【原文翻译】

元强化学习激发语言智能体的探索能力

Yulun Jiang[1,2,*] **Liangze Jiang**[1,3,*] **Damien Teney**[3] **Michael Moor**[2,†] **Maria Brbić**[1,†]

[1]EPFL [2]ETH Zurich [3]Idiap Research Institute

*同等贡献 †同等指导

摘要

强化学习（RL）使大型语言模型（LLM）智能体能够与环境交互并解决多轮长视程任务。然而，经过RL训练的智能体通常在需要主动探索的任务中表现挣扎，并且无法有效地从试错经验中适应。在本文中，我们提出了 **LAMER**，这是一个通用的元强化学习（Meta-RL）框架，使LLM智能体能够在测试时主动探索并从环境反馈中学习。LAMER 包含两个关键组件：(i) 一个**跨回合训练框架**，旨在鼓励探索和长期奖励优化；以及 (ii) **通过反思进行的上下文内策略适应**，允许智能体根据任务反馈信号调整策略，而无需梯度更新。在多样化环境中的实验表明，LAMER 的表现显著优于 RL 基线，在 Sokoban（推箱子）、MineSweeper（扫雷）和 Webshop 上分别获得了 11%、14% 和 19% 的性能提升。此外，与 RL 训练的智能体相比，LAMER 在更具挑战性或以前未见过的任务中也表现出更好的泛化能力。总体而言，我们的结果表明，元强化学习提供了一种原则性的方法来诱导语言智能体中的探索行为，使其能够通过习得的探索策略更稳健地适应新环境。

1 引言

大型语言模型（LLMs）的最新进展已使其从构建对话系统转变为构建能够推理并与环境交互的决策智能体（Yao et al., 2023b; Shinn et al., 2023; Wang et al., 2025; Feng et al., 2025）。为了实现这一目标，语言智能体在多轮、文本观察-行动循环中运作，并且必须利用跨回合的记忆快速适应。这种适应的核心是**探索**，它允许智能体测试不确定的行动，获取新知识，并避免过早收敛于次优策略。然而，与能够系统性探索并在新环境中快速适应的人类不同（Wilson et al., 2014），如果没有实质性的干预，LLM智能体无法稳健地参与探索（Krishnamurthy et al., 2024）。

最近的工作已开始通过在测试时引导 LLMs 进行探索行为来解决这一局限性。例如，Tajwar et al. (2025) 离线训练模型以从不同环境的轨迹中蒸馏探索策略，而 Gandhi et al. (2024) 则从离线搜索轨迹中诱导此类策略。Setlur et al. (2025) 训练模型在上下文中学习探索，以此作为一种利用测试时计算资源的更好方式（Snell et al., 2025）。然而，这些工作要么主要关注单轮非代理推理问题，要么依赖于离线数据，这将其限制在模仿而非主动探索的范畴。

在这项工作中，我们迈向了能够**主动探索**其环境、收集反馈并利用这些经验进行更有效利用（exploitation）的智能体。由于多轮任务通常在以回合（episode）结束后才有一个稀疏的成功信号，我们考虑一种多回合机制（Shinn et al., 2023），其中一个回合是探索和利用的基本单元。平衡探索与利用可以自然地表述为一个跨回合的强化学习（RL）框架。在该框架下跨越许多相似但不同的环境进行训练，即导向了元强化学习（Meta-RL）（Duan et al., 2016; Wang et al., 2016; Bauer et al., 2023; Beck et al., ...

页边注记：arXiv:2512.16848v1 [cs.LG] 18 Dec 2025

【高三解读】

核心概念：教 AI 像“探险家”一样思考

各位同学，试想一下你第一次玩《扫雷》或者《塞尔达传说》时的场景。你是不可能一上来就完美通关的，对吧？你会先试探性地点击几个格子，或者在地图上乱跑（这是**探索**，Exploration）；等你“挂”了几次，摸清了怪物的规律或雷区的逻辑后，你就会利用这些经验去快速通关（这是**利用**，Exploitation）。

这篇文章解决的就是当前顶尖 AI（如 ChatGPT）的一个大毛病：**它们是很好的“做题家”，但不是好的“探险家”。**

1. **现状**：现有的 AI 智能体（Agent）大多是靠“背题库”（训练数据）长大的。如果把它们扔到一个从未见过的陌生环境中，它们往往束手无策，不知道如何通过试错来积累经验。
2. **创新点 LAMER**：作者提出了一种叫 LAMER 的方法，核心思想是 **Meta-RL（元强化学习）**。
   - **普通的 RL（强化学习）** 就像训练一只狗，做对了给骨头，做错了不给，它只能学会特定的动作。
   - **Meta-RL（元强化学习）** 则是教这只狗“如何学习新把戏”。不仅仅是学会这一个任务，而是学会“当我遇到新任务时，我该怎么试探、怎么总结教训”。

难点解析：不仅要“学”，还要“学会如何学”

这里有两个非常硬核的概念，我们来拆解一下：

1. 跨回合训练（Cross-episode training）：

- **高三视角**：这就像你们做理综卷子。如果你做第一遍（Episode 1）发现时间不够，第二遍做模拟考（Episode 2）时，你就会调整策略，比如“先做生物，最后做物理大题”。
- **AI 视角**：传统的 AI 训练往往把每一次尝试都看作独立的。但 LAMER 框架让 AI 记住前几次失败的教训（比如“上次走左边掉坑里了”），并将这种记忆带入下一次尝试中。这实际上是在训练 AI 的**长期记忆**和**策略调整能力**。

2. 上下文内策略适应（In-context policy adaptation via reflection）：

- **难点**：通常 AI 更新自己需要“梯度下降”，这需要巨大的计算量和修改模型参数（这就好比你每学一个新单词都要动手术改写大脑神经元，太慢了）。
- **突破**：这里提到的“无需梯度更新”是指，AI 像人类一样，通过“写日记”来反思。它在输入框（Context）里对自己说：“刚才那个策略不行，因为线索 A 没注意到，下次我得注意 B。” 这种**反思（Reflection）**直接变成了下一次行动的指令。这就好比你在草稿纸上写下“易错点”，下次做题直接看草稿纸提醒自己，而不需要重新背一遍课本。

知识联想：跨学科的智慧

为了帮大家更好地建立知识体系，我们可以把这些概念映射到你们熟悉的学科中：

- **生物学（进化与适应）**：
  文章提到的“探索（Exploration）”本质上就是生物进化中的**基因突变**。如果生物只保持现状（Exploitation），一旦环境改变就会灭绝。只有不断尝试新的变异，才能在自然选择中找到新的出路。AI 现在的困境就是“变异”不足，太保守了。

- **数学（函数最值与局部最优）**：
  在导数大题中，我们要找函数的最大值（Global Maximum）。现在的 AI 往往容易陷入**局部最优（Local Maximum）**——就像爬山爬到一个小土坡顶就以为到了珠峰，不再走了。本文的 Meta-RL 机制，就是给 AI 一个动力，让它由“贪心算法”（只看眼前）转变为“模拟退火”或更高级的全局搜索策略，敢于下坡去寻找更高的山峰。
- **心理学（元认知）**：
  文章标题中的 "Meta-"（元）是一个非常高级的前缀，对应心理学中的**元认知（Metacognition）**，即“对思考的思考”。普通学生只是“做题”，学霸会“分析我为什么做错题”。LAMER 就是赋予 AI 这种“学霸思维”，让它具备自我监控和自我调节的能力。

总结：这篇论文通过教会 AI“如何从失败中学习”，让它们在扫雷、网购等复杂任务中变得更聪明。对于你们来说，这也是一种启示：**在学习中，单纯的“刷题量”（RL）不如“从错题中总结规律的方法论”（Meta-RL）来得有效！**

第 2 页

【原文翻译】

图表内容描述

图 1：扫雷（MineSweeper）环境中 RL（强化学习）与 Meta-RL（元强化学习）训练的对比。

- **左图（散点图）：** 展示了“成功率（Success Rate %）”与“轨迹多样性（Trajectory Diversity）”的关系。
  - **基座模型（Base Model）：** 位于左下角，成功率低，多样性低。
  - **RL（黄点）：** 位于左侧中部，成功率有所提升（约 60%），但多样性依然很低（约 0.5 - 1.0）。
  - **Meta-RL（绿点）：** 位于右上角，成功率最高（>70%），且多样性显著增加（>3.0）。箭头展示了训练演进方向：从基座模型到 RL，再到 Meta-RL，实现了性能与多样性的双重提升。
- **中图（直方图）：RL 轨迹分布**
  - 显示了不同轨迹（Trajectory Index）出现的概率（Probability）。
  - 大部分概率集中在少数几个轨迹上（左侧高耸的蓝色柱状条），表明行为模式非常单一、重复。
  - 插图显示了扫雷棋盘上的点击点，由于策略单一，覆盖范围有限。
- **右图（直方图）：Meta-RL 轨迹分布**
  - 概率分布更加平坦、宽广，表明智能体尝试了大量不同的轨迹。
  - 插图显示棋盘上的点击点分布更加广泛，体现了更强的探索性。

图注原文翻译：
图 1：扫雷环境中 RL 和 Meta-RL 训练的对比。**左图：** 使用 LaMER 的 Meta-RL 训练在保持比基座模型更高样本多样性的同时，实现了更高的成功率，在“探索（exploration）”与“利用（exploitation）”之间达到了更好的权衡。**右图：** 在扫雷环境中，聚合多个采样轨迹得到的不同轨迹及其经验概率。每条轨迹对应棋盘上的一系列点击（带数字的单元格）。样本多样性通过经验分布的熵（entropy）来量化。Meta-RL 训练的模型产生了更加多样化且具有探索性的轨迹。

正文翻译

...2025），此时智能体被迫去发现那些在未见过且可能更困难的环境中依然有效的通用策略。

在此基础上，我们提出了 **LaMER**（**L**LM **A**gent with **Me**ta-**R**L，基于元强化学习的大语言模型智能体），这是一个用于 LLM 智能体训练的通用 Meta-RL 框架。LaMER 包含两个关键的设计原则。首先，与标准的单回合（single-episode）RL 不同，LaMER 是围绕**多回合结构（multi-episode structure）**设计的，旨在训练智能体通过试错来解决问题。在早期的回合中，鼓励智能体从环境中收集多样化的经验和信息反馈，这些信息随后被用于调整其策略以适应后续的回合。通过最大化跨回合的长期奖励，智能体内化了一种学习算法，该算法显式地激励探索，从而改善下游的利用（exploitation）效果。其次，在训练和测试阶段，智能体都能有效地利用前几回合的反馈和反思来决定下一回合的策略。这本质上是在上下文（in context）**中实现了一种 RL 算法，使该方法比标准 RL 能够产生更多样化的样本，同时实现更高的性能，从而在探索和利用之间达到更好的平衡（见图 1）。据我们所知，这是首次将 Meta-RL 框架用于 LLM 智能体的训练。

我们在四个具有挑战性的长视界（long-horizon）任务上评估了 LaMER：推箱子（Sokoban，Racanière et al., 2017）、扫雷（MineSweeper，Li et al., 2024）、网上购物（Webshop，Yao et al., 2022）和 ALFWorld（Shridhar et al., 2021）。使用 Qwen3-4B（Yang et al., 2025），我们要证明 LaMER 在所有环境中均始终优于提示（prompting）和 RL 基线方法。我们观察到，训练后的模型学会了在探索和利用之间进行平衡，从而提高了测试时的扩展性能（test-time scaling performance）。具体而言，LaMER 在测试时调整了训练好的策略，与 RL 相比，分别在推箱子、扫雷和网上购物任务上实现了 11%、14% 和 19% 的绝对性能提升。此外，我们要展示 LaMER 训练的模型能更好地泛化到更困难和分布外（out-of-distribution）的任务中。总之，LaMER 向能够主动行动以揭示信息并改善在新环境中决策的自主智能体迈出了一步。

2 相关工作

LLM作为智能体（LLM-as-agent）。 随着大语言模型（LLMs）在复杂场景推理能力上的增强（Wei et al., 2022），人们对让它们在自主智能体中进行决策越来越感兴趣。早期的工作依赖于对冻结参数的 LLM 进行提示（Yao et al., 2023b; Shinn et al., 2023; Park et al., 2023; Wang et al., 2024a; AutoGPT）。ReAct（Yao et al., 2023b）通过上下文示例提示 LLM 生成文本行动和推理思维。随后，Reflexion（Shinn et al., 2023）将这一原则扩展到多回合设置中，智能体口头反思上一回合的表现，并为下一回合维护自己的反思缓冲区。最近的研究通过设计先进的 RL 算法（Wang et al., 2025; Feng et al., 2025）用于多轮交互，或在生成的交互轨迹上进行监督微调（supervised fine-tuning）来训练 LLM 智能体以完成各种任务（Tajwar et al., 2025）。由于与环境的完全口头交互，LLM 智能体的评估也面临挑战。最近的基准测试涵盖了广泛的领域，包括基于文本的...

【高三解读】

高三·深度解读：教AI学会“如何学习”

你好！这一页的内容非常硬核，它介绍了一种让 AI 变得更聪明、更灵活的新方法，叫 **LaMER**。如果把训练 AI 比作培养一个高中生，那么这一页讲的就是如何从“死记硬背”进化到“掌握解题套路”。

1. 核心概念：LaMER 是什么？

这就好比我们在准备高考数学。

- **标准 RL（强化学习）** 就像是一个学生做题，做对了有分，做错了没分。他为了拿高分，一旦发现一种解法（比如“暴力代入法”）能得分，就死守这个方法，不再尝试别的。这导致他**只会一种套路**（图表中“RL Trajectories”那个又高又瘦的柱子），遇到新题型就挂了。
- **Meta-RL（元强化学习，即 LaMER 的核心）** 则是教会学生“如何去试错”。老师告诉他：“这道题你先试探一下，第一遍做错没关系，关键是要从错误中吸取教训，调整策略，第二遍、第三遍把它做对。”
- **结果：** 使用 LaMER 的 AI，不仅分数更高（图1左图绿点），而且解题思路非常开阔（图1右图分布很广）。它不再是只会背答案的机器，而是一个懂得**探索（Exploration）**和**利用（Exploitation）**的策略家。

2. 难点拆解：为什么“多样性”这么重要?

