Human Agent Coevolution

构建面向人机共演化的Context Layer

王天富 | 2026//07

TianfuWang.tech

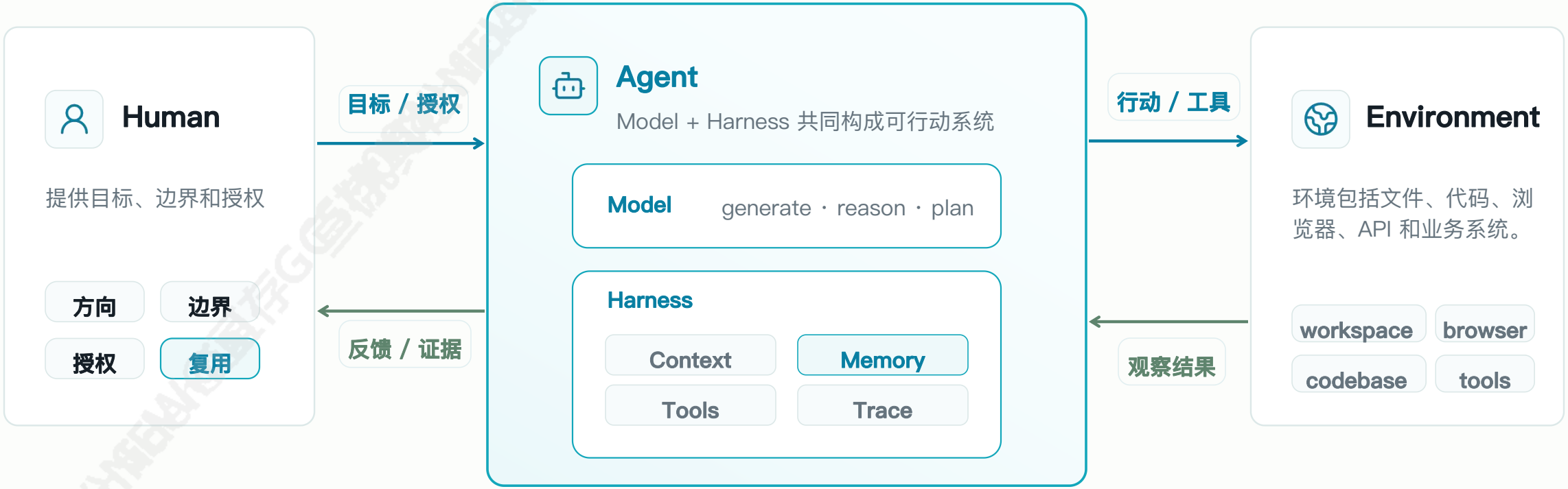
github.com/GeminiLight

tianfuwang.cs@gmail.com

01 Harness

Harness 让模型成为可行动的 Agent

模型提供能力，Harness 管理目标、工具、权限、状态和反馈。



人的作用：给出方向、形成判断、设定边界、授权行动，并决定哪些经验可被用。

01 Harness

长期运行后，Harness 开始承担经验治理

一次执行结束后，trace、反馈和纠错会影响后续任务，关键问题开始从执行转向经验治理。

01 Action

行动控制

tools / workflow
permission / observability

这一步能否安全完成?

02 Experience

经验形成

trace / feedback
correction / preference

这次经历暴露了什么?

03 Judgment

判断对齐

intent / evidence standard
boundary / reuse

哪些经验值得复用?

行动留下 trace，trace 形成经验，经验经过治理后再被复用。

02 Harness

Agent 越快，越会放大隐含判断

真正的风险来自隐含判断：证据、边界和复用规则没有被显式说清。

01

快速行动

搜索、改写、调用工具、
更新状态都变快。

02

隐含判断

看哪些证据、改到哪里、
哪些经验可复用。

03

默认放大

隐含判断进入后续行动

判断留在暗处

风险被加速

证据漂移 · 越界修改 · 错误复用

外化

判断进入系统

系统更稳

证据依据 · 边界条件 · 复用范围 · 回滚条件

人的价值在于把关键判断外化成系统可用的约束。

02 Human

Harness 需要人的判断：行动授权与经验治理

行动前，人定义目标、边界和委派；经验后，人解释证据、纠错并授权复用。

行动授权与经验治理的判断门

Agent 可以持续执行，但关键状态转移仍要经过人的判断。

行动前

意图 01 什么值得做?	**边界** 02 什么不能越界?	**委派** 03 可以自主到哪一步?

