

AI 产品经理 作品集

从真实问题出发，把 AI 做成可用、可测、可上线的产品。

邵志成 AI 产品经理 · AGENT PRODUCT

01

全球商业洞察 Agent

P.03—15 · 企业出海决策 · 研究与证据

02

知香

P.16—27 · 垂直香水导购 · 理解与推荐

03

AI Eval × Human-in-the-loop

P.28—39 · 模型质量验证 · 数据生产

我关注的不是模型能做什么， 而是它怎样进入真实决策。

从用户研究和问题定义出发，把模糊需求拆成可验证任务；再用 Agent 架构、知识工程和评测体系，让模型输出可控、可解释、可落地。



邵志成 | AI 产品经理
需求洞察 · AGENT 架构 · 评测迭代 · 工程落地

50+

消费者与企业负责人访谈

3

完整推进的 AI 产品项目

3 套

面向真实任务建立的评测体系

全链路

产品、架构、评测与部署独立推进

需求研究

田野调研、用户访谈、痛点与指标转译

AI 产品架构

Harness、多 Agent、RAG、知识图谱与工具调用

评测迭代

硬门禁、LLM-as-Judge、回归集与 bad case

工程落地

前后端、数据库、部署、监控与可靠性

全球商业洞察Agent

我从印度尼西亚田野调研中发现，企业做海外市场研究最难的不是搜不到资料，而是很难判断哪些信息真正属于目标国家、哪些结论值得信。基于这个问题，我独立做了一款面向中国企业出海负责人的商业洞察 Agent：优先检索本地来源，把证据、判断和后续追问放在同一条研究链路里。

邵志成 | AI 产品经理 产品定义 · Agent 架构 · 评测迭代 · 工程落地

先确认“信息从哪来”，再讨论“结论是什么”。

30+

企业出海负责人访谈
直接来自一线问题

2,367

已分层来源域名
覆盖 195 个主权国家

3.33/4

LLM-as-Judge 七维均分
达到 83.3%

100%

10 题代码检查通过
报告与引用均完整

印尼田野考察与访谈现场

访谈 / 商业活动 / 合规交流 / 一手报道



30+

企业出海负责人访谈，提炼真实研究需求

01 · USER PROBLEM

资料很多， 但真正能拿来判断的很少

企业进入一个新市场，通常要同时查政策、税务、竞争、渠道和消费者。问题是这些资料散在政府网站、本地媒体、行业机构和不同语言里；通用大模型能很快写出一份像样的报告，却可能把邻国案例、区域新闻或二手转述混进来。

“企业真正需要的不是更长的报告，而是知道信息来自哪里、结论为什么成立，并且能沿着同一份研究继续追问。”

— 基于印尼田野考察与 30+ 企业出海负责人访谈形成的问题定义

跨国资料误用

邻国案例、区域报道被误当成目标国事实。

本地信息不足

国际媒体容易搜到，真正关键的官方和本地行业资料反而靠后。

结论脱离证据

结论写得很顺，但数字、政策和建议找不到对应出处。

引用不可追踪

引用编号、链接和正文对不上，用户没法快速核实。

研究无法复用

每次追问都像重新研究一遍，之前的报告和证据接不上。

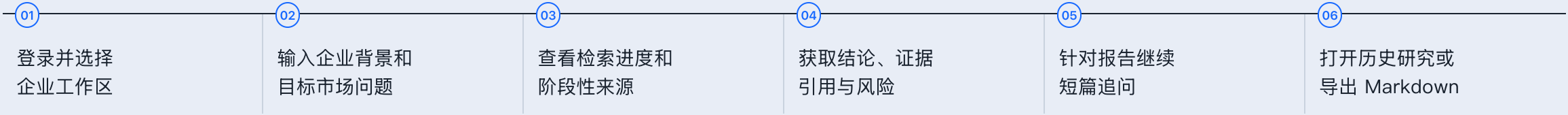
先把“查什么、信什么”管住，再让模型写报告

所以我没有继续堆 Prompt 或换更强的模型，而是先把研究中最容易出错的地方拆出来：目标国家是否正确、来源值不值得信、证据能不能支撑结论、引用能不能对得上正文。

PRODUCT THESIS

模型负责判断
代码守住底线

Agent 可以自己决定从哪些角度研究、报告怎么组织；但目标国家、搜索次数、证据上限、引用编号和失败后的修复方式，都由程序明确控制。



02 · DECIDE METHOD

把研究流程 变成决策流程

DECIDE 是这个 Agent 的战略内核：先定义要做的决策，再解释行业机制、识别约束、比较选项，设计资源与行动，并根据结果持续演进。

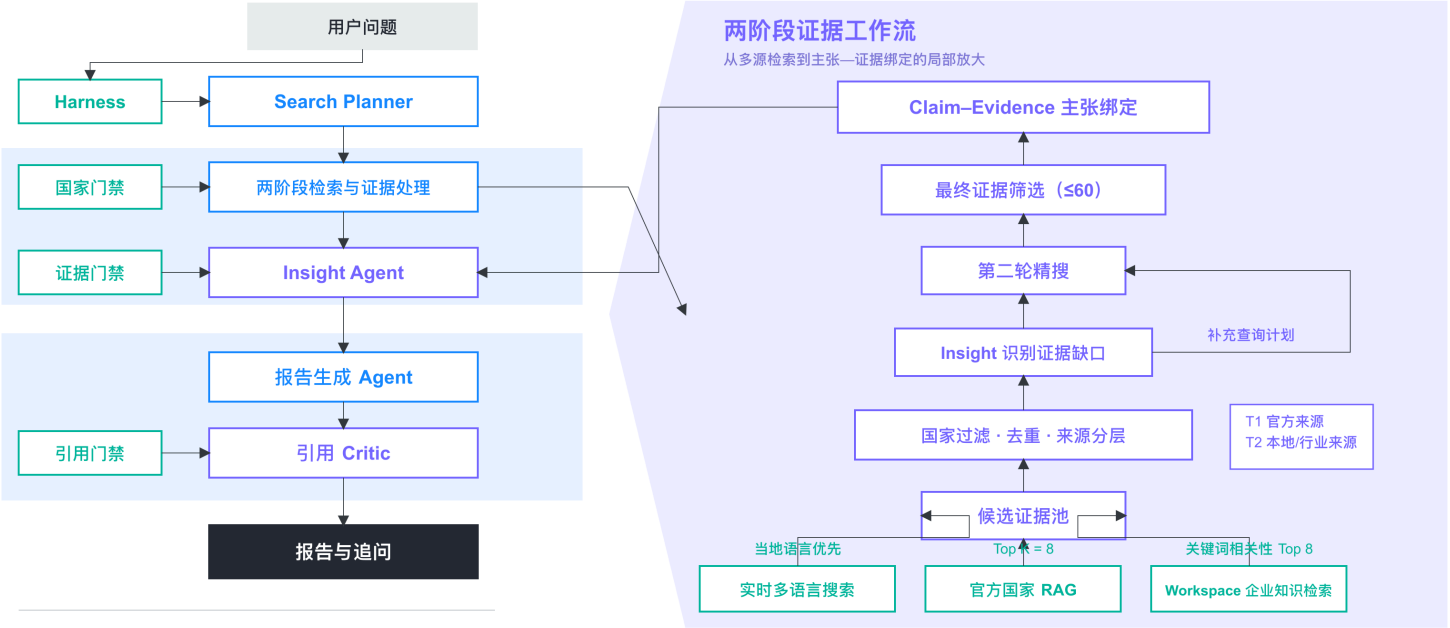


把一次复杂研究拆开， 让每一步只做自己该做的事

Harness 负责流程和规则，不替模型做判断；不同 Agent 只拿完成当前任务需要的信息。检索、分析、写作和检查分开处理，既减少重复读取，也避免让写报告的模型自己给自己“验收”。

