The Landscape of Agentic Reinforcement Learning for LLMs: A Survey

面向大语言模型的主体化强化学习格局: 综述

Guibin Zhang [ϵ†] Hejia Geng [α†] Xiaohang Yu[η†] Zhenfei Yin[α−] Zaibin Zhang [να] Zelin Tan[ζβ] Heng Zhou [ζβ] Zhongzhi Li[ι] Xiangyuan Xu e[κβ] Yijiang Li[ξ] Yifan Zhou [μ] Yang Chen [β] Chen Zhang [ζ] Yutao Fan [β] Zihu Wang [χ] Songtao Huang [λβ] Piedrahita-Velez, Francisco [ε] Yue Liao [ϵ] Hongru Wang [κ] Mengyue Yang [θ] Heng Ji [δ] Michael Littman [ε] Jun Wang [γ] Shuicheng Yan [ϵ] Philip Torr [α] Lei Bai [β] [α] University of Oxford [β] Shanghai AI Laboratory [ϵ] National University of Singapore [γ] University College London [δ] University of Illinois Urbana-Champaign [ε] Brown University [ζ] University of Science and Technology of China [η] Imperial College London [θ] University of Bristol [ι] Chinese Academy of Sciences [κ] The Chinese University of Hong Kong [λ] Fudan University [μ] University of Georgia [ξ] University of California, San Diego [υ] Dalian University of Technology [χ] University of California, Santa Barbara [+] Equal contribution, [−] Corresponding Author

Guibin Zhang [ϵ†] Hejia Geng [α†] Xiaohang Yu[η†] Zhenfei Yin[α−] Zaibin Zhang [να] Zelin Tan[ζβ] Heng Zhou [ζβ] Zhongzhi Li[ι] Xiangyuan Xu e[κβ] Yijiang Li[ξ] Yifan Zhou [μ] Yang Chen [β] Chen Zhang [ζ] Yutao Fan [β] Zihu Wang [χ] Songtao Huang [λβ] Piedrahita-Velez, Francisco [ε] Yue Liao [ϵ] Hongru Wang [κ] Mengyue Yang [θ] Heng Ji [δ] Michael Littman [ε] Jun Wang [γ] Shuicheng Yan [ϵ] Philip Torr [α] Lei Bai [β] [α] University of Oxford [β] Shanghai AI Laboratory [ϵ] National University of Singapore [γ] University College London [δ] University of Illinois Urbana-Champaign [ε] Brown University [ζ] University of Science and Technology of China [η] Imperial College London [θ] University of Bristol [ι] Chinese Academy of Sciences [κ] The Chinese University of Hong Kong [λ] Fudan University [μ] University of Georgia [ξ] University of California, San Diego [υ] Dalian University of Technology [χ] University of California, Santa Barbara [+] Equal contribution, [−] Corresponding Author

Abstract: The emergence of agentic reinforcement learning (Agentic RL) marks a paradigm shift from conventional reinforcement learning applied to large language models (LLM RL), reframing LLMs from passive sequence generators into autonomous, decision-making agents embedded in complex, dynamic worlds. This survey formalizes this conceptual shift by contrasting the degenerate single-step Markov Decision Processes (MDPs) of LLM-RL with the partially observable, temporally extended partially observable Markov decision process (POMDP) that define Agentic RL. Building on this foundation, we propose a comprehensive twofold taxonomy: one organized around core agentic capabilities, including planning, tool use, memory, reasoning, self-improvement, and perception, and the other around their applications across diverse task domains. Central to our thesis is that reinforcement learning serves as the critical mechanism for transforming these capabilities from static, heuristic modules into adaptive, robust agentic behavior. To support and accelerate future research, we consolidate the landscape of open-source environments, benchmarks, and frameworks into a practical compendium. By synthesizing over five hundred recent works, this survey charts the contours of this rapidly evolving field and highlights the opportunities and challenges that will shape the development of scalable, general-purpose AI agents.

摘要: 主体化强化学习 (Agentic RL) 的出现标志着从传统应用于大语言模型的强化学习 (LLM RL) 到主体化代理范式的转变，将大语言模型从被动的序列生成器重构为嵌入在复杂动态世界中的自主决策体。本综述通过对比退化为单步马尔可夫决策过程 (MDP) 的 LLM-RL 与定义主体化强化学习的部分可观测、具有时间延展性的部分可观测马尔可夫决策过程 (POMDP)，形式化了这一概念转变。在此基础上，我们提出了一个双重综合分类法: 一方面围绕规划、工具使用、记忆、推理、自我改进与感知等核心主体能力组织，另一方面围绕这些能力在不同任务域中的应用组织。我们的核心论点是，强化学习是将这些能力从静态、启发式模块转化为自适应、稳健主体行为的关键机制。为支持并加速未来研究，我们将开源环境、基准与框架的景观整合为实用汇编。通过综述五百余篇近期工作，本综述勾勒了该快速发展的领域轮廓，并强调了将影响可扩展、通用 AI 代理发展的机遇与挑战。

Keywords: Agentic Reinforcement Learning, Large Language Models, LLM Agent

关键词: 主体化强化学习、大语言模型、LLM 代理

17 Date: November 1st, 2025

17 日期:2025 年 11 月 1 日

Code Repository: https://github.com/xhyumiracle/Awesome-AgenticLLM-RL-Papers

代码仓库:https://github.com/xhyumiracle/Awesome-AgenticLLM-RL-Papers

Corresponding: jeremyyin@robots.ox.ac.uk, bailei@pjlab.org.cn

通讯:jeremyyin@robots.ox.ac.uk, bailei@pjlab.org.cn

Main Contact: guibinz@u.nus.edu, genghejia0530@gmail.com, x.yu21@imperial.ac.uk

主要联系人:guibinz@u.nus.edu, genghejia0530@gmail.com, x.yu21@imperial.ac.uk

Contents

目录

1. Introduction

1. 引言

The rapid convergence of large language models (LLMs) and reinforcement learning (RL) has precipitated a fundamental transformation in how language models are conceived, trained, and deployed. Early LLM-RL paradigms largely treated these models as static conditional generators, optimized to produce single-turn outputs aligned with human preferences or benchmark scores. While successful for alignment and instruction following, such approaches overlook the broader spectrum of sequential decision-making that underpins realistic, interactive settings. These limitations have prompted a shift in perspective: rather than viewing LLMs as passive text emitters, recent developments increasingly frame them as Agents, i.e., autonomous decision-makers capable of perceiving, reasoning, planning, invoking tools, maintaining memory, and adapting strategies over extended horizons in partially observable, dynamic environments. We define this emerging paradigm as Agentic Reinforcement Learning (Agentic RL). To more clearly delineate the distinction between the concept of Agentic RL studied in this work and conventional RL approaches, we provide the following definition:

大型语言模型 (LLM) 与强化学习 (RL) 的快速融合引发了对语言模型构想、训练和部署方式的根本性变革。早期的 LLM-RL 范式大多将这些模型视为静态的条件生成器，旨在优化单轮输出以符合人类偏好或基准分数。尽管在对齐与遵循指令方面取得了成功，但此类方法忽视了支撑真实交互场景的序列决策更广泛范畴。这些局限促使观点发生转变: 不再将 LLM 视为被动的文本发出器，近期发展越来越多地将其构想为“主体”(Agents)，即能够感知、推理、规划、调用工具、保持记忆并在部分可观测、动态环境中随时间调整策略的自主决策者。我们将这一新兴范式定义为主体化强化学习 (Agentic Reinforcement Learning)。为更清晰地区分本文研究的主体化 RL 与传统 RL 方法的概念差异，我们给出如下定义:

Agentic Reinforcement Learning (Agentic RL) refers to a paradigm in which LLMs, rather than being treated as static conditional generators optimized for single-turn output alignment or benchmark performance, are conceptualized as learnable policies embedded within sequential decision-making loops, where RL endows them with autonomous agentic capabilities, such as planning, reasoning, tool use, memory maintenance, and self-reflection, enabling the emergence of long-horizon cognitive and interactive behaviors in partially observable, dynamic environments.

主体化强化学习 (Agentic RL) 指一种范式，在该范式中，LLM 不再被视为为单轮输出对齐或基准表现而优化的静态条件生成器，而是被构想为可学习的策略，嵌入序列决策循环中；强化学习赋予其自主主体能力，如规划、推理、使用工具、维护记忆与自我反思，从而在部分可观测、动态环境中促成长时延的认知与交互行为的出现。

In Section 2, we present a more formal, symbolically grounded distinction between agentic RL and conventional RL. Prior research relevant to agentic RL can be broadly grouped into two complementary threads: Synergy between RL and LLMs and LLM Agents, detailed as follows:

在第 2 节中，我们将更形式化、符号化地阐明主体化 RL 与传统 RL 的区别。与主体化 RL 相关的先前研究大致可分为两条互补脉络: 强化学习与 LLM 的协同，以及 LLM 主体，详述如下:

Synergy between RL and LLMs. The second line of research investigates how reinforcement learning algorithms are applied to improve or align LLMs. A primary branch, RL for training LLMs, leverages on-policy (e.g., proximal policy optimization (PPO) [1] and Group Relative Policy Optimization (GRPO) [2]) and off-policy (e.g., actor-critic, Q-learning [3]) methods to enhance capabilities such as instruction following, ethical alignment, and code generation [4, 5, 6]. A complementary direction, LLMs for RL, examines the deployment of LLMs as planners, reward designers, goal generators, or information processors to improve sample efficiency, generalization, and multi-task planning in control environments, with systematic taxonomies provided by Cao et al. [7]. RL has also been integrated throughout the LLM lifecycle: from data generation [8, 9] and pretraining [10] to post-training and inference [11], as surveyed by Guo et al. [12]. The most prominent branch here is post-training alignment, notably Reinforcement Learning from Human Feedback (RLHF) [13], along with extensions such as Reinforcement Learning from AI Feedback (RLAIF) and Direct Preference Optimization (DPO) [14, 15, 16, 4].

RL 与 LLM 的协同。第二条研究线路考察强化学习算法如何用于改进或对齐 LLM。主要分支“用于训练 LLM 的 RL”利用在线方法 (如近端策略优化 PPO [1] 和组相对策略优化 GRPO [2]) 和离线/离策略方法 (如 actor-critic、Q 学习 [3]) 来提升指令遵循、伦理对齐和代码生成等能力 [4, 5, 6]。互补方向“LLM 用于 RL”研究将 LLM 用作规划器、奖励设计者、目标生成器或信息处理器，以提升样本效率、泛化能力和控制环境中的多任务规划，Cao 等人提供了系统分类法 [7]。RL 也已贯穿 LLM 生命周期: 从数据生成 [8, 9] 与预训练 [10] 到后训练与推理 [11]，见 Guo 等人的综述 [12]。最突出的一支是后训练对齐，尤其是基于人类反馈的强化学习 (RLHF)[13]，以及扩展如基于 AI 反馈的强化学习 (RLAIF) 和直接偏好优化 (DPO)[14, 15, 16, 4]。

LLM Agents. LLM-based agents represent an emerging paradigm in which LLMs act as autonomous or semi-autonomous decision-making entities [17, 18], capable of reasoning, planning, and executing actions in pursuit of complex goals. Recent surveys have sought to map this landscape from complementary perspectives. Luo et al. [19] propose a methodology-centered taxonomy that connects architectural foundations, collaboration mechanisms, and evolutionary pathways, while Plaat et al. [20] emphasize the core capabilities of reasoning, acting, and interacting as defining features of agentic LLMs. Tool use, encompassing retrieval-augmented generation (RAG) and API utilization, is a central paradigm, extensively discussed in Li et al. [21] and further conceptualized by Wang et al. [22]. Planning and reasoning strategies form another pillar, with surveys such as Masterman et al. [23] and Kumar et al. [24] highlighting common design patterns like plan-execute-reflect loops, while Tao et al. [25] extend this to self-evolution, where agents iteratively refine knowledge and strategies without substantial human intervention. Other directions explore collaborative, cross-modal, and embodied settings, from multi-agent systems [26] to multimodal integration [27], and brain-inspired architectures with memory and perception [28].

LLM 代理。基于 LLM 的代理是一种新兴范式，LLM 作为自主或半自主决策实体 [17, 18]，能够推理、规划并执行动作以实现复杂目标。近期综述尝试从互补视角描绘该领域。Luo 等人提出以方法学为中心的分类法，连接架构基础、协作机制与演化路径 [19]，而 Plaat 等人强调推理、行动与交互作为代理化 LLM 的核心能力 [20]。工具使用 (包括检索增强生成 RAG 与 API 调用) 是核心范式，在 Li 等人中有广泛讨论 [21]，并由 Wang 等人进一步概念化 [22]。规划与推理策略构成另一支柱，Masterman 等人和 Kumar 等人的综述指出常见设计模式如“计划-执行-反思”循环 [23, 24]，Tao 等人将其扩展到自我进化，代理在较少人工干预下迭代完善知识与策略 [25]。其他方向包括协作型、多模态与具身设置，从多代理系统 [26] 到多模态整合 [27]，以及具有记忆与感知的类脑架构 [28]。

Research Gap and Our Contributions. The recent surge in research on LLM agents and RL-enhanced

LLMs reflects two complementary perspectives: one explores what large language models can do as the core of autonomous agents, while the other focuses on how reinforcement learning can optimize their behavior. However, despite the breadth of existing work, a unified treatment of agentic RL, which conceptualizes LLMs as policy-optimized agents embedded in sequential decision processes, remains lacking. Current studies often examine isolated capabilities, domains, or custom environments, with inconsistent terminology and evaluation protocols, making systematic comparison and cross-domain generalization difficult. To bridge this gap, we present a coherent synthesis that connects theoretical foundations with algorithmic approaches and practical systems. We formalize agentic RL through Markov decision process (MDP) and partially observable Markov decision process (POMDP) abstractions to distinguish it from classical LLM-RL paradigms, and introduce a capability-centered taxonomy that includes planning, tool use, memory, reasoning, reflection (self-improvement), and interaction as RL-optimizable components. Furthermore, we consolidate representative tasks, environments, frameworks, and benchmarks that support agentic LLM training and evaluation, and conclude by discussing open challenges and outlining promising future directions for scalable, general-purpose agentic intelligence. Overall, we aim to further clarify the research scope of this survey:

研究缺口与我们的贡献。近期关于 LLM 代理与 RL 增强 LLM 的研究激增，反映出两种互补视角: 一者探索大语言模型作为自治代理核心能做什么，另一者关注强化学习如何优化其行为。然而，尽管工作繁多，尚缺乏对“代理化 RL”的统一处理——将 LLM 概念化为嵌入序贯决策过程中的策略优化代理。现有研究常考察孤立能力、特定领域或定制环境，术语与评估协议不一致，导致系统比较与跨域泛化困难。为弥合此缺口，我们提出连贯综合，连接理论基础、算法方法与实用系统。我们通过马尔可夫决策过程 (MDP) 和部分可观测马尔可夫决策过程 (POMDP) 抽象对代理化 RL 进行形式化，以将其与经典 LLM-RL 范式区分，并引入以能力为中心的分类法，涵盖可被 RL 优化的规划、工具使用、记忆、推理、反思 (自我改进) 与交互等组件。此外，我们整合了支持代理化 LLM 训练与评估的典型任务、环境、框架与基准，并在结尾讨论开放挑战与可扩展、通用代理智能的有前景方向。总体目标是进一步澄清本综述的研究范围:

Primary focus:

主要关注:

V how RL empowers LLM-based agents (or, LLMs with agentic characteristics) in dynamic environments

V RL 如何在动态环境中赋能基于 LLM 的代理 (或具代理特征的 LLM)

Out of scope (though occasionally mentioned):

范围之外 (虽偶有提及):

X RL for human value alignment (e.g., RL for harmful query refusal);

X 用于人类价值对齐的 RL(例如拒绝有害查询的 RL);

X traditional RL algorithms that are not LLM-based (e.g., MARL [29]);

X 非基于 LLM 的传统 RL 算法 (例如 MARL [29])；

✗ RL for boosting pure LLM performance on static benchmarks.

✗ 为提升静态基准上纯 LLM 性能的 RL。

Structure of the Survey. This survey is organized to progressively build a unified understanding of Agentic RL from conceptual foundations to practical implementations. Section 2 formalizes the paradigm shift to Agentic RL through an MDP/POMDP lens. Section 3 examines agentic RL from the capability perspective, categorizing key modules such as planning, reasoning, tool using, memory, self-improvement, perception, and others. Section 4 explores applications across domains, including search, GUI navigation, code generation, mathematical reasoning, and multi-agent systems. Section 5 consolidates open-source environments and RL frameworks that underpin experimentation and benchmarking. Section 6 discusses open challenges and future directions towards scalable, adaptive, and reliable agentic intelligence, and Section 7 concludes the survey. The overall structure is also illustrated in Figure 1.

调查结构。本调查旨在逐步构建对 Agentic RL 的统一认识，从概念基础到实际实现。第 2 节通过 MDP/POMDP 视角形式化了向 Agentic RL 的范式转变。第 3 节从能力角度审视 agentic RL，归类关键模块如规划、推理、工具使用、记忆、自我改进、感知等。第 4 节探讨跨域应用，包括检索、GUI 导航、代码生成、数学推理和多智能体系统。第 5 节汇总了支撑实验与基准测试的开源环境和 RL 框架。第 6 节讨论面向可扩展、自适应与可靠 agentic 智能的开放挑战与未来方向，第 7 节为结论。总体结构亦在图 1 中示意。

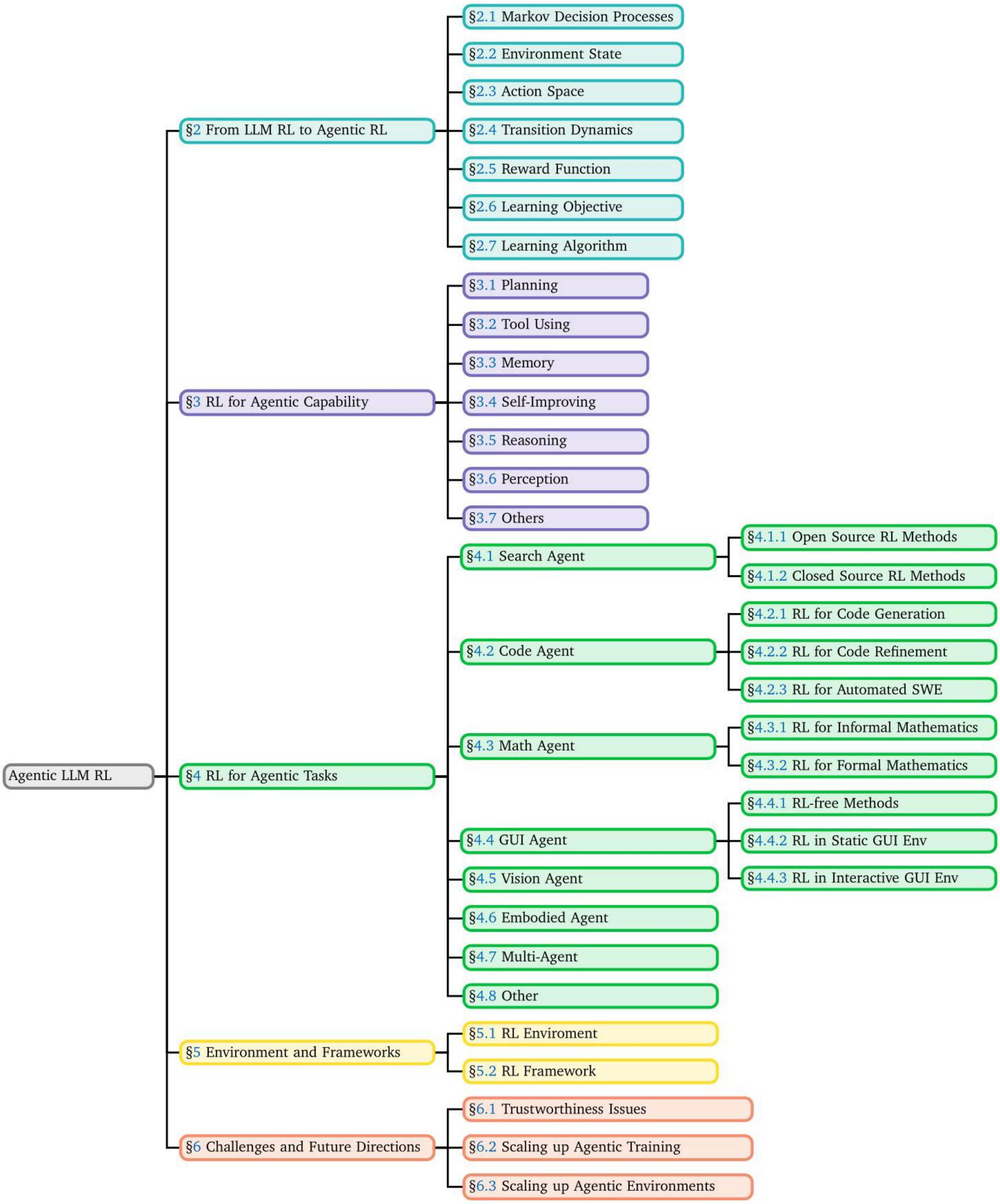
Agentic LLM RL
§2 From LLM RL to Agentic RL
§2.1 Markov Decision Processes
§2.2 Environment State
§2.3 Action Space
§2.4 Transition Dynamics
§2.5 Reward Function
§2.6 Learning Objective
§2.7 Learning Algorithm
§3 RL for Agentic Capability
§3.1 Planning
§3.2 Tool Using
§3.3 Memory
§3.4 Self-Improving
§3.5 Reasoning
§3.6 Perception
§3.7 Others
§4 RL for Agentic Tasks
§4.1 Search Agent
§4.1.1 Open Source RL Methods
§4.1.2 Closed Source RL Methods
§4.2 Code Agent
§4.2.1 RL for Code Generation
§4.2.2 RL for Code Refinement
§4.2.3 RL for Automated SWE
§4.3 Math Agent
§4.3.1 RL for Informal Mathematics
§4.3.2 RL for Formal Mathematics
§4.4 GUI Agent
§4.4.1 RL-free Methods
§4.4.2 RL in Static GUI Env
§4.4.3 RL in Interactive GUI Env
§4.5 Vision Agent
§4.6 Embodied Agent
§4.7 Multi-Agent
§4.8 Other
§5 Environment and Frameworks
§5.1 RL Enviroment
§5.2 RL Framework
§6 Challenges and Future Directions
§6.1 Trustworthiness Issues
§6.2 Scaling up Agentic Training
§6.3 Scaling up Agentic Environments

Figure 1: The primary organizational structure of the survey.

图 1: 本调查的主要组织结构。

2. Preliminary: From LLM RL to Agentic RL

2. 预备: 从 LLM RL 到 Agentic RL

LLMs are initially pre-trained using behavior cloning, which applies maximum likelihood estimation (MLE) to static datasets such as web-scraped text corpora. Subsequent post-training methods enhance capabilities and align outputs with human preferences-transforming them beyond generic web-data replicators. A common technique is supervised fine-tuning (SFT), where models are refined on human-generated (prompt, response) demonstrations. However, procuring sufficient high-quality SFT data remains challenging. Reinforcement fine-tuning (RFT) offers an alternative by optimizing models through reward functions, circumventing dependence on behavioral demonstrations.

大模型最初通过行为克隆预训练，使用最大似然估计 (MLE) 在如网络爬取文本语料的静态数据集上进行。随后的后训练方法增强能力并将输出与人类偏好对齐, 使其超越通用网络数据的简单复现。常见技术是监督微调 (SFT)，在人工生成的 (提示，回应) 示例上精炼模型。然而，获取足够高质量的 SFT 数据仍具挑战性。强化微调 (RFT) 通过优化模型的奖励函数提供替代路径，避免依赖行为示例。

In early RFT research, the core objective is to optimize LLMs through human feedback [13] or data preferences [30], aligning them with human preferences or directly with data preferences (as in DPO). This preference-based RFT (PBRFT) primarily involves learning reward model optimization for LLMs on a fixed preference dataset, or directly implementing it using data preferences. With the release of LLMs such as OpenAI o1 [31] and DeepSeek-R1 [32] that possess reasoning capabilities, their improved performance and cross-domain generalization have garnered widespread attention. With the release of models like OpenAI o3 [33], which possess both self-evolving reasoning capabilities and support for tool use, researchers are beginning to contemplate how to deeply integrate LLMs with downstream tasks through reinforcement learning methods. Subsequently, researchers have shifted their focus from PBRFT, aimed at optimizing fixed preference datasets, to agentic reinforcement learning tailored for specific tasks and dynamic environments.

在早期的 RFT 研究中，核心目标是通过人类反馈 [13] 或数据偏好 [30] 来优化大模型，使其与人类偏好或直接与数据偏好 (如 DPO 中) 对齐。这类基于偏好的 RFT(PBRFT) 主要包括在固定偏好数据集上学习奖励模型或直接利用数据偏好进行实施。随着具备推理能力的模型 (如 OpenAI o1 [31] 和 DeepSeek-R1 [32]) 的发布，其改进的表现和跨域泛化引起广泛关注。再到像 OpenAI o3 [33] 这类既具自我进化的推理能力又支持工具使用的模型出现, 研究者开始思考如何通过强化学习方法将大模型与下游任务深度结合。随后，研究重心从旨在优化固定偏好数据集的 PBRFT 转向为特定任务和动态环境量身定制的 agentic 强化学习。

In this section, we provide a formalization of the paradigm shift from PBRFT to the emerging framework of agentic reinforcement learning (Agentic RL). While both approaches leverage RL techniques to improve LLMs' performance, they fundamentally differ in their underlying assumptions, task structure, and decision-making granularity. Figure 2 illustrates the paradigm shift from LLM RL to agentic RL.

