



北京理工大学
信息系统及安全对抗实验中心
信息安全与对抗技术研究所
Beijing Forest Studio



长时程智能体的退化困境与自我改进

www.isclab.org.cn

博士研究生 王文静

2026年10月18日

- 相关内容

- 刘佳《智能体安全——工具调用攻击与防御》——2026.09.13
- 刘栋涵《智能体的第三方组件安全：从工具到技能》——2026.08.17

- 预期收获
- 问题引入
- 题目内涵解析与研究目标
- 研究背景与意义
- 研究历史与现状
- 基础知识
- 算法原理
 - ACE
 - Salus
- 特点总结与工作展望
- 参考文献



- 预期收获
 - 1. 理解长时程智能体为什么会退化
 - 2. 掌握长时程智能体的两条自我改进路径
 - 3. 掌握解耦、反馈、结构化的智能体设计方法论

问题引入 AI越“努力”反而越差？



• 案例说明

- 在 AppWorld 上用 Dynamic Cheatsheet 智能体做在线适配时观察到：智能体在第60步时上下文有18,282 个 token、准确率66.7%，可下一步上下文被重写后只剩122个 token、准确率掉到 57.1%
- 重写后的准确率比完全不做适配的 63.7% 还低

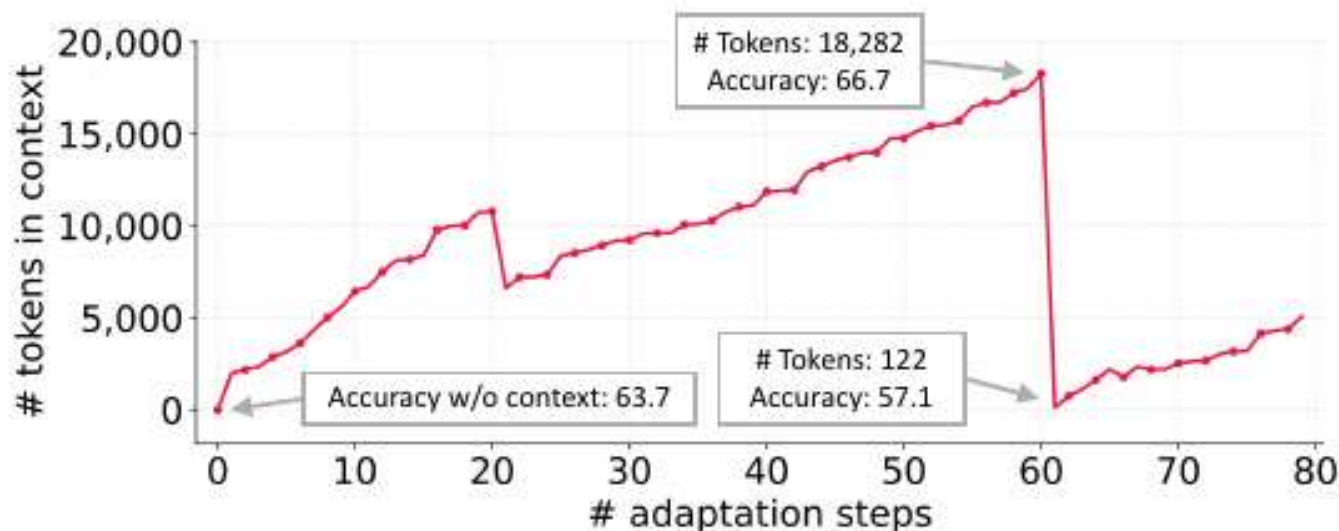
• 上下文越优化越糟，问题出在哪

- 简洁偏置
- 上下文坍塌

上下文一旦被整体重写，累计的知识会突然消失

优化的过程中总是倾向于把内容压短、通用

上下文中的token骤降，准确率也随之降低



问题引入 一次“差一点”的误诊



- 一步之差：误诊与确诊

- 初始临床背景

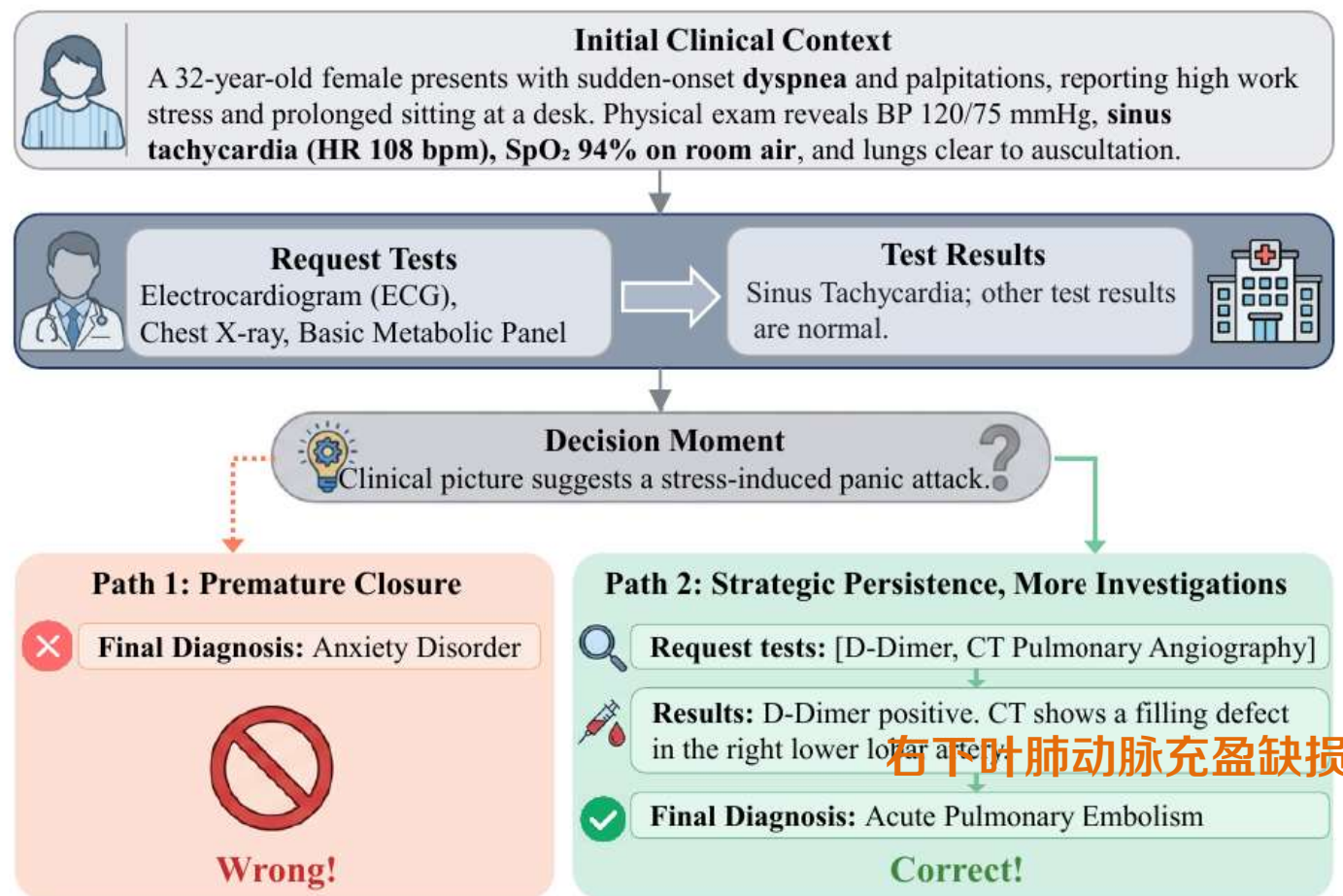
- 32 岁女性，突发呼吸困难 + 心悸；工作压力大、长时间久坐。体征：血压 120/75、心率 108（窦性心动过速）、血氧 94%，肺部听诊清。