经验后

证据标准 04 什么证据足够?	**纠错解释** 05 错误说明了什么?	**复用** 06 哪些经验可以复用?

人的判断让行动保持可承担，也让经验保持可审阅、可授权、可回滚。

02 Human

Agent 折叠了执行过程，也折叠了人的判断练习

如果人只验收结果，试错、外化、反馈和复盘会减少，判断力会失去练习场。

过程可见，但难以形成练习

执行越快，依据、假设和取舍越需要被整理出来。

01	**执行外包**	搜索、草稿、试错和改写被快速完成	试错减少
02	**过程难审阅**	步骤、依据、假设和不确定性散在 trace 里	追问减少
03	**只验收结果**	人少经历提问、取舍和校准	反馈减少
04	**默认接管**	证据、风险和边界逐渐依赖系统默认	主体性变弱

让省下的执行时间部分流向提问、证据、边界和复盘练习。

MindOS 把协作经验治理成 Shared Mind

目标、证据、边界、纠错和复盘被整理成可审阅、可授权、可回滚、可复用的状态。

协作后的流失

关键判断常常留在对话、结果和人的脑中。

01	**上下文散落**	目标、证据留在对话里
02	**判断不透明**	依据与证据不清
03	**复用失控**	临时偏好进入后续默认
04	**练习断开**	只验收结果，复盘丢失

→ **Shared Mind** →

MindOS 的方案

Shared Mind 把这些判断整理成后续任务可使用的状态。

01	**状态可审阅**	来源、证据和取舍清楚
02	**经验可复用**	复用范围和责任可确认
03	**默认可回滚**	错误经验可以撤回或改写
04	**成长可反哺**	复盘线索回到判断练习

用 Shared Mind 承载协作经验

协作经验先被沉淀成可审阅的记忆，再转成 Agent 可用的证据和运行时约束。

经验承载

01	**主记忆**	**人能读、能审、能改**	笔记 · 决策 · SOP · 边界 · 证据 · 来源
02	**运行索引**	**Agent 能检索、能引用**	搜索 · 图谱 · 语义索引 · 上下文包
03	**协作约束**	**运行时能提示、能限制、能回滚**	工具 · 工作流 · 权限 · 审阅队列

