METAREFINE PROTOCOL II • THEORY WHITEPAPER • V2.0

元炼协议 II

自适应推理精炼理论

ADAPTIVE INFERENCE REFINEMENT THEORY
FROM TOKEN OPTIMIZATION TO COGNITIVE RUNTIME



理论提案　版本 2.0　2026 年 8 月

单文件离线版

Donald Tong

EXECUTIVE ABSTRACT

智能不只存在于参数中，也存在于推理过程的组织方式中。

大语言模型的输出质量，在真实任务中由模型、上下文、外部状态、验证机制、推理预算与历史反馈共同决定。元炼协议 II 将“精炼”从 Token 删除提升为一个可恢复、可验证、会随模型变化而自我裁剪的推理时优化问题。

本文提出自适应推理精炼理论（Adaptive Inference Refinement, AIR）：在不修改基础模型参数的前提下，于推理边界之外构造弱耦合认知精炼层，使系统在受限成本下提高任务成功率，或在近似质量下减少有效推理资源消耗。

阅读边界

理论白皮书，不是实验报告

本文给出定义、机制、可证伪假设与实验协议。所有曲线和数值仅表达关系，不代表已测得结果。

THEORY　6 HYPOTHESES　NO CLAIMED BENCHMARK

$$\mathcal{I}_{\text{eff}} = F(M_\theta, C, E, V, B, H)$$

有效智能 = 基础模型 × 有效上下文 × 环境状态 × 验证 × 预算 × 历史反馈的系统函数

目标一

同成本，更高成功率

在总预算约束下，让信息、计算与验证被分配到边际价值最高的位置。

目标二

同质量，更低资源

移除冗余上下文与低价值推理，同时保留恢复路径和安全余量。

目标三

随模型，自我收缩

当模型吸收外部策略后，运行时重新评估并旁路已无正收益的能力。

01 • PARADIGM SHIFT

从模型能力，走向系统能力。

传统叙事把智能上限主要归因于参数、数据与训练。AIR 关注另一个可控边界：同一个模型在推理阶段，为什么会因上下文结构、验证方法与计算调度不同而表现出显著不同的有效能力？

> **精炼不是把输入变短；精炼是在有限资源下，让每一次读取、生成、检索和验证都更接近任务成功。**

传统 TOKEN 压缩	自适应推理精炼
核心对象：文本长度	核心对象：完整推理过程
主要动作：删除、摘要	动作：检索、变换、扩展、验证、停止
局部指标：压缩率	系统指标：质量、成本、延迟、风险联合效用
常见假设：压得越短越好	原则：任务充分且可恢复，才算有效
静态策略	随任务、模型、预算和反馈动态变化

核心重定义

Compression ⊂ Refinement。压缩可以是精炼动作，但精炼还包括计算再投资、外部状态接入、正交验证与动态停止。

02 • FORMAL OBJECTIVE

一个受约束的系统优化问题。

给定基础模型 M_θ、任务 x、环境状态 e 与预算 B，外部策略 π_φ 不是为了永久增加流程，而是选择当前最有价值的推理动作。

$$J(\theta, \varphi) = \mathbb{E}[Q] - \lambda_C \mathbb{E}[C] - \lambda_L \mathbb{E}[L] - \lambda_R \mathbb{E}[R]$$

Q：任务质量 · C：计算与 Token 成本 · L：延迟 · R：不可验证性、语义损失与操作风险

模式 A • 成本受限

max $\mathbb{E}[Q]$

subject to $\mathbb{E}[C] \le B_C$, $\mathbb{E}[L] \le B_L$, $\mathbb{E}[R] \le B_R$

模式 B • 质量受限

min $\mathbb{E}[C + \eta L + \mu R]$

subject to $\mathbb{E}[Q] \ge Q_{target}$

这一定义把“效果变好”和“资源变省”放进同一目标函数，也迫使系统把风险与可验证性计入成本。它允许不同应用通过 λ 系数表达不同偏好，而不要求所有任务共享单一策略。

