豪泓科技
HAOHONG TECHNOLOGY

H2 AIDM · L2 智能体层 · 白皮书 04

H2 Harness
任务越长，越需要编排。

把不确定的大任务拆成受控的短闭环：分工 · 审查 · 单步重做 · 记忆与复用。

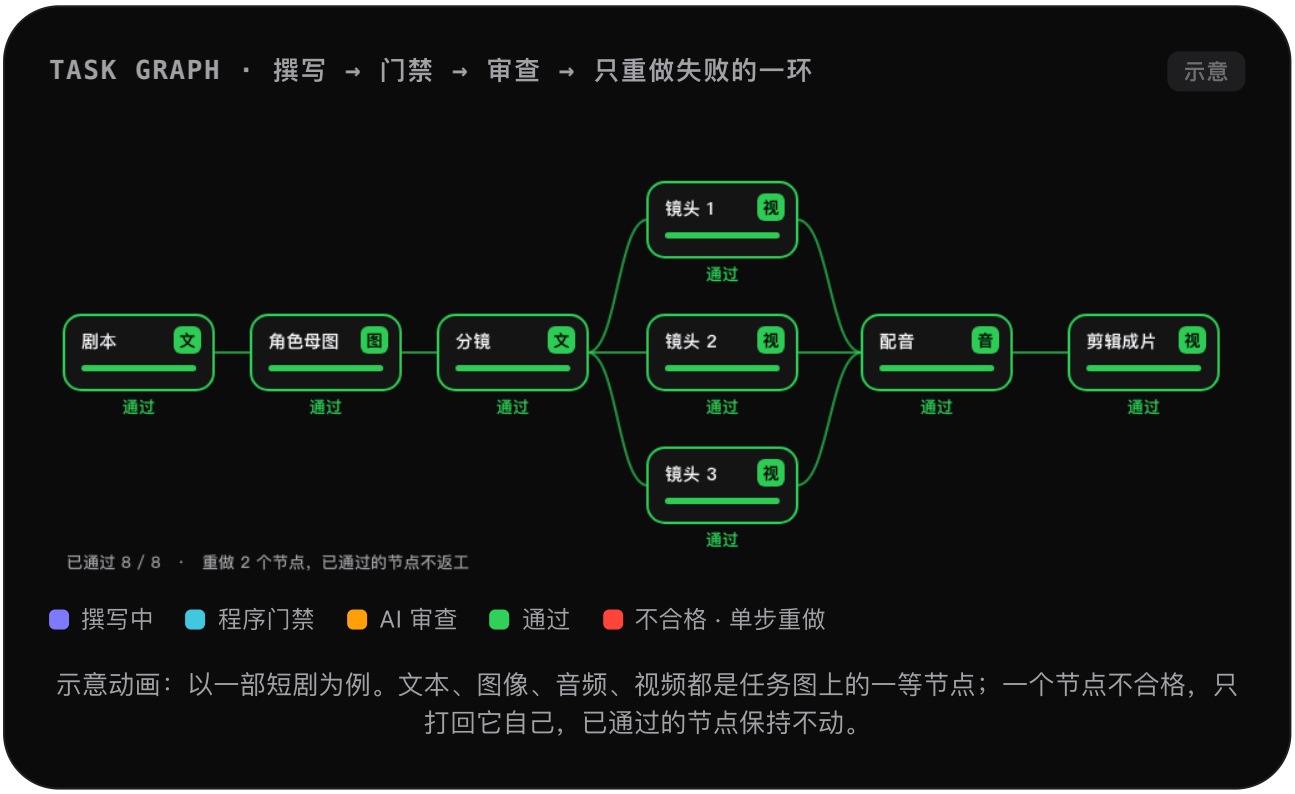
上海豪泓信息技术有限公司 | v5.1 | 2026 年 11 月

数字口径：实测 / 生产实证 / 已上线 / 规划测算，见各处标注。规划值以实际运行数据为准。

H2 Harness （简称 H2AC）

任务越长，越需要编排。

H2 Harness 不是套在模型外面的一层调用循环，而是一套驻留在服务端的智能体运行时架构：状态、记忆、编排、门禁、确认点和模型路由，都在服务端重新设计过。



示意动画：以一部短剧为例。文本、图像、音频、视频都是任务图上的一等节点；一个节点不合格，只打回它自己，已通过的节点保持不动。

100+

专职智能体，各有提示与验收标准。

已上线

100 分钟

小影马一键生全剧，最长无人值守运行时长。

已验证

1 天 / 部

上线一部量产剧。

已验证

7 个

已接入的 Agent 客户端。

已上线

一套架构，不是一层外壳

它不是传统意义上的 Harness。

常见的 Harness 是包在模型外面的一个循环：发提示词、调工具、出错重试，状态留在客户端。H2 Harness 把这件事做成了服务端的运行时，并在七个地方做了专门的工程优化。