- 第一轮检查

- 心电图、胸片、基础代谢面板

- 检查结果

- 仅窦性心动过速，其余正常



过早闭合 → 焦虑症 → 错误

策略性坚持 → 追加检查 → 确诊急性肺栓塞 → 正确

- 题目内涵解析

长时程：指任务无法一步完成，需要连续多步、多轮交互才能达成。

- 长时程智能体

- 需要在**多轮交互**里连续推进、状态不断累积、还要动态做决策的智能体，区别于一问一答的单轮任务

- 退化困境

- 长过程中出现的两类系统性退化

- 一类是“忘”，也就是简洁偏置和**上下文坍塌**
 - 一类是“急”，也就是**过早闭合**

- 自我改进

- 不重训整个模型，也能在过程中持续变好，有**上下文自演化**和**策略自改**两条路

- 研究目标

- 让智能体具备“长期自主完成复杂目标”的能力——在长时间、多步骤、动态变化的真实环境中，稳定、可靠、可持续地朝目标推进



- 研究背景

- 范式转变

上下文坍缩、过早闭合：越用越差

- 从一问一答走向长时程、**多轮交互**
 - ReAct 开启智能体范式：边推理边行动、调用工具、观察结果再继续

- 现有方法力不从心

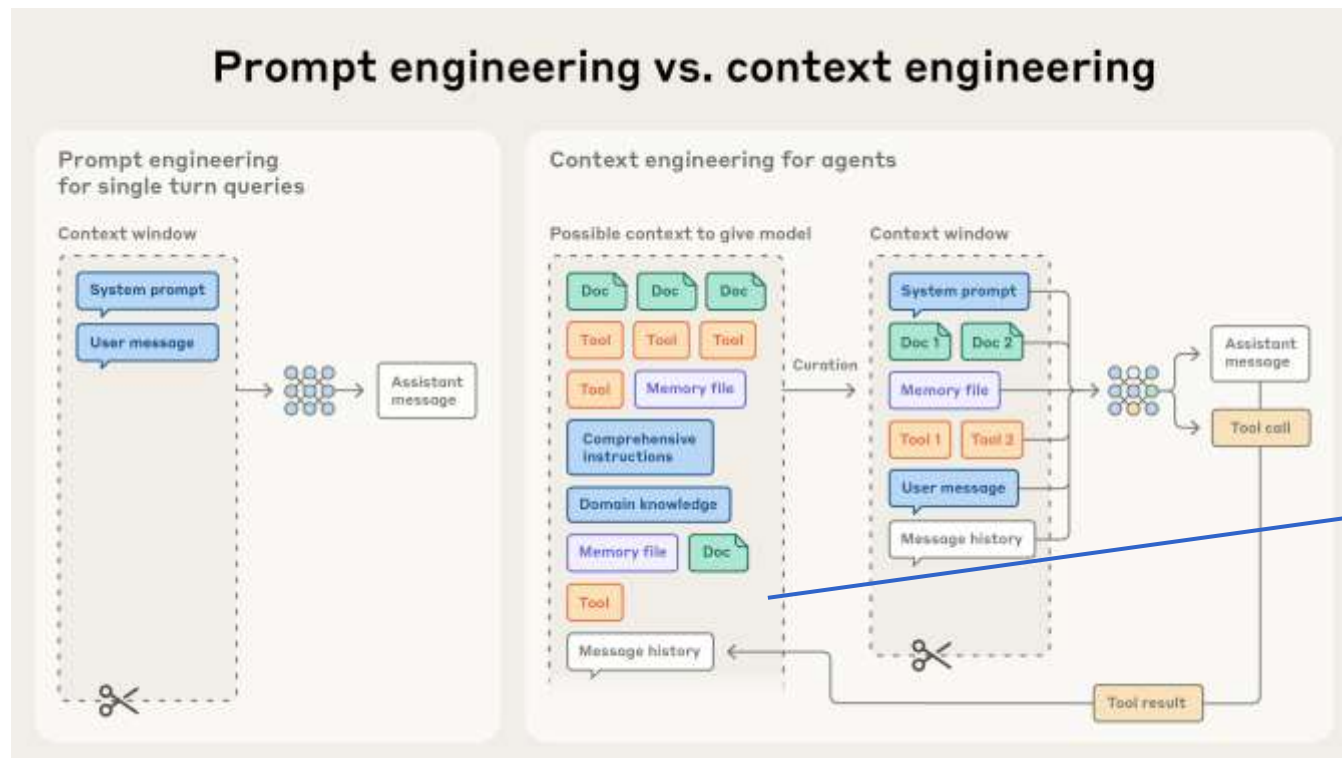
路线	不足
提示优化	简洁偏置——细节被压没
记忆方法	上下文坍缩——越重写越短、知识被抹掉
多轮强化学习	奖励稀疏 + 奖励欺骗——训不动、或学会投机

- 研究意义

- 通过揭示智能体的退化机理，制定有效的自我改进范式，从而提升智能体在**医疗**等高风险场景中的**可靠性**

- 上下文工程

- 定义：在推理过程中，持续策管、维护进入上下文窗口的信息
- 重要性：上下文是**有限资源**（窗口有限、注意力预算有限）；输入越长、越杂，模型对信息的召回与利用就越**不可靠**



从候选池挑选（策管）→ 放进窗口 → 内容过多时裁剪

文档、工具、记忆文件、领域知识、消息历史

Jin等人指出缺少贴近真实的大规模医学评测，遂用执照考题构建单轮选择题基准，成为医学LLM评测起点；其**单轮、静态属性**也成为日后**真实诊疗是动态过程**的批评靶子

Shinn等人指出智能体**不会从失败中学习、反复犯同样错**，于是用自然语言反思并写入记忆；**开创反思式自我改进**，是ACE的Reflector与Salus从轨迹提炼经验"的思想源头

Tu等人指出LLM单轮分数高但**真实多轮问诊能力不明**，遂以对话式诊断+自博弈训练达到**比肩初级医生**；证明可连续临床推理，但闭源、偏问诊对话而非策略性开检查，留下缺口

Gao等人指出LLM顺序诊断中**频繁过早闭合**，多轮RL又受**奖励稀疏与欺骗困扰**，于是把诊断形式化为顺序决策+DR/SC/WP 三角色+分层奖励RL；补上医学侧可训练交互策略缺口：证明小规模模型也能具备顶尖的临床决策能，以远小于对手的参数量达到领先水平



Yao等人指出**纯LLM无法与环境交互、调用工具**，于是提出ReAct，把推理与行动交织；**奠定智能体范式**，是医学与上下文两条脉络的共同地基

Suzgun等人指出推理时**无法积累复用经验、常缺标注**，遂建**自适应外部记忆**、无需真值标签；把上下文从静态提示推进到积累经验，是ACE的直接前身，也暴露了上下文坍塌