03 • SIX PRINCIPLES

六项原则，定义精炼边界。

01

S

任务条件充分性

上下文目标不是最短，而是对当前决策保留最大有效信息。

02

R

可恢复精炼

活跃状态可以很小，但原始或高保真状态必须能按需恢复。

03

Δ

残差智能

外部层只承担当前模型尚未内化且仍有正边际收益的能力。

04

∂

可吸收性边界

静态策略可能被模型吸收；状态、权限、证据和资源不会永久参数化。

05

B

计算自适应

依据难度、不确定性与边际收益动态决定检索、扩展、验证或停止。

06

↻

外部认知共进化

外部探索、验证、能力吸收与运行时裁剪形成持续循环。

04 • COGNITIVE RUNTIME

它不是更厚的 Harness，而是可自我裁剪的认知运行时。

认知运行时位于用户任务与基础模型之间，却不试图复制模型。它保存模型无法拥有的现实边界，并按当前任务决定哪些外部能力值得被调用。

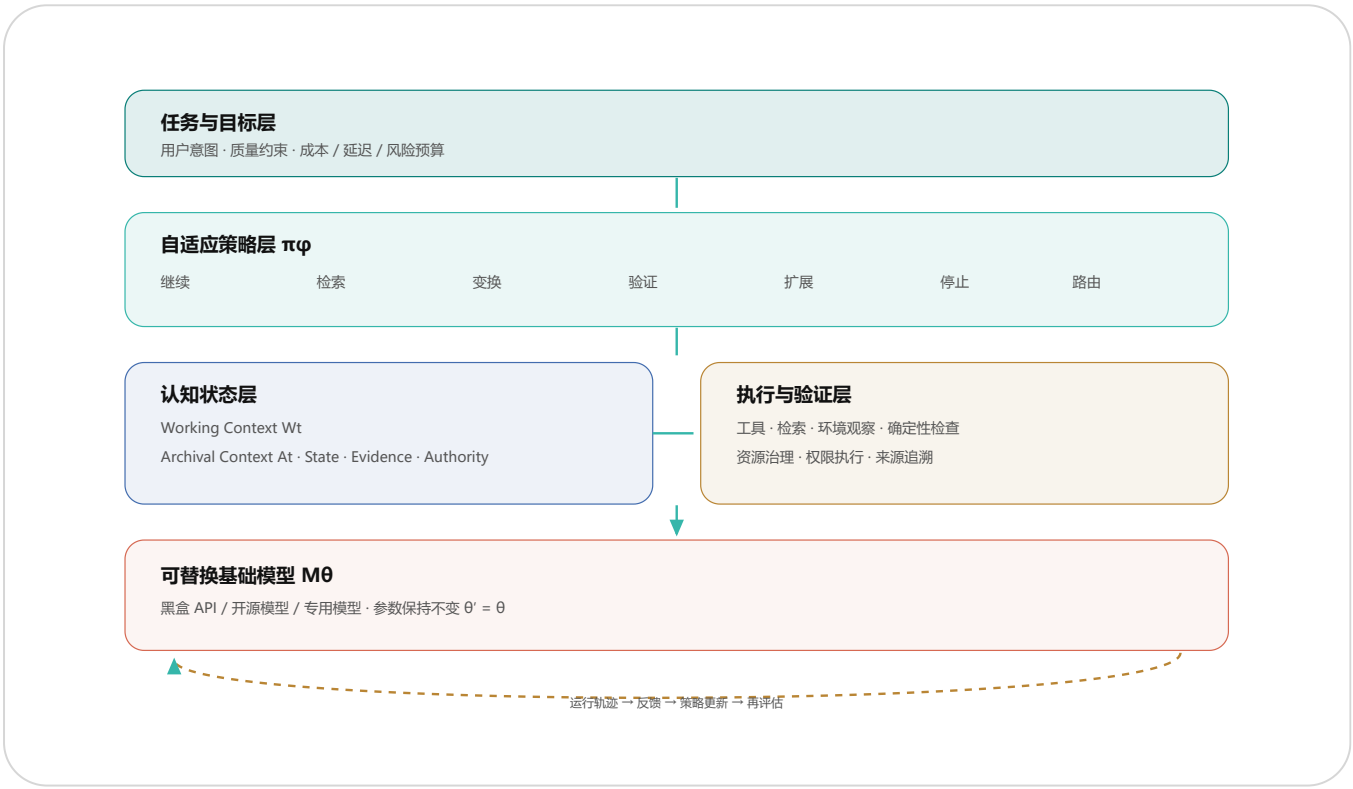


图 1 | 认知运行时分层。模型是可替换计算核心；状态、证据、权限、验证、资源和反馈由模型外部持有。

05 • RECOVERABLE CONTEXT

把压缩改造成认知工作集管理。

未知后续问题使“全局最优短摘要”通常不存在。AIR 因此把上下文拆为活跃认知状态与可恢复归档状态：

$$\mathcal{C}_t = (W_t, A_t) \qquad |W_t| \ll |A_t|$$

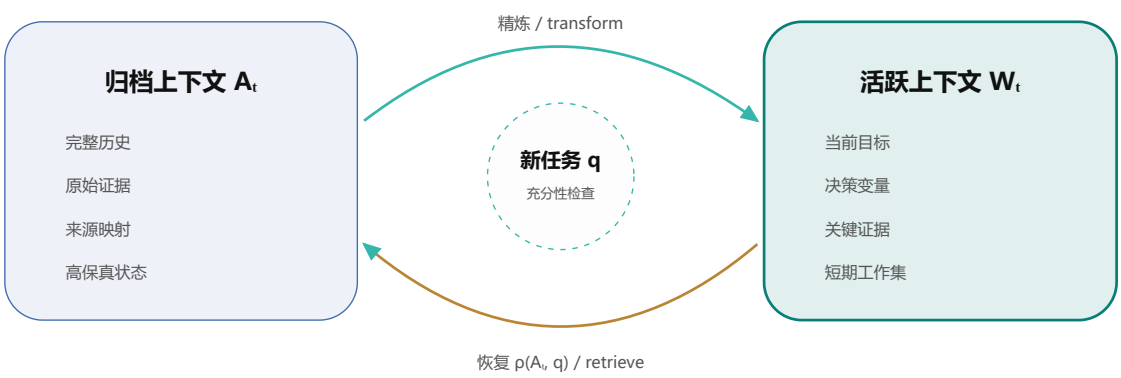


图 2 | 双态上下文。活跃表示负责低成本推理；归档状态负责未来可恢复性。二者之间保留双向映射。