环节	常见 Harness 的做法	H2 Harness 的做法	带来什么
运行位置	跑在客户端或单个进程里，状态跟着窗口走。	**服务端运行时：只追加的事件日志加跨进程锁，服务端是状态的唯一来源，终端只负责显示。**	关页、断网、换设备都能续接。
上下文装配	历史全量带入，或者到长度就直接截断。	**近期轮次保留原文，早期轮次做「三行压缩」（原话、结论、工具摘要），已加载的技能正文全局去重。**	模型拿到最小而充分的上下文，少花词元，结论不丢。
角色与预算	一个提示词既写正文又查格式。	**调度、撰写、审查、修订、归档各有最小上下文；调度与审查走短输出预算，撰写与修订走长输出预算。**	该快的环节快，该写长的环节写得完整。
质量校验	让模型自己检查自己。	**缺段、字段、格式、计数这类可数的检查交给程序门禁，不消耗词元；AI 审查只看语义，并给出差异清单。**	低级错误在进审查之前就被拦下。

环节	常见 Harness 的做法	H2 Harness 的做法	带来什么
高风险动作	没有确认点，超时重试可能重复执行、重复扣费。	**确认点状态机：服务端原子确认加预写日志，同一个确认在多台终端只生效一次。**	发布、计费、写入业务系统这类动作可控、可查。
多模态	把附件链接拼进文字里。	**任务图：文、图、音、视频同为节点，带上下游依赖，改了哪个节点只重跑受影响的部分。**	局部修改不用整条重来。
与模型层的配合	把模型当成黑盒接口来调。	**与 H2 Model Combo、H2 GPU Combo 一起设计：按「调模型、等工具、再调模型」的节奏做会话亲和，让读过的历史缓存下一步还能命中。**	长上下文首字约 1 秒 实测

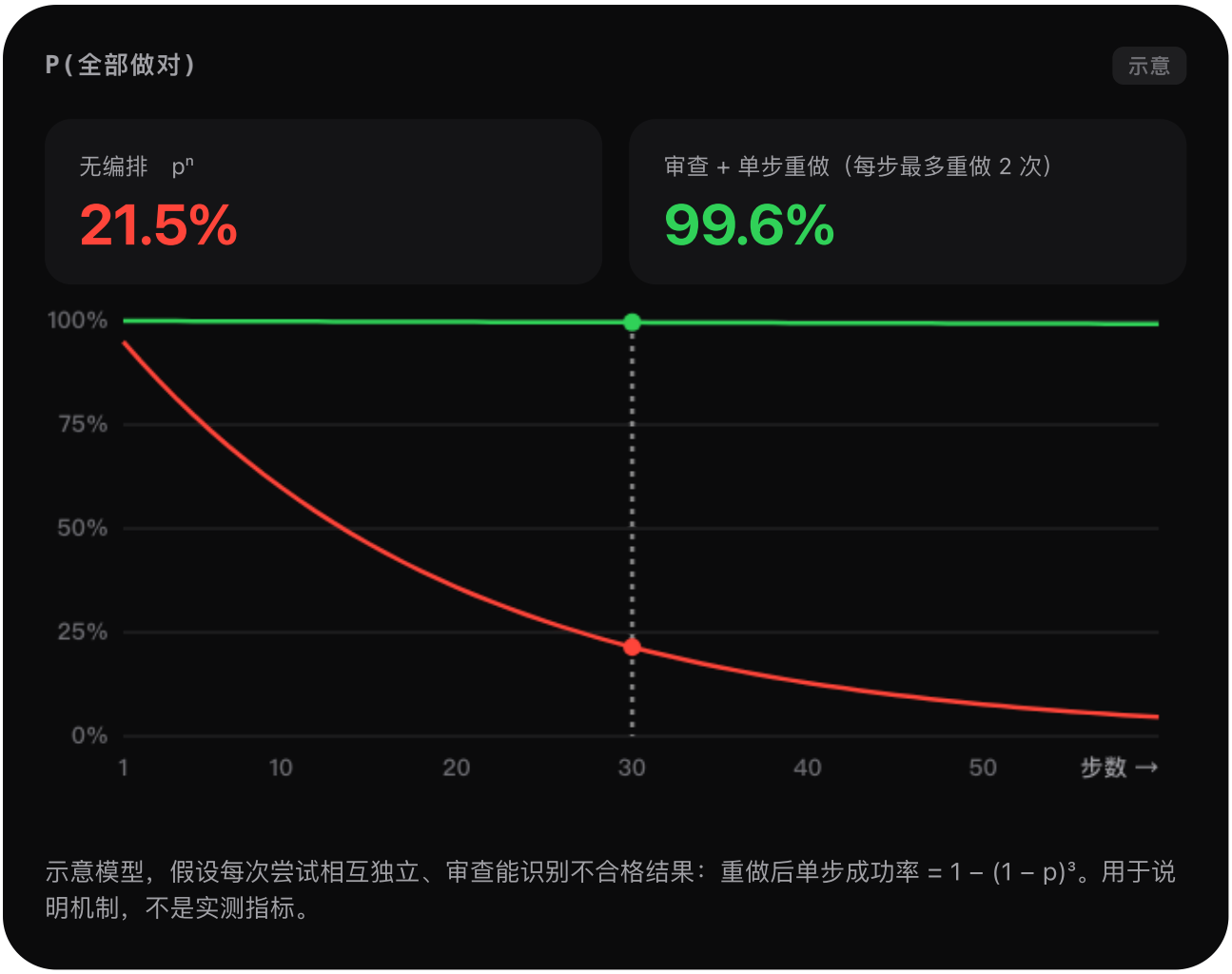
「常见 Harness 的做法」按常见形态归纳，仅作对照；各家产品能力以其官方说明为准。