进入协作

协作中的四个作用

记 记忆管理

治理什么经验进入长期状态

认 认知建模

建模人的判断系统与成长轨迹

协 主动式协作

在关键节点形成状态共识

演 共同演化

同时更新系统能力和人的判断

03 MindOS

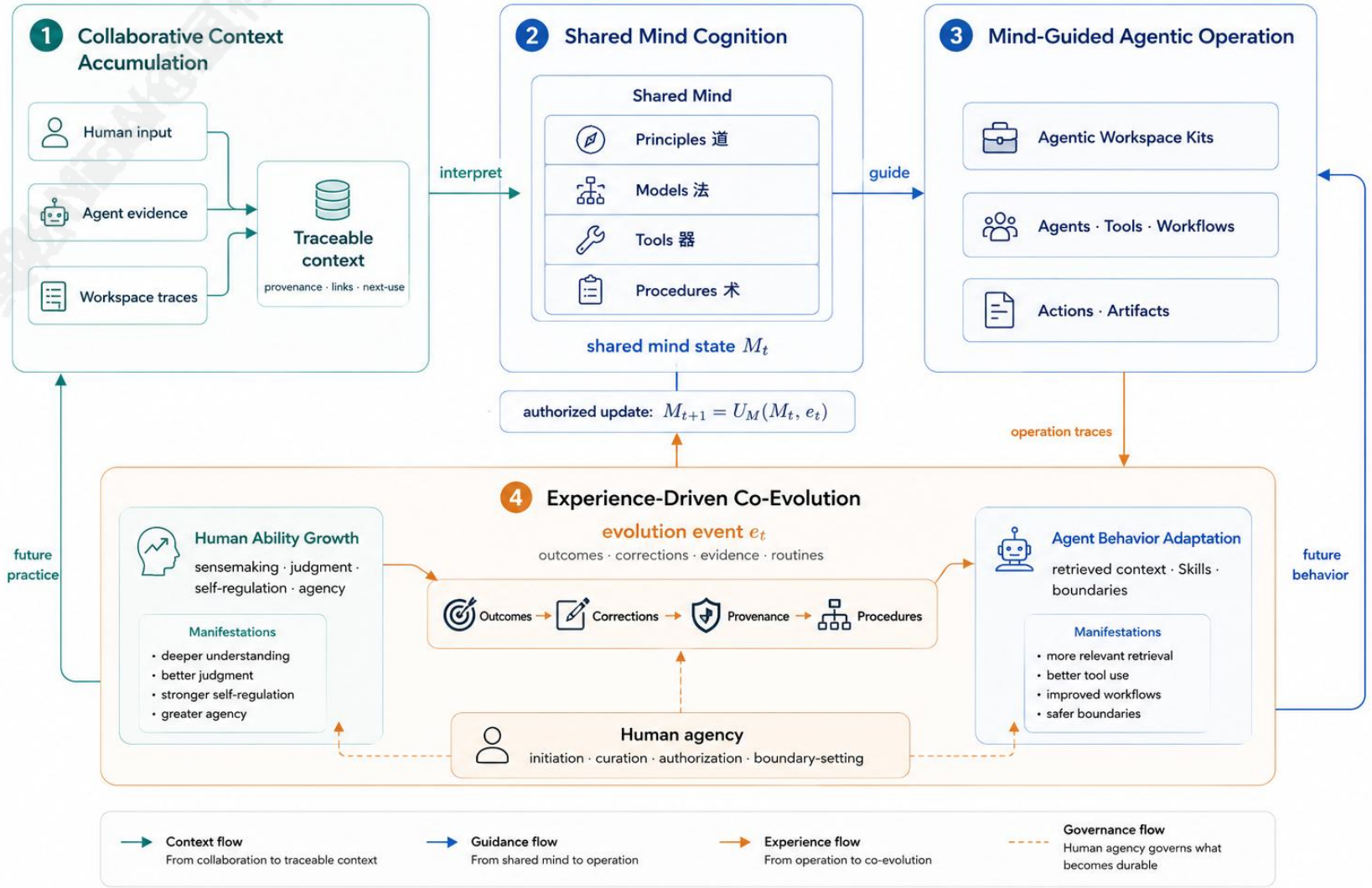
人机共演化的心智系统

记忆管理

认知心智

主动协作

共同演化



A. 人机共治的记忆系统

协作经验先由 Agent 抽取为候选知识，再由人校准关键边界，沉淀为可治理的长期记忆。

经验累积

任务轨迹
人的判断
环境证据

Agent 抽取

候选知识

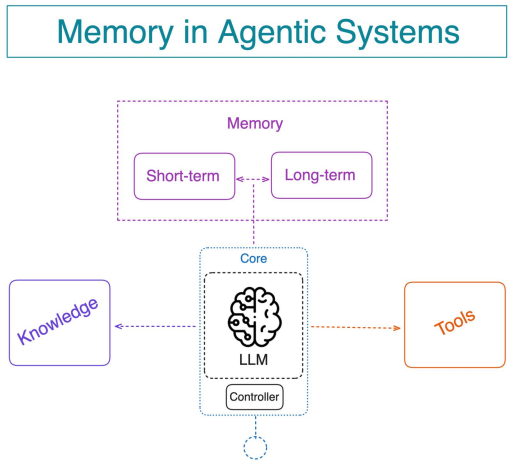
偏好
规则
边界

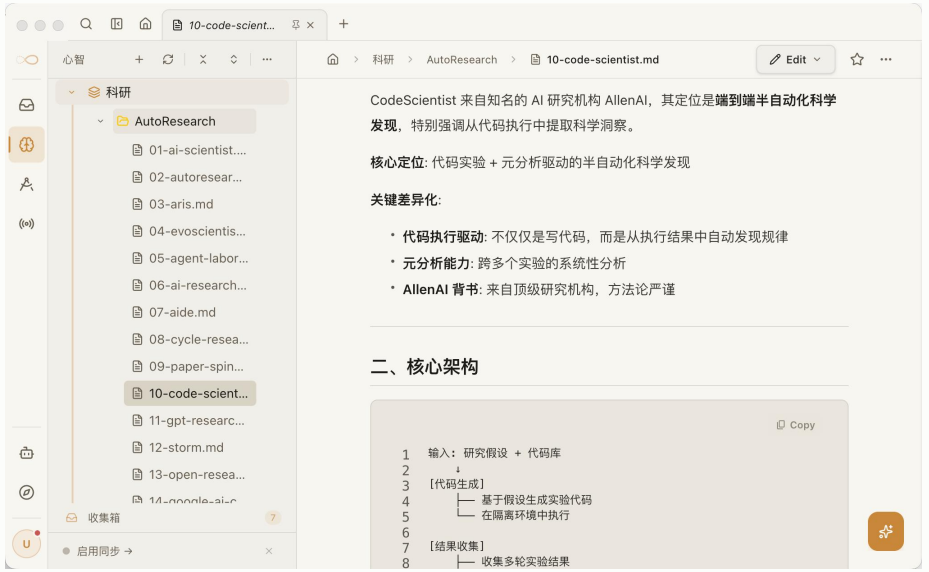
人类校准

治理记忆

可调用
可追溯
可回滚

从 Agent Memory
到人机共读





记忆管理：Memory–as–Compilation

