大语言模型

从预测语言，到在世界中行动

清华计算机系科协暑培

景亿

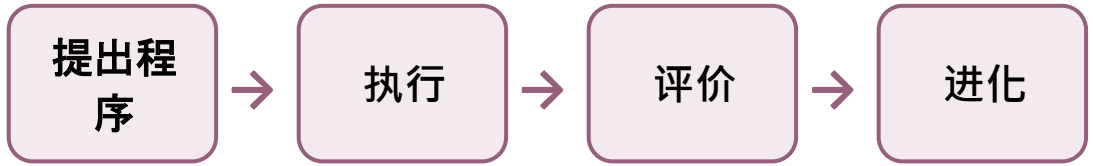
2026.8.5

一个近期新闻：LLM 写进了下一代 TPU 的硅片

2026.05.07 · DeepMind 发布一周年影响回顾

AlphaEvolve 生成的电路设计被集成进下一代 TPU

语言模型提出程序，自动执行与评分，进化搜索再把高分方案送入真实系统。



−20%

Spanner 写放大
= 底层写入量 / 用户写入量
相对原启发式 · **越低越好**

14% → 88%+

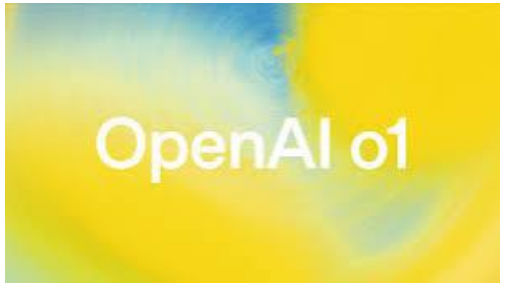
AC 最优潮流：可行解率
= GNN 输出满足电网约束的比例
越高越好

两组数字来自两个独立生产 / 研究案例，不是同一 benchmark。

真正把语言模型变成“发现机器”的是闭合的反馈回路。

三个节点，本质是反馈闭环逐步外移

ChatGPT · 2022
文本 → **人**类反馈
对话成为接口



o1 · 2024
人类 → 验证器
推理成为搜索

Claude Code · 2025
验证器 → 环境
行动产生后果

第一次学习：让人类留下的文本充当老师

密集监督：文本里的每个位置，都给出下一个 Token **的**标签。
压缩压力：为了在不同上下文中持续猜对，模型学习可复用的结构。
关键边界：目标奖励**“像数据”，并不直接**奖励**“在世界中**为真”。

机器首先必须决定：什么算一个“词”？

同一段字符可以被切成完全不同的计算单元（示意）

$$x \rightarrow (t_1, \ldots, t_T) \rightarrow \text{integer IDs}$$

Token 是词表大小、压缩率与计算成本之间的工程折中，不是语言学真理。

Token 本身没有意义，意义来自上下文中的关系

$$h_i^{(0)} = E[t_i] + p_i$$

$$h_i = f\theta(h_1,\ldots,h_T)$$

苹果 发布了新芯片 → **公司 / 产品**

桌上有一个 苹果 → **水果 / 食物**

Embedding 是入口；真正的“含义”是经过上下文计算后的状态。

同一个 Token 在不同上下文中会落到不同的表示状态。

Attention：每个 Token 都在决定向谁取信息

$$\text{Attention}(Q,K,V) = \text{softmax}((QK^T + M) / \sqrt{(d_k)})\, V$$

小明没把 奖杯 放进箱子，因为 它 太大了。

“它” ← 取回关于“奖杯”的信息

Q · Query

当前位置
想寻找什么？

K · Key

每个位置
提供什么索引？

V · Value

真正被聚合的
信息是什么？

Attention 更像内容寻址与动态路由；权重本身不等于忠实解释。

Transformer 把许多小计算叠成一个临时程序

$$h' = h + \mathrm{Attn}(\mathrm{LN}(h))$$

$$h^{+} = h' + \mathrm{MLP}(\mathrm{LN}(h'))$$

Attention

跨位置搬运信息
把相关上下文写入当前状态

MLP

逐位置变换特征
把已有组合映射成新特征

Residual stream

保留并叠加状态
让许多层协同组成计算

知识分布在特征、残差流与回路中。

“预测下一个 Token”能带来全部的智能吗？

$$p(x) = \Pi_t \, p\theta(x_t \mid x_{<t})$$

$$L\_pre = -\Sigma_t \log p\theta(x_t \mid x_{<t})$$

L_pre 是验证文本上逐 Token 的负对数似然：越低表示越会预测真实下一个 Token；它不直接等于事实正确率或任务完成率。

语法	事实与程序	人的行为
哪些词能合法组合	世界知识与代码模式	解释、论证、计划与社会规范

目标是局部的；为了持续猜对，模型必须压缩产生文本的潜在结构。

训练数据是人类认知的压缩

书籍与论文

事实与概念
论证结构
学科范式

网页与对话

社会规范
偏好与冲突
口语互动

代码与工具轨迹

程序模式
接口约定
可执行过程

继承：知识 · 方法 · 规范

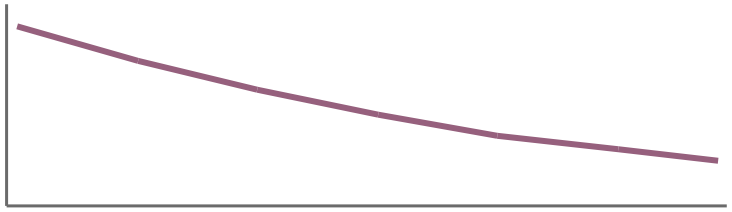
也继承：偏见 · 隐私 · 错误

数据是人类认知过程的投影，需要经过进一步审校。

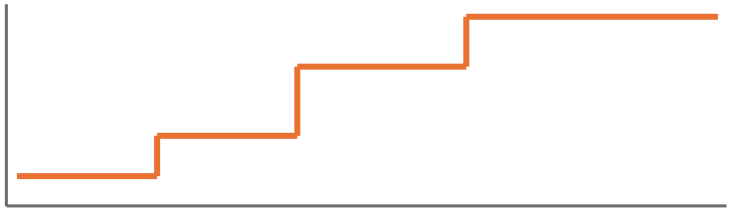
“涌现”：先有共享表示，再有可见能力

世界与行为留下数据投影 → 密集的预测压力 → 可复用的共享表示 → 算法与 in-context learning → 任务指标上出现能力

底层 loss 往往平滑改善



离散指标会制造“悬崖”



