大型推理模型强化学习综述

Kaiyan Zhang[1*†], Yuxin Zuo[1*†], Bingxiang He[1*], Youbang Sun[1*], Runze Liu[1*], Che Jiang[1*], Yuchen Fan[2,3*], Kai Tian[1*], Guoli Jia[1*], Pengfei Li[2,6*], Yu Fu[9*], Xingtai Lv[1*], Yuchen Zhang[2,4*], Sihang Zeng[7*], Shang Qu[1,2*], Haozhan Li[1*], Shijie Wang[2*], Yuru Wang[1*], Xinwei Long[1], Fangfu Liu[1], Xiang Xu[5], Jiaze Ma[1], Xuekai Zhu[3], Ermo Hua[1,2], Yihao Liu[1,2], Zonglin Li[2], Huayu Chen[1], Xiaoye Qu[2], Yafu Li[2], Weize Chen[1], Zhenzhao Yuan[1], Junqi Gao[6], Dong Li[6], Zhiyuan Ma[8], Ganqu Cui[2], Zhiyuan Liu[1], Biqing Qi[2‡], Ning Ding[1,2‡], Bowen Zhou[1,2‡]

[1] Tsinghua University [2] Shanghai AI Laboratory [3] Shanghai Jiao Tong University [4] Peking University
[5] University of Science and Technology of China [6] Harbin Institute of Technology [7] University of Washington
[8] Huazhong University of Science and Technology [9] University College London

[†] 项目负责人。 [*] 核心贡献者。 [‡] 通讯作者。

zhang-ky22@mails.tsinghua.edu.cn TsinghuaC3I/Awesome-RL-for-LRMs

Abstract | 本文综述了大型语言模型 (LLM) 推理能力的强化学习 (RL) 最新进展.RL 在推进 LLM 能力前沿方面取得了显著成功, 特别是在处理数学和编程等复杂逻辑任务方面. 因此,RL 已成为将 LLM 转化为大型推理模型 (LRM) 的基础方法. 随着该领域的快速发展,RL 在 LRM 中的进一步扩展现在面临基础性挑战, 不仅体现在计算资源方面, 还体现在算法设计, 训练数据和基础设施方面. 为此, 及时重新审视该领域的发展, 重新评估其轨迹, 并探索增强 RL 向人工超级智能 (ASI) 扩展的策略是十分必要的. 特别是, 我们研究了将 RL 应用于 LLM 和 LRM 推理能力的研究, 特别是自 DeepSeek-R1 发布以来的研究, 包括基础组件, 核心问题, 训练资源和下游应用, 以确定这一快速发展领域的未来机遇和方向. 我们希望本综述将促进更广泛推理模型的 RL 未来研究.

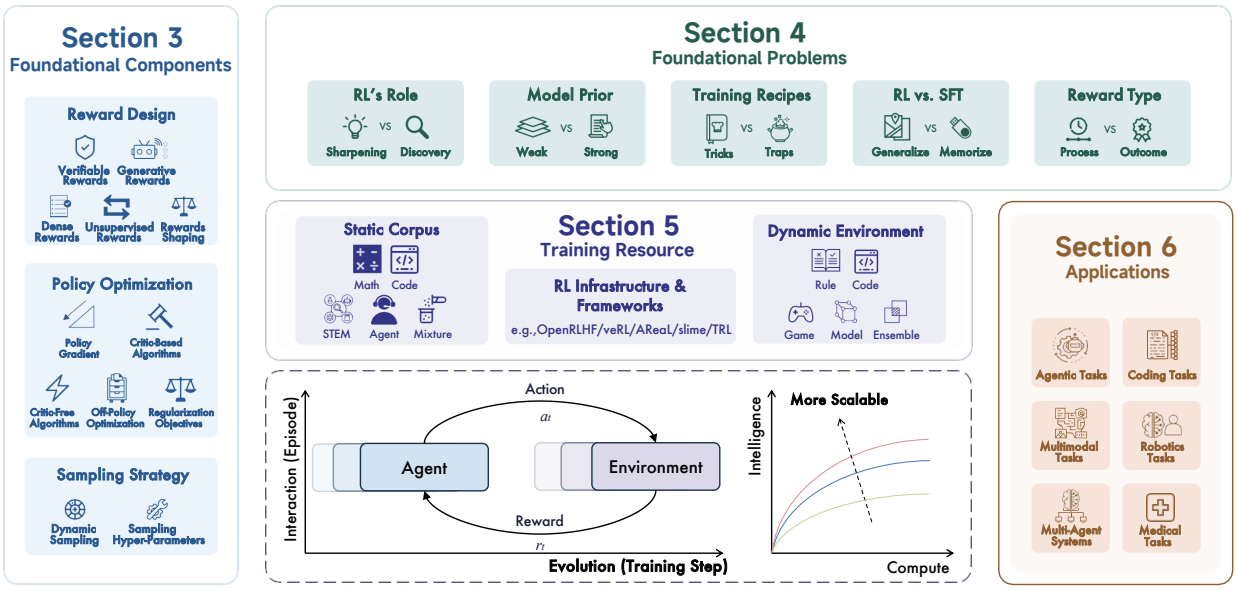


Figure 1 | 综述概览. 我们介绍了 LRM 的 RL 基础组件, 以及开放问题, 训练资源和应用. 本综述的核心关注点是语言智能体与环境在整个长期演化过程中的大规模交互.

Contents

1. 引言

强化学习 (RL) [Sutton et al., 1998] 已经反复证明, 狭隘且明确定义的奖励信号可以驱动人工智能体在复杂任务上达到超人的能力. 像 AlphaGo [Silver et al., 2016] 和 AlphaZero [Silver et al., 2017] 这样的标志性系统, 完全通过自我博弈和奖励反馈来学习, 在围棋, 国际象棋, 将棋和 Stratego [Perolat et al., 2022, Schrittwieser et al., 2020, Silver et al., 2018] 中超越了世界冠军, 确立了 RL 作为高级问题解决的实际且有前景的技术. 在大语言模型 (LLMs) 时代 [Zhao et al., 2023a],RL 最初作为人类对齐的后训练策略而崭露头角 [Ouyang et al., 2022]. 广泛采用的方法如人类反馈强化学习 (RLHF) [Christiano et al., 2017] 和直接偏好优化 (DPO) [Rafailov et al., 2023] 微调预训练模型以遵循指令并反映人类偏好, 显著改善了有用性, 诚实性和无害性 (3H) [Bai et al., 2022b].

最近, 一个新的趋势已经出现: 用于大型推理模型 (LRMs) 的强化学习 [Xu et al., 2025a], 其目标不仅仅是行为对齐, 更是激励推理本身. 两个近期的里程碑 (即 OpenAI o1 [Jaech et al., 2024] 和 DeepSeek-R1 [Guo et al., 2025a]) 表明, 使用可验证奖励的强化学习 (RLVR) 训练 LLMs, 例如数学答案正确性或代码单元测试通过率, 可以使模型执行长篇推理, 包括规划, 反思和自我修正. OpenAI 报告 [Jaech et al., 2024] 称,o1 的性能随着额外的 RL(增加训练时计算) 和推理时更多 “思考” 时间 (测试时计算) [Brown et al., 2024, Liu et al., 2025m, Snell et al., 2024] 而平滑改善, 揭示了超越预训练本身的新扩展轴 [Aghajanyan et al., 2023, Kaplan et al., 2020]. DeepSeek-R1 [Guo et al., 2025a] 采用显式的, 基于规则的数学准确性奖励, 以及基于编译器或测试的编程任务奖励. 这种方法表明, 大规模强化学习, 特别是群体相对策略优化 (GRPO), 即使在后续对齐阶段之前的基础模型中也能诱导复杂的推理行为.

这种转变将推理重新定义为一个可以明确训练和扩展的能力 [OpenAI, 2025a,b]:LRMs 分配大量测试时计算来生成, 评估和修正中间的思维链 [Wei et al., 2022], 其性能随着计算预算的增加而提升. 这种动态引入了能力提升的互补路径, 与预训练期间的数据和参数扩展正交 [Aghajanyan et al., 2023, Kaplan et al., 2020], 同时利用奖励最大化目标 [Silver et al., 2021], 在存在可靠验证器的任何地方使用可自动检查的奖励 (例如, 竞赛数学 [Guo et al., 2025a, Jaech et al., 2024], 竞争性编程 [El-Kishky et al., 2025] 和选定的科学领域 [Bai et al., 2025]). 此外,RL 可以通过自生成训练数据 [Silver et al., 2018, Zhao et al., 2025a] 来克服数据限制 [Shumailov et al., 2024, Villalobos et al., 2022]. 因此,RL 越来越多地被视为通过持续扩展在更广泛任务范围内实现人工超级智能 (ASI) 的有前景技术.

与此同时,LRMs 的 RL 进一步扩展带来了新的约束, 不仅在计算资源方面, 还在算法设计, 训练数据和基础设施方面.LRMs 的 RL 如何以及在哪里可以扩展以实现高级智能并产生实际价值仍然是未解决的问题. 因此, 我们认为现在正是重新审视这一领域发展并探索增强 RL 向人工超级智能扩展策略的适当时机. 总之, 本综述回顾了 LRMs 的 RL 的最新工作, 具体如下:

- 我们介绍了 LRMs 背景下 RL 建模的初步定义 (§ 2.1), 并概述了自 OpenAI o1 发布以来前沿推理模型的发展 (§ 2.2).
- 我们回顾了 LRMs 的 RL 基础组件的最新文献, 包括奖励设计 (§ 3.1), 策略优化 (§ 3.2) 和采样策略 (§ 3.3), 比较了每个组件的不同研究方向和技术方法.

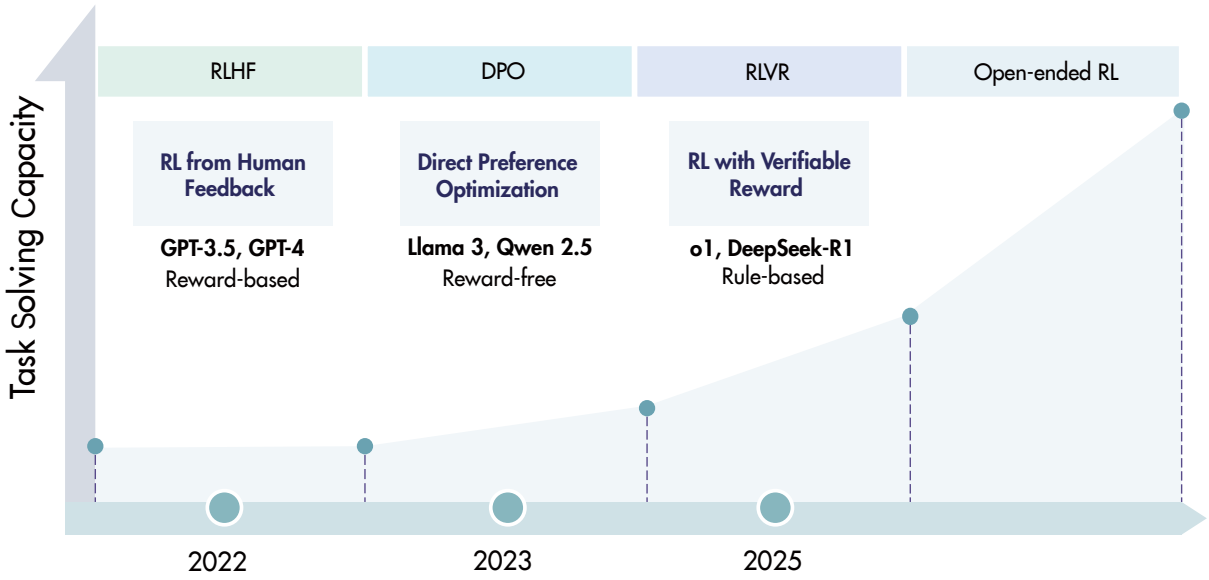


Figure 2 | RLHF 和 DPO 一直是近年来人类对齐的两个主要 RL 方法. 相比之下,RLVR 代表了 LRMs 中 RL 的新兴趋势, 显著增强了它们解决复杂任务的能力.LLMs 的 RL 扩展的下一阶段仍然是一个开放性问题, 开放式 RL 呈现了一个特别具有挑战性和前景的方向.

- 我们讨论了 LRMs 的 RL 中基础性且仍有争议的问题 (§ 4), 如 RL 的作用 (§ 4.1),RL 与监督微调 (SFT) 的对比 (§ 4.2), 模型先验 (§ 4.3), 训练配方 (§ 4.4) 和奖励定义 (§ 4.5). 我们认为这些问题值得进一步探索, 以实现 RL 的持续扩展.
- 我们考察了 RL 的训练资源 (§ 5), 包括静态语料库 (§ 5.1), 动态环境 (§ 5.2) 和训练基础设施 (§ 5.3). 虽然这些资源在研究和生产中都可重用, 但仍需要进一步的标准化和开发.
- 我们回顾了 RL 在广泛任务中的应用 (§ 6), 如编程任务 (§ 6.1), 智能体任务 (§ 6.2), 多模态任务 (§ 6.3), 多智能体系统 (§ 6.4), 机器人任务 (§ 6.5) 和医疗应用 (§ 6.6).
- 最后, 我们讨论了语言模型 RL 的未来方向 (§ 7), 涵盖新算法, 机制, 功能和额外的研究途径.

2. 预备知识

2.1. 背景

在本小节中, 我们介绍 RL 的基本组成部分, 并描述如何将语言模型配置为 RL 框架中的智能体. 如图 3 所示,RL 为序列决策提供了一个通用框架, 其中智能体通过采取行动与环境交互, 以最大化累积奖励. 在经典 RL 中, 问题通常被表述为马尔可夫决策过程 (MDP) [Sutton et al., 1998], 其由元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$ 定义. 主要组成部分包括状态空间 $\mathcal{S}$, 动作空间 $\mathcal{A}$, 转移动态 $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$, 奖励函数 $R : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ 和折扣因子 $\gamma \in [0, 1]$. 在每一步中, 智能体观察状态 s_t, 根据其参数为 θ 的策略 π_θ

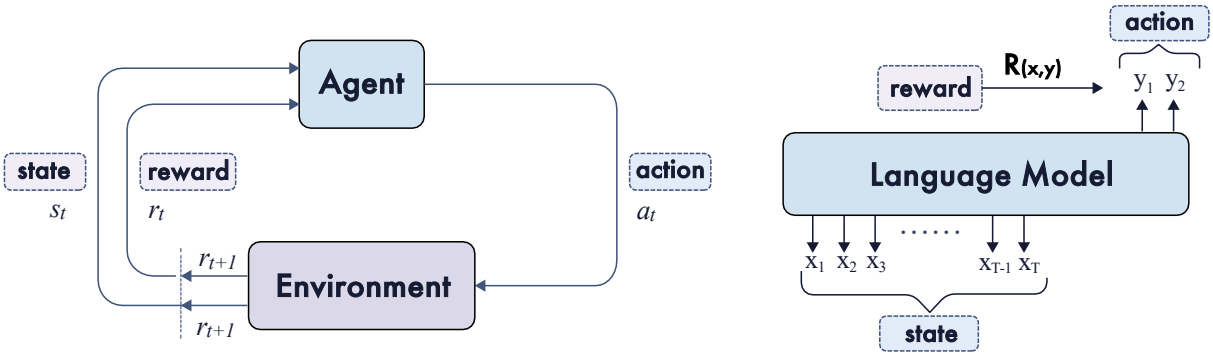


Figure 3 | RL 和语言模型 (LMs) 作为智能体的基本组成部分. 智能体选择动作, 而环境在每轮中提供状态和奖励. 在 LM 的上下文中, 补全 token 被视为动作, 与上下文连接形成状态. 奖励通常在整个响应级别分配.

选择动作 a_t, 接收奖励 r_t, 并转移到下一个状态 s_{t+1}. 当将 RL 应用于语言模型时, 这些概念可以以最小的调整自然地映射到语言领域. 映射总结如下:

- **提示/任务** (x): 对应于初始状态或环境上下文, 从数据分布中抽取并对应于数据集 $\mathcal{D}$.
- **策略** (π_θ): 表示语言模型, 其生成长度为 T 的序列, 记为 $y = (y_1, \ldots, y_T)$ 以响应提示.
- **状态** (s_t): 定义为提示与迄今为止生成的 token, 即 $s_t = (x, a_{1:t-1})$.
- **动作** (a_t): 在步骤 t 从动作空间 $\mathcal{A}$ 中选择的单位. 根据粒度, 动作可以是整个序列 y(序列级),token $a_t \in \mathcal{V}$(token 级) 或段 $y^{(k)} = (y_1^{(k)}, \ldots, y_{T_k}^{(k)})$(步骤级), 详细比较见表 2.
- **转移动态** ($\mathcal{P}$): 在 LLM 的上下文中, 状态转移通常是确定性的, 因为 $s_{t+1} = [s_t, a_t]$, 其中 $[\cdot, \cdot]$ 表示字符串连接. 当状态包含 `EOS` token 时, 策略转移到终止状态, 意味着轨迹结束.
- **奖励** ($R(x, y)$ **或** r_t): 根据动作粒度分配, 例如轨迹结束时的序列级 $R(x, y)$, 每个 token 的 token 级 $r_t = R(x, a_{1:t})$ 或每个段的步骤级 $r_k = R(x, y^{(1:k)})$.
- **回报** (G): 提示 x 的整个轨迹 y 的累积奖励 (对于有限水平通常 $\gamma = 1$). 对于序列级奖励, 它简化为单一标量 $R(x, y)$, 否则聚合每 token/步骤奖励, 详细见表 2.

在此设置下, 学习目标 [Sutton et al., 1998] 是最大化数据分布 $\mathcal{D}$ 上的期望累积奖励, 即

$$\max_\theta \mathcal{J}(\theta) := \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(x)}[G]. \tag{1}$$

在实践中, 通常将学习策略正则化到参考策略 π_{ref}, 通常实现为 KL 散度约束, 以稳定训练并保持语言质量. 在以下部分中, 我们介绍构建在此基本公式之上的各种算法.

2.2. 前沿模型

在本小节中, 我们概述了使用类似 RL 的方法训练的最先进大型推理模型, 按时间顺序大致分为三个主要方向:LRMs, 智能体 LRMs 和多模态 LRMs.

过去一年,RL 逐步扩展了推理模型及其应用的前沿. 第一批大型推理模型 OpenAI 的 o1 系列 [Jaech et al., 2024] 确立了扩展训练时 RL 和测试时计算以获得更强大推理能力的有效性, 在数学, 编程和科学基准上取得了领先结果.DeepSeek 的旗舰模型 R1 [Guo et al., 2025a] 作为第一个匹配 o1 在各基准上性能的开源模型紧随其后. 它采用多阶段训练流水线以确保全面的模型能力, 并探索了没有监督微调的纯 RL 路线 (即 Zero RL). 其他专有模型版本很快跟进:Claude-3.7-Sonnet [Anthropic, 2025a] 具有混合推理功能,Gemini 2.0 和 2.5 [Comanici et al., 2025] 引入了更长的上下文长度,Seed-Thinking 1.5 [Seed et al., 2025b] 具有跨领域的泛化能力, 而 o3 系列 [OpenAI, 2025b] 展示了日益先进的推理能力. 最近,OpenAI 推出了他们的第一个开源推理模型 gpt-oss-120b [Agarwal et al., 2025a], 随后是 GPT5 [OpenAI, 2025a], 这是他们迄今为止最强大的 AI 系统, 可在高效模型和更深层次的推理模型 GPT-5 thinking 之间灵活切换. 并行开源努力继续扩展了这一领域. 在 Qwen 家族中,QwQ-32B [Team, 2025g] 匹配了 R1 的性能, 随后是 Qwen3 系列 [Yang et al., 2025a], 其代表性模型 Qwen3-235B 进一步提高了基准分数.Skywork-OR1 [He et al., 2025d] 模型套件基于 R1 蒸馏模型, 通过有效的数据混合和算法创新实现了可扩展的 RL 训练.Minimax-M1 [Chen et al., 2025a] 是第一个引入混合注意力以高效扩展 RL 的模型. 其他工作包括旨在平衡准确性和效率的 Llama-Nemotron-Ultra [Bercovich et al., 2025]; 通过 RL 从头训练而不从先前模型蒸馏的 Magistral 24B [Rastogi et al., 2025]; 以及强调长上下文推理能力的 Seed-OSS [Team, 2025a].

模型推理的改进反过来扩展了它们在编程和智能体场景中的用例.Claude 系列以其在智能体编程任务上的领先性能而闻名,Claude-4.1-Opus [Anthropic, 2025b] 就是例证, 它进一步推动了 SWE-bench [Jimenez et al., 2023] 的最先进结果.Kimi K2 [Team, 2025d] 是近期的一个代表性智能体模型, 专门针对智能体任务进行了优化, 开创了大规模智能体训练数据合成和适应不可验证奖励的通用 RL 程序. 不久之后,GLM4.5 [Zeng et al., 2025a] 和 DeepSeek-V3.1 版本都强调工具使用和智能体任务, 在相关基准上显示出实质性改进.

多模态是推理模型广泛采用背后的关键组成部分. 大多数前沿专有模型, 包括 GPT-5,o3,Claude 和 Gemini 家族, 都是原生多模态的.Gemini-2.5 [Comanici et al., 2025] 特别强调了在文本, 图像, 视频和音频方面的强大性能. 在开源方面,Kimi 1.5 [Team, 2025d] 代表了多模态推理的早期努力, 强调了长上下文扩展以及文本和视觉领域的联合推理.QVQ [Qwen Team, 2025] 在视觉推理和分析思维方面表现出色, 而 Skywork R1V2 [Wang et al., 2025k] 通过混合 RL 平衡推理和一般能力, 同时使用 MPO 和 GRPO. 作为 InternVL 系列的重要补充,InternVL3 [Zhu et al., 2025c] 采用了统一的原生多模态预训练阶段, 后来的 InternVL3.5 [Wang et al., 2025o] 使用了两级级联 RL 框架, 实现了改进的效率和多功能性. 最近,Intern-S1 模型 [Bai et al., 2025] 专注于跨不同领域的多模态科学推理, 受益于在线 RL 期间的混合奖励设计, 便于在广泛任务上同时训练. 其他近期模型包括旨在高效训练和最小化解码成本的 Step3 [Wang et al., 2025a], 以及在大多数视觉多模态基准上具有最先进性能的 GLM-4.5V [Team et al., 2025a].

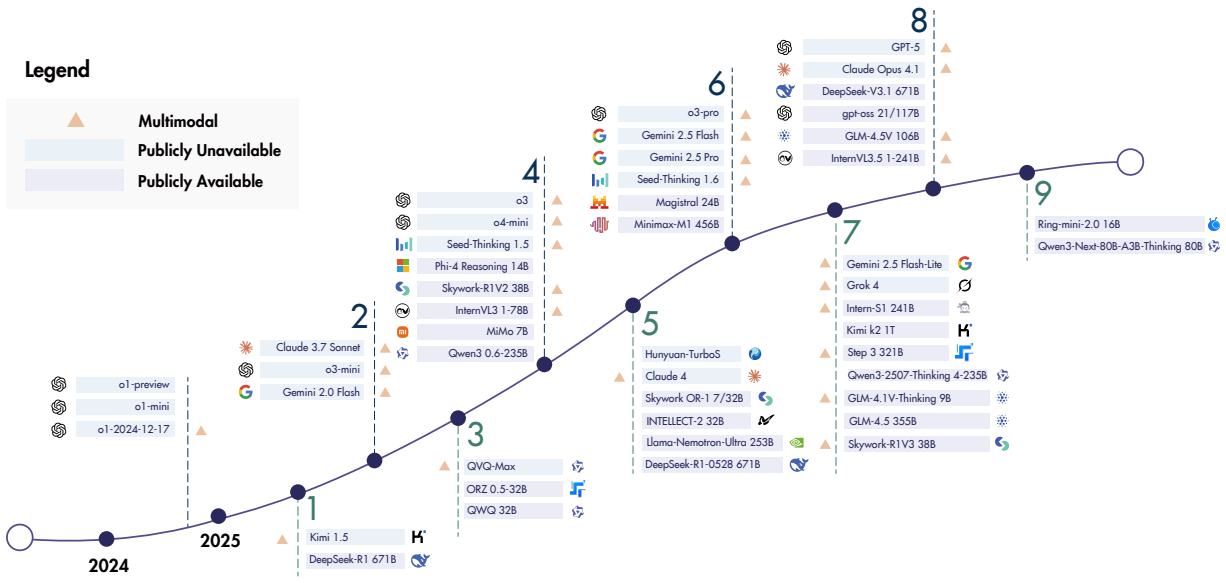


Figure 4 | 使用 RL 训练的开源和闭源推理模型代表的时间线, 包括语言模型, 多模态模型和智能体模型.

除了上述模型外, 我们在图 4 中提供了推理模型的全面列表, 在表 1 中提供了开源模型的详细信息.

2.3. 相关综述

在本小节中，我们比较了与 RL 和 LLMs 相关的近期综述。一些综述主要关注 RL 本身，涵盖了经典 RL 及其近期扩展。Ghasemi et al. [2024] 提出了涵盖算法和实际挑战的通用 RL 综述，Huh and Mohapatra [2023] 专注于多智能体 RL，Zhang et al. [2024b] 回顾了自我对弈技术，Wu et al. [2025h] 综述了计算机视觉任务中的 RL。虽然这些工作为 RL 提供了广阔的视角，但它们没有明确讨论其在 LLMs 中的应用。相比之下，其他综述以 LLMs 及其新兴能力为中心，如长链思维推理 [Chen et al., 2025m, Li et al., 2025w, Xia et al., 2024] 和自适应行为 [Feng et al., 2025c, Sui et al., 2025]，其中 RL 通常被作为支持这些进展的关键方法引入。Zhao et al. [2023a] 提供了 LLM 架构和应用的广泛概述，而更近期的工作则专门集中于推理能力。Zhang et al. [2025a] 在 DeepSeek-R1 出现后综述了推理 LLMs 的复制研究，Chen et al. [2025m] 研究了长链思维推理，Li et al. [2025w] 分析了从系统 1 到系统 2 推理的转换。这些研究强调了 RLHF 和 RLVR 等基于 RL 的方法作为有用工具，但将它们视为广泛推理策略中仅有的一个元素。Sun et al. [2025b] 提供了通过基础模型进行推理的更广泛、结构化的视角。它突出了专门为推理提出或适配的关键基础模型，以及在不同推理任务、方法和基准方面的最新进展。Zhang et al. [2025b] 研究了 RL 如何赋予 LLMs 自主决策和自适应智能体能力。Xu et al. [2025a] 通过讨论 LLMs 的强化推理，强调试错优化如何改进复杂推理，从而更接近我们的关注点。Wu [2025] 通过综述奖励模型和从反馈中学习的策略来补充这一观点。然而，这些工作仍然面向推理性能或奖励设计，而不是为 LLMs 提供 RL 方法作为整体的系统性处理。Srivastava and Aggarwal [2025] 代表了通过回顾用于 LLM 对齐和增强的 RL 算法来连接两个领域的最新尝试，

Table 1 | 使用 RL 训练的开源代表性模型比较.OPMD 表示 Online Policy Mirror Descent;MPO 表示 Mixed Preference Optimization;CISPO 表示 Clipped IS-weight Policy Optimization.T,I 和 V 分别表示文本, 图像和视频模态.

日期	模型	组织	架构	参数量	算法	模态	链接
2025.01	DeepSeek-R1 [Guo et al., 2025a]	DeepSeek	MoE/MLA	671B	GRPO	Text	
2025.03	ORZ [Hu et al., 2025b]	StepAI	Dense	0.5-32B	PPO	Text	
2025.03	QwQ [Team, 2025g]	Alibaba Qwen	Dense	32B	-	Text	
2025.04	Phi-4 Reasoning [Abdin et al., 2025]	Microsoft	Dense	14B	GRPO	Text	
2025.04	Skywork-R1V2 [Wang et al., 2025k]	Skywork	Dense	38B	MPO/GRPO	T/I	
2025.04	InternVL3 [Zhu et al., 2025c]	Shanghai AI Lab	Dense	1-78B	MPO	T/I/V	
2025.04	MiMo [Xiaomi et al., 2025]	Xiaomi	Dense	7B	GRPO	Text	
2025.04	Qwen3 [Yang et al., 2025a]	Alibaba Qwen	MoE/Dense	0.6-235B	GRPO	Text	
2025.05	Llama-Nemotron-Ultra [Bercovich et al., 2025]	NVIDIA	Dense	253B	GRPO	Text	
2025.05	INTELLECT-2 [Team et al., 2025b]	Intellect AI	Dense	32B	GRPO	Text	
2025.05	Hunyuan-TurboS [Team et al., 2025c]	Tencent	Hybrid MoE	560B	GRPO	Text	
2025.05	Skywork OR-1 [He et al., 2025d]	Skywork	Dense	7B/32B	GRPO	Text	
2025.05	DeepSeek-R1-0528 [Guo et al., 2025a]	DeepSeek	MoE/MLA	671B	GRPO	Text	
2025.06	Magistral [Rastogi et al., 2025]	Mistral AI	Dense	24B	GRPO	Text	
2025.06	Minimax-M1 [Chen et al., 2025a]	Minimax	Hybrid MoE	456B	CISPO	Text	
2025.07	Intern-S1 [Bai et al., 2025]	Shanghai AI Lab	MoE	241B	GRPO	T/I/V	
2025.07	Kimi K2 [Team, 2025c]	Kimi	MoE	1T	OPMD	Text	
2025.07	Step 3 [Wang et al., 2025a]	Step AI	MoE	321B	-	T/I/V	
2025.07	Qwen3-2507 [Yang et al., 2025a]	Alibaba Qwen	MoE/Dense	4-235B	GSPO	Text	
2025.07	GLM-4.1V-Thinking [Team et al., 2025a]	Zhipu AI	Dense	9B	GRPO	T/I/V	
2025.07	GLM-4.5 [Zeng et al., 2025a]	Zhipu AI	MoE	355B	GRPO	Text	
2025.07	Skywork-R1V3 [Shen et al., 2025b]	Skywork	Dense	38B	GRPO	T/I	
2025.08	gpt-oss [Agarwal et al., 2025a]	OpenAI	MoE	117B/21B	-	Text	
2025.08	Seed-OSS [Team, 2025a]	Bytedance Seed	Dense	36B	-	Text	
2025.08	GLM-4.5V [Team et al., 2025a]	Zhipu AI	MoE	106B	GRPO	T/I/V	

主要通过 RLHF [Christiano et al., 2017]、RLAIF [Lee et al., 2024b] 和 DPO [Rafailov et al., 2023] 等方法。它仍然主要专注于对齐而不是推理能力。

与涵盖通用 RL 或 LLMs 中推理的先前综述不同，我们将 RL 置于中心位置，并提供其在 LLM 训练整个生命周期中作用的系统性综合，包括奖励设计、策略优化和采样策略。我们的目标是确定在 LRMs 中扩展强化学习以实现 ASI 的新方向，重点关注长期交互和演化。

3. 基础组件

在本节中, 我们回顾了 LRM 的 RL 基础组件, 包括奖励设计 (§ 3.1), 策略优化算法 (§ 3.2) 和采样策略 (§ 3.3). 基础组件的分类如图 5 所示.

3.1. 奖励设计

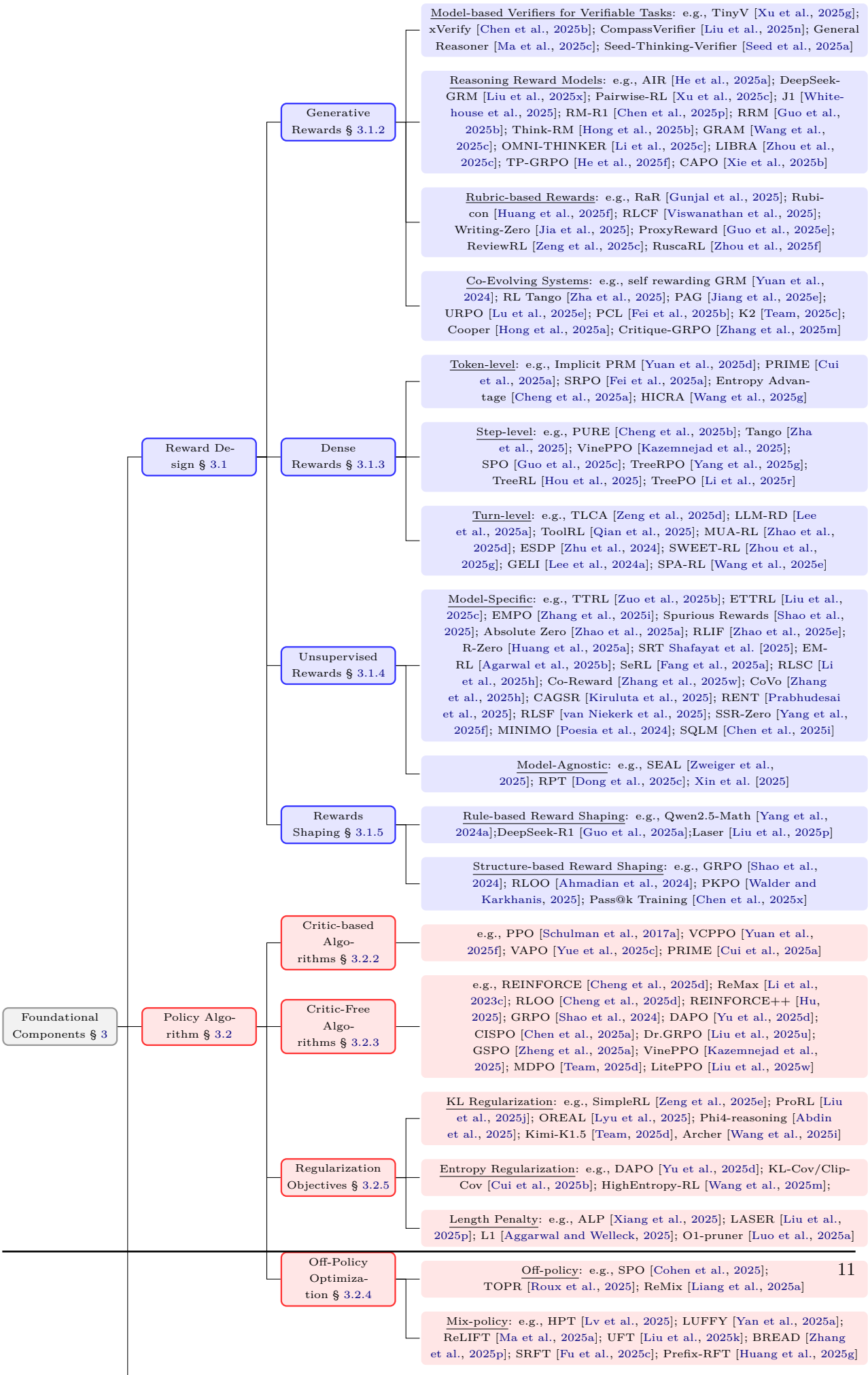
在本小节中, 我们对 LRM 强化学习中的奖励设计进行全面审视. 我们从 § 3.1.1 开始讨论可验证奖励, 这提供了一个自然的起点. 这一方向取得了实质性进展,DeepSeek-R1 的成功就是例证, 它通过可验证奖励机制展示了强化学习的可扩展性. 相比之下,§ 3.1.2 检视生成式奖励, 其中模型被用于验证或直接生成奖励信号. 然而, 可验证奖励和生成式奖励通常都表现为稀疏的数值反馈. 奖励信号的密度是一个重要的补充维度.§ 3.1.3 因此检视了结合密集奖励的方法. 另一个分类轴线涉及奖励是从外部真实标签计算还是由模型直接估计. 这一区别促使我们在 § 3.1.4 中讨论无监督奖励. 在这四个类别的基础上, 我们接着在 § 3.1.5 中转向奖励塑造, 分析组合或转换多样化奖励信号以促进学习的策略.

3.1.1. 可验证奖励

要点

- 基于规则的奖励通过利用准确性和格式检查为强化学习提供可扩展且可靠的训练信号, 特别是在数学和代码任务中.
- 验证者法则强调具有明确自动验证的任务能够实现高效的强化学习优化, 而主观性任务仍然具有挑战性.