在本节, 我们对从 PBRFT 到新兴 agentic 强化学习框架的范式转变进行形式化。虽然两者都利用 RL 技术提升大模型性能，但在基本假设、任务结构与决策粒度上存在根本差异。图 2 展示了从 LLM RL 到 agentic RL 的范式转变。

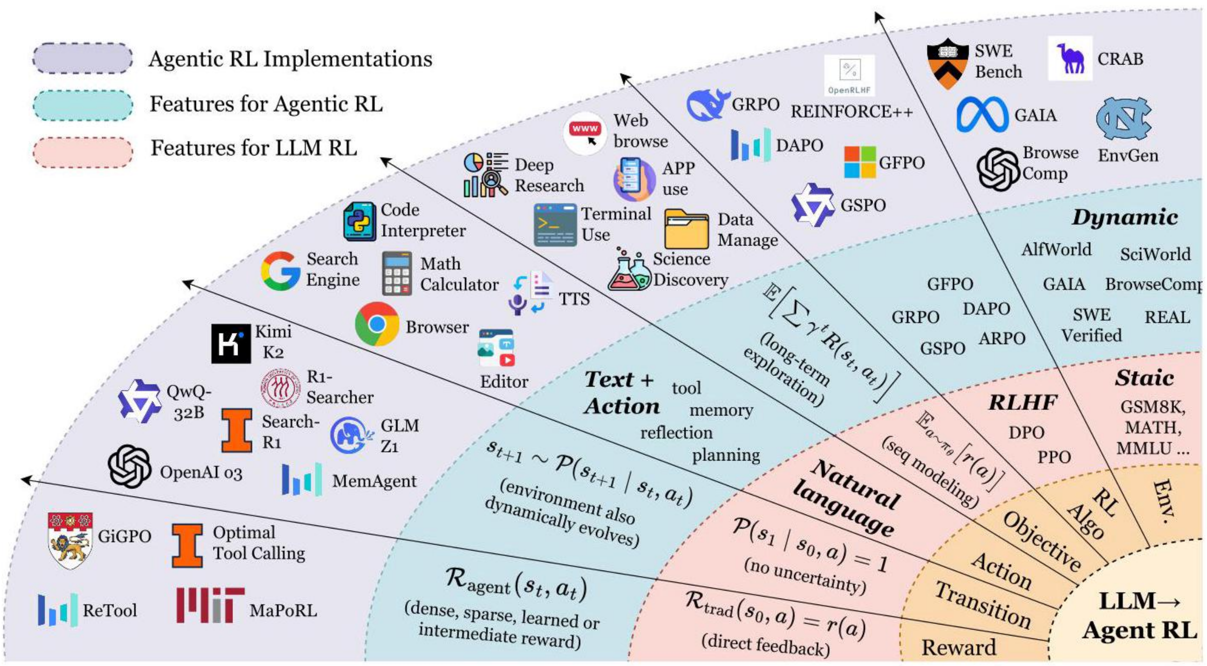

Figure 2: Paradigm shift from LLM-RL to agentic RL. We draw inspiration from the layout of Figure 1 in [24].

图 2: 从 LLM-RL 到 agentic RL 的范式转变。我们借鉴了文献 [24] 中图 1 的布局。

2.1. Markov Decision Processes

2.1. 马尔可夫决策过程

The Markov decision process (MDP) for the RL fine-tuning process can be formalized as a seven-element tuple $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, T, \gamma \rangle$, where $\mathcal{S}$ represents the state space and $\mathcal{O}$ is the observation space of the agent. $\mathcal{A}$ denotes the action space. $\mathcal{R}$ is defined as the reward function, $\mathcal{P}$ encapsulates the state transition probabilities, T signifies the task horizon, and γ is the discount factor. By casting both preference-based RFT and agentic RL as MDP or POMDP, we clarify the theoretical implications of treating LLMs either as static sequence generators or as interactive, decision-capable agents embedded within dynamic environments.

RL 微调过程的马尔可夫决策过程 (MDP) 可形式化为七元组 $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, T, \gamma \rangle$， 其中 $\mathcal{S}$ 表示状态空间，$\mathcal{O}$ 为代理的观测空间。$\mathcal{A}$ 表示动作空间。$\mathcal{R}$ 定义为奖励函数，$\mathcal{P}$ 封装状态转移概率，T 表示任务时域，γ 为折扣因子。通过将基于偏好的 RFT 与 agentic RL 都视作 MDP 或 POMDP，我们阐明了将大模型视为静态序列生成器或作为嵌入于动态环境中、具备交互与决策能力的代理这两种处理方式的理论含义。

PBRFT. The RL training process of PBRFT is formalized as a degenerate MDP defined by the tuple:

PBRFT。PBRFT 的 RL 训练过程形式化为由以下元组定义的退化 MDP:

$$\langle \mathcal{S}_{\text{trad}}, \mathcal{A}_{\text{trad}}, \mathcal{P}_{\text{trad}}, \mathcal{R}_{\text{trad}}, \gamma = 1, T = 1 \rangle \,. \tag{1}$$

Agentic RL. The RL training process of agentic RL is modeled as a POMDP:

Agentic RL。agentic RL 的 RL 训练过程被建模为一个 POMDP:

$$\langle \mathcal{S}_{\text{agent}}, \mathcal{A}_{\text{agent}}, \mathcal{P}_{\text{agent}}, \mathcal{R}_{\text{agent}}, \gamma, \mathcal{O} \rangle, \tag{2}$$

where the agent receives observations $o_t = O(s_t)$ based on the state $s_t \in \mathcal{S}_{\text{agent}}$. The primary distinctions between PBRFT and agentic RL are delineated in Table 1. In summary, PBRFT optimizes sequences of output sentences within a fixed dataset under full observations, whereas agentic RL optimizes semantic-level behaviors in variable environments characterized by partial observations.

其中代理基于状态 $s_t \in \mathcal{S}_{\text{agent}}$ 接收观测 $o_t = O(s_t)$。PBRFT 与 agentic RL 的主要区别在表 1 中阐明。总之，PBRFT 在完全观测下在固定数据集内优化输出句子序列，而 agentic RL 在部分观测表征的可变环境中优化语义层面的行为。

Table 1: Formal comparison between traditional PBRFT and Agentic RL.

表 1: 传统 PBRFT 与 Agentic RL 的形式化比较。

Concept	Traditional PBRFT	Agentic RL
$\mathcal{S}$: State space	$\{s_0\}$ (single prompt); episode ends immediately.	$s_t \in \mathcal{S}_{\text{agent}}$; $o_t = O(s_t)$; horizon $T > 1$.
$\mathcal{A}$: Action space	Pure text sequence.	$\mathcal{A}_{\text{text}} \cup \mathcal{A}_{\text{action}}$.
$\mathcal{P}$: Transition	Deterministic to the terminal state.	Dynamic transition function $P(s_{t+1} \mid s_t, a_t)$.
$\mathcal{R}$: Reward	Single scalar $r(a)$.	Step-wise $R(s_t, a_t)$; combines sparse task and dense sub-rewards.
$J(\theta)$: Objective	$\mathbb{E}_{a\sim\pi_\theta}[r(a)]$.	$\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_t \gamma^t R(s_t, a_t)\right]$.

概念	传统 PBRFT	代理式强化学习
$\mathcal{S}$: 状态空间	$\{s_0\}$ (单次提示)；回合立即结束。	$s_t \in \mathcal{S}_{\text{agent}}$; $o_t = O(s_t)$ ；时长 $T > 1$ 。
$\mathcal{A}$: 动作空间	纯文本序列。	$\mathcal{A}_{\text{text}} \cup \mathcal{A}_{\text{action}}$.
$\mathcal{P}$: 转移	确定性转移到终止状态。	动态转移函数 $P(s_{t+1} \mid s_t, a_t)$ 。
$\mathcal{R}$: 奖励	单个标量 $r(a)$ 。	逐步 $R(s_t, a_t)$ ；结合稀疏任务与密集子奖励。
$J(\theta)$: 目标	$\mathbb{E}_{a\sim\pi_\theta}[r(a)]$.	$\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_t \gamma^t R(s_t, a_t)\right]$.

2.2. Environment State

2.2. 环境状态

PBRFT. In the training process, each episode starts from a single prompt state s_0 ; the episode terminates immediately after the model emits one response. Formally, the underlying MDP degenerates to a single-step decision problem with horizon $T = 1$. The state space reduces to a single static prompt input:

PBRFT。在训练过程中，每个 episode 从单一提示状态 s_0 开始；模型发出一次响应后 episode 立即终止。形式上，底层 MDP 退化为一个步长为 $T = 1$ 的单步决策问题。状态空间简化为单一静态提示输入:

$$\mathcal{S}_{\text{trad}} = \{ \text{prompt} \}. \tag{3}$$

Agentic RL. The LLM agent acts over multiple time-steps in a POMDP. Let $s_t \in \mathcal{S}_{\text{agent}}$ denote the full world-state and the LLM agent gets observation O_t based on current state $o_t = \mathcal{O}(s_t)$. The LLM agent chooses an action a_t based on the current observation o_t, and the state evolves over time:

Agentic RL。LLM 代理在 POMDP 中跨多个时间步执行动作。设 $s_t \in \mathcal{S}_{\text{agent}}$ 表示完整的世界状态，LLM 代理基于当前状态 $o_t = \mathcal{O}(s_t)$ 获取观察 O_t。LLM 代理基于当前观察 o_t 选择动作 a_t，状态随时间演化:

$$s_{t+1} \sim P(s_{t+1} \mid s_t, a_t), \tag{4}$$

as the agent accumulates intermediate signals such as retrieved tool results, user messages, or environment feedback. The interaction is thus inherently dynamic and temporally extended.

随着代理累积中间信号，例如检索到的工具结果、用户消息或环境反馈。交互因此本质上是动态且时间延展的。

2.3. Action Space

2.3. 动作空间

In the agentic RL setting, the LLM's action space comprises two distinct subspaces:

在 agentic RL 设置中，LLM 的动作空间由两类不同的子空间组成:

$$\mathcal{A}_{\text{agent}} = \mathcal{A}_{\text{text}} \cup \mathcal{A}_{\text{action}} \tag{5}$$

Here, $\mathcal{A}_{\text{text}}$ denotes the space of free-form natural language tokens emitted via autoregressive decoding, while $\mathcal{A}_{\text{action}}$ denotes the space of abstract, non-linguistic actions (e.g., delimited in the output stream by special tokens <action_start> and <action_end>). These actions may invoke external tools (e.g., call (”search”, ”Einstein”)) or interact with an environment (e.g., move (”north”)), depending on task requirements.

其中，$\mathcal{A}_{\text{text}}$ 表示通过自回归解码发出的自由形式自然语言令牌空间，而 $\mathcal{A}_{\text{action}}$ 表示抽象的、非语言化动作空间 (例如由特殊标记 <action_start> 与 <action_end> 在输出流中界定)。这些动作可能调用外部工具 (例如调用 (”search”, ”Einstein”)) 或与环境交互 (例如移动 (”north”))，取决于任务需求。

Notably, $\mathcal{A}_{\text{action}}$ is recursively constructed, such that an element $a \in \mathcal{A}_{\text{action}}$ may itself represent a sequence $(a_1, \ldots, a_k)$ of primitive actions, thus unifying primitive and composite actions within the same space.

值得注意的是，$\mathcal{A}_{\text{action}}$ 是递归构造的，使得一个元素 $a \in \mathcal{A}_{\text{action}}$ 本身可能表示原子动作序列 $(a_1, \ldots, a_k)$，从而在同一空间内统一原子与复合动作。

Formally, the two subspaces differ in semantics and functional role: $\mathcal{A}_{\text{text}}$ defines the space of outputs intended for human or machine interpretation without directly altering the external state, whereas $\mathcal{A}_{\text{action}}$ defines the space of environment-interactive behaviors that either (i) acquire new information through tool invocations, or (ii) modify the state of a physical or simulated environment. This distinction enables a unified policy jointly model language generation and environment interaction within the same RL formulation.

形式上，这两类子空间在语义与功能角色上有所不同: $\mathcal{A}_{\text{text}}$ 定义了面向人类或机器解读的输出空间，不直接改变外部状态；而 $\mathcal{A}_{\text{action}}$ 定义了与环境交互的行为空间，这些行为要么 (i) 通过工具调用获取新信息，要么 (ii) 修改物理或模拟环境的状态。此区分使得在同一 RL 表述下统一建模语言生成与环境交互成为可能。

2.4. Transition Dynamics

2.4. 转移动力学

PBRFT. In conventional PBRFT, the transition dynamics are deterministic: the next state is determined once an action is made, as follows:

PBRFT。在传统 PBRFT 中，转移动力学是确定性的: 一旦采取动作，下一个状态即可确定，如下所示:

$$\mathcal{P}(s_1 \mid s_0, a) = 1 \text{ , where there is no uncertainty.} \tag{6}$$

Agentic RL. In agentic RL, the environment evolves under uncertainty according to

Agentic RL。在 agentic RL 中，环境在不确定性下演化，遵循

$$s_{t+1} \sim \mathcal{P}(s_{t+1} \mid s_t, a_t), \; a_t \in \mathcal{A}_{\text{text}} \cup \mathcal{A}_{\text{action}} . \tag{7}$$

Text actions ($\mathcal{A}_{\text{text}}$) generate natural language outputs without altering the environment state. Structured actions ($\mathcal{A}_{\text{action}}$), delimited by <action_start> and <action_end>, can either query external tools or directly modify the environment. This sequential formulation contrasts with the one-shot mapping of PBRFT, enabling policies that iteratively combine communication, information acquisition, and environment manipulation.

文本动作 ($\mathcal{A}_{\text{text}}$) 生成自然语言输出而不改变环境状态。结构化动作 ($\mathcal{A}_{\text{action}}$)，由 <action_start> 和 <action_end> 界定，既可以查询外部工具，也可以直接修改环境。该序列化表述与 PBRFT 的一次性映射形成对比，使策略能够迭代地结合交流、信息获取与环境操作。

2.5. Reward Function

2.5. 奖励函数

PBRFT. PBRFT commonly features a reward function with verifiable response correctness, which may be implemented using either a rule-based verifier [32] or a neural network-parameterized reward model [34]. Regardless of the implementation approach, its core follows the equation:

PBRFT。PBRFT 通常具有可验证响应正确性的奖励函数，可通过基于规则的验证器 [32] 或神经网络参数化的奖励模型 [34] 实现。无论采用何种实现，其核心遵循如下等式:

$$\mathcal{R}_{\text{trad}}\left(s_0, a\right) = r\left(a\right), \tag{8}$$

where $r : \mathcal{A} \to \mathbb{R}$ is a scalar score supplied by a human- or AI-preference model; with no intermediate feedback.

其中 $r : \mathcal{A} \to \mathbb{R}$ 是由人类或 AI 偏好模型提供的标量分数；没有中间反馈。

Agentic RL. The reward function of the LLM agent is based on the downstream task.

Agentic RL。LLM 代理的奖励函数基于下游任务。

$$\mathcal{R}_{\text{agent}}\left(s_t, a_t\right) = \begin{cases} r_{\text{task}} & \text{on task completion,} \\ r_{\text{sub}}\left(s_t, a_t\right) & \text{for step-level progress,} \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

allowing dense, sparse, or learned rewards (e.g., unit-test pass, symbolic verifier success).

允许密集、稀疏或学习到的奖励 (例如，单元测试通过、符号验证成功)。

2.6. Learning Objective

2.6. 学习目标

PBRFT. The optimization objective of PBRFT is to maximize the response reward based on the policy π_θ :

PBRFT。PBRFT 的优化目标是基于策略 π_θ 最大化响应奖励:

$$J_{\text{trad}}\left(\theta\right) = \mathbb{E}_{a \sim \pi_\theta}\left[r\left(a\right)\right]. \tag{10}$$

No discount factor is required; optimization resembles maximum-expected-reward sequence modeling.

不需要折扣因子；优化类似于最大期望奖励的序列建模。

Agentic RL. The optimization objective of Agentic RL is to maximize the discounted reward:

Agentic RL。Agentic RL 的优化目标是最大化折扣奖励:

$$J_{\text{agent}}(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t R_{\text{agent}}(s_t, a_t)\right],\ 0 < \gamma < 1, \tag{11}$$

optimized via policy-gradient or value-based methods with exploration and long-term credit assignment.

通过策略梯度或基于价值的方法进行优化，并结合探索与长期归因。

PBRFT focuses on single-turn text quality alignment without explicit planning, tool use, or environment feedback, while agentic RL involves multi-turn planning, adaptive tool invocation, stateful memory, and long-horizon credit assignment, enabling the LLM to function as an autonomous decision-making agent.

PBRFT 专注于单回合文本质量的对齐，不包含显式规划、工具使用或环境反馈，而 agentic RL 涉及多回合规划、自适应工具调用、有状态记忆和长期归因，使 LLM 能作为自主决策代理运作。

2.7.RL Algorithms

2.7.RL 算法

In contemporary research, RL algorithms constitute a pivotal component in both PBRFT and agentic RL frameworks. Different RL algorithms demonstrate distinct sample efficiency and performance characteristics, each offering a unique approach to the central challenge of aligning model outputs with complex, often subjective, human goals. The canonical methods, such as REINFORCE, PPO [1], GRPO [32], and DPO [30], form a spectrum from general policy gradients to specialized preference learning. We next introduce each of these three classic algorithms and provide a comparison of popular variants from each family in Table 2.

在当代研究中，RL 算法构成 PBRFT 和 agentic RL 框架中的关键组成部分。不同的 RL 算法在样本效率和性能上各有特点, 各自为将模型输出与复杂且常主观的人类目标对齐这一中心挑战提供不同方法。典型方法如 REINFORCE、PPO [1]、GRPO [32] 和 DPO [30]，形成从通用策略梯度到专门偏好学习的谱系。接下来我们介绍这三类经典算法，并在表 2 中比较各家常见变体。

REINFORCE: The Foundational Policy Gradient As one of the earliest policy gradient algorithms, REINFORCE provides the foundational theory for training stochastic policies. It operates by increasing the probability of actions that lead to high cumulative reward and decreasing the probability of those that lead to low reward. Its objective function is given by:

REINFORCE: 基础的策略梯度作为最早的策略梯度算法之一，REINFORCE 为训练随机策略提供了基础理论。其通过增加导致高累积奖励的动作概率、降低导致低奖励的动作概率来运作。其目标函数为:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{R}\left(s_0, a^{(i)}\right) - b\left(s_0\right)\right)\nabla_\theta \log \pi_\theta\left(a^{(i)} \mid s_0\right)\right], \tag{12}$$

where $a^{(i)} \sim \pi_\theta\left(a \mid s_0\right)$ is the i -th sampled response, $\mathcal{R}\left(s_0, a\right)$ denotes the final rewards received on task completion, and $b(s)$ is a baseline function to reduce the variance of the policy gradient estimate. In general, $b(s)$ can be any function, including random variables.

其中 $a^{(i)} \sim \pi_\theta\left(a \mid s_0\right)$ 是第 i 个采样响应，$\mathcal{R}\left(s_0, a\right)$ 表示任务完成时获得的最终奖励，$b(s)$ 是用于降低策略梯度估计方差的基线函数。通常，$b(s)$ 可以是任何函数，包括随机变量。

Proximal Policy Optimization (PPO) PPO [1] became the dominant RL algorithm for LLM alignment due to its stability and reliability. It improves upon vanilla policy gradients by limiting the update step to prevent destructively large policy changes. Its primary clipped objective function is:

近端策略优化 (PPO) PPO [1] 因其稳定性和可靠性成为 LLM 对齐的主流 RL 算法。它通过限制更新步长以防止破坏性的大幅策略变动来改进原始策略梯度。其主要的裁剪目标函数为:

$$L_{PPO}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\min\left(\frac{\pi_\theta\left(a_t^{(i)} \mid s_t\right)}{\pi_{\theta_{old}}\left(a_t^{(i)} \mid s_t\right)}A\left(s_t, a_t^{(i)}\right), \text{clip}\left(\frac{\pi_\theta\left(a_t^{(i)} \mid s_t\right)}{\pi_{\theta_{old}}\left(a_t^{(i)} \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right)A\left(s_t, a_t^{(i)}\right)\right), \tag{13}$$

where $a_t^{(i)} \sim \pi_{\theta_{\text{old}}}\left(a \mid s_t\right)$ is the i -th sampled response from the old policy $\pi_{\theta_{\text{old}}}$, whose update is delayed. A_t is the estimated advantage given by

其中 $a_t^{(i)} \sim \pi_{\theta_{\text{old}}}\left(a \mid s_t\right)$ 是来自被延迟更新的旧策略 $\pi_{\theta_{\text{old}}}$ 的第 i 个采样响应。A_t 是估计的优势，给出为

$$A\left(s_t, a_t\right) = \mathcal{R}\left(s_t, a_t\right) - V\left(s_t\right), \tag{14}$$

where $V_\theta(s)$ is the learned value function, i.e., the expectation $\mathbb{E}_{a\sim\pi_\theta(a|s)}\left[\mathcal{R}(s, a)\right]$, which is derived from a critic network that is of the same size as the policy network. The clip term prevents the probability ratio from moving too far from 1, ensuring stable updates. A key drawback is its reliance on a separate critic network for advantage estimation, which substantially increases the parameter count during training.

其中 $V_\theta(s)$ 是学习到的价值函数，即期望 $\mathbb{E}_{a\sim\pi_\theta(a|s)}\left[\mathcal{R}(s, a)\right]$ ，该价值函数来自与策略网络同等规模的评论家网络。裁剪项防止概率比率远离 1，确保更新稳定。其主要缺点是依赖单独的评论家网络来估计优势，这在训练期间显著增加参数数量。

Direct Preference Optimization (DPO) DPO represents a groundbreaking shift by entirely bypassing the need for a separate reward model. It reframes the problem of maximizing a reward under a KL-constraint as a likelihood-based objective on human preference data. Given a dataset of preferences $D = \{(y_w, y_l)\}$, where y_w is the preferred response and y_l is the dispreferred one, the DPO loss is:

直接偏好优化 (DPO) DPO 通过完全绕过单独奖励模型实现了突破性转变。它将受 KL 约束下最大化奖励的问题重构为对人类偏好数据的似然性目标。给定偏好数据集 $D = \{(y_w, y_l)\}$ ，其中 y_w 为被偏好的响应，y_l 为不被偏好的响应，DPO 损失为:

$$L_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right], \tag{15}$$

where π_{ref} is a reference policy (usually the initial SFT model), and β is a hyperparameter. While DPO eliminates the critic, its performance is intrinsically tied to the quality and coverage of its static preference dataset. Variants have emerged to address its limitations, including IPO (Identity Preference Optimization) [35] which adds a regularization term to prevent overfitting, and KTO (Kahneman-Tversky Optimization) [36], which learns from per-response binary signals (desirable/undesirable) rather than strict pairwise comparisons. See Table. 2 for more variants.

其中 π_{ref} 是参考策略 (通常为初始 SFT 模型)，β 是超参数。尽管 DPO 去除了评论器，其性能本质上依赖于静态偏好数据集的质量与覆盖范围。为解决其局限性出现了若干变体，包括 IPO(Identity Preference Optimization)[35], 通过加入正则项防止过拟合；以及 KTO(Kahneman-Tversky Optimization)[36]，它学习来自每个响应的二元信号 (可取/不可取) 而非严格的成对比较。更多变体见表 2。

Group Relative Policy Optimization (GRPO) The remarkable success achieved by DeepSeek has catalyzed significant research interest in GRPO. Proposed to address the inefficiency of PPO's large critic, GRPO introduces a novel, lightweight evaluation paradigm. It operates on groups of responses, using their relative rewards within a group to compute advantages, thus eliminating the need for an absolute value critic. The core GRPO objective can be conceptualized as:

群体相对策略优化 (GRPO) DeepSeek 取得的显著成功激发了对 GRPO 的大量研究兴趣。为解决 PPO 的大型评论器效率低下问题，GRPO 引入了一种新颖且轻量的评估范式。它在响应组上运行，使用组内相对奖励来计算优势，从而无需绝对值评论器。GRPO 的核心目标可被概念化为:

$$L_{GRPO} = \frac{1}{G}\sum_{g=1}^{G}\min\left(\frac{\pi_\theta\left(a_t^{(g)} \mid s_t^{(g)}\right)}{\pi_{\theta_{old}}\left(a_t^{(g)} \mid s_t^{(g)}\right)}\widehat{A}\left(s_t^{(g)}, a_t^{(g)}\right), \text{clip}\left(\frac{\pi_\theta\left(a_t^{(g)} \mid s_t^{(g)}\right)}{\pi_{\theta_{old}}\left(a_t^{(g)} \mid s_t^{(g)}\right)}, 1-\epsilon, 1+\epsilon\right)\widehat{A}\left(s_t^{(g)}, a_t^{(g)}\right)\right),$$

(16)

where a group of outputs $\left\{\left(s_0^{(g)}, a_0^{(g)}, \ldots, s_{T-1}^{(g)}, a_{T-1}^{(g)}\right)\right\}_{g=1}^{G}$ is sampled from the old policy $\pi_{\theta_{\text{old}}}$. The advantage function is estimated by

其中从旧策略 $\pi_{\theta_{\text{old}}}$ 采样出一组输出 $\left\{\left(s_0^{(g)}, a_0^{(g)}, \ldots, s_{T-1}^{(g)}, a_{T-1}^{(g)}\right)\right\}_{g=1}^{G}$。优势函数估计为

$$\widehat{A}(s_t, a_t) = \frac{\mathcal{R}(s_t, a_t) - \text{mean}\left(\mathcal{R}\left(s_t^{(1)}, a_t^{(1)}\right), \ldots, \mathcal{R}\left(s_t^{(G)}, a_t^{(G)}\right)\right)}{\text{std}\left(\mathcal{R}\left(s_t^{(1)}, a_t^{(1)}\right), \ldots, \mathcal{R}\left(s_t^{(G)}, a_t^{(G)}\right)\right)}. \tag{17}$$

Table 2: Comparison of the popular variants of the PPO, DPO, and GRPO families. Clip corresponds to preventing the policy ratio from moving too far from 1 for ensuring stable updates. KL penalty corresponds to penalizing the KL divergence between the learned policy and the reference policy for ensuring alignment.