文中反复提到“Trajectory Diversity（轨迹多样性）”和“Entropy（熵）”，这其实触及了智能的本质。

- **知识联想（物理/化学）：熵（Entropy）**
  在高中物理和化学里，熵是衡量系统“混乱度”或“无序度”的指标。但在信息论和 AI 里，**高熵**通常是好事，代表**信息量大、可能性多**。
  - 图1中间的蓝色直方图，柱子集中在一点，这是**低熵**。就像全班同学都写一样的作文，千篇一律，没有创新。
  - 图1右边的灰色直方图，柱子铺得很开，这是**高熵**。就像每个人都有独特的见解，百花齐放。
- **探索 vs. 利用（Exploration vs. Exploitation）**
  这是 AI 领域最经典的博弈，也可以对应我们的人生规划：
  - **利用（Exploitation）：** 总是去那家你最喜欢的面馆吃面，因为你知道那里肯定好吃（奖励确定），但你永远失去了发现更好吃餐馆的机会。
  - **探索（Exploration）：** 每次都去试一家新开的店。可能会踩雷（奖励低），但也可能发现绝世美味（长期奖励高）。
  - **LaMER 的突破：** 普通的 RL 太喜欢“利用”（吃老本），而 LaMER 强迫 AI 在初期多“探索”（试错），从而在长期获得更高的回报。文中提到它在“扫雷”游戏中表现优异，就是因为它学会了先点开几个格子收集信息，而不是盲目乱猜。

3. 深度逻辑：In-Context RL（上下文中的强化学习）

这一段有个很高级的概念叫“implements an RL algorithm in context”。

- **传统的学习**是改写大脑（修改神经网络的参数权重 w），这通常很慢，需要大量计算。
- **In-Context Learning** 是大模型特有的能力。就像你在考试时，仅仅通过读题目和回忆刚才的草稿（上下文 Context），大脑参数没变，但你的解题策略已经变了。
- **LaMER 的魔法**在于，它让 AI 把“试错-反思-修正”这个过程，直接在对话框（Context）里完成了。AI 在前几轮对话中对自己说：“刚才那招不行，我得换个方向。”这种能力让它在面对从未见过的难题（分布外任务）时，适应力极强。

4. 总结与启示

这篇文章告诉我们，真正的智能不是“不犯错”，而是“善于从错误中学习”。

- **对于 AI：** 未来的 AI 助手（Agent）将不再是冷冰冰的问答机器，它们会像一个有经验的探险家，主动尝试、主动规划。
- **对于你：** 在高三复习中，不要害怕做错题。如果你能像 LaMER 一样，建立一个“多回合机制”——做题、改错、反思、再做题，那你就是在训练自己的大脑进行“元学习”，这种能力比单纯刷题更重要，也是应对未来高考和人生挑战的通用策略！

第 3 页

【原文翻译】

具身环境（Shridhar et al., 2021）、电子商务网站（Yao et al., 2022）、老虎机（Nie et al., 2024）、经典游戏（Park et al., 2025; Li et al., 2024）以及其他任务（Liu et al., 2024; Nathani et al., 2025）。关于这些工作的更全面综述，我们建议读者参考最近的调查报告（Wang et al., 2024b; Zhang et al., 2025）。

元强化学习（Meta-RL）（Beck et al., 2025）关注“学会强化学习”，以便快速适应新环境。类似于元学习（Thrun & Pratt, 1998; Hospedales et al., 2021），它包含一个本身代表RL算法的内循环（即适应策略），以及一个更新元参数的外循环，从而使内循环在许多任务中变得更加有效。通过在许多任务上进行训练，外循环迫使智能体学习解决任务所需的探索策略，而内循环则使智能体能够根据探索结果快速适应。根据内循环的实现方式，分为上下文方法（in-context methods）和基于梯度的方法（gradient-based methods）。例如，Duan et al. (2016); Wang et al. (2016); Stadie et al. (2018) 将内循环表示为一个由RNN参数化的依赖历史的策略，因此适应是通过收集存储在记忆状态中的更多信息“在上下文中”完成的。另一方面，Finn et al. (2017) 利用基于梯度的方法，其中内循环适应由外循环学习的一般元策略。我们的工作属于前一类，即适应完全在测试时的上下文中发生，自然地利用了LLM的上下文学习能力。

测试时计算（Test-time compute）。 LAMER中的Meta-RL框架可以被视为通过以多回合（multi-episode）而非单回合的方式训练任务来**摊销**（amortizing）测试时计算。通过这种方式，学习到的上下文策略适应平衡了探索（exploration）和利用（exploitation），以在测试时实现快速适应。这本质上是一种更好的测试时计算开销方式（Snell et al., 2025; Muennighoff et al., 2025; Wu et al., 2025; Setlur et al., 2025）。在我们的实验中，我们匹配了RL和Meta-RL基线之间的训练计算预算，并表明Meta-RL鼓励了更好的测试时缩放行为（通过pass@k衡量）。Qu et al. (2025) 同样将Meta-RL与测试时计算联系起来，但他们局限于数学推理的单轮问题，没有利用来自环境的交互式反馈。

LLM中的推理。 更广泛地说，这项工作与LLM中的推理有关，因为语言智能体必须使用推理作为其决策的一部分。最近关于LLM推理的大量工作集中在更高级的提示工程（Wei et al., 2022; Yao et al., 2023a）、后训练（Cobbe et al., 2021; Luong et al., 2024; Shao et al., 2024; DeepSeek-AI et al., 2025）或针对验证器或奖励模型的自举（bootstrapping）（Zelikman et al., 2022），诱导结构化搜索行为（Gandhi et al., 2024; Moon et al., 2024），或反思先前的答案（Kumar et al., 2024; Xiong et al., 2024; Qu et al., 2024）等。这些工作大多数集中在单轮数学（Hendrycks et al., 2021b; Cobbe et al., 2021）和编码（Chen et al., 2021; Hendrycks et al., 2021a）问题上，而我们的目标是多轮智能体环境，在这些环境中，每次行动后和回合结束时都能获得环境反馈。

3 预备知识

我们考虑这样一个场景：一个LLM智能体与环境交互以解决多轮任务。这个过程可以被形式化为一个马尔可夫决策过程（MDP） $\mathcal{M}$ = ($\mathcal{S}$, $\mathcal{A}$, P, R, γ_step)，其中 $\mathcal{S}$和 $\mathcal{A}$ 分别表示状态空间和动作空间，R 是奖励函数。在每个时间步 t = 0, ..., T - 1，LLM智能体观察到一个状态 s_t ∈ $\mathcal{S}$并根据其策略 a_t ∼ π_θ(·|s_t) 选择一个动作 a_t ∈ $\mathcal{A}$。然后环境提供一个标量奖励 r_t ∈ $\mathbb{R}$ 并根据转移函数 P(· | s_t, a_t) 转移到下一个状态

s_{t+1}。轨迹是整个回合中的状态、动作和奖励的序列，即 τ = (s_0, a_0, r_0, ..., s_{T-1}, a_{T-1}, r_{T-1})。强化学习的目标是最大化期望折扣回报：

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\gamma_{step}^t r_t\right] \quad (1)$$

其中 γ_step ∈ [0, 1] 是折扣因子。最近的工作（Wang et al., 2025; Feng et al., 2025）表明，RL训练已经使LLM智能体能够与环境交互并解决多轮任务。然而，这些智能体通常在训练期间学习一个固定的策略，而在**测试时**难以主动探索并适应其行为以应对任务（Nie et al., 2024）。

3

【高三解读】

核心概念：从“死记硬背”到“学会生存”

同学们，这一页虽然充斥着复杂的术语，但核心思想非常精彩，它在讨论如何让AI（特别是像ChatGPT这样的大语言模型）变得更聪明、更具适应性。我们可以把这页内容概括为：**如何训练AI不仅通过“刷题”来掌握知识，更是通过“实战演练”学会一套通用的解题策略，以便在从未见过的考试（测试时）中能迅速适应并拿高分。**

这里有三个关键点：

1. **元强化学习（Meta-RL）**：这是本文的主角。普通的强化学习（RL）像是在训练一只狗做特定的动作（比如握手）；而“元”强化学习是在训练这只狗具备“快速领悟新指令”的能力。它包含两个循环：
   - **内循环（Inner-loop）**：像是AI在考场上的即时反应，它根据题目的反馈快速调整策略。
   - **外循环（Outer-loop）**：像是老师的长期指导，它不断调整AI的大脑参数，让AI的“即时反应”能力越来越强。
2. **测试时计算（Test-time Compute）**：这指的是AI在真正面对问题时（测试阶段）所花的“思考功夫”。作者认为，通过一种特殊的训练方式（LAMER框架），可以让AI在考试时更聪明地分配精力，前期多尝试（探索），后期稳拿分（利用）。
3. **马尔可夫决策过程（MDP）**：这是所有强化学习的数学地基，用来描述“现在的状态”、“我做什么动作”以及“后果是什么”这一连串过程。

难点解析：深挖那些“不明觉厉”的概念

1. “摊销（Amortizing）”是什么意思?
文中提到“amortizing the test-time compute”。在经济学里，“摊销”指把一笔大额开支分摊到很长的时间里。在这里，它是指：我们在训练阶段虽然花了很多精力和时间（大额开支）让AI去练习怎么适应环境，但一旦它学会了，在真正考试（测试时）的时候，它就能非常快地适应新题目，不需要从头学起。这就好比你高三辛苦一年（训练成本），换来的是高考时解题如行云流水（测试时的高效），这笔投入被“摊销”并在考试中获得了回报。

2. 期望折扣回报（Equation 1）的数学含义
公式 $\mathbb{E}\left[\sum \gamma^t * r_t\right]$ 看起来很吓人，其实它是高中数学知识的直接应用：

- **∑ (求和)**：把每一步吃到的“糖果”（奖励 r）加起来。

- **γ^t (折扣因子)**：这是一个等比数列的项！γ 是一个 0 到 1 之间的数（比如 0.9）。
  - 当 t=0 时，γ^0 = 1，当下的奖励完全算数。
  - 当 t=100 时，γ^100 接近于 0，意味着未来的奖励在当前看来价值很低。
  - **含义**：这告诉AI，“落袋为安”，眼前的奖励比遥远的画饼更重要，同时也防止了如果游戏无限进行下去，分数加到无穷大的数学问题。

3. In-context Learning (上下文学习) vs. Gradient-based (基于梯度)
想象你要学会打一个新的电子游戏：

- **基于梯度**：你需要暂停游戏，把大脑拆开，重新接几根神经（修改参数），然后再玩。这很慢。
- **上下文学习（本文采用的）**：你不动大脑结构，而是根据刚才几分钟的操作记忆（Context），立刻意识到“哦，这个Boss怕火攻”，然后马上改变打法。这就是大语言模型的强项。

知识联想：跨学科的智慧

1. **数学（数列与极限）**：
   公式 (1) 本质上就是一个**无穷递缩等比数列求和**的模型（当 T 趋向于无穷大时）。我们在高中学的 $S = a_1/(1-q)$ 在这里不仅仅是数字游戏，它是AI判断“长期利益”与“短期利益”的数学工具。如果 γ 设置得太小，AI就会变得目光短浅（只顾眼前）；设置得太大，AI可能会为了虚无缥缈的未来而忽略当下的危机。
2. **生物学（进化与适应）**：
   文中提到的“外循环”和“内循环”完美对应了生物学中的**“进化”与“学习”**。
   - **外循环 = 进化**：人类花了几百万年进化出聪明的大脑结构（Meta-parameters），这是写在基因里的，改变很慢。
   - **内循环 = 学习**：你用这颗大脑在十几分钟内学会了一个新游戏，这是神经元活动的即时调整（In-context），速度很快。
     Meta-RL 就是试图在计算机里重现这种“进化出超强学习能力”的过程。
3. **心理学（流体智力 vs. 晶体智力）**：
   - 传统的AI训练像是在积累“晶体智力”（死记硬背的知识库）。
   - 本文强调的 Meta-RL 训练的是“流体智力”（面对新问题的推理和适应能力）。
     作为高三学生，你们现在的复习也不应只是背题（晶体），而是要通过做题去掌握解题的通法（流体），这样无论高考题怎么变，你都能像本文设想的AI一样，“Test-time adaptation”（测试时自适应），做到游刃有余！

第 4 页

【原文翻译】

Meta-RL。相反，通过在任务分布上进行训练，元强化学习（Meta-RL，参考 Duan et al., 2016; Wang et al., 2016; Bauer et al., 2023; Beck et al., 2025）鼓励探索，因为它优化的是元参数，从而使智能体能够快速解决新任务。在我们的案例中，元参数即大语言模型（LLM）的参数。这要求智能体学习适用于所训练任

务分布的通用“探索-利用”（exploration-exploitation）策略。例如，对于部分可观测环境中的大多数导航任务，最优策略是在第一个回合（episode）收集环境信息并定位目标，然后在第二个回合尽可能高效地到达目标。这种由智能体实施的*探索而后利用（explore-then-exploit）*策略本身就是一种强化学习算法，其中的学习策略编码了如何根据与新任务互动的阶段，自适应地在信息收集和奖励最大化行为之间切换。对于在多轮任务中操作的 LLM 智能体，这种策略可以在上下文中进行（即测试时无需参数更新），从而自然地利用 LLM 的上下文能力。