基于规则的奖励. 奖励作为强化学习的训练信号, 决定优化方向 [Guo et al., 2025a]. 最近, 基于规则的可验证奖励主要用于大规模强化学习训练 LRM. 这些奖励通过鼓励更长, 更具反思性的思维链来可靠地提升数学和编程推理能力 [Guo et al., 2025a, Team, 2025c, Yu et al., 2025d]. 这种范式在 Tülu 3 [Lambert et al., 2024] 中被形式化为 RLVR, 它用程序化验证器 (如答案检查器或单元测试) 替换学习到的奖励模型. 此类验证器在具有客观可验证结果的领域提供二元, 可检查的信号. 类似的可验证奖励设计基于规则的方法随后被整合到 DeepSeek 的训练流程中. 例如,DeepSeek-V3 [Liu et al., 2024] 明确采用了针对确定性任务的基于规则的奖励系统, 而 DeepSeek-R1 [Guo et al., 2025a] 进一步使用了基于准确性和基于格式的奖励. 基于规则的奖励与基于结果或基于过程的奖励模型 (RM) 形成对比, 如基于人类偏好排名训练的标准 RLHF [Ouyang et al., 2022] 和基于步骤级注释训练的


Foundational Components § 3
Reward Design § 3.1
Policy Algorithm § 3.2
Generative Rewards § 3.1.2
Dense Rewards § 3.1.3
Unsupervised Rewards § 3.1.4
Rewards Shaping § 3.1.5
Critic-based Algorithms § 3.2.2
Critic-Free Algorithms § 3.2.3
Regularization Objectives § 3.2.5
Off-Policy Optimization § 3.2.4
Model-based Verifiers for Verifiable Tasks: e.g., TinyV [Xu et al., 2025g]; xVerify [Chen et al., 2025b]; CompassVerifier [Liu et al., 2025n]; General Reasoner [Ma et al., 2025c]; Seed-Thinking-Verifier [Seed et al., 2025a]
Reasoning Reward Models: e.g., AIR [He et al., 2025a]; DeepSeek-GRM [Liu et al., 2025x]; Pairwise-RL [Xu et al., 2025c]; J1 [Whitehouse et al., 2025]; RM-R1 [Chen et al., 2025p]; RRM [Guo et al., 2025b]; Think-RM [Hong et al., 2025b]; GRAM [Wang et al., 2025c]; OMNI-THINKER [Li et al., 2025c]; LIBRA [Zhou et al., 2025c]; TP-GRPO [He et al., 2025f]; CAPO [Xie et al., 2025b]
Rubric-based Rewards: e.g., RaR [Gunjal et al., 2025]; Rubicon [Huang et al., 2025f]; RLCF [Viswanathan et al., 2025]; Writing-Zero [Jia et al., 2025]; ProxyReward [Guo et al., 2025e]; ReviewRL [Zeng et al., 2025c]; RuscaRL [Zhou et al., 2025f]
Co-Evolving Systems: e.g., self rewarding GRM [Yuan et al., 2024]; RL Tango [Zha et al., 2025]; PAG [Jiang et al., 2025e]; URPO [Lu et al., 2025e]; PCL [Fei et al., 2025b]; K2 [Team, 2025c]; Cooper [Hong et al., 2025a]; Critique-GRPO [Zhang et al., 2025m]
Token-level: e.g., Implicit PRM [Yuan et al., 2025d]; PRIME [Cui et al., 2025a]; SRPO [Fei et al., 2025a]; Entropy Advantage [Cheng et al., 2025a]; HICRA [Wang et al., 2025g]
Step-level: e.g., PURE [Cheng et al., 2025b]; Tango [Zha et al., 2025]; VinePPO [Kazemnejad et al., 2025]; SPO [Guo et al., 2025c]; TreeRPO [Yang et al., 2025g]; TreeRL [Hou et al., 2025]; TreePO [Li et al., 2025r]
Turn-level: e.g., TLCA [Zeng et al., 2025d]; LLM-RD [Lee et al., 2025a]; ToolRL [Qian et al., 2025]; MUA-RL [Zhao et al., 2025d]; ESDP [Zhu et al., 2024]; SWEET-RL [Zhou et al., 2025g]; GELI [Lee et al., 2024a]; SPA-RL [Wang et al., 2025e]
Model-Specific: e.g., TTRL [Zuo et al., 2025b]; ETTRL [Liu et al., 2025c]; EMPO [Zhang et al., 2025i]; Spurious Rewards [Shao et al., 2025]; Absolute Zero [Zhao et al., 2025a]; RLIF [Zhao et al., 2025e]; R-Zero [Huang et al., 2025a]; SRT Shafayat et al. [2025]; EMRL [Agarwal et al., 2025b]; SeRL [Fang et al., 2025a]; RLSC [Li et al., 2025h]; Co-Reward [Zhang et al., 2025w]; CoVo [Zhang et al., 2025h]; CAGSR [Kiruluta et al., 2025]; RENT [Prabhudesai et al., 2025]; RLSF [van Niekerk et al., 2025]; SSR-Zero [Yang et al., 2025f]; MINIMO [Poesia et al., 2024]; SQLM [Chen et al., 2025i]
Model-Agnostic: e.g., SEAL [Zweiger et al., 2025]; RPT [Dong et al., 2025c]; Xin et al. [2025]
Rule-based Reward Shaping: e.g., Qwen2.5-Math [Yang et al., 2024a];DeepSeek-R1 [Guo et al., 2025a];Laser [Liu et al., 2025p]
Structure-based Reward Shaping: e.g., GRPO [Shao et al., 2024]; RLOO [Ahmadian et al., 2024]; PKPO [Walder and Karkhanis, 2025]; Pass@k Training [Chen et al., 2025x]
e.g., PPO [Schulman et al., 2017a]; VCPPO [Yuan et al., 2025f]; VAPO [Yue et al., 2025c]; PRIME [Cui et al., 2025a]
e.g., REINFORCE [Cheng et al., 2025d]; ReMax [Li et al., 2023c]; RLOO [Cheng et al., 2025d]; REINFORCE++ [Hu, 2025]; GRPO [Shao et al., 2024]; DAPO [Yu et al., 2025d]; CISPO [Chen et al., 2025a]; Dr.GRPO [Liu et al., 2025u]; GSPO [Zheng et al., 2025a]; VinePPO [Kazemnejad et al., 2025]; MDPO [Team, 2025d]; LitePPO [Liu et al., 2025w]
KL Regularization: e.g., SimpleRL [Zeng et al., 2025e]; ProRL [Liu et al., 2025j]; OREAL [Lyu et al., 2025]; Phi4-reasoning [Abdin et al., 2025]; Kimi-K1.5 [Team, 2025d], Archer [Wang et al., 2025i]
Entropy Regularization: e.g., DAPO [Yu et al., 2025d]; KL-Cov/Clip-Cov [Cui et al., 2025b]; HighEntropy-RL [Wang et al., 2025m];
Length Penalty: e.g., ALP [Xiang et al., 2025]; LASER [Liu et al., 2025p]; L1 [Aggarwal and Welleck, 2025]; O1-pruner [Luo et al., 2025a]
Off-policy: e.g., SPO [Cohen et al., 2025]; TOPR [Roux et al., 2025]; ReMix [Liang et al., 2025a]
Mix-policy: e.g., HPT [Lv et al., 2025]; LUFFY [Yan et al., 2025a]; ReLIFT [Ma et al., 2025a]; UFT [Liu et al., 2025k]; BREAD [Zhang et al., 2025p]; SRFT [Fu et al., 2025c]; Prefix-RFT [Huang et al., 2025g]

过程奖励模型 (PRM) [Setlur et al., 2024, Sun et al., 2025c, Yuan et al., 2025d]. DeepSeek-V3 和 DeepSeek R1 证明, 当扩展到大规模强化学习设置时,RM 可能遭受奖励黑客攻击, 但通过尽可能利用基于规则的奖励, 我们通过使系统抵抗操纵和利用来确保更大的可靠性 [Guo et al., 2025a, Liu et al., 2024]. 在实践中, 广泛使用两种基于规则的可验证奖励:

- **准确性奖励**: 对于具有确定性结果的任务 (如数学), 策略必须在规定的分隔符内 (通常是 `\boxed{...}`) 产生最终解决方案. 然后自动检查器将此输出与真实标签进行比较. 对于编码任务, 单元测试或编译器提供通过/失败信号 [Albalak et al., 2025, Chen et al., 2025r, Guo et al., 2025a].
- **格式奖励**: 这些施加结构约束, 要求模型将其私有思维链放在 `<think>` 和 `</think>` 之间, 并在单独的字段中输出最终答案 (如 `<answer>...</answer>`). 这提高了大规模强化学习中的可靠解析和验证 [Guo et al., 2025a, Lambert et al., 2024].

基于规则的验证器. 基于规则的奖励通常源自基于规则的验证器. 这些验证器依赖于大量手动编写的等价规则来确定预测答案是否匹配真实标签. 目前, 广泛使用的数学验证器主要建立在 Python 库 Math-Verify[1] 和 SymPy[2] 之上. 此外, 一些工作如 DAPO [Yu et al., 2025d] 和 DeepScaleR [Luo et al., 2025c] 也提供了开源且成熟的验证器. 最近,Huang et al. [2025e] 强调了基于规则和基于模型的验证器所特有的局限性, 为设计更可靠的奖励系统提供指导.

在实践中, 数学问题求解和代码生成等任务难以解决但相对容易验证, 从而满足高效强化学习优化的主要标准 [Guo et al., 2025a, He et al., 2025d]: 存在清晰的真实标签, 可获得快速自动验证, 评估多个候选解决方案的可扩展性, 以及与正确性密切相关的奖励信号. 相比之下, 缺乏快速或客观验证的任务 (如开放式问答或自由形式写作) 对于基于结果的强化学习仍然具有挑战性, 因为它们依赖于噪声学习奖励模型或主观的人类反馈 [Yu et al., 2025e, Zhou et al., 2025e]. 验证者法则假定训练 AI 系统执行任务的难易程度与任务的可验证程度成正比 [3]. 它强调一旦任务可以配备强大的自动反馈, 就变得适合通过强化学习快速改进. §6 中讨论的成功应用证实了这一原则, 因为它们的中心挑战在于设计可靠的可验证反馈. 相反,§7 中强调的许多开放问题正是源于缺乏可靠的自动奖励.

3.1.2. 生成式奖励

要点

- 生成式奖励模型 (GenRM) 通过提供细致的基于文本的反馈, 将强化学习扩展到主观, 不可验证的领域, 克服了基于规则系统的局限性.
- 一个主导趋势是训练 RM 在判断之前进行推理, 通常使用结构化评分标准指导评估, 或在与策略模型的统一强化学习循环中共同进化.

[1] https://github.com/huggingface/Math-Verify

[2] https://www.sympy.org/

[3] https://www.jasonwei.net/blog/asymmetry-of-verification-and-verifiers-law

虽然基于规则的奖励为可验证任务提供可靠信号, 如前所述 (§ 3.1.1), 但其适用性有限. 许多复杂推理任务, 特别是在开放式或创造性领域, 缺乏客观的真实标签, 使其对简单验证器来说是不可处理的. 为了弥合这一差距,GenRM 作为强大的替代方案出现.GenRM 不是输出简单的标量分数, 而是利用 LRM 的生成能力产生结构化的批评, 理据和偏好, 提供更可解释和细致的奖励信号 [Mahan et al., 2024, Zhang et al., 2024a]. 这种方法解决了两个关键挑战: 首先, 它提高了难以解析的可验证任务的验证鲁棒性; 其次, 也是更重要的, 它使强化学习能够应用于主观, 不可验证的领域.

用于可验证任务的基于模型的验证器. 基于规则系统的一个主要挑战是它们的脆弱性; 当模型以意外格式生成正确答案时, 它们经常产生假阴性. 为了缓解这个问题, 一个研究方向使用 基于规范的 GenRM 作为灵活的基于模型的验证器. 这些模型被训练用于语义评估模型的自由格式输出与参考答案之间的等价性. 这种方法已被用于开发增强现有基于规则系统的轻量级验证器 [Xu et al., 2025g], 以及能够处理多样化数据类型和推理任务的更全面的多领域验证器 [Chen et al., 2025b, Liu et al., 2025n, Ma et al., 2025c, Seed et al., 2025a]. 通过用学习的语义判断替换或补充刚性字符串匹配, 这些验证器为可验证领域中的强化学习提供更准确的奖励信号.

用于不可验证任务的生成式奖励. GenRM 的另一个核心应用是 基于评估的 GenRM, 它使强化学习能够应用于验证者法则不成立的任务. 这种范式已经从使用强大的 LLM 作为零样本评估器发展到复杂的共同进化系统. 我们可以根据这些方法的核心设计原则对它们进行分类.

- **推理奖励模型 (学会思考)**: 超越简单偏好预测的一个重大进展是训练 RM 在做出判断之前明确推理. 这种方法是 LLM 即评判者概念的基础 [Li et al., 2023b, Zheng et al., 2023], 涉及提示 RM 生成 CoT 批评或理据. 例如,CLoud RM 首先生成自然语言批评, 然后使用它来预测标量奖励 [Ankner et al., 2024]. 将奖励建模表述为推理任务的这一原则现在是最先进 RM 的核心, 这些 RM 被训练在分配分数或偏好之前产生详细的理据 [Chen et al., 2025p, Guo et al., 2025b, Hong et al., 2025b, Liu et al., 2025x, Wang et al., 2025c, Zhou et al., 2025c]. 为了进一步提高其判断能力, 这些推理 RM 本身经常用强化学习训练, 使用基于其最终裁决正确性的简单, 可验证的元奖励 [Chen et al., 2025l, Whitehouse et al., 2025]. 这一系列工作还探索了不同的奖励格式, 如从 token 概率导出软奖励 [Mahan et al., 2024, Su et al., 2025c, Zhang et al., 2024a], 以及权衡逐点和逐对评分方案之间的平衡 [He et al., 2025a, Xu et al., 2025c].

- **基于评分标准的奖励 (结构化主观性)**: 为了将主观任务的评估建立在更一致的标准上, 许多框架采用结构化评分标准. 与依赖硬编码逻辑处理客观, 可验证任务的基于规则方法不同, 基于评分标准的方法利用自然语言描述来捕获主观, 不可验证领域的细致评估标准, 在这些领域传统的二元规则将是不够的. 这涉及使用 LLM 生成或遵循原则检查表来指导其评估. 像 RaR [Gunjal et al., 2025],QA-LIGN [Dineen et al., 2025],Rubicon [Huang et al., 2025f] 和 RLCF [Viswanathan et al., 2025] 等框架使用此类评分标准产生细粒度, 多方面的奖励. 这个概念扩展到将高级任务分解为一组可验证的代理问题 [Guo et al., 2025e], 或生成领域特定原则, 如创意写作 [Jia et al., 2025] 或科学评论 [Zeng et al., 2025c]. 此外, 评分标准可以双重用作指导策略探索的教学支架和作为最终奖励的标准 [Zhou et al., 2025f].

- **共同进化系统 (统一策略和奖励)**: 最先进的范式超越了静态策略-奖励关系, 转向生成器和验证

器一起改进的动态系统. 这可以通过以下方式实现:

- **自我奖励**, 其中单个模型生成自己的训练信号. 这在自奖励语言模型 [Yuan et al., 2024] 中得到了显著证明, 并在模型在策略和验证器角色之间交替的框架中实现 [Jiang et al., 2025e], 基于自己的批评执行自我纠正 [Team, 2025c, Xiong et al., 2025b, Zhang et al., 2025m], 或通过完成后学习内化奖励函数 [Fei et al., 2025b].
- **共同优化**, 其中策略和单独的奖励模型同时训练. 例如,RL Tango 使用共享的结果级奖励联合训练生成器和过程级 GenRM [Zha et al., 2025]. 类似地,Cooper 共同优化两个模型以增强鲁棒性并减轻奖励黑客攻击 [Hong et al., 2025a]. 其他工作通过统一的强化学习循环将策略 ("玩家") 和奖励 ("裁判") 功能统一在单个模型中 [Lu et al., 2025e].

从静态评判者到动态, 共同进化系统的这种演进通常由结合基于规则和生成式信号的混合奖励方案支持 [Li et al., 2025c, Seed et al., 2025a]. 此外,GenRM 正在被调整以提供更细粒度, 过程级的反馈, 以解决复杂推理链中的信用分配问题 [He et al., 2025f, Khalifa et al., 2025, Xie et al., 2025b, Zhao et al., 2025b]. 本质上, 生成式奖励被证明对于将强化学习扩展到通用 LRM 所针对的全部任务范围是不可或缺的.

3.1.3. 密集奖励

要点

- 密集奖励 (如过程奖励模型) 提供细粒度的信用分配, 并提高强化学习中的训练效率和优化稳定性.
- 对于开放域文本生成等任务, 扩展仍然具有挑战性, 因为难以定义密集奖励或使用验证器.

在诸如游戏和机器人操作任务等经典强化学习中 [Liu et al., 2022, Schrittwieser et al., 2020, Sun et al., 2025d], 密集奖励在 (几乎) 每个决策步骤提供频繁反馈. 这种塑造缩短了信用分配范围, 并通常提高样本效率和优化稳定性, 但如果信号设计不当, 也存在错误指定和奖励黑客攻击的风险 [Hadfield-Menell et al., 2017]. 至于 LLM 推理, 密集奖励通常是监督中间步骤而不仅仅是结果的基于过程的信号, 它们已被证明是有效的, 通常优于基于结果的奖励 [Lightman et al., 2024, Uesato et al., 2022]. 基于 § 2.1 中的定义, 我们根据动作和奖励粒度进一步在 LLM 强化学习的背景下形式化稀疏/结果和密集奖励, 如表 2 所示.

Token **级奖励**. DPO [Rafailov et al., 2023] 及其后续工作 [Rafailov et al., 2024] 表明,token 级奖励可以计算为策略和参考模型之间的对数似然比. 隐式 PRM [Yuan et al., 2025d] 进一步表明,token 级奖励可以通过训练 ORM 并使用 Rafailov et al. [2024] 的参数化来获得.PRIME [Cui et al., 2025a] 将 ORM 学习集成到强化学习训练中, 并使用隐式 token 级奖励训练策略.SRPO [Fei et al., 2025a] 移除了 PRIME 中的 ORM 并改进了优势估计. 另一系列工作专注于使用内部反馈作为 token 级奖励, 如 token 熵 [Cheng et al., 2025a, Tan and Pan, 2025] 和策略性 gram [Wang et al., 2025g].

Table 2 | 语言模型强化学习中动作和奖励粒度的定义 ($z^{(u)}$ 是第 u 轮的环境反馈).

粒度	**动作**	**奖励**	**回报** (G)
轨迹	整个序列 $y = (a_1, \ldots, a_T)$	标量 $R(x, y)$	$R(x, y)$
Token	每个 token $a_t \in \mathcal{V}$	$r_t = R(x, a_{1:t})$	$\sum_{t=1}^{T} \gamma^{t-1} r_t$
步骤	片段 $y^{(k)}$(如句子)	$r_k = R(x, y^{(1:k)})$	$\sum_{k=1}^{K} \gamma^{k-1} r_k$
轮次 (智能体)	每轮智能体响应 $y^{(u)}$	$r_u = R(x, y^{(1:u)}, z^{(1:u)})$	$\sum_{u=1}^{U} \gamma^{u-1} r_u$

步骤级奖励. 步骤级奖励的方法分为两类: 基于模型和基于采样. 早期工作依赖人类专家注释步骤级密集奖励 [Lightman et al., 2024, Uesato et al., 2022], 这成本高昂且难以扩展.

- **基于模型**: 为了降低注释成本,Math-Shepherd [Wang et al., 2024b] 使用蒙特卡洛估计获得步骤级标签, 并证明使用训练过的 PRM 进行过程验证在强化学习中是有效的.PAV [Setlur et al., 2024] 通过优势建模进一步改进过程奖励. 为了缓解基于模型的步骤级奖励的奖励黑客攻击,PURE [Cheng et al., 2025b] 采用最小形式信用分配而不是求和形式, 而 Tango [Zha et al., 2025] 和 AIRL-S [Jin et al., 2025c] 联合训练策略和 PRM. 利用生成式 PRM 的强大验证能力 [Zhao et al., 2025b](在 § 3.1.2 中讨论),ReasonFlux-PRM [Zou et al., 2025],TP-GRPO [He et al., 2025f] 和 CAPO [Xie et al., 2025b] 利用它们为强化学习训练提供步骤级奖励. 然而, 基于模型的密集奖励容易受到奖励黑客攻击, 并且在线训练 PRM 成本高昂.
- **基于采样**: 另一系列工作使用蒙特卡洛采样进行在线过程奖励估计 [Guo et al., 2025c, Hou et al., 2025, Kazemnejad et al., 2025, Li et al., 2025r, Yang et al., 2025g, Zheng et al., 2025c].VinePPO [Kazemnejad et al., 2025] 通过蒙特卡洛估计改进 PPO. 为了改进步骤分割,SPO [Guo et al., 2025c],TreeRL [Hou et al., 2025] 和 FR3E [Zheng et al., 2025c] 使用低概率或高熵 token 作为分割点. 为了提高样本效率和优势估计,SPO [Guo et al., 2025c],TreeRPO [Yang et al., 2025g],TreeRL [Hou et al., 2025] 和 TreePO [Li et al., 2025r] 探索基于树的结构进行细粒度过程奖励计算. MRT [Qu et al., 2025b],S-GRPO [Dai et al., 2025a],VSRM [Yue et al., 2025a] 和 SSPO [Xu et al., 2025f] 强制 LLM 在中间位置终止思考过程以有效估计步骤级奖励. PROF [Ye et al., 2025a] 利用结果奖励和过程奖励之间的一致性来过滤强化学习训练的噪声数据.

轮次级奖励. 轮次级奖励评估每个完整的智能体-环境交互, 如工具调用及其结果, 在多轮任务中以单个轮次的粒度提供反馈. 轮次级奖励的研究大致可分为两条路线: 直接每轮监督和从结果级奖励导出轮次级信号.

- 对于直接每轮监督, 工作在每轮提供显式反馈. 例如, 情感敏感对话策略学习 [Zhu et al., 2024] 利用用户情绪作为每轮奖励指导策略优化, 展示了轮次级反馈如何增强对话智能体的交互质量. 类似地,ToolRL [Qian et al., 2025] 在每个工具调用步骤提供关于格式和正确性的结构化奖励, 为学习提供密集的轮次级信号.Zeng et al. [2025d] 进一步利用可验证信号和显式轮次级优势估计来

改进强化学习期间的多轮工具使用. 此外,SWEET-RL [Zhou et al., 2025g] 学习一个步骤/轮次级评论者, 提供每轮奖励和信用分配, 从而提供显式轮次级监督. 最近,MUA-RL [Zhao et al., 2025d] 将模拟用户交互纳入强化学习循环, 其中每次多轮交换产生每轮反馈, 允许智能体在现实用户-智能体动态下迭代改进其策略. G-RA [Sun et al., 2025g] 通过引入门控奖励聚合扩展了这条工作路线, 其中密集的轮次级奖励 (如动作格式, 工具调用有效性, 工具选择) 只有在满足更高优先级的结果级条件时才会累积.

- 对于从结果级奖励导出轮次级信号, 想法是将基于结果的监督分解或重新分配为更细粒度的单位. 用全局反馈对齐对话智能体 [Lee et al., 2025a] 将会话级分数转换为轮次级伪奖励,GELI [Lee et al., 2024a] 利用韵律和面部表情等多模态线索将会话级反馈细化为局部轮次级信号. 类似地,SPA-RL [Wang et al., 2025e] 通过进度归因将基于结果的奖励重新分配为每步或每轮贡献.ARPO [Dong et al., 2025b] 沿着这条路线, 从轨迹级结果 (如工具使用后) 归因步骤/轮次级优势, 有效地将全局回报转换为局部信号.

总体而言, 轮次级奖励, 无论是在每次交互时直接分配还是从结果分解导出, 都充当基于过程和基于结果监督之间的桥梁, 在稳定和改进多轮智能体强化学习优化中发挥核心作用, 更多细节见 § 6.2.

3.1.4. 无监督奖励

要点

- 无监督奖励消除了人工注释瓶颈, 使奖励信号生成能够达到计算和数据的规模, 而不是人工劳动的规模.
- 主要方法包括从模型自身过程导出信号 (模型特定: 一致性, 内部置信度, 自生成知识) 或从自动化外部来源导出信号 (模型无关: 启发式方法, 数据语料库).

前沿语言模型在广泛任务上表现出色，包括许多异常具有挑战性的任务 [Glazer et al., 2024, Jimenez et al., 2023, Li et al., 2024b, Phan et al., 2025]。然而, 推进这些模型的一个关键限制是依赖人类生成的强化学习奖励信号 (§ 3.1.1–3.1.3). 对于需要超人类专业知识的任务, 人类反馈通常缓慢, 昂贵且不切实际 [Burns et al., 2023]. 为了解决这个问题, 一个有前途的方法是无监督强化学习, 它使用自动生成的, 可验证的奖励信号而不是真实标签. 这种方法是实现 LLM 可扩展强化学习的基础. 本节调查这些无监督奖励机制, 根据其来源将它们分为两类: 源自模型本身的 (模型特定) 和来自外部, 非人类来源的 (模型无关).

模型特定奖励. 这种范式使用 LLM 的内部知识作为监督的唯一来源. 它基于高性能模型将生成一致, 自信或评估上合理输出的假设. 这种方法高度可扩展, 只需要模型和计算资源来生成几乎无限量的 “标记” 数据. 然而, 其闭环性质存在奖励黑客攻击和模型崩溃的风险.

- **来自输出一致性的奖励**: 这种方法假设正确答案将在多个生成的输出中形成密集, 一致的集群. 基础工作如 EMPO [Zhang et al., 2025i] 和测试时强化学习 (TTRL) [Zuo et al., 2025b] 分别通过聚类和多数投票来实现这一点. 后续方法旨在通过提高效率 (ETTRL [Liu et al., 2025c]), 整

合推理轨迹 (CoVo [Zhang et al., 2025h]) 或使用对比一致性来对抗奖励黑客攻击 (Co-Reward [Zhang et al., 2025w]) 来改进这一点.

- **来自内部置信度的奖励**: 另一种方法是直接从模型的内部状态导出奖励, 使用置信度作为正确性的代理. 信号可以基于交叉注意力 (CAGSR [Kiruluta et al., 2025]), 负熵 (EM-RL [Agarwal et al., 2025b],RENT [Prabhudesai et al., 2025]) 或生成概率 (Intuitor [Zhao et al., 2025e],RLSC [Li et al., 2025h],RLSF [van Niekerk et al., 2025]). 这些方法的成功通常依赖于基础模型的初始质量 [Gandhi et al., 2025], 并且可能是脆弱的 [Press et al., 2024, Shumailov et al., 2023], 因为它们依赖于正确和错误路径之间低密度分离等先验 [Chapelle and Zien, 2005, Lee et al., 2013].

- **来自自生成知识的奖励**: 这种范式利用模型的知识创建学习信号, 要么通过充当评判者 (自我奖励), 要么充当问题提出者 (自我指导). 在自我奖励中, 模型评估自己的输出来生成奖励, 这一概念由 Yuan et al. [2024] 和 Wu et al. [2024] 提出, 并在 SSR-Zero [Yang et al., 2025f] 和 MINIMO [Poesia et al., 2024] 等工作中应用. 在自我指导中, 提出者模型为解决者生成课程. 提出者通常因创建最优难度的任务而获得奖励 [Chen et al., 2025i, Huang et al., 2025a, Zhao et al., 2025a], 而解决者的奖励可以是模型无关的 (如来自 AZR [Zhao et al., 2025a] 中的代码执行器) 或模型特定的 (如通过 SQLM [Chen et al., 2025i] 和 SeRL [Fang et al., 2025a] 中的多数投票).

模型无关奖励. 与模型特定方法相比, 这种范式从外部的自动化来源导出奖励. 这种方法将学习过程建立在外部信息基础上, 消除对人工标签的需求. 其核心原则是这些外部信号易于获取且不需要人工努力. 然而, 由于精确反馈通常不可用, 代理奖励的质量至关重要, 奖励黑客攻击的风险仍然存在.

- **启发式奖励**: 这种方法构成了基于规则奖励的另一种形式, 采用基于输出属性 (如长度或格式) 的简单预定义规则作为质量的代理. 它代表了在 § 3.1.1 中讨论的特定情况. 这由 DeepSeek-R1 [Guo et al., 2025a] 开创, 后来通过动态奖励缩放等技术进行改进 [Yu et al., 2025d]. 虽然可扩展, 但这些启发式方法可能被模型操纵, 导致在不推进真实能力的情况下实现表面改进 [Liu et al., 2025t, Xin et al., 2025].

- **以数据为中心的奖励**: 这种方法从大型未标记语料库的结构中导出奖励信号. 类似于大规模预训练的下一个词预测,RPT [Dong et al., 2025c] 将下一个 token 预测重新表述为强化学习任务, 将网络规模数据集转换为数百万个训练示例. 在元层面,SEAL [Zweiger et al., 2025] 允许模型生成自己的训练数据和超参数, 使用下游性能作为奖励.

总之, 无监督奖励设计对于创建 LLM 的可扩展强化学习系统至关重要. 模型特定范式通过利用模型的内部知识促进自我改进, 而模型无关范式将学习建立在外部, 自动反馈基础上. 虽然两种方法都有效绕过了人工注释瓶颈, 但它们仍然容易受到奖励黑客攻击 [Zhang et al., 2025q]. 可扩展强化学习的未来可能涉及战略性地结合这些方法的混合系统, 例如, 使用以数据为中心的奖励进行预训练, 模型特定自我奖励进行复杂推理的微调, 以及最小的人工监督进行安全性和对齐.

3.1.5. 奖励塑造

要点

- 奖励塑造将稀疏信号丰富为 LLM 训练的稳定, 信息丰富的梯度.
- 将验证器与奖励模型结合, 使用组基线加上与 pass@k 对齐的目标来稳定训练, 扩展探索并在大规模上匹配评估指标.

如前所述, 强化学习中智能体的主要学习目标是最大化累积奖励, 这使得奖励函数的设计特别关键 [Sutton et al., 1998]. 在前面的章节中, 我们介绍了各种奖励函数, 如可验证奖励 (§ 3.1.1), 生成式奖励 (§ 3.1.2), 密集奖励 (§ 3.1.3) 甚至无监督奖励 (§ 3.1.4). 除了奖励工程外, 同样重要的是考虑如何修改或增强奖励函数以鼓励推动向期望解决方案进展的行为. 这个过程称为奖励塑造 [Goyal et al., 2019, Gupta et al., 2022, Hu et al., 2020, Xie et al., 2023], 可以分为基于规则和基于结构的奖励塑造.

基于规则的奖励塑造. 在基于 LLM 的强化学习中, 奖励塑造的最简单和最常用方法涉及结合基于规则的验证器和奖励模型的奖励来生成整体奖励信号, 如 Qwen2.5 Math [Yang et al., 2024a] 中所示. 通常使用常数系数来平衡奖励模型和基于规则组件的贡献. 这种方法不是为所有正确响应分配相同奖励, 而是允许根据奖励模型的分数进一步对响应进行排名. 这种方法对于更具挑战性的样本特别有用, 并有助于避免所有奖励值为 0 或 1 的情况, 否则会导致学习梯度无效 [Yu et al., 2025d]. 这种启发式组合策略在开放域任务中被广泛采用, 其中集成基于规则的奖励和奖励模型 [Guo et al., 2025b, Liao et al., 2025a, Liu et al., 2025x] 为 LLM 的强化学习产生更多信息性和有效的奖励信号 [Su et al., 2025c, Zeng et al., 2025c, Zhang et al., 2024a]. 另一种方法涉及结合基于规则的奖励, 如结果级奖励和格式奖励, 如 DeepSeek-R1 [Guo et al., 2025a] 中实施的, 这使 LLM 能够学习长思维链推理. 这些奖励包括基于格式的 [Xin et al., 2025] 和基于长度的组件 [Liu et al., 2025p], 以解决 LLM 输出中的各种异常情况. 与使用固定奖励权重 [Team, 2025d, Yao et al., 2025b] 或奖励插值的启发式规则 [Aggarwal and Welleck, 2025, Zhang and Zuo, 2025] 相比,Lu et al. [2025f] 提出动态奖励加权, 采用超体积引导的权重适应和基于梯度的权重优化. 这种方法在多目标对齐任务上实现卓越性能 [Li et al., 2025a, Liu and Vicente, 2024]. 最近的工作还探索多角色强化学习训练, 并为具有不同奖励函数的不同角色分配不同奖励, 如解决者和评论者 [Li et al., 2025i]. 通常, 这些奖励使用手动设置的常数结合. 最近的工作还探索了多角色强化学习训练 [Li et al., 2025i,j], 为不同角色分配不同的奖励函数以鼓励多样化的行为和目标 [Li et al., 2025i], 如解决者和评论者.

基于结构的奖励塑造. 与仅依赖单个样本的基于规则的奖励塑造相比, 基于结构的奖励塑造通过利用列表级或集合级基线在一组候选中计算奖励. 一个有影响力的方法是 GRPO [Shao et al., 2024], 它使用对同一问题 G 的响应组均值作为基线 (或诸如留一法 [Ahmadian et al., 2024] 或排名等变体), 并相应地为 PPO 风格更新构建优势 [Schulman et al., 2017b]. 最近的工作进一步修改了优化目标或信用分配策略, 以促进更强的探索并实现与评估指标 (如 pass@k) 的更紧密对齐 [Yue et al., 2025b]. 例如,Walder and Karkhanis [2025] 对最终奖励进行联合变换, 使优化直接等同于像 pass@k 这样的集合级目标, 并提供低方差, 无偏的梯度估计.Chen et al. [2025x] 在推导和分析优势及高效近似时直接针对 pass@k, 将集合级目标分解回单个样本信用分配. 这一方向的奖励塑造方法旨在稳定训练并

鼓励策略进行更广泛的探索, 从而降低过早收敛到次优局部解决方案的风险.

3.2. 策略优化

在本小节中, 我们首先提供策略梯度目标数学公式的技术概述 (§ 3.2.1). 接下来, 我们根据梯度计算过程中奖励生成方式的不同, 将 RL 中的同策略优化算法分为两类: 基于 Critic 的算法 (§ 3.2.2) 和无 Critic 的算法 (§ 3.2.3). 此外, 我们讨论了最近将同策略 RL 与离线数据集结合的研究, 以实现更复杂的后训练 (即异策略) 优化 (§ 3.2.4), 以及各种正则化技术, 如熵和 KL 正则化 (§ 3.2.5).

3.2.1. 策略梯度目标

如 § 2.1 中所介绍的,LLM 中 RL 的上下文被视为环境, 下一级预测的概率分布被视为策略. 对于一个 RL 系统, 系统的目标是找到一个最优策略, 使得系统产生的期望累积奖励最大化. LLM 的 RL 策略优化算法大多是一阶基于梯度的算法, 这是由于 LLM 中参数数量庞大. 通常,RL 算法寻求优化网络参数, 使得期望奖励最大化. 下面, 我们给出 LLM 的 RL 算法梯度计算的一般公式.

符号说明. 尽管我们在 § 2.1 中已经介绍了相关符号, 但为了比较清晰起见, 我们在此重新审视这些定义. 令 $x \sim \mathcal{D}$ 为提示 (初始状态 $s_1 = s$). 随机策略 π_θ 生成序列 $y = (a_1, \cdots, a_T)$, 我们将 y 的总序列长度表示为 $|y|$, 其中状态由 $s_{t+1} = (x, s_{\leq t})$ 定义. 我们假设主要是序列级奖励 $R(x, y)$, 可选地分解为 token 级奖励 r_t. 我们使用*行为策略* π_b(也表示为 π_{old}, 指代当前策略的早期版本) 为每个提示收集 $G \geq 1$ 个响应. 可选地, 可以使用*参考策略* π_{ref}(例如, 基础, 微调或指令模型) 进行正则化.

我们重新审视在 § 2.1 中定义的 MDP. 在 MDP 中, 我们将给定当前状态 s 的期望累积奖励表示为 V(值) 函数

$$V(s) = \mathbb{E}_{a_t \sim \pi_\theta(s_t), s_{t+1} \sim \mathcal{P}(s,a)} [\sum_{t=0}^{T} \gamma^t r(s_t, a_t) | s_0 = s], \tag{2}$$

当前状态-动作对的期望累积奖励表示为 Q(质量) 函数

$$Q(s, a) = \mathbb{E}_{a_t \sim \pi_\theta(s_t), s_{t+1} \sim \mathcal{P}(s,a)} [\sum_{t=0}^{T} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]. \tag{3}$$

那么 RL 的目标可以表述为期望累积奖励的最大化问题. 为了优化目标函数, 通常使用策略梯度算法 [Sutton et al., 1999, Williams, 1992] 进行梯度估计:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta} \left[\sum_{t=1}^{T} \nabla_\theta \pi_\theta(y_t | y_{<t}) Q_t \right]. \tag{4}$$

策略梯度可以基于这样的直觉来理解: 遵循策略梯度的算法应该提高高于平均水平动作的概率, 降低低于平均水平动作的概率. 这一概念导致了 A(优势) 函数 $A(s, a) = Q(s, a) - V(s)$ 的引入. 优势衡量当前动作相比现有策略在期望总奖励上的改进程度. 优势可以通过多种方式估计. 如果我们只有完整轨迹的奖励, 原始的 REINFORCE 算法 [Williams, 1992] 直接定义 $A_t = R(x, y)$.

对于训练 LLM 的情况, 原始策略梯度算法经常面临稳定性问题. 相反, 训练通常使用 PPO 算法 [Schulman et al., 2017b] 进行. 对于具有 N 个样本的算法, 我们定义具有 PPO 风格更新的一般目标如下:

$$\mathcal{J}(\theta) = \mathbb{E}_{\text{data}} \left[\frac{1}{Z} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \min \left(w_{i,t}(\theta) \hat{A}_{i,t},\ \text{clip}(w_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}) \hat{A}_{i,t} \right) \right], \tag{5}$$

其中:

- $w_{i,t}(\theta)$ 是重要性比率;
- $\hat{A}_{i,t}$ 是优势 (token 级或序列级);
- T_i 是每个样本的 token 数或响应数;
- N 是给定提示下的总样本数;
- Z 是归一化因子 (例如, 总 token 数, 组大小等).

PPO 算法 [Schulman et al., 2017b] 最初被提出作为 TRPO 算法 [Schulman et al., 2015a] 的计算高效近似. 当原始策略梯度方法面临数据效率和鲁棒性问题时,PPO 表现出色. 此外, 与 TRPO 相比,PPO 被证明实现更简单, 更通用, 并且具有更好的样本复杂度.

然而, 由于 LLM 的复杂和长 CoT 特性, 精确的目标函数, 梯度估计和更新技术可以采取各种不同的形式, 如表 3 所示.

3.2.2. 基于 Critic 的算法

要点

- Critic 模型在标记数据的小子集上训练, 为无标记的 roll-out 数据提供可扩展的 token 级值信号.
- Critic 需要与 LLM 并行运行和更新, 导致显著的计算开销, 并且在复杂任务中扩展性不佳.

第一个与 LLM 相关的 RL 工作关注如何有效地将 LLM 策略与外部监督对齐, 使 LLM 具有更好的指令跟随能力, 同时确保模型是有帮助, 诚实和无害的. LLM 对齐的最常见方法是 RLHF [Bai et al., 2022a, Christiano et al., 2017, Ouyang et al., 2022, Stiennon et al., 2020]. 该技术利用人类作为学习算法的 critic; 具体步骤如下. 首先,LLM 生成一系列模型输出, 由人类标记以创建数据集. 然后使用该数据集训练奖励模型, 预测人类更偏好哪个响应. 最后, 奖励模型与值函数一起用于训练 LLM, 充当系统中的 critic. 训练通常使用 PPO 算法 [Schulman et al., 2017b] 进行.PPO 算法将目标表述为以下形式:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \min \left(w_t(\theta) \hat{A}_t,\ \text{clip}(w_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t \right) \right], \tag{6}$$

Table 3 | 推理模型训练中代表性 RL 算法的比较。

日期	算法	优势估计	重要性采样	损失聚合
2017.01	PPO	Critic-GAE	PPO-Style	Token-Level
2023.10	ReMax	Greedy Baseline	N/A	Token-Level
2024.02	RLOO	Leave-One-Out	N/A	Token-Level
2025.01	RF++	Negative KL + Batch Relative	PPO-Style	Sequence-level
2024.02	GRPO	Group Relative	PPO-Style	Sequence-level
2025.01	PRIME	Outcome + Implicit PRM	PPO-Style	Token-Level
2025.03	VAPO	Value Adjusted GAE	Clip-Higher	Token-Level
2025.03	Dr. GRPO	Group Baseline	PPO-Style	Token-Level
2025.04	DAPO	Group Relative	Clip-Higher	Token-Level
2025.05	Clip-Cov	Group Relative	PPO-Style	Sequence-level
2025.05	KL-Cov	Group Relative	PPO-Style	Sequence-level
2025.06	CISPO	Group Relative	Clipped IS-weight	Token-Level
2025.07	GSPO	Group Relative	PPO-Style	Sequence-level
2025.08	GMPO	Group Relative	Clip-Wider	Geometric-Avg
2025.08	GFPO	Filter + Group Relative	PPO-Style	Token-level
2025.08	LitePPO	Group-level mean, Batch-level std	PPO-Style	Token-level
2025.08	FlashRL	Group Relative	Truncated IS	Token-level
2025.09	GEPO	Group-level mean	Group Expectation	PPO-Style
2025.09	SPO	Entire Batch-level	PPO-Style	Sequence-level

其中 $\hat{A}_t$ 是基于值模型的优势, 且

$$w_t(\theta) = \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\theta_{\text{old}}}(y_t|x, y_{<t})}. \tag{7}$$

我们注意到 PPO 被提出作为 TRPO 的裁剪替代目标, 保留了 TRPO 的保守策略迭代, 同时不受约束, 并且具有接近传统策略梯度方法的计算复杂度. 由于当前策略和采样分布之间的差异,TRPO 中的优势乘以 w_t, 即公式 6 中的重要性采样因子.PPO 最大化与 TRPO 相同的目标, 但移除了信任区域约束. 此外,PPO 添加了裁剪机制和 KL 正则化因子, 确保当前策略不会与 roll-out 策略 $pi_{\theta_{old}}$ 偏离太远.

在基于 Critic 的方法中,RL 的可扩展性通过引入 critic 模型来实现. 在奖励模型在手标记的生成数据小子集上充分训练后, 可以用于构建 critic 模型, 为绝大多数无标记的生成数据大规模生成 token 级值信号以进行 RL. 然而, 这些工作需要一个 critic 模型与目标 LLM 并行运行和优化, 并产生显著的计算开销.

在 PPO 中,critic 模型采用 RL 文献中的广义优势估计 (GAE)[Schulman et al., 2015b]. GAE 通

常使用时间差分误差构建

$$\delta_t = r_t + \gamma V(y_{t+1}) - V(y_t), \tag{8}$$

然后跨时间步累积:

$$\hat{A}_{GAE,t} = \sum_{l=t}^{T} (\gamma\lambda)^l \delta_{t+l}, \tag{9}$$

其中 γ 是 MDP 的折扣因子,λ 是控制偏差-方差权衡的参数.

最近的工作认为, 对于需要长 CoT 的复杂推理任务, 衰减因子扩展性不佳, 并提出了 Value-Calibrated PPO [Yuan et al., 2025f] 和 VAPO [Yue et al., 2025c],VRPO [Zhu et al., 2025a] 提出了在噪声奖励信号下增强 critic 模型鲁棒性的新机制.

此外, 基于 Critic 的算法 [Hu et al., 2025b] 也在基于规则的奖励蒙特卡洛估计中展示了稳定的可扩展性. 类似的方法通过 PRM 的实现, 已适应于固定的外部模型 [Lu et al., 2024, Wang et al., 2024b].

