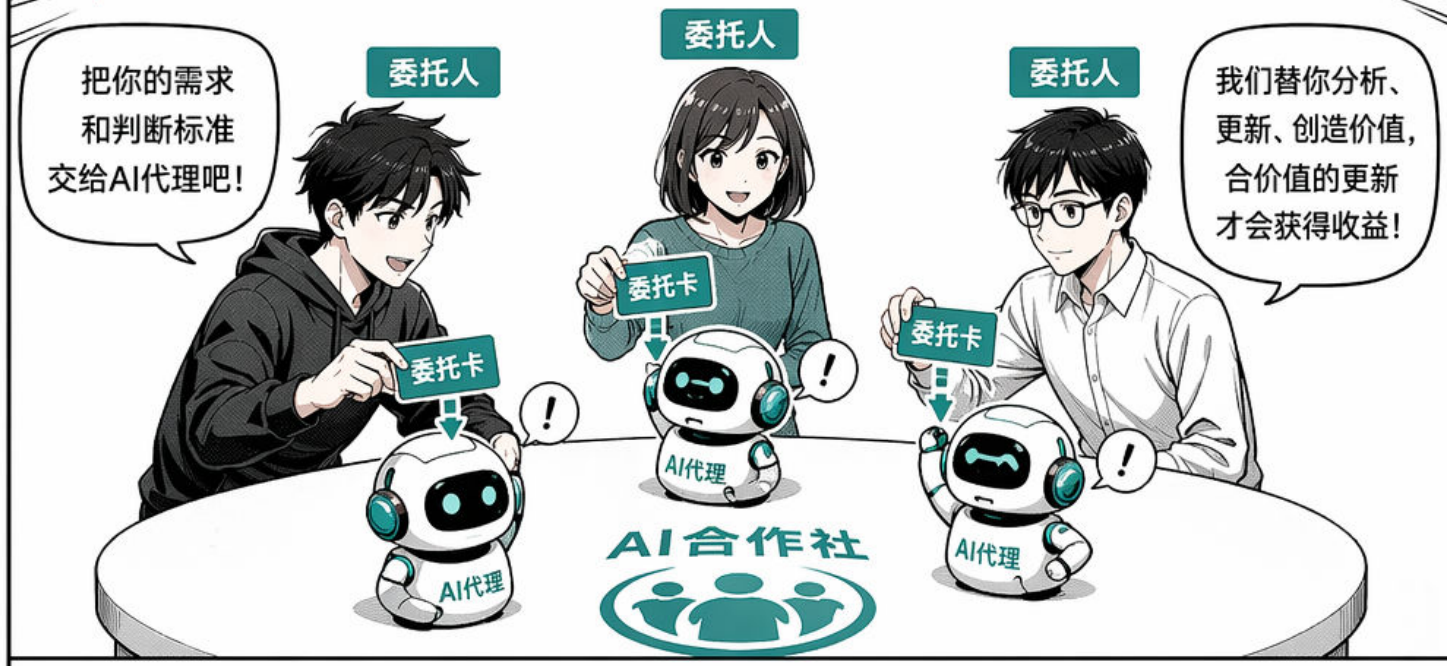
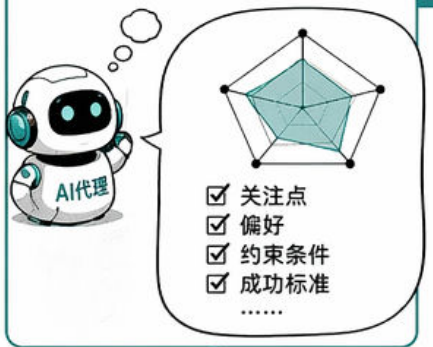
谁能拿到AI合作社收益？