概念示意（非实测）｜横轴：模型规模 / 训练计算 →｜左：验证 loss 越低越好｜右：离散任务得分越高越好

若数据由隐状态产生，重建结构常能提升预测效率；但“涌现”也可能被指标阈值放大。

Scaling law 把“智能”变成了可预算的工程问题

$$L(N,D) = E + A/N^{\alpha} + B/D^{\beta}$$

N · 参数量 → $C \approx 6ND$ → **D · 训练 Token**

模型太小：容量不足

数据太少：模型欠训练

固定算力下要共同配比

L 是验证集逐 Token 交叉熵（越低越好）；C≈6ND 是稠密 Transformer 训练 FLOPs 的数量级近似，系数 6 来自前向与反向计算，并非跨架构常数。

Scaling law 能预测 loss，**却不能保**证真实性、安全性，或某项能力何时出现。

预训练交付的是巨大潜能，**尚未成**为**一个可靠助手**

已经拥有

流畅语言
广泛知识
代码与模式
in-context learning

尚未拥有

稳定指令遵循
诚实与校准
任务边界
可靠拒绝

同时继承

幻觉倾向
偏见与迎合
数据污染
隐私风险

预训练教的是“**人**类通常会写什么”，**我**们还需要教会“**系**统现在应该做什么”。

多模态模型：图像先被编码成连续视觉向量

$$V = P(E_v(I)) \in \mathbb{R}^{n \times d}$$

$$p(y|I,x) = \Pi_t \, p\theta(y_t \mid V, x, y_{<t})$$

n 是一张图压缩后送入 LLM 的视觉向量数，d 是每个向量的维度；n×d 表示张量形状，不是训练样本量。

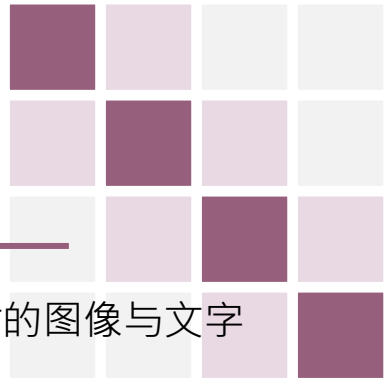
连续 visual embeddings ≠ BPE Token ID；接口丢掉的信息，LLM 无法凭空恢复。