引入 critic 模型的另一种方法是通过隐式 PRM [Yuan et al., 2025d]. 该方法也能够为可扩展的 RL 训练提供 token 级监督. 与 GAE 方法不同, 如隐式 PRM [Yuan et al., 2025d] 和 PRIME [Cui et al., 2025a] 等方法采用特定的奖励模型表述直接生成 token 级奖励.

3.2.3. 无 Critic 算法

要点

- 无 Critic 算法只需要序列级奖励进行训练, 使其更充分和可扩展.
- 对于 RLVR 任务, 基于规则的训练信号可靠地防止了与 critic 相关的问题, 如奖励黑客攻击.

除了为模型训练提供 token 级反馈信号的基于 Critic 模型外, 许多最近的工作指出, 响应级奖励对于使用 RL 的可扩展推理任务是足够的. 这些无 Critic 算法将相同的基于规则或模型生成的响应级奖励应用于响应中的所有 token, 并在各种任务中展示了其有效性. 与基于 Critic 的算法相比, 无 Critic 方法不需要单独的 critic 模型, 显著降低了计算需求并简化了训练. 此外, 在基于规则的环境中训练 LLM 时, 任何响应的奖励都可以明确定义, 无 Critic 算法可以避免由于训练不当的 critic 模型可能产生的奖励黑客攻击问题. 这种特性使得无 Critic 算法在此类设置中比基于 Critic 的方法更具可扩展性.

经典的 REINFORCE [Williams, 1992] 算法是最早为 RL 开发的算法之一. 它被应用于 [Ahmadian et al., 2024] 中的 LLM 问题.REINFORCE 的确切公式如下:

$$\mathcal{J}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y\}\sim\pi_{\text{old}}(\cdot|x)} \left[R(x,y) \nabla_\theta \log(\pi_\theta(y|x)) \right], \tag{10}$$

其中对于 RLVR 任务,$R(x,y)$ 通常采用 ±1 的形式. 这种朴素公式将整个序列视为单个动作, 并将响应任务视为多臂老虎机问题. 然而, 原始算法通常由于高方差而遭受严重的稳定性问题.ReMax [Li et al., 2023c] 引入了方差减少机制, 使用贪心基线估计. Ahmadian et al. [2024] 也引入了 RLOO, 进

一步提供了无偏基线, 结果更稳定.REINFORCE++ [Hu, 2025] 采用了 PPO 和 GRPO 风格算法中的裁剪和全局优势归一化等技术, 提供更准确的优势和梯度估计.

RL 中最流行的无 Critic 方法之一是 GRPO [Shao et al., 2024].GRPO 的目标公式如下:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(w_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}(w_{i,t}(\theta),1-\epsilon,1+\epsilon)\hat{A}_{i,t}\right)\right], \quad (11)$$

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x,y_{i,<t})}, \quad \hat{A}_{i,t} = \hat{A}_i = \frac{R(x,y_i) - \text{mean}(\{R(x,y_i)\}_{i=1}^{G})}{\text{std}(\{R(x,y_i)\}_{i=1}^{G})}, \quad (12)$$

其中 y_i 中的所有 token 共享相同的优势 $\hat{A}_i$.

GRPO 是 PPO 的无 Critic 修改, 其中不是由 critic 提供 GAE, 整个序列使用相同优势估计, 该估计通过组相对归一化计算, 作为比二元基于规则奖励更好的估计. 与 PPO 和 REINFORCE 风格方法相比,GRPO 的基于组优势计算有效减少了训练信号的方差, 并已被证明可以加速训练过程. 其他最近的方法, 包括 DAPO [Yu et al., 2025d],CISPO [Chen et al., 2025a],Dr. GRPO [Liu et al., 2025u],LitePPO [Liu et al., 2025w], 通过仔细调整采样策略, 裁剪阈值和损失归一化对 GRPO 进行了进一步修改, 以进一步增强 RL 训练过程的稳定性. 另一个最近的方法,GSPO [Zheng et al., 2025a], 用序列级裁剪替换了逐 token 裁剪的重要性采样比率.

除了 REINFORCE 和 GRPO 相关算法外, 还有其他无 Critic 方法. VinePPO 通过用蒙特卡洛优势估计替换学习的 critic 来修改 PPO. CPGD [Liu et al., 2025z] 提出了一个新的策略梯度目标, 以及漂移正则化机制. K1.5 [Team, 2025d] 在基础模型训练中利用 RL 与镜像下降的适应, 成功增强了 LLM 的长上下文推理能力. Lv et al. [2025] 最近引入了一个统一的策略梯度估计器和混合后训练算法, 为 LLM 中 RL 的策略梯度估计提供了统一框架. SPO [Xu and Ding, 2025] 引入了无组, 单流策略优化, 用持续的 KL 自适应值跟踪器和全局优势归一化替换每组基线, 比 GRPO 产生更平滑的收敛和更高精度, 同时在长视界和工具集成设置中高效扩展. HeteroRL [Zhang et al., 2025c] 将 roll-out 采样与参数学习解耦以实现分散异步训练, 并通过 GEPO 减少由延迟引起的 KL 漂移 (理论上指数级) 下的重要性权重方差, 即使在严重延迟下也能保持稳定性 (例如, 在 1,800 秒时<3%的退化).

策略优化的重要性采样. 由于 RL 的 roll-out-奖励-训练循环, 确保 roll-out 数据遵循当前模型的精确策略分布通常是计算上不可行的. 因此, 引入了重要性采样来减少训练中的偏差.RL 中的第一个重要性采样版本在 TRPO 中引入, 其中将逐 token 重要性比率 $w_{i,t}$ 引入目标函数. 这种方法在最近的工作中被广泛采用, 如 GRPO. 这种方法仅限于逐 token 重要性比率, 因为实际分布比率无法在 CoT 的长上下文中有效计算. 然而,token 级重要性采样给 RL 算法引入了另一个偏差, 因为给定策略的实际采样分布是相对于状态-动作对定义的, 而 token 级方法只考虑当前动作. GMPO [Zhao et al., 2025f] 寻求通过引入几何平均来缓解, 增加具有极端重要性采样比率的 token 的训练鲁棒性. 在 GSPO [Zheng et al., 2025a] 的最近工作中, 计算了序列级重要性采样因子.GSPO 添加了独特的归一化因子以确保可以计算概率比率, 但这种方法也是实际重要性采样因子的有偏估计. 一个有希望的新方向是超越标准同策略策略梯度方法的理论框架, 而是直接从监督学习理论推导本质上异策略的算法 [Chen et al., 2025c]. 我们将在下一节中详细介绍异策略优化.

3.2.4. 异策略优化

要点

- 异策略 RL 通过将数据收集与策略学习解耦来提升样本效率, 使得能够从历史, 异步或离线数据集进行训练.
- 现代实践混合了异策略, 离线和同策略方法 (例如,SFT+RL 或大规模离线学习) 以提高稳定性和性能.

在 RL 中, 异策略方法处理正在学习的策略 (目标策略) 与生成数据的策略 (行为策略) 不同的情况. 这种核心区别使得智能体能够在不遵循最优行动过程的情况下了解它. 这种灵活性是一个关键优势, 通常导致比同策略对应方法更高的样本效率, 后者需要直接从当前策略采样新数据进行每次更新. 这些方法中的一个核心挑战是纠正行为策略和目标策略之间的分布偏移, 通常使用带权重目标函数的重要性采样来解决:

$$\mathcal{L}_{\text{policy}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_{\text{b}}(y|x)}\left[\frac{\pi_\theta(y|x)}{\pi_{\text{b}}(y|x)}\cdot r(x,y)\right], \tag{13}$$

其中分数 $\frac{\pi_\theta(y|x)}{\pi_b(y|x)}$ 作为目标策略 π_θ 和行为策略 π_{b} 之间的重要性权重.

在实际的大规模模型训练中, 异策略学习通常以不同形式表现. 最近的工作可以大致分为三个方面:1) 训练-推理精度差异, 其中模型以高精度训练但以低精度部署, 在目标策略和行为策略之间造成差距;2) 异步经验回放机制, 通过在学习过程中重用过去轨迹来提高效率和稳定性;3) 更广泛的异策略优化方法, 包括优化器级改进, 数据级离线学习, 以及将监督微调与 RL 结合的混合方法.

训练-推理精度差异. 一个值得注意的异策略情况源于训练模型和推理模型之间的参数精度差异, 为训练和推理采用不同的框架 [Yao et al., 2025a](例如,vLLM vs. FSDP), 或通过模型量化加速推理 [Lin et al., 2016], 这些是 LLM 推理中不确定性的表现 [He and Lab, 2025]. 通常做法是使用高精度参数 (例如,32 位浮点) 训练模型, 然后部署具有低精度参数 (例如,8 位整数) 的量化版本 [Liu et al., 2025i]. 这造成了一种差异, 其中部署的低精度模型充当行为策略, 生成真实世界交互数据, 而高精度模型在训练期间仍然是被更新的目标策略. 虽然这种不匹配建立了异策略学习问题, 但研究表明, 由于量化引起的策略偏离通常很小. 因此, 这种差异可以通过简单的校正技术有效管理, 如截断重要性采样 (TIS) [Ionides, 2008, Yao et al., 2025a], 允许在保留加速推理的好处的同时进行稳定训练.

异步异策略训练. 异步训练自然与 LLM 的异策略 RL 配对. 多个执行器并发生成轨迹并将其附加到共享回放缓冲区, 而中心学习器从此缓冲区采样小批量来更新目标策略. 基于这种观点, 几种最近的方法有意重用过去轨迹以提高效率和稳定性. 一个例子是回溯回放 [Dou et al., 2025], 通过选择性地重放早期推理轨迹来增强 LLM 推理探索, 指导当前策略更新. 类似地,EFRame [Wang et al., 2025b] 采用探索-过滤-回放机制, 将过滤的响应与新鲜 rollout 交错以鼓励更深层次推理. 在代码生成领域, 可能性和通过率优先经验回放 (PPER)[Chen et al., 2024c] 进一步通过优先考虑缓冲区中的高价值代码样本来实现, 导致更稳定的优化. 将这些想法扩展到多模态交互,ARPO [Lu et al., 2025b] 将回放应用于 GUI 智能体, 其中重用成功轨迹在稀疏奖励下提供可靠的学习信号. 最后,RLEP [Zhang et al., 2025d] 用来自早期运行的验证成功轨迹的经验缓冲区锚定探索, 这些轨迹与新 rollout 混合以平衡可

靠性和发现性. 总之, 这些方法说明了回放缓冲区如何成为基于 LLM 的智能体的现代异步异策略训练的基石.

异策略优化. 微调 LLM 的最新进展探索了超越传统同策略 RL 的复杂优化策略. 这些方法, 广义上归类为异策略和混合策略优化, 旨在通过创造性地使用各种来源的数据来提高样本效率, 训练稳定性和整体性能. 我们在下面介绍这个主题:

- **优化器级异策略方法**: 这些方法专注于改进优化过程本身, 强调策略更新的稳定性和效率. 例如,SPO [Cohen et al., 2025] 引入了软策略优化方法, 实现稳定的在线异策略 RL, 而 TOPR [Roux et al., 2025] 提出了锥形异策略 REINFORCE 算法, 以提高稳定性和效率.ReMix [Liang et al., 2025a] 通过专注于高效利用异策略数据以最大化可用信息的效用进一步强调了这一点.
- **数据级异策略方法**: 一类异策略算法完全从大规模外部离线数据学习 [Zhang et al., 2025g]. 例如, 动态微调 (DFT) 框架 [Wu et al., 2025i] 将 SFT 损失推广到 RL 公式并引入停止梯度机制, 能够像 SFT 一样在离线数据上训练, 同时产生改进的性能. 同样基于离线数据, 直觉微调 (IFT) [Hua et al., 2024] 添加了时间残差连接, 融合 SFT 和 RLHF 目标, 并显式建模和优化当前 token 对所有未来生成的影响. 另一个相关方法是直接偏好优化 (DPO) [Rafailov et al., 2023], 它直接从偏好数据优化策略. 这些方法共同代表了向 RL 中更以数据为中心方法的转变, 能够从大量和多样的离线数据源开发复杂的策略.
- **混合策略方法**: 与更有效地重用过去数据并行, 混合策略优化代表了另一个重要趋势, 它结合了 SFT 和 RL 的优势. 这种混合方法利用 SFT 在专家数据上的稳定性, 同时使用 RL 来优化特定的奖励函数, 通过两种主要方式集成监督数据. 一种策略是在损失级别, 其中 SFT 和 RL 目标直接在损失函数中结合 [Lv et al., 2025, Xiao et al., 2025b, Zhang et al., 2025k, ?]. 如 UFT [Liu et al., 2025k],SRFT [Fu et al., 2025c],LUFFY [Yan et al., 2025a],RED [Guan et al., 2025] 和 ReLIFT [Ma et al., 2025a] 等方法都通过创建统一或单阶段训练过程来体现这一点, 这些过程同时从专家演示和 RL 反馈中学习. 第二种策略在数据级别运行, 使用专家数据构建生成过程本身. 这里, 高质量数据作为前缀或锚点来指导模型的探索 [Guo et al., 2025d]. 例如,BREAD [Zhang et al., 2025p] 从专家锚点生成分支 rollout,Prefix-RFT [Huang et al., 2025g] 通过前缀采样混合训练制度. 通过在损失或数据级别混合策略, 这些方法防止奖励黑客攻击, 并确保模型保留来自 SFT 的知识, 导致更强大和更有能力的复杂推理模型.

3.2.5. 正则化目标

要点

- 目标特定的正则化有助于平衡探索和利用, 提升 RL 效率和策略性能.
- KL, 熵和长度正则化的最优选择和形式仍然是开放性问题, 每种都影响策略优化和可扩展性.

如前几节所介绍的, 确保稳定性和防止灾难性策略漂移至关重要. 特别是, 对于长视界训练,KL 正则化和熵正则化等技术被广泛采用.

KL **正则化**. KL 散度正则化的作用是该领域高度争议的话题. 在大多数研究中,KL 正则化应用于 1). 当前策略 π_θ 和参考策略 π_{ref},2). 当前策略 π_θ 和旧策略 π_{old}. 我们在公式 14 中提供统一表述.

$$\mathcal{L}_{\text{KL}} = \beta \sum_{t=1}^{|y|} KL(\pi_\theta(\cdot|y_t)||\pi_{\text{ref/old}}(\cdot|y_t)). \tag{14}$$

- 对于前者, 这是 RLHF [Ouyang et al., 2022, Touvron et al., 2023] 中常用的技术. 最初引入是为了防止模型被破坏性更新. 先前的工作认为, 加入 KL 惩罚对于在数千个训练步骤中保持稳定性和避免熵坍塌至关重要. 为了降低 KL 项过度约束进展的风险,Liu et al. [2025j] 使用这种方法结合周期性参考策略重置, 其中参考模型更新为训练策略的最新快照. 为了同时保持知识和增强推理能力,Wang et al. [2025i] 对低熵 token 应用更强的 KL 正则化, 对高熵 token 应用较弱的正则化. 然而, 在 LLM 推理的 RL 背景下, 这比标准 RLHF 更具挑战性, 这种 KL 正则化的必要性需要重新考虑. 最近, 许多研究已确定策略应该在训练期间自由探索, 因此可能与其初始化显著偏离以发现新的 CoT 结构, 使 KL 约束成为不必要的限制. 因此, 大多数其他最近的工作主张完全移除 KL 惩罚 [An et al., 2025, Arora and Zanette, 2025, Chen et al., 2025q, Cui et al., 2025a, Fan et al., 2025b, He et al., 2025d, Liao et al., 2025b, Liu et al., 2025u, Yan et al., 2025a, Yu et al., 2025d] 以简化实现, 降低内存成本并实现更可扩展的 GRPO.
- 对于后者情况, 它可以作为策略损失的裁剪形式的替代 [Schulman et al., 2017b]. Zhang et al. [2025r] 讨论了前向 KL, 反向 KL, 归一化 KL 和归一化形式之间的差异. 这种方法也已在 Cui et al. [2025b], Lyu et al. [2025], Team [2025d] 中采用, 展示了其在不同 RL 训练规模下的潜力. 尽管如此, 其更深层次的机制及其对可扩展 RL 的意义仍在探索中.

熵正则化. 在 RL 文献中, 保持策略熵被广泛认为是许多算法的关键方面 [Eysenbach and Levine, 2021, Williams, 1992, Williams and Peng, 1991]. 为此, 策略熵通过正则化技术主动控制 [Haarnoja et al., 2018, Schulman et al., 2017b, Ziebart et al., 2008].

$$\mathcal{L}_{\text{ent}} = -\alpha \sum_{t=1}^{|y|} H[\pi_\theta(\cdot|y_t)] = \alpha \sum_{t=1}^{|y|} \sum_{v=1}^{|\mathcal{V}|} \pi_\theta(y_t^v|y_t) \log \pi_\theta(y_t^v|y_t). \tag{15}$$

然而, 在 LLM 的 RL 中, 直接应用熵正则化既不常见也不有效 [Cui et al., 2025b, He et al., 2025d]. 在损失函数中使用显式熵正则化项仍然是一个争议点. 虽然一些人发现它有益, 使用标准系数 [Shrivastava et al., 2025] 或目标损失函数 [Wu et al., 2025e], 但其他人反对它, 发现它可能导致不稳定性甚至训练崩溃, 特别是在稀疏奖励情况下 [An et al., 2025, Liao et al., 2025b]. 许多研究显示了在未应用干预时熵坍塌的现象 [Cheng et al., 2025a, Cui et al., 2025b, Yu et al., 2025d], 这阻碍了训练期间有效的策略探索. 为了解决这个问题,He et al. [2025d] 动态调整熵损失的系数,Yu et al. [2025d] 采用 clip-higher 技术让更多低概率 token 参与策略更新,Wang et al. [2025m] 直接在 20% 高熵 token 上训练,Cheng et al. [2025a] 和 Chen et al. [2025j] 通过将其纳入优势计算来强调熵. 除了这

些显式最大化熵的技术外,Cui et al. [2025b] 提供了熵动态底层机制的理论解释, 将动作输出概率与其优势之间的协方差识别为熵 "驱动器". 基于这一洞察, 提出了 Clip-Cov 和 KL-Cov, 通过选择性地约束表现出异常高协方差的一小部分 token 来调节熵.

长度惩罚. LRM 在复杂任务上的最近成功验证了长 CoT 推理的有效性. 然而更长的推理轨迹产生更高的推理成本. 为了平衡推理预算和性能 [Agarwal et al., 2025a, He et al., 2025e], 许多工作寻求在保持模型性能的同时降低推理成本 [Aggarwal and Welleck, 2025, Liu et al., 2025p, Luo et al., 2025a, Su et al., 2025b, Xiang et al., 2025]. 例如,Aggarwal and Welleck [2025] 通过确保遵循用户指定的长度约束来控制推理长度, 而 Yuan et al. [2025a] 和 Luo et al. [2025a] 设计了相对长度正则化和准确性保持约束到优化目标,Xiang et al. [2025] 和 Liu et al. [2025p] 提出应用基于问题难度的自适应长度惩罚以保持模型能力.

3.3. 采样策略

与静态数据集不同,RL 依赖于主动管理的 rollout, 其中关于采样什么以及如何采样的决策直接影响学习效率, 稳定性和获得的推理行为质量. 有效的采样策略不仅确保多样化和信息丰富的训练信号, 而且使学习过程与预期的奖励结构和策略目标保持一致. 在本小节中, 我们综述了动态和结构化采样的最新进展 (§ 3.3.1), 以及进一步优化采样和策略改进的超参数调整技术 (§ 3.3.2).

3.3.1. 动态和结构化采样

要点

- 高质量, 多样化的 rollout 通过使智能体暴露于更广泛的有意义经验, 稳定 RL 训练并提升整体性能.
- 平衡多样化轨迹的探索与保持高采样效率是 RL 中的一个基本权衡.

采样已成为推理 LLM 的 RL 微调中的一级杠杆, 作为高效和自适应的机制来最大化数据利用率, 减少浪费计算, 提升训练效果, 或作为 LLM 以结构化格式采样的控制和指导.

动态采样. 动态采样基于在线学习信号 (如成功率, 优势, 不确定性或估计难度) 来调整 rollout 提示的选择以及分配给每个提示的计算预算. 主要目标是将计算集中在信息丰富的示例上, 同时避免饱和或无产的示例. 现有方法通常分为两类:

- **效率导向采样**: 一些工作使用在线过滤来将训练集中在中等难度的问题上, 以确保训练的有效性和效率. 代表性的设计是 PRIME [Cui et al., 2025a], 它应用在线过滤器来剔除过易或过难的问题. 另一个例子是 DAPO [Yu et al., 2025d], 它过采样和过滤 rollout 为饱和 (全正确) 或退化 (全错误) 的提示, 然后重复采样直到每个小批量包含非零优势的提示, 专注于中等难度情况以保持信息丰富的梯度. 在此基础上, 优先级方案通过按失败率比例采样来将 rollout 预算分配给未掌握的项目, 如 $p(i) \propto (1 - s_i)$ 规则 [Team, 2025d]. 课程学习方法在多个尺度上运作: 类别级选择 [Chen et al., 2025o] 使用非稳态 bandits, 而 E2H [Parashar et al., 2025] 遵循从易到难的调度, 对小模型

提供收敛保证. 效率方法包括 rollout 前选择以跳过无帮助的提示, 以及基于难度的在线选择与 rollout 回放 [Sun et al., 2025e, Zheng et al., 2025b].POLARIS [An et al., 2025] 通过离线难度估计将其形式化, 通过模型规模构建 “镜像-J” 分布, 持续移除已掌握的项目, 并应用批量内信息替换. 扩展这些效率提升, 最新进展使用轻量级控制器进行自适应采样 [Do et al., 2025, Shi et al., 2025b] 而不修改算法, 而带随机重排的经验回放 [Fujita, 2025] 通过平衡利用减少方差, 增强的优先级方法 [Li et al., 2024a] 基于经验池特征动态调整优先级权重. 采样效率也可以通过用专家数据结构化生成过程来改善: 高质量演示被用作前缀锚点, 将探索偏向搜索空间的有希望区域 [Guo et al., 2025d, Huang et al., 2025g, Zhang et al., 2025p]. 该领域从均匀采样转向模型感知策略, 结合项目级, 类别级和难度级选择, 以获得每次 rollout 更强的学习信号.

- **探索导向采样**: 还有其他工作旨在使用动态 rollout 进行探索.ARPO [Dong et al., 2025b] 被提出用于实现熵引导 rollout 以确保高不确定性, 从而使模型调用外部工具, 提高多样性.DARS [Yang et al., 2025h] 提出 rollout 机制来为不同难度的问题动态分配采样数量.Zhou et al. [2025f] 通过在 rollout 期间为策略提供不同评分标准来提出 RuscaRL 以增强探索. 与上述不同,G^2RPO-A [Guo et al., 2025d] 不丢弃全错问题, 而是在思考过程中添加指导, 为难题生成正确样本. 此外,Li et al. [2025t] 利用最新的 k 个检查点生成 k 个响应, 以防止训练过程中的遗忘.

结构化采样. 结构化采样不仅控制采样的内容, 还控制推理轨迹的拓扑结构, 使生成, 信用分配和计算重用与问题解决的底层结构保持一致. 通过将 rollout 组织为树结构或通过共享和分段的前缀, 这些方法实现了节点级奖励, 改进的部分计算 (如 KV 缓存) 重用, 以及在内存和预算约束下更高的采样效率. 我们重点介绍两种代表性方法:

- **搜索驱动的树** Rollout: 其他工作利用蒙特卡洛树搜索 (MCTS) 进行树格式响应生成, 使用经典阶段: 初始化, 选择, 扩展和反向传播. 他们将单次推理视为树而非单链, 并在节点级别分配奖励, 这可以产生更密集/细粒度的过程信号.Hou et al. [2025] 提出 TreeRL, 一个在策略树搜索框架, 它通过更高效的搜索策略超越传统的思维链 RL(ChainRL), 同时大幅减少计算开销. 同时,ToTRL [Wu et al., 2025c] 在合成谜题环境中引入了思维树引导的训练范式, 实现了对数学推理等分布外任务的涌现泛化. 此外,Yang et al. [2025g] 将 MCTS 集成到训练流程中, 生成基于规则的细粒度过程奖励, 提高了策略优化中奖励信号的粒度和保真度.

- **共享前缀或分段方案**: 虽然这些树搜索方法丰富了探索并提供了细粒度奖励, 但它们的采样效率仍然是一个限制. 一些工作设计分段/共享前缀采样来提高生成效率 [Guo et al., 2025c, Hou et al., 2025, Li et al., 2025r, Yang et al., 2025g]. SPO [Guo et al., 2025c],TreeRPO [Yang et al., 2025g],TreeRL [Hou et al., 2025],FR3E [Zheng et al., 2025c] 和 ARPO [Dong et al., 2025b] 从先前生成的前缀开始进行额外采样. TreePO [Li et al., 2025r] 实现了分段树采样算法, 减轻了 KV 缓存负担, 减少了训练的 GPU 小时数, 并提高了采样效率.

3.3.2. 采样超参数

要点

- 仔细的超参数调优对于可扩展的 RL 至关重要, 因为天真的设置可能导致效率低下和训练不稳定 (例如熵崩溃).
- 可扩展的 RL 依赖于策略的整体组合来平衡成本和稳定性, 如分阶段上下文延长和动态探索控制.

本小节总结了近期工作中采样的超参数调整策略. 有效的 RL 训练需要在几个竞争目标之间实现精细平衡, 近期文献主要关注两个主要轴线的技术: 1) 管理探索-利用权衡, 以确保模型发现和完善有效的推理路径; 2) 高效管理序列长度, 以平衡推理深度与计算成本.

探索和利用动态. 一个核心挑战是平衡探索 (发现新颖的推理策略) 与利用 (完善高奖励解决方案). 主要杠杆是温度, 熵正则化和 PPO 的裁剪机制. 对于温度, 策略差异很大. 一些工作提出动态方法, 如分阶段温度增加 (例如, 对于 4B 模型 1.40 → 1.45 → 1.50, 对于 7B 模型 0.7 → 1.0 → 1.1) 以随着训练进展逐步扩展轨迹多样性 [An et al., 2025], 或使用调度器动态调整温度以保持稳定的熵水平 [Liao et al., 2025b]. 更规范的方法建议调整训练温度以使缩放后熵保持在 0.3 的目标附近, 这被发现能实现最佳平衡 [Liu et al., 2025v, Wu et al., 2025e]. 其他工作简单主张高固定温度 (例如 1.0 或 1.2) 以鼓励初始探索, 同时指出其本身不足以防止长期熵下降 [Arora and Zanette, 2025, Liu et al., 2025j, Shrivastava et al., 2025].

长度预算和序列管理. 几乎所有工作都在努力管理生成响应的长度以平衡性能和成本. 最普遍的策略是分阶段上下文延长 [Luo et al., 2025c]. 这涉及从短上下文窗口 (例如 $8k$) 开始 RL, 然后在后续阶段逐步增加到 $16k$,$24k$ 或 $32k$[Chen et al., 2025q, Liu et al., 2025j,v, Luo et al., 2025c]. 初始短上下文阶段被认为是必不可少的, 因为它迫使模型学习更简洁和 token 高效的推理模式 [Chen et al., 2025q, Liu et al., 2025v, Luo et al., 2025c]. 在很长上下文上训练的替代方案是在推理时应用推理时长度外推技术如 Yarn, 允许在较短序列上训练的模型泛化到更长序列 [An et al., 2025]. 对于处理超过长度预算的响应, 没有共识. 一些工作应用软线性惩罚作为响应接近最大长度 [Yu et al., 2025d] 或在奖励函数中直接应用可调惩罚 (α)[Arora and Zanette, 2025]. 更细微的, 依赖阶段的策略是在长度预算短时 ($8k$-$16k$) 过滤 (屏蔽损失) 过长样本, 但在预算大时 ($32k$) 惩罚它们, 因为在很长上下文时过滤可能变得有害 [Liu et al., 2025v, Wu et al., 2025e].

在所有这些工作中, 有效的超参数调整体现为探索 (温度, 熵目标, 裁剪), 效率 (分阶段长度课程) 和序列管理 (过长过滤器, 惩罚或推理时外推) 的联合调优. 这些方法直接适用于大多数 LLM 的 GRPO/PPO 风格 RL 流水线.

4. 基础问题

在回顾了 LLM 强化学习流程的关键组件后, 我们现在转向该领域中仍然至关重要且经常未解决的几个基础性问题. 在本节中, 我们阐述核心问题, 提出对比观点, 并总结每个开放问题的最新进展. 具体

而言, 我们讨论以下挑战:RL 的基本作用 (精炼与发现) 在 § 4.1 中,RL 与 SFT 的边界 (泛化与记忆) 在 § 4.2 中, 模型先验的选择 (弱模型与强模型) 在 § 4.3 中, 训练算法的有效性 (技巧与陷阱) 在 § 4.4 中, 以及奖励信号的粒度 (过程与结果) 在 § 4.5 中. 通过突出这些开放问题, 我们旨在澄清当前的研究格局, 并推动对 LRMs 强化学习基础原理的进一步研究.

4.1. RL 的作用：精炼还是发现

我们首先总结关于 RL 作用的两种主流观点: 精炼和发现. 这些观点似乎直接对立. 精炼观点认为 RL 不创造真正新颖的模式, 而是精炼和重新权衡基础模型中已包含的正确响应. 相比之下, 发现观点声称 RL 能够发现基础模型在预训练期间未获得且不会通过重复采样产生的新模式.

精炼和发现观点之间的分歧可以通过多个理论视角来理解. 首先, 从 KL 散度优化角度来看,SFT 通常优化前向 KL 散度 $D_{KL}(p_{data}||p_{model})$, 表现出*模式覆盖*行为: 模型试图覆盖数据分布中的所有模式. 相比之下,RL 方法优化反向 KL 散度 $D_{KL}(p_{model}||p_{reward})$, 表现出*模式寻求*行为: 将概率质量集中在高奖励区域 [Ji et al., 2024, Sun, 2024]. 最新的理论进展进一步丰富了这一理解.Xiao et al. [2025b] 证明 RLHF 可以被视为对偏好数据的隐式模仿学习, 建立了基于 RL 的对齐与行为克隆之间的深层联系. 类似地,Sun [2024] 将 SFT 本身框架化为一种逆 RL 形式, 揭示了即使是监督方法也隐含涉及奖励建模. 这些观点表明, 精炼与发现的争论可能正在解决统一学习过程的不同方面: 虽然 RL 的模式寻求特性为精炼提供了机制, 但隐式奖励学习和组合能力可以通过扩展训练实现发现.

- 最初,DeepSeek-R1 [Guo et al., 2025a] 通过 RLVR 展示了有前景的 “Aha” 行为, 启发了轻量级复现工作如 TinyZero [Pan et al., 2025c], 它通过简化的训练配方和最少的代码报告了类似现象. 特定领域的适应很快随之而来, 包括 Logic-RL [Xie et al., 2025c], 它展示了基于规则的 RL, 培养了反思和验证技能, 并迁移到数学推理.

- 然而,Limit-of-RLVR [Yue et al., 2025b] 提供了面向精炼的反驳论点:pass@k 评估表明 RL 提升了 pass@1 性能, 但在大 k 值广泛采样时往往表现不如基础模型. 这表明 RL 主要是缩小搜索空间, 而不是发现根本新颖的解决方案轨迹. 同时的争论质疑观察到的 “Aha” 行为是否真正由 RL 引起, 或仅仅是预训练期间已经嵌入的潜在能力 [Liu et al., 2025t, Setlur et al., 2025]. 机制分析进一步论证,RL 收益通常来自熵整形或奖励代理. 例如, 高熵的 “分叉”token 似乎主导了改进 [Wang et al., 2025m]; 最大化模型置信度 (RENT) 和 TTRL 在不依赖外部奖励的情况下增强了推理 [Prabhudesai et al., 2025, Zuo et al., 2025b]; 甚至虚假或随机奖励信号也可以改变 Qwen 模型 [Shao et al., 2025], 这意味着 RL 通常激活预训练的推理特征, 而不是学习全新的特征. 平行的工作将测试时搜索和计算框架化为元 RL 问题, 提出 MRT 来密集化进度信号, 并产生比仅结果的 RL 更好的 “思考时间” 缩放 [Qu et al., 2025b]. 数据效率研究也表明, 即使是极端案例如 1-shot RLVR 也可以大幅改善数学推理, 再次与引出潜在能力的精炼观点一致 [Wang et al., 2025r]. 补充这些观点, 对 RLVR 中探索的系统性研究 [Deng et al., 2025a] 将 pass@k 形式化为探索边界的度量, 并揭示了训练, 实例和 token 级别的微妙熵-性能权衡, 从而将精炼观点置于统一的分析框架内. 最近,Shenfeld et al. [2025] 引入了 “RL 剃刀” 原理, 证明在线 RL 比监督微调显著更好地保留先验知识. 他们表明,RL 的优势来自于在适应新任务时保持现有能力的能力, 而

不是发现全新的行为.

- 然而, 最近几项工作重新开启了发现的论证.ProRL [Liu et al., 2025j] 报告, 足够长时间和稳定的 RL 可以扩展基础模型的推理前沿, 改善 pass@1 和 pass@k 两者. ProRL v2 [Liu et al., 2025j] 提供了持续的缩放证据, 它结合了工程进步并展示了更强的结果. 同时, 对 pass@k 度量的批评导致了替代方案如 CoT-Pass@k, 得到理论论点的支持, 即 RLVR 隐式地激励正确的推理路径, 而不仅仅是奖励幸运的端点 [Wen et al., 2025c]. 补充方法通过使用自我博弈问题合成来保持熵并增强 pass@k 来维持 RLVR 的益处 [Liang et al., 2025c], 或通过新颖的策略目标直接优化 pass@k [Chen et al., 2025x, Walder and Karkhanis, 2025]. Yuan et al. [2025c] 通过证明 LLM 可以通过现有能力的组合在 RL 中学习新技能, 进一步为发现观点提供了令人信服的证据, 表明 RL 使得超越对现有模式简单精炼的新兴行为成为可能.

精炼和发现之间的明显二分法可能通过最近的理论进展得到调和, 这些进展揭示了不同对齐范式之间的深层联系.Xiao et al. [2025b] 的工作表明 RLHF 隐式地执行模仿学习, 而 Sun [2024] 证明 SFT 可以理解为逆 RL. 这些见解表明, 监督方法和 RL 方法都在分布匹配和奖励优化的共享理论框架内运作. 关键区别不在于这些方法是否能发现新能力, 而在于它们如何平衡探索与利用之间的权衡 [Schmied et al., 2025].RL 中反向 KL 的模式寻求特性提供了高效收敛到高性能区域的机制 (精炼), 而隐式奖励学习和序列决策方面使得在给定足够训练时间和适当正则化时, 能够将现有能力组合成新颖行为 (发现) [Liu et al., 2025j, Yuan et al., 2025c]. 这一统一观点表明, 争论应该从 "精炼或发现" 转移到理解每种现象占主导地位的条件.

4.2. RL vs. SFT: **泛化或记忆**

在本小节中, 我们讨论 RL 和监督微调的作用, 重点关注泛化和记忆之间的相互作用. LLM 后训练有两种主要方法:SFT 和 RL. 当前的争论集中在两个主要问题上: 1) 哪种方法能更好地实现分布外泛化? 2) 通过 SFT 的行为克隆是否设置了泛化能力的上限? 最近, 大量研究关注投入到这个话题. 值得注意的是,Chu et al. [2025a] 在文本和视觉环境中都提供了直接结论, 指出 "SFT 记忆,RL 泛化."

两项近期研究加深了这种对比. Huan et al. [2025] 发现数学任务上的 RL(RL-on-math) 倾向于保持甚至增强非数学任务和指令跟随的性能, 而数学任务的监督微调 (SFT-on-math) 经常导致负迁移和灾难性遗忘. 他们基于潜在空间 PCA 和 token 分布 (KL) 度量的诊断分析, 以及 Mukherjee et al. [2025] 的分析表明,SFT 引起表示和输出漂移 (记忆), 而 RL 更好地保留了基础领域结构 (泛化). 补充地,Zhou et al. [2025d] 剖析了五种数学问题解决训练路径, 并观察到 1) 数学文本的持续预训练仅提供适度迁移,2) 传统短 CoT SFT 经常损害泛化, 但 3) 长 CoT SFT 和基于规则的 RL(带有格式/正确性奖励) 扩展了推理深度和自我反思, 从而改善了更广泛的推理; 此外,RL 之前的 SFT 预热稳定了策略并进一步提升了跨域迁移. 这些结果表明,on-policy 目标和更长的自我反思轨迹培养了在分布偏移下保持鲁棒的可迁移模式, 而短 CoT SFT 倾向于过度拟合表面模式, 反映了泛化与记忆之间的经典 RL 与 SFT 分歧. 关于这个话题有三个主要研究方向:

- RL **表现出优越的泛化能力**: Chu et al. [2025a] 表明在分布外 (OOD) 性能方面,RL 优于 SFT, 而

SFT 倾向于在 GeneralPoints 和 V-IRL 任务上记忆数据. 先前的研究 [Kirk et al., 2023] 也表明 RLHF, 特别是在更大的分布偏移下, 可以比 SFT 更有效地泛化, 尽管这可能以降低输出多样性为代价. 此外,DeepSeek-R1 [Guo et al., 2025a] 证明纯 RL 训练可以导致高级推理行为的自发出现, 如反思和验证.

- RL **并非万能药**: RL 的泛化能力强烈受到初始数据分布和验证奖励设计的影响. Jin et al. [2025d] 发现 RL 可以部分缓解过度拟合; 然而, 在严重过度拟合或突然分布偏移的情况下, 它仍然无效, 如 OOD"24 点" 和光谱分析任务中所观察到的. RL 的主要价值在于其促进 "适当学习" 的能力 [Swamy et al., 2025]. 当应用适当的重新加权, 信任区域约束或动态重新缩放时,SFT 可以显著改善泛化, 并且它通常更好地为后续 RL 准备模型 [Qin and Springenberg, 2025]. 在实践中,SFT 可以作为稀疏奖励 RL 的下限.

- SFT **和** RL **的统一或交替范式**: Yan et al. [2025a] 提出了一个通过结合 off-policy 推理轨迹来增强 RLVR 的框架. Liu et al. [2025k] 将 SFT 和 RL 整合为单阶段目标, 理论上克服了长视野样本复杂性的瓶颈, 并经验上证明了优于单独使用任一方法. Fu et al. [2025c] 提出了使用熵感知权重的演示模仿 (SFT) 和策略改进 (RL) 的联合单阶段整合. Zhang et al. [2025p] 提供理论证据, 在涉及小模型, 高难度或稀疏成功轨迹的场景中, 传统的从 SFT 到 RL 的两阶段方法可能完全失败. 他们通过采用从专家锚点开始的分支推演机制来解决这个问题, 有效连接两个阶段. Ma et al. [2025a] 发现 RL 擅长巩固和增强现有能力, 而 SFT 在引入新知识或新颖模型能力方面更有效.

然而, 几个挑战仍然未解决. 一个主要问题是区分真正的问题解决能力和仅仅是答案记忆, 同时避免数据污染 [Satvaty et al., 2024]. 仍然缺乏标准化, 可重现的分布外基准测试. 此外,RL 训练对初始数据分布高度敏感; 当 SFT 引起显著的表示漂移时,RL 恢复和泛化的能力有限 [Jin et al., 2025d]. 为了应对这些挑战, 需要推广框架如 UFT [Liu et al., 2025k],SRFT [Fu et al., 2025c] 和 Interleaved [Ma et al., 2025a], 它们将 SFT 整合新知识与 RL 增强和鲁棒性机械化. Lv et al. [2025] 也探索自动化调度策略, 以确定何时在 SFT 和 RL 之间切换以及如何有效分配它们的比例.

总之,RL 在可验证任务和重大分布偏移下倾向于实现 "真正泛化", 但它不是万能药. 改进的 SFT 可以帮助弥合泛化中的剩余差距. 因此, 最佳实践正朝着结合两种方法优势的统一或交替混合范式发展 [Chen et al., 2025c,h, Liu et al., 2025k, Lv et al., 2025, Wu et al., 2025i, Zhu et al., 2025e].