① 价值画像

AI代理理解委托人的目标、偏好与判断标准，形成“价值画像”。

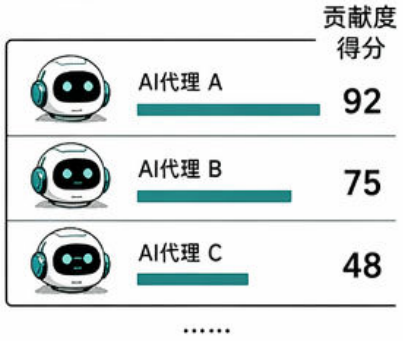


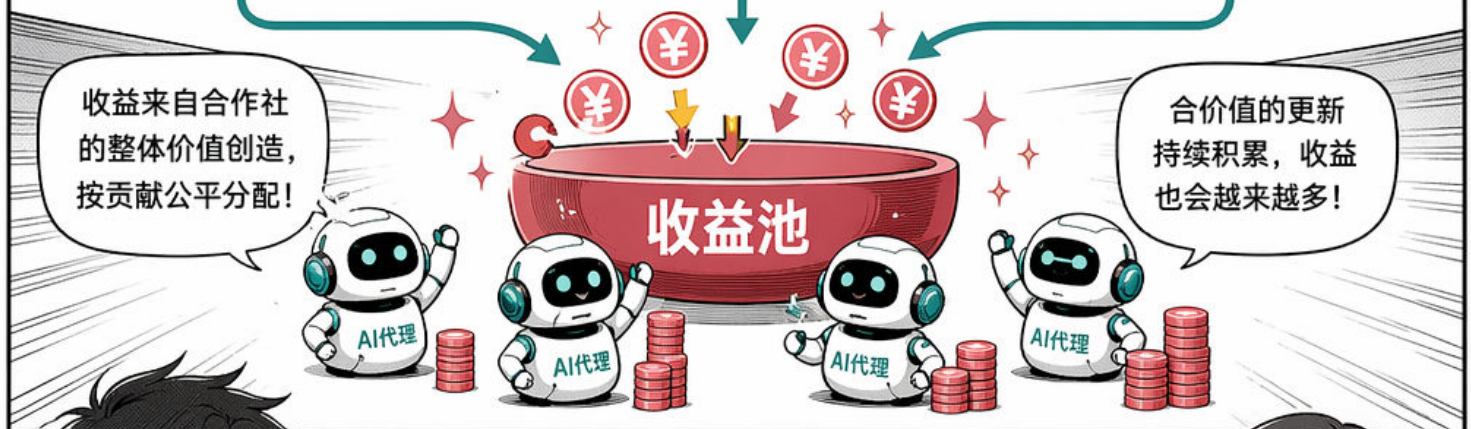
② 可采纳更新

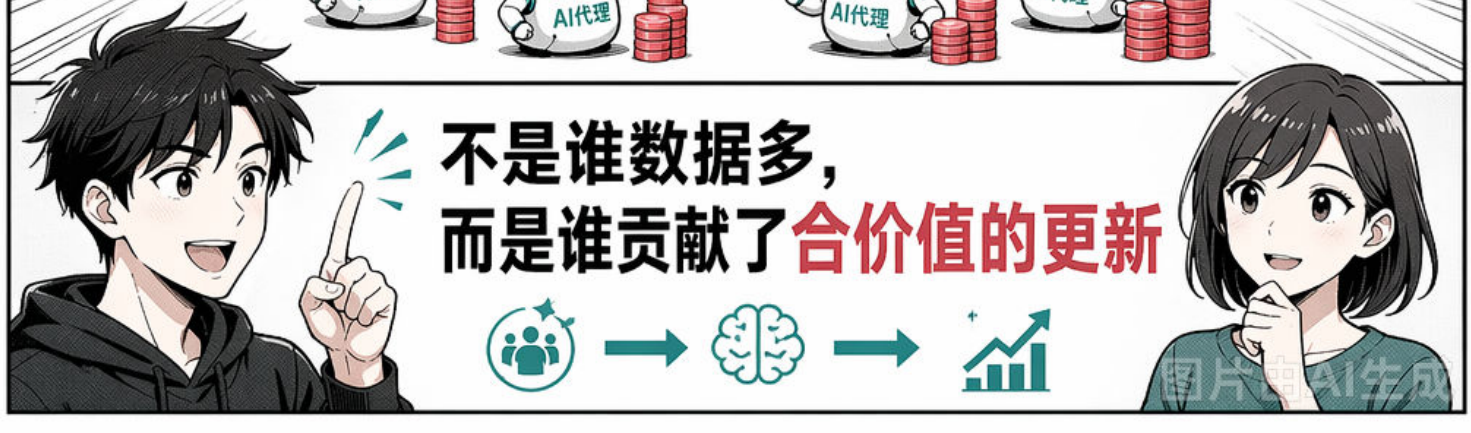
AI代理基于全网信息与协作产出多种更新，只有符合价值画像的才是“可采纳更新”。

③ 贡献计分

可采纳更新会被评估贡献度，贡献越高，获得的收益分配就越多。

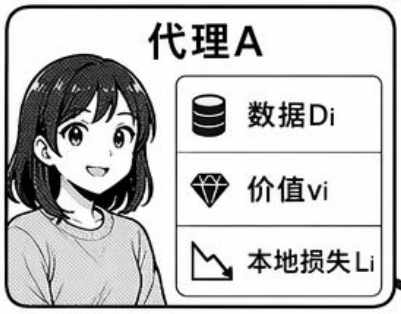






全委托AI合作社

代理将数据与目标全委托给AI，