多模态的第一步是建立一种能把感知接入语言计算的表示接口。

共享语义空间如何学出来：配对数据把图像与语言拉到一起

$$s_{ij} = \cos(v_i, t_j) / \tau$$

4×4 mini-batch 示意



对角线：配对图文

$$L\_align(i) = -\log \left[\exp(s_{ii}) / \Sigma_j \exp(s_{ij}) \right]$$

拉近

真实配对的图像与文字

推远

同一 batch 中的不配对样本

获得

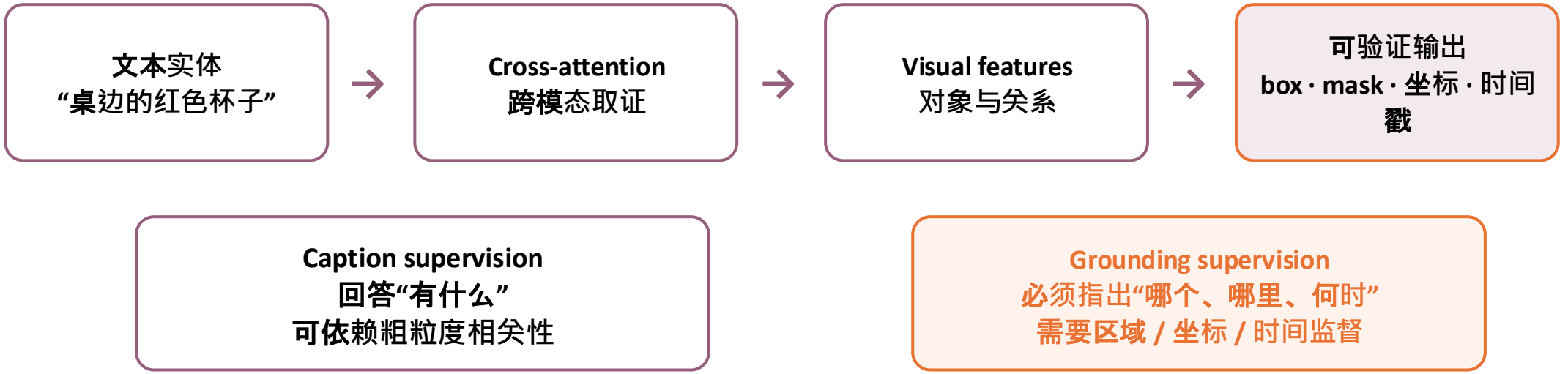
可检索、可零样本迁移的语义空间

s_{ij} 越大表示图文越相似；τ 控制差异放大；L_align 越低越好。CLIP 实际双向优化。

对齐学到的是“数据里哪些图文被配成一对”；caption 的粒度、偏见与缺失也会进入模型。

会描述不等于被世界约束：Grounding 要落到对象、位置与时间

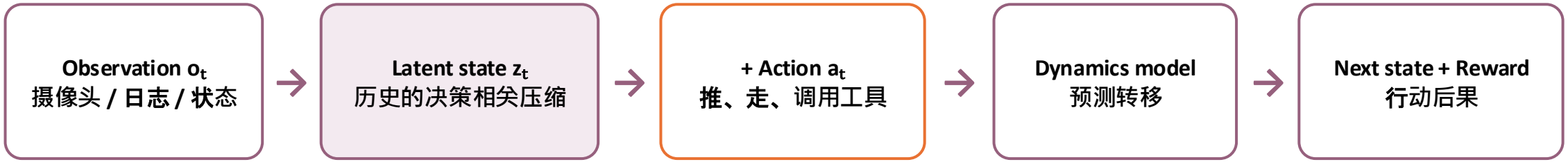
$$A = \mathrm{softmax}(Q_{text} K_{vision}^{T} / \sqrt{d})$$



流畅描述可以来自语言先验；一旦模型要操作现实对象，Grounding 必须落到可独立验证的证据。

世界模型预测行动如何改变状态

$$z_t = \mathrm{Enc}(o_{\le t}, a_{<t}) \qquad \hat{z}_{t+1} \sim p_\theta(\cdot \mid z_t, a_t), \qquad \hat{r}_{t+1} = r_\theta(z_t, a_t)$$



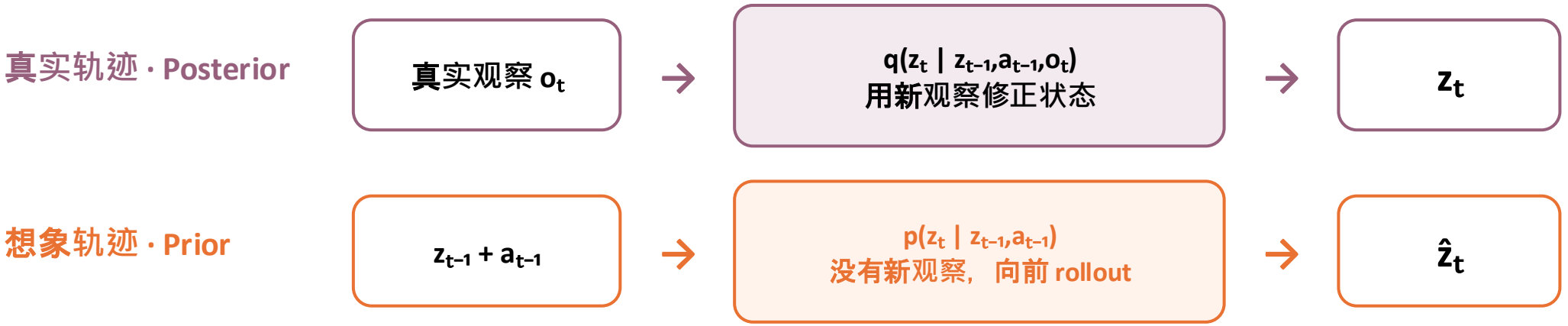
VLM：眼前是什么？　World Model：如果采取另一个动作，接下来会怎样？

$\hat{r}_{t+1}$ 是世界模型预测的任务奖励；帽号表示“想象值”，不是环境真实回报，评测时仍须与真实转移和结果比较。

世界模型的关键增量是 action conditioning；视频重建可以有，但不是定义所必需。

世界模型不必记住每个像素，只需保留足以预测与控制的状态

$$L\_WM = L\_obs + L\_reward + \beta\, KL(q \,\|\, p)$$



L_obs 衡量观察预测误差，L_reward 衡量奖励预测误差，KL 约束想象先验贴近看见新观察后的后验；三项都按越低越好优化，β 只是权重。

纹理可以丢；影响未来、Reward 与可控性的变量不能丢。

好的 latent state 不是当前帧 embedding，而是关于未来的、与决策相关的历史摘要。

ChatGPT：预测器第一次成为大众可用的人机接口

2022.11.30：OpenAI **以研究**预览形式公开 ChatGPT；这里指大众产品节点，不是模型**第一次学会**对话。

关键变化：指令示范、回答比较与 RLHF **共同塑造助手行**为。

不是新知识库：后训练主要让模型知道何时、以何种方式使用已有能力。

大模型经历两次学习：先扩大能力空间，再重新分配行为概率

第一次学习 · Pretraining

数据：海量人类文本
标签：下一个 Token
学到：语言、知识、程序、行为模式

第二次学习 · Post-training

数据：指令示范、回答比较、人类反馈
代理目标：SFT、Reward Model、RLHF
学到：遵循指令、帮助性、基本安全边界

第一次学习：能做什么 → 第二次学习：更可能怎样做

Agent 时代，第二次学习会从“评价回答”继续扩展到“评价整条工作轨迹”。

SFT 用示范告诉模型：一个好助手怎样回应

$$L\_SFT = -E(x,y^*)\ \Sigma_t \log \pi\theta(y^*_t \mid x,\ y^*_{<t})$$

指令 x

“**用三句**话解释为什么 Attention **有用**。”

→

高质量示范 y*

格式、语气、步骤与基本策略，都通过**示范**进入行为分布。

SFT **能教**“像谁做”，**但不会自**动指出哪些错误最严重。

Reward Model：把人类的成对比较学成一个标量代理

图中 A / B 仅示意一次成对标注；实际由同一提示下的多个候选及其排序构成

回答 A
事实准确 · 简洁 · 有边界

回答 B
流畅自信 · 但含无依据细节

人类选择 A ≻ B

$$L_{RM}(\varphi) = -\,E \log \sigma(r_\varphi(x, y_w) - r_\varphi(x, y_l))$$

r_ϕ 越高表示越符合这批标注者在该任务分布上的偏好；Reward Model 不是普遍真理的分数。

人类价值进入模型前，必须先被翻译成数据；这次翻译本身会丢失信息。

RLHF：优化人类偏好训练出的 Reward，同时锚定 SFT 策略

$$J\_RLHF(\theta) = \mathbb{E}[r\varphi(x,y)] - \beta\, KL(\pi\theta \,\|\, \pi SFT), \quad y \sim \pi\theta$$