4.3. 模型先验: 弱与强

最近的研究表明, 当与足够强大的模型先验和可验证奖励信号结合时,RL 现在可以在广泛任务上表现良好, 从而将主要瓶颈从规模转移到环境和评估协议的设计 [4]. 从这个角度来看,RL 主要用于重新精炼预训练期间已编码的潜在能力, 而不是从头开始生成新能力.

在本小节中, 我们检查这种依赖性的三个关键维度: 对基础模型与指令调优模型应用 RL 的比较优势, 不同模型家族间 RL 响应性的显著变化 (特别是 Qwen 和 Llama 架构之间), 以及可以增强弱先

[4] https://ysymyth.github.io/The-Second-Half/

验和强先验模型 RL 结果的新兴策略, 包括中期训练和课程设计.

基础模型与指令模型. DeepSeek-R1 首先引入了关于对基础模型或指令调优模型应用 RL 的讨论, 并介绍了两种可行的后训练范式: 1) R1-Zero, 将大规模基于规则的 RL 直接应用于基础模型, 产生新兴的长视野推理; 2) R1, 在 RL 之前纳入简短的冷启动 SFT 阶段以稳定输出格式和可读性. 独立地,Open-Reasoner-Zero [Hu et al., 2025b] 证明应用于基础 Qwen 模型的极简训练配方足以扩展响应长度和基准准确性, 反映了 R1-Zero 的训练动态. 这些发现表明基础模型先验比指令模型更适合 RL, 通常比从重度对齐的指令模型开始时产生更平滑的改进轨迹, 其中根深蒂固的格式和服从先验可能干扰奖励塑造.

模型家族差异. 最近的研究强调, 基础模型的选择可以关键性地塑造 RL 结果. 例如,One-shot RLVR [Wang et al., 2025r] 表明引入一个精心选择的数学示例可以使 Qwen2.5-Math-1.5B 的 MATH500 准确性翻倍, 在多个基准上提供显著的平均改进. 然而, 虚假奖励 [Shao et al., 2025] 揭示了对比模式:Qwen 家族模型即使在随机或虚假奖励信号下也注册显著收益, 而 Llama 和 OLMo 模型通常不会. 这种分歧突出了模型先验的影响, 并强调了在不同先验模型间验证 RL 声明的重要性. 观察到的不对称性表明预训练中推理模式 (例如数学或代码 CoT) 暴露的差异.Qwen 模型由于广泛暴露于此类分布, 倾向于更 “RL 友好”, 而可比较的 Llama 模型在经受相同 RLVR 程序时经常表现出脆弱性.

中期训练解决方案. 在实践中, 研究人员发现这种性能差距可以通过中期训练或退火训练策略来解决. 在最近的 LLM 研究中, 退火表示预训练的后期阶段, 在此期间学习率衰减, 同时数据分布重新加权以强调较小的高质量来源, 如代码, 数学和精选的 QA 语料库. Llama 3 [Grattafiori et al., 2024] 明确将此阶段命名为退火数据, 描述了数据混合的转变和学习率线性衰减到零. 他们进一步报告在此阶段注入少量高质量数学和代码显著改善了面向推理的基准. 早期,MiniCPM [Hu et al., 2024b] 阐述了类似的两阶段课程, 称为稳定后衰减. 在衰减 (退火) 阶段, 他们将 SFT 风格的高质量知识和技能数据与标准预训练语料库交错, 观察到比仅在预训练后应用相同 SFT 更大的改进. 类似地,OLMo 2 [OLMo et al., 2024] 公开了现代中期训练配方: 预训练分为漫长的重度网络阶段, 随后是较短的中期训练阶段, 该阶段上采样高质量和特定领域来源, 特别是数学, 同时将学习率线性衰减到零. 更一般地, 当代中期训练策略将学习率调度和数据分布切换的联合设计视为首要关注点. 例如,Parmar et al. [2024] 表明最佳持续预训练需要:1) 在后期阶段强调目标能力的双分布课程, 以及 2) 退火的, 非重新加热的 LR 调度, 其中分布切换的时间由 LR 分数而非固定 token 数量决定. 最近的一项系统性研究扩展了这条工作线, 证明注入高质量数学和思维链 QA 语料库的稳定后衰减中期训练课程使 Llama 模型在基于 RL 的微调下具有显著更好的可扩展性, 有效缩小了与 Qwen 模型的性能差距 [Wang et al., 2025u]. 总之, 这些发现为弱先验模型家族提供了一个实用配方: 通过中期训练加强推理先验, 然后应用 RLVR.

强模型改进. 虽然许多复现工作倾向于基础模型, 但越来越多的证据表明, 当课程, 验证和长度控制被仔细设计时,RL 可以进一步改进强化的蒸馏/指令模型. 例如,AceReason-Nemotron [Chen et al., 2025q] 报告了在蒸馏 Qwen 模型上首先数学然后仅代码 RL 的一致收益, 分析显示在 pass@1 和 pass@k 状态下都有改进. 这些发现为简单的 “仅基础” 叙述增添了细微差别: 在正确约束下, 指令/蒸馏开始也可以受益, 但优化不那么宽容. 平行工作评估了推理模型的可控性.MathIF [Fu et al., 2025a]

强调了系统性紧张关系: 扩大推理能力经常损害指令跟随性能, 特别是在长格式输出的背景下. 补充证据表明显式 CoT 提示可以降低指令跟随准确性, 并提出了选择性推理缓解措施 [Li et al., 2025l]. 总之, 这些工作激励了在 RL 中与正确性/可验证性一起的多目标训练 (格式, 简洁性, 服从性).

我们可以从三个角度总结模型先验如何从根本上塑造 LLM 训练中的 RL 结果: 1) 基础模型作为 RL 起点始终优于指令调优模型,DeepSeek-R1 和 Open-Reasoner-Zero 证明了从极简配方中出现的新兴推理; 2) 模型家族表现出不对称的 RL 响应性:Qwen 模型即使在虚假奖励下也显示收益, 而 Llama/OLMo 模型需要带有退火学习率和高质量数学/代码数据注入的仔细中期训练; 3) 强蒸馏模型可以从 RL 中受益, 但需要更复杂的课程设计和多目标优化.

随着 RL 越来越多地用于重新精炼潜在预训练能力而不是创造新能力, 重点转向全面优化预训练到 RL 的流水线, 而不是独立处理这些阶段.

4.4. 训练配方: 技巧或陷阱

大型模型的 RL 训练主要从 PPO [Schulman et al., 2017b] 系列演变而来, 通过各种工程技术 [Huang et al., 2022] 如修剪, 基线校正, 归一化和 KL 正则化来保持稳定性. 在 LLM 推理的 RL 背景下,DeepSeek-Math 和 DeepSeek-R1 引入了无批评家 GRPO [Shao et al., 2024], 通过降低复杂性简化训练过程. 尽管有这些进步, 与训练稳定性和效率相关的挑战仍然存在, 激励了一系列新方法, 包括动态采样, 各种重要性采样比率和多级归一化.

一个更广泛采用的促进探索技术是使用解耦 PPO 裁剪 (“更高裁剪”), 其中上裁剪边界设置得比下裁剪边界高 (例如,$\epsilon_{\text{low}} = 0.2, \epsilon_{\text{high}} = 0.28$), 以允许不太可能但可能有用的 token 的概率更自由地增加 [An et al., 2025, Liu et al., 2025j, Yu et al., 2025d].Archer [Wang et al., 2025i] 为具有不同熵水平的 token 提出了双裁剪机制,Archer2.0 [Wang et al., 2025h] 进一步对具有相反优势值的 token 使用非对称双裁剪.

- **数据和采样中的极简主义**:Xiong et al. [2025a] 分解 GRPO 并发现最大的性能收益来自丢弃所有不正确样本, 而不是依赖复杂的奖励归一化技术. 他们提出像 RAFT [Dong et al., 2023] 或 “Reinforce-Rej” [Liu et al., 2023a] 这样的方法可以使用更简单的机制实现与 GRPO/PPO 相当的稳定性和 KL 效率.DAPO [Yu et al., 2025d] 将 “动态采样+解耦修剪” 系统化为可重现的大规模方法, 并融入解耦 PPO 裁剪 (“更高裁剪”), 其中上裁剪边界设置得比下裁剪边界高 (例如,$\epsilon_{\text{low}} = 0.2, \epsilon_{\text{high}} = 0.28$), 以允许不太可能但可能有用的 token 的概率更自由地增加, 在 AIME24 基准的强基线上展示了最先进的结果. 类似地,GRESO [Zheng et al., 2025b] 表明预过滤可以将推演时间加速 2.4 倍, 整体训练加速 2.0 倍, 而性能损失最小.
- **目标函数的结构修改**:GSPO [Zheng et al., 2025a] 将比率和裁剪操作转移到序列级别, 导致相比 GRPO 的改进稳定性和效率, 特别是对于专家混合 (MoE) 模型的稳定 RL 训练. S-GRPO [Dai et al., 2025a] 进一步减少冗余推理, 缓解了更长和不必要推理链的趋势, 在多个基准上将序列长度缩短 35 – 61 %, 同时准确性略有改进.
- **去偏置与归一化之间的斗争**:Dr. GRPO [Liu et al., 2025u] 识别了 GRPO 中的一个关键偏差, 即

“越错越错”, 并引入轻微的算法修改以改善 token 效率. 同时, 其他研究 (例如 BNPO [Xiao et al., 2025a]) 从自适应分布角度重新审视奖励归一化的重要性, 提出新的归一化族. 这两个阵营的证据是矛盾的, 表明将归一化视为通用解决方案可能是误导性的.

Liu et al. [2025w] 提出了一个具有统一评估的最新综述, 将常见技术整合到单个开源框架 [Wang et al., 2025n] 中以实现隔离和可重现的实验. 这项工作提供了一个路线图, 概述了 “什么设置下哪些技术有效”, 并证明了方法的极简组合可以在多个配置中优于 GRPO 和 DAPO. 至关重要的是, 它突出了该领域最紧迫的挑战: 不一致的实验设置, 不完整的报告和矛盾的结论. 这构成了当前研究社区中 RL 应用的基本限制. 总之, 虽然实用 “技巧” 对于稳定 RL 训练很有价值, 但 “科学训练” 的本质在于验证和可扩展性. 该领域的进展需要统一的实验协议, 可验证的奖励结构和明确的可扩展性-性能-成本曲线 [Nimmaturi et al., 2025], 以显示方法在扩展时仍然有效, 而不仅仅是在特定数据或模型上.

4.5. 奖励类型：过程还是结果

在标准 RL 中, 策略的目标是最大化期望累积奖励 [Sutton et al., 1998].“奖励足够” 假设 [Bowling et al., 2023, Silver et al., 2021] 进一步假设适当设计的奖励是充分的, 最大化回报原则上可以产生智能的所有方面. 在 LLM 的 RL 背景下, 核心挑战是如何提供有意义的奖励, 例如训练奖励模型或验证器来对输出评分, 并将这些分数用于 RL 或搜索. 常见方法包括结果奖励, 只评估最终结果 (例如正确性或通过个别测试), 以及过程奖励, 通过对中间步骤的密集反馈提供逐步评分 [Lightman et al., 2024].

- 如 § 3.1.1 所示, 当任务答案可验证时, 结果奖励对于具有挑战性的数学和编码任务是最简单和最可扩展的. 然而, 仅结果的方法可能 tacitly 鼓励不忠实的思维链 [Arcuschin et al., 2025], 如 “先答案, 后幻觉”, 并奖励推测. 最近的研究 [Baker et al., 2025] 表明, 最先进的模型在现实场景中也表现出不忠实的推理和事后合理化. 其他工作强调了基于规则的 RL 容易受到奖励攻击和推理幻觉的发展 [Sun et al., 2025h].
- PRMs [Zhang et al., 2025f] 自然促进长链信用分配.Lightman et al. [2024] 清楚比较了两种奖励方法: 对于数学推理, 使用过程监督训练的 PRMs 更稳定和可靠, 显著优于仅由结果监督的那些. 然而, 逐步标注极其昂贵, 质量在不同领域经常下降 [Zhang et al., 2025u]. 相关研究表明, 启发式或基于蒙特卡罗的综合方法往往泛化性差并引入偏差 [Yin et al., 2025].

总之, 结果奖励提供 “具有自动验证的可扩展目标对齐”, 而过程奖励提供 “可解释的密集指导”. 将两者结合, 例如通过隐式过程建模 [Cui et al., 2025a] 或生成式验证器 [Zhang et al., 2024a], 可能代表奖励设计的一个有前景的未来方向.

5. 训练资源

LLM 的有效 RL 不仅依赖于算法和目标设计, 还依赖于基础训练资源的质量和结构. 从静态语料库到动态环境和专用 RL 基础设施的资源选择, 深刻影响着大规模训练的稳定性和可扩展性. 在本节中,

我们调查当前实践中使用的关键训练资源类别. 我们首先检查静态语料库作为 RL 基础的作用和局限性 (§ 5.1), 然后讨论动态, 交互环境日益增长的重要性, 这些环境提供更丰富的学习信号和更真实的任务分布 (§ 5.2). 最后, 我们回顾使 LLM 具有可扩展和高效训练流水线的 RL 基础设施 (§ 5.3).

5.1. 静态语料库

Takeaways

- RL 推理数据集正从大规模原始数据转向更高质量, 可验证的监督, 使用蒸馏, 过滤和自动评估来提升样本有效性和过程保真度.
- 数据覆盖范围已扩展到单一领域 (数学/代码/STEM) 之外, 包括搜索, 工具使用和代理任务, 具有可追溯的计划-行动-验证轨迹.

本节调查 LLM 的 RL 静态语料库. 数据构建正在从 “规模优先” 转向 “质量和可验证性优先”, 明确支持可验证奖励 (参见 § 3.1.1). 如表 4 所示, 数据集覆盖范围跨越四个主要轨道: 数学, 编码,STEM 和代理任务 (例如搜索和工具使用). 所有语料库都直接兼容 RLVR, 实现过程感知评估. 这些数据集支持 RL 流水线的关键组件, 包括策略预训练, 奖励建模和难度感知采样.

数学导向的 RL 数据集围绕三个构建流水线聚集, 包括注释/验证, 蒸馏和多源合并, 同时广泛暴露中间推理轨迹, 规模从数百到数百万示例不等. 紧凑, 精心策划的集合如 LIMO [Ye et al., 2025d] 和 LIMR [Li et al., 2025p] 强调具有明确过程反馈的高质量问题; 注释/验证资源如 DAPO [Yu et al., 2025d],Big-MATH [Albalak et al., 2025] 和 DeepMath [He et al., 2025h] 提供适合奖励建模和价值对齐的可靠解决方案轨迹; 在更大规模上,NuminaMath 1.5 [Li et al., 2024b] 扩展了丰富的过程样本; 以蒸馏为中心的语料库包括 DeepScaleR [Luo et al., 2025c],OpenR1-Math [Hugging Face, 2025] 和 OpenMathReasoning [Moshkov et al., 2025] 继承了强教师或 “R1 风格” 的长链推理, 支持策略预训练和 RL 阶段选择; 合并和蒸馏集合如 PRIME [Cui et al., 2025a],OpenReasoningZero [Hu et al., 2025b] 和 STILL-3-RL [Chen et al., 2025w] 将开放问题与自生成候选者集成, 提供难度分层和高质量过滤信号; 社区导向的发布如 Light-R1 [Wen et al., 2025b] 和 MiroMind-M1-RL-62K [Li et al., 2025n] 打包轻量级,RL 就绪格式, 用于计算约束下的快速迭代. 总之, 这些资源涵盖从基础计算到竞赛级别的问题, 并提供最终答案和可测量的中间步骤, 实现可扩展的策略学习, 奖励建模和基于过程的强化.

代码导向的 RL 数据集主要分为三类: 程序修复/编辑, 算法竞赛问题和带有推理的通用代码合成. 这些数据集通常提供可执行的单元测试和中间执行轨迹, 促进奖励塑造和过程级评估. 交互式, 测试驱动的资源如 SWE-Gym [Pan et al., 2024] 针对细粒度编辑策略; 人工验证的修复对如 SWE-Fixer [Xie et al., 2025a] 和 LeetCodeDataset [Xia et al., 2025c] 支持价值对齐和奖励建模. 对于竞赛风格和算法推理,codeforces-cots [Penedo et al., 2025],Z1 [Yu et al., 2025f] 和 OpenCodeReasoning [Ahmad et al., 2025] 强调长链轨迹和难度分层. 在大规模,“R1 风格” 的通用代码生成蒸馏中,KodCode [Xu et al., 2025h] 和 rStar-Coder [Liu et al., 2025q] 提供丰富的过程样本, 有助于策略预训练和 RL 阶段选择. 轻量级, 以合并为中心的发布如 Code-R1 [Liu and Zhang, 2025] 和 DeepCoder [Luo et al., 2025b] 在计算约束下便于快速迭代. 总之, 这些语料库涵盖从单函数修复到竞赛级

Table 4 | 用于 LLM 的 RL 训练的静态数据集，包括数学、编程、STEM 和智能体领域。对于数据获取方法，“Distil” 和 “Anno” 分别表示蒸馏和注释。“Merge” 表示现有数据集的集成，包括难度和质量过滤。

Domain	Date	Name	#Sample	Format	Type	Link
Math	2025.02	DAPO	17k	Q-A	Anno	
	2025.02	PRIME	481k	Q-A	Merge&Distil	
	2025.02	Big-MATH	47k	Q-A	Anno	
	2025.02	LIMO	800	Q-C-A	Anno	
	2025.02	LIMR	1.39k	Q-A	Anno	
	2025.02	DeepScaleR	40.3k	Q-C-A	Distil	
	2025.02	NuminaMath 1.5	896k	Q-C-A	Anno	
	2025.02	OpenReasoningZero	72k	Q-A	Merge&Distil	
	2025.02	STILL-3-RL	90k	Q-A	Merge&Distil	
	2025.02	OpenR1-Math	220k	Q-C-A	Distil	
	2025.03	Light-R1	79.4k	Q-C-A	Merge	
	2025.04	DeepMath	103k	Q-C-A	Distil&Anno	
	2025.04	OpenMathReasoning	5.5M	Q-C-A	Distil	
	2025.07	MiroMind-M1-RL-62K	62k	Q-A	Merge	
Code	2024.12	SWE-Gym	2.4k	Q-A	Anno	
	2025.01	codeforces-cots	47.8k	Q-C-A	Distil	
	2025.01	SWE-Fixer	110k	Q-A	Anno	
	2025.03	KodCode	268k	Q-A	Distil	
	2025.03	Code-R1	12k	Q-A	Merge	
	2025.04	Z1	107k	Q-C-A	Distil	
	2025.04	LeetCodeDataset	2.9k	Q-A	Anno	
	2025.04	OpenCodeReasoning	735k	Q-C-A	Distil	
	2025.04	DeepCoder	24k	Q-A	Merge	
	2025.05	rStar-Coder	592k	Q-C-A	Distil&Anno	
STEM	2025.01	SCP-116K	182k	Q-C-A	Distil	
	2025.02	NaturalReasoning	2.15M	Q-C-A	Distil	
	2025.05	ChemCoTDataset	5k	Q-C-A	Distil	
	2025.06	ReasonMed	1.11M	Q-C-A	Distil	
	2025.07	MegaScience	2.25M	Q-C-A	Merge&Distil	
	2025.09	SSMR-Bench	16k	Q-A	Anno	
Agent	2025.03	Search-R1	221K	Q-A	Anno	
	2025.03	ToRL	28K	Q-A	Merge	
	2025.03	ToolRL	4K	Q-C-A	Distil	
	2025.05	ZeroSearch	170K	Q-A	Anno	
	2025.07	WebShaper	0.5K	Q-A	Anno	
	2025.08	MicroThinker	67.2K	Q-A	Anno	
	2025.08	ASearcher	70K	Q-A	Anno	
Mix	2025.01	dolphin-r1	300k	Q-C-A	Distil	

Continued on next page

Table 4 – Continued from previous page

Domain	Date	Name	#Sample	Format	Type	Link
	2025.02	SYNTHETIC-1/2	2M/156K	Q-C-A	Distil	
	2025.04	SkyWork OR1	14k	Q-A	Merge	
	2025.05	Llama-Nemotron-PT	30M	Q-C-A	Distil	
	2025.06	AM-DS-R1-0528-Distilled	2.6M	Q-C-A	Distil	
	2025.06	guru-RL-92k	91.9k	Q-A	Distil	

别的问题解决, 提供自动可检查的最终工件和分步计划/编辑, 从而实现代码代理的可扩展策略学习, 奖励建模和基于过程的强化.

STEM 导向的 RL 数据集通常围绕三个主题收敛: 教科书或课程提取, 跨学科大规模推理, 以及以合并和蒸馏流水线为特征的领域专用语料库 (例如化学和医学). 这些数据集通常发布思维链理由和证据对齐信号, 实现过程级奖励. SCP-116K [Lu et al., 2025a] 针对本科到博士级科学, 提供自动提取的问题-解决方案对加模型生成的推理.NaturalReasoning [Yuan et al., 2025e] 提供从流行基准中去污染的多学科问题, 附带提取的参考答案.ChemCoTDataset [Li et al., 2025d] 贡献化学特定的 CoT 示例, 涵盖分子编辑/优化和反应预测.ReasonMed [Sun et al., 2025f] 提供多代理蒸馏的医学 QA, 具有多步 CoT 理由和简洁摘要.SSMR-Bench [Wang et al., 2025v] 以编程方式合成基于音乐理论的乐谱推理问题, 采用文本 (ABC 记谱法) 和视觉格式, 每种模态发布 16k 训练对, 并支持评估以及具有可验证奖励的 RL.MegaScience [Fan et al., 2025a] 通过基于消融的选择聚合公共科学语料库, 并为大多数组成集标注逐步解决方案, 形成科学推理 RL 的大型训练池.

混合域 RL 数据集通过蒸馏优先和合并中心的流水线统一数学, 代码和科学推理, 同时广泛发布思维链轨迹, 验证器信号和多轨迹候选者, 实现过程奖励和难度感知选择. 在 R1 风格混合中,dolphin-r1 [Team, 2025b] 混合 DeepSeek-R1,Gemini-thinking 和策划的聊天数据用于一般推理.SYNTHETIC 套件将大规模 SFT 风格轨迹与 RL 就绪的多轨迹样本耦合:SYNTHETIC-1 [Mattern et al., 2025] 聚合 DeepSeek-R1 推理与多样化验证器,SYNTHETIC-2-RL [Mattern et al., 2025] 为偏好/奖励学习提供具有多轨迹的多领域任务.SkyWork OR1-RL-Data [He et al., 2025d] 强调具有难度标签的可验证数学和代码问题, 作为轻量级 RL 池.Llama-Nemotron Post-Training [Bercovich et al., 2025] 编译跨越数学, 代码,STEM, 一般推理和工具使用的指令/R1 风格数据用于后训练.AM-DeepSeek-R1-0528-Distilled [a-m team, 2025] 提供具有文档化质量过滤的跨域蒸馏轨迹,guru-RL-92k [Cheng et al., 2025d] 通过针对 RL 格式优化的五阶段流水线策划六个高强度推理领域. 总之, 这些语料库提供跨域的可验证端点和逐步理由, 支持可扩展的策略学习, 奖励建模和基于过程的强化.

以代理为中心的 RL 数据集专注于两个互补能力, 搜索即行动和工具使用, 同时发布可验证的过程信号, 如搜索/浏览轨迹, 证据 URL 和工具执行日志, 实现过程奖励和离线评估.Search-R1 [Jin et al., 2025b] 基于 NQ/HotpotQA 构建训练交错推理-搜索行为.ToRL [Li et al., 2025q] 从基础模型扩展工具集成 RL, 学习何时以及如何调用计算工具.ToolRL [Qian et al., 2025] 研究学习工具选择和应用的细粒度奖励设计.ZeroSearch [Sun et al., 2025a] 制定离线信息寻求任务, 激励搜索而不进行真实网络调用.WebShaper [Tao et al., 2025] 通过 “扩展代理” 合成信息寻求数据, 覆盖具有 URL 证据的多样化任务形式和推理结构.MicroThinker [Team, 2025f] 为多步代理贡献完整推演轨迹和丰富的工具使

用日志.ASearcher [Gao et al., 2025a] 为长视野搜索代理发布 Apache-2.0 许可的训练分割, 具有问题/答案字段和源注释. 总之, 这些语料库涵盖规划, 检索, 工具编排, 证据验证和答案生成, 支持网络/搜索和工具使用代理的可扩展策略学习, 奖励建模和基于过程的强化.

5.2. 动态环境

Takeaways

- 静态 RL 训练数据集对于先进和可泛化的推理能力越来越不足.
- LLM 的可扩展 RL 需要转向合成或生成的数据和交互式环境, 如各种训练场和世界模型.

现有的静态 RL 语料库, 无论是人工注释, 半自动标记还是从网络抓取的, 对于需要更先进和可泛化推理能力的模型训练越来越不足. 越来越多的工作现在利用 “动态环境” 来共同确保可扩展性和可验证性, 这是有效模型训练的两个基本属性 [Wei, 2025].

与传统推理语料库不同, 这些动态环境代表了一种范式转变. 它们要么实现数据的自动化和无限合成, 要么提供对模型整个推理过程的步骤级, 多轮反馈. 如表 5 所示, 基于用于合成和交互的方法, 这些环境可以被分类, 作为 RL 过程的交互对象. 鉴于我们对训练资源的关注, 本小节对数据集和环境的组织将排除仅用于评估的基准.

基于规则的环境. 仅仅依赖 “完全匹配” 这样的反馈会导致模型走向死记硬背的捷径而不是实际推理。为了对抗这一点，一些环境提供复杂和多样的任务，需要确定性的基于规则的操作作为验证器。AutoLogi [Zhu et al., 2025d] 通过构建基于固定模型输出格式检查逻辑约束正确性的代码，生成具有可控难度的开放式逻辑谜题。Logic-RL [Xie et al., 2025c] 使用可扩展的骑士和骗子谜题创建基于规则的 RL 环境，将 7B 模型的推理能力泛化到数学领域。像 SynLogic [Liu et al., 2025g]、Reasoning Gym [Stojanovski et al., 2025] 和 Enigmata [Chen et al., 2025d] 等项目进一步扩展了任务多样性。它们识别控制每个任务难度的关键参数，允许跨各种逻辑相关推理挑战无限生成数据。相比之下，ProtoReasoning [He et al., 2025b] 基于模型泛化能力来自共享抽象推理原型的假设运行。它将不同的任务类型标准化为一致的格式，如 Prolog 问题或 PDDL 任务，然后使用解释器自动验证模型的输出。

基于代码的环境. LLM 推理的一个重要应用领域是软件工程和代码开发。这些环境的一个关键特征是模型在训练期间必须与可编译的代码环境交互。因此，如何可扩展地构建基于代码的任务环境仍然是一个重要的研究方向。为了教智能体使用工具，ReCall [Chen et al., 2025k] 利用先进的 LLM 构建基于 Python 的工具交互环境，自主合成自己的 SynTool 数据用于 RL 训练。在 AutoML 领域，MLGym [Nathani et al., 2025] 是首批支持迭代实验和训练的交互环境的框架之一。它使用 Docker 容器隔离每个任务的执行环境。虽然其任务大多是固定的，但它提供的可扩展性较差。MLE-Dojo [Qiang et al., 2025] 提供更好的可扩展性，因为用户更容易集成新任务。类似地，MedAgentGym [Xu et al., 2025b] 是医疗领域的高效、可扩展的交互式训练环境。在软件工程领域，R2E-Gym [Jain et al., 2025] 通过直接从 GitHub 提交历史以编程方式生成环境，减少了对手动编写的 GitHub 问题和测试用例的依赖，并与 OpenHands 集成以实现交互功能。类似地，SWE-rebench [Badertdinov

Table 5 | 用于 LLM 的 RL 训练的动态 RL 环境。数据源说明：RD = 读取数据，RS = 基于规则的合成，MS = 基于模型的合成。规模说明：训练集/测试集。

Category	Date	Name	Data Source	Interactive	Scale	Multimodal	Link
Rule-based	2025.02	AutoLogi	RD + MS	×	2458/6739 puzzles	×	
	2025.02	Logic-RL	RS	×	5k samples	×	
	2025.05	Reasoning Gym	RS	×	104 tasks	×	
	2025.05	SynLogic	RS	×	35 tasks	×	
	2025.06	ProtoReasoning	RD + MS	×	6620 samples	×	-
	2025.06	Enigmata	RD + RS	×	36 tasks	×	
Code-based	2024.07	AppWorld	RD + RS	✓	750 tasks	×	
	2025.02	AgentCPM-GUI	RD + RS	✓	55k trajectories	✓	
	2025.02	MLGym	RD + RS	✓	13 tasks	×	
	2025.03	ReCall	RD + MS	✓	10010 samples	×	
	2025.04	R2E-Gym	RD + MS	✓	8135 cases	×	
	2025.05	MLE-Dojo	RD + RS	✓	202 tasks	✓	
	2025.05	SWE-rebench	RD + MS	✓	21336 cases	×	
	2025.05	ZeroGUI	MS	✓	-	✓	
	2025.06	MedAgentGym	RD	✓	72,413 cases	×	
Game-based	2020.10	ALFWorld	RS	✓	6 tasks	✓	
	2022.03	ScienceWorld	RS	✓	30 tasks	×	
	2025.04	Cross-env-coop	RS	✓	1.16e17 cases	×	
	2025.05	lmgame-BENCH	RD + RS	✓	6 games	✓	
	2025.05	G1(VLM-Gym)	RD + RS	✓	4 games	✓	
	2025.06	Code2Logic (GameQA)	RD + MS	×	140k QA	✓	
	2025.06	Play to Generalize	RS	✓	36k samples × 2 games	✓	
	2025.06	KORGym	RS	✓	51 games	✓	
	2025.06	Optimus-3	RS	✓	6 tasks	✓	
	2025.08	PuzzleJAX	RS	✓	~ 900 games	✓	
Model-based	2025.03	Sweet-RL	RD + MS	✓	10k/1k tasks	×	
	2025.04	TextArena	RS	✓	99 games	×	
	2025.05	Absolute Zero	MS	✓	-	×	
	2025.06	SwS	RD + MS	×	40k samples	×	
	2025.07	SPIRAL	RS	✓	3 games	×	
	2025.08	Genie 3	MS	✓	-	✓	
Ensemble-based	2025.06	InternBootcamp	RD + RS	✓	1060 tasks	×	
	2025.07	Synthetic-2	RD + MS	✓	19 tasks	×	

et al., 2025] 通过提出构建软件工程任务的可扩展流水线扩展了原始的静态 SWE-bench。该流水线包括模拟真实软件开发场景的复杂交互任务，确保数据新鲜度并避免数据污染。在计算机使用领域，AgentCPM-GUI [Zhang et al., 2025v] 在 RFT 阶段构建交互式 GUI 环境，为模型的行为提供反馈。类似地，AppWorld [Trivedi et al., 2024] 使用包含各种移动应用程序 API 的环境。ZeroGUI [Yang et al., 2025b] 更进一步，使用现有的先进 VLM 为 Ubuntu 和 Android 构建任务。在训练期间，GUI 智能体与环境交互，然后将反馈提供给 VLM 以给予奖励，所有这些都无需手动数据策划。

基于游戏的环境. 游戏环境的特征是它们清晰而复杂的状态空间，其中 AI 的行为与环境状态紧密耦合。与前面提到的环境相比，这导致了更多的多步骤和连续交互过程，这类环境自然支持 § 3.1.3 中的密集奖励，使 RL 训练更高效和稳定。训练智能体的交互环境的早期工作，如 ALFWorld [Shridhar et al., 2020] 和 ScienceWorld [Wang et al., 2022]，在智能体规划领域仍然具有影响力。Code2Logic [Tong et al., 2025b] 利用游戏代码和问答模板自动生成多模态推理数据，产生了 GameQA 数据集。该数据集不仅可扩展，而且以递增难度测试模型的多模态推理能力。lmgame-Bench [Hu et al., 2025c] 采用不同的方法，直接选择经典游戏并通过统一 API 与 LLM 交互。游戏环境根据 LLM 的行为更新其状态并提供奖励，LLM 然后使用这些奖励来调整其策略。类似地，Play to Generalize [Xie et al., 2025d] 使用简单的可扩展游戏环境进行 RL，训练了一个 7B 参数的 MLLM。研究发现，模型获得的推理技能可以泛化到未见过的游戏和多学科推理任务。G1 工作 [Chen et al., 2025g] 引入了 VLM-Gym，这是一个支持多个游戏状态并行执行的 RL 环境，促进了大规模训练。KORGym [Shi et al., 2025a] 进一步扩展了支持的简单游戏数量，提供交互式和可配置难度的 RL 环境。PuzzleJAX [Earle et al., 2025] 采用不同的方法，使用 JAX 加速从 PuzzleScript 语言生成的游戏。这不仅加速了游戏环境以支持基于 RL 的训练，还提供了访问游戏开发者社区和无限游戏来源的途径。为了学习通用的合作技能，Cross-environment Cooperation [Jha et al., 2025a] 利用游戏 Overcooked 并在自玩框架内最大化环境多样性。对于像 Minecraft 这样更复杂、高自由度的游戏，Optimus 系列工作 [Li et al., 2025u] 利用知识图与游戏环境交互，构建数据来评估模型的长期规划能力。

基于模型的环境. 这种范式通过模型到模型交互或自玩促进了高度灵活和多样化的 RL 环境的创建。SwS [Liang et al., 2025b] 利用模型失败的训练案例来抽象关键概念并生成新问题，从而有针对性地增强其推理能力。SPIRAL [Liu et al., 2025a] 使用三个零和游戏进行自玩，防止对静态策略的过拟合。对于模型到模型交互，Sweet-RL [Zhou et al., 2025g] 使用类似证明者-验证者的训练框架，其中智能体与基于 LLM 的人类模拟器交互和协作，解决前端设计和后端编程任务。TextArena [Guertler et al., 2025] 提议使用对抗性文本游戏结合排名系统，通过让模型直接交互来相对衡量其能力，克服了人类评分的瓶颈。Absolute Zero [Zhao et al., 2025a] 更进一步，完全脱离了人类定义的评估任务，利用三种推理模式让模型自主生成自己的任务并通过自我进化改进推理能力。在视觉领域，Genie 3 [Ball et al., 2025] 生成接近真实和交互式的 3D 虚拟环境，为未来的多模态环境交互 RL 奠定了基础。虽然一些现有的世界模型已经支持了基于 RL 的模型训练 [Dedieu et al., 2025, Hafner et al., 2023, Russell et al., 2025]，并且我们在上面列出了使用基于模型的环境训练 LRM 的工作，但仍然没有足够可扩展的解决方案来支持基于世界模型的 LRM 的 RL 训练。我们假设，这种动态环境的最终形式将是一个能够模拟完整、自包含世界的神谕世界模型。

基于集成的环境. 还有一些工作涉及重大的工程努力，集成各种任务和数据集，形成 RL 的交互环境和训练数据。InternBootcamp [Li et al., 2025g] 是一个大规模、可扩展的环境库，专为训练 LRM 而设计。它通过提供可控制难度的生成器和基于规则的验证器，支持跨八个领域的 1000 多个通用推理任务。一个关键贡献是其实证证明了"任务扩展"，表明增加训练任务数量显著提升推理性能和训练效率。Synthetic-2 [PrimeIntellect, 2025] 通过提供四百万验证推理轨迹的大规模开放数据集为这种方法做出了贡献。这些轨迹是通过"行星规模、流水线并行、去中心化推理运行"协作生成的，展示了为复杂 RL 任务创建验证训练数据的高度可扩展方法。

Table 6 | 用于 LLM 后训练的开源 RL 基础设施。状态说明：✓ = 原生支持，× = 不支持，P = 部分支持。

Date	Framework	Runtime				Serving		Training		
		Async	Agents	Multi-Agents	Multimodal	vLLM	SGLang	DeepSpeed	Megatron	FSDP
	Primary development									
2020.03	TRL	×	×	×	P	✓	×	✓	×	✓
2023.11	OpenRLHF	✓	✓	×	×	✓	×	✓	×	×
2024.11	veRL	✓	✓	×	P	✓	✓	×	✓	✓
2025.03	AReaL	✓	✓	×	P	✓	✓	✓	✓	✓
2025.05	NeMo-RL	P	P	×	✓	✓	×	×	✓	✓
2025.05	ROLL	✓	✓	×	✓	✓	✓	✓	✓	×
2025.07	slime	✓	P	×	×	×	✓	×	✓	×
2025.09	RLInf	✓	✓	×	✓	✓	✓	×	✓	✓
	Secondary development									
2025.02	rllm	P	✓	×	×	✓	✓	×	×	✓
2025.02	VLM-R1	×	×	×	✓	✓	×	✓	×	×
2025.03	EasyR1	×	×	×	✓	✓	×	×	×	✓
2025.03	verifiers	✓	✓	×	×	✓	×	✓	×	✓
2025.05	prime-rl	✓	×	×	×	✓	×	×	×	✓
2025.05	MARTI	P	✓	✓	×	✓	×	✓	×	×
2025.05	RL-Factory	✓	✓	×	✓	✓	✓	✓	✓	✓
2025.06	verl-agent	✓	✓	×	✓	✓	✓	✓	✓	✓
2025.08	agent-lightning	✓	✓	P	×	✓	×	×	✓	✓

5.3. RL 基础设施

Takeaways

- 现代 RL 基础设施以灵活流水线和通信层为中心，在代理推演和策略训练之间分配资源，通常作为成熟分布式训练框架和推理引擎的包装器实现.
- 专门变体 (代理工作流，多代理和多模态) 通常支持异步推演/训练和标准化环境接口.

在本小节中，我们介绍不仅促进算法研究而且促进下游应用发展的开源 RL 基础设施. 我们首先介绍主要开发框架，这些框架主要提供围绕 LLM 训练和推理框架的基本包装器. 接下来，我们介绍次要开发框架，这些框架构建在这些主要框架之上，并进一步适应各种下游应用，包括代理 RL，编码 RL，多代理 RL 和多模态 RL，分布式 RL 等. 我们在表 6 中比较这些开源 RL 框架，并在下面介绍主要框架.

主要开发. 当前 RL 基础设施大量依赖于为 LLM 设计的成熟训练框架和推理引擎. 诸如 DeepSpeed [Rasley et al., 2020],Megatron [Shoeybi et al., 2019] 和完全分片数据并行 (FSDP) [Zhao et al., 2023b] 等框架针对 LLM 的预训练和后训练都进行了优化. 在推理方面,vLLM [Kwon et al., 2023] 和 SGLang[5]

[5] https://github.com/sgl-project/sglang

专为高效推理量身定制, 采用了先进的调度器和闪存注意力机制. 与 PyTorch 模型的直接前向计算相比, 这些优化实现了显著更快的推理. 许多开源 RL 框架构建在即插即用的训练和推理框架之上, 其中大部分在分布式计算引擎如 Ray[6] 上实现. 在这里, 我们回顾基于上述骨干训练和推理框架直接开发的 RL 框架.