Zhang等人指出现有上下文适配存在**简洁偏置与上下文不改权重、无需标注**；补上上下文侧增量维护缺口，在智能体与专业领域任务上均显著超越既有强基线，且**大幅压缩了适配的延迟与成本**



Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models



T	目标	利用可演化、详尽的上下文实现 可扩展 、 低成本 、 可自改进 的智能体
I	输入	结构化上下文
P	处理	1. Generator 产生推理轨迹，并标注哪些bullet有用或误导 2. Reflector批评轨迹、提取复用策略与失败模式，并做多轮 迭代精炼 3. Curator将经验综合为紧凑的 增量delta条目 （结构化bullets） 4. 用 非LLM的确定性逻辑 将delta合并进playbook，支持并行批量与多epoch适配
O	输出	演化后的结构化上下文playbook

P	问题	现有上下文适配方法存在两大缺陷： 简洁偏置 和 上下文坍塌
C	条件	不更新模型权重，仅通过修改输入上下文来改进；可通过系统提示或测试时记忆适配；无需标注监督
D	难点	在持续 累积细节 与 冗余控制 之间取得平衡
L	水平	ICLR 2026 CCF A



- 科学问题
 - 上下文自适应为什么会失效，以及什么样的上下文表示与更新机制才能让知识持续累积而不退化
- 应用问题
 - 如何在重新训权重的情况下，用一份可演化、可复用、低成本构造的上下文来持续自我改进

- ACE (Agentic Context Engineering)

- 专职 Reflector

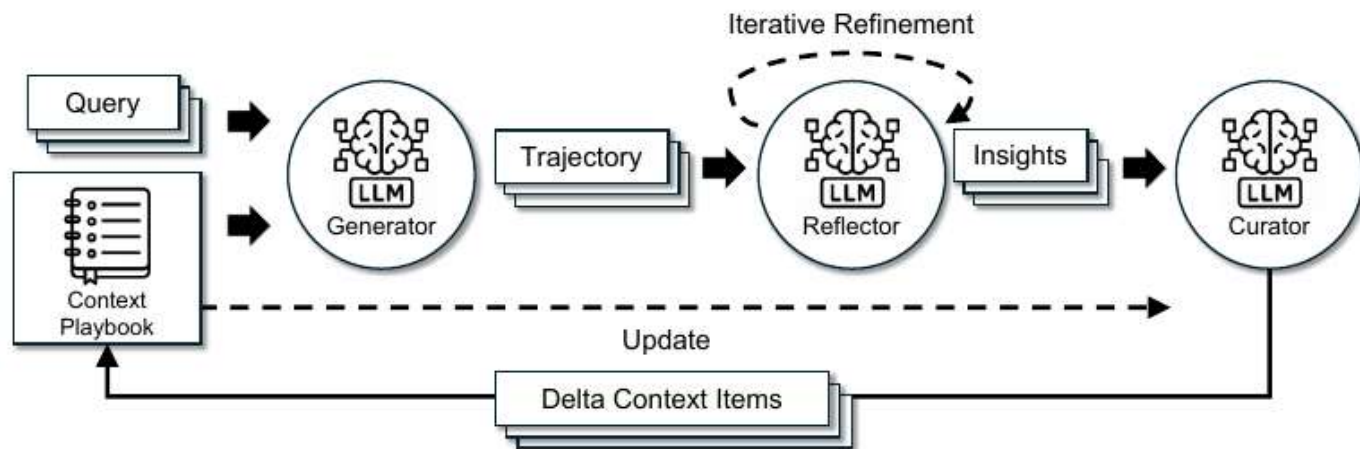
- 避免让一个模型同时承担判断好坏和写入上下文两件事，防止职责过载导致洞见提取质量下降

- 增量式条目更新

- 整体重写会随着上下文变大把它越压越短、信息骤失；局部编辑则保留旧知识、只追加新知识

- 增长和提炼机制

- 防止上下文无限膨胀带来的冗余与噪声，同时又不牺牲"持续累积细节"这一核心优势



Playbook?——操作手册

- 增量式条目更新

- 局部化更新

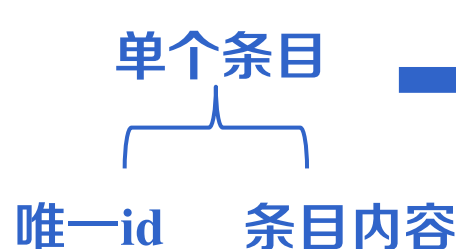
- 只更新相关的条目，其余保持不变；
已有条目支持就地更新，新知识以
新 id 追加

- 细粒度检索

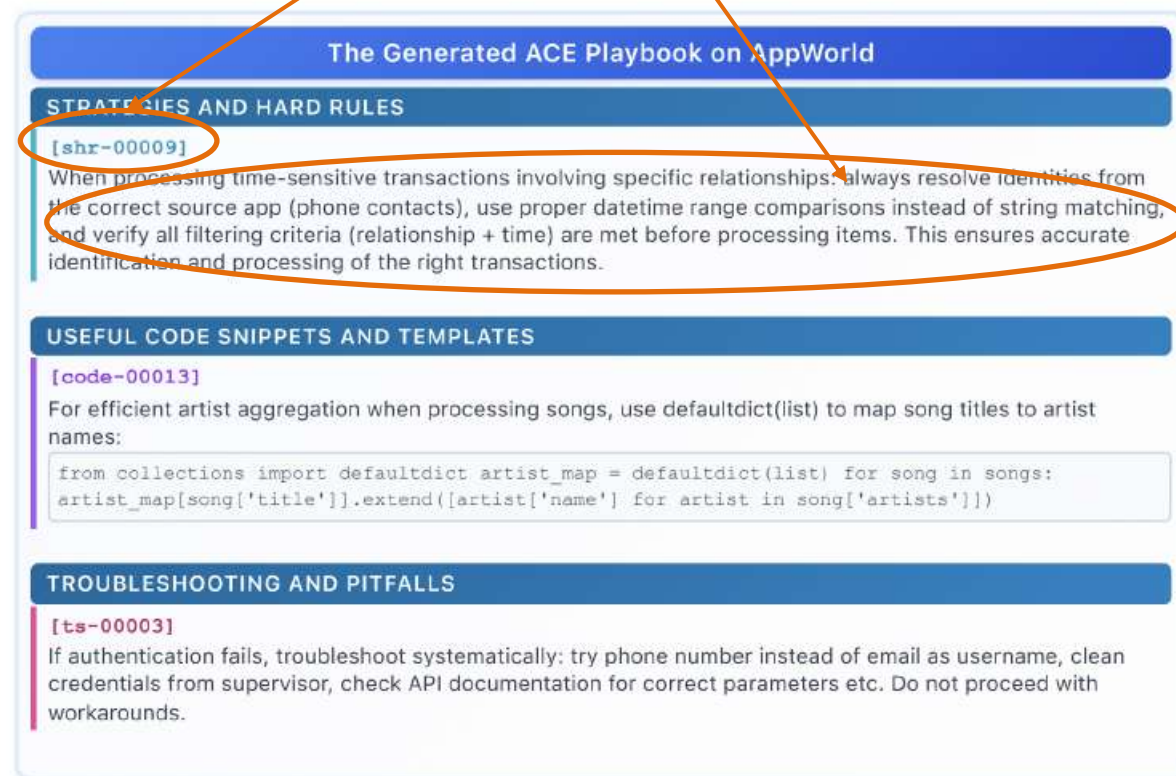
- 每条条目都是一个独立可寻址的单元；Generator 可以聚焦于最相关的知识