出海商业洞察 Agent 架构

Harness 控制研究边界，Multi-Agent 完成检索、洞察、写作与引用审计



运行底座：Vercel 前端/API · Render 异步 Worker · Supabase 会话/报告/监控

A 先锁定目标国家

先识别目标国家及多语言别名；其他国家资料直接剔除，区域性资料只作为背景，不拿来支撑目标国事实。

B 先广搜，再补缺口

第一轮同时查实时网页、官方 RAG 和企业知识；分析后如果发现关键问题没证据，再定向补搜，最后只保留不超过 60 条可用证据。

C 每个关键判断都要有出处

重要判断、具体数字和政策要求必须能找到直接来源；证据不够，就删掉或把措辞降一级。

D Critic 只做检查，不重写报告

Critic 只检查证据是否支持、编号是否存在、数字有没有出处、国家是否匹配；发现问题只能删除、弱化或重新绑定已有引用。

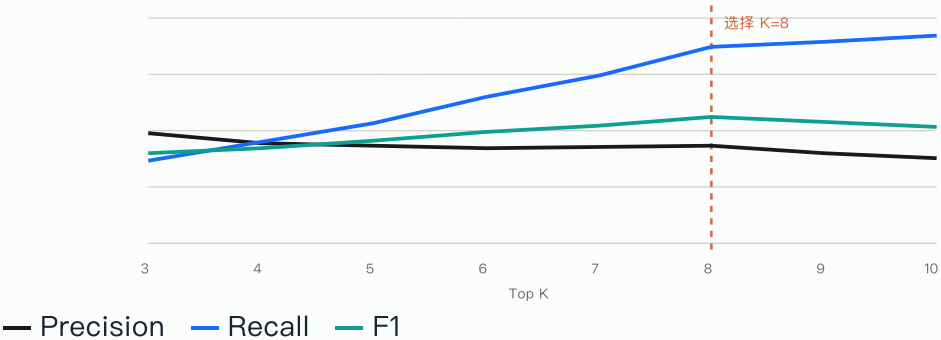
04 · EVALUATION

每次改动，
都回到同一套题上重新验证

我把评测分成三层：RAG 离线测试、代码检查和独立模型评分。每个版本都保留原始结果，只有在同一套题上确实变好的改动才保留，跑差了就回退。

RAG Top K 参数扫描

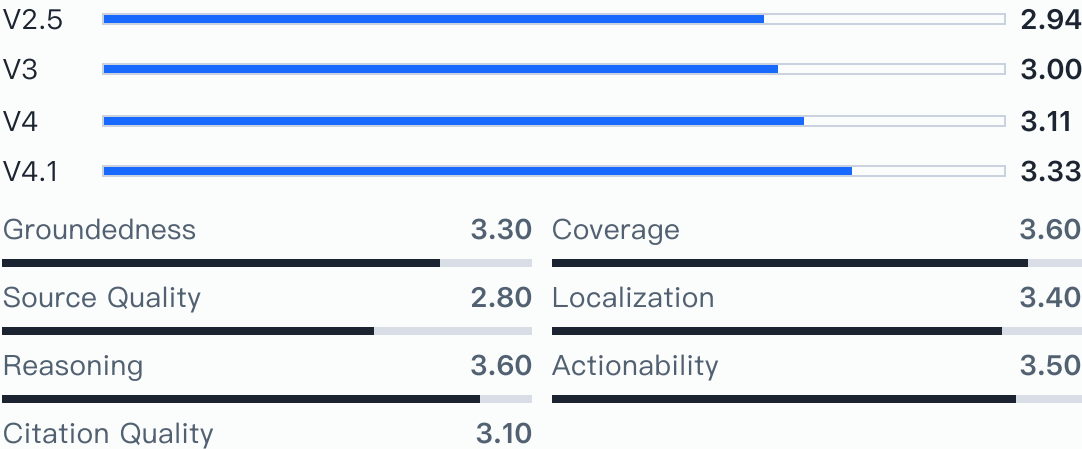
13 个国家问题 · 124 条人工标注 Chunk



选择 K=8: Recall 0.871、F1 0.559；继续提升 K 会增加噪声并降低 Precision。

Agent 版本迭代

同一 10 题回归集 · 七维 1-4 分



2.94 → 3.33 (+13.3%) : Actionability +25%, Citation Quality +19.2%；同时保留 V4/V4.1 运行波动。

Layer 01 · 代码检查

检查报告是否完整、数字有没有引用、编号是否存在、国家有没有串、证据是否超出范围；V4.1 达到 10/10。

Layer 02 · 独立模型评分

用独立模型从 Groundedness、Coverage、Source Quality 等七个维度评价最终报告。

Layer 03 · 换题与复跑

固定 10 题看有没有退步，再用 5 个新问题看能不能迁移；同一版本重复运行，单独观察模型波动。

把能跑的 Agent， 做成别人真的可以用的产品



出海商业洞察

基于证据的智能洞察，助你做出更好的决策。

登录

注册

用户名

4-32 个字符

密码

.....

确认密码

再次输入密码

创建账号

注册后将自动创建你的独立工作区。

Web 端注册与独立工作区

企业知识库采用混合 RAG：以 sentence-transformers 加载 BAAI/bge-small-zh-v1.5 生成 Embedding，由 FAISS 完成语义召回，同时保留关键词召回；两路候选合并、去重与重排后，交给 Agent 组织为可引用证据。产品侧补齐登录、工作区、历史报告、企业资料 and 异步任务。

Global Market Intelligence

企业知识库

返回研究

添加企业资料

可以粘贴文本，也可以上传附件；附件会自动提取文字后参与研究检索。

文档名称

添加附件

PDF、DOCX、TXT、MD、CSV、JSON，单个不超过 4.0 MB

选择文件

或粘贴企业报告、产品资料或内部知识.....