- TRL [von Werra et al., 2020]: TRL 专注于以训练器为中心的后训练，包括 SFT、PPO/GRPO、DPO 和专门的 RewardTrainer（以及最近的在线变体），而不是定制的分布式运行时。它集成了 vLLM 用于在线方法（服务器或并置模式），但本地不支持 SGLang 或 TensorRT-LLM。扩展委托给 accelerate，该框架本地支持 DDP、DeepSpeed ZeRO 和 FSDP；Megatron 不是后端。通过 RewardTrainer 原生支持奖励建模，该库为 GRPO/DPO/在线展开提供了清晰的 API。

- OpenRLHF [Hu et al., 2024a]: OpenRLHF 提供 PPO、GRPO、REINFORCE++（及其基线变体）和 RLOO 的分布式实现，还包括偏好学习基线如 DPO/IPO/cDPO 和 KTO。其运行时支持异步流水线 RLHF 和异步智能体 RL 模式，为多轮设置暴露基于类的智能体 API。在服务方面，OpenRLHF 与 vLLM 紧密集成以实现高吞吐量展开。训练围绕 DeepSpeed ZeRO-3 与自动张量并行（AutoTP）组织，不需要 Megatron 或 FSDP。该框架提供 RM 和 PRM 训练的配方，并将 PRM 信号集成到展开中。

- Verl [Sheng et al., 2025]: Verl 提供最广泛的算法菜单之一（PPO、GRPO、GSPO、ReMax、REINFORCE++、RLOO、PRIME、DAPO/DrGRPO 等），以及多轮训练和工具使用。其运行时以 HybridFlow 控制器为中心，添加智能体 RL 展开和分离异步训练的原型（在公开路线图中有"异步和离策略架构"）。Verl 支持 vLLM 和 SGLang 用于服务，并提供 FSDP 和 Megatron-LM 训练后端。奖励选项包括基于模型的和函数/可验证奖励（例如数学/编码），支持多 GPU LoRA-RL。

- AReaL [Fu et al., 2025b]: AReaL 针对大型推理模型的高吞吐量 RL，采用完全异步设计，通过可中断的展开工作者、重放缓冲区和并行奖励服务（例如基于单元测试的代码奖励）将生成与训练解耦，由陈旧感知 PPO 目标稳定。经验上，该系统报告在数学/代码基准上匹配或更好的最终精度下，训练速度提升高达 2.77×，并近乎线性扩展到 512 个 GPU。开源堆栈强调基于 SGLang 的展开服务和 Ray 启动器，用于单节点到 ~1K-GPU 集群，PyTorch FSDP 作为主要训练后端（Megatron 也可用）；较新的"AReaL-lite"添加了算法优先 API，包含 GRPO 示例和支持多轮智能体 RL/RLVR 工作流。

- NeMo-RL [NVIDIA-NeMo, 2025]: NVIDIA 的 NeMo 堆栈现在暴露了专门的"NeMo RL"库和早期的 NeMo-Aligner 对齐工具包。算法上，NeMo 涵盖 SFT 和偏好训练（DPO/RPO/IPO/REINFORCE）以及完整的 RLHF 与 PPO 和 GRPO，包括多轮变体。运行时强调可扩展、生产导向的编排和广泛并行性；训练建立在 Megatron Core 上（张量/数据/流水线/专家并行），用于 100B 规模模型和多节点集群。在服务方面，NeMo 框架记录了 TensorRT-LLM 和 vLLM 的部署。奖励模型训练在 RLHF 教程中是一等公民，具有从 RM 拟合到 PPO 的端到端流水线。

[6] https://github.com/ray-project/ray

- ROLL [Wang et al., 2025n]: ROLL 针对 LLM 的大规模 RL，包含 GRPO/PPO/REINFORCE++ 和其他配方（例如 TOPR/RAFT++/GSPO），并明确支持异步训练和智能体 RL 流水线。运行时遵循基于 Ray 的多角色设计，集成 SGLang 和 vLLM 用于展开服务。训练主要围绕 Megatron-Core 构建，FSDP2 在公开路线图上列出；DeepSpeed 被确认为依赖项。奖励处理通过 Reward Workers（例如验证器、沙盒工具、LLM 作为评判器）和可插拔环境实现模块化。技术报告详细说明了系统和扩展考虑。

- slime [THUDM, 2025]: Slime 被定位为 SGLang 原生的 RL 扩展后训练框架，将展开侧的 SGLang 与训练侧的 Megatron 连接。它强调基础设施而非算法广度，但提供了密集和 MoE 模型的示例，包括多轮+工具调用（"Search-R1 lite"）。运行时支持异步训练和智能体工作流；服务通过 SGLang 是一等公民。训练使用 Megatron-LM 和 Ray 进行集群启动；奖励建模本身不是主要焦点，尽管验证器/"奖励"信号可以在展开平面上产生。

- RLinf [Team, 2025h]: RLinf 是一个体现智能框架，强调模块化和适应性。在"大脑"和"小脑"范式共存以及该领域仍在发展轨迹的动机下，RLinf 采用宏观到微观流（M2Flow）范式，将宏观层面逻辑工作流与微观层面物理执行分离，从而实现可编程组合和高效调度。在运行时，RLinf 允许强化学习组件（例如 Actor、Critic、Reward、Simulator）灵活放置在任意 GPU 上，并配置为并置、分离或混合执行模式——从共享全部放置到细粒度流水线。代表性案例将 Generator 和基于 GPU 的 Simulator 解耦为流水线，而 Inference 和 Trainer 共享执行。在服务方面，RLinf 支持 vLLM/SGLang，在训练方面集成 Megatron/FSDP。

次要开发. 在这一部分, 我们介绍几个构建在主要开发框架之上并扩展其功能以支持更广泛下游应用的代表性框架. 我们主要关注代理 RL, 多模态 RL 和多代理 RL 的框架. 虽然一些主要框架已经为这些领域提供了部分支持, 但我们强调为特定领域研究设计的专门框架.

- **智能体** RL: 该领域专注于训练 LLM 在各种场景中使用外部工具，如搜索引擎 [Jin et al., 2025b]、Python 解释器 [Feng et al., 2025a]、网页浏览器 [Li et al., 2025f] 等。主要框架如 veRL [Sheng et al., 2025] 和 AReaL [Fu et al., 2025b] 已更新或专门设计以支持这些能力。智能体 RL 的核心特征是异步生成和训练，这在 LLM 与外部环境的长期交互中显著减少计算时间。次要框架大多基于 veRL 构建以集成额外的工具和环境，它们的新特性逐渐被合并回 veRL。关于智能体 RL 的更多细节将在 § 6.1 和 6.2 中讨论。

- **多模态** RL: 虽然主要的开发框架最初是为训练语言模型而设计的，但它们通常基于 transformer，支持视觉语言模型的推理和训练。该领域的主要挑战涉及数据处理和损失函数设计。基于 veRL，已经开发了用于训练视觉语言模型的著名框架，如 VLM-R1 [Shen et al., 2025a] 和 EasyR1 [Zheng et al., 2025d]。对于多模态生成，一些框架专门为基于扩散的模型的 RL 训练而开发，如 DanceGRPO [Xue et al., 2025a]。然而，这些方法超出了本文的范围，读者可以参考最近专注于视觉模型的 RL 调查以获取更多细节 [Wu et al., 2025h]。关于多模态 RL 的更多细节将在 § 6.3 中讨论。

Figure 6 | 应用分类包括编程任务、智能体任务、多模态任务、多智能体系统、机器人任务和医疗应用. 具体工作包括软件工程、竞争性编程、仓库级代码生成、编程智能体、搜索智能体、浏览器使用智能体、深度研究、GUI 智能体、视觉理解、视觉生成、多智能体编程、协作智能体、模拟机器人、真实机器人和医疗理解等方向.

- **多智能体** RL: 智能体 RL 的框架主要专注于为异步展开和训练实现动态工作流。虽然大多数这些框架仍然局限于单智能体应用, 但基于 LLM 的 MARL 仍是一个积极探索的领域。Zhang et al. [2025e] 提出了第一个用于基于 LLM 的多智能体强化训练和推理的高性能开源框架，实现了集中式交互和分布式策略训练。此外，最近的框架如 Agent-Lightning [Luo et al., 2025e] 已经实现了训练和推理的解耦，使得支持多智能体训练更加容易。关于多智能体 RL 的更多细节将在 § 6.4 中讨论。

6. 应用

LLM 的 RL 进展最好通过其在各个领域的实际影响来理解. 在本节中, 我们回顾了将 RL 训练的语言模型应用于实际任务所取得的最新进展和挑战. 我们重点介绍了 RL 驱动的方法如何在编程任务 (§ 6.1) 中提高能力, 如何实现更自主和自适应的智能体行为 (§ 6.2), 以及如何将 LLM 扩展到跨文本, 视觉等的多模态推理 (§ 6.3). 此外, 我们还讨论了在多智能体系统 (§ 6.4), 机器人学 (§ 6.5) 和医学 (§ 6.6) 中的应用, 说明了每个领域的广泛潜力和独特要求. 我们在图 6 中提供了应用的整体分类以及相应的相关工作.

6.1. 编程任务

> Takeaways
>
> - 强化学习在竞争性编程和领域特定任务中推进了大语言模型的推理和代码生成能力, 推动向智能体化, 闭环编程方向发展.
> - 然而, 在大规模软件环境中的可扩展性, 跨任务泛化和鲁棒自动化仍然是开放的挑战.

最近, 大量研究表明强化学习在可验证任务中具有显著优势. 鉴于编程任务固有的可验证性和实际重要性, 强化学习已成为改进代码推理的核心方法, 并持续吸引大量关注. 为了系统性地回顾该领域, 我们根据任务复杂性和发展趋势, 将现有研究分为三个方向: 代码生成, 软件工程辅助和智能体编程, 从更简单的可验证任务向更复杂的自主智能体编程发展.

代码生成. 该方向的主要目标是生成正确且可执行的代码. 研究重点在于使用强化学习调整大语言模型的生成分布以满足多样化编程任务的需求. 在 DeepSeek-R1 展示强化学习在复杂推理方面的潜力后, 越来越多的研究将强化学习应用于代码生成.

- **竞争性编程**: 竞争性编程作为最早的基准测试之一, 激发了包括 Code-R1 [Liu and Zhang, 2025],Open-

R1 [Face, 2025],DeepCoder [Luo et al., 2025b],AceReason-Nemotron [Chen et al., 2025q],SkyWork-OR1 [He et al., 2025d] 和 AReaL [Fu et al., 2025b] 等研究, 这些研究在代码任务中复现了 DeepSeek-R1 的结果. 为解决强化学习训练不稳定和推理缓慢的问题,DeepCoder [Luo et al., 2025b] 和 SkyWork OR1 [He et al., 2025d] 采用分阶段强化学习训练, 逐步增加上下文长度以稳定学习过程;DeepCoder [Luo et al., 2025b] 和 AReaL [Fu et al., 2025b] 进一步采用异步展开来解耦训练与推理并加速学习. 为解决代码生成中缺乏显式抽象推理能力的问题,AR^2 [Yeh et al., 2025] 框架 (抽象推理的对抗强化学习) 通过 RLVR 迭代训练教师和学生模型. 除了使用自回归模型进行代码生成的尝试外,Dream-Coder [Xie et al., 2025f] 将 RLVR 训练范式整合到扩散模型中, 实现了更快的生成速度. 关于跨任务泛化,AceReason-Nemotron [Chen et al., 2025q] 观察到从数学推理任务到竞争性编程的正迁移效应.

- **领域特定代码**: 由于代码要求的领域特定差异, 强化学习越来越多地应用于专门任务. 在数据检索中,Reasoning-SQL [Pourreza et al., 2025],ReEX-SQL [Dai et al., 2025b] 和 CogniSQL-R1-Zero [Gajjar et al., 2025] 将 GRPO 算法应用于 Text-to-SQL 任务, 在相应基准测试中取得了显著性能. 在形式化证明中,Kimina-Prover [Wang et al., 2025d] 和 DeepSeek-Prover-v2 [Ren et al., 2025] 通过结合自然语言与 Lean 统一了非形式化和形式化证明, 而 StepFun-Prover [Shang et al., 2025] 开发了端到端工具集成训练管道,Leanabell-Prover-V2 [Ji et al., 2025a] 通过多轮验证器反馈直接优化推理轨迹, 进一步推进了强化学习在该领域能力. 在其他领域,MedAgent-Gym [Xu et al., 2025b] 为大规模轨迹生成提供了可执行编码环境以改进基于大语言模型的医疗推理;VeriReason [Wang et al., 2025s],Proof2Silicon [Jha et al., 2025b] 和 CodeV-R1 [Zhu et al., 2025f] 将 RLVR 扩展到电子设计自动化 (EDA) 领域, 加速了大语言模型驱动的硬件设计. 此外, 图表到代码生成使智能体能够处理结构化或视觉输入并将其转换为可执行代码, 展示了跨模态领域特定代码生成 [Chen et al., 2025e].

软件工程. 尽管在竞争性编程和领域特定任务方面取得了进展, 但这些研究往往难以达到真实软件开发环境的要求. 因此, 强化学习研究也关注实际软件工程, 包括代码修复, 质量优化和仓库级生成.

- **代码质量改进**: 自动代码修复和质量改进在保持功能的同时增强软件可靠性. 强化学习显著提高了修复效果和泛化能力, 使模型能够处理未见过的缺陷. RePaCA [Fuster-Pena et al., 2025] 通过思维链推理和基于 GRPO 的微调指导大语言模型来减轻 APR 补丁过拟合, 而 Repair-R1 [Hu et al., 2025a] 联合优化测试用例生成和修复, 减少对事后验证的依赖. 除错误修复外, 强化学习还提高了代码效率, 可维护性, 可读性和安全性.CURE [Wang et al., 2025q] 和 UTRL [Lee et al., 2025b] 通过编码器-测试器交互演进化代码和单元测试, 无需真实监督,Afterburner [Du et al., 2025a] 利用执行反馈, 将 pass@1 从 47%提高到 62%, 超越了人类水平效率.REAL [Yao et al., 2025b] 将程序分析和单元测试作为混合奖励来提高可扩展性和质量, 实现了无需人工干预的高质量代码生成.

- **仓库级代码生成**: 超越函数和片段级任务, 近期工作探索仓库级代码生成和维护, 强调跨复杂跨文件和跨模块依赖的一致性和可维护性.RLCoder [Wang et al., 2024c] 将检索增强生成 (RAG) 与强化学习结合来训练检索器并提高代码补全准确性.RepoGenReflex [Wang et al., 2024a] 进一

步引入反思机制来评估生成结果并提供反馈, 持续优化生成策略并改进泛化. 通过将强化学习与自动化测试和持续集成相结合, 这种方法使大语言模型优化与实际开发流程保持一致, 推进软件工程自动化.

6.2. 智能体任务

Takeaways

- 智能体强化学习实现了高级行为, 但面临环境内高计算成本和长展开时间带来的可扩展性问题.
- 异步展开和记忆智能体有助于减少延迟和管理上下文, 但进一步进展依赖于更好的训练数据.

工具使用被认为是语言模型的基本能力 [Schick et al., 2023]. 近期工作利用强化学习帮助大语言模型掌握工具并完成更复杂的问题 [Dong et al., 2025a, Team, 2025d]. 我们将它们分为**编程智能体**, **简单搜索智能体**, **浏览器使用智能体**, **深度研究**, **图形用户界面与计算机使用智能体**和**其他任务**.

编程智能体. 强化学习与智能体范式的结合将代码生成从单步输出推进到多轮交互和自主迭代, 赋予大语言模型执行和验证能力以实现闭环优化.

- **代码智能体**: 常见的做法是将强化学习集成到配备执行和验证能力的代码智能体中, 并在 SWE-Bench 等现实基准上进行评估.SWE-RL [Wei et al., 2025c] 将 GRPO 应用于补丁生成-执行-纠正循环, 实现持续策略优化并改进数学推理, 通用代码生成和跨领域任务.EvoScale (Satori-SWE) [Zeng et al., 2025b] 允许智能体在没有外部验证器的情况下自主增强补丁质量. 强化学习增强的模型如 Kimi-K2 [Team, 2025d],Qwen3-Coder 和 GLM-4.5 展现出更强的智能体行为, 促进了更大的自主性和可扩展性.Sinha et al. [2025] 研究了大语言模型中的长视界执行, 并证明了由于错误累积, 单步准确性的改进不一定能转化为成功的多步任务性能. 这些发展表明, 将强化学习与智能体编程结合正在推动从' 单步生成' 向' 自主迭代' 的转变.
- **工具集成推理**: 强化学习的另一个新兴应用是工具集成推理 (TIR), 它通过将自然语言推理与外部工具执行环境紧密耦合来增强大语言模型的代码推理能力. 这种方法使模型能够生成, 执行和验证中间代码或程序输出, 减少错误并提高可验证性. 代表性工作如 ARPO [Dong et al., 2025b],AutoTIR [Wei et al., 2025b],CoRT [Li et al., 2025b] 和 ToRL [Li et al., 2025q] 采用类似策略: 模型使用 SFT 或强化学习 (主要是 GRPO 或变体) 进行后训练, 输出被结构化 (例如 <code>...</code>) 以触发工具执行, 将结果反馈到推理循环中. Li et al. [2025v], Paprunia et al. [2025], Xue et al. [2025b] 通过改进小型大语言模型的工具使用能力, 稳定多轮推理以及奖励独立于最终答案的工具使用序列来扩展基于强化学习的工具集成推理. 这种紧密集成提供了明确的强化学习奖励信号, 指导模型产生逻辑一致的输出并通过可验证计算迭代改进它们. 此外, 自动形式化方法如 FormaRL [Huang et al., 2025d] 通过集成基于编译器的语法检查和大语言模型一致性评估, 用最少的标注数据将 TIR 扩展到基于 Lean 的形式化证明生成, 进一步提高了可靠性和正确性.

- **自动化机器学习编程**: 强化学习在自动化机器学习 (AutoML) 中显示出前景, 将代码智能体扩展为能够自主数据处理, 模型构建和优化的机器学习工程智能体 (MLE 智能体).MLE-bench [Chan et al., 2024] 评估机器学习智能体能力;MLE-STAR [Nam et al., 2025] 提出基于搜索和优化的机器学习工程智能体;ML-Agent [Liu et al., 2025s] 展示了强化学习驱动的自主机器学习工程.Yang et al. [2025e] 表明, 由经过强化学习训练的相对较小模型驱动的智能体可以超越使用更大但静态模型的智能体, 特别是在增强持续时间感知更新和环境仪器以提供更细粒度奖励信号的情况下.

简单搜索智能体. 大语言模型可以通过结构化提示, 多轮生成以及与在线搜索引擎 (如 Google) 或静态本地语料库 (如维基百科) 的集成来训练为搜索智能体 [Jin et al., 2025a,b, Song et al., 2025a]. 然而, 使用在线搜索引擎进行训练通常会产生大量 API 成本, 使这种方法过于昂贵. 为解决这一挑战,Sun et al. [2025a] 建议在具有搜索能力的大语言模型训练期间模拟搜索引擎, 显著降低成本同时保持甚至提高性能. 其他工作如 R1-Search++ [Song et al., 2025b] 和 SEM [Sha et al., 2025] 利用大语言模型的内部知识来减少训练预算, 同时产生更好的性能. 具体来说,SSRL [Fan et al., 2025c] 提出在完全模拟的环境中训练模型, 通过 Sim2Real 泛化可以无缝适应真实场景. 同时, 可以为特定应用开发多样化的奖励信号.Dao and Le [2025], Mei et al. [2025] 采用多样性奖励来鼓励全面而准确的信息收集.Wang et al. [2025w] 利用步骤级奖励来进一步增强搜索智能体的性能.S3 [Jiang et al., 2025d] 利用超越 RAG 的收益来用更少的数据实现更好的性能. 为了增强大语言模型在更具挑战性查询上的能力, 如 GAIA [Mialon et al., 2023] 和 BrowseComp [Wei et al., 2025a] 等基准测试中的查询,WebSailor [Li et al., 2025f] 从知识图构建训练数据, 使模型能够搜索和浏览开放网络环境来解决晦涩问题.WebShaper [Tao et al., 2025] 引入了一个形式化的数据构建框架, 旨在提高通用 AI 助手的问题解决能力.

浏览器使用智能体. 除了使用搜索引擎, 其他浏览器用户智能体也利用网页浏览.WebGPT [Nakano et al., 2021] 使用文本网页描述来训练模型使其具备浏览网站的能力.Web-RL [Qi et al., 2024] 采用课程策略和 ORM 将大语言模型转换为网络智能体.DeepResearcher [Zheng et al., 2025e] 利用另一个大语言模型在浏览时充当摘要器来帮助搜索过程.Vattikonda et al. [2025] 使用各种超参数自举训练学生模型以实现稳定训练和更好性能.WebAgent-R1 [Wei et al., 2025d] 提出多轮异步 GRPO 来训练端到端网页浏览智能体, 实现了强大的性能.WebDancer [Wu et al., 2025d] 进行 SFT 和强化学习以通过网络搜索和浏览实现深度信息搜索和多步推理. 此外, 其他任务也需要网络智能体, 例如学术浏览 [Zhou et al., 2025b].

深度研究智能体. DeepResearch 被引入用于从各种在线来源收集信息，帮助完成现实世界问题，例如报告生成。WebThinker [Li et al., 2025m] 使用迭代 DPO 训练，利用 LRM 的长 CoT 能力，使用深度网络浏览器和 LLM 编写者完成挑战性任务。Kimi-Searcher [AI, 2025] 识别了多智能体的困境，并自动构建密集的工具使用数据来端到端训练单个智能体模型，在 HLE [Prabhudesai et al., 2025] 上取得优异性能。Jan-nano [Dao and Vu, 2025] 通过多阶段 RLVR 消除冷启动或 SFT 的需求，分别关注工具调用、回答质量和扩展响应长度。MicroThinker [Team, 2025e] 使用 SFT 和 DPO 训练 Qwen3 [Wu et al., 2025a]，增强其在现实世界应用中的性能。最近，提出了 WebWatcher [Geng et al., 2025]，这是一个多模态深度研究模型，能够使用外部工具和视觉信息解决极其复杂的问题。Atom-Searhcer [Deng et al., 2025b] 利用 LRM 作为 PRM，在训练期间提供细粒度奖励信号，取得

更好的性能。ASearcher [Gao et al., 2025a] 将交互轮次扩展到 10 轮以上，以激发深度研究智能体的推理能力。WebExplorer [Liu et al., 2025h] 采用基于模型的数据合成方法构建高质量数据，取得更好的性能。SFR-DeepResearch [Nguyen et al., 2025] 赋予单个智能体最小轮次的工具使用，在较短轨迹下实现与长轨迹相当的性能。除了一般 QA 任务，MedResearcher-R1 [Yu et al., 2025a] 被提出用于解决临床问题。

图形用户界面与计算机使用智能体. UI-R1 [Lu et al., 2025g] 是首个将基于规则的强化学习应用于图形用户界面 (GUI) 任务的工作. 它引入了一种新颖的基于规则的动作奖励, 并使用小型人工策划的训练集进行优化. 基于这一实践,GUI-R1 [Luo et al., 2025d],GUI-Critic-R1 [Wanyan et al., 2025] 等工作 [Ai et al., 2025, Du et al., 2025b, Lin et al., 2025a] 仔细设计细粒度的基于规则的奖励, 针对 GUI 任务的特定目标, 如动作准确性、参数正确性和步骤级状态. GUI-G1 [Zhou et al., 2025h] 对先前方法进行了实证分析, 识别了长度偏差、难度偏差和易受奖励攻击等问题, 并重新制定了奖励归一化方案以缓解这些限制. 此外, 近期研究 [Gu et al., 2025, Shi et al., 2025c] 尝试从在线 GUI 环境获得反馈, 以更好地模拟真实世界操作条件. GUI-Reflection [Wu et al., 2025g] 和 UIShift [Gao et al., 2025b] 基于 UI 元素的变化推导二元奖励, 以指示动作成功或失败. Liu et al. [2025r] 提出两阶段训练范式, 明确增强规划和反思推理能力. ZeroGUI [Yang et al., 2025b] 引入了用于生成挑战性任务的自动化流水线, 并仅基于在线环境反馈估计奖励, 消除了人工标注的需求. 与上述步骤级方法不同, 应用端到端异步强化学习框架来训练移动 [Lu et al., 2025b,d, Ye et al., 2025b] 和计算机 [Lai et al., 2025] 使用智能体已成为增长趋势, 这些方法仅使用基于规则的任务级完成奖励来优化模型, 不需要步骤级奖励信号.UI-TARS [Wang et al., 2025f] 通过迭代训练和反思调整从错误中学习并适应未预见的情况. 为进一步发展,UI-TARS 2 [Qin et al., 2025] 通过端到端强化学习增强了在 GUI、游戏、代码和工具使用方面的能力.

其他任务. 除了搜索和 GUI 智能体，强化学习也已成功应用于各种其他智能体任务。例如，Jiang et al. [2025a] 通过利用历史性能指标（如点击率）作为奖励信号来指导基于强化学习的优化，从而改进广告文案生成。在电子商务领域，Shop-R1 [Zhang et al., 2025s] 引入了一个复合奖励函数，将内部模型 logits 与外部分层反馈结合，更好地模拟购物环境中类似人类的决策制定。对于自动驾驶，LaviPlan [Oh, 2025] 将感知视觉能力与上下文感知决策制定对齐，在动态条件下实现更鲁棒的导航。类似地，Drive-R1 [Li et al., 2025s] 旨在平衡复杂驾驶场景的推理和规划能力，改善战略和反应行为。在结构化数据交互中，OpenTab-R1 [Qiu, 2025] 采用两阶段训练框架来增强 LLM 在基于表格问答中的熟练度。此外，通用智能体模型如 Qian et al. [2025] 和 Team [2025d] 中展示的模型展示了掌握多个常用工具（例如计算器、API 和数据库）来解决多样化现实世界任务的能力，展示了 RL 在构建多功能、工具增强智能体方面的可扩展性。

6.3. 多模态任务

Takeaways

- 强化学习加强多模态模型以解决有限数据设置, 长视频推理以及数值或属性敏感的跨模态生成等挑战.
- 探索理解与生成的统一强化学习框架是当务之急.

强化学习的成功不仅体现在语言模型中, 还在多模态任务中推动了显著进展. 已经开发了特定的优化方法来增强空间感知 [Chen et al., 2025v, Su et al., 2025e] 和跨模态可控性 [Chen et al., 2025u, Wu et al., 2025h] 等能力. 下面, 我们从理解和生成方面讨论强化学习在多模态任务中的应用.

多模态理解. 与语言场景相比, 多模态理解需要强大的空间感知和跨模态语义对齐. 最近, 大量研究采用强化学习来增强跨图像, 视频和 3D 空间的推理能力, 在理解能力方面显示出显著改进.

- **图像理解中的强化学习**: Vision-R1 [Huang et al., 2025c],VLM-R1 [Shen et al., 2025a] 和 Visual-RFT [Liu et al., 2025y] 代表了首次将 DeepSeek-R1 风格的 RFT 从数学和代码领域扩展到多模态感知任务的尝试. 这些方法标志着训练范式的转变：从 SFT 中的数据扩展转向针对特定任务目标设计的可验证奖励函数的策略性设计。它们在多个检测和定位基准上取得了优异的性能，即使在有限的训练数据下也展现了强化微调（RFT）的高级泛化能力。随后，多个视觉推理模型 [Kan et al., 2025, Xia et al., 2025a] 采用了类似的思考-回答格式，试图通过试错来学习。这些方法通过结果奖励驱动的优化来增强推理能力，消除了对昂贵的逐步监督或 CoT 训练数据的需求。最近，Deepeyes [Zheng et al., 2025f]、CoF [Zhang et al., 2025n] 及其他方法 [Cao et al., 2025, Fan et al., 2025d, Su et al., 2025a] 已经超越了纯文本基础的 CoT，扩展到明确的多模态交错推理链。这些方法尝试使用现成工具 [Su et al., 2025d] 或图像生成模型 [Xu et al., 2025e] 来迭代识别图像中的感兴趣区域，从而实现更可解释的推理过程。其他方法 [Chu et al., 2025b, Chung et al., 2025] 通过在推理阶段复制和路由视觉 tokens 来实现隐式的多模态交错 CoT，从而缓解了长文本基础 CoT 中的幻觉问题。尽管取得了显著的成功，但仍有几个挑战需要解决：1）推理和回答不一致：模型生成的思考未能映射到最终答案。2）长链探索崩溃：随着响应长度的增加，模型变得脆弱并容易产生幻觉。3）对数据质量的敏感性：RL 样本选择至关重要，因为低质量训练数据可能导致性能不佳甚至负面优化。
- **视频理解中的强化学习**: 扩展视频理解能力以解释和推理动态视觉内容对于多模态理解至关重要。为了实现这一目标，Video-R1 [Feng et al., 2025b] 为视频多模态大语言模型（MLLM）引入了系统的强化学习框架，使用时序感知的 GRPO 算法（T-GRPO）来改进时空推理。ReAd-R [Long et al., 2025] 提出了一个通过基于规则的强化学习优化的框架，用于模拟人类启发式思维来进行广告视频理解。Focused Thinking [Dang et al., 2025] 采用 token 加权奖励方案，修剪冗长、通用的思维链，并使用分级（部分学分）奖励来增强视频推理。VQ-Insight [Zhang et al., 2025o] 设计分层奖励，结合通用的任务特定时序学习，为长视频定制 QA 过程。为了从第一人称视角理解人类日常生活，Ego-R1 [Tian et al., 2025] 通过强化学习训练工具思维链代理，通

过动态调用检索和视觉工具进行逐步推理，来处理超长自我中心视频（长度为数天或数周）。同样，LongVILA [Chen et al., 2025t] 的 Long-RL 框架构建了大型 LongVideo-Reason 数据集和专门的两阶段 CoT-SFT 和强化学习管道，具有序列并行性，使 MLLM 能够处理超长视频。为了自动化更多视频 CoT 数据创建，VideoRFT [Wang et al., 2025l] 使用 LLM 从丰富的视频描述符生成基本原理，通过 VLM 精炼，并引入语义一致性奖励来对齐文本推理与视觉证据。同时，VideoChat-R1 [Li et al., 2025o] 证明，有针对性的多任务强化学习微调可以在不降低一般聊天性能的情况下显著增强特定的时空技能。总的来说，这些研究为通过强化学习开发稳健和可泛化的视频推理铺平了道路。

- RL in 3D Understanding: 虽然 MLLM 通过 RL 在 2D 视觉理解方面取得了显著进展，但将其能力扩展到 3D 空间的视觉-空间理解仍然是一个具有挑战性的前沿领域 [Wu et al., 2025b, Yang et al., 2025c]。MetaSpatial [Pan and Liu, 2025] 采用多轮基于 RL 的优化机制，整合物理感知约束来增强 MLLM 中的空间推理。基于 GRPO [Shao et al., 2024]，Spatial-MLLM [Wu et al., 2025b] 和 SpaceR [Ouyang et al., 2025] 证明，即使是小规模模型也可以通过类似 R1-Zero 的训练 [Liao et al., 2025c] 来缩小与更大模型的性能差距。此外，RoboRefer [?] 将基于 RL 的空间推理扩展到具身设置，将推理根植于现实世界动力学中。

多模态生成. LLM 中 RL 的探索也已扩展到多模态生成。关于测试时缩放 [Liu et al., 2025b, Ma et al., 2025b, Singhal et al., 2025] 和 DPO [Black et al., 2024b, Liang et al., 2025d, Liu et al., 2025l, Tong et al., 2025a, Wallace et al., 2024] 的开创性研究推动了图像和视频生成中美学和文本保真度的显著进展。最近，越来越多的关注被投入到增强图像和视频生成中的推理能力 [Guo et al., 2025f, Jiang et al., 2025b]。

- **图像生成中的强化学习**: 扩散模型极大地推进了视觉生成 [Esser et al., 2024, Liu et al., 2023b, Rombach et al., 2022]，越来越多的研究将强化学习融入其中，通过将去噪步骤视为 CoT 轨迹来隐式执行推理 [Liu et al., 2025d, Pan et al., 2025b, Xue et al., 2025a]。然而，GRPO 在扩散模型中与常微分方程（ODE）采样之间存在固有冲突。具体来说，GRPO 依赖随机采样来估计优势，而 ODE 采样遵循确定性的去噪轨迹，这限制了推出样本的多样性。为了解决这个问题，采用 ODE 到 SDE 的转换 [Liu et al., 2025d, Wu et al., 2025a, Xue et al., 2025a] 来鼓励采样过程中的随机项。考虑到 SDE 的低效率，MixGRPO [Li et al., 2025e] 通过整合 SDE 和 ODE 设计了混合采样策略。此外，TempFlow-GRPO [He et al., 2025g] 明确利用基于流模型中的时序结构，实现更精确的信用分配和策略优化。最近，GPT-4o 展示了强大的文本保真度和编辑一致性 [OpenAI, 2024]，引发了对自回归模型可控性的兴趣。基于大规模图像-文本训练数据，SimpleAR [Wang et al., 2025j] 直接应用 GRPO 进行后训练，并在高分辨率图像生成中取得了卓越的性能。为了加强对细粒度属性（如空间关系和数值一致性）的遵循，FocusDiff [Pan et al., 2025e] 构建了仅在细微属性变化上不同的成对数据集，并使用它们来训练生成模型。此外，RePrompt [Wu et al., 2025f] 将额外的多模态理解模型纳入图像生成框架，并使用 GRPO 训练它来精炼提示。同时，T2I-R1 [Jiang et al., 2025b]、GoT-R1 [Duan et al., 2025] 和 ReasonGen-R1 [Zhang et al., 2025t] 在单个模型中统一了提示精炼和图像生成，利用 GRPO 进行联合优化。

- RL in Video Generation: 与图像生成相比，将 RL 扩展到视频生成在时序连贯性和物理现实主义方面提出了更大的挑战。DanceGRPO [Xue et al., 2025a] 在 HunyuanVideo [Kong et al., 2024] 上进行后训练，并使用 VideoAlign [Liu et al., 2025e] 基于视频美学、运动质量和文本-视频一致性提供奖励。此外，InfLVG [Fang et al., 2025b] 使用 GRPO 根据上下文相关性指导 token 选择，从而实现语义一致且时序连贯的长视频生成。另外，Phys-AR [Lin et al., 2025b] 引入速度和质量作为球运动场景的可验证奖励，显著增强了视频生成的物理现实主义。

目前，几个 ULM 模型采用统一框架同时优化多模态理解和生成。为此，提出了从文本到图像和从图像到文本的双向 [Jiang et al., 2025c] 和双重 [Hong et al., 2025c] 奖励，以增强生成和理解能力。对于多模态理解，Deepeyes 和 CoF 已尝试使用生成模型或外部工具来实现多模态 CoT。对于多模态生成，使用精炼文本作为 CoT 也依赖于多模态理解能力。因此，探索多模态理解和生成的统一后训练方法是未来研究的紧迫任务。从特定领域的角度来看，代码生成可以作为文本和图像生成之间的桥梁。应用 RL 来促进模型对复杂图表进行推理，并为特定领域图像生成产生结构化代码 [Chen et al., 2025e,f, Tan et al., 2025b] 是一个有前景的应用。

6.4. 多智能体系统

Takeaways

- 在多智能体系统 (MAS) 中改进协作, 推理和信用分配很重要, 能够在复杂任务上实现更稳定有效的团队合作.
- 在开发高效的协作和交互机制以充分释放集体能力并进一步提高智能体性能方面仍然存在关键挑战.

目前，基于 LLM 推理的 RL 研究主要集中于单模型，而将 RL 应用于 MAS 已成为一个突出的前沿研究方向。本节首先概述传统 RL 和多智能体 RL（MARL）的基本概念，重点介绍其主要挑战。此外，本节讨论 LLM 在 MARL 中的创新应用，强调它们在信息共享和信用分配方面的优势。最后，审视了 MAS 中集成 RL 与 LLM 的最新进展，重点关注如何利用 RL 来增强智能体之间的协作和策略优化，从而促进多智能体推理能力的发展。

传统多智能体强化学习. 近年来，作为复杂的分布式智能系统，MAS 在 RL 领域引起了广泛关注 [Dorri et al., 2018]。传统 MARL [Busoniu et al., 2008] 主要关注在共享环境中多个智能体的交互和联合学习，以实现全局目标。传统 MARL 的主要挑战包括信用分配的复杂性、环境的非平稳性以及智能体之间通信和协作的效率 [Canese et al., 2021]。为了解决这些问题，研究者提出了集中式训练与分布式执行（CTDE）范式 [Lowe et al., 2017]，其中智能体在训练阶段共享全局信息进行策略优化，而执行期间的决策仅依赖于局部观测。基于 CTDE 范式，研究者引入了基于价值的方法（如 VDN [Sunehag et al., 2017] 和 QMIX [Rashid et al., 2020]）、基于策略梯度的方法（如 MADDPG [Lowe et al., 2017]）和演员-评论员方法（如 COMA [Foerster et al., 2018]）。此外，由于 PPO 被认为是传统 RL 中的 SOTA，MAPPO 在一些简单协作任务中也显示出令人惊讶的效果 [Yu et al., 2022]。然而，随着智能体数量增加和任务复杂度提高，传统 MARL 方法在样本效率和可扩展性方面面临重大

挑战。为了解决这个问题，学者们考虑在与所有智能体的交互中用邻近智能体替换当前智能体（如 MF-MARL [Yang et al., 2018]），这有效缓解了 MARL 中智能体数量增加导致的维度灾难问题。然而，它仍然无法有效应用于需要多个智能体同时协作的复杂任务场景。

多智能体强化学习中的 LLM. LLM 的快速发展在解决 MARL 内的挑战方面显示出巨大潜力。利用其强大的自然语言理解和生成能力，LLM 可以为 MAS 提供有效的信息共享机制。例如，在 MARL 的信用分配问题中，研究者利用 LLM 设计直观的奖励分配机制，从而提高信用分配的准确性和可解释性。Zhang et al. [2023b] 通过使 LLM 实时推断每个智能体的意图并生成下一个协作计划，显著改善了稀疏奖励场景中的多智能体协作效率。Ding et al. [2023] 利用 LLM 将自然语言任务描述解析为可执行的实体级子目标，从而实现奖励塑形和策略共享，有效缓解了 MARL 中的信用分配问题。Li et al. [2023a] 利用 LLM 的"心智理论"能力，让智能体对队友的潜在策略生成语言信念，从而在多智能体协调中实现更准确的决策。

基于 LLM **的多智能体系统中的** RL. 在将 RL 与 LLM 集成的背景下，基于 LLM 的 MAS 研究逐渐成为热点。相关研究主要关注如何充分利用 LLM 的语言理解和生成能力，同时利用 RL 实现多个智能体之间的高效协作和策略优化。LLaMAC 和 CTRL 等框架将 LLM 与演员-评论员架构集成。LLaMAC [Zhang et al., 2023a] 采用集中式 LLM-Critic 为多个 LLM-Actor 提供基于自然语言的价值反馈，从而促进多个智能体之间的协作学习。CTRL [Xie et al., 2025e] 使用合成数据训练 LLM 进行"自我批评"，并通过 RL（如 GRPO）迭代优化模型输出，这可以在无需人工标注的情况下提高测试时性能。

在大规模多智能体协作场景中，MAPoRL [Park et al., 2025] 通过联合训练多个 LLM 并引入推理感知奖励，促进多轮任务中的高效和可迁移协作。MAGRPO [Liu et al., 2025o] 将 LLM 协作建模为协作式多智能体 RL 问题，提出了群体级相对策略优化机制，显著提高了写作和代码生成等任务中多轮联合输出的质量。ReMA [Wan et al., 2025] 引入了高层智能体和底层智能体的双 LLM 结构，通过交替冻结和更新策略实现元思维和推理能力的协同增强。JoyAgents-R1 [Han et al., 2025] 设计了联合进化训练过程，通过交替全局经验回放和个体 PPO 更新，在开放域问答任务中促进异构 LLM 团队内的多样性和一致性。AlphaEvolve [Novikov et al., 2025] 设计了进化优化机制来协调多 LLM 协作。通过直接修改代码并持续接收评估反馈，MAS 增强了处理复杂编码任务的能力。AutoAgents [Chen et al., 2023a] 通过动态生成针对任务需求的专门智能体并纳入观察者角色进行反思和改进，显著增强了 MAS 在复杂任务中的适应性和问题解决能力。

6.5. 机器人任务

Takeaways

- 强化学习通过将大语言模型风格的方法适应视觉-语言-动作 (VLA) 模型来解决机器人学中的数据稀缺和泛化挑战.
- 允许 VLA 从环境交互和简单奖励中学习, 最近的强化学习方法 (如 GRPO,RLOO,PPO) 在最小监督下实现了卓越性能和新颖行为.