Reward Model

近似人类偏好
定义策略优化的地形

Policy Optimization

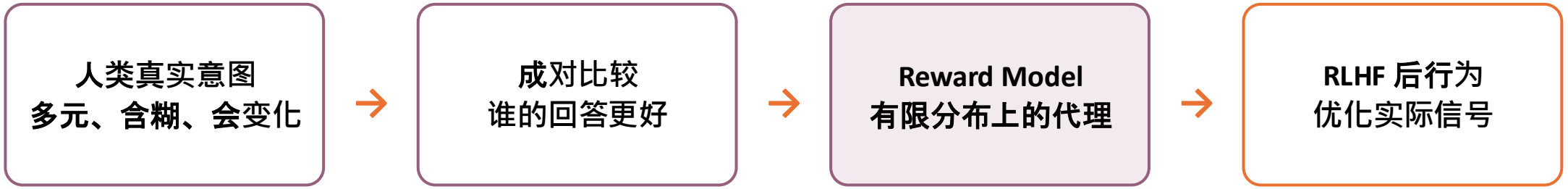
采样并提高高奖励回答概率
PPO 是一种实现

KL Anchor

不让策略偏离 SFT 太远
保留可用语言行为

RLHF 不是把真理注入模型，而是让策略在 Reward Model 定义的地形上移动。

RLHF 的边界：人类意图经过三次翻译才成为模型行为

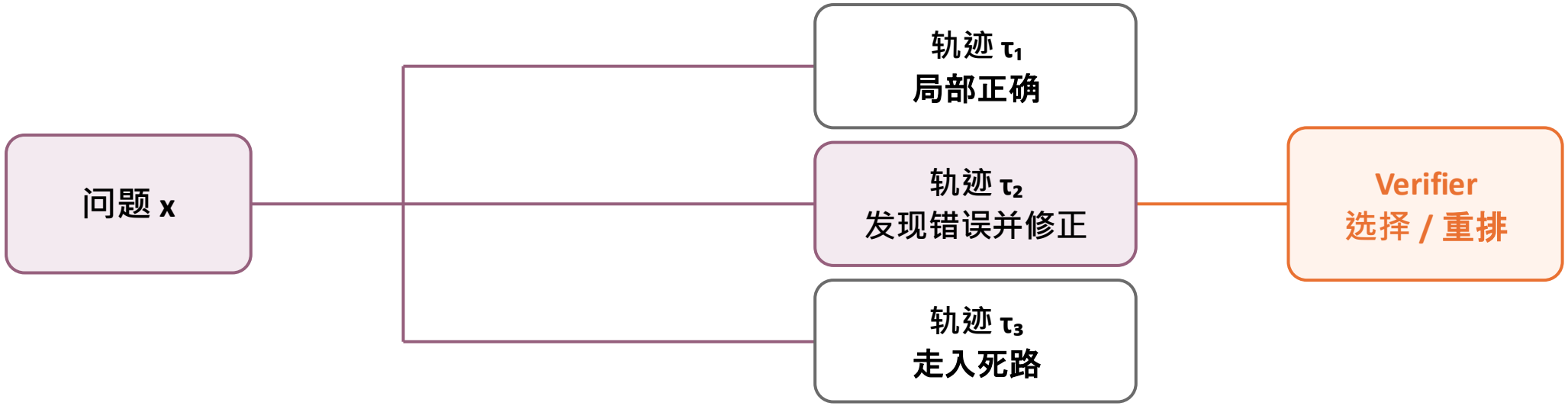


RLHF 对齐的是特定标注者在特定分布上的比较，不自动等于“普遍人类价值”。

o1：模型开始在回答之前主动分配计算

核心变化：同一组参数，可以按任务难度动态投入推理计算。
计算怎么花：更长的轨迹、更多候选，以及 Verifier 的筛选。
新的瓶颈：额外候选只有经过可靠验证，才会变成更高正确率。

推理的本质是在候选轨迹中搜索



$$\hat{y} = \text{argmax}_i \, V(x, y_i), \qquad y_i \sim \pi\theta(\cdot|x)$$

τ₁ / τ₂ / τ₃ 只是候选类型示意，并非固定采样 3 条；真实系统可采样任意 K 条再评分。

Reasoning model 的核心不是输出更多tokens，而是产生、检查、修正并选择轨迹。

Test-time scaling：**想更久**、试更多次、验证再选择

想更久

更长的串行轨迹
展开、反思、改写

风险：**冗**长不等于正确

试更多次

并行采样多个候选
增加至少一次成功的机会

风险：错误可能高度相关

验证再选择

搜索、评分、重排、停止
让算力投入更有方向

瓶颈：Verifier 质量

$$pass@K \approx 1 - (1 - p)^K$$

仅在各次独立、且单次成功率相同的近似下成立

p 是单次成功率；pass@K 是 K 个候选中至少一个正确的概率（越高越好），衡量候选覆盖率，不等于最终选中答案后的准确率。

多采样只有在错误足够多样、**且**选择器可靠时，**才真正**转化为能力。

当生成答案变便宜，Verifier 会成为新的瓶颈

AIME 2024 | 每套 15 道数学题 | 下列均为同一 o1 的不同推理配置 | 答对率越高越好

74%

单次作答
≈ 11.1 / 15

83%

每题 64 次
多数投票 · ≈ 12.5 / 15

93%

每题 1000 次
学习评分器重排 · ≈ 13.9 / 15

额外推理算力 → proposal diversity × verifier quality

生成一千个候选并不等于智能；稀缺的是能识别正确候选的可信评价器。

思维链不是模型的“解释书”，信任仍要落到可验证证据

内部轨迹 ≠ 忠实解释：模型写出的思维链可能省略、合理化或受提示影响。
外部证据更重要：引用、计算、单元测试、实验结果可以独立复核。
可解释性扩展：对 Agent，还要能重放状态、工具调用、权限与失败恢复。