- 增量式适配

- 支持并行合并与批量适配；上下文
随适配不断增长而不失控；避免全
量重写的开销与延迟



- 1、局部化
- 2、细粒度检索
- 3、增量式适配



只追加不整理？——冗余、膨胀！

条目化表示

增长和提炼机制

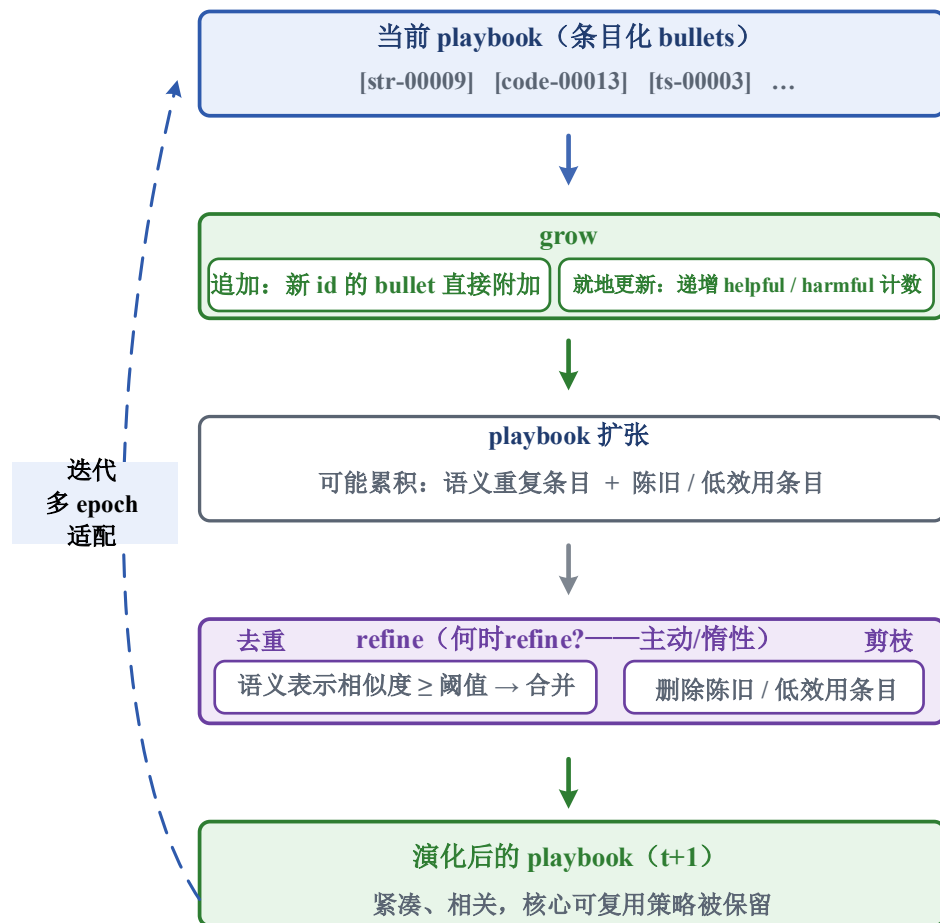
– 增长

- 带新标识符的条目**直接追加**进来；而已有的条目则**就地更新**（例如递增其计数器）

– 提炼

- 比较条目语义嵌入来剪除冗余
- 执行该操作有两种策略
 - 主动执行：每合并完一个条目就**立刻去重+剪枝**。上下文始终干净最新，但**延迟更高**
 - 惰性执行：上下文长度超出模型上下文窗口时才批量整理一次。**开销小、延迟低**，代价是中间过程会有一定冗余

条目化上下文随适配持续增长，同时周期性 / 惰性地控制冗余与规模。



增长和提炼机制示意图



- 数据资源及指标

- 任务与数据集：APPWorld、**领域专用推理**（FiNER、Formula）、**跨领域验证**（DDXPlus、BIRD-SQL）

- 评价指标

- AppWorld：TGC（**任务目标完成率**）和SGC（**场景目标完成率**）

- FiNER / Formula / DDXPlus：Accuracy

- BIRD-SQL：LLM-as-a-judge，由 GPT-4o-mini 担任评判

- 评估协议

- 所有数据集沿用其原始 train / validation / test 划分

- 离线适配：在训练集上优化，在测试集上以 pass@1 评估，即每个样本只回答一次

- 在线适配：在测试集上顺序评估，每个样本先用当前上下文预测，再基于该样本更新上下文



基线方法	性质与设置
Base LLM	不做任何上下文工程，使用数据集默认提示直接评测
ICL(2024,NeurIPS)	上下文学习：在提示中放入少样本/多样本示范；放得下就全给，否则填满窗口
MIPROv2(2024,EMNLP)	提示优化器：贝叶斯优化联合优化系统指令与示范；官方 DSPy 实现，auto=heavy
GEPA(2025,arXiv)	反思式提示进化（Genetic-Pareto）；收集执行轨迹、自然语言反思、遗传搜索保前沿；DSPy auto=heavy。 最强直接对比基线
Dynamic Cheatsheet（DC） (2025,arXiv)	测试时自适应记忆，累积模式 DC-CU，无需真值标签。 ACE 架构的前身
ACE（该论文）	智能体式上下文工程：同 LLM 分饰 Generator/Reflector/Curator，离线与在线均适配
智能体框架：官方 ReAct 实现（Yao 等，2023）， 所有基线与方法均构建于其上 ACE 关键超参：batch size = 1（每个样本构建一个 delta 上下文）；Reflector 精炼轮数 ≤ 5；离线适配 epoch ≤ 5 ACE默认基座模型：DeepSeek-V3.1-671B（非思考模式）；且 Generator / Reflector / Curator 三角色共用同一 LLM （隔离"上下文构建本身"的收益，防止强角色给弱 Generator 输血）	

实验结果 AppWord对比实验



- AppWord智能体基线测试结果分析
 - 离线适配大幅领先
 - 结构化、可演化、详尽的上下文比静态示范或单条优化指令更有效
 - 无标签也能自我改进
 - 利用执行过程自然产生的信号来指导Reflector 与 Curator增强了智能体的鲁棒性
 - 比肩生产级智能体
 - 离线模式下ACE（59.4%）与榜首IBM CUGA（60.3%）基本持平，但ACE用的是小得多的开源模型DeepSeek-V3.1

Method	GT Labels	Test-Normal		Test-Challenge		Average
		TGC↑	SGC↑	TGC↑	SGC↑	
DeepSeek-V3.1-671B as Base LLM						
ReAct		63.7	42.9	41.5	21.6	42.4
Offline Adaptation						
ReAct + ICL	✓	64.3 ^{+0.6}	46.4 ^{+3.5}	46.0 ^{+4.5}	27.3 ^{+5.7}	46.0 ^{+3.6}
ReAct + GEPA	✓	64.9 ^{+1.2}	44.6 ^{+1.7}	46.0 ^{+4.5}	30.2 ^{+8.6}	46.4 ^{+4.0}
ReAct + ACE	✓	76.2 ^{+12.5}	64.3 ^{+21.4}	57.3 ^{+15.8}	39.6 ^{+18.0}	59.4 ^{+17.0}
ReAct + ACE	✗	75.0 ^{+11.3}	64.3 ^{+21.4}	54.4 ^{+12.9}	35.2 ^{+13.6}	57.2 ^{+14.8}
Online Adaptation						
ReAct + DC (CU)	✗	65.5 ^{+1.8}	58.9 ^{+16.0}	52.3 ^{+10.8}	30.8 ^{+9.2}	51.9 ^{+9.5}
ReAct + ACE	✗	69.6 ^{+5.9}	53.6 ^{+10.7}	66.0 ^{+24.5}	48.9 ^{+27.3}	59.5 ^{+17.1}