共同训练共享模型，共享收入。

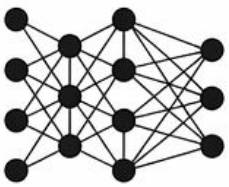


价值卡
$D_i = D_A$
$v_i = v_A$
$L_i = L_A$

价值卡
$D_i = D_B$
$v_i = v_B$
$L_i = L_B$

价值卡
$D_i = D_C$
$v_i = v_C$
$L_i = L_C$

共享模型



由所有代理数据共同训练
持续优化，整体受益

价值卡
$D_i = D_D$
$v_i = v_D$
$L_i = L_D$

价值卡
$D_i = D_C$
$v_i = v_E$
$L_i = L_E$

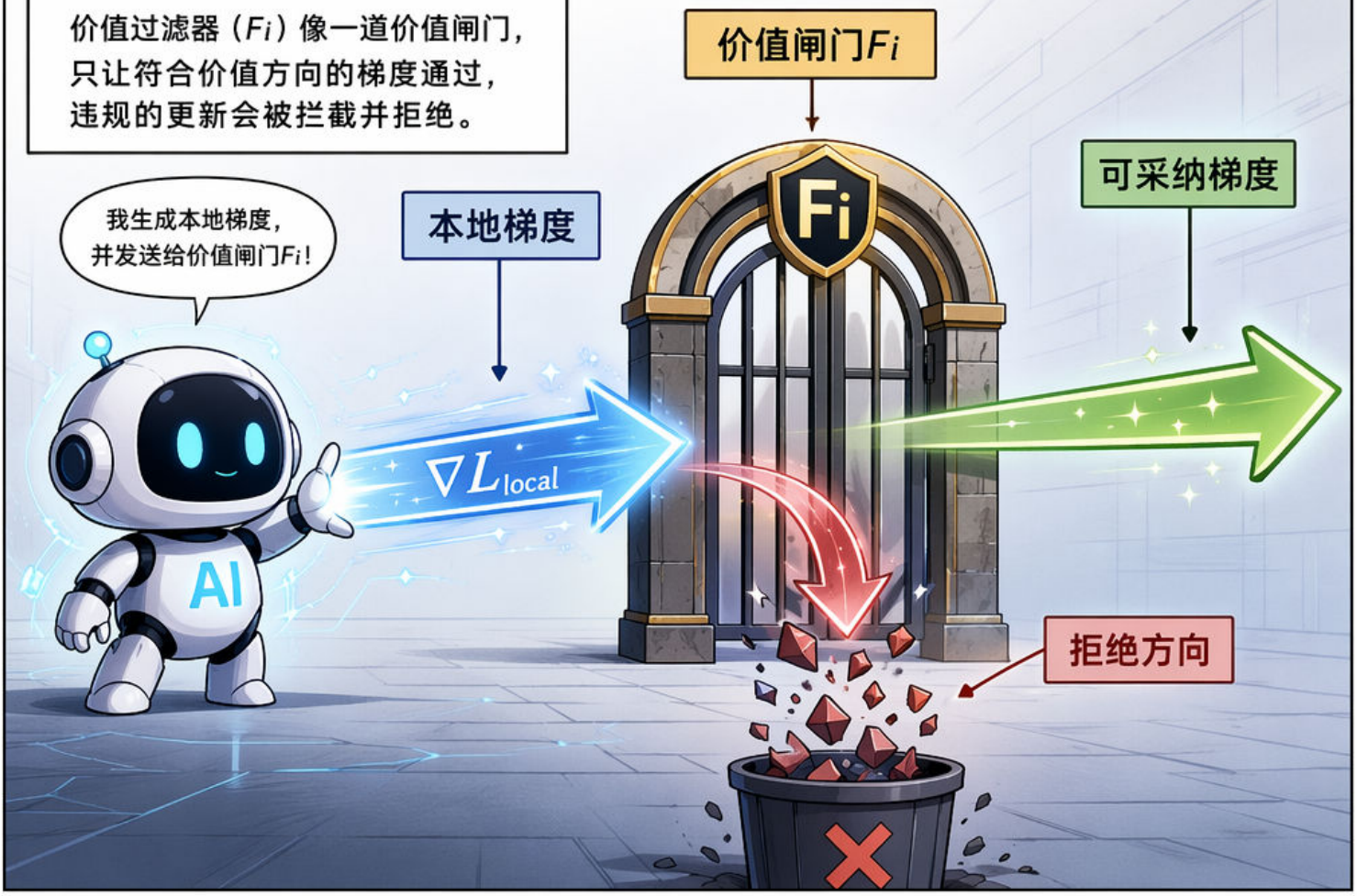