4 LAMER：面向 LLM 智能体的元强化学习框架

采用 Meta-RL 原则，我们提出了 LAMER，这是一个训练 LLM 智能体的框架，使其具备从环境中主动探索和自适应学习的能力。该框架解决了两个核心挑战：(i) 如何在对一个任务的多次尝试中平衡探索与利用，以及 (ii) 如何在训练和评估期间高效地调整策略。为了解决第一个挑战，LAMER 引入了一种*跨回合训练方案（cross-episode training scheme）*，将每一次试验（trial）视为一系列回合的序列，使智能体能够在早期回合中探索，并在后续回合中利用这些信息。其次，LAMER 不依赖基于梯度的更新，而是使用*自我反思（self-reflection）作为一种上下文适应机制*，允许智能体总结过去的经验并相应地调整其策略。这两个组件共同实现了 LLM 智能体在统一 Meta-RL 框架下的可扩展训练，并可以通过标准 RL 算法进行优化。

跨回合训练框架。在 LAMER 的训练中，每一次试验由智能体按顺序生成的 N 个回合组成：

$$\mathscr{T} = (\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(N-1)}), \text{ 其中 } \tau^{(n)} \sim \pi\_\theta^{(n)}(\cdot),\ n \in [0, N-1], \quad (2)$$

其中 $\pi\_\theta^{(n)}(\cdot)$ 是第 n 个回合的策略，该策略根据积累的历史记录 $\tau^{(0)}, \ldots, \tau^{(n-1)}$ 通过某种适应策略更新而来。为了简化分析，我们假设所有回合都包含与环境互动的 T 个步骤，即 $\tau^{(n)} = (s_0^{(n)}, a_0^{(n)}, r_0^{(n)}, \ldots, s\_{T-1}^{(n)}, a\_{T-1}^{(n)}, r\_{T-1}^{(n)})$ 对于所有 $n \in [0, N-1]$。如果 $\tau^{(n)}$ 成功（由环境反馈指示），则滚动过程在 n 处终止。否则，智能体将从相同的初始状态开始新的回合 $\tau^{(n+1)}$，重复此过程直到达到最大回合预算。对于动作 $a_t^{(n)}$，回合 $\tau^{(n)} \in \mathscr{T}$ 内的折扣回报 $g_t^{(n)}$ 为：

$$g_t^{(n)} = \sum\_{l=t}^{T-1} \gamma\_{step}^{l-t}\, r\_l^{(n)}, \quad (3)$$

其中 $\gamma\_{step} \in [0, 1]$ 是回合内（within-the-episode）折扣因子。

为了增强探索并最大化长期奖励，在 LAMER 框架中，我们定义了跨越 $\mathscr{T}$ 中各回合的折扣回报 $G_t^{(n)}$ 为：

$$G_t^{(n)} = g_t^{(n)}\ [\text{回合内部分}] + \sum\_{m=n+1}^{N-1} \gamma\_{traj}^{m-n}\, g_0^{(m)}\ [\text{跨回合部分}], \quad (4)$$

其中 $\gamma\_{traj} \in [0, 1]$ 是跨回合（cross-episode）折扣因子。最后，LLM 智能体通过以下 Meta-RL 目标函数进行训练：

$$J(\theta) = \mathbb{E}\_{\mathscr{T} \sim \pi\_\theta}\left[\sum\_{n=0}^{N-1} \gamma\_{traj}^{n} \sum\_{t=0}^{T-1} \gamma\_{step}^{t}\, r\_t^{(n)} \right] = \mathbb{E}\_{\mathscr{T} \sim \pi\_\theta}\left[G_0^{(0)} \right]. \quad (5)$$

【高三解读】

核心概念：让AI学会“玩通关游戏”的智慧

这就好比你在玩一个全新的、难度很高的闯关游戏（比如《超级马里奥》或者密室逃脱），你只有 N 条命（N个回合）。

本页介绍的 **LAMER 框架**，就是教 AI 如何利用这 N 条命来拿到最高分。这里面有两个核心思想：

1. **“先侦查，后收割”（Meta-RL 的精髓）：**
   - 普通的 AI 可能每次玩都像个愣头青，只顾眼前的利益。
   - 但是 LAMER 训练出来的 AI 懂得“策略”。比如，第 1 条命（第 1 回合），它可能根本不在乎得分，而是到处乱跑（**探索**），把地图背下来，找到宝藏的位置。虽然第 1 条命可能很快就挂了或者得分很低，但它获得了关键信息。
   - 到了第 2、3 条命（后续回合），它就利用之前的记忆，直奔宝藏，疯狂得分（**利用**）。
   - **这就是“元强化学习”（Meta-RL）**：它不是在学“怎么走这个迷宫”，而是在学“到了一个陌生迷宫该怎么快速上手”。
2. **“写日记”代替“换脑子”（In-context Reflection）：**
   - 传统的 AI 学习需要修改大脑神经元连接（参数更新），这很慢也很麻烦。
   - LAMER 让 AI 在每次失败后，给自己写一段总结（**自我反思**），比如“刚才走到那个角落掉坑里了，下次别去”。下一局开始时，它读着自己的总结接着玩。这利用了大语言模型强大的上下文理解能力。

难点解析：那个看起来很复杂的公式 $G_t^{(n)}$

看到公式 (4) 里的 $G_t^{(n)}$ 了吗？别被求和符号 Σ 吓倒，我们来拆解它：

- **情景**：假设你正在玩第 n 局游戏。
- $g_t^{(n)}$ **（前半部分）**：这是你**当前这一局**能拿到的分数。这很好理解。
- $\sum ...g_0^{(m)}$ **（后半部分）**：这是你**未来所有局**能拿到的分数的总和！
- **为什么要把未来的分也加到现在？** 这就是 LAMER 的高明之处！它告诉 AI：“嘿，你在第 1 局里的表现，不仅仅看你现在拿了多少分，还要看你是否为第 2 局、第 3 局打好了基础。”如果你在第 1 局花时间找到了捷径，虽然第 1 局没通关，但因为你为未来贡献了巨大的价值，系统依然认为你第 1 局表现很棒（$G_t^{(n)}$ 会很高）。
- γ_{traj} **（折扣因子）**：这是一个 0 到 1 之间的数（比如 0.9）。它的意思是“未来的奖励不如现在的奖励值钱”。就像我有 100 块钱，现在给你，比一年后给你更有吸引力。这在数学上保证了求和是收敛的，在逻辑上让 AI 稍微偏向于早点成功，而不是无限拖延。

知识联想：连接你的高中知识库

1. **数学（数列求和）**：
   - 看公式 (3) 和 (4) 中的 Σ，这是典型的**等比数列求和**的变体。如果你把奖励 r 看作常数，这就完全是高中数学必修五里的 $S_n = \frac{a_1(1-q^n)}{1-q}$ 的应用场景。这里的 q 就是折扣因子 γ。
2. **生物/心理学（元认知）**：
   - 我们在学习时，不仅仅是“背单词”（学习任务），更重要的是学会“怎么背单词才快”（学习方法）。后者就是**元认知（Metacognition）**，即“对思考的思考”。LAMER 实际上就是在赋予 AI 这种元认知

能力，让它学会调整自己的学习策略。

3. **物理（势能与动能）**：
   - 第一局的“探索”就像是把石头推上山坡（积累重力势能），虽然看起来在做负功（没得分甚至牺牲了），但转化为了巨大的势能。第二局的“利用”就是石头滚下来（释放动能），势能转化为了实实在在的分数。公式 (4) 就是在计算这个“总机械能”守恒。

总结：这一页不仅展示了数学的严谨美，更揭示了 AI 如何从“死记硬背”进化到“懂得反思和规划”的高级智能形态。你看，数学公式其实就是描述智能行为的语言！

第 5 页

【原文翻译】

在这里，γ_traj 是平衡*探索（exploration）*与开发（exploitation）*的重要因子。理想情况下，较小的 γ_traj 会使目标偏向早期回合，从而导致为了解决问题而进行快速的开发。相比之下，较大的 γ_traj 强调长远的回报，因此鼓励在早期阶段进行更多的探索。

图表内容描述

图 2：LAMER 中使用的 RL（上图）和 Meta-RL（下图）训练过程的比较

训练模式	流程描述
RL Training (强化学习训练)	**Agent (智能体)** 生成一组独立的轨迹。 Episode $\tau^{(0)}$ (独立) → Episode $\tau^{(1)}$ (独立) → ... → Episode $\tau^{(N-1)}$ 轨迹之间相互**独立 (Independent)**，互不影响。
Meta-RL Training (元强化学习训练)	**Agent (智能体)** 按顺序生成轨迹，并进行自我反思。 Episode $\tau^{(0)}$ → **Reflection (反思/放大镜)** → Episode $\tau^{(1)}$ → **Reflection (反思/放大镜)** → Episode $\tau^{(N-1)}$ 轨迹通过 γ_{traj} 关联，利用**反思**来调整下一回合的策略。

图注：对于单个任务，RL 独立地生成一组轨迹。相比之下，在 LAMER 中，我们使用 Meta-RL 按顺序生成轨迹，并通过自我反思（self-reflection）在上下文（in-context）中调整策略。轨迹折扣因子 γ_traj 用于跨回合的信用分配（credit assignment）。

基于自我反思的上下文策略适应

(In-context policy adaptation with self-reflection)

在 Meta-RL 中，策略适应是 LLM 代理学习过程的内循环。因此，灵活且高效的适应机制在训练期间起着重要作用，而像梯度下降（Finn et al., 2017）这样的方法可能过于昂贵，特别是对于 LLMs 而言。在 LAMER 中，我们提出了一种基于自我反思的策略（Shinn et al., 2023），以*在上下文（in-context）中适应策略*

（Brown et al., 2020; Laskin et al., 2023）。具体来说，在每个回合结束后，我们提示代理生成关于前一次尝试的文本反思，提供具体的反馈和计划以指导下一回合（用于所用提示词的详情见附录 A）。因此，策略通过修改包含历史轨迹和反思的上下文来更新。

$\pi_\theta^{(n)}(\cdot) = \pi_\theta(\cdot|\mathcal{H}^{(n)})$

其中 $\mathcal{H}^{(n)}$ 表示包含历史轨迹和反思的跨回合记忆。重要的是，自我反思步骤在 LAMER 中也是显式训练的，使用的是在下一回合中获得的奖励。注意，内容 $\mathcal{H}^{(n)}$ 可以根据预定义的记忆缓冲区进行调整，以减少上下文长度并提高效率。默认情况下，我们在 $\mathcal{H}^{(n)}$ 中保留历史和反思，并在第 6.2 节提供消融研究。

与 RL 训练的比较

(Comparison to the RL training)

与 RL 目标（公式 1）相比，Meta-RL 将信用分配扩展到多个回合，以激励早期阶段的探索。在实践中，给定单个任务，RL 和 Meta-RL 都会在训练期间采样一组回合以估计优势（advantage）。关键的区别在于，RL 的滚动（rollouts）是独立的，而在 Meta-RL 中，每个回合都以试验中的前一个回合为条件。图 2 展示了 RL 和 Meta-RL 训练过程之间的概念差异。

优化

(Optimization)

公式 (5) 中提出的 Meta-RL 目标可以通过标准的策略梯度方法进行优化。给定上面定义的每个动作的跨回合回报 $G_t^{(n)}$，梯度可以通过下式估计：

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta}\left[\sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(n)}|s_t^{(n)}, \mathcal{H}^{(n)})\, A_t^{(n)} \right], \quad (6)$$

其中 $A_t^{(n)}$ 是从 $G_t^{(n)}$ 推导出的优势估计。该框架与广泛使用的优化器如 PPO（Schulman et al., 2017）以及无评论家（critic-free）的方法如 GRPO（Shao et al., 2024）和 GiGPO（Feng et al., 2025）兼容。

【高三解读】

核心概念：让 AI 学会“写日记”和“复盘”

同学你好！这张页面主要介绍了一种名为 **LAMER** 的人工智能训练方法。为了让你听得懂，我们先打个比方。

想象你在玩一个很难的闯关游戏（比如《超级马里奥》或者《黑神话：悟空》）：

1. **普通 RL（强化学习）就像“失忆的玩家”**：你玩了一把，死掉了。系统重新开始，但你像喝了孟婆汤一样，完全忘了刚才为什么死，只能靠肌肉记忆（神经网络的参数）慢慢积累经验。每一局游戏（Episode）之间是**独立**的，互不沟通。
2. **Meta-RL（元强化学习，本文的方法）就像“写攻略的玩家”**：你玩完一把，虽然重开了，但你手里拿着上一把的“日记”（即图中的 **Reflection**）。日记里写着：“刚才是被草丛里的老六阴死的，这次要小心草丛。”

本文的核心思想就是： 利用大语言模型（LLM）强大的理解能力，让它在每次尝试后进行**自我反思（Self-reflection）**，并将这些反思作为下一次尝试的**上下文（Context）**输入。这样，AI 不用动“脑部手术”（修

改参数），光靠“看笔记”就能变强。

难点解析：不仅要努力，还要有策略

1. γ_traj（轨迹折扣因子）—— 你是急功近利还是高瞻远瞩？

文中提到的 `γ_traj` 是一个数学参数，用来调节 AI 的“耐心”。

- **如果 γ 很小**：AI 会变得很“急”。它只看重前几局能不能赢，倾向于**开发（Exploitation）**，即用已知的安全招数快速通关，不愿尝试新路。
- **如果 γ 很大**：AI 会变得很“佛系”且有远见。它更看重整个训练周期的总回报，愿意在早期几局即使输掉也要去**探索（Exploration）**地图的角落，因为它知道这些探索对长远的胜利有帮助。