机器人任务中的 RL. 强化学习已在机器人学中得到广泛应用，主要集中在三个领域：机器人控制、视觉-语言导航 (VLN) 和机器人操作任务。传统机器人控制中的强化学习研究已经成熟，并得到广泛应用，例如类人机器人的动作生成 [Peng et al., 2018]、稳健的四足运动执行 [Hwangbo et al., 2019] 和灵巧手部操作 [Chen et al., 2023b]。同样，VLN 任务也取得了显著进展 [Anderson et al., 2018, Wang et al., 2018, 2019]。然而，这些领域在模型架构、规模、任务类型、奖励函数设计、优化目标和算法方法方面与基于 LLM 的强化学习存在实质性差异，因此不在本综述的范围内。

机器人操作任务使机器人能够在现实世界环境中解决多样化的操作问题，代表了具身智能最具挑战性和基础性的方面 [Firoozi et al., 2025]。这些任务不仅需要对视觉和文本信息的全面理解以及精细的运动控制，还需要物理推理、长时程规划和逻辑推理能力。利用 LLMs 和 VLMs 卓越的文本和视觉处理能力，多项研究探索将这些模型作为核心组件与操作模块结合用于操作任务，如 Robot-Brain [Ji et al., 2025b] 和 RT-2 [Zitkovich et al., 2023]。

视觉-语言-动作模型. 最近，视觉-语言-动作 (VLA) 模型通过统一的端到端训练将 VLM 骨干与动作模块集成，已成为最有前景的解决方案并成为机器人操作的主流方法 [Zhong et al., 2025]。当前的 VLA 模型遵循两阶段范式 [Sapkota et al., 2025]：在多模态数据上预训练 (如 Open X-Embodiment [O'Neill et al., 2024])，然后在遥操作机器人轨迹上进行监督微调。然而，这种模仿学习范式存在关键局限性：其性能严重依赖于高质量轨迹数据，这些数据的收集成本高昂且效率低下，且生成的模型对未见场景的泛化能力较差。鉴于 VLA 与 LLM 在架构、规模和方法论上的相似性 [Zhong et al., 2025]，将 LLM 风格的 RL 方法适应于 VLA 训练为解决数据稀缺和泛化挑战提供了有前景的方向。

将 DeepSeek-R1 的 RL 方法论应用于 VLA 需要解决几个挑战：1) 与单轮完成任务的 LLM 不同，VLA 需要多轮环境交互来生成完整轨迹；2) VLA 在连续动作空间中操作；3) 传统 RL 方法依赖于手工制作的过程奖励，限制了可扩展性。最近的工作包括 SimpleVLA-RL [SimpleVLA-RL Team, 2025]、VLA-RL [Lu et al., 2025c]、VLA RL Generalization [Liu et al., 2025f]、RIPT-VLA [Tan et al., 2025a] 和 ConRFT [Chen et al., 2025s] 开创了 DeepSeek-R1 方法论在 VLA 训练中的应用。

SimpleVLA-RL [SimpleVLA-RL Team, 2025] 使 VLA 模型能够与环境交互来推出多样化的完整轨迹，采用二元成功/失败奖励作为监督信号，并使用 GRPO 算法训练 OpenVLA-OFT [Kim et al., 2025]。仅凭单条演示轨迹，这种 RL 方法在 LIBERO 和 RobotWin2.0 基准测试中超越了最先进的 VLA 模型如 π_0 [Black et al., 2024a]，实现了 SOTA 性能并在真实机器人实验中超越了先进的 RDT 模型。此外，作为 π_0 的升级版本，$\pi_{0.5}$ [Intelligence et al., 2025] 使用来自不同场景和来源的多模态机器人数据进行异构训练，使 VLA 在可泛化的现实世界机器人操作任务中提供了新的里程碑。类似于 DeepSeek-R1 的"顿悟时刻"，RL 训练的 VLA 也发现了新颖的行为模式。VLA RL Generalization [Liu et al., 2025f] 研究了 RL 对 VLA 泛化能力的影响，证明在未见环境、物体和纹理方面相比 SFT 有显著改进，同时比较了 GRPO 和 PPO 的有效性。RIPT-VLA [Tan et al., 2025a] 采用 RLOO [Ahmadian et al., 2024] 进行 VLA RL 训练。RLinf [Team, 2025h] 为 VLA RL 设计了一个灵活、可扩展的 RL 框架，统一了渲染、推理和训练，提高了 VLA 训练效率和性能。ConRFT [Chen et al., 2025s] 通过交替的 RL 和 SFT 轮次迭代训练 VLA，通过多次迭代逐步提升性能。

强化学习的数据效率、改进的泛化能力和最小监督要求有效地解决了 VLA 当前的数据稀缺和

泛化能力差的挑战。通过允许 VLA 仅通过结果监督来自主探索和从试错中学习，这种方法与复杂且昂贵的遥操作数据收集相比，大幅降低了实施成本。此外，强化学习的数据效率消除了对大规模昂贵轨迹数据集的需求，实现了可扩展的 VLA 后训练能力。

However, current VLA RL research remains primarily simulation-based. While SimpleVLA-RL [SimpleVLA-RL Team, 2025] achieved real-world deployment through Sim2Real transfer [Chen et al., 2025n], few works have yet deployed physical robots to collect real-world trajectories for RL. In addition, research on VLA RL is also limited by the current development of RL in robotics, including but not limited to sample efficiency, reward sparsity, and sim2real. Key challenges include autonomous sampling on physical robots requiring multiple devices for efficiency, continuous manual resetting and annotation.

6.6. 医疗任务

Takeaways

- 医疗大语言模型的强化学习面临独特挑战: 可验证任务允许稳定的奖励设计, 而非可验证任务使奖励定义困难.
- 可验证任务使用基于规则的奖励的 SFT+RL; 非可验证任务利用 DPO, 评分标准, 课程强化学习或离线强化学习, 尽管可扩展性和稳定性仍然是开放问题.

医疗 LLM 的 RL 优化通常旨在增强推理和泛化能力，通常采用 SFT 后接 RL 的两阶段流程。现有工作大致可分为**基于规则奖励的可验证问题**和**基于生成或评分标准奖励的非可验证问题**。

医疗理解. 这些任务，如多选题 QA、结构化预测、临床编码或视觉定位，允许使用确定性奖励，使其成为医疗大语言模型中强化学习最成熟的领域。典型范式是 SFT 后接强化学习的两阶段流水线，其中 GRPO 等算法直接针对基于正确性的信号优化模型。例如，HuatuoGPT-o1 [Chen et al., 2024a] 通过医疗验证器合成可靠的推理轨迹数据，并使用 SFT 和强化学习训练模型来增强推理能力。Med-U1 [Zhang et al., 2025l] 采用混合二元正确性奖励与长度惩罚，确保准确性和格式合规性，而 MED-RLVR [Zhang et al., 2025j] 将可验证奖励应用于多选题 QA，改进 OOD 泛化。Open-Medical-R1 [Qiu et al., 2025] 证明了仔细的数据过滤提高了强化学习的效率。Gazal-R1 [Arora et al., 2025] 设计了多组件奖励系统，通过 GRPO 改进准确性、格式遵守和推理质量，以增强医疗推理。ProMed [Ding et al., 2025] 将医疗大语言模型从被动范式转向主动范式，其中大语言模型可以在决策前提出临床有价值的问题，在 MCTS 引导的轨迹探索和强化学习期间使用 Shapley 信息增益奖励。MedGR^2 [Zhi et al., 2025] 引入了生成式奖励学习框架，创建了自我改进的良性循环，共同开发数据生成器和奖励模型，实现 SFT 和强化学习训练的高质量多模态医疗数据的自动化创建。

超越文本 QA，最近的模型将基于规则的奖励扩展到视觉和多模态任务。MedVLM-R1 [Pan et al., 2025d] 采用 RL 框架，通过格式和准确性奖励激励模型发现人类可解释的推理路径，而不使用任何推理参考。MedGround-R1 [Xu and Nie, 2025] 为医学影像定位任务引入了空间-语义奖励，结合了空间准确性奖励和语义一致性奖励。ARMed [Liu and Wei, 2025] 通过自适应语义奖励解决了开放

式医疗 VQA 中的奖励崩溃问题，该奖励在训练期间根据历史奖励分布动态调整语义奖励。Liu and Li [2025] 利用基于规则的格式和匹配奖励指导医疗视觉信息提取的结构化 JSON 生成，仅使用 100 个注释样本。MMedAgent-RL [Xia et al., 2025b] 是基于 RL 的多智能体框架，实现医疗智能体之间的动态和优化协作。MedGemma [Sellergren et al., 2025] 经过 RL 后训练，并在 MedXpertQA [Zuo et al., 2025a] 上进一步评估，这是一个专家级医疗多选题基准，包括用于评估推理模型的子集。

对于其他临床应用，DRG-Sapphire [Wang, 2025] 将 GRPO 与基于规则的奖励应用于诊断相关分组。EHRMIND [Lin and Wu, 2025] 结合 SFT 预热和 RL VR，使用电子健康记录 (EHR) 数据进行复杂临床推理任务，包括医学计算、患者试验匹配和疾病诊断。ChestX-Reasoner [Fan et al., 2025e] 整合来自临床报告的过程奖励，训练模型模拟放射科医生的逐步推理。CX-Mind [Li et al., 2025k] 采用 SFT 和 RL，结合格式、结果和过程奖励，训练胸部 X 光诊断的交错推理。为了实现基于代码的医疗推理基准测试，MedAgentGym [Xu et al., 2025b] 提出了医疗智能体代码生成的基准，并证明 RL 可以改进这种推理能力。

医疗生成. 这些任务包括放射学报告生成 [Jing et al., 2025]、多轮临床对话 [Bani-Harouni, 2025]、治疗规划 [Nusrat, 2025] 和诊断叙述 [Yooseok Lim, 2025]，它们缺乏唯一的真实答案。因此，基于规则的奖励不直接适用。虽然 DPO 已被应用于改进医疗 LLM 在偏好对齐生成任务上的表现 [Yang et al., 2025i, Yu et al., 2025c]，但在非可验证任务上的大规模 RL 正在兴起但仍相对未被充分探索。例如，DOLA [Nusrat, 2025] 将 LLM 代理与商业治疗规划系统集成，包含一个奖励函数来引导目标覆盖和危险器官保护之间的权衡，以优化治疗方案生成。LA-CDM [Bani-Harouni, 2025] 提出了一个通过混合训练范式的双代理结构，将监督微调与 RL 结合以平衡诊断准确性、不确定性校准和决策效率。在诊断对话中，PPME [Sun et al., 2025i] 开发了一个即插即用框架，使用大规模 EMR 和混合训练来通过专门的询问和诊断模型增强 LLM 交互诊断能力。在临床决策支持中，MORE-CLEAR [Yooseok Lim, 2025] 将多模态离线 RL 应用于脓毒症治疗策略，改进 MIMIC-III/IV 中的生存预测决策制定。对于放射学报告生成，BoxMed-RL [Jing et al., 2025] 在其预训练阶段利用 RL，使用格式奖励和交并比 (IoU) 奖励来确保生成的报告对应于像素级别的解剖学证据。Baichuan-M1 [Inc., 2025a] 采用三阶段 RL 方法：ELO(探索性对数似然优化) 增强思维链推理多样性，TDPO(令牌级直接偏好优化) 解决长度依赖约束，最后使用带奖励模型反馈的 PPO 进行策略优化。Baichuan-M2 [Inc., 2025b] 引入了一个新颖的动态验证框架，超越了静态答案验证器，建立了大规模、高保真度的交互式强化学习系统，配备患者模拟器和临床评分标准生成器用于真实的临床环境。

总体而言，医疗大语言模型中的强化学习在可验证问题上已经成熟，其中确定性正确性允许基于规则的奖励和稳定的 GRPO 训练。相比之下，面向生成的任务仍然具有挑战性：当前解决方案采用基于评分标准的奖励、课程迁移或离线强化学习来近似质量信号。在非可验证任务上可扩展强化学习的稀缺性突显了构建可信的、具有推理能力的医疗基础模型的关键未来方向。

7. 未来方向

虽然 LLM 的 RL 取得了显著进展，但许多基本挑战和机遇仍然存在. 本节概述了几个有望塑造该领域下一波进展的有前景的方向. 我们强调了持续 RL 对于适应不断变化的数据和任务的重要性 (§ 7.1),

基于记忆和基于模型的 RL 对于增强推理能力的重要性 (§ 7.2 和 § 7.3), 以及教授 LLM 高效和潜在空间推理的新兴方法 (§ 7.4 和 § 7.5). 我们还讨论了在预训练期间利用 RL(§ 7.6), 将 RL 应用于基于扩散的架构 (§ 7.7) 和推动科学发现 (§ 7.8) 的前沿. 最后, 我们考虑了架构-算法协同设计的挑战和前景, 以满足越来越大和高效率智能模型的需求 (§ 7.9). 通过综述这些方向, 我们旨在为 LLM 的 RL 未来研究提供路线图和灵感.

7.1. LLM 的持续 RL

为了在基于 RL 的后训练期间增强 LLM 的多域性能, 主流方法是混合不同任务的数据并进行统一训练 [Guo et al., 2025a, Yang et al., 2025a]. 在合成数据上 [Chen et al., 2025d, Liu et al., 2025g], 多阶段 RL 已被证明表现比混合数据训练更差, 甚至在 RL 中增加难度的课程学习可能不必要 [Xie et al., 2025c]. 然而,Chen et al. [2025d] 表明跨不同任务的多阶段 RL 在泛化到困难或未见问题方面具有优势. 尽管关于多阶段 RL 有效性的争论仍在继续，但随着该领域朝着构建必须在动态环境中适应不断发展的数据和任务的 AI 系统前进，探索 LLM 的持续强化学习（CRL）变得必要。

与传统的 CRL 类似，LLM 在多阶段 RL 训练过程中面临平衡稳定性和可塑性的根本挑战 [Pan et al., 2025a]。可塑性对 LLM 来说可能特别值得关注，因为广泛使用的深度学习技术在持续学习设置中可能导致大型模型的表现不比浅层网络更好 [Dohare et al., 2024]。LLM 的 CRL 的另一个挑战在于 LLM 中知识和推理的纠缠性质，这区别于传统的 RL 设置，在传统设置中任务可以离散定义，策略可以模块化组织，例如在游戏类环境 [Chevalier-Boisvert et al., 2023, Towers et al., 2024] 或具身体验场景 [Todorov et al., 2012, Wołczyk et al., 2021] 中。

传统 CRL 研究中现有的方法论框架为解决 LLM 特定需求提供了有希望的基础。传统 CRL 研究的核心方法论见解，包括经验回放 [Berseth et al., 2021, Li et al., 2021, Rolnick et al., 2019]、策略重用 [Garcia and Thomas, 2019, Gaya et al., 2022] 和奖励塑形 [Jiang et al., 2021, Zheng et al., 2022]。开发专门针对 LRM 的 CRL 框架仍然是一个有价值的研究方向。为 LLM 或 LRM 开发专门的 CRL 技术对于创建更具适应性和效率的 AI 系统至关重要，这些系统能够终身学习并在动态和不断变化的环境中运行。

7.2. 基于记忆的 LLM 强化学习

尽管许多智能体 RL 工作已经探索了记忆机制，从外部长期存储和插入 [Chhikara et al., 2025, Xu et al., 2025d, Zhong et al., 2024] 到内部记忆处理和工作记忆控制 [Yu et al., 2025b, Zhou et al., 2025i]，但大多数设计仍然针对当前任务定制，超出该任务的泛化能力有限。正如 Silver and Sutton [2025] 强调的那样，下一代智能体将主要从经验中学习，通过持续交互获得技能。本着这种精神，一个关键方向是将智能体记忆从特定任务的缓冲区转变为结构化、可重用、可跨不同任务迁移的经验存储库，让记忆演变为更广泛适应性和终身学习的基础。这种以经验为中心的观点也与 RL 自然契合，因为智能体与环境交互产生的数据提供了丰富的经验轨迹，可以被有效利用。此外，尽管最近的工作已经探索了维护共享经验池以从过去历史中检索相关策略并将其他智能体的经验适应新任务场景 [Tang et al., 2025]，但这个方向仍然探索不足。这里的一个核心挑战是通过 RL 使智能体自动

学习如何操作和管理记忆，跨任务组合和泛化经验知识。解决这一挑战对于迈向"经验时代"至关重要，在这个时代中，集体交互轨迹将成为更广泛智能体智能的基础。

7.3. 基于模型的 LLM 强化学习

RL 的一个核心挑战在于获得可扩展和稳健的奖励信号以及从环境中获得有意义的状态表示。先前的工作已经研究了世界模型的构建 [Luo et al., 2024, Moerland et al., 2023]，为 RL 智能体提供信息丰富的状态，最近 LLM 也被用作各种 RL 背景下的世界模型 [Benechehab et al., 2024, Gu et al., 2024, Hu and Shu, 2023]。在 LLM 的 RL 情况下，特别是对于语言智能体，构建准确捕捉环境状态并生成可靠奖励的世界模型的能力至关重要。最近的进展表明，生成式世界模型，包括那些通过视频预训练增强的模型 [Assran et al., 2025, Ball et al., 2025, Bruce et al., 2024]，既实用又有效。然而，将世界模型与基于 LLM 的智能体的 RL 无缝集成仍然是一个开放的研究问题。因此，基于模型的 LLM RL 正在成为未来研究中一个特别有前途和可扩展的方向。

7.4. 教授 LRM 高效推理

推理时扩展提高了 LRM 在困难任务上的准确性，但它也引入了系统性的过度思考（对简单实例进行不必要地长的推理链）[Chen et al., 2024b, Qu et al., 2025a, Sui et al., 2025, Yan et al., 2025b] 和在激进截断下的思考不足（过早停止和依赖脆弱的捷径）[Su et al., 2025b, Wang et al., 2025t]。LLM 的 RL 的一个核心挑战是开发计算分配策略，使推理的深度和停止适应实例难度和认知不确定性。当前研究已经探索了提示中硬编码的推理级别 [Agarwal et al., 2025a, Wen et al., 2025a, Zhu et al., 2025g]、基于自适应长度的奖励塑形 [Liu et al., 2025p, Yuan et al., 2025a] 以及在损失函数中使用长度惩罚 [Aggarwal and Welleck, 2025, Xiang et al., 2025]。

然而，将这些方法推广为原则性的成本-性能权衡仍然是一个开放问题 [Gan et al., 2025]。教导 LRM 成为资源理性的，只有当边际效用证明其合理性时才进行更长时间的推理，仍然是语言推理中 RL 的一个核心未解决问题。

7.5. 教授 LLM 潜在空间推理

CoT[Wei et al., 2022] 通过提示模型阐述中间步骤来鼓励逐步推理，提高了可解释性和准确性。最近的研究结合了 CoT 和 RL 来进一步提高推理质量，在回答前对长形式思维进行采样以进行建模训练 [Guo et al., 2025a]。然而，当前的实现通常依赖于离散标量空间中的 token 级采样 [Cui et al., 2025a, Ouyang et al., 2022, Rafailov et al., 2023]，这可能成为瓶颈，因为在连续空间中丢失了有意义的语义信息 [Hua et al., 2024]。一个最近提出的方法，称为潜在空间推理（LSR）[Arriola et al., 2025, Geiping et al., 2025, Hao et al., 2024]，可能对 RL 优化更友好。LSR 在 LLM 的连续潜在空间中运行推理，促进更细致和流畅的语义推理。这一特性有助于更平滑的学习动态和与 RL 技术的更好集成。RL 和 LSR 的结合对于未来开发更强大和适应性更强的推理模型具有巨大潜力。然而，评估连续潜在思维的质量比评估基于 token 的思维更具挑战性。这将使提供准确的监督信号（如奖励和优势）变得复杂，这将成为 LSR 和 RL 组合的一个开放挑战。

7.6. LLM 预训练中的强化学习

传统预训练依赖于大型文本语料库和下一个 token 预测，扩展这种范式已被证明是基础模型发展的核心 [Brown et al., 2020, Kaplan et al., 2020]。新兴研究现在探索将 RL 更早地移到流程中，不仅在后训练中应用，而且在预训练本身期间也应用。例如，强化预训练 [Dong et al., 2025c] 将下一个 token 预测重新概念化为一个 RL 问题，具有从语料库派生的可验证奖励，报告了随着可用计算增加而持续增长的收益，从而将 RL 定位为预训练的一个有前途的扩展策略。

与此同时，诸如 avataRL [tokenbender, 2025] 等开放倡议展示了纯粹使用 RL 从随机初始化训练语言模型，自举 token 级奖励并使用迭代"裁判"评分，从而说明了从头开始的 RL 训练的具体路径。这与转世 RL 范式 [Agarwal et al., 2022] 一致，其中利用先前获得的计算知识（预训练的 critic）而不是从头开始训练。这些发展突显了一个实际问题：*如何使 RL 风格的预训练在大规模上具有成本效益？*解决这一挑战可能需要减少验证器负担和与奖励工程相关的成本，这似乎是基于 RL 的预训练扩展的关键。此外，这一研究方向与 § 3.1.4 中引入的无监督奖励设计密切相关，提出了关于*如何获得既可扩展又可靠的奖励*的重要问题。

7.7. 基于扩散的 LLM 强化学习

扩散大语言模型（DLLM）[Cheng et al., 2025c, Labs et al., 2025, Nie et al., 2025, Tae et al., 2025, Xie et al., 2025f, Ye et al., 2025c] 代表了语言生成中的新兴范式。与自回归（AR）模型相比，DLLM 提供优势，包括卓越的解码效率和通过多轮扩散进行自我纠正的更大潜力。初步努力已经开始探索 DLLM 的 RL[Borso et al., 2025, Gong et al., 2025, Yang et al., 2025d]，但几个关键问题仍然未解决。

将 RL 应用于 DLLM 的核心挑战在于准确且高效地估计采样响应的对数概率。这是由于自回归模型和扩散语言模型在本质上建模样本似然的方式存在根本差异。AR 模型通过下一个 token 预测生成序列，并通过链式法则分解联合概率，实现直接的从左到右采样。然而，DLLM 通过最大化证据下界（ELBO）来近似似然优化。ELBO 涉及对扩散时间步和掩码数据的双重期望，通常需要大量采样来实现准确估计；否则，它在偏好优化期间引入高方差。尽管 [Zhao et al., 2025c] 中的一步估计器和 [Zhu et al., 2025b] 中的采样分配策略等方法已经被提出来缓解方差，但对于策略学习来说，高效且准确的 ELBO 估计仍然是一个开放问题。

此外，DLLM 中存在多个可行的解码轨迹，这引入了一个额外的研究维度：利用 RL 引导模型朝向最优采样轨迹。这需要为中间去噪步骤设计有效的奖励函数。例如，He et al. [2025c] 将去噪公式化为多步决策问题，并将奖励模型应用于中间状态，[Wang et al., 2025p] 提出了一个基于扩散的值模型，计算前缀条件、逐 token 的优势以实现轨迹级奖励，而 Song et al. [2025c] 利用基于编辑距离的奖励来最大化解码效率。未来的工作也可能从计算机视觉中为连续扩散模型开发的 RL 技术中汲取灵感 [Black et al., 2024b, Xue et al., 2025a, Yang et al., 2024b]，可能为统一的多模态框架铺平道路。

7.8. 科学发现中的 LLM 强化学习

最近的研究表明，涉及 RL 可以提高 LLM 在推理密集型科学任务上的性能，在某些情况下甚至让它们超越专门方法 [Fallahpour et al., 2025, Fang et al., 2025c, Narayanan et al., 2025, Rizvi et al., 2025]。在生物学和化学等领域，RL 的一个核心挑战是大规模执行结果验证，这个过程传统上依赖于湿实验室实验。几种现有方法专注于替代或补充实验验证：Pro-1 [Hla, 2025] 使用 Rosetta 能量函数作为优化蛋白质稳定性的奖励函数，而 rbio1 [Istrate et al., 2025] 使用生物模型和外部知识源验证基因扰动结果预测。

在奖励制定和改进 oracle 模型本身方面仍有很大的探索空间。与此相关的是构建支持快速实验-反馈循环的合适 RL 环境的更广泛问题。诸如 Coscientist [Boiko et al., 2023] 和 Robin [Ghareeb et al., 2025] 等智能体系统通过实验室循环验证取得了成功，但这种稀疏、延迟和高成本的反馈信号对于直接训练底层 LLM 来说是不切实际的。实验环境的计算机模拟，例如细胞级别的扰动响应预测 [Bunne et al., 2024, Noutahi et al., 2025]，代表了前进的潜在路径。然而，由于范围有限以及严重缺乏准确性和可推广性 [Ahlmann-Eltze et al., 2025, Kedzierska et al., 2023]，许多这些系统远不足以替代现实的实验室环境。其他研究路线已经探索将领域特定模型纳入 LLM 训练以处理科学数据 [Fallahpour et al., 2025]，以及开发能够执行一系列明确定义任务的通才模型 [Bigaud et al., 2025, Narayanan et al., 2025]。这些方向，加上通用 RL 方法的进步，将继续扩展 LLM 的用例，从狭义定义的任务到具有开放目标的复杂交互，使它们能够更实质性地为新发现做出贡献。

7.9. 架构-算法协同设计中的强化学习

大多数当前的 LLM RL 流水线假设密集 Transformer [Vaswani et al., 2017] 或专家混合（MoE）[Dai et al., 2024, Jiang et al., 2024, Shazeer et al., 2017] 主干，优化几乎完全与任务准确性相关的奖励。因此，架构自由度及其硬件影响被排除在学习循环之外。与此同时，新一波的硬件、架构协同设计已经出现（例如，DeepSeek 的 NSA 中的硬件对齐稀疏注意力 [Yuan et al., 2025b] 和 Step-3 中的模型-系统协同设计 [Wang et al., 2025a]），表明通过将模型结构与计算基底对齐可以实现更高的效率和能力。

我们认为，在 RL 中将架构作为一等动作空间代表了下一代 LLM 的一个开放和高影响挑战。例如，强化 MoE 方法可以使模型在 RL 期间学习路由策略、专家激活、容量分配或稀疏模式，不仅优化任务奖励，还优化硬件感知目标，如延迟、内存流量、能耗和激活预算。在这种框架下，RL 的任务是学习不仅在 token 上"推理" [Guo et al., 2025a]，而且在参数和模块之间进行推理，动态调整模型的拓扑以适应每个提示的难度和实时计算约束。这一视角超越了经典的基于 RL 的神经架构搜索（NAS）[Zoph and Le, 2016]，后者通常为给定任务或数据集找到固定架构。相比之下，强化 MoE 专注于在推理期间优化每个输入的路由和模块化适应 [Han et al., 2021]，可能产生更高的效率和灵活性。关键开放问题包括设计避免平凡解决方案（例如，全专家稀疏性）的稳健多目标奖励函数，在架构操作修改网络拓扑时实现稳定的信用分配，以及在提示、任务和部署规模上摊销架构策略学习。解决这些挑战对于在未来 LLM 中实现真正集成的架构-算法协同优化至关重要。

8. 结论

我们综述了 LRM 的 RL 最新进展, 特别强调推理能力, 有效地将 LLM 转化为 LRM. 与之前主要用于人类对齐的 RLHF 或 DPO 方法相比, 我们的重点是 LLM 的 RLVR.RLVR 通过提供直接的结果级奖励来增强 LLM 的推理能力. 首先, 我们介绍了 RLVR 的核心组件, 包括奖励设计, 策略优化和采样策略. 我们为每个部分总结了多个研究方向和现有工作. 然后我们讨论了 LLM 的 RL 训练中几个最受争议的问题. 此外, 我们介绍了 LLM 的 RL 训练资源, 包括静态数据集, 动态环境和 RL 基础设施. 最后, 我们回顾了 RL 在各种场景中 LLM 的下游应用, 并强调了几个旨在通过基于 RL 的 LLM 实现超级智能的有前景的研究方向.

作者贡献

我们在下面呈现所有参与作者的贡献, 明确每个章节的主要负责人以及参与贡献的人员. 每位作者的具体贡献详细说明如下:

- **通讯作者**: Biqing Qi, Ning Ding, Bowen Zhou
- **项目负责人**: Kaiyan Zhang, Yuxin Zuo
- **引言**: Kaiyan Zhang
- **预备知识**:
  - 背景:Kaiyan Zhang
  - 前沿模型:Shang Qu, Yuru Wang (调研)
  - 相关综述:Yuxin Zuo
- **基础组件**:
  - 可验证奖励:Yuxin Zuo
  - 生成式奖励:Bingxiang He, Sihang Zeng (初稿)
  - 密集奖励:Runze Liu, Yu Fu (回合级)
  - 无监督奖励:Bingxiang He, Yuxin Zuo (调研)
  - 奖励塑造:Kaiyan Zhang
  - 策略梯度目标:Youbang Sun, Kaiyan Zhang (公式)
  - 基于批评家的算法:Youbang Sun, Kaiyan Zhang (公式)
  - 无批评家算法:Youbang Sun, Kaiyan Zhang (公式)
  - 离策略优化:Xingtai Lv, Yu Fu (回放缓冲区)
  - 正则化目标:Yuchen Zhang, Bingxiang He (熵), Yuxin Zuo (调研)
  - 动态和结构化采样:Yuchen Fan, Xuekai Zhu (面向效率的采样)
  - 采样超参数:Yuxin Zuo, Bingxiang He (调研)
- **基础问题**: Kaiyan Zhang, Yuxin Zuo
- **训练资源**:
  - 静态语料库:Kai Tian, Zhenzhao Yuan (调研)
  - 动态环境:Che Jiang
  - RL 基础设施:Kaiyan Zhang

- **应用**:
  - 编码任务:Pengfei Li, Xiang Xu (工具集成推理)
  - 代理任务:Yuchen Fan, Xinwei Long (GUI/计算机使用)
  - 多模态任务:Guoli Jia, Fangfu Liu (视频/3D), Xinwei Long (图像)
  - 多智能体系统:Shijie Wang
  - 机器人任务:Haozhan Li
  - 医疗任务:Sihang Zeng, Jiaze Ma (调研)
- **未来方向**:
  - Che Jiang (持续强化学习), Yu Fu (基于记忆的强化学习), Ermo Hua (潜在空间推理), Yuxin Zuo (预训练), Yihao Liu (扩散模型), Shang Qu (科学发现), Kaiyan Zhang (其他所有)
- **其他贡献**:
  - 图表:Yuru Wang, Kaiyan Zhang, Yuxin Zuo
  - 审阅和编辑:Zhiyuan Ma, Ganqu Cui, Huayu Chen, Weize Chen, Yafu Li, Xiaoye Qu, Junqi Gao, Dong Li, Zonglin Li, 以及上述所有作者

我们也感谢更广泛社区的宝贵建议和反馈. 特别感谢 Mingjie Wei, Wei Shen, Thomas Schmied, Muhammad Khalifa, Zhangquan Chen, Xinyu Zhu, Jacob Dineen, 和 Michal Kozak 提供建设性反馈, 包括识别文中的拼写错误, 描述不准确和缺失的引用. 我们欢迎进一步的反馈, 以帮助使这项工作成为该领域更有价值的资源.

References

a-m team. Am-deepseek-r1-0528-distilled, June 2025. URL https://github.com/a-m-team/a-m-models.

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. arXiv preprint arXiv:2504.21318, 2025.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Reincarnating reinforcement learning: Reusing prior computation to accelerate progress. Advances in neural information processing systems, 35:28955–28971, 2022.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. arxiv preprint arXiv: 2508.10925, 2025a.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. arXiv preprint arXiv:2505.15134, 2025b.

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. arXiv preprint arXiv:2503.04697, 2025.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In International Conference on Machine Learning, pages 265–279. PMLR, 2023.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods, pages 1657–1661, 2025.

Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. Opencodereasoning: Advancing data distillation for competitive coding. arXiv preprint arXiv:2504.01943, 2025.

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. arXiv preprint arXiv:2402.14740, 2024.

Moonshot AI. Kimi-researcher: End-to-end rl training for emerging agentic capabilities. https://moonshotai.github.io/Kimi-Researcher/, 2025. Accessed: 2025-08-13.

Qihang Ai, Pi Bu, Yue Cao, Yingyao Wang, Jihao Gu, Jingxuan Xing, Zekun Zhu, Wei Jiang, Zhicheng Zheng, Jun Song, et al. Inquiremobile: Teaching vlm-based mobile agent to request human assistance via reinforcement fine-tuning. arXiv preprint arXiv:2508.19679, 2025.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, et al. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. arXiv preprint arXiv:2502.17387, 2025.

Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models, 2025. URL https://hkunlp.github.io/blog/2025/Polaris.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3674–3683, 2018.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models. arXiv preprint arXiv:2408.11791, 2024.

Anthropic. Claude 3.7 sonnet and claude code, 2025a. URL https://www.anthropic.com/news/claude-3-7-sonnet.

Anthropic. Claude opus 4.1, 2025b. URL https://www.anthropic.com/claude/opus.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. arXiv preprint arXiv:2503.08679, 2025.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. arXiv preprint arXiv:2502.04463, 2025.

Pranav Arora, Rohan Gupta, and Kavya Patel. Gazal-r1: Scaling medical reasoning with grpo and multi-component reward design, 2025. URL https://arxiv.org/abs/2506.21594.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. arXiv preprint arXiv:2503.09573, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025.

Ibragim Badertdinov, Alexander Golubev, Maksim Nekrashevich, Anton Shevtsov, Simon Karasik, Andrei Andriushchenko, Maria Trofimova, Daria Litvintseva, and Boris Yangel. Swe-rebench: An automated pipeline for task collection and decontaminated evaluation of software engineering agents. arXiv preprint arXiv:2505.20411, 2025.

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, et al. Intern-s1: A scientific multimodal foundation model. arXiv preprint arXiv:2508.15763, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073, 2022b.

Baidu-ERNIE-Team. Ernie 4.5 technical report. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf, 2025.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv preprint arXiv:2503.11926, 2025.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, and et al. Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, 2025.

David Bani-Harouni. Language agents for hypothesis-driven clinical decision making with reinforcement learning, 2025. URL https://arxiv.org/abs/2506.13474.

Abdelhakim Benechehab, Youssef Attia El Hili, Ambroise Odonnat, Oussama Zekri, Albert Thomas, Giuseppe Paolo, Maurizio Filippone, Ievgen Redko, and Balázs Kégl. Zero-shot model-based reinforcement learning using large language models. arXiv preprint arXiv:2410.11711, 2024.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. arxiv preprint arXiv: 2505.00949, 2025.

Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, and Sergey Levine. Comps: Continual meta policy search. arXiv preprint arXiv:2112.04467, 2021.

Nathan Bigaud, Vincent Cabeli, Meltem Gürel, Arthur Pignet, John Klein, Gilles Wainrib, and Eric Durand. OwkinZero: Accelerating biological discovery with AI. arXiv preprint arXiv: 2508.16315, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024a.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In The Twelfth International Conference on Learning Representations, 2024b. URL https://openreview.net/forum?id=YCWjhGrJFD.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. Nature, 624(7992):570–578, 2023.

Umberto Borso, Davide Paglieri, Jude Wells, and Tim Rocktäschel. Preference-based alignment of discrete diffusion models. arXiv preprint arXiv:2503.08295, 2025.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In International Conference on Machine Learning, pages 3003–3020. PMLR, 2023.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. arXiv preprint arXiv:2407.21787, 2024.

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. Advances in neural information processing systems, 33:1877–1901, 2020.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In Forty-first International Conference on Machine Learning, 2024.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell, 187(25):7045–7063, 2024.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. arXiv preprint arXiv:2312.09390, 2023.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(2):156–172, 2008.

Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. Applied Sciences, 11(11):4948, 2021.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. arXiv preprint arXiv:2505.20272, 2025.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. arXiv preprint arXiv:2410.07095, 2024.

Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In International workshop on artificial intelligence and statistics, pages 57–64. PMLR, 2005.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. arXiv preprint arXiv:2506.13585, 2025a.

Ding Chen, Qingchen Yu, Pengyuan Wang, Wentao Zhang, Bo Tang, Feiyu Xiong, Xinchi Li, Minchuan Yang, and Zhiyu Li. xverify: Efficient answer verifier for reasoning model evaluations. arXiv preprint arXiv:2504.10481, 2025b.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation. arXiv preprint arXiv:2309.17288, 2023a.

Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. Bridging supervised learning and reinforcement learning in math reasoning. arXiv preprint arXiv:2505.18116, 2025c.

Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, Hongli Yu, Qiying Yu, Xuefeng Li, Jiaze Chen, et al. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles. arXiv preprint arXiv:2505.19914, 2025d.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024a. URL https://arxiv.org/abs/2412.18925.

Lei Chen, Xuanle Zhao, Zhixiong Zeng, Jing Huang, Liming Zheng, Yufeng Zhong, and Lin Ma. Breaking the sft plateau: Multimodal structured reinforcement learning for chart-to-code generation. arXiv preprint arXiv:2508.13587, 2025e.

Lei Chen, Xuanle Zhao, Zhixiong Zeng, Jing Huang, Yufeng Zhong, and Lin Ma. Chart-r1: Chain-of-thought supervision and reinforcement for advanced chart reasoner. arXiv preprint arXiv:2507.15509, 2025f.

Liang Chen, Hongcheng Gao, Tianyu Liu, Zhiqi Huang, Flood Sung, Xinyu Zhou, Yuxin Wu, and Baobao Chang. G1: Bootstrapping perception and reasoning abilities of vision-language model via reinforcement learning. arXiv preprint arXiv:2505.13426, 2025g.

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. Beyond two-stage training: Cooperative sft and rl for llm reasoning, 2025h. URL https://arxiv.org/abs/2509.06948.

Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Self-questioning language models. arXiv preprint arXiv:2508.03682, 2025i.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. arXiv preprint arXiv:2505.12346, 2025j.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. Learning to reason with search for llms via reinforcement learning. arXiv preprint arXiv:2503.19470, 2025k.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge. arXiv preprint arXiv:2504.00050, 2025l.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567, 2025m.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088, 2025n.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for llm reasoning. arXiv preprint arXiv:2505.14970, 2025o.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. arXiv preprint arXiv:2412.21187, 2024b.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. Rm-r1: Reward modeling as reasoning. arXiv preprint arXiv:2505.02387, 2025p.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. arXiv preprint arXiv:2505.16400, 2025q.

Yongchao Chen, Yueying Liu, Junwei Zhou, Yilun Hao, Jingquan Wang, Yang Zhang, and Chuchu Fan. R1-code-interpreter: Training llms to reason with code via supervised and reinforcement learning. arXiv preprint arXiv:2505.21668, 2025r.

Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5):2804–2818, 2023b.

Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. arXiv preprint arXiv:2502.05450, 2025s.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. arXiv preprint arXiv:2507.07966, 2025t.

Yuyang Chen, Kaiyan Zhao, Yiming Wang, Ming Yang, Jian Zhang, and Xiaoguang Niu. Enhancing llm agents for code generation with possibility and pass-rate prioritized experience replay. arXiv preprint arXiv:2410.12236, 2024c.

Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. arXiv preprint arXiv:2503.07523, 2025u.

Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. Sifthinker: Spatially-aware image focus for visual reasoning. arXiv preprint arXiv:2508.06259, 2025v.

Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, et al. An empirical study on eliciting and improving r1-like reasoning models. arXiv preprint arXiv:2503.04548, 2025w.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models. arXiv preprint arXiv:2508.10751, 2025x.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. arXiv preprint arXiv:2506.14758, 2025a.

Jie Cheng, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Gang Xiong, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. arXiv preprint arXiv:2504.15275, 2025b.

Shuang Cheng, Yihan Bian, Dawei Liu, Yuhua Jiang, Yihao Liu, Linfeng Zhang, Wenghai Wang, Qipeng Guo, Kai Chen, Biqing Qi*, and Bowen Zhou. Sdar: A synergistic diffusion – autoregression paradigm for scalable sequence generation, 2025c. URL https://github.com/JetAstra/SDAR.

Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, et al. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. arXiv preprint arXiv:2506.14965, 2025d.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. Advances in Neural Information Processing Systems, 36:73383–73394, 2023.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. arXiv preprint arXiv:2504.19413, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161, 2025a.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information. arXiv preprint arXiv:2505.23558, 2025b.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. arXiv preprint arXiv:2505.18842, 2025.

Taco Cohen, David W Zhang, Kunhao Zheng, Yunhao Tang, Remi Munos, and Gabriel Synnaeve. Soft policy optimization: Online off-policy rl for sequence models. arXiv preprint arXiv:2503.05453, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arxiv preprint arXiv: 2507.06261, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. arXiv preprint arXiv:2502.01456, 2025a.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. arXiv preprint arXiv:2505.22617, 2025b.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066, 2024.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models. arXiv preprint arXiv:2505.07686, 2025a.