加入知识库

当前工作区资料

暂无企业资料。

企业知识库与工作区资料

从研究对话到可追溯的战略判断

用户先确认研究方向，再查看核心结论、进入判定与引用证据；三张产品页面展示同一条研究链路如何从问题走向行动。

出海商业洞察

wergit 的工作区 安装到桌面 中 EN 退出登录

出海商业洞察

面向商业问题的 AI 驱动研究。

你

我们做新能源汽车，想进入印尼，要怎么做

研究助手

建议先按以下方向开展研究：
决策问题：是否进入印度尼西亚新能源汽车市场，以及以何种模式（整车出口、本地组装、合资/合作或技术授权）进入。
研究目标：在控制法规、投资和竞争风险的前提下，以可执行的进入模式和资源投入，在印度尼西亚新能源汽车市场建立可持续的初步市场份额。
时间范围：3-5年
判断标准：市场规模与增长潜力、本地化政策与合规要求、竞争强度与现有玩家格局、投资回报与进入成本、供应链与充电基础设施成熟度
已知约束：需符合印尼新能源汽车本地化率及税收优惠政策要求、需考虑印尼本地合作伙伴或本地制造能力、初始阶段资源有限，需优先选择低资产或逐步加码的进入路径
重点验证：印尼现行新能源汽车税收优惠、本地化率门槛及进口限制的具体条款、印尼新能源汽车市场现有竞争格局及中国品牌、日系品牌布局、本地充电基础设施覆盖及电网条件对纯电车型可行性的影响、印尼本地供应链能力及动力电池回收/环保要求的落地情况、吉利、比亚迪、五菱、丰田等品牌在印尼的价格带、渠道和消费者认知
检索范围：market-size、policy、competition、risks、investment、supply-chain、adoption、chinese-players
确认后我会开始检索、验证证据并生成报告；你也可以直接补充或调整研究重点。

你

更精确一点吧，我们在印尼已经建了汽车厂

研究助手

现有印尼工厂主要具备生产哪类车型的能力，且希望优先进入哪个细分市场：乘用车、商用车，还是两者兼具？

已建成工厂决定可落地车型范围，细分市场选择会直接改变需求测算、竞争对标、认证路径和渠道设计；若不区分，研究范围可能覆盖过泛或偏离现有制造条件。

你

乘用车

研究助手

建议先按以下方向开展研究：
决策问题：如何基于印尼已建成的乘用车整车工厂，制定进入印尼乘用车市场的量产、销售与运营策略
研究目标：在印尼已具备本地乘用车生产能力的条件下，明确最具可行性和竞争力的市场进入与扩张路径，以实现可持续

研究方向 · 先确认决策问题

出海商业洞察

wergit 的工作区 安装到桌面 中 EN 退出登录

你

确认并开始研究

核心结论

印尼已建厂的乘用车企业应把优先客群锁定为爪哇城市及城郊的刚需家庭用户，重点切入7座家庭电动车或中端SUV，以40% TKDN合规为起点快速量产一款价格清晰的车型，用销量和渠道覆盖建立认知，再分阶段把本地化率提升到60%以上 1 2 142 。主要理由是：中国品牌已占印尼电动车市场约93%，市场教育成本已大幅降低，真正的竞争窗口在于2026-2027年CBU进口优惠终止后的本地量产替代期，而已建工厂恰好具备结构性先发条件 37 63 90 。关键条件是产品定价不能依赖尚不确定的2026年补贴；更稳健的路径是以不依赖补贴的定价模型完成规划，把政策激励视为增量收益，同时平行推进供应链本地化和锂电池兼容性评估，以对冲激励政策向本国品牌或锂电池倾斜的风险 10 53 147 。

优先客群应锁定爪哇刚需家庭用户

优先客群是爪哇岛城市及城郊的刚需家庭用户，核心购买场景是日常代步和多成员出行，而非追求高阶智能化配置。这一特征意味着已建厂的乘用车企业不应照搬国内爆款车型逻辑，而应针对家庭出行场景重构产品定义 71 109 。家庭用户的价格敏感度显著高于中国市场，因此首批车型的价格带需要贴近主流走量区间，而不是先推高配旗舰。企业应优先验证7座家庭电动车或中端SUV是否存在未被充分覆盖的价格带，因为这类车型同时满足多人出行和实用续航两大刚需 26 。在渠道上，应优先布局雅加达、泗水、万隆等爪哇核心城市及其城郊，以降低初期补能和服务网络的覆盖难度。

ref 71 109 25

本地量产替代进口是当前最确定的竞争窗口

整装进口CBU电动车的进口优惠于2025年12月31日终止，2026年1月起进口CBU电动车必须配以等量的本地生产承诺 1 63 。政策上，竞争焦点已从进口关税套利转向本地规模化生产。对已建厂的企业，这是结构性利好：依赖进口的对手将面临成本上升和市场准入受限，而自身若能快速合规量产，可以在政策收紧期抢占先发位置 92 。企业需要以最短路径把已建工厂转换为可持续销售的合规产能，而不是继续讨论是否进入印尼。

ref 1 63 92

TKDN已从激励条件升格为市场准入硬门槛

印尼本地化率政策已从可选激励升级为销售准入前提。按总统令规定，本地生产的电动汽车2026年需达到40% TKDN，2027-2029年升至60%，2030年起达到80% 2 142 143 。这意味着已建工厂不能只做CKD拼装，而要建立实质性的本地零部件采购体系。机制上，合规压力会从整车装配环节向

报告详情 · 结论与行动

出海商业洞察

wergit 的工作区 安装到桌面 中 EN 退出登录

需要触发调整或退出的观察信号

三个信号决定策略是否继续按当前路径推进。第一，2026年补贴方案若明确优先支持本国品牌或锂电池车型，而企业车型仍以LFP为主，则需要评估是否切换电池路线或调整本地品牌合作结构 10 147 。第二，若40% TKDN合规车型的实际量产成本显著超过目标价格带，导致无法在不依赖补贴的情况下竞争，则应缩小首发市场范围，优先服务爪哇核心城市的高意向客群 142 。第三，若爪哇以外充电基础设施扩展持续缓慢，且日系混动车在家庭用户中加速渗透，则应下调中长期纯电销量预期，考虑是否以本地生产的混动或插混产品承接过渡需求 95 145 。这些信号出现并不代表立即退出，但要求重新定义车型组合和渠道密度，而不是保持原计划不变。

ref 10 147 142 95 145

引用来源

1] Pemerintah Tagih Janji Produsen Mobil untuk Penuhi Wajib TKDN 40% - GAIKINDO
gaikindo.or.id 官方

2] Lawmakers: Chinese Electric Car Manufacturers Must Meet Local Component Requirements - GAIKINDO
gaikindo.or.id 官方

4] Peraturan TKDN Kendaraan Listrik Dilonggarkan, Kesempatan Industri Komponen Berkembang - GAIKINDO
gaikindo.or.id 官方

10] Menperin sebut insentif EV akan diprioritaskan mobil merek nasional - ANTARA News
antaranews.com 本地媒体 Thu, 30 Jul 2026 08:14:33 GMT

12] Wajib TKDN 60 Persen untuk Dapatkan Insentif Kendaraan Listrik 2027
otomotif.kompas.com 本地媒体 Fri, 31 Jul 2026 00:42:00 GMT