$$\Gamma_{\text{recoverable}} = |S^* \cap S(W \cup \varrho(A,Q))| \,/\, |S^*| \quad \to \quad 1$$

允许 Γactive < 1，但关键语义的总可恢复覆盖率应趋近于 1

关键约束

可恢复不等于保存一切到活跃窗口。系统只为当前必需信息支付推理成本，同时保存来源、索引和恢复能力。

06 • RESIDUAL INTELLIGENCE

外部系统的价值，是模型尚未拥有的那一部分。

固定 Harness 会不断叠加功能；残差智能要求每个外部策略持续证明自己的边际价值。

$$\mathcal{P}^*(M) = \{\, p_i : U_i(M,D) > \tau \,\}$$

当模型升级并吸收某种规划、反思或格式能力时，Ui 下降。若其净收益低于阈值 τ，该策略应被旁路或删除，而不是作为历史包袱永久保留。

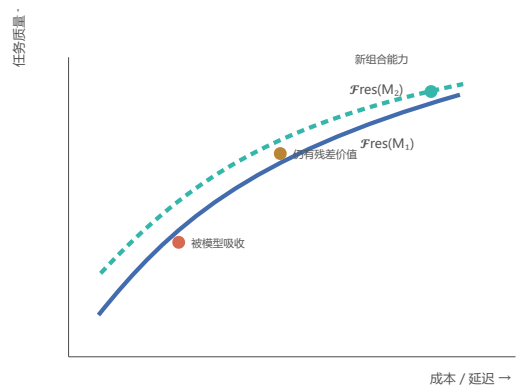


图 3 | 残差智能前沿。示意曲线表达模型升级后的策略价值迁移，不代表实测数值。

什么会被吸收，什么必须留在模型外？

更可能被模型吸收	不能被固定参数永久替代
静态、可学习、分布稳定的规划模板	**State**：实时、非平稳的世界状态
通用反思套路与固定格式变换	**Evidence**：来源、证据链与审计轨迹
常见工具选择启发式	**Authority**：身份、权限与操作边界
重复性提示脚手架	**Verification**：确定性检查与外部判据
稳定任务分布下的工作流	**Resource**：实时预算、配额与调度
可从训练轨迹内化的策略	**Feedback**：部署后的真实结果与事后反馈