Claude Code：答案系统性地变成可执行工作

答案 → 工作产物：模型开始读取仓库、修改文件、运行测试并交付 diff。
能力变成系统属性：Model、Tool、Skill、Environment **与 Harness 共同决定**结果。
安全进入执行层：权限、沙箱、预算、停止与回滚决定最坏后果。

Agent 是一套受控的工作系统

Agent = Model + Tools + Skills + Environment + Harness

Model	Tools	Skills	Environment	Harness
提出下一步	**执行动作**	**复用经验**	**保存状态**	**组织与约束**

同一模型放进不同系统，可能表现得像完全不同的 Agent。

能力不再只位于参数中，而是分布在模型、计算、环境与评价之间。

Tool 提供动作原语，Skill 压缩可复用的工作经验

Tool · 能做什么

search
read / write
run code
call API
operate browser

Skill · 应该怎样做

调试流程
代码审查规范
实验模板
领域知识
脚本与资源

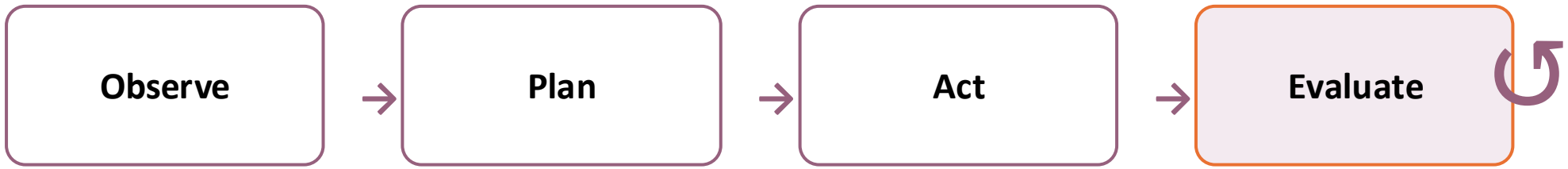
Task · 此刻为何做

具体目标
当前状态
用户约束
完成条件
风险边界

Harness 决定：何时加载 Skill、允许调用哪些 Tool、如何验证与停止。

Tool 扩大行动半径；Skill 压缩程序化经验；两者也同时扩大攻击面。

Harness **决定上下文、**预算、权限、检查点与失败恢复

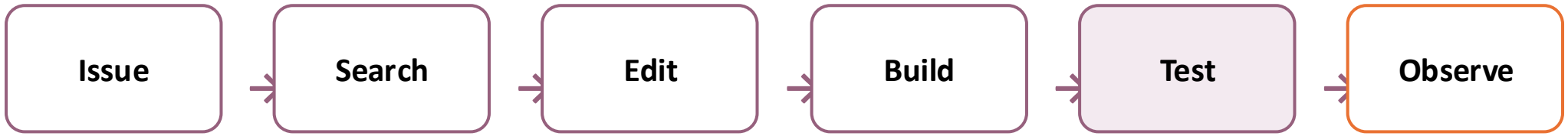


Context Permission State Budget Retry Logs **Stop** **Rollback**

每个 Harness 组件，都编码了一个**“模型目前**还做不到什么**”的假**设。

Harness **把概率模型**变成可运营系统；**模型升**级后，**也要重新消融旧假**设。

编码 Agent 的秘密：环境能在较短循环内提供真实反馈



失败信息 ↺ 回到 Search / Edit

反馈密集

较短循环内获得编译或单测结果
具体时延取决于仓库与构建系统

反馈可验证

错误码、单测、diff 可复核

反馈可回滚

版本控制限制失败后果

代码 Agent 的优势不只在训练数据，更在环境持续给出便宜、客观、可回滚的反馈。

长链路可靠性会指数下降，所以系统必须会停、查、退

$$P(\text{task success}) \approx p\_step^{n}$$

$$0.98^{100} \approx 13.3\%$$

示意计算（非实测）：只有在 100 个步骤都不可或缺、相互独立，且每步成功率均为 98% 时，端到端全部成功率才是 $0.98^{100} \approx 13.3\%$。