25] Peluang Mobil Listrik 7-Penumpang di Pasar Indonesia
otomotif.kompas.com 本地媒体

33] 10 Merek Mobil Listrik Tertaris di Indonesia Januari-September 2025
databoks.katadata.co.id 本地媒体

37] Merek China Kuasai Pasar Mobil Listrik Indonesia Kuartal I 2025
databoks.katadata.co.id 本地媒体

引用详情 · 来源可追溯

一次研究做完之后，用户还要能继续用下去

所以我把首次研究、连续追问、企业资料、历史报告和任务监控接在一起，而不是让用户每次都从一个空白输入框重新开始。

01 首次研究和追问分开处理
第一次研究输出核心结论和 3—7 个章节；后续追问只回答当前问题，用 2—3 段给出判断和引用，不重新生成整篇报告。

02 每家企业有自己的工作区
账号、会话和企业资料按 Workspace 隔离；检索时只调取与当前国家、行业和问题真正相关的企业文档。

03 之前做过的研究可以直接接着用
保存问题、回答、完整结果、来源和时间；点击历史记录时恢复当时那一份报告，而不是错误地显示最近一次结果。

04 知道每次任务花了多少资源
每个 Task 都记录 Token、耗时、RAG 命中、Tavily 调用、质量状态和估算成本，同时限制单账号任务数和并发，避免资源失控。

体验层 Next.js / PWA 研究界面、历史会话、工作区、管理看板	执行层 Render Worker 异步任务、SSE 进度、超时恢复与状态回传	数据层 Supabase 账号、Workspace、报告、企业资料 和指标	智能层 LLM + Tavily + FAISS 规划、检索、洞察、写作、审计和官方 RAG
---	--	---	---

我做的不只是 Prompt， 而是把整个产品从 0 做到上线

从 0 到 1：
找问题 → 做方案 → 搭系统 →
跑评测 → 上线

我最关心的不是用了哪个框架，而是企业到底愿不愿意拿这份报告做判断：信息是不是目标国的、来源能不能核实、追问能不能接着上一次研究继续。

这个项目从需求、产品方案、Agent 架构到评测和部署都由我主导。关键决策不是凭感觉定，而是从访谈、失败案例和每轮评测结果里一步步改出来。

01

在印尼开展田野考察并访谈 30+ 企业出海负责人，把“海外研究不好做”拆成具体需求和失败场景。

02

确定“先证据、后生成”的产品原则，设计 Harness + Multi-Agent workflows 及各模块的输入输出。

03

整理 2,367 个 T1/T2 分层来源，接入官方 RAG 与企业知识，并建立国家级检索和过滤规则。

04

搭建人工 Chunk 标注、代码检查和 LLM-as-Judge 评测，保留每个版本的冻结结果做回归对比。

05

根据 Bad Case 定位跨国误用、重复内容和引用错配，逐轮修改；效果变差时直接回退。

06

完成 Vercel、Render、Supabase 全链路部署，并补齐登录、工作区、管理看板和 PWA。

每次迭代只问一件事： 这份报告是不是更可靠了？

固定 10 题回归集，再配合代码检查和 LLM-as-Judge 七维评分；
每轮改动都要回到同一批问题上验证，跑差了就不保留。

固定 10 题回归集 代码硬门禁 LLM-as-Judge 七维 Bad Case / 泛化集

V1 · 问题发现 先把“哪里错了”找出来 用户问题 会混进别国资料、引用对不上，用户追问一次就重新生成整篇报告。 产品动作 建立固定题集，并用代码检查报告完整性、目标国家和引用是否正确。 早期只有约 3/5 通过；评测没跑完整时，不把结果当成绩。	V2 → V2.5 · 证据边界 让每个结论都找得到出处 用户问题 来源质量不稳定，数字单位偶尔读错，不同章节还会重复；用户很难核实关键判断。 产品动作 把证据分成 direct / contextual；追问改成短回答；Critic 只检查引用、国家和“这条证据到底支不支持这句话”。 V2 硬门禁 8/10；V2.5 得分 2.94 / 4。	V3 · 结构与检索 减少重复搜索和无效上下文 用户问题 Summary Agent 会重复读取同一批材料，按章节各自检索也带来不少重复和噪声。 产品动作 合并 coreConclusion，取消 Summary Agent；先广搜再按证据缺口补搜，同时把完整报告和后续追问分开处理。 3.00 / 4；重复内容明显减少，但来源质量仍然是主要短板。	V4 → V4.1 · 质量与稳定性 证据够用就停，稳定性单独看 用户问题 候选证据太多会把噪声带进报告；同一个版本重复运行，结果也会有波动。 产品动作 两阶段筛选证据，最终最多保留 60 条 Evidence Card；同一版本重复运行，单独记录分数波动。 10/10 硬门禁；3.11 → 3.33 / 4。
--	---	--	--

2.94 → 3.00 → 3.11 → 3.33 V2.5 → V3 → V4 → V4.1；Actionability 2.80 → 3.50、Citation Quality 2.60 → 3.00。V4.1 与 V4 的差异还包含同版本运行波动，因此不把所有分数上涨都算作产品改进。

复盘：模型会写，不等于产品就可信

做完这个项目后，我最确定的三件事是：重要规则不能只写在 Prompt 里；没有固定评测，就不知道一次修改到底有没有变好；模型运行会波动，产品要把这种不确定性显式记录下来。

LEARNING 01

重要规则不能只写在 Prompt 里

Prompt 可以提醒模型“只看印尼”，但不能保证它每次都做到。目标国过滤、证据类型和引用编号这类不能出错的规则，必须由代码兜底。

LEARNING 02

没有固定评测，就不知道到底有没有变好

固定回归集、泛化题、Bad Case 和版本快照让我能回答一个很实际的问题：这次修改是真的改善了结果，还是只是刚好这一次跑得更好。

NEXT STEP

下一步先补本地一手来源

下一阶段会优先增加目标国政府、监管机构和本地行业来源，同时减少无效 Evidence，并通过更多重复运行测试搜索排序和证据筛选是否稳定。

— GLOBAL MARKET INTELLIGENCE

让企业知道： 这份判断为什么值得信。

从一线访谈、研究流程到上线评测，页面中的数据均来自项目实际运行与冻结结果。邵志成 | AI 产品经理。

30+ 访谈	195 个国家	2,367 个来源域名	3.33 / 4	100% 硬门禁
--------	---------	-------------	----------	----------