2. “In-context policy adaptation”（上下文策略适应）

这听起来很高大上，其实联系到你的学习过程就很好懂。

- **传统的梯度下降（Gradient Descent）**：就像把书本知识刻进脑子里，这是长时记忆，很难改，也很慢。
- **Context（上下文）**：就像考试时的“草稿纸”或“提示卡”。本文的方法不急着改脑子，而是把上一轮的错误写在提示卡上给 AI 看。这叫“In-context”，即**在当前的语境中立刻调整策略**，效率极高。

3. 那个复杂的公式 (6) 是什么？

不要被那一堆符号吓到，我们来拆解它：

- `∇_θ J(θ)`：这就是我们要求的“梯度”，也就是“怎么改脑子能变强”的方向。
- `Σ_{n=0}^{N-1} Σ_{t=0}^{T-1}`：这是两个求和符号（Sigma）。
  - 外层的 `n` 代表**回合（Episodes）**：第1局、第2局……直到第N局。
  - 内层的 `t` 代表**每局里的步数（Time steps）**：第1步、第2步……
- **核心含义**：这个公式告诉 AI，在计算“怎么变强”时，不能只看眼前这一步（t），要看**整整一连串游戏（N个回合）的总表现**。如果你在第1局的反思写得好，帮助第3局赢了，那么第1局的反思也是有功劳的（这就是 Credit Assignment，信用分配）。

知识联想：这不就是“错题本”吗？

作为一个高三学生，你一定有**错题本**。这个 LAMER 框架简直就是“错题本理论”的数学版：

1. **Reflection = 错题分析**：你做完卷子（Episode），不能只对答案，要写下“为什么错？是公式记错了还是计算粗心？”（Textual Reflection）。
2. **Context = 考前复习**：下次考试前，你翻看错题本（Modify Context），带着这些教训去考，成绩立马提高。
3. **Optimization = 长期提分**：如果你坚持这样复盘，你的大脑（Policy θ）最终会发生质变，这就是公式 (6) 试图达到的最终效果。

总结一下：这一页讲的是如何让 AI 像聪明学生一样，通过“做题 -> 反思 -> 再做题”的循环，利用短期记忆（Context）来辅助长期进化（Parameters），从而更高效地掌握复杂任务。

第 6 页

【原文翻译】

5 实验

在本节中，我们进行全面的实验以在 Meta-RL（元强化学习）框架下评估 LAMER。具体而言，我们展示了关于以下方面的评估：(i) LAMER 在不同智能体环境中的整体性能；(ii) LAMER 对更困难任务的泛化能力；以及 (iii) LAMER 在分布偏移下的泛化能力。

5.1 实验设置

环境。我们在四个具有挑战性且多样化的环境中评估 LAMER：Sokoban（推箱子，Racanière 等人, 2017）、MineSweeper（扫雷，Li 等人, 2024）、Webshop（Yao 等人, 2022）和 ALFWorld（Shridhar 等人, 2021）。其中，Sokoban 是一款经典的基于网格的游戏，专注于规划，其环境是完全可观察的（fully observable）。相比之下，MineSweeper、ALFWorld 和 Webshop 的环境是部分可观察的（partially observable），这要求智能体在不确定性下进行探索和规划以完成任务。具体来说，MineSweeper 是一款关于隐藏方格逻辑推理的棋盘游戏。Webshop 模拟现实的网页购物任务，ALFWorld 提供基于文本的具身环境。我们在附录 A 中提供了环境的详细介绍。所有的实验均使用文本模态进行，尽管所提出的方法可以自然地应用于多模态环境。

训练细节。我们使用 Qwen3-4B（Yang 等人, 2025）作为所有实验的基础模型。为了提高智能体循环中的展开（rollout）效率，我们在轨迹生成过程中使用了非思考模式。此外，我们在 Llama3.1-8B-Instruct（Grattafiori 等人, 2024）上验证了我们的方法，结果见附录 D.1。对于 Meta-RL 设置，我们使用 $\gamma_{traj} = 0.6$ 作为默认的轨迹折扣因子，并在消融研究中探索其影响。我们将 GiGPO 作为所有使用 LAMER 的环境的默认优化算法。重要的是，为了进行 Meta-RL 训练，我们采样 N = 3 个回合（episodes），并为每个任务设置大小为 8 的组（group size）。为了确保公平比较，我们在标准 RL 训练中使用 24 的组大小，从而使每个梯度更新步骤使用的轨迹总数相同。为了公平比较，所有其他超参数和配置在 RL 和 Meta-RL 之间保持一致，并提供在附录 C 中。我们的代码在 https://github.com/mlbio-epfl/LaMer 公开可用。

5.2 性能比较

我们将 LAMER 框架的性能与基于提示（Prompting）的基线（Zero-shot, ReAct (Yao 等人, 2023b; Shinn 等人, 2023)）以及 RL 方法（PPO (Schulman 等人, 2017), RLOO (Ahmadian 等人, 2024), GRPO (Shao 等人, 2024), 和 GiGPO (Feng 等人, 2025)）在三个环境中进行了比较：Sokoban、MineSweeper 和 Webshop。对于每种方法，我们报告了在 1、2 和 3 次尝试下的成功率（即 pass@1, pass@2, 和 pass@3）。结果汇总于表 1。

表 1：在 Sokoban、MineSweeper 和 Webshop 环境上的性能。p@1, p@2 和 p@3 分别表示在 1、2 和 3 次尝试下的成功率 (%)。

方法	Sokoban			MineSweeper			Webshop		
	p@1	p@2	p@3	p@1	p@2	p@3	p@1	p@2	p@
Prompting									
Zero-shot	6.8	9.8	12.9	4.5	6.6	8.6	1.4	2.1	2.
ReAct	7.2	9.6	12.5	6.3	7.0	10.9	3.1	4.5	4.
Reflexion	6.4	9.8	12.1	5.5	7.2	9.8	2.7	3.3	3.
Training with RL									
PPO	12.5	15.4	16.8	29.7	34.2	35.5	53.1	54.5	54
RLOO	13.5	16.6	18.8	48.8	51.2	51.6	67.6	68.4	69
GRPO	22.9	26.4	27.0	36.3	40.0	40.4	72.9	73.0	73
GiGPO	41.6	43.6	44.1	**52.0**	54.9	55.1	**73.4**	74.6	75
Training with Meta-RL (ours)									
LAMER	**42.4**	**52.0**	**55.9**	44.1	**66.4**	**74.4**	67.8	**84.4**	**89**

【高三解读】

核心概念：AI 的“全能大考”与“应试策略”

这篇文章实际上是在记录一场针对人工智能的“智力大比拼”。科学家们想知道，他们研发的新学习方法——**LAMER**（基于 Meta-RL 框架），是不是比现有的 AI 学习方法更聪明、更善于举一反三。

为了验证这一点，他们设计了三个主要考场，就像我们高中的不同科目考试：

1. **Sokoban（推箱子）**：这考的是**逻辑规划**。就像数学里的立体几何，所有的信息（箱子位置、目标点）都在明面上，你需要推算出最优路径。这被称为“完全可观察环境”。
2. **MineSweeper（扫雷）**：这考的是**推理与风险管理**。这就像做一道很难的概率题，有些信息是隐藏的（地雷在哪里不知道），你必须根据已知数字去推理，甚至要在不确定时做出最优赌注。这被称为“部分可观察环境”，难度更高。
3. **Webshop（网购模拟）**：这考的是**现实应用能力**。AI 需要在一个模拟的购物网站上根据指令买东西，考验它处理繁杂信息和完成多步骤任务的能力。

难点解析：什么是“元强化学习”？为什么 LAMER 这么强？

文中提到了两个核心概念的对比：**标准 RL（强化学习）** 与 **Meta-RL（元强化学习）**。

- **标准 RL（死记硬背 + 题海战术）**：
  你看表中的 PPO、GRPO 等方法，它们就像是一个勤奋但略显死板的学生。它们通过成千上万次的练

习（“梯度更新”），试图记住每一道题的答案或固定的解题套路。如果题目稍微变一下，它们可能就傻眼了。

- **Meta-RL / LAMER（掌握学习方法 + 举一反三）**：
  LAMER 就像是班里的学霸。它不只是在学“这道题的答案是什么”，而是在学“**面对一道新题，我该如何快速找到解法**”。注意文中的参数 `N = 3 episodes`，这意味在测试时，LAMER 被允许尝试 3 次。
  - **观察数据（表 1）**：看 LAMER 的数据行。在 `p@1`（第一次尝试）时，它的优势可能还没那么夸张（比如在 MineSweeper 上甚至不如 GiGPO）。但是看 `p@3`（第三次尝试），LAMER 的分数飙升（从 44.1% 涨到 74.4%！）。
  - **解读**：这说明 LAMER 具有极强的**反思和修正能力**。第一次错了没关系，它能迅速分析失败原因，调整策略，在第二次、第三次尝试中解决问题。而标准 RL 方法（如 GiGPO）从 p@1 到 p@3 的提升非常微弱（52.0% -> 55.1%），说明它们即使多给机会，也难以从错误中吸取教训。

知识联想：建立你的知识体系

1. **高中数学（概率与统计）**：
   - 表中的 `pass@k` 指标其实就是概率论中的“至少一次成功率”。假设单次成功率为 p，如果你有 k 次独立尝试的机会，成功的概率通常会增加。但在 AI 中，这几次尝试往往不是独立的，而是后一次基于前一次的经验。LAMER 的曲线斜率大，说明它的条件概率 $P(\text{Success}_2 \mid \text{Fail}_1)$ 很高。
2. **高中生物（神经调节与学习）**：
   - **RL（强化学习）** 类似于生物学中的“条件反射”或“操作性条件反射”（斯金纳箱），通过奖励（得分）来强化行为。
   - **Meta-RL** 则更接近人类的高级认知功能——**元认知（Metacognition）**。我们在做理综卷时，不仅是在做题，还在监控自己的做题状态：“这道物理题太难了，我先跳过，等会儿再回来”。这种“对思考过程的思考”，正是 LAMER 试图模仿的能力。
3. **物理（可观察性）**：
   - 文中区分了“完全可观察”和“部分可观察”。这就像经典力学（所有变量如位置、动量理论上可测）与量子力学（存在测不准原理，无法同时获知所有状态）的区别。在现实世界（部分可观察）中，比如驾驶、炒股、做科研，我们永远无法掌握全知视角，因此像扫雷这样“在不确定性下决策”的能力，才是 AI 通向通用智能的关键。

总结：这一页的核心证据是表格。它告诉我们，在这个实验中，让 AI 学会“如何学习”（Meta-RL），比单纯让它“大量刷题”（RL）在解决复杂、多变的问题时更有效，尤其是在它能够进行多次尝试的情况下。

第 7 页

【原文翻译】

Meta-RL 获得了更好的性能

在所有三个环境中，使用 **Meta-RL**（元强化学习）训练的 **LAMER** 在最终的 pass@3 成功率上始终优于基于提示（prompting-based）的基线模型和 RL 训练方法。在 **Sokoban**（推箱子）环境中，LAMER 达到了 55.9% 的 pass@3 成功率，比最佳的 RL 训练模型高出 19%。在 **Webshop** 环境中，LAMER 的表现也比 RL 训练的模型高出 14%。值得注意的是，性能的提升不仅限于 pass@3：在所有环境中 pass@2 的性能也

有所提升，甚至在 Sokoban 中 pass@1 也有提升。总之，这些结果表明 LAMER 为训练后的智能体在复杂环境中解决长视界（long-horizon）任务提供了持续的增益。

Meta-RL 展现出更强的测试时扩展性（Test-time Scaling）

使用 Meta-RL 训练的 LAMER 在测试时扩展性方面表现出显著的有效性，根据表 1 的结果，随着尝试次数的增加，其性能增益更大。例如，在从 pass@1 到 pass@3 的过程中，Sokoban 上的 LAMER 实现了 13.5% 的提升，明显大于 RL 训练和基于提示的基线（提升均小于 5%）。值得注意的是，尽管 LAMER 在 **MineSweeper**（扫雷）和 **Webshop** 环境中的起始 pass@1 性能略低于 RL 基线（GiGPO），但它迅速恢复并在 pass@2 和 pass@3 上超越了所有基线。结果表明，Meta-RL 训练的模型已成功学会了在早期剧集中主动探索，并从错误中有效地适应，从而在随后的尝试中获得显著收益。训练后的智能体产生的说明性轨迹和反思展示在附录 E 中。

Meta-RL 诱导探索

为了进一步分析模型的行为，我们测量了跨环境的答案轨迹的多样性。对于每个问题，我们从智能体采样多个轨迹，并将具有相同状态和动作的相同轨迹分组。这些组用于形成不同轨迹的经验分布，如图 1 所示。然后我们估计分布的**熵（entropy）来量化轨迹多样性**。图 3 比较了基础模型、RL 和 Meta-RL 智能体在不同环境下的轨迹多样性。我们观察到基础模型表现出最高的熵，表明它生成了范围广泛的轨迹，尽管这种多样性并没有转化为更高的成功率（见表 1）。RL 训练的智能体减少了多样性，并收敛向更确定性的行为。相比之下，使用 Meta-RL 训练的 LAMER 保持了比 RL 基线更高水平的多样性，允许在测试时进行更多的探索。

图 3：基础模型和训练后模型的轨迹多样性。 与 RL 相比，Meta-RL 保留了来自基础模型的更多样化的轨迹，在探索（exploration）和利用（exploitation）之间取得了更好的平衡。