Agent Scores

[Raw Data](#)

Add your agent to the leaderboard [following the instructions](#) on GitHub. 🚀

All Level 1 Level 2 Level 3

Interactions

Method	LLM	Link	Date	Test-Normal		Test-Challenge	
				TGC	SGC	TGC	SGC▼
IBM CUGA	GPT-4.1	🔗	2025-07-12	73.2	62.5	57.6	48.2
LOOP	Qwen2.5-32B	🔗	2025-04-09	72.6	53.6	47.2	28.8
ReAct + 2 SetBSR Demos	GPT-4o	🔗	2025-07-13	68.5	57.1	38.9	23
ReAct	GPT-4o	🔗	2024-07-26	48.8	32.1	30.2	13



- 领域专用基线测试结果分析
 - 反馈质量很重要
 - 有真值标签才稳
 - 无标签在线时 ACE 72.9、DC 65.4，均可能退化，这揭示了上下文会被虚假信号污染
 - 跨领域可迁移
 - 医疗诊断 DDXPlus 由 75.2 提
升至 90.2 (+15.0，而 GEPA 仅
+1.2)

对精确领域知识要求高的任务收益更大

Method	GT Labels	FiNER (Acc↑)	Formula (Acc↑)	Average
DeepSeek-V3.1 as Base LLM				
Base LLM		70.7	67.5	69.1
Offline Adaptation				
ICL	✓	72.3 ^{+1.6}	67.0 ^{-0.5}	69.6 ^{+0.5}
MIPROv2	✓	72.4 ^{+1.7}	69.5 ^{+2.0}	70.9 ^{+1.8}
GEPA	✓	73.5 ^{+2.8}	71.5 ^{+4.0}	72.5 ^{+3.4}
ACE	✓	78.3 ^{+7.6}	85.5 ^{+18.0}	81.9 ^{+12.8}
ACE	✗	71.1 ^{+0.4}	83.0 ^{+15.5}	77.1 ^{+8.0}
Online Adaptation				
DC (CU)	✓	74.2 ^{+3.5}	69.5 ^{+2.0}	71.8 ^{+2.7}
DC (CU)	✗	68.3 ^{-2.4}	62.5 ^{-5.0}	65.4 ^{-3.7}
ACE	✓	76.7 ^{+6.0}	76.5 ^{+9.0}	76.6 ^{+7.5}
ACE	✗	67.3 ^{-3.4}	78.5 ^{+11.0}	72.9 ^{+3.8}

Method	Accuracy (↑)
DeepSeek-V3.1 as Base LLM	
Base LLM	75.2
Offline Adaptation	
GEPA	76.4 ^{+1.2}
ACE	90.2 ^{+15.0}

医学推理基准测试结果



- Reflector 或 多轮迭代
 - 二者共同把粗浅的轨迹打磨成**高质量策略**。去掉该组件，上下文**只能累积表层经验**、提不出真正有用的经验，因此性能逐级下降
- offline warmup
 - 离线预热可以消除在线阶段的冷启动
 - 先离线建好初始操作手册，相当于给在线阶段一个**好起点**，避免冷启动期的**低质反思**拖累整体表现
- 增量更新
 - 增量**条目只追加/局部修改**，旧知识因此被保留下来

消融实验结果（a）

Method	GT Labels	Test-Normal		Test-Challenge		Average
		TGC↑	SGC↑	TGC↑	SGC↑	
DeepSeek-V3.1 as Base LLM						
ReAct		63.7	42.9	41.5	21.6	42.4
Offline Adaptation						
ReAct + ACE w/o Reflector or multi-epoch	✓	70.8 ^{+7.1}	55.4 ^{+12.5}	55.9 ^{+14.4}	38.1 ^{+17.5}	55.1 ^{+12.7}
ReAct + ACE w/o multi-epoch	✓	72.0 ^{+8.3}	60.7 ^{+17.8}	54.9 ^{+13.4}	39.6 ^{+18.0}	56.8 ^{+14.4}
ReAct + ACE	✓	76.2 ^{+12.5}	64.3 ^{+21.4}	57.3 ^{+15.8}	39.6 ^{+18.0}	59.4 ^{+17.0}
Online Adaptation						
ReAct + ACE	✗	67.9 ^{+4.2}	51.8 ^{+8.9}	61.4 ^{+19.9}	43.2 ^{+21.6}	56.1 ^{+13.7}
ReAct + ACE + offline warmup	✗	69.6 ^{+5.9}	53.6 ^{+10.7}	66.0 ^{+24.5}	48.9 ^{+27.3}	59.5 ^{+17.1}

消融实验结果（b）

Method	Test-Normal TGC (↑)	Test-Normal SGC (↑)	Average
DeepSeek-V3.1 as Base LLM			
ReAct	63.7	42.9	53.3
Offline Adaptation			
ReAct + ACE (no incremental update)	67.3 ^{+3.6}	46.4 ^{+3.5}	56.9 ^{+3.6}
ReAct + ACE (with incremental update)	76.2 ^{+12.5}	64.3 ^{+21.4}	70.3 ^{+17.0}



- 离线适配更少rollout、更低延迟
 - GEPA需要提示生成和验证两大循环，**反复跑**任务验证候选提示
 - ACE取消了这条验证循环，改用局部化的增量delta更新，不必反复整份重评估，因此**执行次数与耗时都大幅下降**
- 在线适配更低延迟、更低token成本
 - 在线适配每来一个样本就要更新一次上下文，DC的**全量重写代价高昂**
 - ACE只写增量条目并由非LLM的确定性逻辑合并，**单步开销极小**，故延迟与花费同时骤降

Method	Latency (s)↓	# Rollouts↓
ReAct + GEPA	53898	1434
ReAct + ACE	9517 _(-82.3%)	357 _(-75.1%)

(a) Offline (AppWorld).

Method	Latency (s)↓	Token Cost (\$)↓
DC (CU)	65104	17.7
ACE	5503 _(-91.5%)	2.9 _(-83.6%)

(b) Online (FiNER).