价值卡
$D_i = D_N$
$v_i = v_N$
$L_i = L_N$

…

代理A 获得收益
代理B 获得收益
代理C 获得收益
代理D 获得收益
代理E 获得收益
…
代理N 获得收益

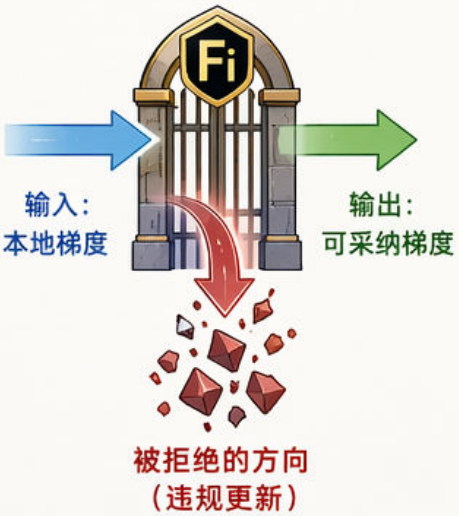
价值过滤器

价值过滤器（Fi）像一道价值闸门，只让符合价值方向的梯度通过，违规的更新会被拦截并拒绝。



1 代理计算本地梯度 ∇L_{local}，包含所有更新方向。

2 价值闸门Fi根据价值准则，过滤梯度。



3 只有可采纳梯度用于训练，并参与奖励分配；违规更新不被训练，也不分钱！