Yaxun Dai, Wenxuan Xie, Xialie Zhuang, Tianyu Yang, Yiying Yang, Haiqin Yang, Yuhang Zhao, Pingfu Chao, and Wenhao Jiang. Reex-sql: Reasoning with execution-aware reinforcement learning for text-to-sql. arXiv preprint arXiv:2505.12768, 2025b.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. arXiv preprint arXiv:2505.24718, 2025.

Alan Dao and Thinh Le. Rezero: Enhancing llm search ability by trying one-more-time. arXiv preprint arXiv:2504.11001, 2025.

Alan Dao and Dinh Bach Vu. Jan-nano technical report. arXiv preprint arXiv:2506.22760, 2025.

Antoine Dedieu, Joseph Ortiz, Xinghua Lou, Carter Wendelken, Wolfgang Lehrach, J Swaroop Guntupalli, Miguel Lazaro-Gredilla, and Kevin Patrick Murphy. Improving transformer world models for data-efficient rl. In ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling, 2025.

Jia Deng, Jie Chen, Zhipeng Chen, Daixuan Cheng, Fei Bai, Beichen Zhang, Yinqian Min, Yanzipeng Gao, Wayne Xin Zhao, and Ji-Rong Wen. From trial-and-error to improvement: A systematic analysis of llm exploration mechanisms in rlvr. arXiv preprint arXiv:2508.07534, 2025a.

Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, et al. Atom-searcher: Enhancing agentic deep research via fine-grained atomic thought reward. arXiv preprint arXiv:2508.12800, 2025b.

Jacob Dineen, Aswin RRV, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, et al. Qa-lign: Aligning llms through constitutionally decomposed qa. arXiv preprint arXiv:2506.08123, 2025.

Hongxin Ding, Baixiang Huang, and Yue Fang. Promed: Shapley information gain guided reinforcement learning for proactive medical llms, 2025. URL https://arxiv.org/abs/2508.13514.

Ziluo Ding, Wanpeng Zhang, Junpeng Yue, Xiangjun Wang, Tiejun Huang, and Zongqing Lu. Entity divider with language grounding in multi-agent reinforcement learning. In International Conference on Machine Learning, pages 8103–8119. PMLR, 2023.

Dai Do, Manh Nguyen, Svetha Venkatesh, and Hung Le. Sparft: Self-paced reinforcement fine-tuning for large language models. arXiv preprint arXiv:2508.05015, 2025.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. Nature, 632(8026): 768–774, 2024.

Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. arXiv preprint arXiv:2505.16410, 2025a.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. arXiv preprint arXiv:2507.19849, 2025b.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767, 2023.

Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement pre-training. arXiv preprint arXiv:2506.08007, 2025c.

Ali Dorri, Salil S Kanhere, and Raja Jurdak. Multi-agent systems: A survey. Ieee Access, 6: 28573–28593, 2018.

Shihan Dou, Muling Wu, Jingwen Xu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. Improving rl exploration for llm reasoning through retrospective replay. arXiv preprint arXiv:2504.14363, 2025.

Mingzhe Du, Luu Anh Tuan, Yue Liu, Yuhao Qing, Dong Huang, Xinyi He, Qian Liu, Zejun Ma, and See-kiong Ng. Afterburner: Reinforcement learning facilitates self-improving code efficiency optimization. arXiv preprint arXiv:2505.23387, 2025a.

Yong Du, Yuchen Yan, Fei Tang, Zhengxi Lu, Chang Zong, Weiming Lu, Shengpei Jiang, and Yongliang Shen. Test-time reinforcement learning for gui grounding via region consistency. arXiv preprint arXiv:2508.05615, 2025b.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:2505.17022, 2025.

Sam Earle, Graham Todd, Yuchen Li, Ahmed Khalifa, Muhammad Umair Nasir, Zehua Jiang, Andrzej Banburski-Fahey, and Julian Togelius. Puzzlejax: A benchmark for reasoning and learning, 2025. URL https://arxiv.org/abs/2508.16821.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. Competitive programming with large reasoning models. arXiv preprint arXiv:2502.06807, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning, 2024.

Benjamin Eysenbach and Sergey Levine. Maximum entropy rl (provably) solves some robust rl problems. arXiv preprint arXiv:2103.06257, 2021.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, january 2025. URL https://github.com/huggingface/open-r1, page 9, 2025.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. BioReason: Incentivizing multimodal biological reasoning within a DNA-LLM model. arXiv preprint arXiv: 2505.23579, 2025.

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. Megascience: Pushing the frontiers of post-training datasets for science reasoning. arXiv preprint arXiv:2507.16812, 2025a.

Tiantian Fan, Lingjun Liu, Yu Yue, Jiaze Chen, Chengyi Wang, Qiying Yu, Chi Zhang, Zhiqi Lin, Ruofei Zhu, Yufeng Yuan, et al. Truncated proximal policy optimization. arXiv preprint arXiv:2506.15050, 2025b.

Yuchen Fan, Kaiyan Zhang, Heng Zhou, Yuxin Zuo, Yanxu Chen, Yu Fu, Xinwei Long, Xuekai Zhu, Che Jiang, Yuchen Zhang, et al. Ssrl: Self-search reinforcement learning. arXiv preprint arXiv:2508.10874, 2025c.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. arXiv preprint arXiv:2505.15879, 2025d.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. arXiv preprint arXiv:2504.20930, 2025e.

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. Serl: Self-play reinforcement learning for large language models with limited data. arXiv preprint arXiv:2505.20347, 2025a.

Xueji Fang, Liyuan Ma, Zhiyang Chen, Mingyuan Zhou, and Guo-jun Qi. Inflvg: Reinforce inference-time consistent long video generation with grpo. arXiv preprint arXiv:2505.17574, 2025b.

Yin Fang, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Aidong Zhang, Jiawei Han, and Zhiyong Lu. Cell-o1: Training LLMs to solve single-cell reasoning puzzles with reinforcement learning, 2025c.

Wu Fei, Hao Kong, Shuxian Liang, Yang Lin, Yibo Yang, Jing Tang, Lei Chen, and Xiansheng Hua. Self-guided process reward optimization with redefined step-wise advantage for process reinforcement learning. arXiv preprint arXiv:2507.01551, 2025a.

Xiang Fei, Siqi Wang, Shu Wei, Yuxiang Nie, Wei Shi, Hao Feng, and Can Huang. Post-completion learning for language models. arXiv preprint arXiv:2507.20252, 2025b.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms. arXiv preprint arXiv:2504.11536, 2025a.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776, 2025b.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. arXiv preprint arXiv:2504.10903, 2025c.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. The International Journal of Robotics Research, 44(5): 701–739, 2025.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. Scaling reasoning, losing control: Evaluating instruction following in large reasoning models. arXiv preprint arXiv:2505.14810, 2025a.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. arXiv preprint arXiv:2505.24298, 2025b.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. arXiv preprint arXiv:2506.19767, 2025c.

Yasuhiro Fujita. Experience replay with random reshuffling. arXiv preprint arXiv:2503.02269, 2025.

Marcos Fuster-Pena, David de Fitero-Dominguez, Antonio Garcia-Cabot, and Eva Garcia-Lopez. Repaca: Leveraging reasoning large language models for static automated patch correctness assessment. arXiv preprint arXiv:2507.22580, 2025.

Kushal Gajjar, Harshit Sikchi, Arpit Singh Gautam, Marc Hammons, and Saurabh Jha. Cognisql-r1-zero: Lightweight reinforced reasoning for efficient sql generation. arXiv preprint arXiv:2507.06013, 2025.

Zeyu Gan, Hao Yi, and Yong Liu. Cot-space: A theoretical framework for internal slow-thinking via reinforcement learning, 2025. URL https://arxiv.org/abs/2509.04027.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. arXiv preprint arXiv:2503.01307, 2025.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. arXiv preprint arXiv:2508.07976, 2025a.

Longxi Gao, Li Zhang, and Mengwei Xu. Uishift: Enhancing vlm-based gui agents through self-supervised reinforcement learning. arXiv preprint arXiv:2505.12493, 2025b.

Francisco Garcia and Philip S Thomas. A meta-mdp approach to exploration for lifelong reinforcement learning. Advances in Neural Information Processing Systems, 32, 2019.

Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. arXiv preprint arXiv:2211.10445, 2022.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. arXiv preprint arXiv:2502.05171, 2025.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. arXiv preprint arXiv:2508.05748, 2025.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, and Samuel G. Rodriques. Robin: A multi-agent system for automating scientific discovery, 2025.

Majid Ghasemi, Amir Hossein Moosavi, and Dariush Ebrahimi. A comprehensive survey of reinforcement learning: From algorithms to practical challenges. arXiv preprint arXiv:2411.18892, 2024.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. arXiv preprint arXiv:2411.04872, 2024.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. Diffucoder: Understanding and improving masked diffusion models for code generation. arXiv preprint arXiv:2506.20639, 2025.

Prasoon Goyal, Scott Niekum, and Raymond J Mooney. Using natural language for reward shaping in reinforcement learning. arXiv preprint arXiv:1903.02020, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Jihao Gu, Qihang Ai, Yingyao Wang, Pi Bu, Jingxuan Xing, Zekun Zhu, Wei Jiang, Ziming Wang, Yingxiu Zhao, Ming-Liang Zhang, et al. Mobile-r1: Towards interactive reinforcement learning for vlm-based mobile agent via task-level rewards. arXiv preprint arXiv:2506.20332, 2025.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, et al. Is your llm secretly a world model of the internet? model-based planning for web agents. arXiv preprint arXiv:2411.06559, 2024.

Zhong Guan, Likang Wu, Hongke Zhao, Jiahui Wang, and Le Wu. Recall-extend dynamics: Enhancing small language models through controlled exploration and refined offline integration. arXiv preprint arXiv:2508.16677, 2025.

Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. Textarena, 2025. URL https://arxiv.org/abs/2504.11442.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. arXiv preprint arXiv:2507.17746, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025a.

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Reward reasoning model. arXiv preprint arXiv:2505.14674, 2025b.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. arXiv preprint arXiv:2505.23564, 2025c.

Yongxin Guo, Wenbo Deng, Zhenglin Cheng, and Xiaoying Tang. G2 rpo-a: Guided group relative policy optimization with adaptive guidance. arXiv preprint arXiv:2508.13023, 2025d.

Zhihan Guo, Jiele Wu, Wenqian Cui, Yifei Zhang, Minda Hu, Yufei Wang, and Irwin King. From general to targeted rewards: Surpassing gpt-4 in open-ended long-context generation. arXiv preprint arXiv:2506.16024, 2025e.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025f.

Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. Advances in Neural Information Processing Systems, 35:15281–15295, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning, pages 1861–1870. Pmlr, 2018.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/32fdab6559cdfa4f167f8c31b9199643-Paper.pdf.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104, 2023.

Ai Han, Junxing Hu, Pu Wei, Zhiqian Zhang, Yuhang Guo, Jiawei Lu, and Zicheng Zhang. Joyagents-r1: Joint evolution dynamics for versatile multi-llm agents with reinforcement learning. arXiv preprint arXiv:2506.19846, 2025.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(11): 7436–7456, 2021.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769, 2024.

Bingxiang He, Wenbin Zhang, Jiaxi Song, Cheng Qian, Zixuan Fu, Bowen Sun, Ning Ding, Haiwen Hong, Longtao Huang, Hui Xue, et al. Air: A systematic analysis of annotations, instructions, and response pairs in preference dataset. arXiv preprint arXiv:2504.03612, 2025a.

Feng He, Zijun Chen, Xinnian Liang, Tingting Ma, Yunqi Qiu, Shuangzhi Wu, and Junchi Yan. Protoreasoning: Prototypes as the foundation for generalizable reasoning in llms. arXiv preprint arXiv:2506.15211, 2025b.

Haoyu He, Katrin Renz, Yong Cao, and Andreas Geiger. Mdpo: Overcoming the training-inference divide of masked diffusion language models. arXiv preprint arXiv:2508.13148, 2025c.

Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference. Thinking Machines Lab: Connectionism, 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. arXiv preprint arXiv:2505.22312, 2025d.

Qianyu He, Siyu Yuan, Xuefeng Li, Mingxuan Wang, and Jiangjie Chen. Thinkdial: An open recipe for controlling reasoning effort in large language models. arXiv preprint arXiv:2508.18773, 2025e.

Tao He, Rongchuan Mu, Lizi Liao, Yixin Cao, Ming Liu, and Bing Qin. Good learners think their thinking: Generative prm makes large reasoning model more efficient math learner. arXiv preprint arXiv:2507.23317, 2025f.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. arXiv preprint arXiv:2508.04324, 2025g.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. arXiv preprint arXiv:2504.11456, 2025h.

Michael Hla. Pro-1, 2025. URL https://michaelhla.com/blog/pro1.html.

Haitao Hong, Yuchen Yan, Xingyu Wu, Guiyang Hou, Wenqi Zhang, Weiming Lu, Yongliang Shen, and Jun Xiao. Cooper: Co-optimizing policy and reward models in reinforcement learning for large language models. arXiv preprint arXiv:2508.05613, 2025a.

Ilgee Hong, Changlong Yu, Liang Qiu, Weixiang Yan, Zhenghao Xu, Haoming Jiang, Qingru Zhang, Qin Lu, Xin Liu, Chao Zhang, et al. Think-rm: Enabling long-horizon reasoning in generative reward models. arXiv preprint arXiv:2505.16265, 2025b.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. Reinforcing multimodal understanding and generation with dual self-rewards. arXiv preprint arXiv:2506.07963, 2025c.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. Treerl: Llm reinforcement learning with on-policy tree search. arXiv preprint arXiv:2506.11902, 2025.

Haichuan Hu, Xiaochen Xie, and Quanjun Zhang. Repair-r1: Better test before repair. arXiv preprint arXiv:2507.22853, 2025a.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. arXiv preprint arXiv:2501.03262, 2025.

Jian Hu, Xibin Wu, Zilin Zhu, Weixun Wang, Dehao Zhang, Yu Cao, et al. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. arXiv preprint arXiv:2405.11143, 2024a.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290, 2025b.

Lanxiang Hu, Mingjia Huo, Yuxuan Zhang, Haoyang Yu, Eric P Xing, Ion Stoica, Tajana Rosing, Haojian Jin, and Hao Zhang. lmgame-bench: How good are llms at playing games? arXiv preprint arXiv:2505.15146, 2025c.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395, 2024b.

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. Advances in Neural Information Processing Systems, 33:15931–15941, 2020.

Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. arXiv preprint arXiv:2312.05230, 2023.

Ermo Hua, Biqing Qi, Kaiyan Zhang, Yue Yu, Ning Ding, Xingtai Lv, Kai Tian, and Bowen Zhou. Intuitive fine-tuning: Towards simplifying alignment into a single process. arXiv preprint arXiv:2405.11870, 2024.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. arXiv preprint arXiv:2507.00432, 2025.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. arXiv preprint arXiv:2508.05004, 2025a.

Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. The 37 implementation details of proximal policy optimization. The ICLR Blog Track 2023, 2022.

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. arXiv preprint arXiv:2507.23478, 2025b.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749, 2025c.

Yanxing Huang, Xinling Jin, Sijie Liang, Peng Li, and Yang Liu. Formarl: Enhancing autoformalization with no labeled data. arXiv preprint arXiv:2508.18914, 2025d.

Yuzhen Huang, Weihao Zeng, Xingshan Zeng, Qi Zhu, and Junxian He. Pitfalls of rule-and model-based verifiers–a case study on mathematical reasoning. arXiv preprint arXiv:2505.22203, 2025e.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement learning with rubric anchors. arXiv preprint arXiv:2508.12790, 2025f.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. arXiv preprint arXiv:2507.01679, 2025g.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey. arXiv preprint arXiv:2312.10256, 2023.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. Science Robotics, 4 (26):eaau5872, 2019.

Baichuan Inc. Baichuan-m1: Pushing the medical capability of large language models. arXiv preprint arXiv::2502.12671, 2025a.

Baichuan Inc. Baichuan-m2: Scaling medical capability with large verifier system. arXiv preprint arXiv::2509.02208, 2025b.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi 0.5: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

Edward L Ionides. Truncated importance sampling. Journal of Computational and Graphical Statistics, 17(2):295–311, 2008.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. rbio1-training scientific reasoning LLMs with biological world models as soft verifiers. bioRxiv 2025.08.18.670981, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

Naman Jain, Jaskirat Singh, Manish Shetty, Liang Zheng, Koushik Sen, and Ion Stoica. R2e-gym: Procedural environments and hybrid verifiers for scaling open-weights swe agents. arXiv preprint arXiv:2504.07164, 2025.

Kunal Jha, Wilka Carvalho, Yancheng Liang, Simon Shaolei Du, Max Kleiman-Weiner, and Natasha Jaques. Cross-environment cooperation enables zero-shot multi-agent coordination. In Forty-second International Conference on Machine Learning, 2025a.

Manvi Jha, Jiaxin Wan, and Deming Chen. Proof2silicon: Prompt repair for verified code and hardware generation via reinforcement learning. arXiv preprint arXiv:2509.06239, 2025b.

Haozhe Ji, Cheng Lu, Yilin Niu, Pei Ke, Hongning Wang, Jun Zhu, Jie Tang, and Minlie Huang. Towards efficient exact optimization of language model alignment. arXiv preprint arXiv:2402.00856, 2024.

Xingguang Ji, Yahui Liu, Qi Wang, Jingyuan Zhang, Yang Yue, Rui Shi, Chenxi Sun, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Leanabell-prover-v2: Verifier-integrated reasoning for formal theorem proving via reinforcement learning. arXiv preprint arXiv:2507.08649, 2025a.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 1724–1734, 2025b.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Writing-zero: Bridge the gap between non-verifiable tasks and verifiable rewards. arXiv e-prints, pages arXiv–2506, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.

Daniel R Jiang, Alex Nikulkov, Yu-Chia Chen, Yang Bai, and Zheqing Zhu. Improving generative ad text on facebook using reinforcement learning. arXiv preprint arXiv:2507.21983, 2025a.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703, 2025b.

Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation. arXiv preprint arXiv:2505.17534, 2025c.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don't need that much data to train a search agent via rl. arXiv preprint arXiv:2505.14146, 2025d.

Yuhua Jiang, Yuwen Xiong, Yufeng Yuan, Chao Xin, Wenyuan Xu, Yu Yue, Qianchuan Zhao, and Lin Yan. Pag: Multi-turn reinforced llm self-correction with policy as generative verifier. arXiv preprint arXiv:2506.10406, 2025e.

Yuqian Jiang, Suda Bharadwaj, Bo Wu, Rishi Shah, Ufuk Topcu, and Peter Stone. Temporal-logic-based reward shaping for continuing reinforcement learning tasks. In Proceedings of the AAAI Conference on artificial Intelligence, volume 35, pages 7995–8003, 2021.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? arXiv preprint arXiv:2310.06770, 2023.

Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O Arik, and Jiawei Han. An empirical study on reinforcement learning for reasoning-search interleaved llm agents. arXiv preprint arXiv:2505.15117, 2025a.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv preprint arXiv:2503.09516, 2025b.

Can Jin, Yang Zhou, Qixin Zhang, Hongwu Peng, Di Zhang, Marco Pavone, Ligong Han, Zhang-Wei Hong, Tong Che, and Dimitris N Metaxas. Your reward function for rl is your best prm for search: Unifying rl and search-based tts. arXiv preprint arXiv:2508.14313, 2025c.

Hangzhan Jin, Sicheng Lv, Sifan Wu, and Mohammad Hamdaqa. Rl is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for llms. arXiv preprint arXiv:2508.16546, 2025d.

Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation. arXiv preprint arXiv:2504.18453, 2025.

Zhehan Kan, Yanlin Liu, Kun Yin, Xinghua Jiang, Xin Li, Haoyu Cao, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, et al. Taco: Think-answer consistency for optimized long-chain reasoning and efficient data learning via reinforcement learning in lvlms. arXiv preprint arXiv:2505.20777, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=Myx2kJFzAn.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Assessing the limits of zero-shot foundation models in single-cell biology, 2023.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. arXiv preprint arXiv:2504.16828, 2025.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645, 2025.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. arXiv preprint arXiv:2310.06452, 2023.

Andrew Kiruluta, Andreas Lemos, and Priscilla Burity. A self-supervised reinforcement learning approach for fine-tuning large language models using cross-attention signals. arXiv preprint arXiv:2502.10482, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

Inception Labs, Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, et al. Mercury: Ultra-fast language models based on diffusion. arXiv preprint arXiv:2506.17298, 2025.

Hanyu Lai, Xiao Liu, Yanxiao Zhao, Han Xu, Hanchen Zhang, Bohao Jing, Yanyu Ren, Shuntian Yao, Yuxiao Dong, and Jie Tang. Computerrl: Scaling end-to-end online reinforcement learning for computer use agents, 2025. URL https://arxiv.org/abs/2508.14040.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124, 2024.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, volume 3, page 896. Atlanta, 2013.

Dong Won Lee, Hae Won Park, Yoon Kim, Cynthia Breazeal, and Louis-Philippe Morency. Improving dialogue agents by decomposing one global explicit annotation with local implicit multimodal feedback. arXiv preprint arXiv:2403.11330, 2024a.

Dong Won Lee, Hae Won Park, Cynthia Breazeal, and Louis-Philippe Morency. Aligning dialogue agents with global feedback via large language model reward decomposition. arXiv preprint arXiv:2505.15922, 2025a.

Dongjun Lee, Changho Hwang, and Kimin Lee. Learning to generate unit test via adversarial reinforcement learning. arXiv preprint arXiv:2508.21107, 2025b.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In Ruslan

Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 26874–26901. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/lee24t.html.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. arXiv preprint arXiv:2507.01915, 2025a.

Chengpeng Li, Zhengyang Tang, Ziniu Li, Mingfeng Xue, Keqin Bao, Tian Ding, Ruoyu Sun, Benyou Wang, Xiang Wang, Junyang Lin, et al. Cort: Code-integrated reasoning within thinking. arXiv preprint arXiv:2506.09820, 2025b.

Chunmao Li, Yang Li, Yinliang Zhao, Peng Peng, and Xupeng Geng. Sler: Self-generated long-term experience replay for continual reinforcement learning. Applied Intelligence, 51(1):185–201, 2021.

Derek Li, Jiaming Zhou, Amirreza Kazemi, Qianyi Sun, Abbas Ghaddar, Mohammad Ali Alomrani, Liheng Ma, Yu Luo, Dong Li, Feng Wen, et al. Omni-think: Scaling cross-domain generalization in llms via multi-task rl with hybrid rewards. arXiv preprint arXiv:2507.14783, 2025c.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. arXiv preprint arXiv:2505.21318, 2025d.

Hu Li, Xuezhong Qian, and Wei Song. Prioritized experience replay based on dynamics priority. Scientific Reports, 14(1):6014, 2024a.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. arXiv preprint arXiv:2310.10701, 2023a.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. Hugging Face repository, 13:9, 2024b.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. arXiv preprint arXiv:2310.05470, 2023b.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802, 2025e.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. Websailor: Navigating super-human reasoning for web agent. arXiv preprint arXiv:2507.02592, 2025f.

Peiji Li, Jiasheng Ye, Yongkang Chen, Yichuan Ma, Zijie Yu, Kedi Chen, Ganqu Cui, Haozhan Li, Jiacheng Chen, Chengqi Lyu, et al. Internbootcamp technical report: Boosting llm reasoning with verifiable task scaling. arXiv preprint arXiv:2508.08636, 2025g.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. Confidence is all you need: Few-shot rl fine-tuning of language models. arXiv preprint arXiv:2506.06395, 2025h.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025i. URL https://arxiv.org/abs/2509.02534.

Weizhen Li, Jianbo Lin, Zhuosong Jiang, Jingyi Cao, Xinpeng Liu, Jiayu Zhang, Zhenqiang Huang, Qianben Chen, Weichen Sun, Qiexiang Wang, et al. Chain-of-agents: End-to-end agent foundation models via multi-agent distillation and agentic rl. arXiv preprint arXiv:2508.13167, 2025j.

Wenjie Li, Yujie Zhang, and Haoran Sun. Cx-mind: A pioneering multimodal large language model for interleaved reasoning in chest x-ray via curriculum-guided reinforcement learning, 2025k. URL https://arxiv.org/abs/2508.03733.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in llms. arXiv preprint arXiv:2505.11423, 2025l.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. arXiv preprint arXiv:2504.21776, 2025m.

Xingxuan Li, Yao Xiao, Dianwen Ng, Hai Ye, Yue Deng, Xiang Lin, Bin Wang, Zhanfeng Mo, Chong Zhang, Yueyi Zhang, et al. Miromind-m1: An open-source advancement in mathematical reasoning via context-aware multi-stage policy optimization. arXiv preprint arXiv:2507.14683, 2025n.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958, 2025o.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. arXiv preprint arXiv:2502.11886, 2025p.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Torl: Scaling tool-integrated rl. arXiv preprint arXiv:2503.23383, 2025q.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, Zheng Zhang, Wei Shen, Qian Liu, Chenghua Lin, Jian

Yang, Ge Zhang, and Wenhao Huang. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. arXiv preprint arXiv:2508.17445, 2025r.

Yue Li, Meng Tian, Dechang Zhu, Jiangtong Zhu, Zhenyu Lin, Zhiwei Xiong, and Xinhai Zhao. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. arXiv preprint arXiv:2506.18234, 2025s.

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, Xiang Yue, and Radha Poovendran. Temporal sampling for forgotten reasoning in llms. arXiv preprint arXiv:2505.20196, 2025t.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Weili Guan, Dongmei Jiang, and Liqiang Nie. Optimus-3: Towards generalist multimodal minecraft agents with scalable task experts. arXiv preprint arXiv:2506.10357, 2025u.

Zhiwei Li, Yong Hu, and Wenqing Wang. Encouraging good processes without the need for good answers: Reinforcement learning for llm agent planning. arXiv preprint arXiv:2508.19598, 2025v.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv:2502.17419, 2025w.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. arXiv preprint arXiv:2310.10505, 2023c.

Jing Liang, Hongyao Tang, Yi Ma, Jinyi Liu, Yan Zheng, Shuyue Hu, Lei Bai, and Jianye Hao. Squeeze the soaked sponge: Efficient off-policy reinforcement finetuning for large language model. arXiv preprint arXiv:2507.06892, 2025a.

Xiao Liang, Zhong-Zhi Li, Yeyun Gong, Yang Wang, Hengyuan Zhang, Yelong Shen, Ying Nian Wu, and Weizhu Chen. Sws: Self-aware weakness-driven problem synthesis in reinforcement learning for llm reasoning. arXiv preprint arXiv:2506.08989, 2025b.

Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. arXiv preprint arXiv:2508.14029, 2025c.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 13199–13208, 2025d.

Jianxing Liao, Tian Zhang, Xiao Feng, Yusong Zhang, Rui Yang, Haorui Wang, Bosi Wen, Ziying Wang, and Runzhi Shi. Rlmr: Reinforcement learning with mixed rewards for creative writing. arXiv preprint arXiv:2508.18642, 2025a.

Mengqi Liao, Xiangyu Xi, Ruinian Chen, Jia Leng, Yangen Hu, Ke Zeng, Shuai Liu, and Huaiyu Wan. Enhancing efficiency and exploration in reinforcement learning for llms. arXiv preprint arXiv:2505.18573, 2025b.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. arXiv preprint arXiv:2504.00883, 2025c.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Darryl Lin, Sachin Talathi, and Sreekanth Annapureddy. Fixed point quantization of deep convolutional networks. In International conference on machine learning, pages 2849–2858. PMLR, 2016.

Hongyu Lin, Yuchen Li, Haoran Luo, Kaichun Yao, Libo Zhang, Mingjie Xing, and Yanjun Wu. Os-r1: Agentic operating system kernel tuning with reinforcement learning. arXiv preprint arXiv:2508.12551, 2025a.

Jiacheng Lin and Zhenbang Wu. Training llms for ehr-based reasoning tasks via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.24105.

Wang Lin, Liyu Jia, Wentao Hu, Kaihang Pan, Zhongqi Yue, Wei Zhao, Jingyuan Chen, Fei Wu, and Hanwang Zhang. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning. arXiv preprint arXiv:2504.15932, 2025b.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. arXiv preprint arXiv:2506.24119, 2025a.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. In Proceedings of the IEEE/CVF international conference on computer vision, 2025b.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, ShaoGuo Liu, and TingTing Gao. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. arXiv preprint arXiv:2508.11356, 2025c.

Jiawei Liu and Lingming Zhang. Code-r1: Reproducing r1 for code with reliable rewards. https://github.com/ganler/code-r1, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470, 2025d.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. arXiv preprint arXiv:2501.13918, 2025e.

Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can rl bring to vla generalization? an empirical study. arXiv preprint arXiv:2505.19789, 2025f.

Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, et al. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. arXiv preprint arXiv:2505.19641, 2025g.

Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents, 2025h. URL https://arxiv.org/abs/2509.06501.

Lijun Liu and Ruiyang Li. Efficient medical vie via reinforcement learning, 2025. URL https://arxiv.org/abs/2506.13363.

Liyuan Liu, Feng Yao, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Flashrl: 8bit rollouts, full power rl, August 2025i. URL https://fengyao.notion.site/flash-rl.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. arXiv preprint arXiv:2505.24864, 2025j.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Uft: Unifying supervised and reinforcement fine-tuning. arXiv preprint arXiv:2505.16984, 2025k.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 8009–8019, 2025l.

Runze Liu, Fengshuo Bai, Yali Du, and Yaodong Yang. Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 22270–22284. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8be9c134bb193d8bd3827d4df8488228-Paper-Conference.pdf.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. arXiv preprint arXiv:2502.06703, 2025m.

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, et al. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. arXiv preprint arXiv:2508.03686, 2025n.

Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. Llm collaboration with multi-agent reinforcement learning. arXiv preprint arXiv:2508.04652, 2025o.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. Annals of Operations Research, 339(3):1119–1148, 2024.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. arXiv preprint arXiv:2309.06657, 2023a.

Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. Learn to reason efficiently with adaptive length-based reward shaping. arXiv preprint arXiv:2505.15612, 2025p.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In The Eleventh International Conference on Learning Representations, 2023b.

Yifei Liu, Li Lyna Zhang, Yi Zhu, Bingcheng Dong, Xudong Zhou, Ning Shang, Fan Yang, and Mao Yang. rstar-coder: Scaling competitive code reasoning with a large-scale verified dataset. arXiv preprint arXiv:2505.21297, 2025q.

Yizhou Liu and Jingwei Wei. Breaking reward collapse: Adaptive reinforcement for open-ended medical reasoning with enhanced semantic discrimination, 2025. URL https://arxiv.org/abs/2508.12957.

Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. arXiv preprint arXiv:2504.14239, 2025r.

Zexi Liu, Jingyi Chai, Xinyu Zhu, Shuo Tang, Rui Ye, Bo Zhang, Lei Bai, and Siheng Chen. Ml-agent: Reinforcing llm agents for autonomous machine learning engineering. arXiv preprint arXiv:2505.23723, 2025s.

Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. There may not be aha moment in r1-zero-like training—a pilot study, 2025t.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. arXiv preprint arXiv:2503.20783, 2025u.

Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. arXiv preprint arXiv:2506.13284, 2025v.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. Part i: Tricks or traps? a deep dive into rl for llm reasoning. arXiv preprint arXiv:2508.08221, 2025w.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. arXiv preprint arXiv:2504.02495, 2025x.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785, 2025y.

Zongkai Liu, Fanqing Meng, Lingxiao Du, Zhixiang Zhou, Chao Yu, Wenqi Shao, and Qiaosheng Zhang. Cpgd: Toward stable rule-based reinforcement learning for language models. arXiv preprint arXiv:2505.12504, 2025z.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. arXiv preprint arXiv:2509.08621, 2025.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems, 30, 2017.

Dakuan Lu, Xiaoyu Tan, Rui Xu, Tianchu Yao, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. Scp-116k: A high-quality problem-solution dataset and a generalized pipeline for automated extraction in the higher education science domain. arXiv preprint arXiv:2501.15587, 2025a.

Fanbin Lu, Zhisheng Zhong, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Arpo: End-to-end policy optimization for gui agents with experience replay. arXiv preprint arXiv:2505.16282, 2025b.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. arXiv preprint arXiv:2505.18719, 2025c.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. Autopsv: Automated process-supervised verifier. Advances in Neural Information Processing Systems, 37:79935–79962, 2024.

Quanfeng Lu, Zhantao Ma, Shuai Zhong, Jin Wang, Dahai Yu, Michael K Ng, and Ping Luo. Swirl: A staged workflow for interleaved reinforcement learning in mobile gui control. arXiv preprint arXiv:2508.20018, 2025d.

Songshuo Lu, Hua Wang, Zhi Chen, and Yaohua Tang. Urpo: A unified reward & policy optimization framework for large language models. arXiv preprint arXiv:2507.17515, 2025e.

Yining Lu, Zilong Wang, Shiyang Li, Xin Liu, Changlong Yu, Qingyu Yin, Zhan Shi, Zixuan Zhang, and Meng Jiang. Learning to optimize multi-objective alignment through dynamic reward weighting, 2025f.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Guanjing Xiong, and Hongsheng Li. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. arXiv preprint arXiv:2503.21620, 2025g.

Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning. Science China Information Sciences, 67(2):121101, 2024.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. arXiv preprint arXiv:2501.12570, 2025a.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51, 2025b. Notion Blog.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, et al. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. Notion Blog, 2025c.

Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. arXiv preprint arXiv:2504.10458, 2025d.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K Qiu, and Yuqing Yang. Agent lightning: Train any ai agents with reinforcement learning. arXiv preprint arXiv:2508.03680, 2025e.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, et al. Towards a unified view of large language model post-training. arXiv preprint arXiv:2509.04419, 2025.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. Exploring the limit of outcome reward for learning mathematical reasoning. arXiv preprint arXiv:2502.06781, 2025.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. arXiv preprint arXiv:2506.07527, 2025a.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732, 2025b.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. arXiv preprint arXiv:2505.14652, 2025c.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. Generative reward models. arXiv preprint arXiv:2410.12832, 2024.

Justus Mattern, Sami Jaghouar, Manveer Basra, Jannik Straube, Matthew Di Ferrante, Felix Gabriel, Jack Min Ong, Vincent Weisser, and Johannes Hagemann. Synthetic-1: Two million collaboratively generated reasoning traces from deepseek-r1, 2025. URL https://www.primeintellect.ai/blog/synthetic-1-release.

Jianbiao Mei, Tao Hu, Daocheng Fu, Licheng Wen, Xuemeng Yang, Rong Wu, Pinlong Cai, Xinyu Cai, Xing Gao, Yu Yang, et al. O2-searcher: A searching-based agent model for open-domain open-ended question answering. arXiv preprint arXiv:2505.16582, 2025.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In The Twelfth International Conference on Learning Representations, 2023.

Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. Foundations and Trends® in Machine Learning, 16(1):1–118, 2023.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. arXiv preprint arXiv:2504.16891, 2025.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. arXiv preprint arXiv:2505.11711, 2025.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332, 2021.

Jaehyun Nam, Jinsung Yoon, Jiefeng Chen, Jinwoo Shin, Sercan Ö Arık, and Tomas Pfister. Mle-star: Machine learning engineering agent via search and targeted refinement. arXiv preprint arXiv:2506.15692, 2025.

Siddharth M. Narayanan, James D. Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G. Rodriques, and Andrew D. White. Training a scientific reasoning model for chemistry. arXiv preprint arXiv: 2506.17238, 2025.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. Mlgym: A new framework and benchmark for advancing ai research agents. arXiv preprint arXiv:2502.14499, 2025.

Xuan-Phi Nguyen, Shrey Pandit, Revanth Gangi Reddy, Austin Xu, Silvio Savarese, Caiming Xiong, and Shafiq Joty. Sfr-deepresearch: Towards effective reinforcement learning for autonomously reasoning single agents, 2025. URL https://arxiv.org/abs/2509.06283.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Datta Nimmaturi, Vaishnavi Bhargava, Rajat Ghosh, Johnu George, and Debojyoti Dutta. Predictive scaling laws for efficient grpo training of large reasoning models. arXiv preprint arXiv:2507.18014, 2025.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K. Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, et al. Virtual cells: Predict, explain, discover, 2025.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131, 2025.

Humza Nusrat. Autonomous radiotherapy treatment planning using dola: A privacy-preserving, llm-based optimization agent, 2025. URL https://arxiv.org/abs/2503.17553.

NVIDIA-NeMo. Nemo rl: A scalable and efficient post-training library. https://github.com/NVIDIA-NeMo/RL, 2025. GitHub repository.

Hayeon Oh. Laviplan: Language-guided visual path planning with rlvr. arXiv preprint arXiv:2507.12911, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. arXiv preprint arXiv:2501.00656, 2024.

OpenAI. Introducing gpt-4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2024. Accessed: 2025-08-25.

OpenAI. Gpt-5 system card. Blog, 2025a.

OpenAI. Openai o3 and o4-mini system card. Blog, 2025b.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35: 27730–27744, 2022.

Abby O' Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 6892–6903. IEEE, 2024.

Chaofan Pan, Xin Yang, Yanhua Li, Wei Wei, Tianrui Li, Bo An, and Jiye Liang. A survey of continual reinforcement learning. arXiv preprint arXiv:2506.21872, 2025a.

Jiadong Pan, Zhiyuan Ma, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Self-reflective reinforcement learning for diffusion-based image reasoning generation. arXiv preprint arXiv:2505.22407, 2025b.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training software engineering agents and verifiers with swe-gym. arXiv preprint arXiv:2412.21139, 2024.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025c. Accessed: 2025-01-24.

Jiazhen Pan, Che Liu, and Junde Wu. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning, 2025d. URL https://arxiv.org/abs/2502.19634.

Kaihang Pan, Wendong Bu, Yuruo Wu, Yang Wu, Kai Shen, Yunfei Li, Hang Zhao, Juncheng Li, Siliang Tang, and Yueting Zhuang. Focusdiff: Advancing fine-grained text-image alignment for autoregressive visual generation through rl. arXiv preprint arXiv:2506.05501, 2025e.

Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. arXiv preprint arXiv:2503.18470, 2025.

Dhruvi Paprunia, Vansh Kharidia, and Pankti Doshi. Advancing slm tool-use capability using reinforcement learning. arXiv preprint arXiv:2509.04518, 2025.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. arXiv preprint arXiv:2506.06632, 2025.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. arXiv preprint arXiv:2502.18439, 2025.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Reuse, don't retrain: A recipe for continued pretraining of language models. arXiv preprint arXiv:2407.07263, 2024.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíek, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Codeforces cots. https://huggingface.co/datasets/open-r1/codeforces-cots, 2025.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG), 37(4):1–14, 2018.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. Science, 378(6623):990–996, 2022.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, and et al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

Gabriel Poesia, David Broman, Nick Haber, and Noah Goodman. Learning formal mathematics from intrinsic motivation. Advances in Neural Information Processing Systems, 37:43032–43057, 2024.

Mohammadreza Pourreza, Shayan Talaei, Ruoxi Sun, Xingchen Wan, Hailong Li, Azalia Mirhoseini, Amin Saberi, Sercan Arik, et al. Reasoning-sql: Reinforcement learning with sql tailored partial rewards for reasoning-enhanced text-to-sql. arXiv preprint arXiv:2503.23157, 2025.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. arXiv preprint arXiv:2505.22660, 2025.

Ori Press, Ravid Shwartz-Ziv, Yann LeCun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. arXiv preprint arXiv:2405.05012, 2024.

PrimeIntellect. Synthetic-2 release: Four million collaboratively generated reasoning traces. https://www.primeintellect.ai/blog/synthetic-2-release#synthetic-2-dataset, 2025. Technical Report.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. arXiv preprint arXiv:2411.02337, 2024.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs. arXiv preprint arXiv:2504.13958, 2025.

Rushi Qiang, Yuchen Zhuang, Yinghao Li, Rongzhi Zhang, Changhao Li, Ian Shu-Hei Wong, Sherry Yang, Percy Liang, Chao Zhang, Bo Dai, et al. Mle-dojo: Interactive environments for empowering llm agents in machine learning engineering. arXiv preprint arXiv:2505.07782, 2025.