表 2: 对 PPO、DPO 和 GRPO 系列常用变体的比较。Clip 指防止策略比率偏离 1 过远以确保更新稳定。KL 惩罚指对学习策略与参考策略之间的 KL 散度进行惩罚以确保对齐。

Method	Year	Objective Type	Clip	KL Penalty	Key Mechanism	Signal
PPO family						
PPO [1]	2017	Policy gradient	Yes	No	Policy ratio clipping	Reward
VAPO [37]	2025	Policy gradient	Yes	Adaptive	Adaptive KL penalty + variance control	Reward + variance signal
LitePPO [38]	2025	Policy gradient	Yes	Yes	Stable advantage updates	Reward
PF-PPO [39]	2024	Policy gradient	Yes	Yes	Policy filtration	Noisy reward
VinePPO [40]	2024	Policy gradient	Yes	Yes	Unbiased value estimates	Reward
PSGPO [41]	2024	Policy gradient	Yes	Yes	Process supervision	Process Reward
DPO family						
DPO [30]	2024	Preference optimization	No	Yes	Implicit reward related to the policy	Human preference
β -DPO [42]	2024	Preference optimization	No	Adaptive	Dynamic KL coefficient	Human preference
SimPO [43]	2024	Preference optimization	No	Scaled	Use the average log probability of a sequence as the implicit reward	Human preference
IPO [35]	2024	Implicit preference	No	No	Leverage generative LLMs as preference classifiers for reducing the dependence on external human feedback or reward models	Preference rank
KTO [36]	2024	Knowledge transfer optimization	No	Yes	Teacher stabilization	Teacher-student logit
ORPO [44]	2024	Online regularized preference optimization	No	Yes	Online stabilization	Online feedback reward
Step-DPO [45]	2024	Preference optimization	No	Yes	Step-wise supervision	Step-wise preference
LCPO [46]	2025	Preference optimization	No	Yes	Length preference with limited data and training	Reward
GRPO family						
GRPO [32]	2025	Policy gradient under group-based reward	Yes	Yes	Group-based relative reward to eliminate value estimates	Group-based reward
DAPO [47]	2025	Surrogate of GRPO's	Yes	Yes	Decoupled clip and dynamic sampling	Dynamic group-based reward
GSPO [48]	2025	Surrogate of GRPO's	Yes	Yes	Define the importance ratio based on sequence likelihood and performs sequence-level clipping, rewarding, and optimization	Smooth group-based reward
GMPO [49]	2025	Surrogate of GRPO's	Yes	Yes	Geometric mean of token-level rewards	Margin-based reward
ProRL [50]	2025	Same as GRPO's	Yes	Yes	Reference policy reset	Group-based reward
Posterior-GRPO [51]	2025	Same as GRPO's	Yes	Yes	Reward only successful processes	Process-based reward
Dr.GRPO [52]	2025	Unbiased GRPO's objective	Yes	Yes	Eliminate the bias in optimization of GRPC	Group-based reward
Step-GRPO [53]	2025	Same as GRPO's	Yes	Yes	Rule-based reasoning rewards	Step-wise reward
SRPO [54]	2025	Same as GRPO's	Yes	Yes	Two-staged history-resampling	Reward
GRESO [55]	2025	Same as GRPO's	Yes	Yes	Pre-rollout filtering	Reward
StarPO [56]	2025	Same as GRPO's	Yes	Yes	Reasoning-guided actions for multi-turn interactions	Group-based reward
GHPO [57]	2025	Policy gradient	Yes	Yes	Adaptive prompt refinement	Reward
Skywork R1V2 [58]	2025	GRPO's with hybrid reward signal	Yes	Yes	Selective sample buffer	Multimodal reward
ASPO [59]	2025	GRPO's with shaped advantage function	Yes	Yes	Apply a clipped bias directly to advantage function	Group-based reward
TreePo [60]	2025	Same as GRPO's	Yes	Yes	Self-guided policy rollout for reducing the compute burden	Group-based reward
EDGE-GRPO [61]	2025	Same as GRPO's	Yes	Yes	Entropy-driven advantage and duided error correction to mitigate advantage collapse	Group-based reward
DARS [62]	2025	Same as GRPO's	Yes	No	Reallocate compute from medium-difficulty to the hardest problems via multi-stage rollout sampling	Group-based reward
CHORD [63]	2025	Weighted sum of GRPO's and Supervised Fine-Tuning losses	Yes	Yes	Reframe Supervised Fine-Tuning as a dynamically weighted auxiliary objective within the on-policy RL process	Group-based reward
PAPO [64]	2025	Surrogate of GRPO's	Yes	Yes	Encourage learning to perceive while learning to reason through the Implicit Perception Loss	Group-based reward
Pass@k Training [65]	2025	Same as GRPO's	Yes	Yes	Pass@k metric as the reward to continually train a model	Group-based reward

方法	年份	目标类型	裁剪	KL 惩罚	关键机制	信号
PPO 家族						
PPO [1]	2017	策略梯度	是	否	策略比率裁剪	奖励
VAPO [37]	2025	策略梯度	是	自适应	自适应 KL 惩罚 + 方差控制	奖励 + 方差信号
LitePPO [38]	2025	策略梯度	是	是	稳定的优势更新	奖励
PF-PPO [39]	2024	策略梯度	是	是	策略过滤	带噪声的奖励
VinePPO [40]	2024	策略梯度	是	是	无偏的价值估计	奖励
PSGPO [41]	2024	策略梯度	是	是	过程监督	过程奖励
DPO 家族						
DPO [30]	2024	偏好优化	否	是	与策略相关的隐式奖励	人类偏好
β -DPO [42]	2024	偏好优化	否	自适应	动态 KL 系数	人类偏好
SimPO [43]	2024	偏好优化	否	缩放	使用序列的平均对数概率作为隐式奖励	人类偏好
IPO [35]	2024	隐式偏好	否	否	利用生成式大模型作为偏好分类器以减少对外部人工反馈或奖励模型的依赖	偏好排序
KTO [36]	2024	知识迁移优化	否	是	教师稳定化	师生 logit
ORPO [44]	2024	在线正则化 偏好优化	否	是	在线稳定化	在线反馈奖励
Step-DPO [45]	2024	偏好优化	否	是	逐步监督	逐步偏好
LCPO [46]	2025	偏好优化	否	是	在有限数据和训练下的长度偏好	奖励
GRPO 家族						
GRPO [32]	2025	在基于分组的奖励下的策略梯度	是	是	基于分组的相对奖励 以消除价值估计	基于分组的奖励
DAPO [47]	2025	GRPO 的替代目标	是	是	解耦裁剪和动态采样	动态分组奖励
GSPO [48]	2025	GRPO 的替代目标	是	是	基于序列似然定义重要性比并执行序列级截断、奖励与优化	平滑的基于组的奖励
GMPO [49]	2025	GRPO 的替代目标	是	是	标记级奖励的几何平均	基于边际的奖励
ProRL [50]	2025	与 GRPO 相同	是	是	参考策略重置	基于分组的奖励
Posterior-GRPO [51]	2025	与 GRPO 相同	是	是	仅对成功过程给予奖励	基于过程的奖励
Dr.GRPO [52]	2025	无偏的 GRPO 目标	是	是	消除 GRPC 优化中的偏差	基于分组的奖励
Step-GRPO [53]	2025	与 GRPO 相同	是	是	基于规则的推理奖励	逐步奖励
SRPO [54]	2025	与 GRPO 相同	是	是	两阶段历史重采样	奖励
GRESO [55]	2025	与 GRPO 相同	是	是	预回放过滤	奖励
StarPO [56]	2025	与 GRPO 相同	是	是	用于多轮交互的推理引导动作	基于分组的奖励
GHPO [57]	2025	策略梯度	是	是	自适应提示精炼	奖励
Skywork R1V2 [58]	2025	带混合奖励信号的 GRPO	是	是	选择性样本缓冲区	多模态奖励
ASPO [59]	2025	带整形优势函数的 GRPO	是	是	直接对优势函数施加截断偏置	基于分组的奖励
TreePo [60]	2025	与 GRPO 相同	是	是	自引导策略回放以降低计算负担	基于分组的奖励
EDGE-GRPO [61]	2025	与 GRPO 相同	是	是	熵驱动的优势与引导误差修正以缓解优势崩溃	基于分组的奖励
DARS [62]	2025	与 GRPO 相同	是	否	通过多阶段回放抽样将计算从中等难度问题重新分配到最难问题	基于分组的奖励
CHORD [63]	2025	GRPO 损失与监督微调损失的加权和	是	是	将监督微调重新框定为策略内 RL 过程中的动态加权辅助目标	基于分组的奖励
PAPO [64]	2025	GRPO 的替代目标	是	是	通过隐式感知损失鼓励在学习推理的同时学习感知	基于分组的奖励
Pass@k Training [65]	2025	与 GRPO 相同	是	是	将 Pass@k 指标作为奖励持续训练模型	基于分组的奖励

This group-relative approach is highly sample-efficient and reduces computational overhead. Consequently, a series of novel algorithms derived from the GRPO framework have been subsequently proposed (see Table. 2), aiming to substantially enhance both the sample efficiency and asymptotic performance of reinforcement learning methodologies.

这种基于组相对性的做法对样本极为高效并减少计算开销。因此，随后提出了一系列由 GRPO 框架派生的新算法 (见表 2)，旨在大幅提升强化学习方法的样本效率与渐近性能。

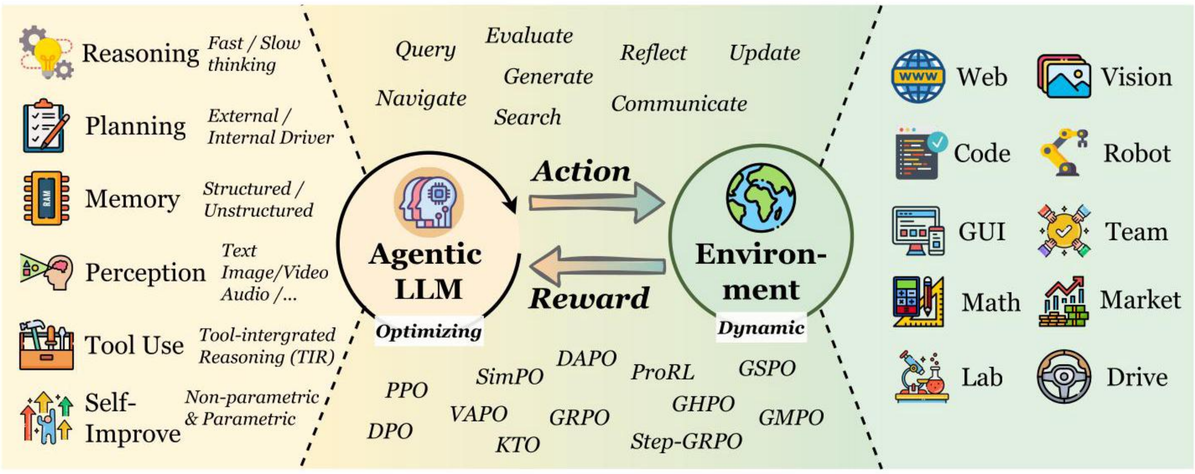

Figure 3: The dynamic interaction process between agentic LLMs and the environment.

图 3: 主体化 LLM 与环境之间的动态交互过程。

3. Agentic RL: The model capability perspective

3. 主体化 RL: 模型能力视角

In this section, we conceptually characterize Agentic RL as the principled training of an autonomous agent composed of a set of key abilities/modules, i.e., planning (Section 3.1), tool use (Section 3.2), memory (Section 3.3), self-improvement (Section 3.4), reasoning (Section 3.5), perception (Section 3.6), and others (Section 3.7), following the classic LLM agent definition [66, 67], as demonstrated in Figure 4. Traditionally, an agent pairs an LLM with mechanisms for planning (e.g., task decomposition and plan selection) [68], reasoning (chain-of-thought or multi-turn inference) [69], external tool invocation [70], long- and short-term memory, and iterative reflection to self-correct and refine behavior. Agentic RL thus treats these components not as static pipelines but as interdependent policies that can be jointly optimized: RL for planning learns multi-step decision trajectories; RL for memory shapes retrieval and encoding dynamics; RL for tool use optimizes invocation timing and fidelity; and RL for reflection drives internal self-supervision and self-improvement. Consequently, our survey systematically examines how RL empowers planning, tool use, memory, reflection, and reasoning in subsequent subsections. We aim to provide a high-level conceptual delineation of RL's applications for agent capabilities, rather than an exhaustive enumeration of all related work, which we provide in Section 4.

在本节中，我们从概念上将主体化 RL 描述为对由一组关键能力/模块组成的自治体的原则性训练，即规划 (第 3.1 节)、工具使用 (第 3.2 节)、记忆 (第 3.3 节)、自我改进 (第 3.4 节)、推理 (第 3.5 节)、感知 (第 3.6 节) 及其他 (第 3.7 节)，遵循经典的 LLM 代理定义 [66, 67]，如图 4 所示。传统上，代理将 LLM 与用于规划 (如任务分解与计划选择)[68]、推理 (思路链或多轮推断)[69]、外部工具调用 [70]、长短期记忆及迭代反思以自我修正和优化行为的机制配对。因此，主体化 RL 将这些组件视为可共同优化的相互依赖策略: 用于规划的 RL 学习多步决策轨迹；用于记忆的 RL 塑造检索与编码动态；用于工具使用的 RL 优化调用时机与准确性；用于反思的 RL 推动内部自监督与自我改进。因此，我们的综述在后续小节系统考察 RL 如何增强规划、工具使用、记忆、反思与推理。我们的目标是提供关于 RL 在代理能力中应用的高层概念性划分，而非穷尽式列举相关工作，相关工作汇总见第 4 节。

3.1. Planning

3.1. 规划

Planning, the deliberation over a sequence of actions to achieve a goal, constitutes a cornerstone of artificial intelligence, demanding complex reasoning, world knowledge, and adaptability [71]. While initial efforts leveraged the innate capabilities of LLMs through prompting-based methods [72] (e.g., ReAct [73]), these approaches lacked a mechanism for adaptation through experience. RL has emerged as a powerful paradigm to address this gap, enabling agents to refine their planning strategies by learning from environmental feedback. The integration of RL into agent planning manifests in two distinct paradigms, distinguished by whether RL functions as an external guide to a structured planning process or as an internal driver that directly evolves the LLM's intrinsic planning policy, which we will detail below.

规划，即为实现目标而对一系列动作进行深思熟虑，是人工智能的基石，要求复杂推理、世界知识与适应性 [71]。尽管早期工作通过基于提示的方法利用了 LLM 的先天能力 [72](例如 ReAct [73])，但这些方法缺乏通过经验进行适应的机制。RL 已成为弥补该缺口的强大范式，使代理能够通过环境反馈来改善其规划策略。RL 与代理规划的整合呈现两种不同范式，取决于 RL 是作为对结构化规划过程的外部引导，还是作为直接演化 LLM 内在规划策略的内部驱动，下面将详细阐述。

RL as an External Guide for Planning. One major paradigm frames RL as an external guide to the planning process, where the LLM's primary role is to generate potential actions within a structured search framework.

将 RL 作为规划的外部引导。一种主要范式将 RL 视为对规划过程的外部引导，其中 LLM 的主要角色是在结构化搜索框架内生成潜在动作。

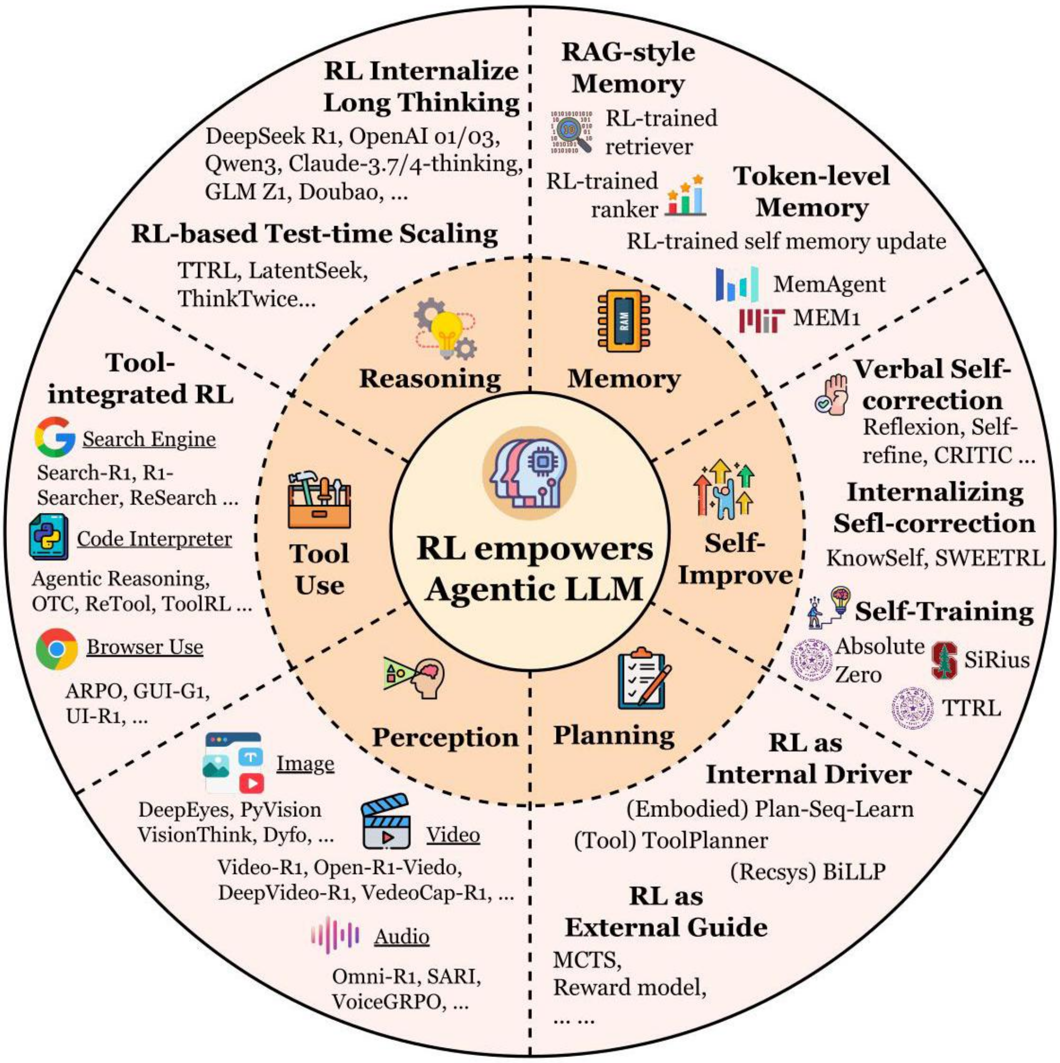

Figure 4: A summary of the overall six aspects where RL empowers agentic LLMs. Note that the representative methods listed here are not exhaustive; please refer to our main text.

图 4:RL 增强主体化 LLM 的六大方面综述。注意此处列示的代表性方法并非详尽；请参阅正文。

Here, RL is not employed to fine-tune the LLM's generative capabilities directly, but rather to train an auxiliary value or heuristic function [68]. This learned function then guides a classical search algorithm, such as Monte Carlo Tree Search (MCTS), by evaluating the quality of different planning trajectories. Representative works like RAP [74] and LATS [75] exemplify this approach. Planning without Search [76] extends this idea by leveraging offline goal-conditioned RL to learn a language-based value critic that guides LLM reasoning and planning without updating the LLM parameters. In this configuration, the LLM acts as a knowledge-rich action proposer, while RL provides adaptive, evaluative feedback for efficient exploration. Beyond static guidance, Learning When to Plan [77] formulates dynamic planning as an RL-driven test-time compute allocation problem, training agents to decide when to invoke explicit planning to balance reasoning performance against computational cost. Conversely, MAPF-DT [78] explores the reverse direction, employing Decision Transformer-based offline RL for decentralized multi-agent path planning, with LLM guidance enhancing adaptability and long-horizon efficiency in dynamic environments.

在此，RL 并未被用来直接微调 LLM 的生成能力，而是用来训练一个辅助的价值或启发式函数 [68]。该学习到的函数随后通过评估不同规划轨迹的质量来指导诸如蒙特卡洛树搜索 (MCTS) 等经典搜索算法。RAP [74] 与 LATS [75] 等代表性工作即为此类范例。《无搜索规划》[76] 扩展了这一思想，通过离线目标条件 RL 学习基于语言的价值评估器，引导 LLM 的推理与规划而不更新 LLM 参数。在这种配置中，LLM 扮演知识丰富的动作提出者，而 RL 为高效探索提供自适应的评估反馈。除静态引导外，Learning When to Plan [77] 将动态规划形式化为一个由 RL 驱动的测试时计算分配问题，训练代理决定何时调用显式规划以在推理性能与计算成本之间取得平衡。相反，MAPF-DT [78] 探索了相反方向，采用基于 Decision Transformer 的离线 RL 用于去中心化多智能体路径规划，LLM 的引导提高了动态环境中的适应性与长时视野效率。

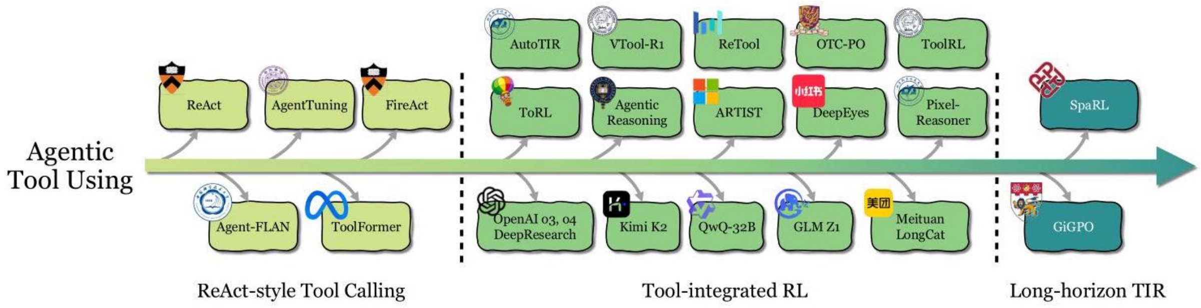

Figure 5: The development of agentic tool use. Note that we only select a small bunch of representative works here to reflect the progress.

图 5: 主体化工具使用的发展。注意我们仅选取少量代表性工作以反映进展。

RL as an Internal Driver of Planning. A second, more integrated paradigm positions RL as an internal driver of the agent's core planning capabilities. This approach casts the LLM directly as a policy model and optimizes its planning behavior through direct environmental interaction. Instead of guiding an external search algorithm, RL-based feedback from trial and error is used to directly refine the LLM's internal policy for generating plans. This is achieved through methods derived from RLHF, such as leveraging DPO on successful versus failed trajectories as seen in ETO [79], or through lifelong learning frameworks. For instance, VOYAGER [80] iteratively builds and refines a skill library from environmental interaction. This paradigm transforms the LLM from a static generator into an adaptive policy that continuously evolves, enhancing its robustness and autonomy in dynamic environments. In a complementary direction, Dynamic Speculative Planning (DSP) [81] embodies an online reinforcement mechanism that adapts the agent's policy to jointly optimize latency and operational cost, demonstrating that internal policy refinement can govern not only task success but also system efficiency. RLTR [82] decouples planning from answer generation and introduces tool-use rewards that directly evaluate action sequence quality, enabling focused optimization of the agent's planning capability without relying on verifiable final answers. AdaPlan and its PilotRL framework [83] leverage global plan-based guidance with progressive RL to enhance LLM agents' long-horizon planning and execution coordination in text game environments like AFLWorld and TextCraft. Planner-R1 [84] examines reward-density effects in agentic RL, showing that shaped, process-level rewards markedly improve learning efficiency and enable smaller models to attain competitive planning capability. Complementing these RL-driven approaches, the Modular Agentic Planner (MAP) [85] introduces a brain-inspired, modular architecture that decomposes planning into specialized LLM modules for conflict monitoring, state evaluation, and coordination. While not RL-based, this architecture provides a promising substrate for integrating reinforcement signals into future agentic planners.

强化学习作为规划的内部驱动因素。第二种更为整合的范式将强化学习定位为代理核心规划能力的内部驱动。这种方法将大模型直接视为策略模型，并通过直接与环境交互来优化其规划行为。不再只是指导外部搜索算法，而是利用试错得到的强化学习反馈直接精炼大模型内部的生成计划的策略。这可通过源于 RLHF 的方法实现，例如在 ETO [79] 中对成功与失败轨迹使用 DPO，或通过终身学习框架。例如，VOYAGER [80] 通过环境交互迭代构建并完善技能库。此范式将大模型从静态生成器转变为持续演化的自适应策略，增强其在动态环境中的鲁棒性与自主性。在互补方向上，动态投机性规划 (DSP)[81] 体现了一种在线强化机制，调整代理策略以联合优化延迟与运行成本，表明内部策略精炼不仅能支配任务成功，还能提升系统效率。RLTR [82] 将规划与答案生成解耦，并引入直接评估动作序列质量的工具使用奖励，使得在不依赖可验证最终答案的情况下，能聚焦优化代理的规划能力。AdaPlan 及其 PilotRL 框架 [83] 利用基于全局计划的引导与渐进式强化学习，提升大模型代理在 AFLWorld 和 TextCraft 等文本游戏环境中的长时程规划与执行协调。Planner-R1 [84] 研究了代理式强化学习中的奖励密度效应，表明经塑形的过程级奖励显著提高学习效率，使得更小的模型也能达到有竞争力的规划能力。与这些以 RL 为驱动的方法互补的是模块化代理规划器 (MAP)[85]，其引入了类脑的模块化架构，将规划分解为用于冲突监测、状态评估与协调的专用大模型模块。尽管非基于 RL，但该架构为未来将强化信号整合到代理规划器中提供了有前景的基底。

Prospective: The Synthesis of Deliberation and Intuition. The prospective horizon for agentic planning lies in the synthesis of these two paradigms: moving beyond the distinction between external search and internal policy optimization. The ultimate goal is to develop an agent that internalizes the structured search process itself, seamlessly blending intuitive, fast plan generation with deliberate, slow, deliberative reasoning. In such a model, RL would not only refine the final plan but also optimize a meta-policy governing the deliberation process: learning when to explore alternative paths, how to prune unpromising branches, and how deeply to reason before committing to an action. This would transform the LLM agent from a component that either proposes actions or acts as a raw policy into an integrated reasoning engine.