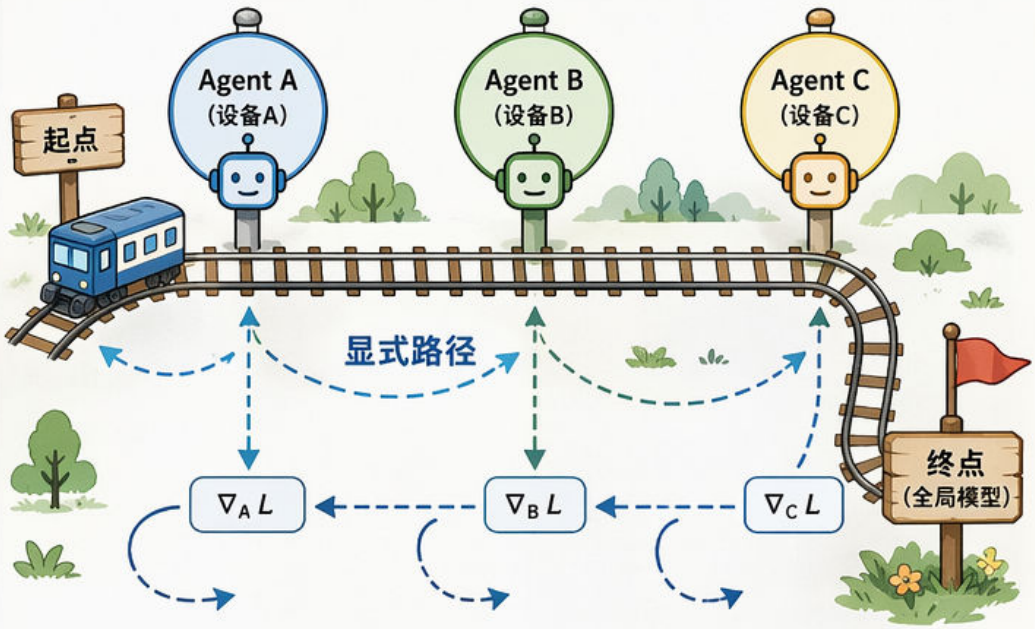
✓ 可采纳梯度
→ 用于训练
→ 参与分钱

✗ 违规更新
→ 不训练
→ 不分钱

违规更新不能训练，也不能分钱

为什么用TL？

用TL：显式路径的遍历学习



遍历学习：前进+回传，路径清晰可追踪

显式路径
学习路径看得见，
每一步去向明确。

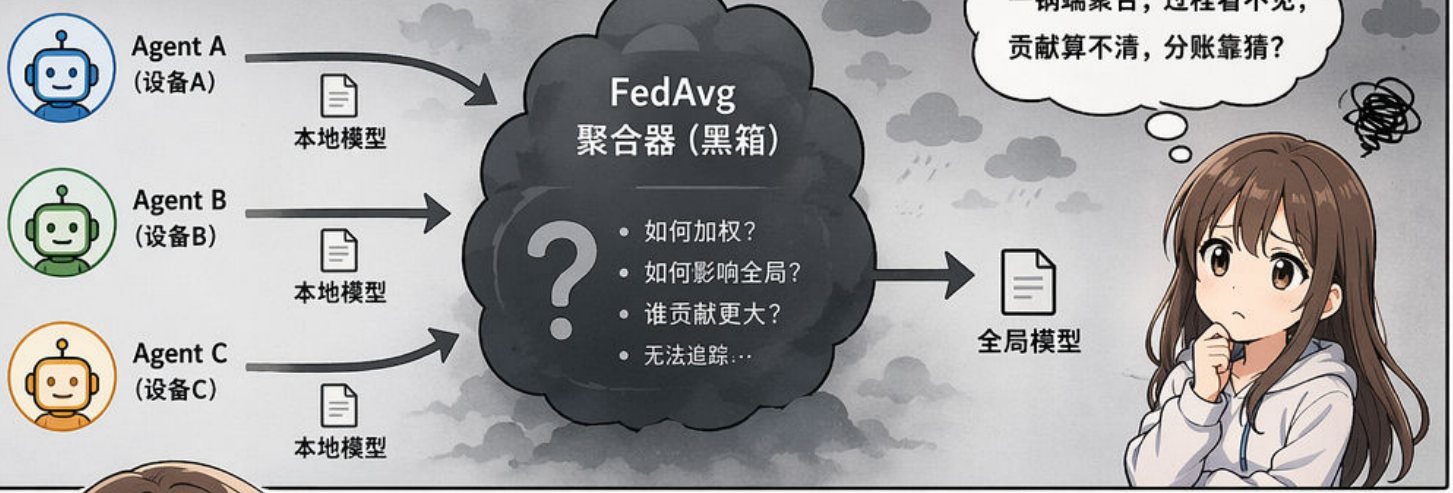
梯度流
梯度逐跳回传，
来源清晰可追踪。

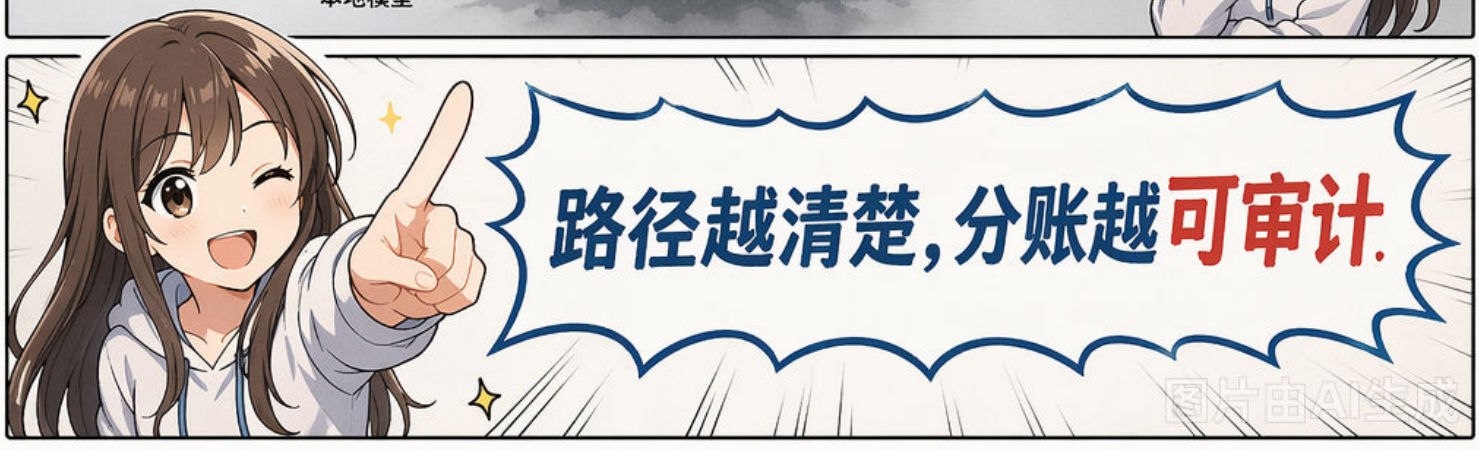
验证改善
梯度可验证效果，
及时发现并改进。

贡献归因
每个节点的贡献
可量化、可分账。

不是黑箱聚合
拒绝一锅端，
透明可审计！