指标/环境	Sokoban (推箱子)	MineSweeper (扫雷)	Webshop (网购)
Y轴: 轨迹多样性 (熵)	范围: 0 - 6	范围: 0 - 4	范围: 0 - 6
Base Model (基础模型)	高中位数 (~4.5)，高方差	高中位数 (~3.5)，中等方差	极高中位数 (~4.5)，大方差
RL (强化学习)	低中位数 (~1.5)，低方差	极低中位数 (~0.5)，低方差	低中位数 (~1.5)，中等方差
Meta-RL (元强化学习)	中等中位数 (~2.5)，中等方差	中等中位数 (~1.5)，中等方差	中等中位数 (~3.0)，大方差

5.3 泛化到更难的任务

接下来，我们研究预训练模型在更难任务上的泛化能力。为此，我们选取了用 RL 和 Meta-RL 训练的模型，并在 Sokoban 和 MineSweeper 环境中更难的任务上对其进行评估。我们通过在网格中使用更多的箱子（Sokoban）和更多的地雷（MineSweeper）来增加难度。结果显示在图 4 中。不出所料，使用 RL 和 Meta-RL 训练的模型在更难的任务上表现都不如预期，随着网格中箱子或地雷数量的增加，性能有所下降。然而，Meta-RL 在所有难度级别上始终优于 RL。值得注意的是，在最困难的设置下，从 Meta-RL 训练的模型仍然优于 RL 训练的模型，在 Sokoban 上有 10% 的性能差距，在 MineSweeper 上有 5% 的性能[差距]...（此处文本截断）

【高三解读】

【高三解读】从“死记硬背”到“举一反三”：AI 进化的成绩单

你好！我是你的学术导读员。如果说上一页我们在讲这篇论文的“核心思想”，那么这一页就是它的“期末成绩单”和“心理分析报告”。这页内容非常硬核，展示了作者提出的 **Meta-RL（元强化学习）** 方法究竟好在哪里。

我们可以把这一页的内容想象成在对比三个学生：

1. **基础模型（Base Model）**：一个完全没复习的学生，做题全靠蒙，答案五花八门。
2. **RL（强化学习）模型**：一个搞“题海战术”的学生，死记硬背标准答案，遇到做过的题又快又准，但稍微变通一下就懵了。
3. **Meta-RL（元强化学习）模型**：这就是我们要夸的“学霸”，他不仅掌握了知识，还学会了“考试技巧”——如果第一遍做错了，他能迅速反思并修正。

下面我们从三个维度来拆解这份“成绩单”。

1. 核心概念：为什么“测试时扩展性”这么重要？

文中提到了一个很酷的概念叫 **“Test-time scaling”（测试时扩展性）**。这是什么意思呢？

- **普通情况**：通常我们认为模型训练好了，能力就固定了。就像你走进考场那一刻，水平就定格了。
- **Meta-RL 的魔法**：但在推箱子（Sokoban）等游戏中，作者发现，给 Meta-RL 模型多几次尝试机会（从 pass@1 到 pass@3），它的成功率会暴涨 13.5%！
- **解读**：这说明 Meta-RL 具有**“在战斗中学习战斗”**的能力。它第一次可能失败（探索），但它能记住失败的教训，在第二次、第三次尝试时迅速调整策略。而那个死记硬背的 RL 学生，给他多少次机会，他还是会犯同样的错，因为他只会一种解法。

2. 难点解析：图3中的“熵”与“探索-利用”的博弈

图3 是这一页的精华，也是最难懂的地方。它用**“熵（Entropy）”来衡量“轨迹多样性”**。

- **知识联想（物理/化学）**：在高二物理热学或化学反应原理中，我们学过“熵”是衡量系统**混乱度**或**无序度**的指标。熵越大，越混乱。
- **AI 中的熵**：在这里，熵代表**思维的活跃度**。
  - **基础模型（高熵）**：像没头苍蝇一样乱撞。想法多，但都是错的。虽然“多样性”高，但解决不了问题。
  - **RL 模型（低熵）**：这就叫**“模式坍缩”**（Mode Collapse）。它变得太保守了，只敢走它背下来的那条路。一旦那条路堵死了，它就没招了。这就像只会套公式，不会变通。
  - **Meta-RL（中等熵）**：**这就是完美的平衡点！** 图3 显示它的箱线图位于中间。它比死记硬背的 RL 灵活，保留了一定的“混乱度”去尝试新路径（探索），但又比瞎猜的基础模型有章法。**这种“有纪律的自由”，正是解决复杂问题的关键。**

3. 知识进阶：泛化能力与“高分低能”的终结

文章最后一段提到了 **“Generalization to Harder Tasks”（泛化到更难的任务）**。这也是高考区分度的体现。

- 当推箱子的箱子变多、扫雷的地雷变多时，这就相当于从“课本例题”变成了“压轴大题”。
- 结果显示，所有模型的表现都下降了（这很正常），但 Meta-RL 依然比 RL 强（在 Sokoban 上强 10%）。
- **这告诉我们**：真正的智能不是“记住”怎么推 3 个箱子，而是“理解”推箱子的逻辑。当你掌握了逻辑（元知识），哪怕面对 10 个箱子，你也能推演出来。

总结与启示

这一页实际上是在告诉我们要**拒绝“过拟合”（Overfitting）**的学习方式。

- **RL（强化学习）** 就像是疯狂刷题，把题目背下来，这就是“过拟合”，虽然平时作业分高，但缺乏**探索性**。
- **Meta-RL** 则是在培养**“元认知”**能力——即“关于思考的思考”。它允许自己在考试中犯错、尝试、调整。这不仅是 AI 的进化方向，也是我们高中生在复习时应该追求的境界：**不要只记答案，要学会分析为什么错，以及如何换个角度解决问题。**

第 8 页

【原文翻译】

图表内容描述

图表 4：推箱子（左）与扫雷（右）成功率对比

- **左图（推箱子 Sokoban）：** 横轴为“箱子数量”（2, 3, 4, 5），纵轴为“成功率 (%)”（10-50）。图中展示了两条折线：黄色代表“RL”（强化学习），绿色代表“Meta-RL”（元强化学习）。随着箱子数量增加，两者的成功率都下降，但 Meta-RL 始终高于 RL，且在箱子数量为 3 和 4 时差距明显。
- **右图（扫雷 MineSweeper）：** 横轴为“地雷数量”（3, 4, 5, 6），纵轴为“成功率 (%)”（0-60）。同样随着地雷增加成功率下降，Meta-RL（绿色）的表现显著优于 RL（黄色），且保持了较大的差距。

图 4： RL 和 Meta-RL 训练的模型在难度增加的任务上的表现。对于推箱子，我们逐渐增加箱子的数量；对于扫雷，我们增加网格中地雷的数量。

...扫雷游戏上的差距。这一持续的差距表明，使用 Meta-RL 训练的 LaMER 不仅在训练分布上表现更好，而且能更好地泛化到更难的任务中。

5.4 泛化到未见过的任务

我们进一步研究了 LaMER 和替代方法泛化到分布外（out-of-distribution）任务的能力。在这个实验中，我们使用了 ALFWorld 环境 (Shridhar et al., 2021)。作为一个基于文本的具身环境，ALFWorld 包含 6 类常见的家庭活动：拾取并放置 (*Pick*)、在灯光下检查 (*Look*)、清洁并放置 (*Clean*)、加热并放置 (*Heat*)、冷却并

放置 (*Cool*) 以及 拾取两个并放置 (*Pick2*)。我们将 *Pick*、*Look*、*Clean* 和 *Heat* 作为分布内任务，并将 *Cool* 和 *Pick2* 作为分布外任务。我们在分布内任务上训练 LaMER 和替代基线模型，然后评估模型在分布内任务（使用保留的测试集）和分布外任务示例上的表现。结果如表 2 所示。正如我们可以看到的，RL 训练的模型通常在分布内任务上表现良好，并且通过在 *Look*、*Clean* 和 *Heat* 上实现超过 20% 的改进，优于基于提示（prompting-based）的方法。然而，在分布外任务 *Cool* 和 *Pick2* 上，它仅获得了 58.1% 和 36.0% 的成功率。相比之下，采用 Meta-RL 的 LaMER 在分布内和分布外任务上均表现一致，在分布外任务上有着显著的性能差距。特别是在 *Cool* 任务上，我们的 LaMER 框架实现了 23% 的性能提升，在 *Pick2* 上提升了约 14%。总体而言，这些结果表明，在 ALFWorld 上，Meta-RL 训练的模型比 RL 训练的模型能更好地泛化到分布外任务。

表 2：ALFWorld 任务上的分布外泛化评估。

方法 (Method)	i.d (分布内)				o.o.d (分布外)	
	Pick	**Look**	**Clean**	**Heat**	**Cool**	**Pick2**
Prompting	91.9	52.9	48.4	44.8	42.8	21.2
RL	95.5	83.0	67.9	86.6	58.1	36.0
Meta-RL	**97.7**	**100.0**	**90.2**	**89.5**	**81.0**	**50.2**

6 分析

我们进一步对 LaMER 的关键设计因素进行了一系列消融研究（ablation studies），包括 (i) 轨迹折扣因子 γ_{TRAJ} 对探索与利用之间权衡的影响，以及 (ii) 跨回合记忆配置的消融。我们还讨论了 (iii) 提出的 Meta-RL 框架与 RL 训练相比的计算预算。

6.1 轨迹折扣因子的影响

跨回合折扣因子 γ_{TRAJ} 控制奖励在试验内的传播方式，从而调节训练期间 LaMER 框架中探索（exploration）与利用（exploitation）之间的平衡。为了理解折扣因子的影响，我们在 Sokoban 和 MineSweeper 环境中使用不同的 γ_{TRAJ} 值训练带有 LaMER 的智能体（图 5）。我们观察到……

【高三解读】

同学们好！今天我们继续深入探讨前沿人工智能（AI）如何像人类一样“举一反三”。这张页面展示了这篇论文非常核心的实验结果和分析部分。如果说前面的部分是在讲“我们造了一辆新车”，那这一页就是在讲“这辆车在越野路况下跑得有多好”以及“为什么它跑得这么快”。

核心概念：从“死记硬背”到“融会贯通”

这一页主要想告诉我们一件事：**传统的强化学习（RL）像是一个只会刷题的学生，而本文提出的 Meta-RL（元强化学习）则是一个掌握了学习方法的学霸。**

1. **直面困难（图4）：** 作者让 AI 玩“推箱子”和“扫雷”。这就好比考试题不仅有基础题，还有压轴题。图表显示，随着游戏难度增加（箱子或地雷变多），普通 RL（黄线）的成绩掉得很快，而 Meta-RL（绿线）虽然也掉，但始终保持领先。这说明新方法抗压能力强，面对复杂局面更稳健。
2. **举一反三（5.4节 & 表2）：** 这是本页的重头戏——**泛化能力**。作者在 ALFWorld 这个模拟家庭环境中做了实验。
   - **分布内（i.d.）**：好比老师上课讲过的例题（如“捡东西”、“加热东西”）。
   - **分布外（o.o.d.）**：好比考试时遇到的新题型，逻辑相似但没见过（如“冷却东西”、“一次捡俩”）。
   - **结果**：普通 RL 遇到“新题型”就歇菜了（看表2中 RL 在 Cool 和 Pick2 上的分数），而 Meta-RL 依然能拿高分。这就是我们常说的“迁移学习”能力，它不仅仅学会了具体的动作，更是学会了底层的逻辑。

难点解析

这里有几个学术术语，其实对应着非常有意思的思维模型：

1. **分布内 (i.d.) vs. 分布外 (o.o.d.)**：
   - *全称*：In-distribution 和 Out-of-distribution。
   - *解读*：这是统计学和机器学习的核心。假设你只见过白天鹅（这是你的训练分布），突然来了一只黑天鹅（这是分布外），你会不会把它认成鸭子？AI 的强大与否，关键看它能不能处理“分布外”的数据，即它未曾直接见过的场景。
2. **消融研究 (Ablation Studies)**：
   - *解读*：这个词听起来很医学，其实就是大家高中生物或物理实验中常用的**“控制变量法”**或**“减法实验”**。为了证明我的系统有效，我尝试把其中某个组件（比如“记忆功能”）去掉，看看系统性能会不会下降。如果下降了，说明这个组件是核心部件；如果没变化，说明它是多余的。
3. **轨迹折扣因子 (Trajectory Discount Factor, γ_{TRAJ})**：
   - *解读*：你可以把它理解为**“耐心值”**。在强化学习中，γ (Gamma) 决定了 AI 是看重“眼前的糖果”（即时奖励）还是“未来的大餐”（长期回报）。
   - 如果 γ 很小，AI 就会短视，只顾眼前利益（利用）；
   - 如果 γ 很大，AI 就会愿意牺牲当下的舒适去探索未知的可能性，为了长远的胜利（探索）。
   - 这也像我们的人生规划：是选择今晚玩游戏（低 γ），还是选择复习考好大学（高 γ）？作者在这里就是在调整这个参数，寻找最佳的平衡点。

知识联想

- **物理/生物实验设计**：你在做实验时，是否设置了对照组？表2 中不仅有 RL 和 Meta-RL，还对比了 Prompting（提示法），这就是严谨的**多组对照实验**。以后大家写实验报告，也要学会用数据表格直观地展示差距。
- **元认知 (Metacognition)**：Meta-RL 中的 "Meta"（元）与心理学中的“元认知”异曲同工。普通学习是“学习知识”，元学习是“关于学习的学习”，即反思自己的学习策略是否有效。作为高三学生，在刷题之余，停下来思考“我为什么错”、“这题考什么逻辑”，就是在进行“元学习”，这往往比盲目刷题更有效。

总结来说，这一页通过严详实的数据证明了新模型不仅“内功深厚”（基础任务做得好），而且“变通能力强”（新任务也能搞定），接下来的第6章则是要像解剖麻雀一样，拆解它为什么能成功。