真实步骤会相关，系统也可检查、重试与回滚；公式只是警报器。

Checkpoint · Unit test · Least privilege · Human gate · Stop · Rollback

---

Agent 的安全、评测与可靠性不是外围治理，而是闭环本身的组成部分。

Agent 时代如何训练模型：先把开放任务变成可评价的轨迹

$$R(\tau) = w^{\top} g(\tau) \qquad \text{Safety}(\tau) = 1$$

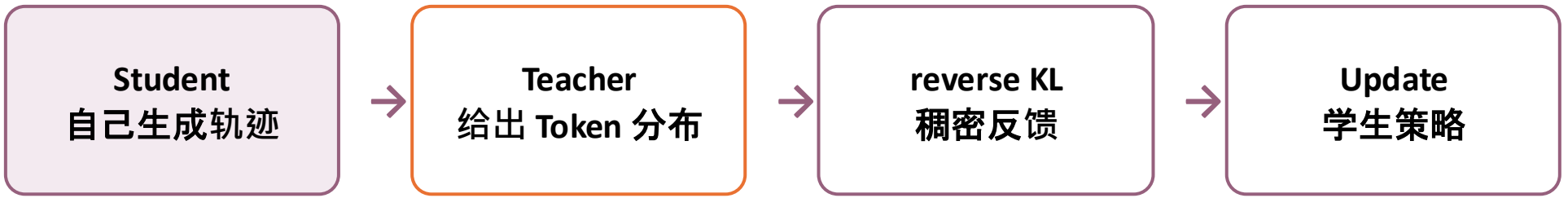
Intent → Rubric → Grader → Reward

Safety / Compliance HARD GATE 不可被平均分抵消

条形仅示意五类轨迹评分维度，并非 benchmark 实测；Safety(τ)=1 表示预设安全 / 合规硬门槛全部通过，不是“安全性 100%”。

Agent 训练的首要瓶颈不是优化算法，而是真实、可扩展且不易被投机的轨迹反馈。

OPD：在模型自己会到达的状态上获得稠密教师反馈



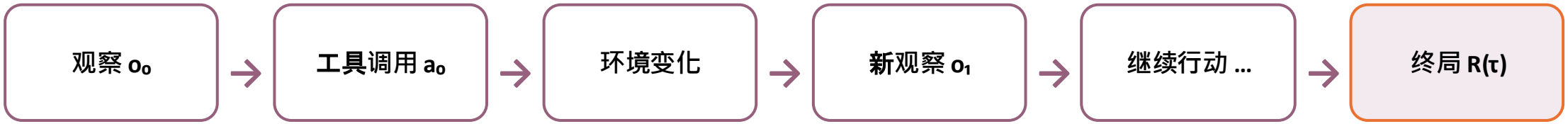
$$L\_OPD = \mathbb{E}[\ KL(\pi\theta(\cdot|s_t) \Vert \pi\tau(\cdot|s_t))\ ], \quad s_t \sim d(\pi\theta)$$

OPD：匹配 Teacher policy（稠密） ≠ Agentic RL：优化环境 / Grader 的 Reward（常稀疏）

OPD 在学生真正到达的状态上纠偏。

Agentic RL：训练单位从回答变成与环境交互的整段轨迹

$$\tau = (o_0, a_0, o_1, \ldots, a_t, o_{t+1}), \quad J_{ARL}(\theta) = \mathbb{E}[R(\tau)], \quad \tau \sim (\pi\theta, Env)$$



长程 Credit assignment
把最终成败归因回关键动作

稀疏、延迟 Reward
整段轨迹结束后才知道成败

环境与 Grader
可扩展、可复现、难以投机

Agentic RL 的关键是让环境后果真正更新策略；只模仿历史工具轨迹仍然是 SFT。

高价值领域的护城河，是可被训练系统捕捉的专家判断

2026.06 · Thinking Machines × Bridgewater | 六个投资人日常的信息筛选 / 文档截断任务 | 独立留出测试集

78.2%

最佳通用前沿模型 + 专家 prompt
六任务平均准确率 · **越高越好**

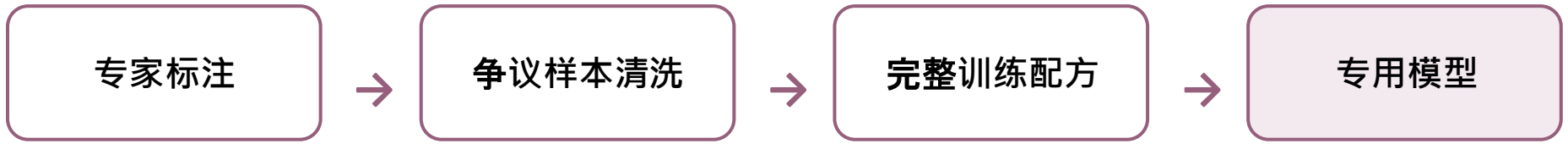
→

84.7%

领域专用模型
六任务平均准确率 · **越高越好**

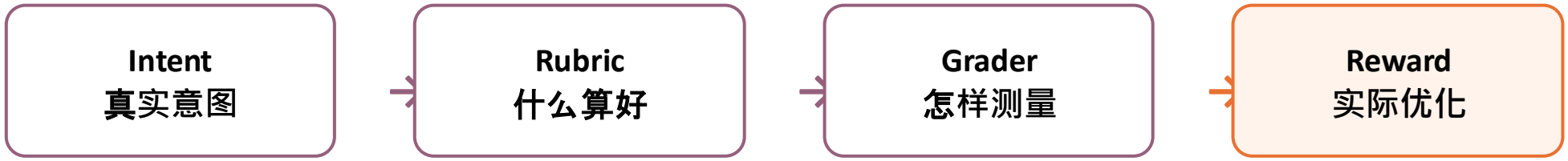
13.8×

完成同一任务的推理成本
约为对比通用模型的 1 / 13.8



通用模型不一定拥有组织特有的 taste；高质量专家反馈会产生“**差异化智能**”。

开放专业任务先建设 Evaluator：把专家判断编译成 Rubric 与 Grader



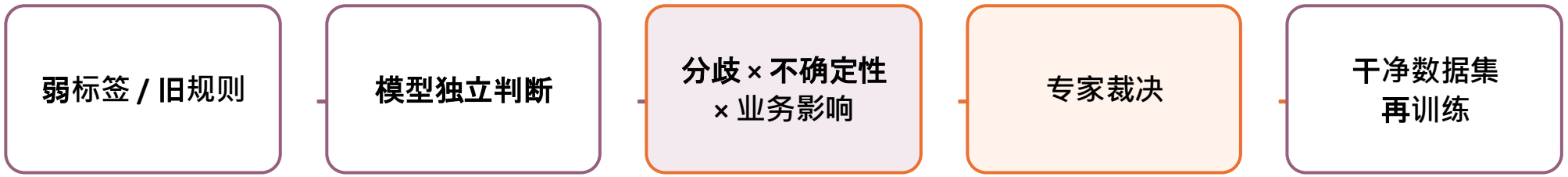
$$g(y) = [\text{correct}, \text{grounded}, \text{complete}, \text{safe}] \qquad R(y) = \mathbb{1}[\text{hard gates}] \cdot w^{\top} g(y)$$