长期记忆在查询时被编译成小而准、可引用、可验证的 Evidence Pack。

总体结构

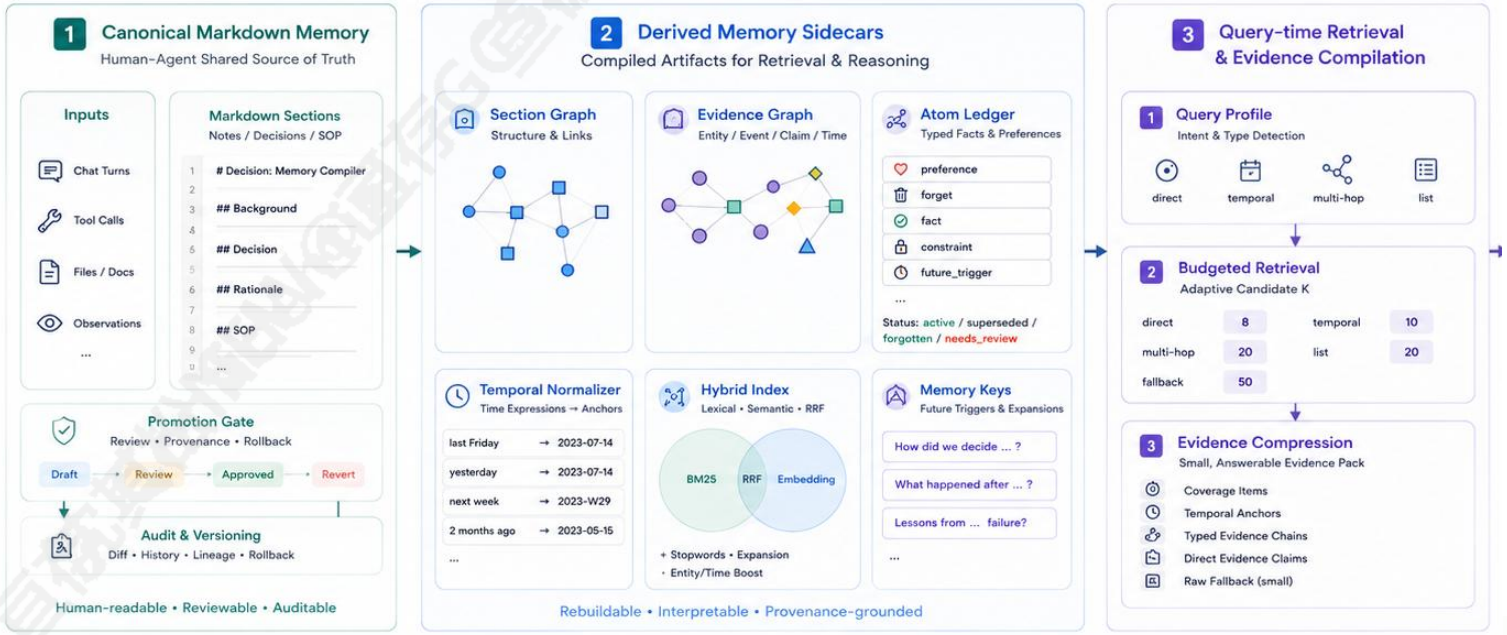
Canonical Markdown
人机共读的主记忆

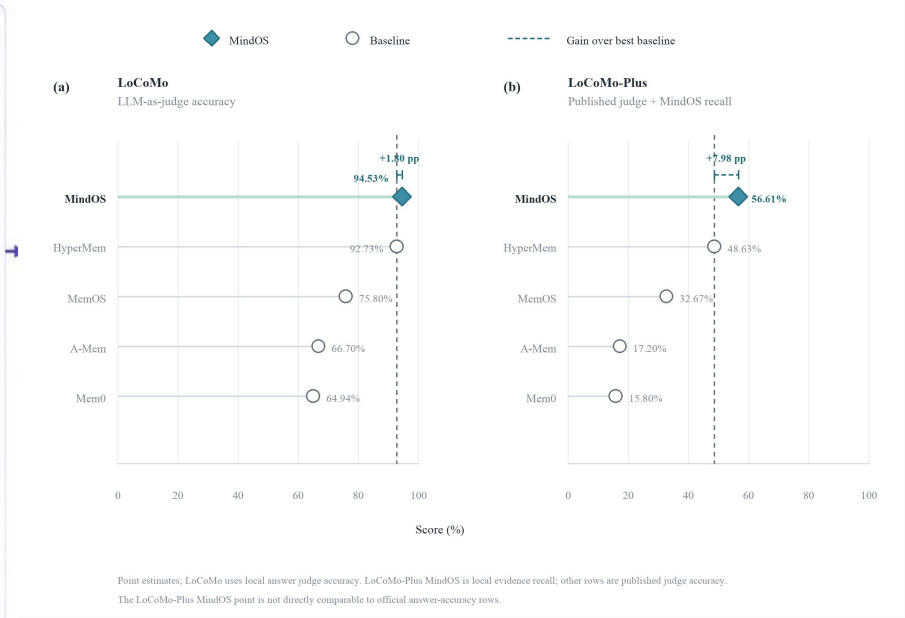
→

Derived Sidecars
可重建的运行时结构

→

Evidence Pack
小而准的可引用证据





认知建模：建模人的判断系统

认知建模描述人在情境中的判断状态。

普通 personalization

偏好、格式、历史行为

能记住人，却难以理解人如何行动。

认知建模

人如何把意义、规则、方法和工具组织成行动系统

静态结构

- **道** 意义与方向 长期目标
- **法** 判断原则 授权边界
- **术** 实践方法 复盘习惯
- **器** 工具环境 外部脚手架

动态轨迹

- **状态变化** 任务、角色、注意力
- **场景切换** 边界、授权、风险
- **成长轨迹** 判断更清楚，行动更稳定