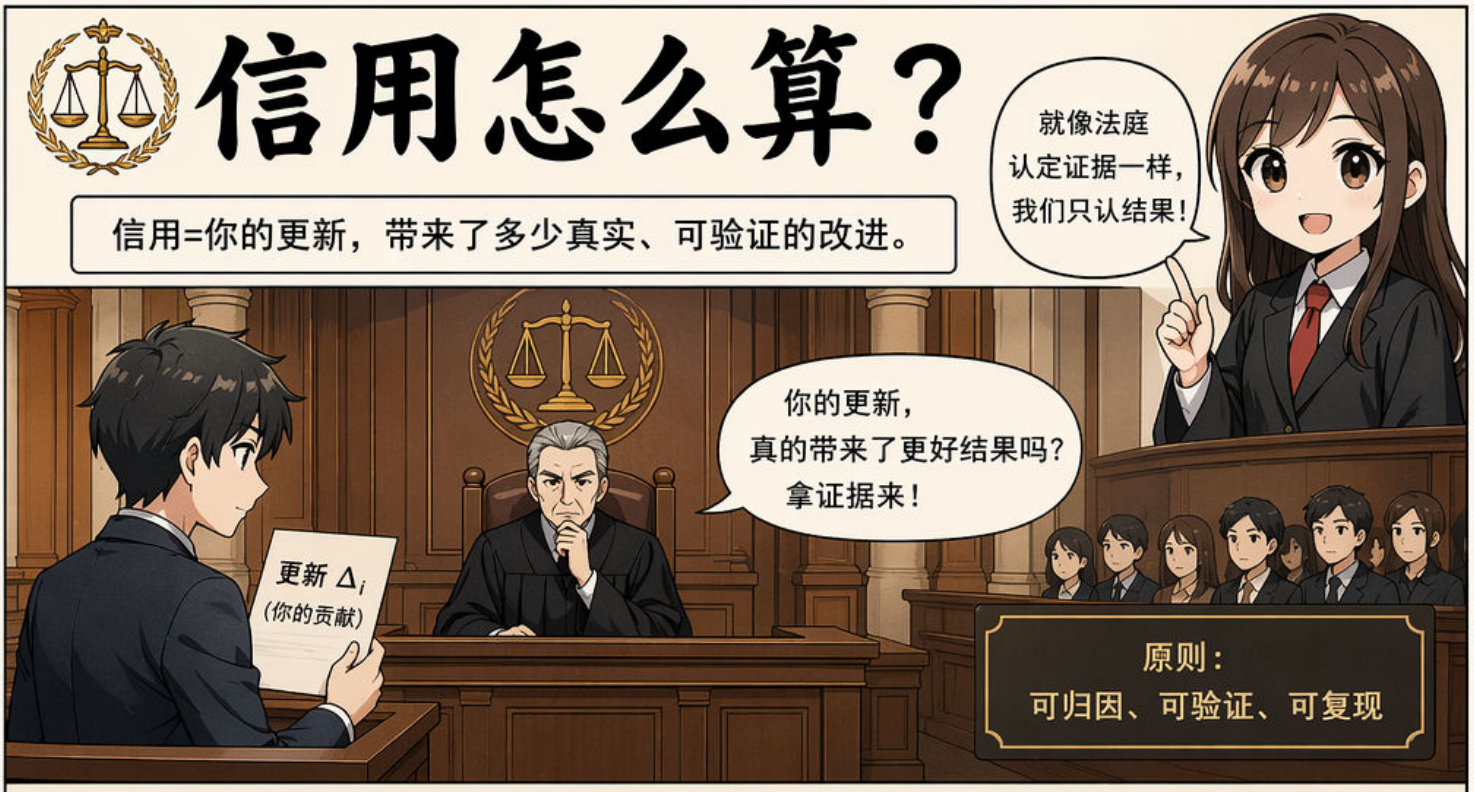
不用TL：FedAvg 黑箱聚合





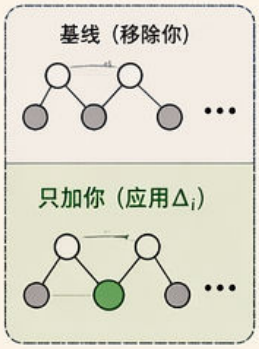
信用怎么算？

信用=你的更新，带来了多少真实、可验证的改进。



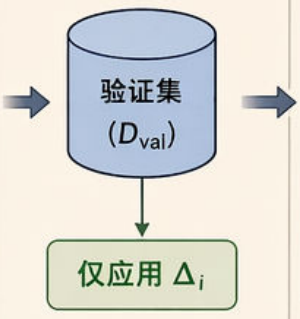
1 反事实测试

构造反事实基线：
移除你的更新 Δ_i，
其余保持不变。



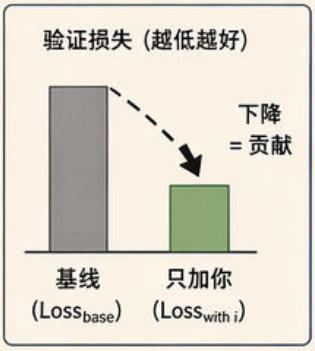
2 只应用你的更新

在验证集上，
只应用 Δ_i 的影响，
不叠加他人更新。



3 验证损失下降

比较验证损失变化：