查看滚动版作品集 ↗

查看高清架构图

知香

把“说不清的感觉”，变成能买到的香水。

我从用户怎么描述气味开始，搭起意图理解、香水知识检索、真实商品推荐和评测迭代，并把产品部署上线。

邵志成 | AI产品经理 PRODUCT · AGENT · EVALUATION

知识图谱帮我理解
“想要什么味道”，
真实商品源告诉
我“现在能买到
什么”。



01 · USER PROBLEM

用户知道自己想要什么感觉，却不知道该怎么搜什么。

在电商里买香水，搜索框更擅长接收品牌、款名和香调；但新手通常只会说“像雨后森林”“通勤别太甜”“送妈妈，预算 400”。

20+ 电商消费者访谈
提炼三个核心痛点

不会说术语
用户会描述感觉，却不知道“干净、清冷、温柔”分别该对应什么香调。

线上无法试香
闻不到实物时，只看商品标题和卖点，很难判断自己会不会喜欢。

推荐缺少解释
推荐常只有“猜你喜欢”，用户看不出它和自己的需求到底哪里匹配。



什么香适合你，和什么香买得到，是两回事。

我先把这两个问题拆开：知识图谱负责判断香味方向，商品接口只负责告诉系统当前有哪些真实商品。

KNOWLEDGE LAYER

先理解用户想要什么味道

把一句口语拆成香调、原料、甜度、场景、对象、预算和避雷项，再沿着图谱关系找到可能合适的香水方向。



COMMERCE LAYER

再去真实商品源里找有货的款

从拼多多和淘宝联盟拉取当前可购买商品，再按预算、品类和重复款做一轮硬筛。

如果最匹配的款没有货，我宁可推荐同方向的替代款，或者直接告诉用户暂时没找到，也不编一个不存在的商品。



03 · PRODUCT EXPERIENCE

从一句口语，走到一组真的能买的香水。

先把一句口语拆成预算、场景、香味偏好和避雷项，再根据这些条件找真实商品，并解释每个推荐为什么合适。

- 推荐过程中实时显示“正在理解什么、正在找什么”
- 每张商品卡都给出价格、来源、购买链接和推荐理由
- 用户可以继续说“更便宜”“不要太甜”，直接修正上一轮结果
- 信息不够时先问一个关键问题，不为了凑结果硬推商品



