豪泓科技
HAOHONG TECHNOLOGY

H2 AIDM · L3 应用层 · 白皮书 06

H2 Code
长任务也不半路走偏的桌面 Agent。

写材料、查资料、做表。嵌入 H2 Harness 编排内核，长任务首字约 1 秒，可私有化部署、换成企业自己的品牌。

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。

L3 · 应用层

H2 Code

长任务也不半路走偏的桌面 Agent。

写材料、查资料、做表。H2 Code 是 L3 应用，建在下面三层之上：内嵌 H2 Harness，直连 H2 Model Combo。

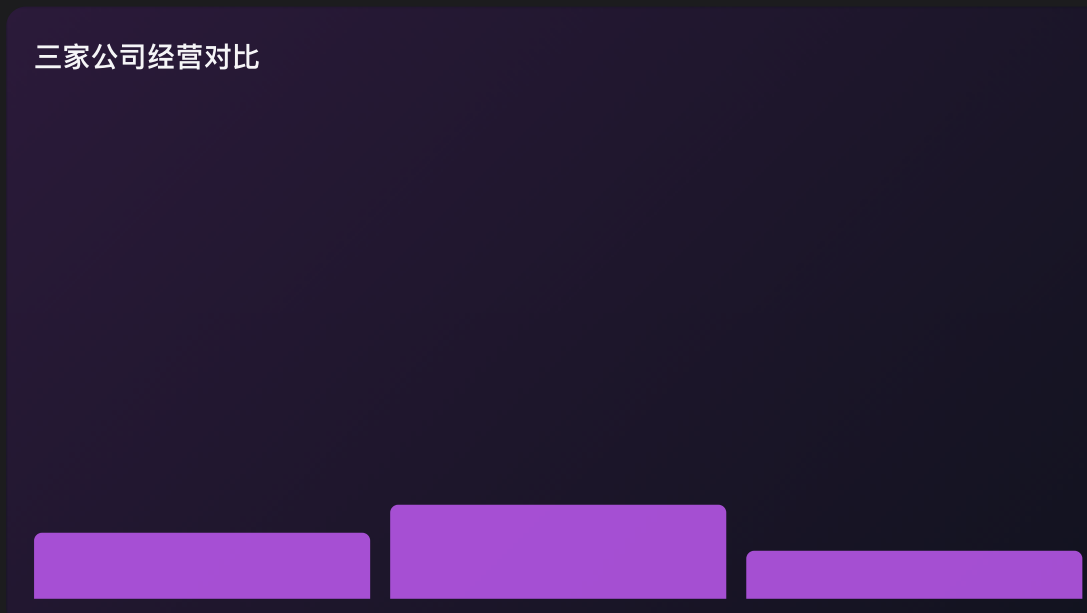
H2 Code

整理三份财报，做一张对比表，再出一页 PPT 发给财务总监。

- 读取三份财报 PDF
- 抽取营收、毛利率、现金流
- 交叉核对口径与单位
- 生成对比表
- 审查：逐项核对数字
- 生成一页 PPT
- 确认点：对外发送前请批准
- 发送给财务总监