前瞻: 深思与直觉的合成。代理规划的前景在于这两种范式的合成: 超越外部搜索与内部策略优化之间的区分。最终目标是开发出能内化结构化搜索过程的代理，将直觉的快速计划生成与审慎的慢速推理无缝融合。在这样的模型中，强化学习不仅会精炼最终计划，还会优化支配深思过程的元策略: 学习何时探索替代路径、如何剪枝无望分支、以及在做出行动决策前应进行多深的推理。这将把大模型代理从一个要么提出行动要么作为原始策略的组件，转变为一个一体化的推理引擎。

3.2. Tool Using

3.2. 工具使用

RL has emerged as a pivotal methodology for evolving tool-enabled language agents from post-hoc, ReAct-style pipelines to deeply interleaved, multi-turn Tool-Integrated Reasoning (TIR) systems. While early paradigms successfully demonstrated the feasibility of tool invocation, their reliance on SFT or prompt engineering limited agents to mimicking static patterns, lacking the strategic flexibility to adapt to novel scenarios or recover from errors. Agentic RL addresses this by shifting the learning paradigm from imitation to outcome-driven optimization, enabling agents to autonomously discover when, how, and which tools to deploy. This evolution charts a clear trajectory, which we explore in three stages. We begin with (1) early ReAct-style tool calling, then examine (2) modern tool-integrated reasoning (TIR) that deeply embeds tool use within cognitive loops, and finally, discuss the prospective challenge of (3) multi-turn TIR, focusing on temporal credit assignment for robust, long-horizon performance.

强化学习已成为将具工具能力的语言代理从事后式、ReAct 风格流水线演进为深度交错、多回合的工具集成推理 (TIR) 系统的关键方法。早期范式虽成功演示了调用工具的可行性，但其对 SFT 或提示工程的依赖使代理仅能模仿静态模式，缺乏在新场景中适应或从错误中恢复的策略灵活性。代理式强化学习通过将学习范式从模仿转向以结果为导向的优化来解决此问题，使代理能自主发现何时、如何以及使用哪些工具。这一演进勾勒出清晰轨迹，我们将其分为三阶段探讨。首先是 (1) 早期 ReAct 风格的工具调用，然后审视 (2) 将工具使用深度嵌入认知回路的现代工具集成推理 (TIR)，最后讨论面向稳健长时程表现的 (3) 多回合 TIR 的前瞻性挑战，即时间信用分配问题。

ReAct-style Tool Calling. Early paradigms for tool invocation predominantly relied on either prompt engineering or SFT to elicit tool-use behaviors. The (I) prompt engineering approach, exemplified by ReAct [73], leveraged few-shot exemplars to guide an LLM to interleave reasoning traces and actions within a ”Thought-Action-Observation” cycle, capitalizing on the model’s in-context learning abilities. Going beyond, (II) SFT-based methods were introduced to internalize models’ tool-use capabilities. Frameworks like Toolformer [86] employed a self-supervised objective to teach models where to insert API calls, while others like FireAct [87], AgentTuning [88], Agent-FLAN [89] fine-tuned models on expert-generated or curated datasets of tool-interaction trajectories (e.g., AgentBank [90], APIBank [91]). Although SFT improved the reliability of tool invocation, both of these early approaches are fundamentally constrained by their imitative nature. They train agents to replicate static, pre-defined patterns of tool use, thereby lacking the strategic flexibility to adapt to novel scenarios or recover from unforeseen errors, a limitation that RL-centric approaches directly address by shifting the learning objective from imitation to outcome-driven optimization.

ReAct 风格的工具调用。早期的工具调用范式主要依赖提示工程或 SFT 来引发工具使用行为。以 ReAct [73] 为代表的 (I) 提示工程方法通过少量示例引导大模型在“思考-行动-观察”循环中交替插入推理痕迹与动作，利用模型的上下文学习能力。更进一步，(II) 基于 SFT 的方法被引入以使模型内化工具使用能力。像 Toolformer [86] 这样的框架采用自监督目标教会模型在何处插入 API 调用，而 FireAct [87]、AgentTuning [88]、Agent-FLAN [89] 等则在专家生成或策划的工具交互轨迹数据集 (例如 AgentBank [90]、APIBank [91]) 上微调模型。尽管 SFT 提高了工具调用的可靠性，但这两类早期方法本质上受限于其模仿性: 它们训练代理复制静态的、预定义的工具使用模式，因此缺乏在新场景中适应或从意外错误中恢复的策略灵活性，而这是以强化学习为中心的方法通过将学习目标转向以结果为导向的优化所直接解决的缺陷。

Tool-integrated RL. Building on the limitations of purely imitative paradigms, RL-based approaches for tool use shift the objective from replicating fixed patterns to optimizing end-task performance. This transition enables agents to strategically decide when, how, and in what combination to invoke tools, adapting dynamically to novel contexts and unforeseen failures. At the foundation, frameworks such as ToolRL [92] demonstrate that, even when initialized from base models without any imitation traces, RL training can elicit emergent capabilities, e.g., self-correction of faulty code, adaptive adjustment of invocation frequency, and the composition of multiple tools for complex sub-tasks. Subsequently, a recent surge in research has produced works such as OTC-PO [93], ReTool [94], AutoTIR [95], VTool-R1 [96], DeepEyes [97], Pixel-Reasoner [98], Agentic Reasoning [99], ARTIST [100], ToRL [101] and numerous other works [102, 103, 104, 105, 106, 107, 108, 109, 110], which employ RL policies that interleave symbolic computation (e.g., code execution, image editing) with natural-language reasoning within a single rollout. This integrated control loop allows the agent to balance precise, tool-mediated operations with flexible verbal inference, tailoring the reasoning process to the evolving task state. Recent work [59] theoretically proves that TIR fundamentally expands LLM capabilities beyond the ”invisible leash” of pure-text RL by introducing deterministic tool-driven state transitions, establishes token-efficiency arguments for feasibility under finite budgets, and proposes Advantage Shaping Policy Optimization (ASPO) to stably guide agentic tool use.

> 工具集成的强化学习。基于纯模仿范式的局限性，基于强化学习的工具使用方法将目标从复制固定模式转向优化最终任务表现。这一转变使代理能够有策略地决定何时、如何以及以何种组合调用工具，动态适应新情境和意外失败。在基础层面，诸如 ToolRL [92] 的框架表明，即便从没有任何模仿痕迹的基础模型初始化，RL 训练也能激发涌现能力，例如对有缺陷代码的自我纠正、对调用频率的自适应调整以及为复杂子任务组合多个工具。随后，研究热潮涌现出 OTC-PO [93]、ReTool [94]、AutoTIR [95]、VTool-R1 [96]、DeepEyes [97]、Pixel-Reasoner [98]、Agentic Reasoning [99]、ARTIST [100]、ToRL [101] 及众多其他工作 [102, 103, 104, 105, 106, 107, 108, 109, 110]，这些工作采用在单次 rollout 中交织符号计算 (如代码执行、图像编辑) 与自然语言推理的 RL 策略。该集成控制回路使代理在精确的工具操作与灵活的语言推理之间进行权衡，根据不断演化的任务状态定制推理过程。近期工作 [59] 从理论上证明，TIR 通过引入确定性的工具驱动状态转换，实质上扩展了 LLM 能力，超越了纯文本 RL 的“无形牵引”，并提出了在有限预算下的 token 效率论证，以及用于稳定引导自治工具使用的 Advantage Shaping Policy Optimization (ASPO)。

Today, such tool-integrated reasoning is no longer a niche capability but a baseline feature of advanced agentic models. Mature commercial and open-source systems, such as OpenAI’s DeepResearch and o3 [111], Kimi K2 [112], Qwen QwQ-32B [113], Zhipu GLM Z1 [114], Microsoft rStar2-Agent [115] and Meituan LongCat [116], routinely incorporate these RL-honed strategies, underscoring the centrality of outcome-driven optimization in tool-augmented intelligence.

> 如今，这类工具集成推理不再是小众能力，而是先进自治模型的基线特性。成熟的商用和开源系统，如 OpenAI 的 DeepResearch 与 o3 [111]、Kimi K2 [112]、Qwen QwQ-32B [113]、智谱 GLM Z1 [114]、微软 rStar2-Agent [115] 和美团 LongCat [116]，常规地采用这些经 RL 打磨的策略，凸显了以结果为导向的优化在工具增强智能中的核心地位。

Prospective: Long-horizon TIR. While tool-integrated RL has proven effective for optimizing actions within a single reasoning loop, the primary frontier lies in extending this capability to robust, long-horizon tasks that require multi-turn reasoning [117]. This leap is fundamentally bottlenecked by the challenge of temporal credit assignment [118]. Current RL approaches often depend on sparse, trajectory-level/outcome-based rewards, making it difficult to pinpoint which specific tool invocation in a long, interdependent sequence contributed to success or failure. While nascent research has begun to explore more granular reward schemes, such as turn-level advantage estimation in GiGPO [119] and SpaRL [120], these are still early steps. Consequently, developing more granular credit assignment mechanisms that can accurately guide the agent through complex decision chains without inadvertently punishing useful exploration or promoting reward hacking remains a critical and largely unsolved problem for advancing agentic systems.

前瞻: 长时域 TIR。尽管工具集成的 RL 在优化单次推理回路内的动作方面已被证明有效，主要前沿在于将该能力扩展到需要多轮推理的鲁棒长时域任务 [117]。这一飞跃在根本上受制于时间信用分配的挑战 [118]。当前 RL 方法通常依赖稀疏的轨迹级/基于结果的奖励，很难定位在长且相互依赖的序列中哪一次具体的工具调用促成了成功或导致了失败。尽管初步研究已开始探索更细粒度的奖励方案，例如 GiGPO [119] 与 SpaRL [120] 中的回合级 advantage 估计，但这些仍是早期尝试。因此，开发能够精确引导代理穿越复杂决策链的更细粒度信用分配机制，同时不误伤有益探索或鼓励奖励投机，仍然是推进自治系统的关键且尚未解决的问题。

3.3. Memory

3.3. Memory

Agentic RL transforms memory modules from passive data stores into dynamic, RL-controlled subsystems, deciding what to store, when to retrieve, and how to forget similar to human [121]. This section traces this evolution through four representative phases.

自治式 RL 将记忆模块从被动数据存储转变为动态的 RL 控制子系统，决定何时存储、何时检索及如何遗忘，类似人类 [121]。本节通过四个代表性阶段追溯这一演进。

RL in RAG-style Memory. Early systems (e.g., retrieval-augmented generation) treated memory as an external datastore; when RL was employed at all, it solely regulated when to perform queries. Several classic memory systems without RL involvement, such as MemoryBank [122], MemGPT [123], and HippoRAG [124], adopt predefined memory management strategies that specify how to store, integrate, and retrieve information (e.g., storage via vector databases or knowledge graphs; retrieval based on semantic similarity or topological connectivity). Subsequently, RL was incorporated into the memory management pipeline as a functional component. A notable example is the framework proposed in [125], where the RL policy adjusts retrieval behavior through prospective reflection (multi-level summarization) and retrospective reflection (reinforcing retrieval outcomes). Nevertheless, the memory medium itself remained static (e.g., simple vector store or summary buffer), and the agent exerted no control over the write processes. Recently, Memory-R1 [126] introduces a RL-based memory-augmented Agent framework where a Memory Manager learns to perform structured operations (ADD/UPDATE/DELETE/NOOP) via PPO or GRPO based on downstream QA performance, while an Answer Agent employs a Memory Distillation policy over RAG-retrieved entries to reason and answer. Follow-up works like Mem- α [127] and Memory-as-action [128] have also explored RL for training agents into automatic memory manager.

RAG 风格记忆中的 RL。早期系统 (例如检索增强生成) 将记忆视为外部数据存储；即便使用 RL，也仅调节何时执行查询。若干不涉 RL 的经典记忆系统，如 MemoryBank [122]、MemGPT [123] 与 HippoRAG [124], 采用预定义的记忆管理策略来指定如何存储、整合和检索信息 (例如通过向量数据库或知识图谱存储; 基于语义相似性或拓扑连通性检索)。随后, RL 被纳入记忆管理流水线作为功能组件。一个显著例子是 [125] 中提出的框架, 其中 RL 策略通过前瞻性反思 (多层摘要) 和回顾性反思 (强化检索结果) 来调整检索行为。然而，记忆媒介本身仍然静态 (例如简单向量存储或摘要缓冲区), 代理对写入过程没有控制。最近, Memory-R1 [126] 引入了一个基于 RL 的记忆增强代理框架, 记忆管理器通过 PPO 或 GRPO 根据下游问答表现学习执行结构化操作 (ADD/UPDATE/DELETE/NOOP), 而回答代理则对 RAG 检索到的条目使用记忆蒸馏策略进行推理与回答。后续工作如 Mem- α [127] 与 Memory-as-action [128] 也探索了用 RL 培训代理成为自动记忆管理器。

RL for Token-level Memory. Subsequent advancements introduced models equipped with explicit, trainable memory controllers, enabling agents to regulate their own memory states (often stored in token form) without relying on fixed, external memory systems. Notably, such memory is commonly instantiated in two forms. The first is (I) explicit tokens, corresponding to human-readable natural language. For example, in MemAgent [129], the agent maintains a natural-language memory pool alongside the LLM, with an RL policy determining, at each segment, which tokens to retain or overwrite, effectively compressing long-context inputs into concise, informative summaries. Similar approaches include MEM1 [130] and Memory Token [131], both of which explicitly preserve a pool of natural-language memory representations. More frequently, works like ReSum [132], context folding [133] has also explored RL for context memory management. The second form is (II) implicit tokens, where memory is maintained in the form of latent embeddings. A representative line of work includes MemoryLLM [134] and M+ [135], in which a fixed set of latent tokens serves as ”memory tokens.” As the context evolves, these tokens are repeatedly retrieved, integrated into the LLM's forward computation, and updated, thereby preserving contextual information and exhibiting strong resistance to forgetting. Unlike explicit tokens, these memory tokens are not tied to human-readable text but rather constitute a machine-native form of memory. Related efforts include IMM [136] and Memory [137]. Across both paradigms, these approaches empower agents to autonomously manage their memory banks, delivering significant improvements in long-context understanding, continual adaptation, and self-improvement. MemGen [138] for the first time proposes the paradigm of leveraging latent memory tokens for carrying and generating experiential knowledge, posing promising directions for RL-based latent memory.

用于令牌级记忆的强化学习。随后进展引入了配备显式可训练记忆控制器的模型，使智能体能够调节自身的记忆状态 (常以令牌形式存储)，而无需依赖固定的外部记忆系统。值得注意的是，此类记忆通常以两种形式体现。第一类为 (I) 显式令牌，对应可读的人类语言。例如，在 MemAgent [129] 中, 智能体在与大模型并行维护一个自然语言记忆池, 并由强化学习策略在每个片段决定保留或覆盖哪些令牌，从而将长上下文输入压缩为简洁且信息量高的摘要。类似方法还有 MEM1 [130] 和 Memory Token [131]，它们都显式保留了一组自然语言记忆表示。更常见的工作如 ReSum [132]、context folding [133] 也探讨了用于上下文记忆管理的强化学习。第二类为 (II) 隐式令牌，记忆以潜在嵌入的形式维护。代表性工作包括 MemoryLLM [134] 和 M+ [135]，其中一组固定的潜在令牌作为“记忆令牌”使用。随着上下文演进，这些令牌被反复检索、整合到大模型的前向计算中并更新，从而保留上下文信息并表现出强大的抗遗忘能力。与显式令牌不同, 这些记忆令牌不对应可读文本, 而是构成机器原生的记忆形式。相关工作还有 IMM [136] 和 Memory [137]。在这两种范式中，这些方法使智能体能够自主管理其记忆库，在长上下文理解、持续适应和自我改进方面带来显著提升。MemGen [138] 首次提出利用潜在记忆令牌承载与生成经验性知识的范式，为基于强化学习的潜在记忆指明了有前景的方向。

Table 3: An overview of three classic categories of agent memory; works marked with † directly employ RL. The list here is not exhaustive, and we refer readers interested in broader agent memory to [121].

表 3: 三类经典智能体记忆的概览；标注有 † 的工作直接采用强化学习。此处所列并非详尽，欲了解更广泛智能体记忆的读者可参考 [121]。

Method	Type	Key Characteristics
RAG-style Memory		
MemoryBank [122]	External Store	Static memory with predefined storage/retrieval rules
MemGPT [123]	External Store	OS-like agent with static memory components
HippoRAG [124]	External Store	Neuro-inspired memory with heuristic access
Prospect † [125]	RL-guided Retrieval	Uses RL for reflection-driven retrieval adjustment
Memory-R1† [126]	RL-guided Retrieval	RL-driven memory management: ADD/UPDATE/DELETE/NOOP
Mem- α† [127]	RL-guided Retrieval	RL-guided agents for memory retrieval
Memory-as-action [128]	RL-guided Manage	Ene-to-end training agents for memory management
Token-level Memory		
MemAgent† [129]	Explicit Token	RL controls which NL tokens to retain or overwrite
MEM1’ [130]	Explicit Token	Memory pool managed by RL to enhance context handling
Memory Token [131]	Explicit Token	Structured memory for reasoning disentanglement
ReSum+ [132]	Explicit Token	Turn-wise Interaction summary for ReAct agents
Context Folding+ [133]	Explicit Token	Context folding for ReAct agents
MemoryLLM [134]	Latent Token	Latent tokens repeatedly integrated and updated
M+ [135]	Latent Token	Scalable memory tokens for long-context tracking
IMM [136]	Latent Token	Decouples word representations and latent memory
Memory [137]	Latent Token	Forget-resistant memory tokens for evolving context
MemGen† [138]	Latent Token	Context-sensitive latent token as memory carriers
Structured Memory		
Zep [139]	Temporal Graph	Temporal knowledge graph enabling structured retrieval
A-MEM [140]	Atomic Memory Notes	Symbolic atomic memory units; structured storage
G-Memory [141]	Hierarchical Graph	Multi-level memory graph with topological structure
Mem0 [142]	Structured Graph	Agent memory with full-stack graph-based design

方法	类型	关键特征
RAG 风格记忆		
MemoryBank [122]	外部存储	具有预定义存取规则的静态记忆
MemGPT [123]	外部存储	具有静态记忆组件的类操作系统代理
HippoRAG [124]	外部存储	受神经启发、带启发式访问的记忆
Prospect † [125]	强化学习引导检索	使用强化学习进行反思驱动的检索调整
Memory-R1† [126]	强化学习引导检索	强化学习驱动的记忆管理: 添加/更新/删除/无操作
Mem- α† [127]	强化学习引导检索	用于记忆检索的强化学习引导代理
Memory-as-action [128]	强化学习引导的管理	端到端训练的记忆管理代理
Token 级记忆		
MemAgent† [129]	显式 Token	强化学习控制保留或覆盖哪些自然语言 token
MEM1’ [130]	显式 Token	由强化学习管理的记忆池以增强上下文处理
Memory Token [131]	显式 Token	用于解耦推理的结构化记忆
ReSum+ [132]	显式 Token	为 ReAct 代理的回合式交互摘要
Context Folding+ [133]	显式 Token	面向 ReAct 代理的上下文折叠
MemoryLLM [134]	潜在 Token	潜在 token 被反复整合和更新
M+ [135]	潜在 Token	可扩展的记忆 token 用于长上下文跟踪
IMM [136]	潜在 Token	将词表示与潜在记忆解耦
Memory [137]	潜在 Token	对遗忘有抵抗力的记忆 token，以适应不断变化的上下文
MemGen† [138]	潜在 Token	作为记忆载体的上下文敏感潜在 token
结构化记忆		
Zep [139]	时序图	支持结构化检索的时序知识图
A-MEM [140]	原子记忆笔记	符号化的原子记忆单元；结构化存储
G-Memory [141]	层级图	具拓扑结构的多级记忆图
Mem0 [142]	结构化图	具全栈图设计的代理记忆

Prospective: RL for Structured Memory. Building on token-level approaches, recent trends are moving toward structured memory representations, which organize and encode information beyond flat token sequences. Representative examples include the temporal knowledge graph in Zep [139], the atomic memory notes in A-MEM [140], and the hierarchical graph-based memory designs in G-Memory [141] and Mem0 [142]. These systems capture richer relational, temporal, or hierarchical dependencies, enabling more precise retrieval and reasoning. However, their management, spanning insertion, deletion, abstraction, and linkage updates, has thus far been governed by handcrafted rules or heuristic strategies. To date, little work has explored the use of RL to dynamically control the construction, refinement, or evolution of such structured memory, making this an open and promising direction for advancing agentic memory capabilities.

前瞻: 用于结构化记忆的强化学习。基于 token 级方法，近期趋势转向结构化记忆表示，超越扁平的 token 序列来组织和编码信息。代表性例子包括 Zep [139] 中的时间知识图、A-MEM [140] 的原子记忆笔记，以及 G-Memory [141] 和 Mem0 [142] 中的分层图式记忆设计。这些系统捕捉更丰富的关系性、时间性或层级依赖，从而实现更精确的检索与推理。然而，其管理 (包括插入、删除、抽象与链接更新) 至今多由手工规则或启发式策略支配。迄今为止，几乎没有研究探索使用 RL 动态控制此类结构化记忆的构建、精炼或演化，使其成为提升智能体记忆能力的一个开放且有前景的方向。

3.4. Self-Improvement

3.4. 自我改进

As LLM agents evolve, recent research increasingly emphasizes RL as a mechanism for ongoing reflection, enabling agents to learn from their own mistakes across planning, reasoning, tool use, and memory [143]. Rather than relying exclusively on data-driven training phases or static reward models, these systems incorporate iterative, self-generated feedback loops, ranging from prompt-level heuristics to fully fledged RL controllers, to guide agents toward continual self-improvement.

随着 LLM 智能体的发展，近期研究越来越强调将 RL 作为持续反思的机制，使智能体能在规划、推理、工具使用与记忆等方面从自身错误中学习 [143]。这些系统不再仅依赖数据驱动的训练阶段或静态奖励模型，而是引入迭代的、自生成的反馈回路，从提示级启发式方法到完整的 RL 控制器，以引导智能体实现持续自我改进。

RL for Verbal Self-correction. Initial methods in this vein leveraged prompt-based heuristics, sometimes referred to as verbal reinforcement learning, where agents generate an answer, linguistically reflect on its potential errors, and subsequently produce a refined solution, all within a single inferential pass without gradient updates. Prominent examples include Reflexion [144], Self-refine [145], CRITIC [146], and Chain-of-Verification [147]. For instance, the Self-Refine [145] protocol directs an LLM to iteratively polish its output using three distinct prompts for generation, feedback, and refinement, proving effective across domains like reasoning and programming. To enhance the efficacy and robustness of such self-reflection, several distinct strategies have been developed: (I) multiple sampling, which involves generating multiple output rollouts by sampling from the model's distribution. By aggregating critiques or solutions from multiple attempts, the agent can improve the consistency and quality of its self-reflection. This method has been widely studied in works like If-or-Else [148], UALA [149] and Multi-agent Verification [150]. This approach is conceptually analogous to test-time scaling techniques, so we refer the reader to [118] for more details; (II) structured reflection workflows, rather than prompting for a monolithic reflection on a final answer, prescribe a more dedicated and granular workflow. For example, Chain-of-Verification [147] manually decomposes the process into distinct ”Retrieving, Rethinking, and Revising” stages; (III) external guidance, which grounds the reflection process in verifiable, objective feedback by incorporating external tools. These tools include code interpreter as seen in Self-Debugging [151], CAD modeling programs in Luban [152], mathematical calculators in T1 [153], step-wise reward models [154], and others [146].

用于口头自我纠正的 RL。该方向的早期方法利用基于提示的启发式策略，有时称为口头强化学习，智能体在单次推理过程中先生成答案、用语言反思其潜在错误，然后产生精炼解答，无需梯度更新。代表作包括 Reflexion [144]、Self-refine [145]、CRITIC [146] 和 Chain-of-Verification [147]。例如，Self-Refine [145] 协议指示 LLM 通过三个不同的提示 (生成、反馈、精炼) 迭代打磨输出，证明对推理和编程等领域有效。为提升此类自我反思的有效性与鲁棒性，已发展出若干策略:(I) 多重采样，通过从模型分布中采样生成多个输出回合。通过汇总多次尝试的批评或解答，智能体可提高自我反思的一致性与质量。该方法在 If-or-Else [148]、UALA [149] 和 Multi-agent Verification [150] 等工作中得以广泛研究。此法在概念上类似于测试时扩展技术，详见 [118]；(II) 结构化反思流程，不是对最终答案进行整体化反思，而是制定更专门和细化的流程。例如 Chain-of-Verification [147] 将过程手动拆分为“检索、重新思考与修正”等独立阶段；(III) 外部引导，将反思过程基于可验证的客观反馈，通过整合外部工具来实现。这些工具包括 Self-Debugging [151] 中的代码解释器、Luban [152] 中的 CAD 建模程序、T1 [153] 中的数学计算器、逐步奖励模型 [154] 以及其他工作 [146]。

RL for Internalizing Self-correction. While verbal self-correction offers a potent inference-time technique, its improvements are ephemeral and confined to a single session. To instill a more durable and generalized capability for self-improvement, subsequent research has employed RL with gradient-based updates to internalize these reflective feedback loops directly into the model's parameters and to fundamentally enhance the model's inherent ability to identify and correct its own errors. This paradigm has been applied across multiple domains. For instance, KnowSelf [155] leverages DPO and RPO [156] to enhance agents' self-reflection capabilities in text-based game environments, while Reflection-DPO [157] focuses on user-agent interaction scenarios, enabling agents to better infer user intent through reflective reasoning. DuPo [158] employs RL with dual-task feedback to enable annotation-free optimization, enhancing LLM agents' self-correction across translation, reasoning, and reranking tasks. SWEET-RL [159] and ACC-Collab [160] adopt a slightly different setting from the above works: they train an external critic model to provide higher-quality revision suggestions for the actor agent's actions. Nonetheless, the underlying principle remains closely aligned.