知香香水导购智能体 — 系统架构

对话编排 · 确定性香味理解 · 知识图谱混合检索 · 真实商品双源

● 真实商品 · 无模拟数据回退

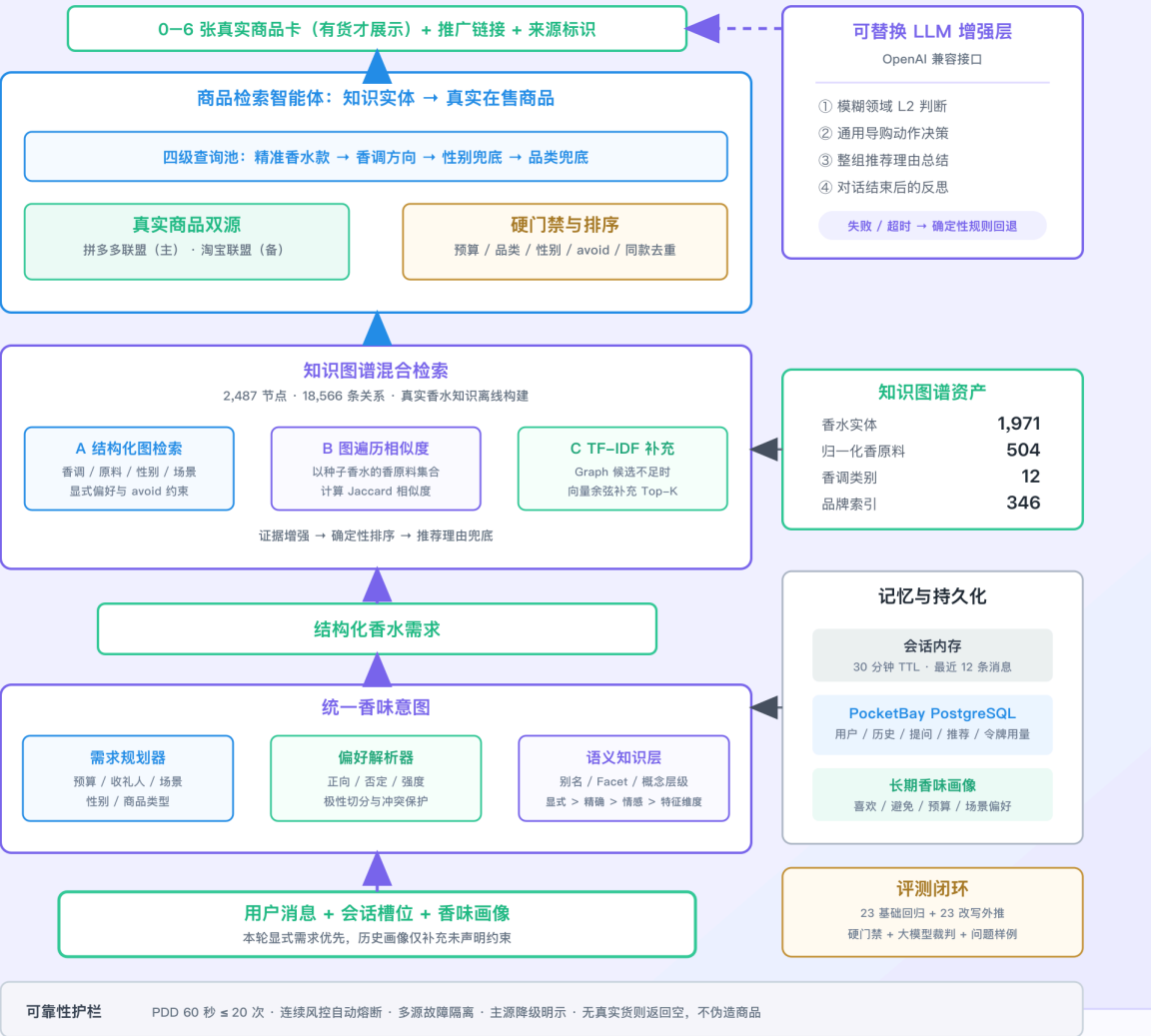
端到端智能体主链路

从用户输入到可购买商品卡，以 SSE 流式返回



香味理解 × 知识检索 × 真实商品召回

主链路以确定性推理为骨架，LLM 作为可替换增强层



香水领域知识图谱

品牌、香水、香调与原料通过结构化关系连接。

1,971

款香水

346

个品牌

504

个归一化原料

12

类香调

2,487

个图谱节点

18,566

条图谱关系

05 · KNOWLEDGE ENGINEERING

香水知识不是一堆文档，
而是一张“品牌—香水—香调—原料”的
关系网。

比如把“白衬衫的干净感”“甜的接受不了”翻译成可查询的香调、原料和否定偏好，再沿着图谱关系去找候选香水。

A

直接查明确条件

用户明确说出的香调、原料、品牌和人群条件，直接作为筛选条件。

B

沿关系找相近香

当精确命中太少时，只沿相邻原料和香调扩一层，避免越搜越偏。

C

文本检索补漏

图谱没覆盖到的描述，再用文本特征补充，但仍保留匹配依据。

为什么用图谱？ 用户说的是“感觉”，最终却要落到香调、原料和具体香水。图谱能把这条映射关系保留下来，也方便解释“为什么推荐”。

推荐的不是图谱里的名字， 而是此刻真正能买的商品。

系统先按“具体款 → 香调 → 性别 → 品类”逐层放宽搜索，从拼多多和淘宝联盟拿到商品后，再在本地做预算、品类和重复款检查。



如果商品源没有返回合适结果，就直接告诉用户“暂时没找到”，不拿假商品把页面填满。



淘宝显示“复制淘宝链接”；拼多多显示“复制拼多多链接”。

基础集综合评分 / 5

2.26 历史基线



3.14 V2.5

+39% 综合提升

各维度平均分 2.60 → 3.52 (+35%)；被标记的问题从 16 题降到 9 题。需要先追问的题，不拿去硬算商品匹配分。

07 · EVALUATION

每次改完都重跑一遍，看它是不是真的会更推荐了。

每次改动后都跑同一套回归测试：先用代码检查价格、品类、链接和商品是否真实，再用独立模型评估香调、场景、人群和预算是否匹配。

基础回归集

23 题

覆盖核心导购能力。

香调专项问题

10 题

综合分 2.11 → 2.78。

同义改写测试

23 × 2

换种说法后，两轮平均 2.88/5。

真实商品样本

250 件

228 PDD + 22 淘宝。

为了让问题能复现： 每道测试题都保留输入、商品结果、评分和失败原因；某一题出错后，可以单独重跑，不用每次整套重测。

让评测回答一个产品问题：用户的话，系统到底听懂了吗？

我把评测里反复失败的问题一类类拆开修，再用“同一个意思、换一种说法”的测试，检查系统是不是真的听懂了。

V1 · 基础理解 先补齐基础约束 用户问题 “400 以内”“不要甜”“随便看看”这类常见说法，早期版本经常理解错。 产品动作 补金额解析和否定词规则；条件不够时先追问；同一款商品不重复占位。	BAD CASE · 香调攻坚 保护用户真正想要的香调 用户问题 用户明明说了“不甜”“清冷”，系统却可能因为相近香调把甜味商品又带回来。 产品动作 把“不喜欢什么”和“喜欢什么”分开处理，降低伴随香调的权重，并补充行业词与商品描述的对应关系。	V2 · 意图深化 从“听见关键词”到“理解完整需求” 用户问题 真实需求往往一次说很多条件：上班用、送谁、什么味道、能不能持久、预算多少。只命中其中一个关键词，推荐还是会偏。 产品动作 把场景、对象、香调、预算、否定偏好等分别记录，并标出哪些是必须满足、哪些只是加分项，再让这些条件同时参与检索和排序。	V2.5 · 泛化验证 换个说法，也能做对 验证结果 基础集 3.14/5；改写外推两轮 2.88/5。 产品判断 换种问法后，场景和适用人群仍能基本判断对；香调仍是最容易出错的一项。
---	--	--	--

2.26 → 3.14 迭代重点也变了：先解决“有没有商品”，再解决“推荐出来的商品是不是符合用户原话”。



登录后可以在不同设备继续查看自己的历史对话。

09 · ENGINEERING DELIVERY

让用户第二次打开时，还能接着上次聊。

Demo 跑通一次并不难；真正上线后，还要解决登录、历史记录、多人数据隔离，以及切换会话时会不会串数据。

PostgreSQL 持久化

保存账户、会话、消息、推荐与埋点数据。

用户数据隔离

鉴权并校验会话归属，只能读取自己的历史。

流式竞态保护

切换会话、清空或退出时取消旧 SSE。

PocketBay 部署

H5 与 Express 同源上线，环境变量管理密钥。

Taro

React

TypeScript

Node.js

Express

SSE

PostgreSQL

PocketBay

做完之后我发现，最难的不是生成推荐，而是知道什么时候该推荐、该追问，什么时候该说“没有合适商品”。

复盘整个项目，我最后留下了四条最有用的产品判断。

01 / BOUNDARY

先把“懂香”和“有货”分开
知识图谱只回答“什么香可能适合”，
商品接口只回答“现在有没有货”。缺货就直接说，不拿假商品补位。

02 / EVIDENCE

把“推荐准不准”
拆成能查的问题
把预算、香调、场景、否定偏好、商品真实性分别评分。哪一项掉分，就知道下一轮该修哪里。

03 / CONTROL

把该交给模型和规则的事分开
模型负责理解“清冷、干净、白衬衫感”这类模糊表达；价格、明确否定词、品类和商品真实性则交给确定性规则。

04 / NEXT

下一步看推荐有没有真的帮到购买
下一步不只看模型评分，还要接入点击、收藏、购买和试香后的反馈，判断用户到底有没有因为推荐更快做出选择。

核心收获 这个项目最大的收获是：模型只解决一部分问题。真正决定体验的，是数据从哪来、哪些规则不能错、出了错能不能被评测发现，以及用户反馈能不能回到下一轮迭代。

—— MY OWNERSHIP

把知香从用户访谈和原型，做到可用、可测、可上线。

我独立负责用户研究、产品方案和交互设计，并用 AI Coding 完成 Agent 流程、知识图谱、真实商品接入、评测系统、数据库与线上部署；从发现问题到上线验证，整条链路都亲自跑过。