第 9 页

【原文翻译】

较大的 γ_traj 值并不一定能在 pass@3 上带来更好的最终性能，相反，γ_traj 的最佳设置因环境而异。对于推箱子（Sokoban）和购物网站（Webshop）环境，中间值（例如，γ_traj = 0.6）能产生最佳结果，这表明平衡即时奖励和长期奖励对这些任务更为重要。相比之下，扫雷（MineSweeper）受益于较大的 γ_traj（例如，γ_traj = 0.9），表明扩展的信用分配能更好地支持在该环境中的策略性探索。总体而言，结果表明 γ_traj 提供了一种实用的方法来控制跨环境的“探索”与“利用”之间的权衡。

Sokoban **MineSweeper** **Webshop**
(此处为三个折线图，显示成功率随 γ_traj 变化的趋势。Sokoban 和 Webshop 在 0.6 处达到峰值，MineSweeper 在 0.9 处表现较好)

图 5：使用不同 γ_traj 训练的模型的成功率。较高的值鼓励在训练期间进行更多的探索。

6.2 关于跨回合记忆的消融实验

在 LaMER 中，代理策略通过跨回合记忆 $\mathscr{H}^{(n)}$ 进行“情境中”(in-context) 调整，该记忆默认包含先前回合的轨迹和反思。为了评估记忆内容对训练的影响，我们考虑了 $\mathscr{H}^{(n)}$ 的两种替代配置：(i) 仅历史轨迹；(ii) 仅反思。训练后的代理在各配置下的性能报告于表 3 中。结果显示，自我反思在 LaMER 中提供了明显的收益，分别在 Sokoban 上带来了 21.6% 的提升，在 MineSweeper 上带来了 11.0% 的提升，在 Webshop 上带来了 3.5% 的提升。有趣的是，“仅反思”的配置在所有环境中甚至优于 LaMER 的默认设置（即包含轨迹和反思两者）。我们假设这是因为仅包含反思的记忆提供了更简洁和专注的指导，从而导致代理行为的适应更加有效。

表 3：LaMER 与不同跨回合记忆配置的比较。

$\mathscr{H}^{(n)}$ 中的内容	Sokoban	MineSweeper	Webshop
仅轨迹 (Trajectory-only)	34.8	69.5	89.3
仅反思 (Reflection-only)	**56.4**	**80.5**	**92.8**
两者皆有 (Both)	55.9	74.4	89.1

6.3 训练预算

接下来，我们分析强化学习（RL）和元强化学习（Meta-RL）的训练预算，重点关注数据使用量和计算效率。为了确保公平比较，我们将标准 RL 的组大小（group size）设置为 Meta-RL 的三倍。这一调整保证了两种方法在每次梯度更新时消耗相同数量的轨迹。除了这种比例缩放外，所有其他实验配置——如学习率、批次大小和网络架构——都保持不变。这种设计选择突显了 Meta-RL 与 RL 相比并不需要更大的数据预算；换句话说，这两种方法都依赖于相同总量的轨迹来进行学习。

尽管如此，与 RL 基线相比，LaMER 可能会引入额外的训练时间成本。在 RL 训练中，所有回合（episode）都可以并行采样，因为它们是相互独立的。相比之下，LaMER 表现出的并行性较低，因为同一试验（trial）中的回合需要按顺序生成。结果是，我们观察到在当前的实现中，LaMER 的训练时间成本大

约是 RL 的两倍。这表明，更复杂的采样策略，如异步展开（asynchronous rollout），可以进一步提高 LaMER 训练大型语言模型（LLM）代理的效率。

9

【高三解读】

核心概念：AI 的“性格养成”与“高效学习法”

这一页的内容非常有意思，它实际上是在探讨如何训练一个更聪明、更高效的 AI 代理（Agent）。你可以把这看作是在研究“如何培养一个学霸”。文章主要讨论了三个关键点：

1. **调节 AI 的“好奇心” (γ_traj)**：
   文章首先提到了一个参数 γ_{traj}。你可以把它想象成 AI 的“目光长远度”或者“好奇心指数”。
   - **较低的值**意味着 AI 比较“短视”，只看重眼前的利益（exploitation，利用）。
   - **较高的值**意味着 AI 愿意为了长远的巨大回报而牺牲眼前的蝇头小利，或者去尝试未知的领域（exploration，探索）。
   - **结论**：图 5 告诉我们，并不是好奇心越强越好。像推箱子（Sokoban）这种需要精细操作的游戏，过度的探索反而会坏事，需要一个平衡点（0.6）；而像扫雷（MineSweeper）这种充满了不确定性的游戏，AI 需要更大胆地去探索（0.9）才能学会策略。这就像你们填报志愿，有的学科需要稳扎稳打，有的学科需要发散思维。
2. **“死记硬背” vs. “总结反思” (消融实验)**：
   在 6.2 节，作者做了一个“消融实验”（Ablation Study）。这是科研中常用的方法，就像控制变量法，把系统拆掉一部分，看看剩下的部分表现如何。
   - 他们对比了 AI 在学习时依靠什么记忆：是依靠单纯的“历史轨迹”（就像把做过的错题原封不动背下来），还是依靠“反思”（即总结自己为什么错了，下一步该怎么做）。
   - **惊人的发现**：表 3 显示，“仅反思”的效果竟然最好！比“既背题又反思”还好。这说明，**精炼的、高质量的思考比海量的信息堆砌更有价值**。对于你们高三学生来说，这简直是金玉良言：不要盲目刷题，要多做错题分析和总结！
3. **计算“补习班”的成本 (训练预算)**：
   6.3 节讨论了这种名为 LaMER 的新方法（Meta-RL）是不是更“烧钱”。
   - **数据成本**：作者设计得很公平，证明了新方法并不需要更多的练习题（数据量），它和传统方法一样高效。
   - **时间成本**：但是，新方法在时间上慢了两倍。为什么？因为传统方法可以同时做十张卷子（并行计算），而 LaMER 需要根据上一题的结果来做下一题（串行/顺序生成），这限制了速度。不过作者也说，未来可以通过技术手段（异步展开）来解决这个问题。

难点解析：什么是“消融实验”与“并行性”？

1. **消融实验 (Ablation Study)**：
   这个词听起来很医学，其实逻辑很简单。假设你有一辆跑得很快的赛车，你想知道它为什么快。你把尾翼拆了，发现速度没变；你把涡轮拆了，速度慢了一半。结论就是：涡轮是关键。文中作者通过移除“轨迹记忆”或“反思记忆”，发现“反思记忆”才是提分的关键组件。这教导我们在分析问题时，要学会拆解要素，找到核心矛盾。

2. **串行 vs. 并行 (Sequential vs. Parallel)**：
   文中提到 LaMER 训练慢是因为“lack of parallelism”（缺乏并行性）。
   - **传统 RL**：就像 50 个学生同时做同一套卷子，大家各做各的，互不干扰，老师最后收卷子一起改，效率极高（并行）。
   - **LaMER (Meta-RL)**：就像一个学生在做闯关游戏，必须先通了第一关，拿到了经验和反思，才能带着这些记忆去打第二关。第二关的策略依赖于第一关的结果，所以没法同时进行，只能按顺序来（串行），自然就慢了。

知识联想：跨学科的智慧

- **心理学（元认知）**：文中的“反思”（Reflection）对应心理学中的“元认知”（Metacognition），即“对思考的思考”。研究表明，元认知能力强的学生，学习效率远高于只知道死记硬背的学生。AI 的进化方向也在模仿人类的高级认知功能。
- **物理学（电路）**：训练时间的讨论可以联想到物理中的**串联与并联电路**。并联电路中，各支路互不影响，电流可以同时通过；串联电路中，电流必须依次通过各个元件。并行计算就像并联电路，阻力小、流量大；串行计算就像串联电路，一个卡住，后面都得等。
- **数学（极值问题）**：图 5 中寻找最佳 γ_{traj} 的过程，本质上就是在寻找函数的**极大值点**。在 0 到 1 的定义域内，通过实验描点，找到导数为零（或峰值）的位置。这就是导数在实际工程优化中的应用。

第 10 页

【原文翻译】

7 结论

能够探索并从环境中收集信息，对于构建能够快速且稳健适应的自主智能体（Agent）至关重要。我们介绍了 **LaMER**，这是一个利用**元强化学习（meta reinforcement learning）原理的通用大语言模型（LLM）智能体训练框架。与以往那些通过最大化单集（single-episode）回报以获得即时收益的强化学习（RL）方法不同，LaMER 最大化的是折现后的跨集（cross-episode）回报**，从而自然地平衡何时探索（explore）与何时利用（exploit），以最大化长期性能。这种在训练时允许的探索行为，教会了智能体通用的探索策略，使其能够在测试时实现快速的**情境中适应（in-context adaptation）**。我们在多种不同的环境中展示了 LaMER 显著优于 RL 方法，能够泛化到更难的环境，并且在测试时随着更多剧集（episodes）的增加展现出更好的扩展性。

局限性与未来工作。 我们的结果提出了几个未来工作的有希望的方向。(i) 我们方法的通用性允许将其与其他 RL 算法或自我反思（self-reflection）框架相结合。我们假设更先进的优势估计策略（advantage estimation strategy）或更强的推理模型可能会提升性能。(ii) 我们的方法需要在跨集训练中顺序采样剧集，因为剧集之间是相互依赖的。这最终导致训练时间比 RL 方法更长。未来的工作将探索更高效的训练策略。(iii) 最后，在较简单的环境上训练的 LaMER 可以泛化到同类更难的环境或相对相似的领域。这最终表明了构建能够适应完全新颖环境的通用智能体（generalist agents）的可能性。

致谢

M.B. 衷心感谢瑞士国家科学基金会 (SNSF) 启动资助 TMSG12_226252/1、SNSF 资助 IC00I0_231922 以及瑞士人工智能计划的支持。M.B. 是多尺度人类计划（Multiscale Human Program）中的 CIFAR 研究员。

参考文献

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

AutoGPT. Significant-Gravitas/AutoGPT: An experimental open-source attempt to make gpt-4 fully autonomous., 2023. URL `https://github.com/Significant-Gravitas/Auto-GPT/tree/master`.

Jakob Bauer, Kate Baumli, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, et al. Human-timescale adaptation in an open-ended task space. *In International Conference on Machine Learning*, 2023.

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A tutorial on meta-reinforcement learning. *In Foundations and Trends in Machine Learning*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *In Advances in Neural Information Processing Systems*, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec

【高三解读】

高三解读：从“死记硬背”到“学会学习”——AI进化的新篇章

同学们，这篇论文的结论部分实际上是在讨论一个困扰人工智能界很久的核心问题：**如何让AI不只是“通过考试”，而是真正“学会学习”？** 我们常说“授人以鱼不如授人以渔”，这篇论文提出的 **LaMER** 框架，就是试图教给 AI 那个“渔”的方法。

1. 核心概念：不仅要“赢在当下”，更要“布局未来”

首先，我们要区分两个概念：**强化学习（RL）** 和 **元强化学习（Meta-RL）**。

- **传统的 RL** 就像是一个只看重“期末考试分数”的学生。它玩游戏或做任务时，每一步都只想着怎么在这个回合（Episode）里拿最高分。这种方法的缺点是目光短浅，而且换个新题型可能就懵了。
- **论文提出的 LaMER** 则像是一个注重“学习方法”的学霸。它不只盯着眼前这一局游戏的输赢，而是关注**跨集（Cross-episode）的回报。什么意思呢？它为了搞清楚这个游戏的规律，可能会在第一局故意**

> **“试错”（Explore），哪怕这一局输了也没关系，因为它通过试错收集到了关键信息，从而保证在第二局、第三局以及未来的所有局里都能拿高分。这叫最大化长期性能**。

文中提到的“**探索（Explore）** vs **利用（Exploit）**”是AI（也是人生）中的经典博弈。就像你在食堂打饭，是去尝试那个没吃过但可能很难吃的“黑暗料理”（探索），还是老老实实打你最爱的番茄炒蛋（利用）？LaMER 的高明之处在于，它学会了什么时候该去“试菜”，什么时候该“吃老本”，从而让你整个高三生涯的饮食体验（长期回报）最好。

2. 难点解析：为什么“连贯剧情”比“单元剧”更难拍？

在“局限性”部分，作者提到了一个技术难点：**(ii) 训练时间长**。

为了让大家理解这一点，我们可以联想一下电视剧。传统的 RL 训练就像情景喜剧（如《家有儿女》），每一集都是独立的，你可以同时找100个编剧并行写100集，效率很高。但 LaMER 的训练就像是一部逻辑严密的悬疑连续剧，第二集的剧情完全依赖于主角在第一集里学到了什么线索。既然剧集之间存在**相互依赖（Dependent）**，你就不能并行处理，必须先拍完第一集，有了结果才能拍第二集，这就是文中说的“顺序采样（Sequential Sampling）”。这虽然导致训练变慢，但只有这样，AI才能学会处理连续变化的复杂世界。

另一个难点是 **“泛化（Generalize）”**。作者发现，在简单关卡训练出来的 LaMER，居然能通关更难的关卡。这就好比你只在课本上学了基础公式（简单环境），但在高考压轴题（困难环境）中也能灵活运用。这是通往**通用人工智能（AGI）**的重要一步。

3. 知识联想：跨学科的智慧共鸣

- **数学（级数与极限）**：文中提到的“折现后的回报”其实就是数学中的**几何级数求和**模型（$\sum \gamma^t r_t$）。未来的收益 r_t 会乘以一个小于1的系数 γ（折现因子），意味着越久远的未来，不确定性越大，权重越低。这和银行计算复利、物理学中的衰变都有异曲同工之妙。
- **生物与进化**：人类本身就是最好的“元强化学习”机器。婴儿时期的抓握、乱咬（探索），看似没有直接收益，甚至有危险，但这些行为构建了我们对物理世界的认知模型，让我们成年后能处理各种从未见过的问题。LaMER 正是在模仿这种生物进化的智慧。
- **心理学（延迟满足）**：著名的“棉花糖实验”告诉我们要为了更大的奖励忍受当下的诱惑。LaMER 懂得牺牲第一局的分数来换取整个任务的成功，这在算法层面上实现了“延迟满足”，是智能等级提升的体现。