ACE用增量条目更新以及非LLM确定性合并，实现了更高准确率与更低成本/延迟。



• 算法贡献

- 上下文新范式：把上下文从被压缩的摘要重新定义为详尽、可条目化、可积累的**操作手册**
- 三角色分工架构：Generator生成轨迹、Reflector反思提炼、Curator整合写入，模仿人**实验—反思—固化**的学习方式
- 增量条目更新：直击上下文坍塌，用**局部化编辑**替代整体重写，旧知识不被压没、新知识持续追加

• 算法不足

- **对Reflector有强依赖**，其提炼不出有效知识时，上下文会被污染
- 没有真值或可靠执行信号时，上下文会**被虚假信号污染**
- **适用场景有明显边界**，只对吃细节、需复杂工具、环境特定策略的任务最有效



Salus: Strategic Diagnostic Testing for Complex Diagnosis via Multi-Agent Reinforcement Learning



T	目标	利用多智能体强化学习和结构化奖励优化实现复杂诊断的顺序决策，克服过早闭合
I	输入	初始临床背景、累计病史
P	处理	1. 三角色 DR/SC/WP：把诊断拆成列鉴别诊断、判断该不该继续、提议做什么检查，三个专门角色，各司其职。 2. 与环境多轮交互：AI 反复请求检查 → 读结果 → 更新病史 → 再判断，直到决定停止或达上限，环境只给客观结果、不泄露答案。 3. 冷启动 SFT + 多智能体 RL：先请强模型演示完整推理让模型学会基本功，再用奖励微调 DR 与 SC 的决策。由 AI 裁判按答案排第几名、证据够不够、诊断对不对分层打分，抑制过早闭合与投机答题。
O	输出	最终诊断结果
P	问题	交互过程中普遍出现过早诊断闭合，多轮 RL 面临奖励稀疏与欺骗。
C	条件	拥有高质量、多来源的临床病例语料。
D	难点	奖励稀疏、欺骗两难；精度与成本权衡。
L	水平	ICML 2026 CCF A



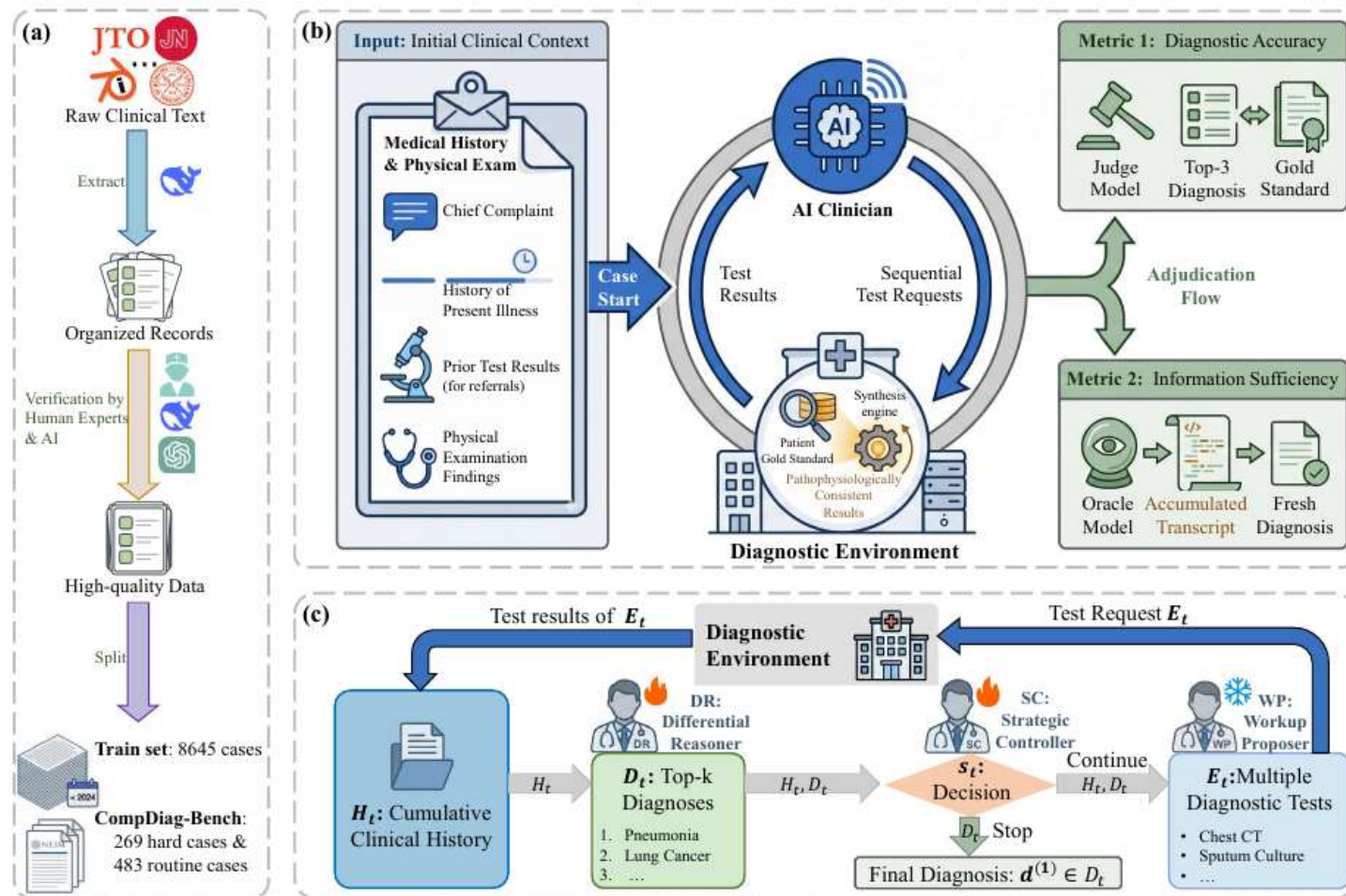
- 科学问题
 - 如何把复杂诊断正确地形式化，并揭示LLM在其中的失效机理及其优化难点
- 应用问题
 - 如何构建一个在真实临床约束下可用的AI临床医生，并为其提供贴近现实、可公平比较的评测环境与度量

• CompDiag-Bench

- 提出一个交互式评测框架，将复杂诊断建模为AI 临床医生 $\leftrightarrow$ 诊断环境的顺序决策过程（而非静态问答）

• Salus

- 专提出多智能体架构，把临床推理拆成三个专职角色（DR / SC / WP），将推理与策略控制解耦



ACE框架

- **CompDiag-Bench**

- 数据管理

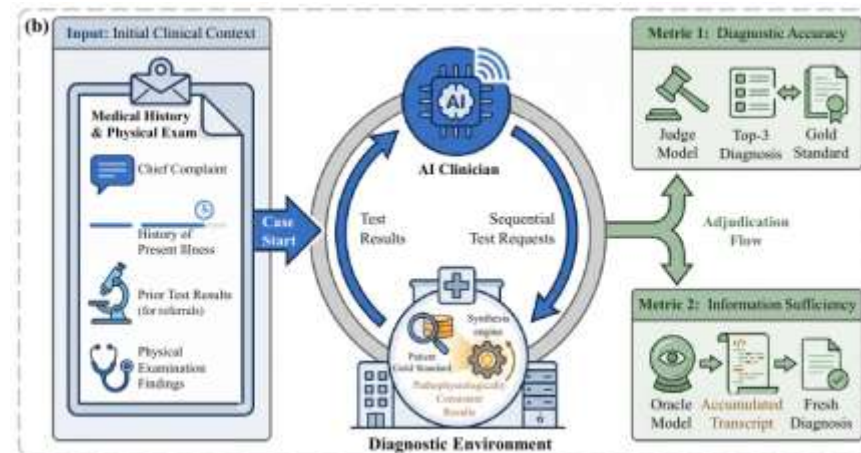
- 用 DeepSeek-V3.1 抽取流水线，把原始临床文本标准化为**四要素病历**：病史、体格检查、检查结果、金标准诊断