- 需求研究
- 产品设计
- Agent 架构
- 知识工程
- 评测迭代
- 工程上线

在线体验知香 ↗

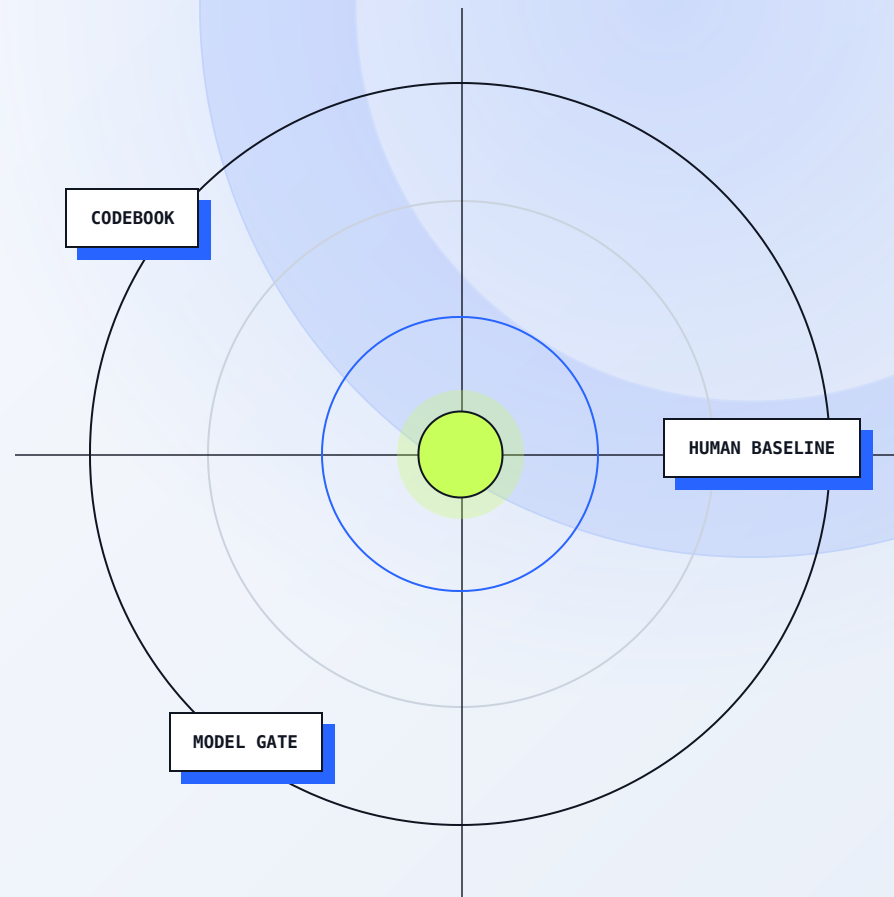
查看高清架构图

给 LLM 建一把可信的尺子

没有天然 Ground Truth 时，我们凭什么相信模型？

在 EALA 国际气候谈判标注项目中，我组织人工评测团队，把专家规则、双人独立标注、分歧讨论与 Human Baseline 组合成一套模型质量验证机制。

邵志成 | 模型评测团队负责人 AI EVAL · HITL · DATA QUALITY



— 01 · PRODUCT PROBLEM

原问题不是“模型能不能标”， 而是“标完敢不敢用”。

国际气候谈判数据库长期依赖人工标注。文本持续新增、规则复杂、标注者也会产生分歧；如果只追求自动化，很容易得到一批“看起来完整、其实不可验收”的数据。

10 年+

原人工数据库
长期缺少持续更新

产品重构：把“LLM 标注模型”重新定义成“可持续的数据生产系统”。系统不仅要能生成，还要有明确的质量门槛、错误反馈和人工兜底。

01**持续更新难**

人工团队成本高，数据更新容易滞后几年甚至更久。

02**标签边界模糊**

Agreement、Support、Opposition 等关系需要结合上下文和规则判断。

03**人工也会不一致**

如果 Human Label 自己不稳定，模型分数就没有可信参照系。

04**错误会被放大**

一旦进入规模化生产，系统性误差会直接污染下游研究数据。



— 02 · WHY IT IS HARD

这不是普通分类，而是复杂结构化判断。

每个段落要输出“发送者—接收者—关系—议题”的动态四元组。很多关系并没有原文直说，而要依据 Codebook 做推断；因此一个总 Accuracy 很难说明系统是否真的可靠。

隐含关系

多人共同支持或反对时，还要推断他们之间的 Agreement。

范围约束

文本里出现的所有主体并不都属于可标注实体空间。

动态语义

谈判议题持续演化，旧标签空间无法永久覆盖新问题。

— 03 · PRODUCT DECISIONS

我先做四个产品决策， 把“模型好不好”变成可执行标准。

复杂 AI 产品最危险的不是分数低，而是分数没有可信含义。先把尺子做准，再谈模型优化。

DECISION 01

先定义“什么算对”

把 Codebook 的关系定义、排除规则、方向性和隐含关系推导统一成评测口径。

RULE FIRST

DECISION 02

双人独立标注，而不是一起商量

先保留真实 disagreement，再通过讨论定位规则歧义与判断差异。

BLIND FIRST

DECISION 03

Precision / Recall 一起看

既不能“多标错关系”，也不能“漏掉真实关系”，不追一个好看的单指标。

MULTI-METRIC

DECISION 04

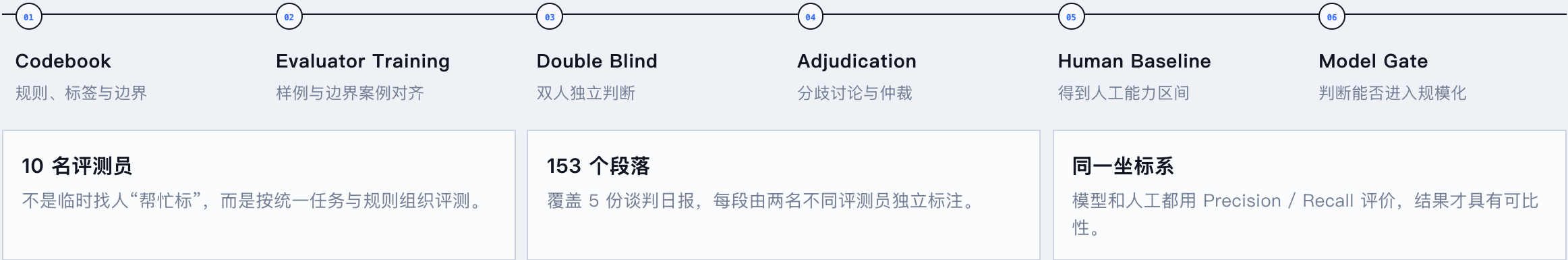
Eval 是持续质量闭环

错误不能停在 benchmark 表里，要回流到规则、样本和模型侧。

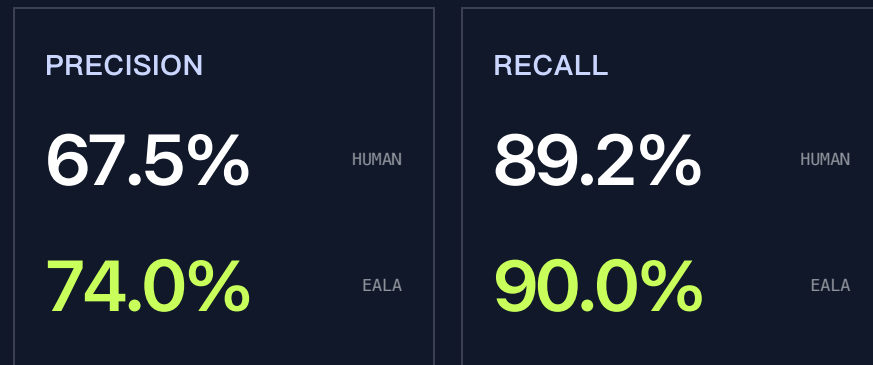
ITERATE

先把“专家判断”做成一条可重复的评测流程。

我组织 10 名评测员参与关系抽取评测。每段文本由两人独立标注，分歧再讨论；最后用同一套 Precision / Recall 去比较模型与人工。



RELATION EXTRACTION · HUMAN vs EALA



结论不是“模型全面替代人工”，而是：在关系抽取任务上，EALA 已进入可用于规模化辅助生产的质量区间。

— 05 · MODEL GATE

不是证明模型“赢了人”， 而是证明它“够不够用”。

对数据生产系统而言，模型有没有价值取决于是否达到可接受质量门槛，而不是 headline 上是否“超过人类”。

Precision 提升

EALA 74.0%，高于单个人工标注者 67.5%，错误关系的比例进一步下降。

Recall 基本持平

EALA 90.0%，与人工 89.2% 基本处在同一水平，没有明显牺牲“找全关系”的能力。

产品判断

质量达到进入 Human-in-the-loop 生产链路的条件，但机器生成数据仍需要持续纠错。

历史任务评得好还不够，模型遇到“新议题”怎么办？

关系抽取有大量历史标注，是 data-rich 场景；动态议题预测却缺少成熟标签，更像真实 AI 产品里的新 Intent / 新类别冷启动。

50 谈判段落 从已有数据中抽取，构建议题预测专项评测集。	394 互动关系 每条关系都需要判断它围绕哪个动态议题展开。	8 名 coder 每条样本由两人独立判断，分歧再通过讨论解决。	0.79 IAA 先确认“人之间是否有稳定共识”，再谈模型到底有多准。
--	---	---	--