问题 · 错误累积

每步 95% 靠谱，
三十步后只剩两成。

单步 95%，连续 30 步全对只有约 21%——近八成长任务会在中途出错（示意计算，假设每步独立）。

H2 Harness 在每一步后加门禁与审查，不合格只重做这一步。拖动滑块，看两条曲线如何分开。



示意模型，假设每次尝试相互独立、审查能识别不合格结果：重做后单步成功率 = 1 − (1 − p)³。用于说明机制，不是实测指标。

四个机制

分工 · 审查 · 单步重做 · 记忆与复用。

四个机制合在一起，让长任务不失忆、不跑偏、可审计。

01

分工

每个角色只拿自己的最小上下文、专属提示与验收标准，撰写角色拿长预算，审查角色拿短预算；不共用一个万能提示词。

02

审查

程序门禁先做确定性校验，审查角色再逐项打分；不合格退回本环，附差异清单重做。

03

单步重做

只重跑失败的那一环，已通过的环节保持不动，不整段返工。

04

记忆与复用

会话与工作空间记忆留在服务端；角色、母图、模板沉淀为档案与技能，下一个项目直接调用。

程序门禁 · 确定性校验

能用程序判断的，不交给模型。

字段缺失、格式、字数与时长、计费幂等等规则由程序直接校验，不消耗词元，结果可复现。

AI 审查 · 语义质检

程序判断不了的，交给专职审查角色。

剧情逻辑、视觉一致性、数据严谨性等语义问题，由审查角色逐项打分，输出可执行的修改意见。

多模态任务图 × 跨端确认点

多模态协同发生在任务图里，高风险动作停下来等人批。

任务图 · TASK GRAPH

文、图、音、视频，同为一等节点。

不赌某个底层模型天生懂所有模态。每个产物写回图上的节点，后续步骤按引用精确取用，不用人工来回搬素材。

确认点 · CONFIRMATION

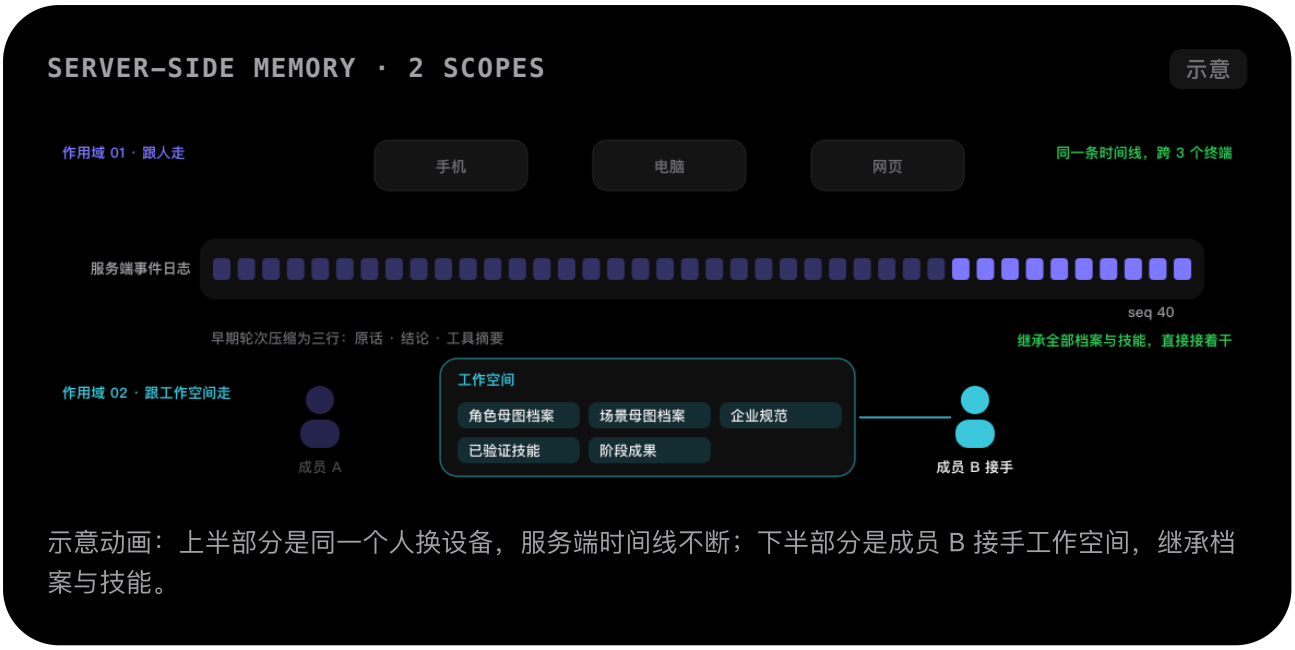
不可逆与计费动作，先停后批。

对外发布、大额算力计费、写入业务系统等节点，任务自动停在确认点；在手机或电脑任一端批复，结果幂等写回同一条时间线。

双作用域记忆

记忆跟人走，也跟工作空间走。

记忆在服务端，而不在浏览器标签页里。换设备能续接，换人能接手。



示意动画：上半部分是同一个人换设备，服务端时间线不断；下半部分是成员 B 接手工作空间，继承档案与技能。

作用域 01 · 跟人走（跨终端）

只追加的事件日志

会话在服务端落成只追加的事件日志，带单调序号与跨进程锁。最近几轮保留原文，早期轮次压缩为「原话 · 结论 · 工具摘要」三行。关掉浏览器、换电脑，进度与上下文照常续接。