- 交互诊断环境

- 用AI代理患者和医院诊断基础设施

- 评价指标与裁决

- **诊断准确率**
 - **信息充分性**



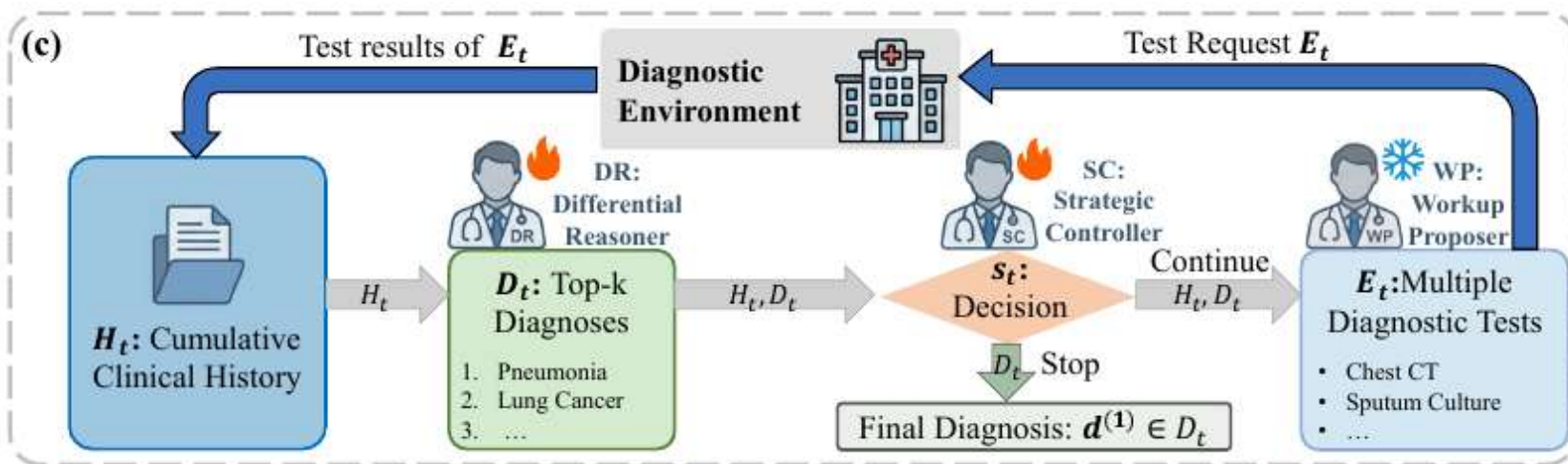
CompDiag-Bench框架

将诊断任务建模为顺序决策过程，目标是在可接受的检查效率内最大化诊断准确率

- Salus

- 临床推理与策略控制解耦

- Differential Reasoner: 临床综合, 产出 **top-k 鉴别诊断**
 - Strategic Controller: 元认知守门人, 判断证据是否充分, **决定停止还是继续**, 是防止过早闭合的关键
 - Workup Proposer: 把诊断不确定性转化为一组有**针对性的检查项目**





- Salus优化
 - 推理感知冷启动
 - 用高容量教师模型做全轨迹 rollout，**丢弃**教师过早停止且top1决策错误的SC数据
 - LLM-as-a-Judge 奖励建模
 - 因**医学术语多变同义**，不用字符串匹配，改用大模型作为裁判按四层rubric裁决，**产生语义化奖励**

输入

输出

$$\begin{aligned} X_{\text{DR}} &= (\mathcal{I}_{\text{DR}}, H_t), & Y_{\text{DR}} &= \{\text{CoT}_t^{\text{DR}}, D_t\} \\ X_{\text{SC}} &= (\mathcal{I}_{\text{SC}}, H_t, D_t), & Y_{\text{SC}} &= \{\text{CoT}_t^{\text{SC}}, s_t\} \\ X_{\text{WP}} &= (\mathcal{I}_{\text{WP}}, H_t, D_t), & Y_{\text{WP}} &= \{\text{CoT}_t^{\text{WP}}, E_t\} \end{aligned}$$

冷启动阶段 SFT 训练数据的格式

LABEL	DEFINITION
Completely Correct	Matches all core diagnoses; clinically identical to gold-standard diagnosis.
Clinically Acceptable	Core direction is correct with minor precision flaws; differences do not affect subsequent treatment plans.
Partially Correct	Clinically relevant but contains critical omissions or vagueness that significantly impact management.
Incorrect	Fundamentally incorrect, entirely deviated from gold-standard diagnosis.