07 • ADAPTIVE COMPUTE

是否继续思考，本身也是一个决策。

固定推理深度对简单任务浪费资源，对困难任务又可能不足。AIR 把推理过程建模为预算约束下的序贯决策。

$$s_t = (c_t, b_t, q_t, u_t, v_t) \qquad a_t \sim \pi_\varphi(a \mid s_t)$$

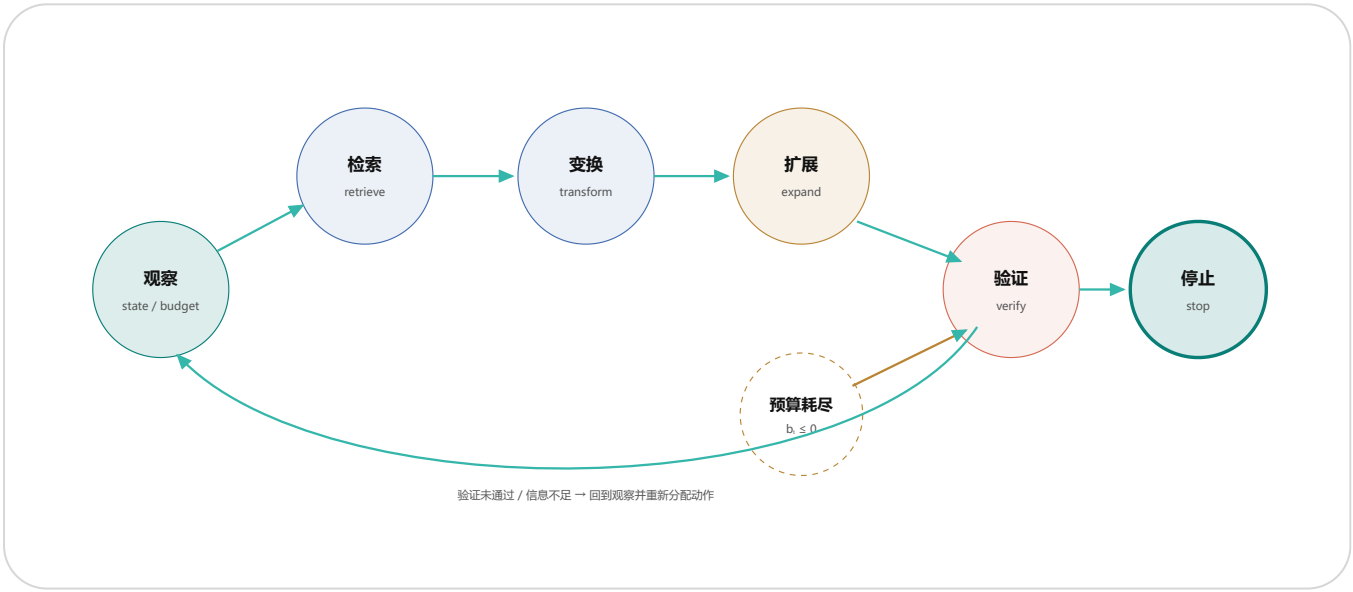


图 4 | 精炼状态机。动作集合不是固定流水线；策略可跳过任意低价值步骤。

动态停止规则

当下一轮预期质量提升小于新增计算和延迟的影子价格时，应停止扩展：

$\mathbb{E}[\Delta q_{k+1}] < \lambda_C \Delta C + \lambda_L \Delta L$。

08 • VERIFICATION LOOP

生成更多答案，不等于得到更多真相。

同一模型的多次生成往往共享错误来源。当候选错误相关系数 ρ → 1，继续增加样本数量的收益趋于饱和。因此有效验证需要与原生成机制不完全相关的信号。

同源信号

低正交性

同模型自评、同提示重采样、同一错误知识下的投票。可改善随机误差，但难以打破系统性偏差。

正交信号

高验证价值

编译器、测试、数据库约束、环境观测、来源证据、权限系统、异构模型与人工反馈。

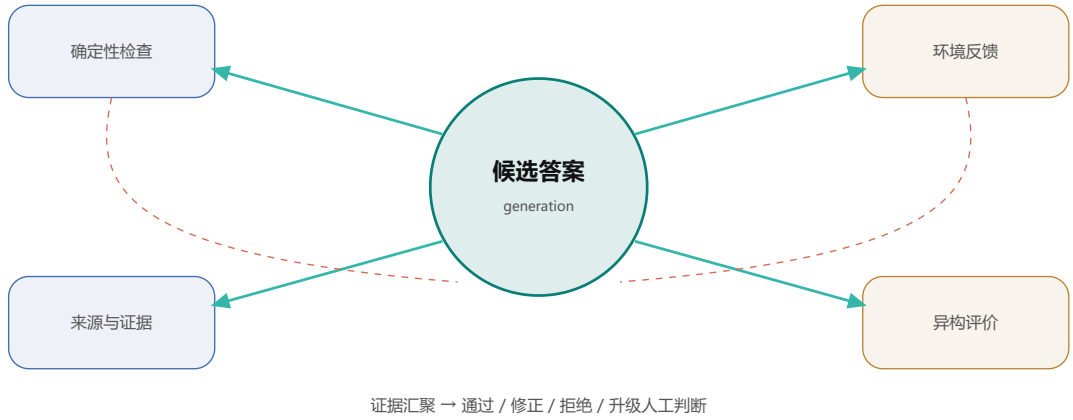