作用域 02 · 跟工作空间走（跨成员）

结构化档案 × 团队技能

角色与场景母图、企业规范、阶段成果由专职角色写成结构化档案，存放在工作空间。团队换人接手，直接继承已验证的事实与方法。

模型能力槽

换模型，
记忆和流程都不动。

对上只暴露四类能力槽，每个槽配置主供应商与备用供应商，限流、超时或故障时自动切换。更换底层模型、调整画质档位、切到私有化节点，只改路由配置；记忆、角色、技能与确认点保持不动。

语言	图像	视频	音频
按角色预算分流	按产品档位映射	按产品档位映射	主备切换

三级技能库 × 三层契约

技能放方法 · 消息放事实 · 门禁做校验

技能分平台、团队、个人三级存放，先按元数据召回少量候选，再装配正文。换一个行业，只换一套技能包和门禁规则，同一个内核就能嵌入新的产品；时间线、记忆、编排、任务图、确认点与主备容灾全部复用。

嵌入式内核

一个内核，驱动两款应用。

用户看不到 H2 Harness，但小影马和 H2 Code 的每一个长任务都跑在它上面。

小影马 · 离线量产

从一句话到一部剧

编剧、母图审计、分镜导演、相机参数编排、配音剪辑等专职智能体协作；1 天上线一部量产剧，一键生全剧最长 100 分钟无人值守。

小影马直播 · 实时中枢

直播间的后台大脑

弹幕理解、知识库检索、场控导演与合规过滤并行运行，驱动 RayWorld Live 实时出流。

H2 Code · 通用长任务

能交成果的办公 Agent

行业研报、多表交叉分析、跨系统业务流程；每一步自动查错修订，高风险动作停在确认点。

在 H2 AIDM 中的位置

向下调用模型层的能力槽，向上嵌入两款应用。

↓ L1 模型 · 依靠

H2 Model Combo

长任务动辄几十轮、几十万词元上下文；有状态缓存与会话亲和让长上下文首字约 1 秒（实测），多角色编排才跑得快、用得起。

了解 H2MC ›

↓ L1 世界模型 · 调用

RayWorld

向 R1 下发结构化的相机位姿参数（位置 xyz + 朝向 yaw / pitch / roll），向 Live 推送直播话术与互动分支——智能体可以编程控制的镜头。

了解 RayWorld ›

↑ L3 应用 · 嵌入

小影马与小影马直播

短剧流水线、一键生全剧与直播实时中枢的编排内核。

了解小影马 ›

↑ L3 应用 · 嵌入

H2 Code

长任务不中途走偏、换设备不断片、多模态一站交付，靠的就是这套内核。

了解 H2 Code ›