求（境内），境外涉及 RIA / SEC 等注册制度。输出区间与情景分析不等于自动合规，销售前必须核对。数据方面：行情数据的使用需遵守来源条款，语料仅使用公开新闻，不引入客户数据。

13.2 方法论风险

- **过拟合。** 30 个事件太少，调参必须限定在训练集，测试集只测不调；否则 82.2% 会变成自欺。
- **判分器漂移。** LLM 判分不是人工校准，必须补 κ；否则所有下游数字继承判分器的偏差。
- **模型被当成预言机。** 输出必须始终附带局限页，任何删除局限页的定制需求都要拒绝或加价。

13.3 明确不做

- 不做“看起来很准”的点位预测器：随机挑参数、事后展示命中样本，三天能做出 demo，两周会失去第一个客户。
- 不追 star：品类全 GitHub 只有 12 个仓库，自然增长天花板是个位数。
- 不把物理数值交给语言模型：没有仿真器就不交付数字，只交付“不可知”的说明。

附录 A 已交付清单（0.4.0）

模块	内容
crh 命令	version / meta / config / stats / events / event / evidence / pack / rehearse / demo / bench / forecast / site / serve
HTTP 接口	healthz / readyz / metrics / meta / stats / events / events/{id} / seed / evidence / rehearse
引擎	offline（确定性规则）、llm（可选，默认拒绝运行）
数值层	区间、方向概率、样本外覆盖、发布门禁（退出码 3）
工程	113 项测试、wheel、Dockerfile、compose、Makefile、CI 矩阵 3.9–3.12
文档	README、ARCHITECTURE、API、OPERATIONS、ENGINE、NUMERIC、SECURITY、CHANGELOG

表 16 当前仓库交付面

附录 B 命令速查

```
crh demo --out demo              # 生成三份自包含档案（离线）
crh stats                        # 重算 82.2% 与置信区间
crh bench                        # 严格 / 宽松两档基线
crh forecast --csv data/market/nvda.csv
crh rehearse --event 1 --out out/
crh serve --port 8080            # API + Web UI + /metrics
```

式 5 常用命令

附录 C 指标口径

指标	口径	常见误读
82.2%	事件均值的分级命中率，30 事件 × 3 跑	不是二值命中率，不是预测准确率
95% CI	t 分布，df = 29	统计单位是事件，不是 90 次运行
跨跑一致 5/30	三次运行得分完全相同	不是“稳定”，是“极不稳定”
24.4% / 45.6%	关键词判分，与 82.2% 不可比	不能与 LLM 判分横向比较
3.50 美元	NVDA 次日收盘朴素预测的平均绝对误差	不是模型误差下限，是噪声量级

表 17 对外引用时必须带上的口径

— 完 —