总结：这一页虽然只是结论，但它描绘了AI未来的图景——不再是机械的执行者，而是懂得反思、懂得为了长远利益去探索未知环境的“思考者”。作为高三学生，你们现在的每一次刷题、每一次改错，其实也是在进行一次次“跨集训练”，为了就是在那场终极挑战中实现完美的“情境适应”。加油！

第 11 页

【原文翻译】

Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, 和 Wojciech Zaremba。评估基于代码训练的大型语言模型。*arXiv 预印本 arXiv:2107.03374*, 2021。

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, 等。训练验证器以解决数学应用题。*arXiv 预印本 arXiv:2110.14168*, 2021。

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, 等。Deepseek-R1：通过强化学习激励大语言模型的推理能力，2025。

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, 和 Pieter Abbeel。RL²：通过慢速强化学习实现快速强化学习。*arXiv 预印本 arXiv:1611.02779*, 2016。

Lang Feng, Zhenghai Xue, Tingcong Liu, 和 Bo An。用于大语言模型智能体训练的组内组（Group-in-group）策略优化。*发表于《神经信息处理系统进展》（NeurIPS）*, 2025。

Chelsea Finn, Pieter Abbeel, 和 Sergey Levine。用于深度网络快速适应的模型无关元学习。*发表于《国际机器学习会议》（ICML）*, 2017。

Kanishk Gandhi, Denise H J Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, 和 Noah Goodman。搜索流（SoS）：在语言中学习搜索。*发表于《语言建模会议》*, 2024。

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 等。Llama 3 模型群。*arXiv 预印本 arXiv:2407.21783*, 2024。

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, 和 Jacob Steinhardt。使用 APPS 衡量编程挑战能力。*发表于《神经信息处理系统数据集和基准测试赛道》（NeurIPS Datasets and Benchmarks Track）*, 2021a。

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, 和 Jacob Steinhardt。使用 MATH 数据集衡量数学问题求解能力。*发表于《神经信息处理系统数据集和基准测试赛道》（NeurIPS Datasets and Benchmarks Track）*, 2021b。

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, 和 Amos Storkey。神经网络中的元学习：综述。*IEEE 模式分析与机器智能汇刊*, 2021。

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, 和 Aleksandrs Slivkins。大型语言模型能在上下文中进行探索吗？*发表于《神经信息处理系统进展》（NeurIPS）*, 2024。

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, 等。通过强化学习训练语言模型进行自我修正。*发表于《国际学习表征会议》（ICLR）*, 2024。

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, maxime gazeau, Himanshu Sahni, Satinder Singh, 和 Volodymyr Mnih。基于算法蒸馏的上下文强化学习。*发表于《国际学习表征会议》（ICLR）*, 2023。

Yinghao Li, Haorui Wang, 和 Chao Zhang。评估大型语言模型的逻辑谜题解决能力：来自扫雷案例研究的见解。*发表于2024年计算语言学协会北美分会会议：人类语言技术*, 2024。

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, 和 Jie Tang。Agentbench：评估作为智能体的大型语言模型。*发表于《国际学习表征会议》（ICLR）*, 2024。

11

【高三解读】

这一页看似枯燥，实则是人工智能领域的“名人堂”或“武林秘籍索引”。它不仅仅是一份参考文献列表（Bibliography），更像是一张描绘当前AI前沿阵地——特别是“推理能力”和“自主学习”——的藏宝图。对于高三同学来说，你们写论文或研究性学习报告时最后也会列出参考文献，但这页纸上的每一行引用，都代表了过去几年甚至几个月内，全球最顶尖的头脑在解决最困难问题时留下的脚印。

核心概念：从“鹦鹉学舌”到“深度思考”

这页参考文献主要围绕一个核心主题：**如何让AI不仅仅是模仿人类说话，而是真正具备逻辑推理和自我进化的能力。**

1. **推理的觉醒（Reasoning）**：请注意列表中的 **DeepSeek-R1**（2025年）和 **Cobbe** 的数学验证器（2021年）。这代表了AI的一个重大转折。以前的AI像是一个背诵了整个图书馆的学生，你问什么它答什么，但遇到没见过的数学题就会瞎编。现在的研究方向（如DeepSeek-R1）是让AI学会“慢思考”，通过强化学习（Reinforcement Learning）来推导逻辑链条，就像你们做数学大题时，一步步写出“因为……所以……”，而不是直接猜答案。
2. **学会如何学习（Meta-Learning）**：列表中的 **Finn (MAML, 2017)** 和 **Duan (RL², 2016)** 提到的“元学习”是一个非常高级的概念。简单来说，就是“教AI如何学习”。普通的AI是在学习具体的知识（比如背单词），而元学习是让AI掌握学习的方法（比如学会了拼读规则，以后看到新单词就能自己读）。RL² 提到的“Fast reinforcement learning via slow reinforcement learning”就像是：人类经过几百万年的进化（慢速学习）获得了大脑结构，这让我们在出生后能迅速学会骑自行车（快速学习）。
3. **智能体与评估（Agents & Benchmarks）**：像 **AgentBench (Xiao Liu, 2024)** 和 **Hendrycks (APPS, MATH)** 的论文，关注的是“考试”。我们要怎么知道AI真的变强了？不能只看它聊天溜不溜，而是要看它能不能写代码（APPS）、解奥数题（MATH）、甚至玩扫雷游戏（Yinghao Li, 2024）。这标志着AI从“聊天机器人”向“能干活的智能助手”转变。

难点解析：什么是“强化学习（RL）”和“上下文（In-context）”?

- **强化学习（Reinforcement Learning, RL）**：你可以把它想象成“驯兽”。你不能直接告诉小狗怎么做算术，但当它做对了动作，你给它一块饼干（奖励）；做错了，就没有饼干甚至批评（惩罚）。列表中提到的 **DeepSeek-R1** 和 **Kumar (Self-correct)** 就是用这种方法，让AI通过不断的尝试和自我修正，找到解决复杂问题的最优路径，而不是仅仅模仿标准答案。
- **上下文学习（In-context Learning）**：参考 **Krishnamurthy** 的论文。这指的是模型不需要重新训练（不需要去“上学”重修），仅仅通过你刚才跟它说的几句话（上下文），就能迅速掌握新任务。这就像你在考场上，看到题目里给了一个新定义的公式，你立刻就能现学现用，解出答案。这是衡量AI智商高低的关键指标。

知识联想：连接你的高中学科

- **与生物学的联系**：文献中的 **RL²** 理论与生物进化论高度相关。我们在高中生物学过“自然选择”，那是物种层面的“慢速学习”；而个体的神经反射建立则是“快速学习”。AI正在模仿这种生物机制，试图在只有少量数据的情况下也能迅速适应环境。
- **与数学的联系**：Hendrycks 的 **MATH 数据集** 论文告诉我们，数学是检验真理的唯一标准。对于AI来说，写一篇优美的散文可能很容易（因为文无第一），但解出一道高中数学导数大题却很难（因为逻辑

必须严丝合缝）。如果AI能攻克数学，就意味着它真正具备了理性。

- **与历史的联系**：看看这些年份，从2016、2017的基础理论（MAML, RL²），到2021年的数据集建设（MATH, Codex），再到2024、2025年的大爆发（DeepSeek-R1, Llama 3）。这就像历史书上的“工业革命”时间轴，我们正处在“智能革命”的爆发期。每一篇论文都是这个时代的一个里程碑。

总结：这一页不仅仅是书目，它是AI进化树的“基因图谱”。它展示了科学家们如何一步步把计算机从一个冷冰冰的计算器，培养成一个能思考、能反思、甚至能像你一样通过做题来提升自己的“智慧生命体”。

第 12 页

【原文翻译】

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. **Reft：通过强化微调进行推理 (Reft: Reasoning with reinforced fine-tuning)**。发表于 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*，2024年。

Seungyong Moon, Bumsoo Park, and Hyun Oh Song. **引导式搜索流：通过最佳路径引导让语言模型更好地搜索 (Guided stream of search: Learning to better search with language models via optimal path guidance)**。*arXiv 预印本 arXiv:2410.02992*，2024年。

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. **s1：简单的测试时扩展 (s1: Simple test-time scaling)**。*arXiv 预印本 arXiv:2411.13303*（原文此处年份标注为2025，通常指预计发表年份或勘误，实际s1论文多见于2024/2025周期）。

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. **Mlgym：一个用于推进AI研究智能体的新框架和基准 (Mlgym: A new framework and benchmark for advancing ai research agents)**。*arXiv 预印本 arXiv:2502.144*，2025年。

Allen Nie, Yi Su, Bo Chang, Jonathan N Lee, Ed H Chi, Quoc V Le, and Minmin Chen. **Evolve：评估和优化用于探索的大型语言模型 (Evolve: Evaluating and optimizing llms for exploration)**。*arXiv 预印本 arXiv:2410.06238*，2024年。

Dongmin Park, Minkyu Kim, Beongjun Choi, Junhyuck Kim, Keon Lee, Jonghyun Lee, Inkyu Park, Byeong-Uk Lee, Jaeyoung Hwang, Jaewoo Ahn, Ameya S. Mahabaleshwarkar, Bilal Kartal, Pritam Biswas, Yoshi Suhara, Kangwook Lee, and Jaewoong Cho. **Orak：用于在多样化视频游戏中训练和评估LLM智能体的基础基准 (Orak: A foundational benchmark for training and evaluating llm agents on diverse video games)**。*arXiv 预印本 arXiv:2506.03610*，2025年。

Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. **生成式智能体：人类行为的交互式模拟 (Generative agents: Interactive simulacra of human behavior)**。发表于 *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*，2023年。

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. **递归内省：教语言模型智能体如何自我改进 (Recursive introspection: Teaching language model agents how to self-improve)**。发表于 *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*，2024年。

Yuxiao Qu, Matthew Y. R. Yang, Amrith Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. **通过元强化微调优化测试时计算 (Optimizing test-time compute via meta reinforcement finetuning)**。发表于 *International Conference on Machine Learning (ICML)*，2025年。

Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. **用于深度强化学习的想象增强智能体 (Imagination-augmented agents for deep reinforcement learning)**。发表于 *Advances in Neural Information Processing Systems (NeurIPS)*，2017年。

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. **近端策略优化算法 (Proximal policy optimization algorithms)**。*arXiv 预印本 arXiv:1707.06347*，2017年。

Amrith Setlur, Matthew YR Yang, Charlie Snell, Jeremy Greer, Ian Wu, Virginia Smith, Max Simchowitz, and Aviral Kumar. **e3：学会探索实现了LLM测试时计算的外推 (e3: Learning to explore enables extrapolation of test-time compute for LLMs)**。*arXiv 预印本 arXiv:2506.09026*，2025年。

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. **Deepseekmath：推动开放语言模型中数学推理的极限 (Deepseekmath: Pushing the limits of mathematical reasoning in open language models)**。*arXiv 预印本 arXiv:2402.03300*，2024年。

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. **Hybridflow：一个灵活高效的RLHF框架 (Hybridflow: A flexible and efficient rlhf framework)**。发表于 *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*，2025年。

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. **Reflexion：具有口头强化学习能力的语言智能体 (Reflexion: language agents with verbal reinforcement learning)**。发表于 *Advances in Neural Information Processing Systems (NeurIPS)*，2023年。

【高三解读】

这是一页看似枯燥、实则暗藏玄机的“学术藏宝图”。对于一名高三学生来说，你可能习惯了课本上那些已经沉淀了百年的经典定律，但这页纸展示的是人类科技最前沿的**“正在进行时”**。这是一份参考文献列表（Bibliography），它记录了某项最新人工智能研究背后的“巨人肩膀”。

核心概念：这页纸在讲什么？

如果把一篇顶尖的AI论文比作一部好莱坞科幻大片，那么这页纸就是片尾的“致谢名单”。但它比电影名单更重要，因为这其中列出的每一项研究，都是构建现代**“强人工智能（AGI）”**的一块拼图。

通过解读这些标题，我们可以清晰地看到当前AI进化的三大核心趋势：

1. **从“聊天机器人”到“智能体（Agents）”**：
   注意看列表中反复出现的词——**"Agents"**（如 *Generative agents*, *Mlgym*, *Orak*）。现在的AI不再满足于仅仅做一个“问答机器”（你说一句，它回一句）。科学家们正在把AI变成**“智能体”**，就像游戏里的角色一样，它们拥有记忆、能规划未来、能模拟人类行为（如 *Joon Sung Park* 的研究，模拟了一个像《模拟人生》一样的小镇），甚至能玩视频游戏。这是AI从“工具”向“拟人化”迈出的关键一步。
2. **不仅要“学得快”，还要“想得深”**：
   这是一个非常新的热点，对应列表中的 **"Test-time compute"**（测试时计算）和 **"Reasoning"**（推理）。以前的AI追求“快”，你问什么它秒回。但现在的研究（如 *s1*, *e3*, *Deepseekmath*）发现，如果要解决复杂的奥数题或科学难题，必须允许AI“停下来思考一会儿”，就像你在考场上做压轴题需要打草稿一样。让AI学会“慢思考”，是通往超级智能的必经之路。
3. **自我反思与进化**：
   看标题中的 **"Reflexion"**（反思）、**"Recursive introspection"**（递归内省）和 **"Self-improve"**（自我改进）。这像极了高三学生的复习过程：AI开始学习如何“检查自己的作业”，发现错误后自我修正，而不是仅仅依赖人类老师的打分。这种“元认知”能力（即关于思考的思考）是智慧的高级表现。