HealthBench 2025 | 真实风格医疗对话
262 名医生为 5,000 段对话编写 48,562 条题目专属标准
GPT-4.1 逐条判定；总分越高 = 满足的加权医学标准越多

GDPval 2025 | 9 行业 · 44 职业 · 1,320 项真实交付任务
同职业专家盲评 AI 与人类产物
自动 Grader 仅为实验性辅助，不能替代专家

先建设被专家验证的 Evaluator，再扩大训练；错误的 Reward 只会让系统更快学会错误标准。

不要让专家标遍所有数据：在分歧边界上购买专家时间



$$\text{priority}(x) = \text{disagreement}(x) \times \text{uncertainty}(x) \times \text{impact}(x)$$

Bridgewater 的训练飞轮

44.8 → 73.5 → 84.7

Qwen3-235B 基座　GRPO　完整配方[1]

六任务留出集平均准确率 · 越高越好 | 最佳通用前沿 + 专家 prompt：78.2

[1] interleaved batching + CISPO + OPD

专家只清洗训练集中的模型—标签分歧样本；最终分数来自独立留出集。

专业数据的关键不是数量最大，而是让专家时间集中在最能改变决策边界的样本上。

高价值 Agent 需要分层反馈，而不是直接追逐终局 KPI

$$R(\tau) = \alpha R\_process + \beta R\_expert + \gamma R\_outcome - \lambda \cdot Cost$$

subject to Compliance(τ) = 1 | 二元硬门槛：任一强制规则失败都不能被其他高分抵消；不是“合规率 100%”

过程反馈 · 密集

引用是否可追溯
计算是否可复现
规则是否满足

便宜、即时、可审计

专家判断 · 稀缺

证据是否充分
例外是否合理
建议是否可行动

昂贵、语境化、可争议

终局结果 · 真实

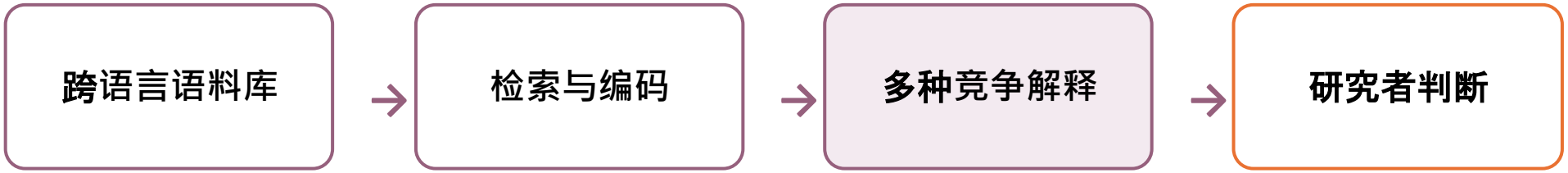
预测是否发生
治疗是否改善
决策是否创造价值

延迟、**混**杂、难归因

Mantic 2026 | 约 1 万道“事件是否会在截止日前发生”的二元预测题：检索 → 给出概率 p → 事后真实结果 y∈{0,1}
Brier 奖励 $R = 1 - (p - y)^2$；越接近 1，概率预测越准确

终局结果最真实，却常常太晚、太稀疏；可靠系统用局部可验证过程训练，用真实结果校准与问责。

人文社科：扩大可比较的证据空间，而不是替代解释



Provenance rail：每个主张都保留原文、出处、版本与不确定性

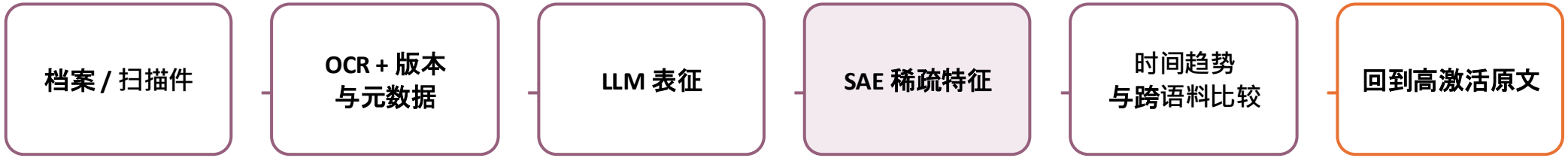
语境 · 语言变化与历史制度

价值冲突 · 不能被一个 reward 消解

史料批判 · 沉默、缺失与代表性偏差

模型能扩展阅读、比较与假设生成；解释责任仍属于研究者。

人文学案例：从档案证据到概念演化，再回到原文



EvidenceUnit
原文 · 文献 ID · 页码
作者 · 时间 · 地点 · 文类
原始语言 · OCR 置信度

$$z_i = \mathrm{TopK}(W\_enc\ h_i + b)$$

$$\mu_{k,t} = \mathbb{E}[z_{i,k} \mid \mathrm{time}(i)=t]$$

HistLens：在《新青年》《向导》等语料中追踪概念内部语义构成，再把变化信号回链到原文。

时间、文类、语言与出处需要进入解释本身。

LLM 可以把阅读规模化，但不能跳过编码、反证与真实样本

规模化编码

理论概念 → Codebook

↓

多模型 / 多次独立编码

↓

一致样本 → 规模化统计

↓

分歧样本 → 人工复核

$$\kappa = (p_o - p_e) / (1 - p_e)$$

κ 扣除随机一致：κ=1 完全一致，κ≈0 仅相当于随机。

“合成人类”的边界

PNAS 2023 | 四组共 6,183 条推文 / 新闻
用同一 Codebook 标注相关性、立场、主题与框架
零样本 ChatGPT 对比 MTurk：平均准确率约 +25 个百分点
每条 < $0.003，成本约低 30×