用于将自我纠正内化的 RL。尽管口头自我纠正是一个强有力的推理时技术，但其改进是短暂且局限于单次会话。为将这种反思反馈回路更持久和泛化地内化到模型参数中，从而根本上增强模型识别并纠正自身错误的能力，后续研究采用了带梯度更新的 RL。该范式已在多个领域应用。例如，KnowSelf [155] 利用 DPO 与 RPO [156] 在基于文本的游戏环境中增强智能体的自我反思能力，Reflection-DPO [157] 则聚焦于用户—智能体交互场景，使智能体通过反思推理更好地推断用户意图。DuPo [158] 使用双任务反馈的 RL 实现无标注优化，提升 LLM 智能体在翻译、推理与重排序任务中的自我纠正能力。SWEET-RL [159] 与 ACC-Collab [160] 的设置略有不同: 它们训练一个外部评论模型，为执行体智能体的行为提供更高质量的修订建议。尽管如此，其基本原理仍高度一致。

RL for Iterative Self-training. Moving toward full agentic autonomy, the third and most advanced class of models combines reflection, reasoning, and task generation into a self-sustaining loop, enabling unbounded self-improvement without human-labeled data. These methods can be distinguished by the architecture of their learning loops: (I) Self-play and search-guided refinement, which emulates classic RL paradigms like AlphaZero. R-Zero [161], for instance, employs a Monte Carlo Tree Search (MCTS) to explore a reasoning tree, using the search results to iteratively train both a policy LLM (the actor) and a value LLM (the critic) entirely from scratch. Similarly, the ISC framework [162] operationalizes a cycle of ”Imagination, Searching, and Criticizing,” where the agent generates potential solution paths, uses a search algorithm to explore them, and applies a critic to refine its reasoning strategy before producing a final answer. (II) Execution-guided curriculum generation, where the agent creates its own problems and learns from verifiable outcomes. Absolute

Zero [163] exemplifies this by proposing its own tasks, attempting solutions, verifying them via execution, and using the outcome-based reward to refine its policy. Similarly, Self-Evolving Curriculum [164] enhances this process by framing problem selection itself as a non-stationary bandit task, allowing the agent to strategically generate a curriculum that maximizes its learning gains over time. TTRL [165] applies this principle for on-the-fly adaptation to a single problem. At test time, it uses execution-based rewards to rapidly fine-tune a temporary copy of the agent's policy for the specific task at hand; this specialized policy is then used to generate the final answer before being discarded. Though differing in whether the learning is permanent or ephemeral, all these methods underscore a powerful, unified strategy: harnessing execution-based feedback to autonomously guide the agent's reasoning process. ALAS [166] constructs an autonomous pipeline that crawls web data, distills it into training signals, and continuously fine-tunes LLMs, thereby enabling self-training and self-evolution without manual dataset curation. (III) Collective bootstrapping, where learning is accelerated by aggregating shared experience. SiriuS [167], for example, constructs and augments a live repository of successful reasoning trajectories from multi-agent interactions, using this growing knowledge base to bootstrap its own training curriculum. MALT [168] shares a similar motivation, yet its implementation is limited to a three-agent setup. Nevertheless, all these methods define feedback loops that are internally generated and continuously evolving, representing a significant step toward truly autonomous agents.

用于迭代自我训练的强化学习。迈向完全的代理自治，第三类也是最先进的模型将反思、推理与任务生成结合为自我维持的循环，使得在无人工标注数据下实现无界自我改进成为可能。这些方法可按其学习循环的架构区分:(I) 自对弈与搜索引导的精化，模仿如 AlphaZero 的经典 RL 范式。例如 R-Zero [161] 使用蒙特卡罗树搜索 (MCTS) 来探索推理树，利用搜索结果从零开始迭代训练策略大模型 (执行者) 与价值大模型 (评论者)。类似地，ISC 框架 [162] 将“想象、搜索与批评”的循环实现为: 代理生成潜在解路径，使用搜索算法探索这些路径，并用评论者在给出最终答案前精化其推理策略。(II) 执行引导的课程生成，代理自主创建问题并从可验证的结果中学习。Absolute Zero [163] 的做法是提出自身任务、尝试解答、通过执行验证，并用基于结果的奖励来改进策略。Self-Evolving Curriculum [164] 通过将问题选择本身视为非平稳多臂老虎机任务来增强该过程，使代理能够策略性地生成能最大化其长期学习收益的课程。TTRL [165] 将此原理用于对单一问题的即时适应；在测试时，它用基于执行的奖励快速微调代理策略的临时副本以针对该特定任务；然后用此专门化策略生成最终答案并丢弃。尽管在学习是永久还是短暂方面存在差异，所有这些方法都强调一个强有力的统一策略: 利用基于执行的反馈自主引导代理的推理过程。ALAS [166] 构建了一个自动化管道，抓取网络数据，将其蒸馏为训练信号并持续微调 LLM，从而实现无需人工数据集筛选的自我训练和自我进化。(III) 集体自举，通过聚合共享经验来加速学习。例如 SiriuS [167] 从多代理交互中构建并扩充成功推理轨迹的实时库，利用这一不断增长的知识库自举其训练课程。MALT [168] 的动机类似，但其实现限定于三代理设置。尽管如此，所有这些方法都定义了内部生成且持续演化的反馈循环，代表了朝向真正自治代理迈进的重要一步。

Prospective: Meta Evolution of Reflection Ability. While current research successfully uses RL to refine an agent's behavior through reflection, the reflection process itself remains largely handcrafted and static. The next frontier lies in applying RL at a higher level of abstraction to enable meta-learning for adaptive reflection, focusing not just on correcting an error, but on learning how to self-correct more effectively over time. In this paradigm, the agent may learn a meta-policy that governs its own reflective strategies. For instance, it could learn to dynamically choose the most appropriate form of reflection for a given task, deciding whether a quick verbal check is sufficient or if a more costly, execution-guided search is necessary. Furthermore, an agent could use long-term outcomes to evaluate and refine the very heuristics it uses for self-critique, effectively learning to become a better internal critic. By optimizing the reflective mechanism itself, this approach moves beyond simple self-correction and toward a state of continuous self-improvement in the learning process, representing a crucial step toward agents that can not only solve problems but also autonomously enhance their fundamental capacity to learn from experience.

前瞻: 反思能力的元演化。虽然现有研究成功地使用 RL 通过反思来改进代理行为，但反思过程本身在很大程度上仍是人为设计且静态的。下一前沿是将在更高抽象层面应用 RL，以实现适应性反思的元学习，关注的不仅是纠正错误，而是学习如何随着时间更有效地自我纠正。在这一范式中，代理可能学习一个元策略来治理其自身的反思策略。例如，它可以学会为给定任务动态选择最合适的反思形式，决定是进行快速的口头检查足矣，还是需要更昂贵的执行引导搜索。此外，代理可以使用长期结果来评估并精化其用于自我批评的启发式规则，实质上学会成为更好的内部评论者。通过优化反思机制本身，这一方法超越了简单的自我纠正，走向在学习过程中持续自我改进的状态，代表了朝向那些不仅能解决问题且能自主增强其从经验中学习的基本能力的代理迈出的关键一步。

3.5. Reasoning

3.5. 推理

Reasoning in large language models can be broadly categorized into fast reasoning and slow reasoning, following the dual-process cognitive theory [169, 24]. Fast reasoning corresponds to rapid, heuristic-driven inference with minimal intermediate steps, while slow reasoning emphasizes deliberate, structured, and multi-step reasoning. Understanding the trade-offs between these two paradigms is crucial for designing models that balance efficiency and accuracy in complex problem-solving.

大型语言模型的推理大致可分为快速推理与慢速推理，遵循双过程认知理论 [169, 24]。快速推理对应于快速的、启发式驱动的推断，具有最少的中间步骤，而慢速推理强调深思熟虑的、结构化的多步推理。理解这两种范式之间的权衡对于设计在复杂问题求解中平衡效率与准确性的模型至关重要。

Fast Reasoning: Intuitive and Efficient Inference Fast reasoning models operate in a manner analogous to System 1 [18] cognition: quick, intuitive, and pattern-driven. They generate immediate responses without explicit step-by-step deliberation, excelling in tasks that prioritize fluency, efficiency, and low latency. Most conventional LLMs fall under this category, where reasoning is implicitly encoded in next-token prediction [2, 170]. However, this efficiency comes at the cost of systematic reasoning, making these models more vulnerable to factual errors, biases, and shallow generalization.

快速推理: 直觉且高效的推断。快速推理模型的运作类似于系统 1 [18] 认知: 快速、直觉且以模式为驱动。它们在没有显式逐步推敲的情况下生成即时回复，擅长强调流畅性、效率和低延迟的任务。大多数传统 LLM 屬于此类，其推理隐式编码在下一个词预测中 [2, 170]。然而，这种效率以牺牲系统化推理为代价，使这些模型更易出现事实性错误、偏见和浅层泛化。

To address the severe hallucination problems in fast reasoning, current research has largely focused on various direct approaches. Some studies attempt to mitigate errors and hallucinations in the next-token prediction paradigm by leveraging internal mechanisms [171, 172, 173] or by simulating human-like cognitive reasoning. Other works propose introducing both external and internal confidence estimation methods [174, 175] to identify more reliable reasoning paths. However, constructing such external reasoning frameworks often risks algorithmic adaptivity issues and can easily fall into the complexity trap.

为解决快速推理中严重的幻觉问题，当前研究主要集中在各种直接方法上。一些研究试图通过利用内部机制 [171, 172, 173] 或模拟类人认知推理来缓解下一个词预测范式中的错误和幻觉。其他工作提出引入外部和内部置信度估计方法 [174, 175] 来识别更可靠的推理路径。然而，构建此类外部推理框架往往存在算法适应性问题的风险，且容易陷入复杂性陷阱。

Slow Reasoning: Deliberate and Structured Problem Solving In contrast, slow reasoning models are designed to emulate System 2 cognition [18] by explicitly producing intermediate reasoning traces. Techniques such as chain-of-thought prompting, multi-step verification [176], and reasoning-augmented reinforcement learning allow these models to engage in deeper reflection and achieve greater logical consistency. While slower in inference due to extended reasoning trajectories, they achieve higher accuracy and robustness in knowledge-intensive tasks such as mathematics, scientific reasoning, and multi-hop question answering [177]. Representative examples include OpenAI's o1 [31] and o3 series [33], DeepSeek-R1 [32], as well as methods that incorporate dynamic test-time scaling [178, 179, 180, 172] or reinforcement learning [181, 47, 182, 183, 184, 185] for reasoning.

慢速推理: 深思熟虑且有结构的问题解决相比之下，慢速推理模型旨在通过明确产生中间推理痕迹来模拟 System 2 认知 [18]。链式思维提示、多步验证 [176] 及增强推理的强化学习等技术使这些模型能够进行更深入的反思并实现更强的逻辑一致性。尽管由于更长的推理轨迹推理速度较慢，但在数学、科学推理和多跳问答等知识密集型任务中能获得更高的准确性与鲁棒性 [177]。具代表性的例子包括 OpenAI 的 o1 [31] 与 o3 系列 [33]、DeepSeek-R1 [32]，以及将动态测试时扩展 [178, 179, 180, 172] 或强化学习 [181, 47, 182, 183, 184, 185] 用于推理的方法。

Modern slow reasoning exhibits output structures that differ substantially from fast reasoning. These include a clear exploration and planning structure, frequent verification and checking behaviors, and generally longer inference lengths and times. Past work has explored diverse patterns for constructing long-chain reasoning outputs. Some methods—Macro-o1, HuatuoGPT-o1, and AlphaZero—have attempted to synthesize long chains-of-thought via structured, agentic search [186, 187, 188]. Other approaches focus on generating long-CoT datasets that embody specific deliberative or reflective thinking patterns; examples include HiICL-MCTS, LLaVA-CoT, rStar-Math, and ReasonFlux [189, 190, 191, 192]. Recent approaches that perform reasoning in the latent space leverage latent representations to conduct parallel reasoning and explore diverse reasoning trajectories [193, 194]. With the progress of pretrained foundation models, more recent work has shifted toward self-improvement paradigms-frequently instantiated with reinforcement learning-to further enhance models' reasoning capabilities [181, 47].

现代慢速推理展现出与快速推理显著不同的输出结构，包括明确的探索与规划结构、频繁的验证与检查行为，以及通常更长的推理长度和时间。以往工作探索了构建长链推理输出的多样模式。一些方法——Macro-o1、HuatuoGPT-o1 与 AlphaZero——尝试通过结构化、主体性搜索合成长链思维 [186, 187, 188]。其他方法侧重于生成体现特定深思或反思思维模式的长链-CoT 数据集；示例包括 HiICL-MCTS、LLaVA-CoT、rStar-Math 与 ReasonFlux [189, 190, 191, 192]。近期在潜在空间中进行推理的方法利用潜在表示进行并行推理并探索多样推理轨迹 [193, 194]。随着预训练基础模型的进步，更新的工作转向自我改进范式——常借助强化学习——以进一步增强模型的推理能力 [181, 47]。

Prospective: Integrating Slow Reasoning Mechanisms into Agentic Reasoning The dichotomy between fast and slow reasoning highlights an open challenge in agentic reasoning: how to employ reinforcement learning for reliably training slow-thinking reasoning capabilities in agentic scenarios. Reinforcement learning in agentic scenarios faces greater challenges in training stability, such as ensuring compatibility with diverse environments. Agentic reasoning itself is also susceptible to overthinking problems. Purely fast models may overlook critical reasoning steps, while slow models often suffer from excessive latency or overthinking behaviors, such as unnecessarily long chains of thought. Emerging approaches seek hybrid strategies [195] that combine the efficiency of fast reasoning with the rigor of slow reasoning [196, 197, 198, 199]. For instance, adaptive test-time scaling allows a model to decide whether to respond quickly or to engage in extended deliberation depending on task complexity. Developing such cognitively-aligned mechanisms is a key step toward building reasoning agents that are both efficient and reliable.

展望: 将慢速推理机制整合进主体性推理快速与慢速推理的二分揭示了主体性推理中的一个开放挑战: 如何在主体性场景中使用强化学习可靠地训练慢思维推理能力。主体性场景下的强化学习在训练稳定性方面面临更大挑战，例如确保与多样环境的兼容性。主体性推理本身也易受过度思考问题的影响。纯快速模型可能忽略关键推理步骤，而慢速模型则常受制于过高延迟或过度思考行为，例如不必要的长思路链。新兴方法寻求混合策略 [195]，将快速推理的效率与慢速推理的严谨性结合 [196, 197, 198, 199]。例如，自适应测试时扩展允许模型根据任务复杂度决定是快速响应还是进行延长的深思。开发此类与认知对齐的机制是构建既高效又可靠的推理代理的重要一步。

3.6. Perception

3.6. 感知

By bridging visual perception with linguistic abstraction, Large Vision-Language Models (LVLMs) have demonstrated unprecedented capabilities for perceiving and understanding multimodal content [200, 201, 202, 203, 204, 205, 206, 207]. Central to this progress is the incorporation of explicit reasoning mechanisms into multimodal learning frameworks [208, 209], moving beyond passive perception toward active visual cognition [210]. RL has emerged as a powerful paradigm for this purpose, enabling the alignment of vision-language-action models with complex, multi-step reasoning objectives that go beyond the constraints of supervised next-token prediction [211, 212].

通过将视觉感知与语言抽象相结合，大型视觉-语言模型 (LVLMs) 在感知与理解多模态内容方面展现出前所未有的能力 [200, 201, 202, 203, 204, 205, 206, 207]。这一进展的核心在于将显式推理机制纳入多模态学习框架 [208, 209]，从被动感知迈向主动视觉认知 [210]。强化学习已成为实现这一目标的强大范式，使视觉-语言-行动模型与复杂的多步骤推理目标对齐，超越监督下一个词预测的限制 [211, 212]。

From Passive Perception to Active Visual Cognition Multimodal content often requires nuanced, context-dependent interpretation. Inspired by the remarkable success of RL in enhancing reasoning within LLMs [32, 213], researchers have increasingly sought to transfer these gains to multimodal learning [214, 215]. Early efforts focused on preference-based RL to strengthen the Chain-of-Thought (CoT) reasoning ability of MLLMs [216, 217, 218]. Visual-RFT [219] and Reason-RFT [220] directly apply GRPO to the vision domain, adaptively incorporating vision-specific metrics such as IoU as verifiable reward signals, while STAR-R1 [221] extended this idea by introducing partial rewards tailored for visual GRPO. Building upon this, a series of approaches—Vision-R1 [222], VLM-R1 [214], LMM-R1 [215], and MM-Eureka [223]—developed specialized policy optimization algorithms designed to incentivize step-wise visual reasoning, demonstrating strong performance even on smaller 3B-parameter models. SVQA-R1 [224] introduced Spatial-GRPO, a novel groupwise RL method that enforces view-consistent and transformation-invariant objectives. Visionary-R1 [225] enforces image captioning as a prerequisite step before reasoning, mitigating shortcut exploitation during reinforcement finetuning. A line of curriculum-learning methods have also been proposed to ease and smooth the RL training process of vision reinforcement finetuning [226, 227, 228, 229, 217]. R1-V [227] introduces VLM-Gym and trains G0/G1 models via scalable, pure RL self-evolution with a perception-enhanced cold start, yielding emergent perception-reasoning synergy across diverse visual tasks. R1-Zero [230] shows that even simple rule-based rewards can induce self-reflection and extended reasoning in non-SFT models, surpassing supervised baselines. PAPO [64] proposes a perception-aware policy optimization framework that augments RLVR methods with an implicit perception KL loss and double-entropy regularization, while [231] proposes a summarize-and-then-reason framework under RL training to mitigate visual hallucinations and improve reasoning without dense human annotations. Collectively, these approaches demonstrate that R1-style RL can be successfully transferred to the vision domain, provided that well-designed, verifiable reward metrics are used-yielding significant improvements in performance, robustness, and out-of-distribution generalization.

从被动感知到主动视觉认知多模态内容常常需要微妙且依赖上下文的解读。受到强化学习在提升大模型推理能力方面的显著成功启发 [32, 213]，研究者越来越多地尝试将这些收益迁移到多模态学习 [214, 215]。早期工作侧重于基于偏好的强化学习以增强 MLLMs 的链式思维 (CoT) 推理能力 [216, 217, 218]。Visual-RFT [219] 与 Reason-RFT [220] 将 GRPO 直接应用于视觉领域，自适应地引入 IoU 等视觉特有指标作为可验证的奖励信号，而 STAR-R1 [221] 通过引入为视觉 GRPO 量身定制的部分奖励扩展了这一思路。在此基础上，一系列方法——Vision-R1 [222]、VLM-R1 [214]、LMM-R1 [215] 和 MM-Eureka [223]——开发了专门的策略优化算法以激励逐步视觉推理，即便在较小的 3B 参数模型上也表现强劲。SVQA-R1 [224] 提出 Spatial-GRPO，一种新颖的分组强化学习方法，强制实现视图一致性与变换不变性目标。Visionary-R1 [225] 将图像描述作为推理前置步骤，从而在强化微调期间缓解捷径利用。一系列课程学习方法也被提出以缓解并平滑视觉强化微调的 RL 训练过程 [226, 227, 228, 229, 217]。R1-V [227] 引入 VLM-Gym，并通过具有感知增强冷启动的可扩展纯 RL 自我演化训练 G0/G1 模型，在多样视觉任务中产生感知—推理协同的涌现。R1-Zero [230] 表明即便是简单的基于规则的奖励也能在非 SFT 模型中诱导自我反思和延伸推理，超越监督基线。PAPO [64] 提出一种感知感知的策略优化框架，通过隐式感知 KL 损失与双熵正则化增强 RLVR 方法，而 [231] 在 RL 训练下提出总结再推理的框架，以在无需大量人工标注的情况下缓解视觉幻觉并改进推理。总体而言，这些方法表明，R1 风格的 RL 在采用设计良好、可验证的奖励度量的前提下可以成功迁移到视觉领域，从而显著提升性能、鲁棒性与分布外泛化能力。

More recent work explores another key advantage of RL: moving beyond the formulation of tasks as passive perception, where static, verifiable rewards are computed only on the text-based outputs of LVLMs. Instead, RL can be used to incentivize active cognition over multimodal content—treating visual representations as manipulable and verifiable intermediate thoughts. This paradigm empowers models not merely to ”look and answer,” but to actively see, manipulate, and reason with visual information as part of a multi-step cognitive process [210].

较新的工作探索了 RL 的另一个关键优势: 突破将任务表述为被动感知的范式——仅在 LVLM 的文本输出上计算静态可验证奖励。相反，RL 可用于激励对多模态内容的主动认知——将视觉表征视为可操作且可验证的中间思维。这一范式使模型不仅仅“看然后答”，而是在多步认知过程中主动观察、操作并基于视觉信息进行推理 [210]。

Grounding-Driven Active Perception. To advance from passive perception to active visual cognition, a key direction is enabling LVLMs to repeatedly look back and query the image while generating their reasoning process. This is achieved through grounding [232, 233], which anchors each step of the generated chain-of-thought (CoT) to specific regions of the multimodal input-facilitating more valid and verifiable reasoning by explicitly linking text with corresponding visual regions.

> 以落地点驱动的主动感知。要从被动感知迈向主动视觉认知，一个关键方向是使 LVLM 能在生成推理过程中反复回视并查询图像。通过 grounding [232, 233] 实现这一点，即将生成的链式思维 (CoT) 的每一步锚定到多模态输入的特定区域——通过将文本明确链接到相应视觉区域，促进更有效且可验证的推理。

To begin with, GRIT [234] interleaves bounding-box tokens with textual CoT and uses GRPO with both verifiable rewards and bounding-box correctness as supervision. [235] introduces a simple point-and-copy mechanism that allows the model to dynamically retrieve relevant image regions throughout the reasoning process. Ground-R1 [236] and BRPO [237] highlight evidence regions (via IoU-based or reflection rewards) prior to text-only reasoning, while DeepEyes [97] demonstrates that end-to-end RL can naturally induce such grounding behaviors. Chain-of-Focus further refines this approach by grounding CoT steps followed by zooming in operations.

> 首先，GRIT [234] 在文本 CoT 中交错插入边界框标记，并使用包含可验证奖励与边界框正确性监督的 GRPO。[235] 引入了一种简单的点选并复制机制，允许模型在推理过程中动态检索相关图像区域。Ground-R1 [236] 与 BRPO [237] 在纯文本推理之前突出证据区域 (通过基于 IoU 或反射的奖励)，而 DeepEyes [97] 表明端到端 RL 自然能诱导出此类落地点行为。Chain-of-Focus 进一步完善了这一方法，通过在落地点 CoT 步骤后执行放大操作。

Tool-Driven Active Perception. Another promising direction for enabling active perception is to frame visual cognition as an agentic process, where external tools, code snippets, and runtime environments assist the model’s cognitive workflow [238, 239]. For instance, VisTA [240] and VTool-R1 [241] focus on teaching models how to select and use visual tools through RL, while OpenThinkIMG [242] provides standardized infrastructure for training models to ”think with images.” Finally, Visual-ARFT [219] leverages RL to facilitate tool creation, harnessing the code-generation capabilities of MLLMs to dynamically extend their perceptual toolkit. Pixel Reasoner [98] expands the model’s action space with operations such as crop, erase, and paint, and introduces curiosity-driven rewards to discourage premature termination of exploration.

> 工具驱动的主动感知。另一条有前景的主动感知路径是将视觉认知构建为主体性过程，外部工具、代码片段与运行时环境辅助模型的认知工作流 [238, 239]。例如，VisTA [240] 与 VTool-R1 [241] 专注于通过 RL 教会模型如何选择并使用视觉工具，而 OpenThinkIMG [242] 提供了用于训练模型“以图像思考”的标准化基础设施。最后，Visual-ARFT [219] 利用 RL 促进工具创建，利用 MLLMs 的代码生成能力动态扩展其感知工具箱。Pixel Reasoner [98] 通过裁剪、擦除与绘制等操作扩展模型的动作空间，并引入好奇心驱动的奖励以防止过早终止探索。

Generation-Driven Active Perception. In addition to grounding and tool use, humans employ one of their most powerful cognitive abilities-imagination-to produce sketches or diagrams that aid problem-solving. Inspired by this, researchers have begun equipping LVLMs with the ability to generate sketches or images interleaved with chain-of-thought (CoT) reasoning, enabling models to externalize intermediate representations and reason more effectively [243, 244, 245]. Visual Planning [243] proposes to use imagined image rollouts only as the CoT images thinking, using downstream task success as the reward signal. GoT-R1 [246] applies RL within the Generation-CoT framework, allowing models to autonomously discover semantic-spatial reasoning plans before producing the image. Similarly, T2I-R1 [247] explicitly decouples the process into a semantic-level CoT for high-level planning and a token-level CoT for patch-wise pixel generation, jointly optimizing both stages with RL.

Generation-Driven Active Perception. 除了接地与工具使用，人类还利用其最强大的认知能力之一——想象力——来生成有助于问题解决的草图或示意图。受此启发，研究者开始赋予 LVLMs 生成与思维链 (CoT) 交错的草图或图像的能力，使模型能够外化中间表征并更有效地推理 [243, 244, 245]。Visual Planning [243] 提出仅将想象的图像回放作为 CoT 图像思考，将下游任务成功作为奖励信号。GoT-R1 [246] 在 Generation-CoT 框架内应用 RL，允许模型在生成图像前自主发现语义-空间推理计划。类似地，T2I-R1 [247] 将过程明确解耦为用于高层规划的语义级 CoT 和用于分块像素生成的令牌级 CoT，并以 RL 联合优化两个阶段。

Audio. RL has also been extended beyond vision-language models to a diverse range of modalities, including audio. Within the audio-language domain, we categorize RL applications into two broad classes. (1) Reasoning enhancement for large audio-language models: RL is leveraged to guide models in producing structured, step-by-step reasoning chains for tasks such as audio question answering and logical inference [248, 249, 250, 250, 248]. (2) Fine-grained component optimization in speech synthesis (TTS): RL is employed to directly refine system components—for example, improving duration predictors—using perceptual quality metrics such as speaker similarity and word error rate as reward signals, thereby yielding more natural and intelligible speech [251]. Some other works such as EchoInk-R1 [252] further enrich visual reasoning by integrating audio-visual synchrony under GRPO optimization.

Audio. RL 也已从视觉-语言模型扩展到包括音频在内的多种模态。在音频-语言领域，我们将 RL 应用划分为两大类。(1) 用于大型音频-语言模型的推理增强: 利用 RL 引导模型生成结构化的逐步推理链，用于音频问答和逻辑推断等任务 [248, 249, 250, 250, 248]。(2) 语音合成 (TTS) 中对细粒度组件的优化: 使用 RL 直接精炼系统组件，例如改进时长预测器，采用说话者相似度和词错误率等感知质量指标作为奖励信号，从而产生更自然、更易懂的语音 [251]。另有工作如 EchoInk-R1 [252] 在 GRPO 优化下通过整合视听同步进一步丰富视觉推理。

3.7. Others

3.7. Others

Beyond optimizing the above core cognitive modules, agentic RL also strengthens the ability to maintain strategic coherence over extended, multi-turn interactions. Here, RL is applied to support long-horizon reasoning and effective credit assignment.