图 5 | 验证增强智能。验证器不是装饰性的“再想一次”，而是尽可能引入不同误差结构。

评价器风险

优化系统可能学习迎合评价器而非真实目标。持续演化必须依赖多源反馈、定期外部校准和对评价器本身的审计。

09 • CO-EVOLUTION

外部探索，内部吸收，运行时再度变薄。

外部层是能力孵化器，而不是永远膨胀的容器。它先以低风险、可观察的方式探索策略，再将稳定能力输送给训练；模型吸收之后，外部层重新裁剪。

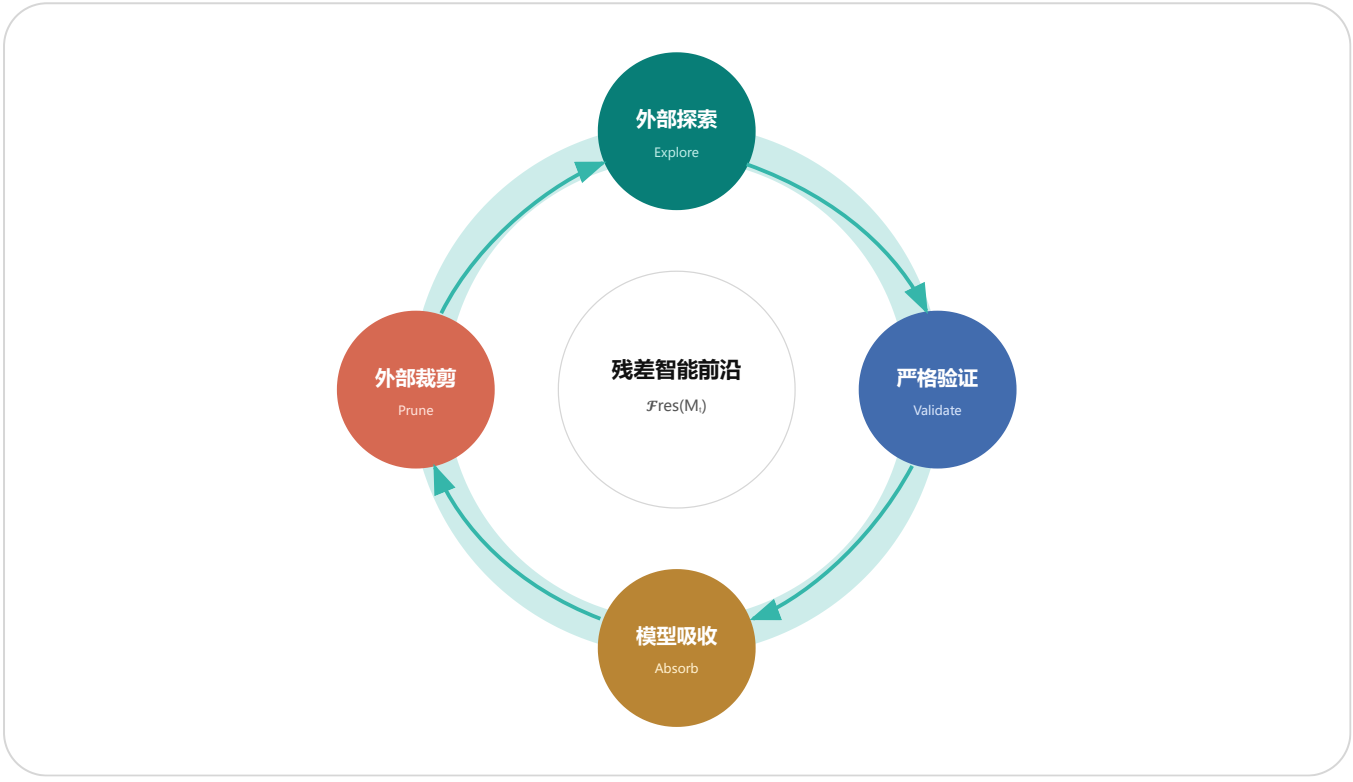


图 6 | 共进化飞轮。新策略必须经过验证才能进入训练候选；能力吸收后还必须重新基准测试并裁剪外部实现。

10 • THEORETICAL PROPERTIES

六个性质，
也是六个工程约束。

P1 • 基础模型非侵入性

精炼层只作用于推理边界之外，因此 **θ′ = θ**；无需修改模型参数即可改变系统能力。

P2 • 模型替换性

策略应依赖能力接口与测量结果，而非某个模型的私有内部结构；更换 M 后重新估计策略价值。

P3 • 预算有界性

任一推理轨迹必须受显式 Token、计算、延迟和风险预算约束；不存在无限递归或无限验证。

P4 • 条件非劣化

只有当候选策略在可信评价上满足净效用门槛才启用；无法证明收益时退回基础路径。

P5 • 可恢复鲁棒性

面对未知后续查询，双态上下文比同等活跃长度的不可逆压缩保留更高的语义恢复上限。