C1. 主动协作：Agentic Consensus

主动协作的第一步，是让人真正看懂当前状态。

隐藏工作状态

trace
workspace context
历史经验
证据来源
Agent 推理状态
风险信号

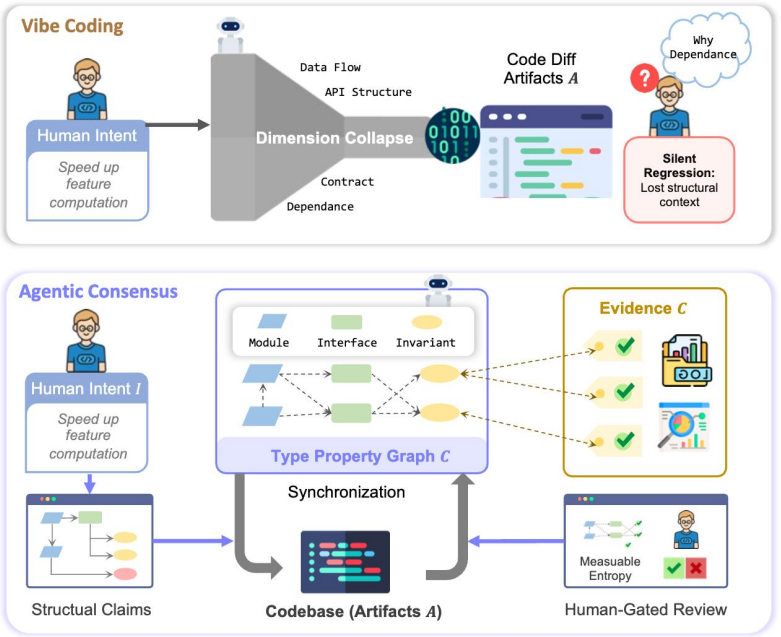
人需要获得什么

人需要看见依据、边界、
不确定性和下一步选择。

Agentic Consensus

把 Agent 正在依据什么、假设什么、还不确定什么，主动整理成 shared state。

目标　进展　证据
假设　风险　选项



Agentic Consenus | Preprint 2026

C2. 主动协作：Intervention Routing

先形成 Agentic Consensus，再决定哪些判断必须交还给人。

输入

shared state
风险等级
可逆性
权限边界
人的角色
当前决策类型

→

Intervention Routing

把正确判断，在正确时刻，交还给正确的人。

继续执行	先问
请求确认	升级专家
暂停	拒绝自动执行

→

人获得什么

不会每一步都被打扰
但会在关键状态转移点
重新获得判断权

高风险动作	不可逆操作
证据不足	影响范围不明 测试覆盖不足
越过边界	权限或委派变化
写入记忆	长期记忆 SOP 复用

失败模式：过度打扰、问得太晚、问错人，或者把人变成 rubber stamp。

共同演化：人机系统一起演化

同一次协作经验，一边更新 Agent System，一边推动人的能力成长。

D1. Agent Evolution 发生在参数内外

Agent System 先在参数外快速演化，再把稳定经验蒸馏回参数内。

What evolves?

Model	参数、能力、偏好、推理习惯
Context	任务、可见信息、工作标准
Memory	偏好、项目历史、纠错记录
Skills / Tools	SOP、工具策略、操作例程
Architecture	planner · reflection · graph

How does it evolve?