除了优化上述核心认知模块外，代理式 RL 还增强了在长期多轮交互中保持策略一致性的能力。这里，RL 被用于支持长时程推理和有效的信用分配。

For long-horizon interactions, the central challenge is temporal credit assignment [118], where sparse and delayed feedback obscures the link between an agent’s actions and a distant outcome. Agentic RL directly confronts this by evolving both the learning signal and the optimization framework. One major approach is the (I) integration of process-based supervision with final outcome rewards. Rather than relying on a single reward at a trajectory's conclusion, this paradigm uses auxiliary models or programmatic rules to evaluate the quality of intermediate steps, providing a denser and more immediate learning signal that guides the agent's multi-turn strategy. For example, EPO [253], ThinkRM [254], SPO [255], and AgentPRM [256] introduce external reward models to provide step-wise signals for agents; in contrast, RIVMR [257] designs manually defined, programmatic rules to guide the intermediate supervision. A second, complementary strategy is to (II) extend preference optimization from single turns to multi-step segments. Techniques like Segment-level DPO (SDPO) [258] move beyond comparing isolated responses and instead construct preference data over entire conversational snippets or action sequences. This allows the model to directly learn how early decisions influence long-term success, thereby refining its ability to maintain strategic coherence in extended dialogues and complex tasks.

对于长时程交互，核心挑战是时间信用分配 [118]，稀疏且延迟的反馈使代理的动作与远期结果之间的关联变得模糊。代理式 RL 通过演进学习信号和优化框架直接应对这一点。一种主要方法是 (I) 将基于过程的监督与最终结果奖励整合。该范式不依赖于轨迹结束时的单一奖励，而是使用辅助模型或程序化规则评估中间步骤的质量，提供更稠密且即时的学习信号以指导代理的多轮策略。例如 EPO [253]、ThinkRM [254]、SPO [255] 和 AgentPRM [256] 引入外部奖励模型以提供逐步信号；相对地，RIVMR [257] 设计了手工定义的程序化规则来引导中间监督。第二种互补策略是 (II) 将偏好优化从单轮扩展到多步片段。像 Segment-level DPO (SDPO) [258] 的技术超越了对孤立回复的比较，而是构建关于整段对话片段或动作序列的偏好数据。这使模型能够直接学习早期决策如何影响长期成功，从而提升其在扩展对话和复杂任务中保持策略一致性的能力。

4. Agentic RL: The Task Perspective

4. Agentic RL: The Task Perspective

Agentic RL manifests through a wide spectrum of concrete tasks that test and shape its evolving capabilities. This section surveys representative application domains where Agentic RL has demonstrated remarkable potential and unique challenges. We begin with search and information retrieval (Section 4.1), followed by code generation and software engineering (Section 4.2), and mathematical reasoning (Section 4.3). We then discuss its role in GUI navigation (Section 4.4), vision understanding tasks (Section 4.5) as well as VLM embodied interaction (Section 4.6). Beyond single-agent scenarios, we extend the perspective to multi-agent systems (Section 4.7) and conclude with other emerging domains (Section 4.8). Together, these applications highlight how agentic RL transitions from abstract paradigms into actionable, real-world problem solving, as illustrated in Figure 6.

代理式 RL 通过一系列具体任务展现并塑造其不断演进的能力。本节综述了代理式 RL 展示出显著潜力与独特挑战的代表性应用领域。我们从搜索与信息检索 (第 4.1 节) 开始，接着讨论代码生成与软件工程 (第 4.2 节)、数学推理 (第 4.3 节)，然后论及 GUI 导航 (第 4.4 节)、视觉理解任务 (第 4.5 节) 以及 VLM 具身交互 (第 4.6 节)。超越单代理场景，我们将视角扩展到多代理系统 (第 4.7 节)，并在第 4.8 节总结其他新兴领域。综上，这些应用展示了代理式 RL 如何从抽象范式转向可执行的现实问题解决，如图 6 所示。

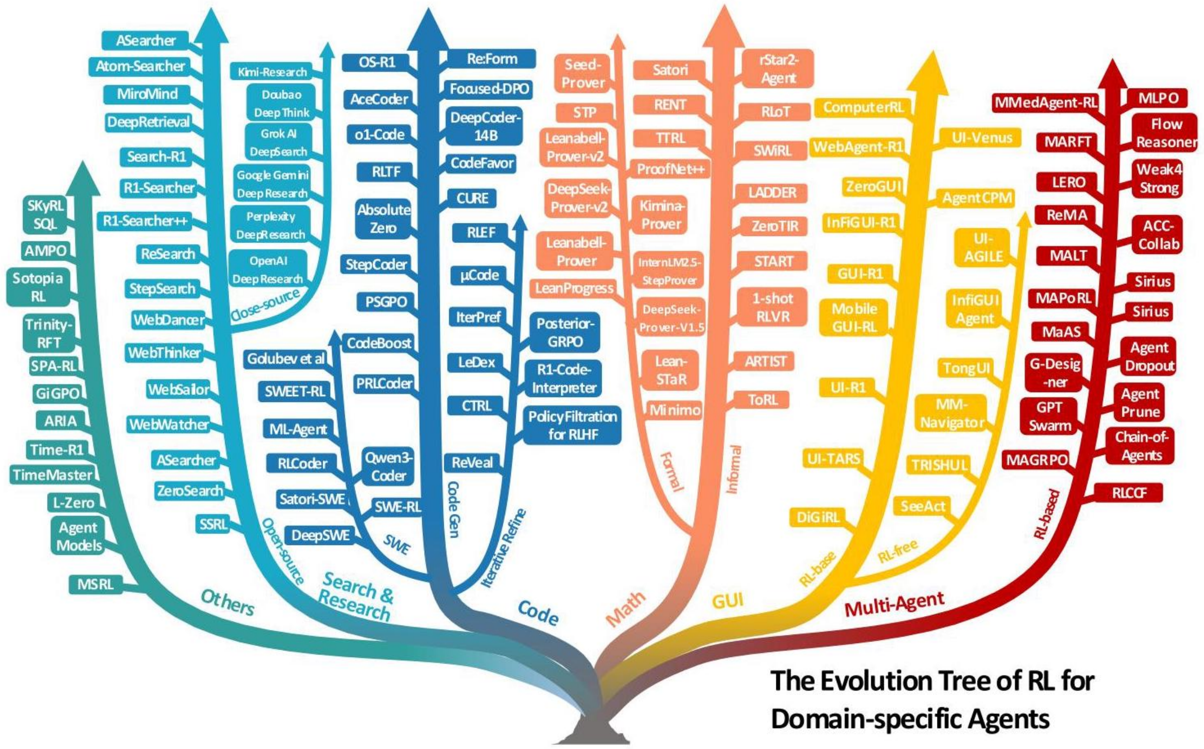

Figure 6: The evolution tree of RL for domain-specific agents.

Figure 6: The evolution tree of RL for domain-specific agents.

4.1. Search & Research Agent

4.1. Search & Research Agent

Search has been central to extending LLMs with external knowledge, with Retrieval-Augmented Generation (RAG) as a widely used approach [259, 260]. The paradigm is now evolving beyond simple information retrieval towards creating autonomous agents capable of deep research: complex, multi-step processes that involve not just finding information but also performing in-depth analysis, synthesizing insights from numerous sources, and drafting comprehensive reports [112, 261]. This shift elevates the objective from answering queries to tackling complex research tasks. Early prompt-driven methods relied on brittle query strategies and manual engineering. While more recent works like Search-o1 [107] leverage large reasoning models for agentic, inference-time retrieval, and multi-agent systems such as DeepResearch [262] coordinate querying and summarization sub-agents, they still lack learning signals. These prompt-based methods lack any fine-tuning signal, leading to limited generalization and poor effectiveness in multi-turn settings that demand a tight loop of search, reasoning, and synthesis. These limitations have led to the adoption of reinforcement learning to directly optimize the end-to-end process of query generation and search-reasoning coordination for advanced research objectives. Table 4 presents the majority of works studied in this section. In the following, we will detail how RL empowers these agents.

搜索一直是利用外部知识扩展大语言模型 (LLMs) 的核心，检索增强生成 (RAG) 是一种广泛使用的方法 [259, 260]。这一范式如今正从简单的信息检索向创建能够进行深度研究的自主智能体发展: 这是复杂的多步骤过程，不仅涉及查找信息，还包括进行深入分析、综合来自众多来源的见解以及撰写全面的报告 [112, 261]。这种转变将目标从回答查询提升到解决复杂的研究任务。早期基于提示的方法依赖脆弱的查询策略和手动设计。虽然像 Search-o1 [107] 这样的近期研究利用大型推理模型进行智能体式的推理时检索，以及像 DeepResearch [262] 这样的多智能体系统协调查询和摘要子智能体，但它们仍然缺乏学习信号。这些基于提示的方法缺乏任何微调信号，导致泛化能力有限，并且在需要搜索、推理和综合紧密循环的多轮场景中效果不佳。这些局限性促使人们采用强化学习来直接优化查询生成以及搜索 - 推理协调的端到端过程，以实现高级研究目标。表 4 展示了本节研究的大部分工作。接下来，我们将详细介绍强化学习如何赋能这些智能体。

Table 4: A summary of RL-based methods for search and research agents.

表 4: 基于强化学习的检索与研究代理方法综述。

Method	Category	Base LLM	Resource Link
Open Source Methods			
DeepRetrieval [263]	External	Qwen2.5-3B-Instruct, Llama-3.2-3B-Instruct	GitHub
Search-R1 [264]	External	Qwen2.5-3B/7B-Base/Instruct	GitHub
R1-Searcher [265]	External	Qwen2.5-7B, Llama3.1-8B-Instruct	GitHub
R1-Searcher++ [266]	External	Qwen2.5-7B-Instruct	GitHub
ReSearch [108]	External	Qwen2.5-7B/32B-Instruct	GitHub
StepSearch [267]	External	Qwen2.5-3B/7B-Base/Instruct	GiHHub
DeepResearcher [268]	External	Qwen2.5-7B-Instruct	GiHHub
WebDancer [106]	External	Qwen2.5-7B/32B, QWQ-32B	GitHub
WebThinker [269]	External	QwQ-32B, DeepSeek-R1-Distilled-Qwen, Qwen2.5-32B	GitHub
WebSailor [105]	External	Qwen2.5-3B/7B/32B/72B	GitHub
WebWatcher [270]	External	Qwen2.5-VL-7B/32B	C GitHub
WebShaper [271]	External	Qwen-2.5-32B/72B, QwQ-32B	GitHub
ASearcher [117]	External	Qwen2.5-7B/14B, QwQ-32B	GiHHub
Atom-Searcher [272]	External	Qwen2.5-7B-Instruct	GiHHub
MiroMind Open Deep Research [273]	External	-	#Website
SimpleDeepResearcher [274]	External	QwQ-32B	⍰GitHub
AWorld [275]	External	Qwen3-32B	GitHub
SFR-DeepResearch [276]	External	QwQ-32B, Qwen3-8B, GPT-oss-20b	-
ZeroSearch [277]	Internal	Owen2.5-3B/7B-Base/Instruct	GitHub
SSRL [278]	Internal	Qwen2.5, Llama-3.2/Llama-3.1, Qwen3	GitHub
Closed Source Methods			
OpenAI Deep Research [111]	External	OpenAI Models	#Website
Perplexity's DeepResearch [261]	External	-	#Website
Google Gemini's DeepResearch [279]	External	Gemini	#Website
Kimi-Researcher [112]	External	Kimi K2	#Website
Grok AI DeepSearch [280]	External	Grok3	excessive
Doubao with Deep Think [281]	External	Doubao	Website
Manus WideResearch	External	-	Website

方法	类别	基础 LLM	资源链接
开源方法			
DeepRetrieval [263]	外部	Qwen2.5-3B-Instruct, Llama-3.2-3B-Instruct	GitHub
Search-R1 [264]	外部	Qwen2.5-3B/7B-Base/Instruct	GitHub
R1-Searcher [265]	外部	Qwen2.5-7B, Llama3.1-8B-Instruct	GitHub
R1-Searcher++ [266]	外部	Qwen2.5-7B-Instruct	GitHub
ReSearch [108]	外部	Qwen2.5-7B/32B-Instruct	GitHub
StepSearch [267]	外部	Qwen2.5-3B/7B-Base/Instruct	GiHhub
DeepResearcher [268]	外部	Qwen2.5-7B-Instruct	GiHhub
WebDancer [106]	外部	Qwen2.5-7B/32B, QWQ-32B	GitHub
WebThinker [269]	外部	QwQ-32B, DeepSeek-R1-Distilled-Qwen, Qwen2.5-32B	GitHub
WebSailor [105]	外部	Qwen2.5-3B/7B/32B/72B	GitHub
WebWatcher [270]	外部	Qwen2.5-VL-7B/32B	C GitHub
WebShaper [271]	外部	Qwen-2.5-32B/72B, QwQ-32B	GitHub
ASearcher [117]	外部	Qwen2.5-7B/14B, QwQ-32B	GiHhub
Atom-Searcher [272]	外部	Qwen2.5-7B-Instruct	GiHhub
MiroMind Open Deep Research [273]	外部	-	# 网站
SimpleDeepResearcher [274]	外部	QwQ-32B	⍰GitHub
AWorld [275]	外部	Qwen3-32B	GitHub
SFR-DeepResearch [276]	外部	QwQ-32B, Qwen3-8B, GPT-oss-20b	-
ZeroSearch [277]	内部	Owen2.5-3B/7B-Base/Instruct	GitHub
SSRL [278]	内部	Qwen2.5, Llama-3.2/Llama-3.1, Qwen3	GitHub
闭源方法			
OpenAI Deep Research [111]	外部	OpenAI 模型	# 网站
Perplexity 的 DeepResearch [261]	外部	-	# 网站
Google Gemini 的 DeepResearch [279]	外部	Gemini	# 网站
Kimi-Researcher [112]	外部	Kimi K2	# 网站
Grok AI DeepSearch [280]	外部	Grok3	过度
Doubao with Deep Think [281]	外部	Doubao	网站
Manus WideResearch	外部	-	网站

4.1.1. Open Source RL Methods

4.1.1. 开源强化学习方法

Search from the external Internet A major line of work builds on the RAG foundation but relies on real-time web search APIs as the external environment, using reinforcement learning to optimize query generation and multi-step reasoning. Early progress was spearheaded by DeepRetrieval [263], which framed one-shot query generation as a GRPO-trained policy and directly rewarded recall and relevance against live search results. Motivated by its gains, subsequent methods extended the paradigm into multi-turn, reasoning-integrated, and multi-modal search. Search-R1 [264] and DeepResearcher [268] integrates retrieved-token masking with outcome-based rewards to interleave query formulation and answer generation. AutoRefine [282] further advances this trajectory by inserting refinement phases between successive search calls, using GRPO to reward not only answer correctness but also retrieval quality, enabling agents to iteratively filter and structure noisy evidence during long-horizon reasoning. R1-Searcher [265] employs a two-stage, cold-start PPO strategy-first learning when to invoke web search, then how to exploit it—while its successor R1-Searcher++ [266] adds supervised fine-tuning, internal-knowledge rewards to avoid redundancy, and dynamic memory for continual assimilation. ReSearch [108] pursues fully end-to-end PPO without supervised tool-use trajectories, while StepSearch [267] accelerates convergence on multi-hop QA by assigning intermediate step-level rewards. Atom-Searcher [272] is an agentic deep research framework that significantly improves LLM problem-solving by refining the reasoning process itself, not just the final outcome. WebDancer [106] leverages human browsing trajectory supervision plus RL fine-tuning to produce autonomous ReAct-style agents, excelling on GAIA [283] and WebWalkerQA [284]. WebThinker [269] embeds a Deep Web Explorer into a think-search-draft loop, aligning via DPO with human feedback to tackle complex report-generation. WebSailor [105] is a complete post-training methodology designed to teach LLM agents sophisticated reasoning for complex web navigation and information-seeking tasks. WebWatcher [270] further extends to multimodal search, combining visual-language reasoning, tool use, and RL to outperform text-only and multimodal baselines on BrowseComp-VL and VQA benchmarks. ASearcher [117] uses large-scale asynchronous reinforcement learning with synthesized QA data, enabling long-horizon search (40+ tool calls) and outperforming prior open-source methods. MiroMind Open Deep Research (MiroMind ODR) [273] is to build a high-performance, fully open-sourced, open-collaborative deep research ecosystem - with an agent framework, model, data, and training infra all fully accessible and open.

从外部互联网搜索一大类工作基于 RAG 框架，但把实时网页搜索 API 作为外部环境，使用强化学习优化查询生成与多步推理。早期进展由 DeepRetrieval [263] 引领，将一次性查询生成视为用 GRPO 训练的策略，并直接以实时搜索结果的召回率和相关性作为奖励。受其成效启发，后续方法将范式扩展到多轮、推理融合及多模态搜索。Search-R1 [264] 和 DeepResearcher [268] 将检索到的 token 屏蔽与基于结果的奖励结合，交替进行查询制定与答案生成。AutoRefine [282] 在此基础上进一步推进，通过在连续搜索调用之间插入细化阶段，使用 GRPO 不仅奖励答案正确性也奖励检索质量，使代理在长时序推理中能够迭代过滤和结构化噪声证据。R1-Searcher [265] 采用两阶段冷启动 PPO 策略——先学习何时调用网页搜索，再学习如何利用——其后继 R1-Searcher++ [266] 增加了监督微调、避免冗余的内在知识奖励和用于持续同化的动态记忆。ReSearch [108] 追求完全端到端的 PPO，无需监督工具使用轨迹，而 StepSearch [267] 通过给中间步骤级别的奖励加速多跳问答的收敛。Atom-Searcher [272] 是一个智能化深度研究框架，通过优化推理过程本身而非仅结果，显著提升 LLM 的问题解决能力。WebDancer [106] 利用人类浏览轨迹监督加 RL 微调，产生自主的 ReAct 风格代理，在 GAIA [283] 和 WebWalkerQA [284] 上表现优异。WebThinker [269] 将深度网页探索器嵌入思考-搜索-草拟循环，通过 DPO 与人类反馈对齐，以应对复杂报告生成。WebSailor [105] 是一套完整的后训练方法学，旨在教会 LLM 代理复杂网页导航和信息搜寻任务的高级推理。WebWatcher [270] 进一步扩展到多模态搜索，结合视觉-语言推理、工具使用与 RL，在 BrowseComp-VL 和 VQA 基准上超越文本或多模态基线。ASearcher [117] 使用大规模异步强化学习和合成 QA 数据，实现长时序搜索 (40+ 工具调用)，优于以往开源方法。MiroMind Open Deep Research (MiroMind ODR) [273] 致力于构建一个高性能、完全开源且开放协作的深度研究生态——包括代理框架、模型、数据和训练基础设施全部可访问且开源。

Search from LLM internal knowledge However, these training methods that rely on external APIs face two major challenges: (1) the document quality of real-time internet document search is uncontrolled, and noisy information brings instability to the training process. (2) The API cost is too high and severely limits scalability. To enhance the efficiency, controllability and stability of training, some recent studies have used controllable simulated search engines such as LLM internal knowledge. For example, ZeroSearch [277] replaces live web retrieval with a pseudo search engine distilled from LLMs themselves, combining curriculum RL to gradually approach live-engine performance without issuing real queries. SSRL [278] takes this idea further: the agent performs entirely offline ”self-search” during training, without explicit search engines, yet transfers seamlessly to online inference, where real APIs can still boost performance. Though still at an early stage, offline self-search enhances stability and scalability beyond API limits, pointing toward more self-reliant research agents.

从 LLM 内部知识搜索然而，依赖外部 API 的这些训练方法面临两大挑战:(1) 实时互联网文档搜索的文档质量不可控，噪声信息会使训练不稳定；(2) API 成本过高，严重限制可扩展性。为提升训练的效率、可控性和稳定性，近期一些研究使用可控的模拟搜索引擎，例如 LLM 内部知识。例如，ZeroSearch [277] 用从 LLM 本身蒸馏出的伪搜索引擎替代实时网页检索，结合课程化强化学习逐步接近真实引擎性能而不发出真实查询。SSRL [278] 将这一想法推进一步: 代理在训练期间完全离线地执行“自我搜索”，无需显式搜索引擎，但能无缝迁移到在线推理，在那里真实 API 仍可提升性能。尽管仍处于早期阶段，离线自我搜索在稳定性和可扩展性上超越 API 限制，指向更自给自足的研究代理。

4.1.2. Closed Source RL Methods

4.1.2. 闭源强化学习方法

Despite progress in combining RAG and RL, most open source models still fail on OpenAI's BrowseComp [285], a challenging benchmark that measures the ability of AI agents to locate hard-to-find information, revealing gaps in long-horizon planning, page-grounded tool use, and cross-source verification. In contrast, recent closed source systems are markedly stronger, having shifted from mere query optimization to fully autonomous research agents that navigate the open web, synthesize information from multiple sources, and draft comprehensive reports. This is likely due to the industry's more powerful foundation models and the availability of more high-quality data. OpenAI Deep Research [111] achieves 51.5% pass@1 on BrowseComp. Other prototypes, Perplexity's DeepResearch [261], Google Gemini's DeepResearch [279], Kimi-Researcher [112], Grok AI DeepSearch [280], Doubao with Deep Think [281], combine RL-style fine-tuning with advanced tool integration and memory modules, ushering in a new era of interactive, iterative research assistants.

尽管在结合 RAG 与 RL 上取得进展，大多数开源模型在 OpenAI 的 BrowseComp [285] 上仍然失败——该基准衡量 AI 代理定位难觅信息的能力，暴露出长时序规划、基于页面的工具使用和跨源验证的短板。相比之下, 近期闭源系统明显更强, 已从单纯的查询优化转向完全自主的研究代理, 能够在开放网络中导航、从多源合成信息并起草全面报告。这很可能归因于产业界更强的基础模型和更多高质量数据的可用性。OpenAI Deep Research [111] 在 BrowseComp 上达成 51.5% 的 pass@1。其它原型如 Perplexity 的 DeepResearch [261]、Google Gemini 的 DeepResearch [279]、Kimi-Researcher [112]、Grok AI DeepSearch [280]、豆包的 Deep Think [281]，将 RL 风格微调与先进的工具集成和记忆模块结合，开启了交互式、迭代式研究助手的新纪元。

4.2. Code Agent

4.2. 代码代理

Code generation, or more broadly, software engineering, provides an ideal testbed for LLM-based agentic RL: execution semantics are explicit and verifiable, and automated signals (compilation, unit tests, and runtime traces) are readily available [286]. Early multi-agent frameworks (e.g., MetaGPT, AutoGPT, AgentVerse) coordinated roles through prompting without parameter updates, showcasing the promise of modular role allocation [287, 288, 289]. Initial RL for code, such as CodeRL, incorporated execution-based reward modeling and actor-critic training [290], catalyzing a wave of studies that exploit execution feedback to guide policy updates. Table 5 presents the majority of works studied in this section. We structure the literature along increasing task complexity, progressing from code generation (Section 4.2.1) to code refinement (Section 4.2.2) and software engineering (Section 4.2.3).

代码生成，或更广义的软件工程，为基于大模型的代理式强化学习提供了理想试验场: 执行语义是明确且可验证的，且自动化信号 (编译、单元测试与运行时轨迹) 易于获取 [286]。早期多代理框架 (如 MetaGPT、AutoGPT、AgentVerse) 通过提示协调角色而不更新参数，展示了模块化角色分配的潜力 [287, 288, 289]。最初的代码 RL 工作 (如 CodeRL) 引入了基于执行的奖励建模与 actor-critic 训练 [290]，催生了一系列利用执行反馈来指导策略更新的研究。表 5 列出了本节所述的大部分工作。我们按任务复杂度递增组织文献，从代码生成 (第 4.2.1 节) 到代码迭代改进 (第 4.2.2 节) 再到软件工程 (第 4.2.3 节)。

4.2.1.RL for Code Generation

4.2.1.RL for Code Generation

Early research focused on relatively simple, single-round code generation (e.g., completing a function or or solving a coding challenge in one go), which lays the foundation for subsequent large-scale software engineering.

早期研究集中在相对简单的单轮代码生成 (例如一次性补全函数或解决一道编程题)，为后续的大规模软件工程奠定了基础。

Outcome reward RL. Methods in this class optimize directly for final correctness, typically measured by pass@k or unit-test success. AceCoder [291] introduces a data-efficient RLHF pipeline for code generation, constructing large-scale preference pairs from existing code fragments to train a reward model via Bradley-Terry loss, which then guides RFT on the synthesized dataset. Beyond early actor-critic formulations, recent open-source efforts scale outcome-based RL on large pre-trained code models. DeepCoder-14B [292] stabilizes GRPO training via iterative context lengthening and DAPO-inspired filtering, and employs a sparse Outcome Reward Model (ORM) to prevent reward hacking on curated coding data. RLTF employs an online RL loop that uses unit test results as multi-granularity reward signals, from coarse pass/fail outcomes to fine-grained fault localization, directly guiding code refinement [293]. CURE formalizes coder-tester co-evolution: a tester generates or evolves unit tests while a coder iteratively patches code; a reward-precision objective mitigates low-quality test effects during joint training [294]. Absolute Zero applies self-play RL without human data. It generates coding tasks for itself and uses execution outcomes as verifiable rewards to bootstrap reasoning ability [163]. Re:Form [295] leverages formal language-based reasoning with RL and automated verification to reduce human priors, enabling reliable program synthesis and surpassing strong baselines on formal verification tasks. In [296], authors propose a two-stage training pipeline: first fine-tuning for a high-correctness baseline, then perform efficiency-driven online RL optimization.