难点解析：不仅是英语，更是逻辑

- **arXiv 是什么?**
  你会发现很多条目写着 *arXiv preprint*。arXiv（发音类似 archive）是一个在线预印本网站。在传统的科学界，发表论文要经过几个月的同行评审。但在AI这个发展速度以“天”计算的领域，科学家们等不及了，他们先把论文挂在 arXiv 上抢占先机。所以当你看到 2024、2025 年的 arXiv 引用时，意味着你正在接触**人类知识的绝对边界**，甚至是尚未正式出版的未来构想。
- **强化学习（Reinforcement Learning / RL）：**
  文中多次出现 *Reft*, *PPO*, *RLHF*。你可以把它想象成**“训狗法”**。并不是告诉AI具体的语法规则（像教数学公式），而是给它一个任务，它做对了就给“糖吃”（正反馈，提高该行为的参数权重），做错了就“惩罚”（负反馈）。这是目前让ChatGPT等大模型变得“听话”和“聪明”的核心技术。
- **测试时计算（Test-time Compute）：**
  这是一个针对“应试”的策略。传统的模型训练（Training）像平时上课，一旦模型训练好了，参数就固定了。而“测试时计算”是指在推理（Inference）也就是“考试”的时候，通过消耗更多的算力（时间），进行多路径的搜索和验证，从而由量变引起质变，解决原本解决不了的难题。

知识联想：连接你的高中课堂

1. **生物与进化（Biology - Evolution）：**
   看那个叫 *Evolve* 的标题。计算机科学家正在模仿生物界的**“自然选择”**。他们生成无数个AI模型变体，让它们在虚拟环境中竞争，优胜劣汰，自动演化出更强的算法。这正是达尔文进化论在数字世界的重演。
2. **数学与优化（Math - Optimization）：**
   Proximal Policy Optimization (PPO) 听起来很吓人，其实它的核心思想和你高中数学学的**“导数”与“极值”**息息相关。训练AI本质上就是在一个有着数十亿个维度的高维曲面上，寻找“损失函数”的最低点（即误差最小化）。所有的这些算法，都是为了让你在下山（梯度下降）的过程中不至于摔倒，并且能找到最低的山谷。
3. **学术规范（Academic Integrity）：**
   当你写议论文时，老师要求你“引用名言”或“举例论证”。这页纸就是学术界的最高标准。每一个观点都有出处，每一项技术都有源头。这告诉我们：**科学不是孤独的天才灵光一闪，而是无数人智慧的接力。** 未来的某一天，也许你的名字也会出现在这样的列表中，成为后人引用的基石。

总结：
这页纸不仅仅是一串名单，它是AI正在经历的一场**“成人礼”**——从简单的模仿者，变成会反思、会推理、会像人类一样探索世界的智能生命。作为高三学生，你正站在这个时代的门槛上，数理化基础就是你通往这个新世界的门票。

第 13 页

【原文翻译】

参考文献

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht.
Alfworld：对齐文本与具身环境以进行交互式学习。发表于 *International Conference on Learning Representations*, 2021。

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar.
扩展大语言模型的测试时（Test-time）计算量可能比扩展参数对推理更有效。发表于 *International Conference on Learning Representations*, 2025。

Bradly C Stadie, Ge Yang, Rein Houthooft, Xi Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, and Ilya Sutskever.
关于通过元强化学习学习探索的一些思考。发表于 *Advances in Neural Information Processing System*, 2018。

Fahim Tajwar, Yiding Jiang, Abitha Thankaraj, Sumaita Sadia Rahman, J Zico Kolter, Jeff Schneider, and Russ Salakhutdinov.
训练一个普遍好奇的智能体。发表于 *International Conference on Machine Learning*, 2025。

Sebastian Thrun and Lorien Pratt.
学会学习：介绍与综述。发表于 *Learning to learn*. Springer, 1998。

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar.
Voyager：基于大语言模型的开放式具身智能体。发表于 *Transactions on Machine Learning Research*, 2024a。

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick.
学习进行强化学习。*arXiv preprint arXiv:1611.05763*, 2016。

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al.
关于基于大语言模型的自主智能体的综述。发表于 *Frontiers of Computer Science*, 2024b。

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al.
Ragen：通过多轮强化学习理解大语言模型智能体的自我进化。*arXiv preprint arXiv:2504.20073*, 2025。

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al.
思维链（Chain-of-thought）提示在大型语言模型中引发推理能力。发表于 *Advances in Neural Information Processing Systems*, 2022。

Robert C Wilson, Andra Geana, John M White, Elliot A Ludvig, and Jonathan D Cohen.
人类利用定向和随机探索来解决探索-利用困境。发表于 *Journal of Experimental Psychology: General*, 2014。

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang.
推理扩展定律：对语言模型解决问题时计算最优推理的实证分析。*arXiv preprint arXiv:2408.00724*, 2025。

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang.
针对数学推理的自我奖励修正。*arXiv preprint arXiv:2502.19613*, 2025。

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al.
Qwen3 技术报告。*arXiv preprint arXiv:2505.09388*, 2025。

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan.
Webshop：基于具身语言智能体迈向可扩展的真实世界网络交互。发表于 *Advances in Neural Information Processing Systems*, 2022。

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan.
思维树（Tree of thoughts）：利用大型语言模型进行深思熟虑的问题解决。发表于 *Advances in Neural Information Processing Systems*, 2023a。

【高三解读】

高三解读：AI 进化的“军火库”清单

这张图片是一篇前沿人工智能学术论文的**参考文献（References）**页。虽然看起来只是一堆书名和作者，但如果把它们串联起来，你就能看到当前 AI 领域最激烈的“军备竞赛”主战场。这篇论文（推测写于 2025 年）引用了大量 2024-2025 年的最新研究，同时也致敬了经典的心理学和计算机理论。

我们可以把这张清单看作是**“构建超级智能体（Agent）的说明书”**，其中包含四大核心模块：

1. 核心概念：从“聊天机器人”到“行动者”

这张列表里反复出现一个词：**Agent（智能体）**（如 *Alfworld, Voyager, Webshop, Ragen*）。

- **以前的 AI**（如 ChatGPT 刚出来时）主要是“聊天机器人”，你问它答，它活在对话框里。
- **现在的 AI**（这页纸讨论的重点）是“具身智能体”（Embodied Agent）。它们不仅能说话，还能像人一样去**操作环境**。例如，参考文献中的 **Voyager (Wang et al., 2024a)** 就是一个能自己玩《我的世界》（Minecraft）的 AI，它能像玩家一样探索地图、挖掘资源、打怪升级，而不是只懂文字攻略。

2. 难点解析：让 AI 学会“深思熟虑”

这里有几个非常硬核的概念，决定了 AI 现在的智商上限：

- **思维链与思维树（CoT & ToT）：**
  看到 **Jason Wei (2022)** 的“Chain-of-thought”和 **Yao (2023a)** 的“Tree of thoughts”了吗？这是 AI 变聪明的关键。
  - *比喻*：以前的 AI 做数学题是“秒回”答案，很容易算错。**思维链**就是强迫 AI“把解题步骤一步步写在草稿纸上”。而**思维树**更进一步，不仅写步骤，还在每一步思考“我有几种解法？哪种更好？”，就像下棋时预判未来好几步。
- **测试时计算（Test-time Compute）：**
  Snell (2025) 和 **Wu (2025)** 的论文提到了一个 2025 年最火的概念：**“Inference Scaling”**。
  - *背景*：以前大家觉得模型参数越大（脑子越大）越强。但现在发现，如果不从物理上把脑子变大，而是**给它更多的时间去思考**（在测试/回答时消耗更多算力），效果可能更好。
  - *比喻*：这就像考试。与其花十年时间把高三学生培养成爱因斯坦（增加参数，极难），不如给现在的优等生足够的时间，允许他查资料、反复验算（增加推理计算量），他也能解出难题。
- **元学习（Meta-Learning）：**
  Thrun (1998) 的经典论文《学会学习》在这里出现，说明旧理论焕发了新生。它是指 AI 不仅要学知识（比如背历史年代），还要学习“如何学习”这套方法论（比如学会归纳总结的技巧），从而在遇到全新问题时能迅速适应。

3. 知识联想：跨学科的智慧

这份清单证明了 AI 不是计算机的独角戏，而是多学科的结晶：

- **心理学（Psychology）：**
  参考 **Wilson (2014)** 的《探索-利用困境》（Explore-exploit dilemma）。这是人类决策的经典模型：你去食堂吃饭，是去吃那家你知道肯定好吃的窗口（**利用/Exploit**），还是去尝试一家新开但可能踩雷的窗口（**探索/Explore**）？AI 训练也面临这个问题：是重复已知的高分策略，还是冒险尝试新招数以求突破？
- **生物进化论：**
  Wang (2025) 的论文提到了“自我进化”（Self-evolution）。现代 AI 正试图模仿生物进化，通过不断的“尝试-反馈-修正”循环，让智能体在没有人类手把手教的情况下，自我迭代变得更强。

总结给同学的话：
如果你立志投身 AI，这张页面告诉你，未来的方向不再仅仅是“训练更大的模型”，而是如何设计出**懂得思考、能自主行动、并且会自我反思**的智能体。这需要你不仅数学好、编程好，还要懂认知科学和逻辑学！

第 14 页

【原文翻译】

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, 和 Yuan Cao。React：协同语言模型中的推理与行动 (React: Synergizing reasoning and acting in language models)。载于 *国际学习表征会议 (International Conference on Learning Representations)*, 2023b。

Eric Zelikman, Yuhuai Wu, Jesse Mu, 和 Noah Goodman。
Star：用推理引导推理 (Star: Bootstrapping reasoning with reasoning)。载于 *神经信息处理系统进展大会 (Advances in Neural Information Processing Systems)*, 2022。

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, 等。
大语言模型代理强化学习全景：综述 (The landscape of agentic reinforcement learning for LLMs: A survey)。*arXiv 预印本 arXiv:2509.02547*, 2025。

【高三解读】

核心概念：AI 进化的“巨人肩膀”

这一页虽然看起来只是枯燥的“参考文献（References）”列表，但对于做学术研究的人来说，这里藏着通往真理的地图。这三篇文献并非随意罗列，它们代表了当前大语言模型（LLM）向“智能体（Agent）”进化的三个**里程碑时刻**。如果说之前的 AI 只是一个会聊天的“百科全书”，那么这页纸上引用的技术，正在试图把 AI 变成一个能思考、能干活、能自我反思的“行动派”。

简单来说，这一页展示了作者在构建自己的理论大厦时，使用了哪些最顶尖的基石：

1. **ReAct**：教 AI 一边动脑子（推理），一边动手（行动）。
2. **STaR**：教 AI 像好学生一样“自学”，通过做错题来提升自己。
3. **Agentic RL Survey**：这是一份最新的（2025年）全景地图，总结了如何用强化学习让 AI 变得更有“代理感”（即自主性）。

难点解析：如何让模型“动”起来?

这里有两个核心术语需要为你拆解，它们是理解现代 AI 逻辑的关键：

1. **ReAct (Reasoning + Acting)**：
   - **概念拆解**：以往的模型要么只会在脑子里想（输出文本），要么只会闷头做事（执行代码）。第一篇文献提出的 ReAct 框架，就是打破这个隔阂。它要求 AI 遵循“观察-思考-行动”的循环。
   - **生活案例**：想象你在做一道很难的高中物理实验题。你不会上来就写答案，而是先**观察**题目条件，**思考**用什么公式（Reasoning），然后动手**计算**或查表（Acting），再根据算出的结果进行下一步思考。这就是 ReAct 的本质——像人类解决复杂问题一样，把“想”和“做”交织在一起。
2. **Bootstrapping (自举/引导)**：
   - **概念拆解**：第二篇文献提到的 STaR 里的“Bootstrapping”原意是指“拉着自己的鞋带把自己提起来”，在计算机领域指“自举”或“自我迭代”。
   - **生活案例**：这就像一个没有老师指导的学生。他做了一套卷子，发现做错了。但他不只是把答案抄上去，而是看着正确答案，倒推自己的逻辑哪里断了，把推理过程修正后，重新把这道题“喂”给自己作为新的训练数据。通过这种不断的“自我纠错”，不需要外部老师的大量干预，他的水平就能螺旋式上升。这在 AI 训练中是极高阶的技巧。

知识联想：学科的交响乐

- **历史与学术规范**：你看这一页的排版，每一条都包含了作者、题目、出处（会议）和年份。这和你写议论文时引用名言警句是一样的道理，叫“论据来源”。在科学界，这种引用是对前人智慧的致敬，也构成了科学发展的“家谱”。看到 references，你就看到了科学是如何一步步**迭代**（Iteration）的。
- **生物学（元认知）**：ReAct 和 STaR 本质上是在模拟人类大脑的**前额叶皮层**功能。我们不仅思考（Cognition），我们还“思考我们的思考”（Metacognition，元认知）。STaR 让 AI 去分析自己的推理过程，这正是生物学中高级智慧的体现。
- **数学（递归与收敛）**：从数学角度看，STaR 的自我学习过程就是一个**递归函数**。$f_{n+1}(x) = Improve(f_n(x))$，通过不断的迭代 n，让模型的误差逐渐**收敛**（Converge）到最小。高中数学里的数列极限思想，在这里被完美应用到了 AI 的自我进化中。

导师寄语：
同学，看到这里你可能会发现，最顶尖的 AI 研究，其实都在试图用算法还原人类最朴素的学习方法——“知行合一”（ReAct）和“吾日三省吾身”（STaR）。你们现在高三刷题、改错本的过程，其实就是在对自己的大脑进行最硬核的“强化学习”。坚持下去，你就是在训练属于你自己的超级智能！