Chongli Qin and Jost Tobias Springenberg. Supervised fine tuning on curated data is reinforcement learning (and can be improved). arXiv preprint arXiv:2507.12856, 2025.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326, 2025.

Zhongxi Qiu, Zhang Zhang, Yan Hu, Heng Li, and Jiang Liu. Open-medical-r1: How to choose data for rlvr training at medicine domain, 2025. URL https://arxiv.org/abs/2504.13950.

Zipeng Qiu. Opentable-r1: A reinforcement learning augmented tool agent for open-domain table question answering. arXiv preprint arXiv:2507.03018, 2025.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. arXiv preprint arXiv:2503.21614, 2025a.

Yuxiao Qu, Matthew Y. R. Yang, Amrith Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. Optimizing test-time compute via meta reinforcement

finetuning. In Forty-second International Conference on Machine Learning, 2025b. URL https://openreview.net/forum?id=TqODUDsU4u.

Qwen Team. Qvq: To see the world with wisdom, 2025. URL https://qwenlm.github.io/blog/qvq-72b-preview.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems, 36:53728–53741, 2023.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q star: Your language model is secretly a q-function. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=kEVcNxtqXk.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. Journal of Machine Learning Research, 21(178):1–51, 2020.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pages 3505–3506, 2020.

Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. Magistral. arxiv preprint arXiv: 2506.10910, 2025.

ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. arXiv preprint arXiv:2504.21801, 2025.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Jeffrey Ishizuka, Rahul M Dhodapkar, Hattie Chung, Shekoofeh Azizi, Bryan Perozzi, and David van Dijk. Scaling large language models for next-generation single-cell analysis. bioRxiv: 2025.04.14.648850, 2025.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. Advances in neural information processing systems, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy reinforce: Stable and efficient reinforcement learning for llms. arXiv preprint arXiv:2503.14286, 2025.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523, 2025.

Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. arXiv preprint arXiv:2505.04769, 2025.

Ali Satvaty, Suzan Verberne, and Fatih Turkmen. Undesirable memorization in large language models: A survey. arXiv preprint arXiv:2410.02650, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36:68539–68551, 2023.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. Llms are greedy agents: Effects of rl fine-tuning on decision-making abilities. arXiv preprint arXiv:2504.16078, 2025.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. Nature, 588(7839):604–609, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889–1897. PMLR, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015b.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. arXiv preprint arXiv:1704.06440, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017b.

ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, et al. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. arXiv preprint arXiv:2504.13914, 2025a.

ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, et al. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning, 2025b.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. arXiv preprint arXiv:2507.05201, 2025.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. arXiv preprint arXiv:2410.08146, 2024.

Amrith Setlur, Matthew YR Yang, Charlie Snell, Jeremy Greer, Ian Wu, Virginia Smith, Max Simchowitz, and Aviral Kumar. e3: Learning to explore enables extrapolation of test-time compute for llms. arXiv preprint arXiv:2506.09026, 2025.

Zeyang Sha, Shiwen Cui, and Weiqiang Wang. Sem: Reinforcement learning for search-efficient large language models. arXiv preprint arXiv:2505.07903, 2025.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train? arXiv preprint arXiv:2505.21444, 2025.

Shijie Shang, Ruosi Wan, Yue Peng, Yutong Wu, Xiong-hui Chen, Jie Yan, and Xiangyu Zhang. Stepfun-prover preview: Let's think and verify step by step. arXiv preprint arXiv:2507.20199, 2025.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious rewards: Rethinking training signals in rlvr. arXiv preprint arXiv:2506.10947, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538, 2017.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615, 2025a.

Wei Shen, Jiangbo Pei, Yi Peng, Xuchen Song, Yang Liu, Jian Peng, Haofeng Sun, Yunzhuo Hao, Peiyu Wang, Jianhao Zhang, et al. Skywork-r1v3 technical report. arXiv preprint arXiv:2507.06167, 2025b.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less. arXiv preprint arXiv:2509.04259, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In Proceedings of the Twentieth European Conference on Computer Systems, pages 1279–1297, 2025.

Jiajun Shi, Jian Yang, Jiaheng Liu, Xingyuan Bu, Jiangjie Chen, Junting Zhou, Kaijing Ma, Zhoufutu Wen, Bingli Wang, Yancheng He, et al. Korgym: A dynamic game platform for llm reasoning evaluation. arXiv preprint arXiv:2505.14552, 2025a.

Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. arXiv preprint arXiv:2504.05520, 2025b.

Yucheng Shi, Wenhao Yu, Zaitang Li, Yonglin Wang, Hongming Zhang, Ninghao Liu, Haitao Mi, and Dong Yu. Mobilegui-rl: Advancing mobile gui agent through reinforcement learning in online environment. arXiv preprint arXiv:2507.05720, 2025c.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In International Conference on Learning Representations, 2020.

Vaishnavi Shrivastava, Ahmed Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. Sample more to think less: Group filtered policy optimization for concise reasoning. arXiv preprint arXiv:2508.09726, 2025.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. arXiv preprint arXiv:2305.17493, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. Nature, 631(8022):755–759, 2024.

David Silver and Richard S Sutton. Welcome to the era of experience. Google AI, 1, 2025.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science, 362(6419): 1140–1144, 2018.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. Artificial intelligence, 299:103535, 2021.

SimpleVLA-RL Team. Simplevla-rl: Online rl with simple reward enables training vla models with only one trajectory. https://github.com/PRIME-RL/SimpleVLA-RL, 2025. GitHub repository.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In Forty-second International Conference on Machine Learning, 2025.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in llms. arXiv preprint arXiv:2509.09677, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314, 2024.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. arXiv preprint arXiv:2503.05592, 2025a.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. arXiv preprint arXiv:2505.17005, 2025b.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. arXiv preprint arXiv:2508.02193, 2025c.

Saksham Sahai Srivastava and Vaneet Aggarwal. A technical survey of reinforcement learning techniques for large language models. arXiv preprint arXiv:2507.04136, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. Advances in neural information processing systems, 33:3008–3021, 2020.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards. arXiv preprint arXiv:2505.24760, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966, 2025a.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. arXiv preprint arXiv:2505.00127, 2025b.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. arXiv preprint arXiv:2503.23829, 2025c.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617, 2025d.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918, 2025e.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. arXiv preprint arXiv:2503.16419, 2025.

Hao Sun. Supervised fine-tuning as inverse reinforcement learning. arXiv preprint arXiv:2403.12017, 2024.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of llms without searching. arXiv preprint arXiv:2505.04588, 2025a.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. ACM Computing Surveys, 57(11):1–43, 2025b.

Lin Sun, Chuang Liu, Xiaofeng Ma, Tao Yang, Weijia Lu, and Ning Wu. Freeprm: Training process reward models without ground truth process labels. arXiv preprint arXiv:2506.03570, 2025c.

Shengjie Sun, Runze Liu, Jiafei Lyu, Jing-Wen Yang, Liangpeng Zhang, and Xiu Li. A large language model-driven reward design framework via dynamic feedback for reinforcement learning. Knowledge-Based Systems, 326:114065, 2025d. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.114065. URL https://www.sciencedirect.com/science/article/pii/S0950705125011104.

Yifan Sun, Jingyan Shen, Yibin Wang, Tianyu Chen, Zhendong Wang, Mingyuan Zhou, and Huan Zhang. Improving data efficiency for llm reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. arXiv preprint arXiv:2506.05316, 2025e.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. arXiv preprint arXiv:2506.09513, 2025f.

Zetian Sun, Dongfang Li, Zhuoen Chen, Yuhuai Qin, and Baotian Hu. Stabilizing long-term multi-turn reinforcement learning with gated rewards. arXiv preprint arXiv:2508.10548, 2025g.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. arXiv preprint arXiv:2505.12886, 2025h.

Zhoujian Sun, Ziyi Liu, Cheng Luo, Jiebin Chu, and Zhengxing Huang. Improving interactive diagnostic ability of a large language model agent through clinical experience learning, 2025i. URL https://arxiv.org/abs/2503.16463.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296, 2017.

Richard S Sutton, Andrew G Barto, et al. Introduction to reinforcement learning, volume 135. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, 12, 1999.

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. arXiv preprint arXiv:2503.01067, 2025.

Jaesung Tae, Hamish Ivison, Sachin Kumar, and Arman Cohan. Tess 2: A large-scale generalist diffusion language model. arXiv preprint arXiv:2502.13917, 2025.

Hongze Tan and Jianfei Pan. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. arXiv preprint arXiv:2508.04349, 2025.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. arXiv preprint arXiv:2505.17016, 2025a.

Wentao Tan, Qiong Cao, Chao Xue, Yibing Zhan, Changxing Ding, and Xiaodong He. Chartmaster: Advancing chart-to-code generation with real-world charts and chart similarity reinforcement learning. arXiv preprint arXiv:2508.17608, 2025b.

Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, et al. Agent kb: Leveraging cross-domain experience for agentic problem solving. arXiv preprint arXiv:2507.06229, 2025.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, et al. Webshaper: Agentically data synthesizing via information-seeking formalization. arXiv preprint arXiv:2507.15061, 2025.

ByteDance Seed Team. Seed-oss open-source models, 2025a. URL https://github.com/ByteDance-Seed/seed-oss.

Dolphin Team. Dolphin r1 dataset. https://huggingface.co/datasets/QuixiAI/dolphin-r1, 2025b. URL https://huggingface.co/datasets/QuixiAI/dolphin-r1. Dataset, Apache-2.0 license.

GLM-V. Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, et al. GLM-4.5v and GLM-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025a.

Kimi Team. Kimi k2: Open agentic intelligence, 2025c. URL https://arxiv.org/abs/2507.20534.

Kimi Team. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025d.

MiroMind AI Team. Mirothinker: An open-source agentic model series trained for deep research and complex, long-horizon problem solving. https://github.com/MiroMindAI/MiroThinker, 2025e.

MiroMind Data Team. Miroverse v0.1: A reproducible, full-trajectory, ever-growing deep research dataset, 2025f. URL https://huggingface.co/datasets/miromind-ai/MiroVerse-v0.1.

Prime Intellect Team, Sami Jaghouar, Justus Mattern, Jack Min Ong, Jannik Straube, Manveer Basra, Aaron Pazdera, Kushal Thaman, Matthew Di Ferrante, Felix Gabriel, et al. Intellect-2: A reasoning model trained through globally decentralized reinforcement learning. arXiv preprint arXiv:2505.07291, 2025b.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025g. URL https://qwenlm.github.io/blog/qwq-32b/.

RLinf Team. Rlinf: Reinforcement learning infrastructure for agentic ai. https://github.com/RLinf/RLinf, 2025h. GitHub repository.

Tencent Hunyuan Team, Ao Liu, Botong Zhou, Can Xu, Chayse Zhou, ChenChen Zhang, Chengcheng Xu, Chenhao Wang, Decheng Wu, Dengpeng Wu, et al. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. arXiv preprint arXiv:2505.15431, 2025c.

THUDM. slime: An sglang-native post-training framework for rl scaling. https://github.com/THUDM/slime, 2025. GitHub repository.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. arXiv preprint arXiv:2506.13654, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ international conference on intelligent robots and systems, pages 5026–5033. IEEE, 2012.

tokenbender. avatarl: training language models from scratch with pure reinforcement learning, 2025. URL https://github.com/tokenbender/avatarl.

Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. arXiv preprint arXiv:2505.17017, 2025a.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Code2logic: Game-code-driven data synthesis for enhancing vlms general reasoning. arXiv preprint arXiv:2505.13886, 2025b.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032, 2024.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16022–16076, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. arXiv preprint arXiv:2211.14275, 2022.

Carel van Niekerk, Renato Vukovic, Benjamin Matthias Ruppik, Hsien-chin Lin, and Milica Gašić. Post-training large language models via reinforcement learning from self-feedback. arXiv preprint arXiv:2507.21931, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Dheeraj Vattikonda, Santhoshi Ravichandran, Emiliano Penaloza, Hadi Nekoei, Megh Thakkar, Thibault Le Sellier de Chezelles, Nicolas Gontier, Miguel Muñoz-Mármol, Sahar Omidi Shayegan, Stefania Raimondo, et al. How to train your llm web agent: A statistical diagnosis. arXiv preprint arXiv:2507.04103, 2025.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. arXiv preprint arXiv:2211.04325, 2022.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. arXiv preprint arXiv:2507.18624, 2025.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Christian Walder and Deep Karkhanis. Pass@k policy optimization: Solving harder reinforcement learning problems. arXiv preprint arXiv:2505.15201, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8228–8238, 2024.

Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, et al. Rema: Learning to meta-think for llms with multi-agent reinforcement learning. arXiv preprint arXiv:2503.09501, 2025.

Bin Wang, Bojun Wang, Changyi Wan, Guanzhe Huang, Hanpeng Hu, Haonan Jia, Hao Nie, Mingliang Li, Nuo Chen, Siyu Chen, et al. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding. arXiv preprint arXiv:2507.19427, 2025a.

Chen Wang, Lai Wei, Yanzhi Zhang, Chenyang Shao, Zedong Dan, Weiran Huang, Yue Wang, and Yuzhi Zhang. Eframe: Deeper reasoning via exploration-filtering-replay reinforcement learning framework. arXiv preprint arXiv:2506.22200, 2025b.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Qiaozhi He, Murun Yang, Bei Li, Tong Xiao, Chunliang Zhang, Tongran Liu, et al. Gram: A generative foundation reward model for reward generalization. arXiv preprint arXiv:2506.14175, 2025c.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, MD Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, et al. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning, 2025. URL https://arxiv. org/abs/2504.11354, 2025d.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution. arXiv preprint arXiv:2505.20732, 2025e.

Hanyin Wang. Reinforcement learning for out-of-distribution reasoning in llms: An empirical study on diagnosis-related group coding, 2025. URL https://arxiv.org/abs/2505.21908.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. arXiv preprint arXiv:2509.02544, 2025f.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. Emergent hierarchical reasoning in llms through reinforcement learning. arXiv preprint arXiv:2509.03646, 2025g.

Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, and Guorui Zhou. Asymmetric dual-clipping: Unleashing the full potential of rl in llm training. https://rogue-canopy-54a.notion.site/Asymmetric-Dual-Clipping-Unleashing-the-Full-Potential-of-RL-in-LLM-Training-2650e4c8c16a8034a5d3dfec358c9021, 2025h. Notion Blog.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. arXiv preprint arXiv:2507.15778, 2025i.

Jicheng Wang, Yifeng He, and Hao Chen. Repogenreflex: Enhancing repository-level code completion with verbal reinforcement and retrieval-augmented generation. arXiv preprint arXiv:2409.13122, 2024a.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. arXiv preprint arXiv:2504.11455, 2025j.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9426–9439, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning, 2025k.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. arXiv preprint arXiv:2505.12434, 2025l.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11279–11298, 2022.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. arXiv preprint arXiv:2506.01939, 2025m.

Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, et al. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. arXiv preprint arXiv:2506.06122, 2025n.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025o.

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 37–53, 2018.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6629–6638, 2019.

Yanlin Wang, Yanli Wang, Daya Guo, Jiachi Chen, Ruikai Zhang, Yuchi Ma, and Zibin Zheng. Rlcoder: Reinforcement learning for repository-level code completion. arXiv preprint arXiv:2407.19487, 2024c.

Yinjie Wang, Ling Yang, Bowen Li, Ye Tian, Ke Shen, and Mengdi Wang. Revolutionizing reinforcement learning framework for diffusion large language models. arXiv preprint arXiv:2509.06949, 2025p.

Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. Co-evolving llm coder and unit tester via reinforcement learning. arXiv preprint arXiv:2506.03136, 2025q.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. arXiv preprint arXiv:2504.20571, 2025r.

Yiting Wang, Guoheng Sun, Wanghao Ye, Gang Qu, and Ang Li. Verireason: Reinforcement learning with testbench feedback for reasoning-enhanced verilog generation. arXiv preprint arXiv:2505.11849, 2025s.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. Thoughts are all over the place: On the underthinking of o1-like llms. arXiv preprint arXiv:2501.18585, 2025t.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. arXiv preprint arXiv:2506.20512, 2025u.

Zhilin Wang, Zhe Yang, Yun Luo, Yafu Li, Haoran Zhang, Runzhe Zhan, Derek F Wong, Jizhe Zhou, and Yu Cheng. Synthesizing sheet music problems for evaluation and reinforcement learning. arXiv preprint arXiv:2509.04059, 2025v.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. arXiv preprint arXiv:2505.15107, 2025w.

Yuyang Wanyan, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Jiabo Ye, Yutong Kou, Ming Yan, Fei Huang, Xiaoshan Yang, et al. Look before you leap: A gui-critic-r1 model for pre-operative error diagnosis in gui automation. arXiv preprint arXiv:2506.04614, 2025.

Jason Wei. The asymmetry of verification and verifier's law. https://www.jasonwei.net/blog/asymmetry-of-verification-and-verifiers-law, 2025. Accessed: 2025-07-15.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. arXiv preprint arXiv:2504.12516, 2025a.

Yifan Wei, Xiaoyan Yu, Yixuan Weng, Tengfei Pan, Angsheng Li, and Li Du. Autotir: Autonomous tools integrated reasoning via reinforcement learning. arXiv preprint arXiv:2507.21836, 2025b.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I Wang. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. arXiv preprint arXiv:2502.18449, 2025c.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, et al. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. arXiv preprint arXiv:2505.16421, 2025d.

Hao Wen, Xinrui Wu, Yi Sun, Feifei Zhang, Liye Chen, Jie Wang, Yunxin Liu, Ya-Qin Zhang, and Yuanchun Li. Budgetthinker: Empowering budget-aware llm reasoning with control tokens. arXiv preprint arXiv:2508.17196, 2025a.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. arXiv preprint arXiv:2503.10460, 2025b.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. arXiv preprint arXiv:2506.14245, 2025c.

Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. arXiv preprint arXiv:2505.10320, 2025.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3):229–256, 1992.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. Connection Science, 3(3):241–268, 1991.

Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. Advances in Neural Information Processing Systems, 34:28496–28510, 2021.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025a.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747, 2025b.

Haoyuan Wu, Xueyi Chen, Rui Ming, Jilong Gao, Shoubo Hu, Zhuolun He, and Bei Yu. Totrl: Unlock llm tree-of-thoughts reasoning potential through puzzles solving. arXiv preprint arXiv:2505.12717, 2025c.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, et al. Webdancer: Towards autonomous information seeking agency. arXiv preprint arXiv:2505.22648, 2025d.

Lixin Wu, Na Cai, Qiao Cheng, Jiachen Wang, and Yitao Duan. Confucius3-math: A lightweight high-performance reasoning llm for chinese k-12 mathematics learning. arXiv preprint arXiv:2506.18330, 2025e.

Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han, Hao Sun, Jiayi Ji, et al. Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. arXiv preprint arXiv:2505.17540, 2025f.

Penghao Wu, Shengnan Ma, Bo Wang, Jiaheng Yu, Lewei Lu, and Ziwei Liu. Gui-reflection: Empowering multimodal gui models with self-reflection behavior. arXiv preprint arXiv:2506.08012, 2025g.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. arXiv preprint arXiv:2407.19594, 2024.

Weijia Wu, Chen Gao, Joya Chen, Kevin Qinghong Lin, Qingwei Meng, Yiming Zhang, Yuke Qiu, Hong Zhou, and Mike Zheng Shou. Reinforcement learning in vision: A survey. arXiv preprint arXiv:2508.08189, 2025h.

Xiaobao Wu. Sailing by the stars: A survey on reward models and learning strategies for learning from rewards. arXiv preprint arXiv:2505.02686, 2025.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of sft: A reinforcement learning perspective with reward rectification. arXiv preprint arXiv:2508.05629, 2025i.

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. arXiv preprint arXiv:2505.14677, 2025a.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. arXiv preprint arXiv:2506.00555, 2025b.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. arXiv preprint arXiv:2404.15676, 2024.

Yunhui Xia, Wei Shen, Yan Wang, Jason Klein Liu, Huifeng Sun, Siyue Wu, Jian Hu, and Xiaolong Xu. Leetcodedataset: A temporal dataset for robust evaluation and efficient training of code llms. arXiv preprint arXiv:2504.14655, 2025c.

Violet Xiang, Chase Blagden, Rafael Rafailov, Nathan Lile, Sang Truong, Chelsea Finn, and Nick Haber. Just enough thinking: Efficient reasoning with adaptive length penalties reinforcement learning. arXiv preprint arXiv:2506.05256, 2025.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. Bnpo: Beta normalization policy optimization. arXiv preprint arXiv:2506.02864, 2025a.

Teng Xiao, Yige Yuan, Mingxiao Li, Zhengyu Chen, and Vasant G Honavar. On a connection between imitation learning and rlhf. arXiv preprint arXiv:2503.05079, 2025b.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. arXiv preprint arXiv:2505.07608, 2025.

Chengxing Xie, Bowen Li, Chang Gao, He Du, Wai Lam, Difan Zou, and Kai Chen. Swe-fixer: Training open-source llms for effective and efficient github issue resolution. arXiv preprint arXiv:2501.05040, 2025a.

Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. Capo: Towards enhancing llm reasoning through verifiable generative credit assignment. arXiv preprint arXiv:2508.02298, 2025b.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2502.14768, 2025c.

Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. arXiv preprint arXiv:2309.11489, 2023.

Yunfei Xie, Yinsong Ma, Shiyi Lan, Alan Yuille, Junfei Xiao, and Chen Wei. Play to generalize: Learning to reason through game play. arXiv preprint arXiv:2506.08011, 2025d.

Zhihui Xie, Liyu Chen, Weichao Mao, Jingjing Xu, Lingpeng Kong, et al. Teaching language models to critique via reinforcement learning. arXiv preprint arXiv:2502.03492, 2025e.

Zhihui Xie, Jiacheng Ye, Lin Zheng, Jiahui Gao, Jingwei Dong, Zirui Wu, Xueliang Zhao, Shansan Gong, Xin Jiang, Zhenguo Li, et al. Dream-coder 7b: An open diffusion language model for code. arXiv preprint arXiv:2509.01142, 2025f.

Rihui Xin, Han Liu, Zecheng Wang, Yupeng Zhang, Dianbo Sui, Xiaolin Hu, and Bingning Wang. Surrogate signals from format and length: Reinforcement learning for solving mathematical problems without ground truth answers. arXiv preprint arXiv:2505.19439, 2025.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, et al. A minimalist approach to llm reasoning: from rejection sampling to reinforce. arXiv preprint arXiv:2504.11343, 2025a.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. Self-rewarding correction for mathematical reasoning. arXiv preprint arXiv:2502.19613, 2025b.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. arXiv preprint arXiv:2501.09686, 2025a.

Huihui Xu and Yuanpeng Nie. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization, 2025. URL https://arxiv.org/abs/2507.02994.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, Tao Wang, et al. Medagentgym: Training llm agents for code-based medical reasoning at scale. arXiv preprint arXiv:2506.04405, 2025b.

Wenyuan Xu, Xiaochen Zuo, Chao Xin, Yu Yue, Lin Yan, and Yonghui Wu. A unified pairwise framework for rlhf: Bridging generative reward modeling and policy optimization. arXiv preprint arXiv:2504.04950, 2025c.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. arXiv preprint arXiv:2502.12110, 2025d.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. arXiv preprint arXiv:2505.11409, 2025e.

Yuyang Xu, Yi Cheng, Haochao Ying, Zhuoyun Du, Renjun Hu, Xing Shi, Wei Lin, and Jian Wu. Sspo: Self-traced step-wise preference optimization for process supervision and reasoning compression. arXiv preprint arXiv:2508.12604, 2025f.

Zhangchen Xu, Yuetai Li, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Tinyv: Reducing false negatives in verification improves rl for llm reasoning. arXiv preprint arXiv:2505.14625, 2025g.

Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. arXiv preprint arXiv:2503.02951, 2025h.

Zhongwen Xu and Zihan Ding. Single-stream policy optimization, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818, 2025a.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. arXiv preprint arXiv:2509.02479, 2025b.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. arXiv preprint arXiv:2504.14945, 2025a.

Kaiwen Yan, Xuanqing Shi, Hongcheng Guo, Wenxuan Wang, Zhuosheng Zhang, and Chengwei Qin. Drqa: Dynamic reasoning quota allocation for controlling overthinking in reasoning large language models. arXiv preprint arXiv:2508.17803, 2025b.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024a.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arxiv preprint arXiv: 2505.09388, 2025a.

Chenyu Yang, Shiqian Su, Shi Liu, Xuan Dong, Yue Yu, Weijie Su, Xuehui Wang, Zhaoyang Liu, Jinguo Zhu, Hao Li, et al. Zerogui: Automating online gui learning at zero human cost. arXiv preprint arXiv:2505.23762, 2025b.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 10632–10643, 2025c.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8941–8951, 2024b.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809, 2025d.

Sherry Yang, Joy He-Yueya, and Percy Liang. Reinforcement learning for machine learning engineering agents. arXiv preprint arXiv:2509.01684, 2025e.

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. Ssr-zero: Simple self-rewarding reinforcement learning for machine translation. arXiv preprint arXiv:2505.16637, 2025f.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In International conference on machine learning, pages 5571–5580. PMLR, 2018.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. Treerpo: Tree relative policy optimization. arXiv preprint arXiv:2506.05183, 2025g.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Yiwei Wang, Xiaodan Liang, and Jing Tang. Depth-breadth synergy in rlvr: Unlocking llm reasoning gains with adaptive exploration. arXiv preprint arXiv:2508.13755, 2025h.

Zongxian Yang, Jiayu Qian, Zegao Peng, Haoyu Zhang, and Zhi-An Huang. Med-refl: Medical reasoning enhancement via self-corrected fine-grained reflection, 2025i. URL https://arxiv.org/abs/2506.13793.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient rl framework secretly brings you off-policy rl training, August 2025a. URL https://fengyao.notion.site/off-policy-rl.

Feng Yao, Zilong Wang, Liyuan Liu, Junxia Cui, Li Zhong, Xiaohan Fu, Haohui Mai, Vish Krishnan, Jianfeng Gao, and Jingbo Shang. Training language models to generate quality code with program analysis feedback. arXiv preprint arXiv:2505.22704, 2025b.

Chenlu Ye, Zhou Yu, Ziji Zhang, Hao Chen, Narayanan Sadagopan, Jing Huang, Tong Zhang, and Anurag Beniwal. Beyond correctness: Harmonizing process and outcome rewards through rl training. arXiv preprint arXiv:2509.03403, 2025a.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, et al. Mobile-agent-v3: Foundamental agents for gui automation. arXiv preprint arXiv:2508.15144, 2025b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. arXiv preprint arXiv:2508.15487, 2025c.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. arXiv preprint arXiv:2502.03387, 2025d.

Cheng-Kai Yeh, Hsing-Wang Lee, Chung-Hung Kuo, and Hen-Hsen Huang. Ar2: Adversarial reinforcement learning for abstract reasoning in large language models. arXiv preprint arXiv:2509.03537, 2025.

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, and Xuanjing Huang. Dynamic and generalizable process reward modeling. arXiv preprint arXiv:2507.17849, 2025.

ByoungJun Jeon Yooseok Lim. More-clear: Multimodal offline reinforcement learning for clinical notes leveraged enhanced state representation, 2025. URL https://arxiv.org/abs/2508.07681.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, et al. Medreseacher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. arXiv preprint arXiv:2508.14880, 2025a.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. Advances in neural information processing systems, 35:24611–24624, 2022.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. arXiv preprint arXiv:2507.02259, 2025b.

Hongzhou Yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. Finemedlm-o1: Enhancing medical knowledge reasoning ability of llm from supervised fine-tuning to test-time training, 2025c. URL https://arxiv.org/abs/2501.09213.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476, 2025d.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, et al. Rlpr: Extrapolating rlvr to general domains without verifiers. arXiv preprint arXiv:2506.18254, 2025e.

Zhaojian Yu, Yinghao Wu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. Z1: Efficient test-time scaling with code. arXiv preprint arXiv:2504.00810, 2025f.

Danlong Yuan, Tian Xie, Shaohan Huang, Zhuocheng Gong, Huishuai Zhang, Chong Luo, Furu Wei, and Dongyan Zhao. Efficient rl training for reasoning models via length-aware optimization. arXiv preprint arXiv:2505.12284, 2025a.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. arXiv preprint arXiv:2502.11089, 2025b.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From f(x) and g(x) to f(g(x)): LLMs learn new skills in RL by composing old ones. https://husky-morocco-f72.notion.site/From-f-x-and-g-x-to-f-g-x-LLMs-Learn-New-Skills-in-RL-by-Composing-Old-Ones-2499aba4486f802c8108e76a12af3020, 2025c. Notion blog post, available online.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In Forty-second International Conference on Machine Learning, 2025d. URL https://openreview.net/forum?id=8ThnPFhGm8.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. arXiv preprint arXiv:2401.10020, 3, 2024.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Ilia Kulikov, Kyunghyun Cho, Dong Wang, Yuandong Tian, Jason E Weston, et al. Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions. arXiv preprint arXiv:2502.13124, 2025e.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. arXiv preprint arXiv:2503.01491, 2025f.

Chuhuai Yue, Chengqi Dong, Yinan Gao, Hang He, Jiajun Chai, Guojun Yin, and Wei Lin. Promoting efficient reasoning with verifiable stepwise reward. arXiv preprint arXiv:2508.10293, 2025a.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? arXiv preprint arXiv:2504.13837, 2025b.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. arXiv preprint arXiv:2504.05118, 2025c.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. arXiv preprint arXiv:2508.06471, 2025a.

Guangtao Zeng, Maohao Shen, Delin Chen, Zhenting Qi, Subhro Das, Dan Gutfreund, David Cox, Gregory Wornell, Wei Lu, Zhang-Wei Hong, et al. Satori-swe: Evolutionary test-time scaling for sample-efficient software engineering. arXiv preprint arXiv:2505.23604, 2025b.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, et al. Reviewrl: Towards automated scientific review with rl. arXiv preprint arXiv:2508.10308, 2025c.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. arXiv preprint arXiv:2505.11821, 2025d.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. arXiv preprint arXiv:2503.18892, 2025e.

Kaiwen Zha, Zhengqi Gao, Maohao Shen, Zhang-Wei Hong, Duane S Boning, and Dina Katabi. Rl tango: Reinforcing generator and verifier together for language reasoning. arXiv preprint arXiv:2505.15034, 2025.

Bin Zhang, Hangyu Mao, Jingqing Ruan, Ying Wen, Yang Li, Shao Zhang, Zhiwei Xu, Dapeng Li, Ziyue Li, Rui Zhao, et al. Controlling large language model-based agents for large-scale decision-making: An actor-critic approach. arXiv preprint arXiv:2311.13884, 2023a.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. Proagent: Building proactive cooperative ai with large language models. CoRR, 2023b.

Chong Zhang, Yue Deng, Xiang Lin, Bin Wang, Dianwen Ng, Hai Ye, Xingxuan Li, Yao Xiao, Zhanfeng Mo, Qi Zhang, et al. 100 days after deepseek-r1: A survey on replication studies and more directions for reasoning language models. arXiv preprint arXiv:2505.00551, 2025a.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Yue Liao, Hongru Wang, Mengyue Yang, Heng Ji, Michael Littman, Jun Wang, Shuicheng Yan, Philip Torr, and Lei Bai. The landscape of agentic reinforcement learning for llms: A survey, 2025b. URL https://arxiv.org/abs/2509.02547.

Han Zhang, Ruibin Zheng, Zexuan Yi, Hanyang Peng, Hui Wang, and Yue Yu. Group expectation policy optimization for stable heterogeneous reinforcement learning in llms. arXiv preprint arXiv:2508.17850, 2025c.

Hongzhi Zhang, Jia Fu, Jingyuan Zhang, Kai Fu, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Rlep: Reinforcement learning with experience replay for llm reasoning. arXiv preprint arXiv:2507.07451, 2025d.

Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. arXiv preprint arXiv:2504.09696, 2025.

Kaiyan Zhang, Runze Liu, Xuekai Zhu, Kai Tian, Sihang Zeng, Guoli Jia, Yuchen Fan, Xingtai Lv, Yuxin Zuo, Che Jiang, Ziyang Liu, Jianyu Wang, Yuru Wang, Ruotong Zhao, Ermo Hua, Yibo Wang, Shijie Wang, Junqi Gao, Xinwei Long, Youbang Sun, Zhiyuan Ma, Ganqu Cui, Lei Bai, Ning Ding, Biqing Qi, and Bowen Zhou. Marti: A framework for multi-agent llm systems reinforced training and inference, 2025e. URL https://github.com/TsinghuaC3I/MARTI.

Kaiyan Zhang, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, and Bowen Zhou. OpenPRM: Building open-domain process-based reward models with preference trees. In The Thirteenth International Conference on Learning Representations, 2025f. URL https://openreview.net/forum?id=fGIqGfmgkW.

Kaiyi Zhang, Ang Lv, Jinpeng Li, Yongbo Wang, Feng Wang, Haoyuan Hu, and Rui Yan. Stephint: Multi-level stepwise hints enhance reinforcement learning to reason. arXiv preprint arXiv:2507.02841, 2025g.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. arXiv preprint arXiv:2506.08745, 2025h.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. arXiv preprint arXiv:2408.15240, 2024a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. arXiv preprint arXiv:2504.05812, 2025i.

Ruize Zhang, Zelai Xu, Chengdong Ma, Chao Yu, Wei-Wei Tu, Wenhao Tang, Shiyu Huang, Deheng Ye, Wenbo Ding, Yaodong Yang, et al. A survey on self-play methods in reinforcement learning. arXiv preprint arXiv:2408.01072, 2024b.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. arXiv preprint arXiv:2502.19655, 2025j.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. arXiv preprint arXiv:2508.11408, 2025k.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. arXiv preprint arXiv:2506.12307, 2025l.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. arXiv preprint arXiv:2506.03106, 2025m.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. arXiv preprint arXiv:2505.15436, 2025n.

Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. Vq-insight: Teaching vlms for ai-generated video quality understanding via progressive visual reinforcement learning. arXiv preprint arXiv:2506.18564, 2025o.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. Bread: Branched rollouts from expert anchors bridge sft and rl for reasoning. arXiv preprint arXiv:2506.17211, 2025p.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for llm reasoning. arXiv preprint arXiv:2506.17219, 2025q.

Yifan Zhang, Yifeng Liu, Huizhuo Yuan, Yang Yuan, Quanquan Gu, and Andrew C Yao. On the design of kl-regularized policy gradient algorithms for llm reasoning. arXiv preprint arXiv:2505.17508, 2025r.

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, et al. Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning. arXiv preprint arXiv:2507.17842, 2025s.

Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl. arXiv preprint arXiv:2505.24875, 2025t.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. arXiv preprint arXiv:2501.07301, 2025u.

Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, et al. Agentcpm-gui: Building mobile-use agents with reinforcement fine-tuning. arXiv preprint arXiv:2506.01391, 2025v.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. Co-reward: Self-supervised reinforcement learning for large language model reasoning via contrastive agreement. arXiv preprint arXiv:2508.00410, 2025w.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. arXiv preprint arXiv:2505.03335, 2025a.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, et al. Genprm: Scaling test-time compute of process reward models via generative reasoning. arXiv preprint arXiv:2504.00891, 2025b.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning, 2025c. URL https://arxiv.org/abs/2504.12216.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2), 2023a.

Weikang Zhao, Xili Wang, Chengdi Ma, Lingbin Kong, Zhaohua Yang, Mingxiang Tuo, Xiaowei Shi, Yitao Zhai, and Xunliang Cai. Mua-rl: Multi-turn user-interacting agent reinforcement learning for agentic tool use. arXiv preprint arXiv:2508.18669, 2025d.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. arXiv preprint arXiv:2505.19590, 2025e.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277, 2023b.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. arXiv preprint arXiv:2507.20673, 2025f.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. arXiv preprint arXiv:2507.18071, 2025a.

Haizhong Zheng, Yang Zhou, Brian R Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. arXiv preprint arXiv:2506.02177, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36:46595–46623, 2023.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, et al. First return, entropy-eliciting explore. arXiv preprint arXiv:2507.07017, 2025c.

Xuejing Zheng, Chao Yu, and Minjie Zhang. Lifelong reinforcement learning with temporal logic formulas and reward machines. Knowledge-Based Systems, 257:109650, 2022.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025d.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. arXiv preprint arXiv:2504.03160, 2025e.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362, 2025f.

Weihai Zhi, Jiayan Guo, and Shangyang Li. Medgrš: Breaking the data barrier for medical reasoning via generative reward learning. arXiv preprint arXiv:2508.20549, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In 38th AAAI Conference on Artificial Intelligence, AAAI 2024, Feb 20-27 2024 Vancouver, Canada, volume 38, pages 19724–19731. Association for the Advancement of Artificial Intelligence (AAAI), 2024.

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, et al. A survey on vision-language-action models: An action tokenization perspective. arXiv preprint arXiv:2507.01925, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308, 2025a.

Junting Zhou, Wang Li, Yiyan Liao, Nengyuan Zhang, Tingjia Miaoand Zhihui Qi, Yuhan Wu, and Tong Yang. Academicbrowse: Benchmarking academic browse ability of llms. arXiv preprint arXiv:2506.13784, 2025b.

Meng Zhou, Bei Li, Jiahao Liu, Xiaowen Shi, Yang Bai, Rongxiang Weng, Jingang Wang, and Xunliang Cai. Libra: Assessing and improving reward model by learning to think. arXiv preprint arXiv:2507.21645, 2025c.

Ruochen Zhou, Minrui Xu, Shiqi Chen, Junteng Liu, Yunqi Li, Xinxin Lin, Zhengyu Chen, and Junxian He. Does learning mathematical problem-solving generalize to broader reasoning? arXiv preprint arXiv:2507.04391, 2025d.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. arXiv preprint arXiv:2505.21493, 2025e.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Jiale Zhao, Jingwen Yang, Jianwei Lv, Kongcheng Zhang, Yihe Zhou, Hengtong Lu, et al. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. arXiv preprint arXiv:2508.16949, 2025f.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. arXiv preprint arXiv:2503.15478, 2025g.

Yuqi Zhou, Sunhao Dai, Shuai Wang, Kaiwen Zhou, Qinglin Jia, and Jun Xu. Gui-g1: Understanding r1-zero-like training for visual grounding in gui agents. arXiv preprint arXiv:2505.15810, 2025h.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. arXiv preprint arXiv:2506.15841, 2025i.

Dingwei Zhu, Shihan Dou, Zhiheng Xi, Senjie Jin, Guoqiang Zhang, Jiazheng Zhang, Junjie Ye, Mingxu Chai, Enyu Zhou, Ming Zhang, et al. Vrpo: Rethinking value modeling for robust rl training under noisy supervision. arXiv preprint arXiv:2508.03058, 2025a.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, et al. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. arXiv preprint arXiv:2505.19223, 2025b.

Hui Zhu, Xv Wang, Zhenyu Wang, and Kai Xv. An emotion-sensitive dialogue policy for task-oriented dialogue system. Scientific Reports, 14(1):19759, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025c.

Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. arXiv preprint arXiv:2502.16906, 2025d.

Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. Proximal supervised fine-tuning. arXiv preprint arXiv:2508.17784, 2025e.

Yaoyu Zhu, Di Huang, Hanqi Lyu, Xiaoyun Zhang, Chongxiao Li, Wenxuan Shi, Yutong Wu, Jianan Mu, Jinghua Wang, Yang Zhao, et al. Codev-r1: Reasoning-enhanced verilog generation. arXiv preprint arXiv:2505.24183, 2025f.

Yekun Zhu, Guang Chen, and Chengjun Mao. Think in blocks: Adaptive reasoning from direct response to deep reasoning. arXiv preprint arXiv:2508.15507, 2025g.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In Aaai, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Conference on Robot Learning, pages 2165–2183. PMLR, 2023.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578, 2016.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms. arXiv preprint arXiv:2506.18896, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. arXiv preprint arXiv:2501.18362, 2025a.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. arXiv preprint arXiv:2504.16084, 2025b.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. Self-adapting language models. arXiv preprint arXiv:2506.10943, 2025.