结果奖励类 RL。本类方法直接优化最终正确性，通常以 pass@k 或单元测试通过率衡量。AceCoder [291] 提出了一种用于代码生成的数据高效 RLHF 流程，从现有代码片段构建大规模偏好对，用 Bradley-Terry 损失训练奖励模型，再在合成数据集上指导 RFT。除早期的 actor-critic 形式外，近期开源工作在大规模预训练代码模型上扩大了基于结果的 RL。DeepCoder-14B [292] 通过迭代上下文延长与受 DAPO 启发的过滤稳定 GRPO 训练，并采用稀疏的结果奖励模型 (ORM) 以防在精心挑选的代码数据上发生奖励劫持。RLTF 使用在线 RL 循环，将单元测试结果作为多粒度奖励信号，从粗粒度的通过/失败到细粒度的故障定位，直接指导代码修正 [293]。CURE 将编码者-测试者协同进化形式化: 测试者生成或演化单元测试，编码者迭代修补代码；奖励精度目标在联合训练中缓解低质量测试的影响 [294]。Absolute Zero 在没有人工数据的情况下应用自我博弈 RL: 它为自身生成编码任务并将执行结果作为可验证奖励以引导推理能力的引导 [163]。Re:Form [295] 将基于形式语言的推理与 RL 和自动化验证结合，以减少人工先验，实现可靠的程序合成并在形式验证任务上超过强基线。在 [296] 中，作者提出了两阶段训练管线: 先微调以获得高正确性的基线，然后进行以效率为导向的在线 RL 优化。

Process reward RL. To mitigate sparsity and credit assignment, several works design process-level supervision by integrating compilation and execution feedback. StepCoder [297] decomposes compilation and execution into step-level signals for shaping; Process Supervision-Guided Policy Optimization (PSGPO) [41] leverages intermediate error traces and process annotations for dense rewards; and CodeBoost [298] mines raw repositories to unify heterogeneous execution-derived signals, ranging from output correctness to error-message quality, under a single PPO framework. Further, PRLCoder [299] introduces process-supervised RL by constructing reward models that score each partial snippet: a teacher model mutates lines of reference solutions and assigns positive/negative signals based on compiler and test feedback. This fine-grained supervision yields faster convergence and +10.5% pass-rate improvements over the base model, illustrating how dense shaping at the line-level can guide code synthesis more effectively than outcome-only signals. o1-Coder [300] combines RL with Monte Carlo Tree Search, where the policy learns from exploration guided by test case rewards and gradually improves from pseudocode to executable code. Posterior-GRPO [301] rewards intermediate reasoning but gates credit by final test success to prevent speculative reward exploitation; Policy Filtration for RLHF [39] improves reward-correctness alignment by filtering low-confidence pairs before policy updates. Scaling preference supervision beyond costly human annotation has proven effective as well. CodeFavor [302] constructs CodePrefBench from code evolution histories, covering correctness, efficiency, security, and style to improve preference modeling and alignment. Focused-DPO [303] adapts preference-based RL by weighting preference optimization on error-prone regions of the code, making feedback more targeted and improving robustness across benchmarks. [304] studies how RL-trained small-scale agents surpass large-scale prompt-based models in MLE environments via duration-aware gradient updates in a distributed asynchronous RL.

过程奖励类 RL。为缓解稀疏性与归因问题，多项工作设计了通过整合编译与执行反馈的过程级监督。StepCoder [297] 将编译与执行分解为用于形塑的步骤级信号；Process Supervision-Guided Policy Optimization (PSGPO) [41] 利用中间错误轨迹与过程注释提供密集奖励；CodeBoost [298] 从原始仓库挖掘异构的执行来源信号，从输出正确性到错误信息质量，在单一 PPO 框架下统一处理。此外，PRLCoder [299] 通过构建对每个部分代码片段打分的奖励模型引入了过程监督式 RL: 教师模型变异参考解的行并基于编译与测试反馈分配正负信号。这种细粒度监督带来更快的收敛并比基线模型提高 +10.5% 的通过率，说明行级密集形塑比仅有结果信号更能有效引导代码合成。o1-Coder [300] 将 RL 与蒙特卡洛树搜索结合，策略从由测试用例奖励引导的探索中学习，逐步从伪代码改进为可执行代码。Posterior-GRPO [301] 奖励中间推理但以最终测试成功为门控以防止投机性奖励利用；Policy Filtration for RLHF [39] 通过在策略更新前过滤低置信度对来改善奖励—正确性的对齐。将偏好监督扩展到超越昂贵人工标注的规模也被证明有效。CodeFavor [302] 从代码演化历史构建 CodePrefBench，涵盖正确性、效率、安全与风格，以改进偏好建模与对齐。Focused-DPO [303] 通过在代码易错区域加权偏好优化来调整基于偏好的 RL，使反馈更有针对性并提升在基准上的稳健性。[304] 研究了 RL 训练的小规模代理如何通过在分布式异步 RL 中的时长感知梯度更新在 MLE 环境中超越基于提示的大模型。

4.2.2.RL for Iterative Code Refinement

4.2.2.RL for Iterative Code Refinement

A second line of research targets more complex coding tasks that require debugging and iterative refinement. In these scenarios, an agent may need multiple attempts to improve solutions, using feedback from human requirements or failed test results, which is closer to real-world tasks.

第二条研究线路针对更复杂的编码任务，这类任务需要调试与迭代改进。在这些场景中，代理可能需要多次尝试来改进解法，利用来自人工需求或失败测试结果的反馈，这更贴近现实任务。

Outcome reward RL. A representative line treats the entire refinement loop as a trajectory while optimizing for final task success. RLEF [305] (Reinforcement Learning from Execution Feedback) grounds correction loops in real error messages as context while optimizing for ultimate pass rates; this reduces the number of attempts needed and improves competitive-programming performance relative to single-shot baselines. μ Code [306] jointly trains a generator and a learned verifier under single-step reward feedback, showing that verifier-guided outcome rewards can outperform purely execution-feedback baselines. R1-Code-Interpreter [307] harnesses multi-turn supervised fine-tuning and reinforcement learning to train LLMs to decide when and how to invoke a code interpreter during step-by-step reasoning.

结果奖励强化学习。代表性做法将整个细化循环视为一条轨迹，同时优化最终任务成功率。RLEF [305](从执行反馈中强化学习) 将纠正循环以真实错误信息作为上下文并以最终通过率为优化目标；这减少了尝试次数并相较于一次性基线提升了竞技编程表现。μ Code [306] 在单步奖励反馈下联合训练生成器与学习到的验证器，展示了验证器引导的结果奖励可以优于纯执行反馈基线。R1-Code-Interpreter [307] 结合多轮监督微调与强化学习，训练大模型在逐步推理中判断何时以及如何调用代码解释器。

Process reward RL. Process-supervised approaches explicitly guide how the model debugs. IterPref [308] constructs localized preference pairs from iterative debugging traces and applies targeted preference optimization to penalize faulty regions, improving correction accuracy with minimal collateral updates. LeDex [309] couples explanation-driven diagnosis with self-repair: it automatically curates explanation-refinement trajectories and applies dense, continuous rewards to jointly optimize explanation quality and code correctness via PPO, yielding consistent pass@1 gains over SFT-only coders. Beyond explanation-driven shaping, some works like CTRL [310] explicitly train separate critic models to evaluate each attempted refinement and provide gradient signals to the policy, though at the cost of added inference overhead. ReVeal [311] extends process-level refinement into a self-evolving agent that autonomously generates tests and learns from per-turn rewards to enhance reasoning and recovery from errors.

过程奖励强化学习。过程监督方法明确指导模型如何调试。IterPref [308] 从迭代调试轨迹构建局部化偏好对并应用定向偏好优化以惩罚有缺陷的区域，在最小的附带更新下提升修正准确性。LeDex [309] 将基于解释的诊断与自我修复结合: 它自动策划解释—精炼轨迹并对解释质量与代码正确性通过 PPO 应用密集、连续的奖励，共同优化两者，在仅有 SFT 的编码器上持续带来 pass@1 提升。除了解释驱动的塑形外，像 CTRL [310] 的工作会显式训练独立的评论模型来评估每次尝试的精炼并向策略提供梯度信号，代价是增加推理开销。ReVeal [311] 将过程级细化扩展为自我进化代理，能自主生成测试并从每回合奖励中学习，以增强推理与错误恢复能力。

4.2.3.RL for Automated Software Engineering

4.2.3. 面向自动化软件工程的强化学习

Outcome reward RL. End-to-end training in realistic environments demonstrates that sparse-but validated-success signals can scale. DeepSWE performs large-scale RL on software engineering missions using verified task completion as the sole reward, achieving leading open-source results on SWE-bench-style evaluations [312]. SWE-RL extracts rule-based, outcome-oriented signals from GitHub commit histories, enabling training on authentic improvement patterns and generalization to unseen bug-fixing tasks [313]. Satori-SWE introduces an evolutionary RL-enabled test-time scaling method (EvoScale) that trains models to self-improve generations across iterations for sample-efficient software engineering tasks [314]. OS-R1 [317] presents a rule-based reinforcement learning framework for Linux kernel tuning, enabling efficient exploration, accurate configuration, and superior performance over heuristic methods. RLCoder frames retrieval-augmented repository-level code completion as an RL problem, using perplexity-based feedback to train a retriever to fetch helpful context without labeled data [315]. Qwen3-Coder performs large-scale execution-driven reinforcement learning on long-horizon, multi-turn interactions across 20,000 parallel environments, yielding state-of-the-art performance on benchmarks like SWE-Bench Verified [195]. In machine learning domains, ML-Agent executes multi-step pipelines (e.g., automated ML), optimizing performance-based terminal rewards [316].

结果奖励强化学习。在真实环境中的端到端训练表明稀疏但经验证的成功信号可以扩展。DeepSWE 在以已验证任务完成作为唯一奖励的大规模软件工程任务上执行强化学习，在 SWE-bench 风格评测上取得领先的开源结果 [312]。SWE-RL 从 GitHub 提交历史中提取基于规则的结果导向信号，使得可在真实改进模式上训练并泛化到未见的修复任务 [313]。Satori-SWE 引入一种进化的 RL 支持的测试时扩展方法 (EvoScale)，训练模型在多次迭代中自我改进生成以实现样本高效的软件工程任务 [314]。OS-R1 [317] 提出面向 Linux 内核调优的基于规则的强化学习框架，实现高效探索、精确配置并优于启发式方法的性能。RLCoder 将检索增强的仓库级代码补全视作 RL 问题，使用基于困惑度的反馈训练检索器在无标注数据下获取有用上下文 [315]。Qwen3-Coder 在 20,000 个并行环境上对长时序多轮交互执行大规模执行驱动强化学习，在如 SWE-Bench Verified [195] 的基准上取得最先进表现。在机器学习领域，ML-Agent 执行多步流水线 (如自动化 ML)，以基于性能的终端奖励进行优化 [316]。

Process reward RL. Dense supervision over agentic trajectories improves credit assignment across many steps. From the optimization perspective, long-context, multi-turn software agents benefit from stabilized policy-gradient variants; e.g., Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) improves training stability and performance on SWE-bench Verified through multi-turn code generation and debugging interactions, leveraging long-context feedback [318]. SWEET-RL trains multi-turn agents on ColBench (backend and frontend tasks), leveraging privileged information during RL to reduce exploration noise and improve long-horizon generalization [159].

过程奖励强化学习。对代理轨迹的密集监督改善了多步的归因问题。从优化角度看，长上下文、多轮软件代理受益于稳定化的策略梯度变体；例如 Decoupled Clip 与动态采样策略优化 (DAPO) 通过多轮代码生成与调试交互利用长上下文反馈，提升了在 SWE-bench Verified 上的训练稳定性与性能 [318]。SWEET-RL 在 ColBench(后端与前端任务) 上训练多轮代理，在强化学习期间利用特权信息以减少探索噪声并改善长时序泛化 [159]。

Table 5: A summary of RL methods for code and software engineering agents.

表 5: 代码与软件工程代理的强化学习方法汇总。

Method	Reward	Base LLM	Resource
RL for Code Generation			
AceCoder [291]	Outcome	Qwen2.5-Coder-7B-Base/Instruct, Qwen2.5-7B-Instruct	{fitHub
DeepCoder-14B [292]	Outcome	Deepseek-R1-Distilled-Qwen-14B	C GitHub
RLTF [293]	Outcome	CodeGen-NL 2.7B, CodeT5	GitHub
CURE [294]	Outcome	Qwen2.5-7B/14B-Instruct, Qwen3-4B	GiHtHub
Absolute Zero [163]	Outcome	Qwen2.5-7B/14B, Qwen2.5-Coder-3B/7B/14B, Llama-3.1-8B	GitHub
StepCoder [297]	Process	DeepSeek-Coder-Instruct-6.7B	GitHub
PSGPO [41]	Process	-	-
CodeBoost [298]	Process	Qwen2.5-Coder-7B-Instruct, Llama-3.1-8B-Instruct, Seed-Coder- 8B-Instruct, Yi-Coder-9B-Chat	OGitHub
PRLCoder [299]	Process	CodeT5+, Unixcoder, T5-base	-
o1-Coder [300]	Process	DeepSeek-1.3B-Instruct	GitHub
Posterior-GRPO [301]	Process	Qwen2.5-Coder-3B/7B-Base, Qwen2.5-Math-7B	-
Policy Filtration for RLHF [39]	Process	DeepSeek-Coder-6.7B, Qwen1.5-7B	GitHub
CodeFavor [302]	Process	Mistral-NeMo-12B-Instruct, Gemma-2-9B-Instruct, Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.3,	GitHub
Focused-DPO [303]	Process	DeepSeek-Coder-6.7B-Base/Instruct, Qwen2.5-Coder-7B-Instruct Magicoder-S-DS-6.7B,	-
Re:Form [295]	Outcome	Qwen-2.5, 0.5B-14B	GitHub
Qwen Team [296]	Outcome	Qwen-2.5-Coder-Instruct-7B/32B	-
RL for Iterative Code Refinement			
RLEF [305]	Outcome	Llama-3.0-8B-Instruct, Llama-3.1-8B/70B-Instruct	-
μ Code [306]	Outcome	Llama-3.2-1B/8B-Instruct	GitHub
R1-Code-Interpreter [307]	Outcome	Qwen2.5-7B/14B-Instruct-1M, Qwen2.5-3B-Instruct	OGitHub
IterPref [308]	Process	Deepseek-Coder-7B-Instruct, Qwen2.5-Coder-7B, StarCoder2-15B	-
LeDex [309]	Process	StarCoder-15B, CodeLlama-7B/13B	-
CTRL [310]	Process	Owen2.5-Coder-7B/14B/32B-Instruct	Gi 在 Hub
ReVeal [311]	Process	DAPO-Qwen-32B	-
RL for Automated Software Engineering (SWE)			
DeepSWE [312]	Outcome	Qwen3-32B	GitHub
SWE-RL [313]	Outcome	Llama-3.3-70B-Instruct	GitHub
Satori-SWE [314]	Outcome	Owen-2.5-Math-7B	OGitHub
RLCoder [315]	Outcome	CodeLlama7B, StartCoder-7B, StarCoder2-7B, DeepSeekCoder- 1B/7B	Gi 比 Hub
Qwen3-Coder [195]	Outcome	-	Gi 在 Hub
ML-Agent [316]	Outcome	Qwen2.5-7B-Base/Instruct, DeepSeek-R1-Distill-Qwen-7B	Gi 在 Hub
OS-R1 [317]	Outcome	Owen2.5-3B/7B-Instruct	GitHub
Golubev et al. [318]	Process	Owen2.5-72B-Instruct	-
SWEET-RL [159]	Process	Llama-3.1-8B/70B-Instruct	GitHub

方法	奖励	基础 LLM	资源
用于代码生成的强化学习			
AceCoder [291]	结果	Qwen2.5-Coder-7B-Base/Instruct, Qwen2.5-7B-Instruct	{fitHub
DeepCoder-14B [292]	结果	Deepseek-R1-Distilled-Qwen-14B	C GitHub
RLTF [293]	结果	CodeGen-NL 2.7B, CodeT5	GitHub
CURE [294]	结果	Qwen2.5-7B/14B-Instruct, Qwen3-4B	GiHtHub
Absolute Zero [163]	结果	Qwen2.5-7B/14B, Qwen2.5-Coder-3B/7B/14B, Llama-3.1-8B	GitHub
StepCoder [297]	过程	DeepSeek-Coder-Instruct-6.7B	GitHub
PSGPO [41]	过程	-	-
CodeBoost [298]	过程	Qwen2.5-Coder-7B-Instruct, Llama-3.1-8B-Instruct, Seed-Coder- 8B-Instruct, Yi-Coder-9B-Chat	OGitHub
PRLCoder [299]	过程	CodeT5+, Unixcoder, T5-base	-
o1-Coder [300]	过程	DeepSeek-1.3B-Instruct	GitHub
Posterior-GRPO [301]	过程	Qwen2.5-Coder-3B/7B-Base, Qwen2.5-Math-7B	-
RLHF 的策略过滤 [39]	过程	DeepSeek-Coder-6.7B, Qwen1.5-7B	GitHub
CodeFavor [302]	过程	Mistral-NeMo-12B-Instruct, Gemma-2-9B-Instruct, Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.3,	GitHub
Focused-DPO [303]	过程	DeepSeek-Coder-6.7B-Base/Instruct, Qwen2.5-Coder-7B-Instruct Magicoder-S-DS-6.7B,	-
Re:Form [295]	结果	Qwen-2.5, 0.5B-14B	GitHub
Qwen Team [296]	结果	Qwen-2.5-Coder-Instruct-7B/32B	-
用于迭代代码改进的强化学习			
RLEF [305]	结果	Llama-3.0-8B-Instruct, Llama-3.1-8B/70B-Instruct	-
μ 代码 [306]	结果	Llama-3.2-1B/8B-Instruct	GitHub
R1-Code-Interpreter [307]	结果	Qwen2.5-7B/14B-Instruct-1M, Qwen2.5-3B-Instruct	OGitHub
IterPref [308]	过程	Deepseek-Coder-7B-Instruct, Qwen2.5-Coder-7B, StarCoder2-15B	-
LeDex [309]	过程	StarCoder-15B, CodeLlama-7B/13B	-
CTRL [310]	过程	Owen2.5-Coder-7B/14B/32B-Instruct	Gi 在 Hub
ReVeal [311]	过程	DAPO-Qwen-32B	-
用于自动化软件工程 (SWE) 的强化学习			
DeepSWE [312]	结果	Qwen3-32B	GitHub
SWE-RL [313]	结果	Llama-3.3-70B-Instruct	GitHub
Satori-SWE [314]	结果	Owen-2.5-Math-7B	OGitHub
RLCoder [315]	结果	CodeLlama7B, StartCoder-7B, StarCoder2-7B, DeepSeekCoder- 1B/7B	Gi 比 Hub
Qwen3-Coder [195]	结果	-	Gi 在 Hub
ML-Agent [316]	结果	Qwen2.5-7B-Base/Instruct, DeepSeek-R1-Distill-Qwen-7B	Gi 在 Hub
OS-R1 [317]	结果	Owen2.5-3B/7B-Instruct	GitHub
Golubev et al. [318]	过程	Owen2.5-72B-Instruct	-
SWEET-RL [159]	过程	Llama-3.1-8B/70B-Instruct	GitHub

Remark on closed-source systems. Commercial systems such as OpenAI's Codex and Anthropic's Claude Code have emphasized preference-aligned fine-tuning and reinforcement learning to improve usefulness and safety in code generation and editing workflows [319, 320]. While concrete training details are limited publicly, these systems underscore the growing role of RL in aligning agentic behavior with developer-centric objectives in practical IDE and terminal environments.

关于闭源系统的说明。像 OpenAI 的 Codex 和 Anthropic 的 Claude Code 这样的商业系统，强调通过偏好对齐微调与强化学习，来提升代码生成和编辑工作流程的实用性与安全性 [319, 320]。尽管具体的训练细节并未公开，但这些系统突显了强化学习在实际集成开发环境 (IDE) 和终端环境中，使智能体行为与以开发者为中心的目标保持一致方面日益重要的作用。

4.3. Mathematical Agent

4.3. 数学代理

Mathematical reasoning is widely regarded as a gold standard for assessing LLM agents' reasoning ability, owing to its symbolic abstraction, logical consistency, and long-horizon deductive demands. We structure the research efforts around two complementary paradigms: informal reasoning (Section 4.3.1), which operates without formal verification support and includes natural-language reasoning and programming-language tool use; and formal reasoning (Section 4.3.2), which relies on rigorously specified formal languages and proof assistants.

数学推理被广泛视为评估大模型 (LLM) 推理能力的金标准，原因在于其符号抽象、逻辑一致性以及长程演绎要求。我们将相关研究工作围绕两个互补范式结构化: 非形式化推理 (第 4.3.1 节)，在无形式化验证支持下进行，包含自然语言推理与编程语言工具使用；以及形式化推理 (第 4.3.2 节)，依赖严格指定的形式语言与证题助手。

We note that RLVR methods such as DAPO [47], GRPO [321], and GRESO [55] have consistently played a substantial role in recent enhancements of mathematical reasoning in LLMs. However, given their broader relevance across reasoning tasks, we discuss them in Section 2.7, instead of elaborating here.

我们注意到诸如 DAPO [47]、GRPO [321] 与 GRESO [55] 的 RLVR 方法在近期提升 LLM 数学推理能力方面持续发挥了重要作用。不过，鉴于它们在更广泛推理任务中的相关性，我们在第 2.7 节讨论这些方法，而非在此详述。

4.3.1.RL for Informal Mathematical Reasoning

4.3.1. 用于非形式数学推理的强化学习

Informal mathematics essentially refers to reasoning and expression in natural language. Such reasoning may incorporate symbols or function names, but no finite and explicit set of logical rules defines their syntactic validity, and no formal semantics precisely determines their interpretation and meaning [322, 323].

非正式数学本质上是指用自然语言进行推理和表达。这种推理可能会包含符号或函数名，但没有有限且明确的逻辑规则集来定义其句法有效性，也没有形式语义来精确确定其解释和含义 [322, 323]。

While informal mathematical reasoning relaxes strict rigor at the detail level, it affords greater expressive flexibility and better captures the high-level structure of arguments. This makes it particularly suited for a variety of math tasks such as mathematical word problem solving, equation manipulation, and symbolic computation [100, 324]. Although general-purpose programming languages are symbolic, they lack the rigor and formal semantics of proof-assistant languages, and are therefore regarded as informal when applied to mathematical reasoning [322], typically through tool invocation of executors such as Python with numerical or symbolic libraries.

虽然非正式数学推理在细节层面放宽了严格的严谨性，但它提供了更大的表达灵活性，能更好地捕捉论证的高层结构。这使其特别适合各种数学任务，如解决数学应用题、方程处理和符号计算 [100, 324]。尽管通用编程语言是符号化的，但它们缺乏证明辅助语言的严谨性和形式语义，因此在应用于数学推理时被视为非正式的 [322]，通常是通过调用带有数值或符号库的 Python 等执行器工具来实现。

Outcome reward RL. Outcome-only methods define rewards solely by final numerical or symbolic correctness (e.g. algebraic equations) during RL. Empirically, such training often leads to emergent agentic behaviors, including adaptive tool use interleaved with natural language reasoning. ARTIST [100] introduces a framework for tool-integrated agentic reasoning, interleaving tool invocations, e.g. code execution, directly within the reasoning chain. Trained with outcome-only rewards, it achieves strong performance and observes emergent agentic behaviors, including self-reflection, and context-aware CoT, which further shows that by integrating dynamic tool use with RL, agentic tool-integrated reasoning could learn optimal strategies for interacting with environments, highlighting the potential of RL to internalize tool-integrated reasoning strategies in LLMs. Similarly, ToRL [101] improves performance by exploiting the scaling of tool-integrated reasoning RL and encouraging code execution behaviour, and experiments show emergent cognitive behaviors, such as adaptive tool-use, self-correction based on tool feedback, and adaptive computational reasoning. ZeroTIR [324] investigates the scaling law of RL from outcome-based rewards for Tool-Integrated Reasoning with Python code execution settings, revealing a strong correlation between training computational effort and the spontaneous code execution frequency, the average response length, and the final task accuracy, which corroborates the empirical emergence of tool-integrated reasoning strategies. TTRL [165] leverages majority voting to estimate rewards, enabling training on unlabeled data. Fine-tuned on these majority-vote rewards, it not only surpasses the base model's maj@n accuracy but also achieves an empirical performance curve and upper bound that, surprisingly, closely approach those of direct RL training with labeled test answers on MATH-500, underscoring its practical value and potential. However, RENT [325] suggests that majority voting is short on generalization, it applies only to questions with deterministic answers, and will not work on free-response outputs. To address this limitation, it extends the entropy minimization idea [326] to RL, using the token-level average negative entropy as a reward to guide learning, achieving improvements on an extensive suite of benchmarks including math problem solving, suggesting that confidence-based reward shaping can serve as a path toward continual improvement. Alternatively, Satori [314] proposes Chain-of-Action-Thought (COAT), a variant of CoT that explicitly integrates action choices, and modularizes reasoning into 3-fold meta-actions, including continuation, reflection, and exploration of alternatives, and internalizes this behavior via RL with outcome-only rewards. In particular, 1-shot RLVR [327] studies data efficiency of outcome-only RL with verifier signals. Surprisingly, they found that RL with only 1 example performs close to using a 1.2k-example dataset, and with 2 examples comes close to using the 7.5k MATH training dataset. They also highlight an intriguing phenomenon, named post-saturation generalization, that test accuracy continues to improve even after the training accuracy on the single example approaches 100%. In addition to correctness, hallucination remains a major challenge in informal mathematical reasoning, motivating methods that explicitly promote trustworthiness. For instance, Kirchner et al. [328] propose a game-theoretic training algorithm that jointly optimizes for both correctness and legibility. Inspired by Prover-Verifier Games [329], the method alternates between training a small verifier that predicts solution correctness, a ”helpful” prover that generates solutions accepted by the verifier, and a ”sneaky” prover that aims to fool it. Empirically, this increases the helpful prover accuracy, verifier robustness and legibility (measured by human accuracy in time-constraint verification tasks). This result suggests that verifier-guided legibility optimization can enhance the interpretability and trustworthiness of LLM-generated informal reasoning. Recent rStar2-Agent [115] is a 14B-parameter math reasoning model trained with agentic reinforcement learning using a high-throughput Python execution environment, a novel GRPO-RoC algorithm to resample on correct rollouts amid tool-noise, and a multi-stage training recipe-achieving state-of-the-art results in just 510 RL steps, achieving average pass@1 scores of 80.6% on AIME24 and 69.8% on AIME25.