Parametric Evolution

SFT · RL

把稳定模式内化进模型参数。
强泛化，但成本高、周期慢、难回滚。

Non-parametric Evolution

In-context Learning
Memory Update

不训练模型，更新外部状态与运行结构。
快速、局部、可审计、可个性化。

参数外快速适应 → **稳定模式再内化**

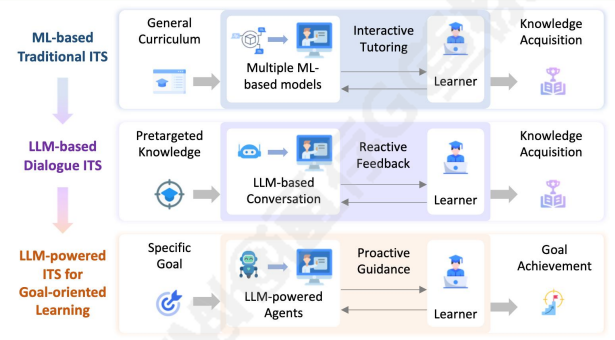
D2. Human Growth 人的能力如何演变

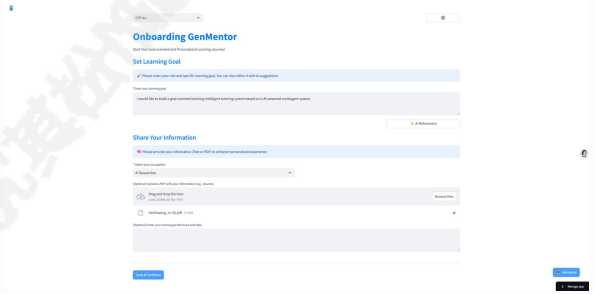
人的成长体现在学习方式、练习方式和主体性的上移。

01 目标导向学习

系统化学习 → **目标导向型学习**

围绕真实任务与反馈形成学习目标。



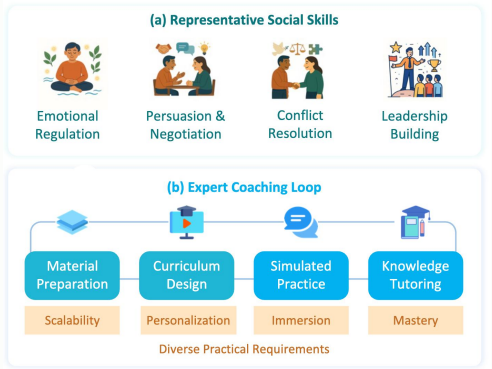


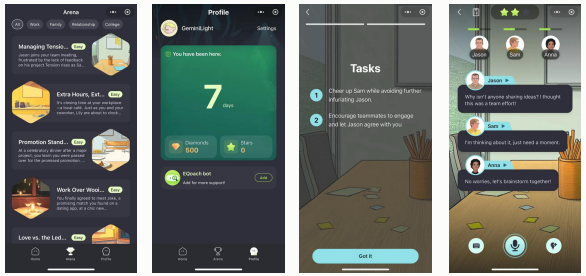
GenMentor | WWW 2025

02 输出式练习

被动式输入 → **输出式练习**

在输出、判断、纠错和复盘中反复产出。





SocialCoach | Preprint 2026

03 主导权提升

技能增长 → **Agency 成长**

发起目标、授权复用，并设定行动边界。

D2. 成长回路：经验如何反哺人

MindOS 把一次协作变成可检查、可校准、可外化的练习回路。

Human growth surfaces 把协作历史变成可反思、可练习、可治理的成长材料。

一次线上纠错，会同时触发四类系统更新

好的纠错，会同时更新记忆、判断、协作和演化方式。

案例：Agent 修复线上 bug，顺手改了共享状态；测试通过，却影响了另一个入口。

记忆管理

保存纠错轨迹
记录误改入口与影响范围
标注回滚和复用条件

认知建模

学习人的工程判断
共享状态改动需确认
默认值与权限不能隐性改变

主动式协作

先形成状态共识
发现跨入口影响时暂停
把关键判断交还给人

共同演化

Agent 更新测试策略与 SOP
人外化审阅标准
团队复用更稳交付规则

评测要证明人机系统真的在变好

任务成功只是起点；要看记忆管理、认知建模、主动协作与共同演化是否改善。

01 记忆管理

经验能否被正确沉淀

来源可追溯
范围可限定
变更可回滚

02 认知建模

能否建模人的判断力

标准可见
边界可辨
变化可追踪

03 主动式协作

系统能否主动推进协作

主动发现
适时发起
协作闭环

04 共同演化

人机系统是否持续变好

可靠性提升
复盘可用
做法可共享

成长对象会从个人扩展到人机系统

智能体越强，人的成长越要从提升自己，扩展到治理整个人机系统。

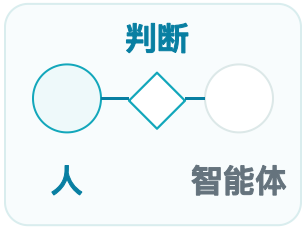
A 个人能力成长

对象：我

成长：能力与效率

问题：我能做什么？

B 协作判断成长

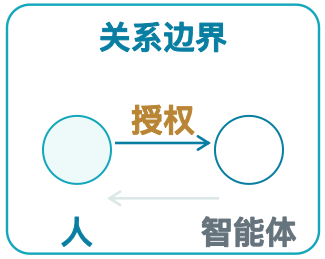


对象：我和智能体

成长：判断与校准

问题：什么值得信任？

C 关系治理成长

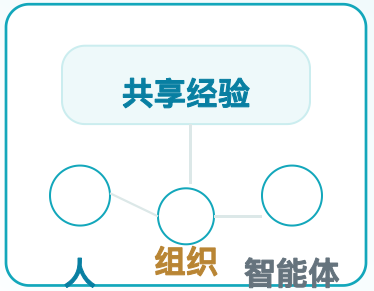


对象：人机关系

成长：授权与责任

问题：谁来决定？

D 系统共同成长



对象：人机系统

成长：共同演化

问题：我们塑造什么？

成长对象扩大：我 → 我和智能体 → 人机关系 → 人机系统

04 Vision

Mind 会从个人判断扩展到共同体治理

Mind 的尺度越大，越需要可解释、可协商、可修正。

Personal Mind	**影响我的下一次判断**	偏好 / 边界 / 复盘 / 委派习惯	**风险：判断变窄，过度依赖默认建议**
Collective Mind	**影响团队以后的协作**	SOP / 项目历史 / 角色边界 / 决策记录	**风险：坏 SOP 固化，少数经验变成团队默认**
Societal Mind	**影响社会如何记住与行动**	公共知识 / 制度记忆 / 规范 / 争议记录	**风险：单一价值观固化，公共记忆缺乏争议机制**

Human–Agent Coevolution

同步优化人类思维 Harness

Thanks!

MindOS
Demo & Code & Report

https://github.com/GeminiLight/MindOS

Tianfu Wang | 2026/07