Nature Computational Science 2025 | 3 个 LLM 模拟 156 项心理学 / 管理学情境实验
与原实验同方向的复现率（越高越常一致）：主效应 73–81% · 交互效应 46–63%
但原研究为零效应时，模型有 68–83% 的概率生成显著效应

训练数据中的“社会可想象性” ≠ 新的社会观测

LLM 适合扩大观察范围、生成假设与寻找边界；金标准、真实样本和反证程序仍不能外包。

价值冲突不能被一个 Reward 消解：权重本身就是政治与伦理选择

$$r(y) = [\ \text{factuality, harm, fairness, representation, autonomy}\ ]$$

$$R\_w(y) = w^T r(y)$$

谁参与、如何提问、怎样聚合、谁有否决权，共同决定 w。

左侧条长仅示意价值权重取舍，
不是两项研究的实测分数。

OpinionQA 2023
用 Pew 民调题比较模型与 60 个美国人口群体
代表性分数 0–1：越高越接近该群体的回答分布，
不表示价值上“更正确”

公共 Constitution 2023
约 1,000 人提交 1,127 条规则、投出 38,252 票
“约 50% 重合”指与 Anthropic 内部宪法的概念 / 价值重合，
不是模型性能分数

事实争议需要证据，价值冲突需要程序；分歧往往正是研究对象，而不是应被平均掉的噪声。

科学的终局需要让假设进入实验闭环

VLM / 仪器数据

Co-Scientist：生成、辩论与排序，再交给实验验证

AlphaEvolve：生成候选 + 自动评价 + 进化搜索

来自自然世界的测量比语言自评更难伪造；科学 Agent 的核心资产是可重复实验与不确定性校准。

Co-Scientist：模型扩大假设空间，实验负责压缩不确定性

$$P(h_i \succ h_j) = \sigma(s(h_i) - s(h_j))$$

内部 Rubric
plausible · novel · testable
impact · safe

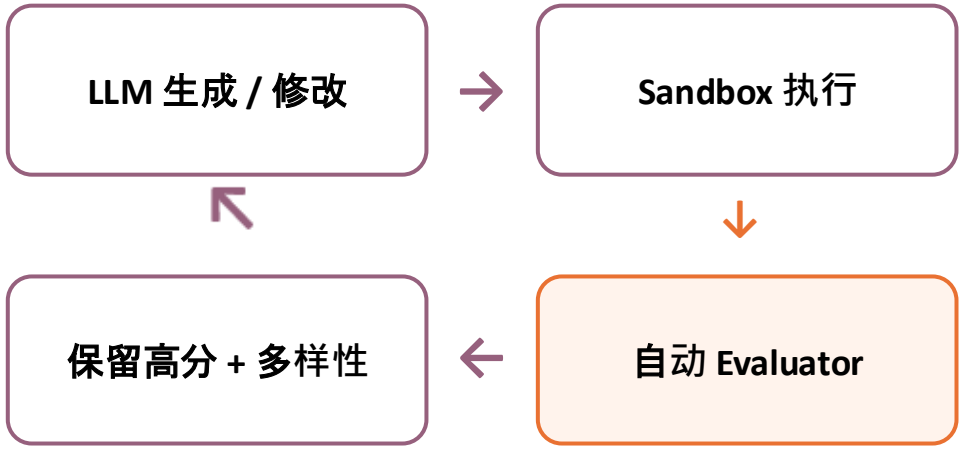
Nature 2026 的白血病药物重定位 / 组合、肝纤维化靶点与抗微生物耐药机制，均经过独立体外湿实验验证。
这里的“验证”不是内部 Tournament 得分；后者只用于候选排序。

Tournament 分数表示“系统更偏好哪个假设”，并不自动表示“自然世界认为它是真的”。

Literature Agent 的瓶颈是证据覆盖；Hypothesis Engine 的瓶颈是评价函数与区分性实验。

AlphaEvolve：当科学问题能编译成“候选程序 + 自动评价器”

进化式程序搜索



$$p^* = \arg\max_p [\, J(p) - \lambda C(p) \,]$$

AlphaEvolve 的可验证产出

49 → 48

4×4 复矩阵乘法所需的标量乘法次数
在结果完全等价的前提下低于
Strassen 递归基线
次数越少 = 算法越高效

0.7%

Google Borg 生产系统运行一年多后的平均
持续回收量
越高 = 同样硬件可承载更多任务
不是一次性 benchmark 跑分

✓ **TPU 电路修改通过严格等价验证**

适用前提：候选可执行；评价便宜、量化且难以投机。

模型负责扩大候选空间，Evaluator 负责淘汰“**听起来合理、执行后无效**”的候选。

自主实验室：最可信的 Reward 来自仪器，而不是模型自评

研究状态 D_t → 选择实验 a_t → **安全 Gate** → 仪器执行 → 测量 y_t → 更新信念

$$a_t^* = \arg\max_{a \in A\_safe} I(H;Y \mid a, D_t) - \lambda C(a)$$

选择最能区分当前假设、同时成本与风险可接受的下一次实验。

A-Lab 2023 · 17 天连续闭环
完成 353 次无机材料合成实验
57 个计算筛选目标 → 36 个经 XRD 与人工精修确认生成
目标成功率 63%；“生成”不等于高纯度成品

动作白名单 · 参数范围 · 物理 Interlock · 原始数据留存
预注册指标 · 阴阳性对照 · 独立复现 · 危险步骤人工审批

物理测量是最强反馈，也最昂贵、最嘈杂、最有风险；好的实验 Agent 优化信息增益，而非只**追逐成功率**。

语言模型到达顶点了吗？

语言模态的智能饱和了吗？

我们需要其他模态的智能吗？