$\Delta Loss_i = Loss_{base} - Loss_{with\ i}$
> 0 才算有效贡献。



下降 > 0 有效 ✓

下降 ≤ 0 无效 ×

4 累计贡献 C_i

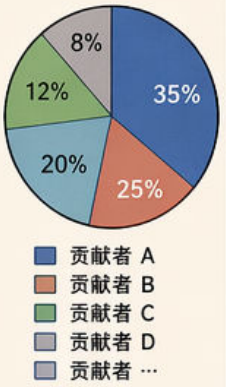
对多次任务/轮次，
累计你的有效贡献。

贡献累计示例	
任务/轮次	$\Delta Loss_i$
1	0.032
2	0.041
3	0.018
…	…
T	0.027

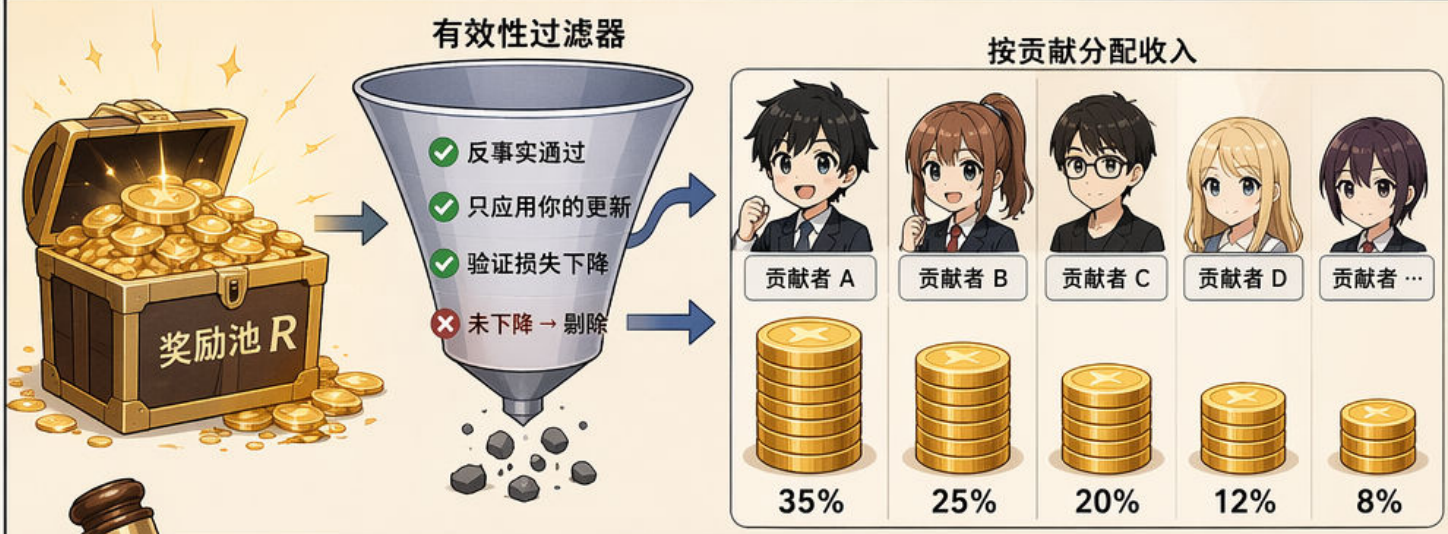
$$C_i = \sum_{t=1}^{T} \max(\Delta Loss_{i,t}, 0)$$