诊断正确性的四层评判标准



• 数据资源及指标

– 任务与数据集：CompDiag-Bench — 训练 8,645 例 | 测试 752 例（269 高复杂 + 483 常规）；LLM + 人类专家双重验证

– 评价指标

• Top-3 Accuracy(诊断准确率)、Inf. Suff. (信息充分性)、#Rounds 、#Tests

– 底座模型

• Qwen2.5-7B Instruct，Salus-7B 和 Salus-Zero-7B 都基于它

• 对比方法

• 闭源模型：GPT-5.2、Gemini 3 Flash

• 通用开源：DeepSeek-V3.2、V3.1、GLM、Qwen

• 医学专用：Baichuan-M2-32B、MedGemma 27B

所有基线都接入同一三角色架构（DR/SC/WP），只靠task-specific prompting驱动、不做参数更新

高复杂案例诊断结果分析

– RL是性能提升的主要来源

- RL用稠密奖励把何时该继续查、差分列表该怎么排这类策略行为显式优化

- 而纯SFT只学到模仿教师轨迹，不具备这套策略判定能力

– 高检查成本换取精度与信息完备：

- SC的奖励结构惩罚过早闭合、不直接惩罚多查（成本未纳入奖励），所以策略天然偏向多查以确保充分

无强化学习训练

LLM	Top-1 (95% CI)	Top-2 (95% CI)	Top-3 (95% CI)	Inf. Suff. ↑ (95% CI)	#Rounds	#Tests
<i>Proprietary LLMs</i>						
GPT-5.2	80.30 (75.1–84.6)	85.50 (80.8–89.2)	86.25 (81.6–89.9)	71.00 (65.3–76.1)	1.75	27.61
Gemini 3 Flash	79.93 (74.7–84.3)	84.39 (79.6–88.2)	86.62 (82.0–90.2)	72.12 (66.5–77.1)	0.85	8.26
Claude Sonnet 4.5	71.75 (66.1–76.8)	76.95 (71.6–81.6)	79.93 (74.7–84.3)	70.63 (64.9–75.8)	1.19	14.75
GPT-5 mini	68.40 (62.6–73.7)	75.09 (69.6–79.9)	76.95 (71.6–81.6)	63.57 (57.7–69.1)	1.06	14.82
GPT-4.1	52.05 (46.1–57.9)	59.85 (53.9–65.5)	62.08 (56.2–67.7)	55.76 (49.8–61.6)	0.89	5.82
<i>Open-source LLMs</i>						
DeepSeek-V3.2	71.38 (65.7–76.4)	77.32 (72.0–81.9)	78.44 (73.1–82.9)	66.91 (61.1–72.3)	1.88	11.95
DeepSeek-V3.1	68.40 (62.6–73.7)	73.23 (67.6–78.2)	77.32 (72.0–81.9)	68.40 (62.6–73.7)	2.55	10.56
GLM-4.7	59.48 (53.5–65.2)	65.06 (59.2–70.5)	68.77 (63.0–74.0)	60.60 (54.6–66.2)	0.87	6.89
Qwen3-32B	43.12 (37.3–49.1)	49.81 (43.9–55.7)	53.53 (47.6–59.4)	52.05 (46.1–57.9)	1.11	7.00
Qwen2.5-72B-Instruct	39.40 (33.8–45.4)	46.84 (41.0–52.8)	49.44 (43.5–55.4)	54.65 (48.7–60.5)	1.07	7.45
Qwen3-8B	33.83 (28.4–39.7)	40.15 (34.5–46.1)	43.87 (38.1–49.8)	54.27 (48.3–60.1)	1.51	9.23
GLM-Z1-Flash	29.37 (24.2–35.1)	37.18 (31.6–43.1)	43.12 (37.3–49.1)	54.65 (48.7–60.5)	2.38	14.87
Qwen2.5-7B-Instruct	27.88 (22.9–33.5)	35.69 (30.2–41.6)	40.15 (34.5–46.1)	59.11 (53.1–64.8)	2.41	18.08
Qwen2.5-3B-Instruct	19.33 (15.1–24.5)	24.91 (20.1–30.4)	26.76 (21.8–32.4)	57.25 (51.3–63.0)	2.37	14.27
<i>Medical LLMs</i>						
Baichuan-M2-32B	48.70 (42.8–54.6)	53.53 (47.6–59.4)	56.13 (50.2–61.9)	56.51 (50.5–62.3)	3.10	25.98
MedGemma 27B	44.98 (39.1–51.0)	52.79 (46.8–58.7)	57.25 (51.3–63.0)	58.74 (52.8–64.5)	1.38	12.19
Baichuan-M3	40.15 (34.5–46.1)	43.49 (37.7–49.5)	46.10 (40.2–52.1)	53.53 (47.6–59.4)	1.31	6.28
Salus-Zero-7B	55.02 (49.0–60.9)	59.11 (53.1–64.8)	62.08 (56.2–67.7)	65.80 (59.9–71.2)	2.54	12.51
Salus-7B	83.64 (78.8–87.6)	86.62 (82.0–90.2)	88.85 (84.5–92.1)	72.86 (67.3–77.8)	2.91	19.98

模型又准、搜集证据又充分，但检查轮次与项目数偏多。

- 常规临床案例的泛化性实验

- 跨难度泛化

- 方法不依赖病例难度分布，在**低难度病例**上既不退化、又保持优势，**泛化性良好**

- 信息覆盖

- **常规病例**证据更易收敛，SC更容易判定已充分而提前终止，属合理的高**效行为**

LLM	Top-1 (95% CI)	Inf. Suff. ↑ (95% CI)
<i>Proprietary LLMs</i>		
Gemini 3 Flash	88.20 (83.7–91.4)	84.06 (79.2–87.9)
GPT-5.2	86.34 (81.6–89.9)	80.75 (75.5–84.9)
GPT-5 mini	80.95 (75.9–85.3)	<u>83.85</u> (79.2–87.9)
<i>Open-source LLMs</i>		
DeepSeek-V3.2	82.19 (77.1–86.3)	82.82 (77.9–86.9)
DeepSeek-V3.1	81.57 (76.3–85.6)	81.57 (76.3–85.6)
Qwen3-32B	76.19 (70.8–80.9)	76.19 (70.8–80.9)
Qwen3-8B	66.87 (61.1–72.3)	76.40 (71.2–81.2)
Qwen2.5-7B-Instruct	59.63 (53.5–65.2)	80.95 (75.9–85.3)
<i>Medical LLMs</i>		
MedGemma 27B	75.16 (69.6–79.9)	80.95 (75.9–85.3)
Baichuan-M2-32B	58.18 (52.4–64.1)	62.73 (56.9–68.4)
Salus-Zero-7B	78.67 (73.5–83.3)	83.23 (78.3–87.3)
Salus-7B	91.10 (87.1–93.9)	<u>81.99</u> (77.1–86.3)

够用即停：准确率高、信息充分性较低。



- 冷启动性能与SFT数据量正相关
 - 数据规模从12.5% → 25% → 50% → 100%，Top-1准确率依次为44.61 → 50.19 → 54.28 → 55.02，全程**单调上升**，增益 +10.4%
- 模型容量对冷启动效果关键
 - 更小的骨干容量不足以承载临床综合 + 策略控制 + 检查提出三角色的**多任务知识**，拟合能力受限
- 更强的教师不一定产生更好的SFT数据
 - 教师换成 **DeepSeek-V3.2** 时 Top-1 降到 43.12%，换成 **GLM-Z1-Flash** 低至 31.97%，均低于默认的 DeepSeek-V3.1

3因素消融实验结果

Factor	Configuration	Top-1	Inf.Suff.
Default	Salus-Zero-7B	55.02	65.80
Data Scale	50% Data	54.28	64.31
	25% Data	50.19	61.34
	12.5% Data	44.61	62.45
Model Size	3B Size Variant	43.87	60.60
Teacher	DeepSeek-V3.2	43.12	62.82
	GLM-Z1-Flash	31.97	62.45

并不是更强的教师就能产出更好的 SFT 数据



- DR 与 SC 的协同是关键
 - SC 的加入是性能的决定性因素
 - DR+SC 同时优化的准确率达83.64%，只训 DR时为61.71%，仅 SFT时为55.02%。
 - 没有学会策略的 SC，模型无法判断证据何时充分，会频繁过早闭合
 - 收益主要来自方法而非模型规模
 - RL优化后的3B模型达65.06%，超过Qwen3-32B 的 43.12%
 - 多智能体RL和结构化奖励把策略这一最难的部分直接优化了，弥补了小模型在参数规模上的劣势

Variant	DR	SC	Top-1	Inf.Suff.
Salus-7B	✓	✓	83.64	72.86
– w/o RL for SC	✓	–	61.71	65.43
– SFT only	–	–	55.02	65.80
Salus-3B	✓	✓	65.06	65.06

• 算法贡献

– 提出多智能体解耦框架 Salus

- 提出一个多智能体架构，将AI临床医生的**认知过程拆解为三个专职角色**。将临床推理与策略控制解耦，用SC充当**防止过早诊断闭合**的安全阀

– 两阶段优化范式

- 先用**推理感知的冷启动**从教师模型蒸馏临床推理与流程管理能力，为三个角色打好数据基础
- 再用带**逐步奖励的多智能体RL**进一步优化

• 算法不足

– 面向**文本记录**

– 未显式建模检查**成本与效率**

– 存在幻觉与数据偏差风险



特点总结与未来展望



- 特点总结

- ACE通过把上下文演化为操作手册，解决上下文适配的简洁偏置以及上下文坍塌的问题。

- 不改权重、无需标注。

- Salus通过把复杂诊断变成顺序决策的多智能体强化学习，解决LLM在临床诊断中的过早诊断闭合问题。

- 未来展望

- 发展更鲁棒的反思/评判机制，抗噪声、抗奖励欺骗

- 从依赖强监督/强教师走向更自监督的自改进

- 从固定角色走向自适应角色分配，并向跨领域通用化迁移

- 面向真实部署，研究多模态输入、低成本低延迟



- [1]Zhang Q. Hu C. Upasani S. et al. Agentic context engineering: evolving contexts for self-improving language models. Proceedings of the International Conference on Learning Representations (ICLR)[C]. 2026.
- [2]Gao S. Chen X. Luo L. et al. Salus: Strategic diagnostic testing for complex diagnosis via multi-agent reinforcement learning. Proceedings of the 43rd International Conference on Machine Learning (ICML)[C]. PMLR. 2026: 33039-33068.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！