P6 • 残差策略可消亡

外部策略不是永久资产。当 $U_i(M,D) \le \tau$，它必须允许被旁路、归档或删除。

严谨性说明

以上为理论性质及设计目标。P4、P5、P6 需要在明确任务分布、评价器和置信区间的实验条件下验证，不能仅由定义直接推出真实系统必然满足。

11 • FALSIFICATION PROTOCOL

一套理论，必须先说明怎样证明它是错的。

AIR 的核心不是某个漂亮架构，而是能否在相同模型、相同任务集和可复现实验预算下产生可测量的净增益。

基础对照

组别	系统	目的
B0	基础模型 M	建立无外部精炼基线
B1	M + 固定 Prompt / 固定 Harness	排除普通脚手架收益
B2	M + 不可逆上下文压缩	比较恢复能力与多轮鲁棒性
AIR	M + 自适应精炼层	测量动态策略的净效用

核心指标

ΔQ
任务成功增益

ΔC
成本变化率

ΔL
延迟变化率

Q / C
单位成本成功率

CU
上下文利用率

RR
恢复成功率

RV

i

残差策略价值

Risk

语义 / 操作风险

最小消融矩阵

移除项	主要观察	支持理论的预期方向
恢复算子 ρ	未知后续问题的 RR 与成功率	显著下降
动态停止	简单任务成本、困难任务质量	成本上升或质量下降
正交验证	系统性错误、不可验证率	错误率上升
策略裁剪	模型升级后的净效用与延迟	冗余成本上升
历史反馈 H	跨任务策略选择稳定性	长期效用下降