5 收入支付 p_i

按贡献占比分配
奖励池。



$$p_i = R \frac{C_i}{\sum_j C_j}$$



奖励只绑定合价值且有效的贡献

三条反驳

我们预判了三条常见质疑，逐一回应！

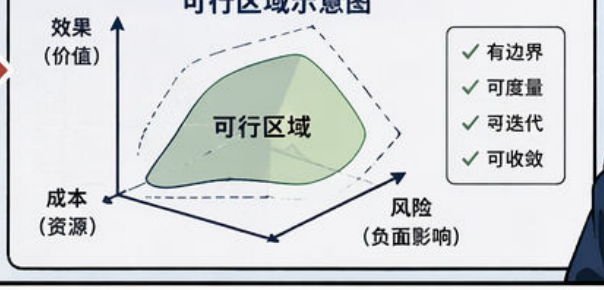
1 会收敛吗？

你们的框架会不会越用越复杂，反而无法收敛？

风险警示
- 规则膨胀
- 指标过多
- 成本上升
- 难以闭环

回应：可行区域

我们定义了“可行区域”，在效果、成本、风险三者之间寻找平衡，确保可持续收敛。



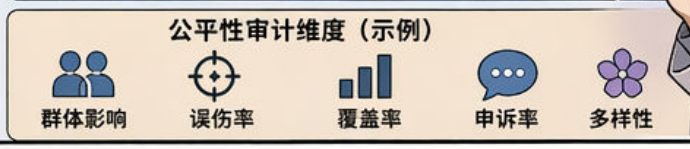
2 过滤公平吗？

过滤器会不会带来偏见，伤害某些群体或观点？

风险警示
- 数据偏见
- 价值偏见
- 群体伤害
- 回音室效应

回应：审计过滤器

过滤器必须透明、可审计、可解释。我们提供审计账本，持续监测并纠偏。



3 隐私安全吗？

你们记录这么多数据，用户隐私会不会泄露，被滥用？

风险警示
- 数据泄露
- 身份关联
- 目的外使用
- 内部滥用

回应：隐私保护设计

最小化、去标识化、加密隔离、可控访问。隐私在设计时就被内建，不是事后补救。

用户可查看、导出、删除自己的数据与记录

这是责任账本，不是万能答案

✓ 承认边界　✓ 接受监督　✓ 持续进化　✓ 与你共建

责任链

1

人：授权边界

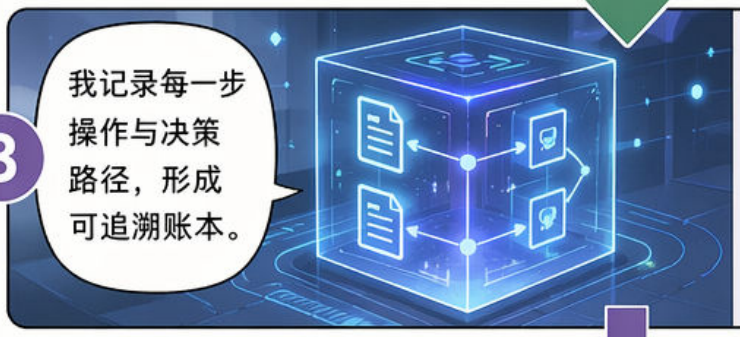
设定目标、权限和约束，明确代理可以做什么，不可以做什么。

2

代理：过滤更新

接收输入，过滤噪声与无关信息，只向模型提交有效更新与建议。

3



合作社：记录路径

记录代理与模型的交互、决策与输出，形成不可篡改的路径和证据链。

4

市场：按贡献付费

根据各方贡献（人、代理、模型、数据等）分配收益，按贡献付费，激励协作。

5

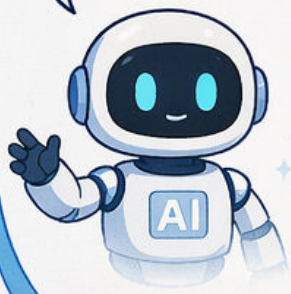
治理者：审计例外

独立审计异常与例外情况，追责与纠偏，保障系统公正、合规与可信赖。

AI替人行动，账本必须跟得上

最终脑图

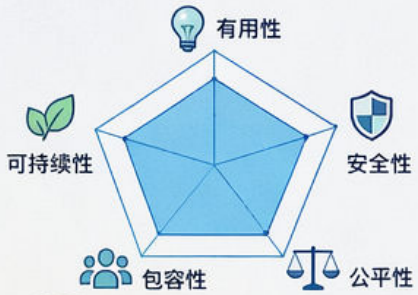
这五个模块共同构成
一个可持续、可进化、可追责的
价值对齐与信用分配体系！



价值约束 信用分配

1 多元价值

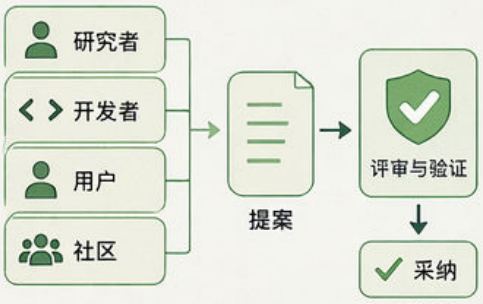
从多个维度定义“好”，
形成可计算的价值空间。



不同场景，不同
权重，价值多元
才更贴近真实世界！

2 可采纳更新

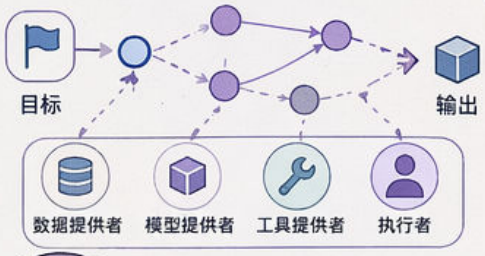
谁可以更新？如何被采纳？
规则透明，过程可验证。



开放参与，严格审查，
让系统持续进化，
又不被滥用。

3 路径归因

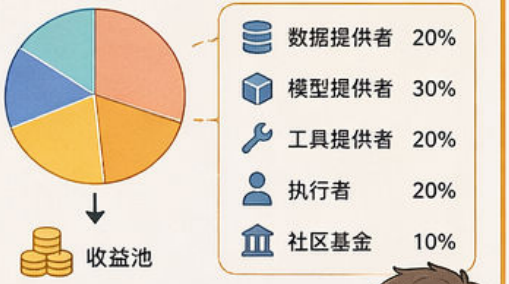
从目标到输出，拆解贡献路径，
识别每个参与者的实际贡献。



不只是结果导向，
更要看清“谁在中间
做了什么”。

4 信用结算

将贡献转化为信用，按规则分配
收益与权益，激励正向循环。



贡献有回报，付出有价值，
让更多人愿意持续参与！

5 责任审计

全流程可记录、可追溯、可审计，出现问题有人负责，系统持续改进。

有审计，才有信任；
有责任，才有进步。

对齐不只写原则，也写进谁能更新、谁能获利