公司	营收（亿元）	毛利率	经营现金流
A	—	—	—
B	—	—	—
C	—	—	—

三家公司经营对比



已发送 · 全过程可回放

示意 任务被拆成多步，每一步有门禁与审查；凡是对外发送、计费或写入业务系统的动作，都会停在确认点，等人批准。

约 1 秒

长上下文智能体任务的首字时间（H200 实测）。

实测

200–300

H200 单路实测出字速度（词元/秒），优化前约 90。

实测

约 100 万

上下文词元上限，与官方同档。

与官方同档

2 台起

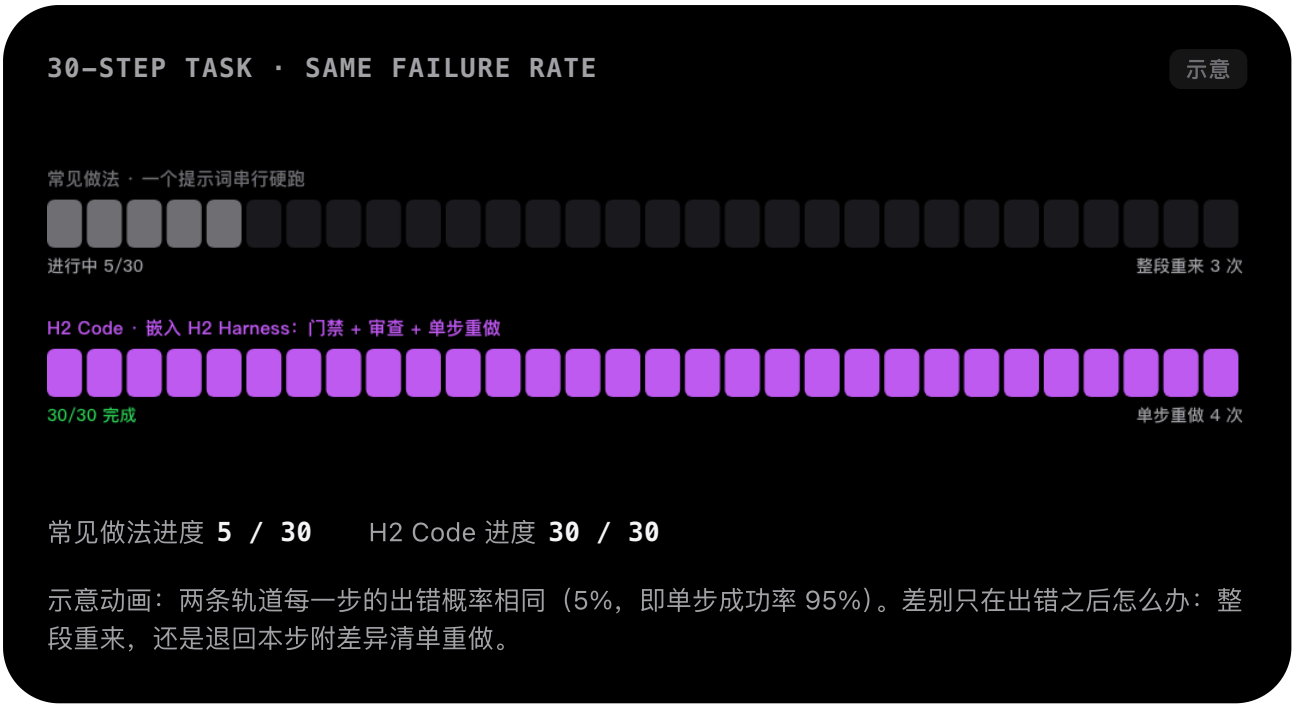
私有化部署的平台只需 2 台普通服务器，数据不出域。

已上线

长任务 · 嵌入 H2 Harness

出错时，只重做那一步。

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 Code 把任务拆成短闭环，不合格只退回那一步。



长上下文 · 直连 H2 Model Combo

读过的，不再重读。

不做缓存时，智能体每走一步都要把读过的历史重算一遍，上下文越长等得越久。提示缓存已是通行做法，难的是在自有算力、高并发下让它稳定命中。

H2 Code 的会话由 L0 调度固定在持有其缓存的通道上（会话亲和），下一步只算新增内容：首字约 1 秒。



五个工程差异

差别不在界面，
在下面有没有整套工程。

维度	常见做法	H2 Code 的做法
长任务稳定性	一个提示词串行执行，步数一多容易偏题、遗漏，出错往往整段重来。	**内嵌 H2 Harness：角色分工、程序门禁、AI 审查，只重做不合格的那一步。**
长上下文响应	不做缓存时，每一步都要重算全部历史，上下文越长等得越久。	**直连 H2 Model Combo：有状态缓存 + 会话亲和（通道锁定），首字约 1 秒；单路出字 200–300 词元/秒** 实测
记忆与协作	状态留在浏览器或本机，换设备就断片，同事难以接手。	**服务端双作用域记忆：个人进度跟人走、跨设备续接；项目档案与技能跟工作空间走、团队可继承。**
多模态交付	以文字为主，图表、文档、图片要人工来回搬运。	**统一多模态任务图：文本、表格、PPT、图片、音频乃至视频都是任务节点，直接交付成品。**

维度	常见做法	H2 Code 的做法
成本与数据	依赖公有云按量调用，长任务账单高；业务数据需要上传第三方。	**按量为官方价 3 折；平台 2 台普通服务器起即可私有化，数据不出域；高风险动作设确认点。**

场景

不只是聊天，是直接交成果。

场景 01

深度研究与材料撰写

检索公开资料与企业知识库，交叉比对多份 PDF、财报、合同，经「撰写 → 核验 → 格式门禁」后输出报告、研报或宣讲 PPT。

场景 02

数据分析与制表

读取 Excel、CSV 与业务数据库，自动清洗、核算、出图，生成经营复盘；数字由程序计算并逐项核对，而不是让模型心算。

场景 03

跨系统业务流程

通过技能包对接 ERP、CRM、OA，完成审批预处理、合规稽核、客户跟进与定时汇总；写入业务系统前停在确认点，手机或电脑一键批准。

交付方式

开箱即用，也能变成企业自己的产品。

个人与团队

按席位月费

开箱即用，团队共享工作空间档案与技能。

企业与机构

白标 + 私有化 + 定制技能包

换成企业自己的名字、图标与价目；部署在客户机房，平台 2 台普通服务器起；按企业工作流开发专属技能包与系统对接。

在 H2 AIDM 中的位置

它的能力，
来自下面三层。

↓ L2 智能体 · 内核

H2 Harness

长任务编排、服务端双作用域记忆、撰写 / 审查 / 门禁闭环与确认点，都来自嵌入的 H2 Harness。

了解 H2 Harness ›

↓ L1 模型 · 推理

H2 Model Combo

首字约 1 秒、单路出字 200–300 词元/秒、约 100 万词元上下文，以及语言、图像、视频、音频四类能力槽。

了解 H2MC ›

↓ L0 算力 · 底座

H2 GPU Combo

用客户的存量显卡或数据中心节点做私有化部署，按卡定制方案，让企业用得起。

了解 H2GC ›

↔ L3 应用 · 同层

小影马

同一套 H2 Harness 与能力槽：小影马把短剧工作流交给智能体，H2 Code 把通用知识工作交给智能体。两边的真实负载都回流优化下面三层。

了解小影马 ›