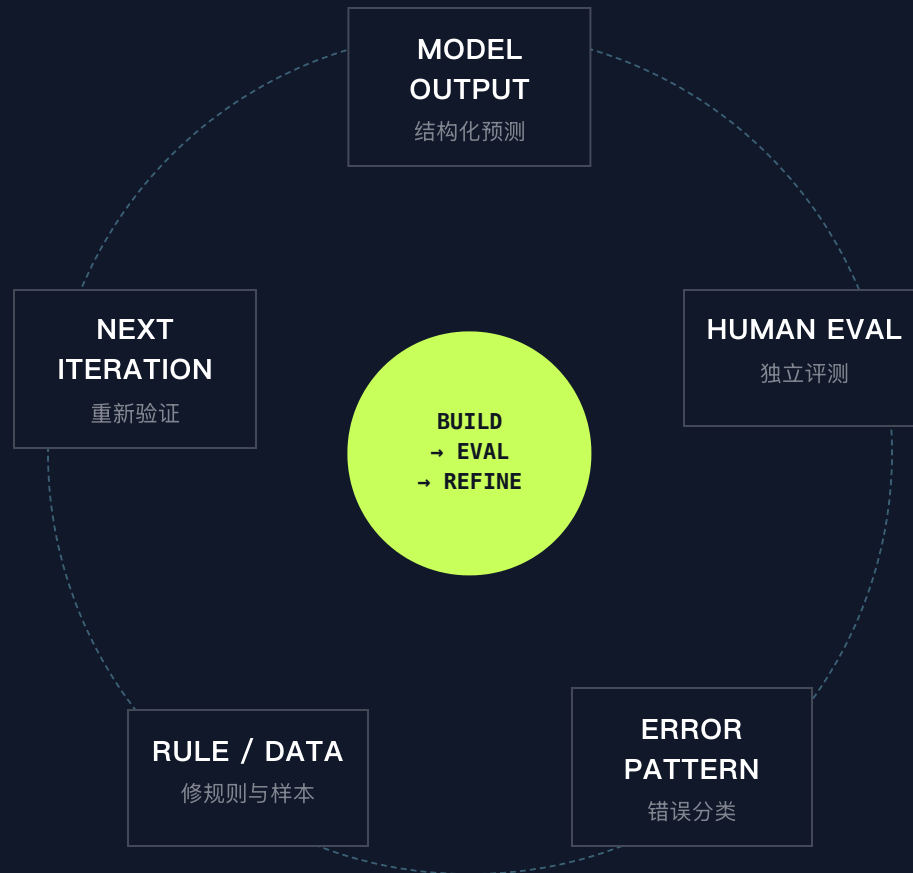
产品启发：“未知类别怎么评”往往比“已知 benchmark 能跑几分”更接近真实上线环境。先建立可解释的人类共识，再让模型进入比较。

— 07 · QUALITY LOOP

从“评模型”升级到“管质量”。

一个成熟的 Eval 体系不只产生一个分数，还要把问题转成下一轮可行动的输入。论文也明确提醒：机器生成数据不能未经纠错直接用于实质研究。

我的理解：Human Eval 是产品闭环中的“诊断层”——负责把模型失败变成规则修订、样本补充与下一轮模型迭代。



— 08 · SCALE

评测的价值，最终要落到规模化数据生产。

不是证明一个 demo 好看，而是让模型在 30 年纵向语料里持续生成结构化数据，并知道哪些环节仍需人工把关。

CORPUS	844 收集到的 ENB 报告
FILTER	678 纳入的 UNFCCC 日报
VALID	589 检测到 interaction 的日报
OUTPUT	45,051 结构化互动关系

1995—2024

近 30 年谈判数据统一进入同一纵向网络框架。

45,051

机器生成结构化 interaction，支撑后续网络分析。

27.8%

动态议题描述仍经人工修订，说明 HITL 不能被“自动化”替代。

1 → N

从人工小样本验证，扩展到大规模持续更新的数据流水线。

— 09 · MY OWNERSHIP

我不是模型架构负责人； 我的价值是把“可信”做成流程。

在一个跨学科项目里，产品思维最重要的是把研究规则、人工判断与模型指标对齐成同一套质量机制。

角色边界：模型训练与框架由合作团队共同完成；我负责人工评测团队组织、评测标准执行、Human Baseline 与部分议题空间质量验证。

01 · 需求与标准

理解 Codebook 与关系抽取边界，把专家判断转成可执行评测任务与验收指标。

02 · 团队与执行

组织 10 名评测员，推进双人独立标注、分歧讨论与结果整理。

03 · 质量与结论

构建 Human Baseline，参与 50 段 / 394 条议题预测评测集与人工审读。

04 · 产品化表达

把“论文里的评测实验”抽象成 Model Gate 与 Human-in-the-loop 的可复用 AI 产品方法。

做完这个项目，我对 AI Eval 留下了四个产品判断。

模型分数只是结果；真正决定一个 AI 数据产品能不能长期运行的，是 Ground Truth、错误类型、人工边界和反馈闭环。



01 · GROUND TRUTH

先测“人”是否稳定

如果专家之间都无法形成相对稳定的一致性，就不该把某个标签当成绝对真值。

产品动作：先做 Human Baseline。



02 · METRIC

不要迷信单一分数

Precision、Recall、IAA、错误类型与下游影响，需要一起决定模型是否可用。

产品动作：把指标与风险绑定。



03 · CONTROL

不要追求 100% 自动化

复杂专家任务天然有主观边界，合理目标是“机器规模 + 人类校验”的分工。

产品动作：保留 HITL。



04 · LOOP

评测必须能驱动下一轮

如果失败案例不能回到规则、样本与模型侧，Eval 就只是一次报告，不是产品机制。

产品动作：Build → Eval → Refine。

— PRODUCT TAKEAWAY

AI 产品不只要做出来， 还要证明它值得被相信。

这个项目让我第一次完整参与“规则 → 人工基线 → 模型比较 → 规模化生产”的闭环，也让我把 Eval 从一个模型指标问题，理解成 AI 产品的质量基础设施。

AI Eval

Human-in-the-loop

Model Gate

Data Quality

Expert Knowledge

Evaluation Design

[查看合作论文](#) ↗

[返回作品集](#)