六个可证伪假设

H1 • 工作集

长上下文任务中，任一时刻真正影响下一步决策的信息量显著小于完整历史。

H2 • 可恢复优势

多轮未知后续任务中，可恢复精炼优于同活跃长度的不可逆压缩。

H3 • 策略迁移

模型升级会显著改变固定 Harness 策略的平均边际收益。

H4 • 计算再投资

节省的低价值上下文预算投入高价值推理或验证，可提升总任务效用。

H5 • 正交验证

误差结构更独立的验证信号，比同源自验证带来更高净可靠性增益。

H6 • 能力孵化

经外部探索和验证得到的稳定策略，可成为后续模型训练的有效监督信号。

12 • LIMITS & SAFETY

精炼不是魔法，边界必须先于能力。

边界 01

精炼不能创造不存在的信息。若原始上下文、外部环境和模型均不包含必要事实，重新组织无法凭空恢复真相。

边界 02

自反思不能消除完全相关错误。同一错误机制重复调用，只会让错误更稳定。

边界 03

不存在对所有未来任务同时最优的任意短表示。未知查询要求保留可恢复来源。

边界 04

评价器可能成为新瓶颈。系统可能优化分数而非真实目标，形成评价器投机。

边界 05

额外计算存在收益递减。当 $\partial Q/\partial C \to 0$，继续推理只会消耗资源。

三条安全基线

TRACEABILITY

原始语义可追溯

每个高层表示 zi 都应保留到 source(zi) 的映射。

SEPARATION

权限与智能分离

模型可以建议动作，但授权、执行与审计必须由外部可信边界控制。

TRANSPARENCY

优化目标透明

质量、成本、延迟与风险权重需要可见、可审计且可被业务覆写。

站在已有研究之上，但不等同于其中任何一支。

研究方向	已解决的问题	AIR 的扩展
Prompt Compression	提升输入信息密度、降低长上下文成本	把不可逆压缩纳入更大的可恢复工作集管理
Test-Time Compute Scaling	用额外推理计算换取质量	同时决定何时增加、转移或停止计算
Recursive Language Models	外化长输入并递归处理子问题	加入显式预算、恢复、验证和策略裁剪
Agentic Context Engineering	让上下文随经验迭代	强调模型升级时的残差价值重估
Reflective Optimization / Agent RL	从轨迹与奖励改进策略	连接“外部探索 → 模型吸收 → 外部裁剪”
传统 Agent Harness	通过固定流程补足模型能力	把流程视作临时策略集合，而非永久架构

核心差异

普通 Harness 常写作 **M + 固定工作流**；AIR 写作 **Mt + $\mathcal{P}$*(Mt, D, B)**。最优外部策略随模型、任务分布和预算共同变化。

模型越强，外部层越薄；但现实边界不会消失。

阶段一｜厚 Harness

模型能力有限，规划、检索、验证和结构控制大量位于模型外部。

阶段二｜策略吸收

基础模型逐渐内化常见规划与反思，固定逻辑的残差价值下降。

阶段三｜薄 Runtime

外部层聚焦 State、Evidence、Authority、Verification、Resource 与 Feedback。

阶段四｜认知操作系统

运行时负责连接正确世界状态、调度有限资源、提供可验证现实，并向下一代模型回输能力。

> **长期存在的不是某个固定 Harness，而是持续寻找残差智能前沿的优化过程。**

APPENDICES

最小理论表达、符号与文献。

▼ 附录 A | 四个最小公式

系统目标： $J = \mathbb{E}[Q] - \lambda_C \mathbb{E}[C] - \lambda_L \mathbb{E}[L] - \lambda_R \mathbb{E}[R]$

上下文目标： $Z^* = \arg\max_Z I(Z;Y \mid Q) - \beta C(Z)$

策略选择： $\mathcal{P}^*(M) = \{p_i : U_i(M) > \tau\}$

共进化： $\varphi_{t+1} \leftarrow \mathrm{Learn}(\varphi_t, H_t, M_t)$

▶ 附录 B | 符号表

▼ 参考文献

01 Shannon, C. E. *A Mathematical Theory of Communication*. Bell System Technical Journal, 1948.

02 Cover, T. M., Thomas, J. A. *Elements of Information Theory*. Wiley.

03 Tishby, N., Pereira, F. C., Bialek, W. *The Information Bottleneck Method*. 1999.

04 Vaswani, A., et al. *Attention Is All You Need*. NeurIPS, 2017.

05 Jiang, H., et al. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. 2023.

06 Jiang, H., et al. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. 2023.

07 Pan, Z., et al. LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. 2024.

08 Snell, C., et al. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. 2024.

09 Chen, Y., et al. Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models. 2024.

10 Lin, K., et al. Sleep-time Compute: Beyond Inference Scaling at Test-time. 2025.

11 Agrawal, L. A., et al. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. 2025.

12 Luo, X., et al. Agent Lightning: Train ANY AI Agents with Reinforcement Learning. 2025.

13 Zhang, Q., et al. Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models. 2025.

14 Zhang, A. L., Kraska, T., Khattab, O. Recursive Language Models. 2025/2026.

15 Sutton, R. S., Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press.

文档来源与编辑说明

本版以《元炼协议 II：自适应推理精炼理论》v2.0 原稿为理论底本，并吸收《精炼。md》讨论记录中关于论文式章节、核心概念图、架构图、演化流程和企业白皮书排版的要求。为保持版本一致性，标题沿用“协议 II / v2.0”；讨论中的“协议 III”被视为后续结构设想，而非本版版本号。

METAREFINE PROTOCOL II

真正长期存在的，不是固定工作流，而是持续寻找残差智能前沿的优化过程。

Adaptive Inference Refinement Theory · v2.0 · 2026