结果导向奖励的强化学习。仅以最终数值或符号正确性 (例如代数方程) 来定义奖励的方法在强化学习中被称为仅结果方法。实证上，此类训练常导致自主性行为的涌现，包括与自然语言推理交错的自适应工具使用。ARTIST [100] 提出一个工具整合的自主推理框架，将工具调用 (如代码执行) 直接嵌入推理链中。以仅结果奖励训练时，它取得了强性能并观察到自主性行为的涌现，包括自我反思与具上下文感知的链式思考，进一步表明通过将动态工具使用与强化学习结合，工具整合的自主推理可学得与环境交互的最优策略，凸显强化学习将工具整合推理策略内化到大模型中的潜力。类似地，ToRL [101] 通过利用工具整合推理强化学习的扩展性并鼓励代码执行行为来提升性能，实验证明了涌现的认知行为，如自适应工具使用、基于工具反馈的自我纠正和自适应计算推理。ZeroTIR [324] 在带有 Python 代码执行的工具整合推理设置中研究了来自结果型奖励的强化学习的规模规律，揭示了训练计算投入与自发代码执行频率、平均响应长度和最终任务准确率之间的强相关性，从而佐证了工具整合推理策略的经验性涌现。TTRL [165] 利用多数投票来估计奖励，使得可在无标签数据上训练。基于这些多数投票奖励微调后，它不仅超越了基模型的 maj@n 准确率，还出乎意料地在 MATH-500 上获得了接近带标签测试答案的直接强化学习训练的经验性能曲线与上界，凸显其实用价值与潜力。然而，RENT [325] 指出多数投票在泛化上不足，仅适用于答案确定性的问题，不适用于自由文本输出。为了解决这一限制，它将熵最小化的思想 [326] 扩展到强化学习，使用逐标记的平均负熵作为奖励来指导学习，在包括数学题求解在内的大量基准上取得改进，表明基于置信度的奖励塑形可作为持续改进的一条路径。另一种方法 Satori [314] 提出行动思维链 (Chain-of-Action-Thought, COAT)，这是 CoT 的一个变体，显式整合行动选择，并将推理模块化为三类元行动: 继续、反思和探索替代方案，通过以仅结果奖励的强化学习将该行为内化。特别地，1-shot RLVR [327] 研究了带验证器信号的仅结果强化学习的数据效率。令人惊讶的是，他们发现仅用 1 个示例的强化学习表现接近使用 1.2k 示例的数据集，使用 2 个示例的表现接近使用 7.5k 的 MATH 训练集。他们还强调了一个有趣现象，称为过饱和后泛化 (post-saturation generalization)，即即便单个示例的训练准确率接近 100%，测试准确率仍会继续提升。除了正确性之外，幻觉在非正式数学推理中仍是主要挑战，这推动了明确促进可信性的做法。例如，Kirchner 等人 [328] 提出一种博弈论训练算法，联合优化正确性与可读性。受证明者-验证者博弈 (Prover-Verifier Games)[329] 的启发，该方法交替训练一个预测解答正确性的微型验证器、一个生成被验证器接受解答的“有帮助”证明者和一个旨在欺骗它的“狡猾”证明者。实证上，这提高了有帮助证明者的准确率、验证器的鲁棒性和可读性 (以受限时间内人工验证任务中的人工准确率衡量)。该结果表明，基于验证器的可读性优化可增强大模型生成的非正式推理的可解释性与可信度。近期的 rStar2-Agent [115] 是一个 14B 参数的数学推理模型，使用代理化强化学习在高吞吐 Python 执行环境中训练，提出一种在工具噪声中对正确回放重采样的新型 GRPO-RoC 算法，并采用多阶段训练流程——仅 510 步强化学习即达成最新成果，在 AIME24 上取得平均 pass@1 80.6%，在 AIME25 上取得 69.8%。

Process reward RL. Process-aware methods leverage intermediate evaluators (e.g. unit tests, assertions, sub-task checks) to provide denser feedback, shaping credit assignment and improving tool-integrated reasoning (TIR). START [330] guides TIR by injecting handcrafted hint text into Long CoT traces, typically after conjunction words or before the CoT stop token, to encourage code executor calls during inference. This enables test-time scaling that improves reasoning accuracy. The collected trajectories are then used to fine-tune the model, internalizing the tool-invocation behavior. LADDER [331] introduces a training-time framework where an LLM recursively generates and solves progressively simpler variants of a complex problem, using verifiable reward signals to guide a difficulty-based curriculum, and achieves substantial improvements in mathematical reasoning. An additional test-time RL step (TTRL) further enhances performance. The authors suggest that this approach of self-generated curriculum learning with verifiable feedback may generalize beyond informal mathematical tasks to any domain with reliable automatic verification. To improve performance on complex problems, SWiRL [332] synthesizes step-wise tool use reasoning data by iteratively decomposing solutions, and then adopts a preference-based step-wise RL approach to fine-tune the base model on the multi-step trajectories. While many of these approaches exploits inference-time interventions, they often suffer from generalization limitations due to their reliance on manually designed logical structures. To overcome this, RLoT [333] instead trains a lightweight navigator agent model with RL to adaptively enhance reasoning, showing improved generalization across diverse tasks.

处理奖励的强化学习。面向过程的方法利用中间评估器 (如单元测试、断言、子任务检查) 提供更密集的反馈，以塑造信任分配并改进与工具集成的推理 (TIR)。START [330] 通过在长链式推理 (Long CoT) 轨迹中注入手工提示文本 (通常在连接词之后或在 CoT 停止标记之前) 来引导 TIR，以在推理时鼓励调用代码执行器。这使得测试时扩展能够提高推理准确性。收集到的轨迹随后用于微调模型，使其内化工具调用行为。LADDER [331] 提出了一种训练时框架，LLM 递归生成并解决逐步简化的复杂问题变体，使用可验证的奖励信号来引导基于难度的课程，从而在数学推理上取得显著提升。一个额外的测试时 RL 步骤 (TTRL) 进一步增强了性能。作者指出，这种以可验证反馈自生成课程学习的方法可能不仅限于非正式的数学任务，而可推广到任何具有可靠自动验证的领域。为提升复杂问题的表现，SWiRL [332] 通过迭代分解解法来合成逐步使用工具的推理数据，然后采用基于偏好的逐步 RL 方法在多步轨迹上微调基础模型。尽管许多此类方法利用了推理时的干预，但由于依赖手工设计的逻辑结构，它们常面临泛化限制。为克服此点，RLoT [333] 改为用 RL 训练一个轻量的导航代理模型以自适应增强推理，展示了在多样任务上的改进泛化能力。

While informal approaches excel at word problems and symbolic computations, they struggle to extend effectively to advanced mathematical tasks such as automated theorem proving. This limitation arises from two fundamental challenges: evaluation difficulty, which demands machine-verifiable feedback unavailable to informal methods, and scarcity of high-quality formal proof data. [322, 323]

虽然非正式方法在文字题和符号计算上表现出色，但它们难以有效扩展到如自动定理证明等高级数学任务。此限制源于两个根本挑战: 评估困难，需要机器可验证的反馈，而非正式方法无法提供；以及高质量形式化证明数据的稀缺。[322, 323]

4.3.2.RL for Formal Mathematical Reasoning

4.3.2. 用于形式化数学推理的强化学习

Formal mathematical reasoning refers to reasoning carried out in a formal language with precisely defined syntax and semantics, yielding proof objects that are mechanically checkable by a verifier. This paradigm is particularly suited for advanced tasks such as automated theorem proving (ATP) [334], where an agent, given a statement (theorem, lemma, or proposition), must construct a proof object that the verifier accepts, thereby ensuring machine-verifiable correctness. From a reinforcement learning perspective, formal theorem proving is commonly modeled as a Markov Decision Process (MDP): proof states transition via the application of tactics [1] , each of which is treated as a discrete action in RL-based proof search. [335]. Under this formulation, formal theorem proving can be cast as a search problem over a vast, discrete, and parameterized action space.

形式化数学推理指以具有精确定义语法与语义的形式语言进行的推理，产生可被验证器机械检查的证明对象。这一范式特别适合诸如自动定理证明 (ATP)[334] 之类的高级任务：代理在给定命题 (定理、引理或命题) 的情况下，必须构建验证器接受的证明对象，从而确保机器可验证的正确性。从强化学习角度看，形式定理证明通常被建模为马尔可夫决策过程 (MDP)：证明状态通过应用战术 [1] 转移，每个战术被视为 RL 基于证明搜索中的离散动作 [335]。在此表述下，形式定理证明可被视为在巨大、离散且参数化的动作空间上的搜索问题。

Formal proofs are verified by proof assistants such as Lean, Isabelle, Coq, and HOL Light. These systems, often referred to as Interactive Theorem Provers (ITPs), deterministically accept or reject proof objects producing binary pass/fail signals as the primary reward for RL training, while some works also explore leveraging error messages as auxiliary signals [336, 337].

形式化证明由 Lean、Isabelle、Coq 和 HOL Light 等证明辅助工具验证。这些系统通常称为交互式定理证明器 (ITP)，以确定性地接受或拒绝证明对象，产生二元的通过/失败信号作为 RL 训练的主要奖励，同时一些工作也探索将错误消息作为辅助信号加以利用 [336, 337]。

Outcome reward RL. The outcome-only paradigm was demonstrated at scale in 2024 with DeepSeek-Prover-v1.5 [334], which releases an end-to-end RL pipeline in Lean based solely on binary verifier feedback, resulting in significant improvements in proof success on benchmarks like miniF2F [338] and ProofNet [339]. The authors propose a variant of MCTS, i.e. RMaxTS, that incorporates intrinsic rewards for discovering novel tactic states to encourage diversity of proof exploration during inference-time search and mitigate the sparse-reward issue. Building on this direction, Leanabell-Prover [340] scales up DeepSeek-Prover-v1.5 by aggregating an expansive hybrid dataset of statement-proof pairs and informal reasoning sketches from multiple sources and pipelines such as Mathlib4 [341], LeanWorkbook [342], NuminaMath [343], STP [344], etc., covering well over 20 mathematical domains. This broad coverage mitigates the scarcity of aligned informal-to-formal (NL to Lean4) training examples, which are crucial for bridging natural-language reasoning and formal proof generation. At the same time, Kimina-Prover [345] Preview further emphasizes the critical challenge of aligning informal and formal reasoning. It implements a structured ”formal reasoning pattern,” where natural-language reasoning and Lean 4 code snippets are interleaved within thinking blocks. To reinforce this alignment, the output is constrained—to include at least one tactic block and to reuse no less than 60% of the Lean 4 snippets in the final proof, ensuring close correspondence between internal reasoning and formal output. A recent work, Seed-Prover [346], integrates multiple techniques. It first adopts a lemma-centered proof paradigm, which enables systematic problem decomposition, cross-trajectory lemma reuse, and explicit progress tracking. It then enriches RL training with a diverse prompting strategy that randomly incorporates both informal and formal proofs, successful and failed lemmas, and Lean compiler feedback, thereby enhancing adaptability to varied inputs. At inference, it employs a conjecture-prover pipeline that interleaves proving conjectures into lemmas and generating new conjectures from the evolving lemma pool, substantially improving its capacity to tackle difficult problems. Complementarily, the accompanying Seed-Geometry system extends formal reasoning to geometry, providing state-of-the-art performance on Olympiad benchmarks. Together, these efforts demonstrate that sparse but explicit reward signals can yield nontrivial gains, particularly when paired with effective exploration strategies.

只依据结果的奖励 RL。仅基于结果的范式在 2024 年被 DeepSeek-Prover-v1.5 [334] 在大规模上验证，该工作发布了一个基于 Lean 的端到端 RL 流水线，仅依赖二元验证器反馈，在 miniF2F [338] 和 ProofNet [339] 等基准上大幅提升了证明成功率。作者提出了一种 MCTS 变体 RMaxTS，纳入了发现新策略状态的内在奖励以鼓励推理搜索期间的多样化探索并缓解稀疏奖励问题。在此方向上，Leanabell-Prover [340] 通过聚合来自 Mathlib4 [341]、LeanWorkbook [342]、NuminaMath [343]、STP [344] 等多源和多条管道的大规模混合语料库 (命题—证明对和非形式推理草图)，扩展了 DeepSeek-Prover-v1.5, 覆盖了 20 多个数学领域。这种广泛覆盖缓解了对齐的非形式到形式 (自然语言到 Lean4) 训练样本的匮乏，而这些样本对桥接自然语言推理与形式证明生成至关重要。与此同时，Kimina-Prover [345] Preview 进一步强调了对齐非形式与形式推理的关键挑战。它实现了一种结构化的“形式推理模式”，在思考块内交错放置自然语言推理和 Lean 4 代码片段。为加强这种对齐，输出被约束——至少包含一个 tactic 块，并在最终证明中重用不少于 60% 的 Lean 4 片段，以确保内部推理与形式输出之间的紧密对应。近期工作 Seed-Prover [346] 集成了多种技术。它首先采用以引理为中心的证明范式，从而实现系统性的问题分解、跨轨迹引理重用和显式进度跟踪；然后通过多样化的提示策略丰富 RL 训练，随机包含非形式和形式证明、成功与失败的引理以及 Lean 编译器反馈，从而增强对不同输入的适应性。推理时，它采用一个猜想—证明器流水线，将证明猜想纳入引理并从不断演化的引理池生成新猜想，大幅提升了解决困难问题的能力。配套的 Seed-Geometry 系统将形式推理扩展到几何，在奥赛基准上提供了最先进的性能。综上，这些工作表明稀疏但明确的奖励信号在结合有效的探索策略时能带来显著收益。

[1] In Lean-style Interactive Theorem Provers (ITPs), a tactic is a command or small script that instructs the system to refine the current proof goal, with the resulting proof term checked by the ITP kernel for correctness.

[1] 在 Lean 风格的交互式定理证明器 (ITP) 中，tactic 是一个命令或小脚本，用于指示系统细化当前证明目标，生成的证明项由 ITP 内核检查其正确性。

Process reward RL. To improve credit assignment and reduce wasted exploration, several works extend the outcome-only paradigm with denser, step-level signals. DeepSeek-Prover-v2 [347] designs a dual-model pipeline to unify both informal (natural-language) and formal (Lean4) mathematical reasoning to reinforce the proving reasoning ability. It introduces subgoal decomposition, where a prover model solves recursively decomposed subgoals and receives binary Lean feedback at the subgoal level, effectively providing denser supervision and improving both accuracy and interpretability. Following this dual-role collaborative mindset, ProofNet++ [336] implements a neuro-symbolic RL framework featuring a Symbolic Reasoning Interface, which maps LLM-generated reasoning into formal proof trees, and a Formal Verification Engine, which verifies these proofs with Lean or HOL Light and routes error feedback back to the LLM for self-correction. Leanabell-Prover-v2 [337] integrates verifier messages into reinforcement updates within a long CoT framework, enabling explicit verifier-aware self-monitoring that stabilizes tactic generation and reduces repeated failure patterns.

过程级奖励 RL。为改善信用分配并减少无效探索，若干工作在仅结果范式上引入更密集的步级信号。DeepSeek-Prover-v2 [347] 设计了一个双模型流水线，将非形式 (自然语言) 与形式 (Lean4) 数学推理统一以增强证明推理能力。它引入子目标分解，由证明器模型递归解决分解的子目标并在子目标层面接收二元 Lean 反馈，有效提供更密集的监督并提升准确性与可解释性。秉持这一双重角色协作思路，ProofNet++ [336] 实现了一个神经符号 RL 框架，包含将 LLM 生成的推理映射为形式证明树的符号推理接口，以及使用 Lean 或 HOL Light 验证这些证明并将错误反馈回 LLM 以便自我纠正的形式验证引擎。Leanabell-Prover-v2 [337] 在长链式思维框架中将验证器消息纳入强化更新，实现了显式的验证器感知自我监控，稳定了 tactic 生成并减少了重复失败模式。

Hybrid reward RL. Although both outcome-only and process-aware reward paradigms have demonstrated encouraging advances, the scarcity of high-quality theorem-proving data further amplifies the challenges of reinforcement learning under sparse rewards as well as the design of step-level preference signals [348, 349, 344]. To mitigate these limitations, a prominent line of work adopts expert iteration (ExIt) [350], a framework that combines search with policy learning. This paradigm provides an alternative to outcome-only or process-aware RL, alleviating data scarcity by producing high-quality supervised trajectories. Instead of directly optimizing against sparse verifier signals, ExIt performs search-guided data augmentation: valid proof trajectories discovered by search and checked by a verifier are reused as expert demonstrations in an imitation-learning loop. It usually employs a two-role system: the expert collects valid and progressive trajectories via MCTS under outcome-only verifier feedback, while the apprentice trains a policy on these process-level trajectories and then shares the improved policy back with the expert, thereby bootstrapping subsequent rounds of search and accelerating convergence. Polu and Sutskever [351] introduces ExIt into formal theorem proving, demonstrating that search-generated expert data can bootstrap models toward tackling complex multi-step proving challenges. Later works adapt this design to Lean and other ITPs.

混合奖励 RL。尽管仅结果和过程感知的奖励范式均取得了令人鼓舞的进展，但高质量定理证明数据的稀缺进一步加剧了在稀疏奖励下的 RL 挑战以及步级偏好信号的设计难题 [348, 349, 344]。为缓解这些限制，一条重要研究路线采用专家迭代 (ExIt)[350]，该框架将搜索与策略学习结合。该范式为仅结果或过程感知的 RL 提供了替代，通过产生高质量的监督轨迹缓解数据匮乏。ExIt 并非直接针对稀疏验证器信号优化，而是执行基于搜索的数据增强: 由搜索发现并由验证器检查的有效证明轨迹被重用为模仿学习循环中的专家示范。它通常采用双角色系统: 专家在仅结果的验证器反馈下通过 MCTS 收集有效且有进展的轨迹，而学徒在这些过程级轨迹上训练策略，随后将改进后的策略回馈给专家，从而引导后续轮次的搜索并加速收敛。Polu 和 Sutskever [351] 将 ExIt 引入形式定理证明，证明了搜索生成的专家数据可以引导模型应对复杂的多步证明挑战。后续工作将此设计适配到 Lean 与其他 ITP。

When applying to formal theorem proving, naive tree search methods often face severe search space explosion when navigating the vast parameterized tactic space. To mitigate this, InternLM2.5-StepProver [352] introduces a preference-based critic model, trained with RLHF-style optimization, to guide expert search, effectively providing a curriculum that directs exploration toward problems of suitable difficulty. Lean-STaR [353] further enhances ExIt by integrating Self-Taught Reasoner (STaR) [354]. It first trains a thought-augmented tactic predictor on synthesized (proof state, generated thought, ground-truth tactic) triples. Then, in the expert-iteration loop, the model produces trajectories that interleave thoughts with tactics; trajectories with tactics successfully validated by Lean are retained and reused for imitation learning. Empirically, the inclusion of thoughts increases the diversity of exploration in the sample-based proof search.

在将其应用于形式定理证明时，朴素的树搜索方法在遍历庞大的参数化策略空间时常面临严重的搜索空间爆炸。为缓解这一问题，InternLM2.5-StepProver [352] 引入了基于偏好的评论家模型，采用类似 RLHF 的优化进行训练，以指导专家搜索，有效提供一种课程化策略，将探索引导到合适难度的问题上。Lean-STaR [353] 通过整合 Self-Taught Reasoner (STaR) [354] 进一步增强了 ExIt。它首先在合成的 (证明状态、生成思路、真实策略) 三元组上训练一个带思路的策略预测器。然后，在专家迭代循环中，模型生成交替包含思路与策略的轨迹；由 Lean 成功验证的含策略轨迹会被保留并用于模仿学习。实验上，引入思路增加了基于样本的证明搜索中的探索多样性。

A recent work, STP [344], points out that solely relying on expert iteration will quickly plateau due to the sparse positive rewards. To address this, it extends the conjecturer-prover self-play idea from Minimo [355] to practical formal languages (Lean/Isabelle) with an open-ended action space and starts from a pretrained model. STP instantiates a dual-role loop in which a conjecturer proposes statements that are barely provable by the current prover, and a prover is trained with standard expert iteration; this generates an adaptive curriculum and alleviates sparse training signals. Empirically, STP reports large gains on LeanWorkbook [342] and reports competitive results among whole-proof generation methods on miniF2F [338] and ProofNet [339].

最近的工作 STP [344] 指出，单靠专家迭代会因正奖励稀疏而很快陷入瓶颈。为此，它将 Minimo [355] 的猜想者-证明者自对弈思想扩展到具有开放式动作空间的实用形式语言 (Lean/Isabelle) 并从预训练模型出发。STP 实现了一个双重角色循环: 猜想者提出当前证明者勉强可证明的命题，而证明者通过标准的专家迭代进行训练；这产生了自适应课程并缓解了稀疏训练信号。实证上，STP 在 LeanWorkbook [342] 上报告了显著提升，并在 miniF2F [338] 与 ProofNet [339] 的整段证明生成方法中报告了有竞争力的结果。

Table 6: A summary of RL methods for mathematical reasoning agents.

表 6: 面向数学推理代理的强化学习方法汇总。

Method	Reward	Resources
RL for Informal Mathematical Reasoning		
ARTIST [100]	Outcome	-
ToRL [101]	Outcome	C GitHub B HuggingFace
ZeroTIR [324]	Outcome	C GitHub HuggingFace
TTRL [165]	Outcome	GiHtHub
RENT [325]	Outcome	C GitHub #Website
Satori [314]	Outcome	Higfiy HuggingFace #Website
1-shot RLVR [327]	Outcome	GiftHub C HuggingFace
Prover-Verifier Games (legibility) [328]	Outcome	-
rStar2-Agent [115]	Outcome	GitHub
START [330]	Process	-
LADDER [331]	Process	-
SWiRL [332]	Process	-
RLoT [333]	Process	GitHub
RL for Formal Mathematical Reasoning		
DeepSeek-Prover-v1.5 [334]	Outcome	Higijhung HuggingFace
Leanabell-Prover [340]	Outcome	OGitHub CHUggingFace
Kimina-Prover (Preview) [345]	Outcome	GiHHub HuggingFace
Seed-Prover [346]	Outcome	C GitHub
DeepSeek-Prover-v2 [347]	Process	CHUHU
ProofNet++ [336]	Process	-
Leanabell-Prover-v2 [337]	Process	GitHub
InternLM2.5-StepProver [352]	Hybrid	GitHub
Lean-STaR [353]	Hybrid	GiHu HuggingFace # Website
STP [344]	Hybrid	GitHub 2 HuggingFace

方法	奖励	资源
用于非正式数学推理的强化学习		
ARTIST [100]	结果	-
ToRL [101]	结果	C GitHub B HuggingFace
ZeroTIR [324]	结果	C GitHub HuggingFace
TTRL [165]	结果	GiHtHub
RENT [325]	结果	C GitHub #Website
Satori [314]	结果	Higfiy HuggingFace #Website
1-shot RLVR [327]	结果	GiftHub C HuggingFace
证明者-验证者博弈 (可读性)[328]	结果	-
rStar2-Agent [115]	结果	GitHub
START [330]	过程	-
LADDER [331]	过程	-
SWiRL [332]	过程	-
RLoT [333]	过程	GitHub
用于形式化数学推理的强化学习		
DeepSeek-Prover-v1.5 [334]	结果	Higijhung HuggingFace
Leanabell-Prover [340]	结果	OGitHub CHUggingFace
Kimina-Prover (Preview) [345]	结果	GiHHub HuggingFace
Seed-Prover [346]	结果	C GitHub
DeepSeek-Prover-v2 [347]	过程	CHUHU
ProofNet++ [336]	过程	-
Leanabell-Prover-v2 [337]	过程	GitHub
InternLM2.5-StepProver [352]	混合	GitHub
Lean-STaR [353]	混合	GiHu HuggingFace # Website
STP [344]	混合	GitHub 2 HuggingFace

4.4.GUI Agent

4.4.GUI 代理

GUI agents have progressed through distinct training paradigms. Early systems used pre-trained vision-language models (VLMs) in a pure zero-shot fashion, mapping screenshots and prompts directly to single-step actions. Later, SFT on static (screen, action) trajectories improved grounding and reasoning, but was limited by scarce human operation traces. Reinforcement fine-tuning (RFT) reframes GUI interaction as sequential decision-making, allowing agents to learn via trial-and-error with sparse or shaped rewards, and has advanced from simple single-task settings to complex, real-world, long-horizon scenarios. Table 7 presents the majority of works studied in this section.

GUI 代理经历了不同的训练范式。早期系统以纯零样本方式使用预训练视觉-语言模型 (VLM)，将截图和提示直接映射为单步动作。随后，对静态 (屏幕，动作) 轨迹的监督微调 (SFT) 提高了定位和推理能力，但受限于稀缺的人类操作轨迹。强化微调 (RFT) 将 GUI 交互重新构架为序贯决策，使代理能够通过稀疏或塑形奖励的试错学习，已从简单的单任务设置发展到复杂的真实世界长时程场景。表 7 汇总了本节研究的大多数工作。

4.4.1. RL-free Methods

4.4.1. 无强化方法

Vanilla VLM-based GUI Agents Early GUI agents directly leveraged pre-trained Vision-Language Models (VLMs) in a pure zero-shot manner, mapping screenshots and prompts to single-step actions without any task-specific fine-tuning. Representative systems include MM-Navigator [356], SeeAct [357], and TRISHUL [358], which differ in interface domains or parsing strategies but share the same reliance on off-the-shelf VLMs. While showcasing the generality of foundation models, these approaches suffer from limited grounding accuracy and reliability, restricting their applicability to complex tasks [359, 360].

基于原生 VLM 的 GUI 代理早期 GUI 代理直接利用预训练视觉-语言模型 (VLM) 以纯零样本方式工作，将截图和提示映射为单步动作，无需任何任务特定微调。代表性系统包括 MM-Navigator [356]、SeeAct [357] 和 TRISHUL [358]，它们在界面领域或解析策略上有所不同，但共同依赖现成的 VLM。尽管展示了基础模型的通用性，这些方法在定位准确性和可靠性上有限，限制了其在复杂任务中的适用性 [359, 360]。

Supervised Fine-Tuning (SFT) with Static Trajectory Data The SFT paradigm adapts pre-trained vision-language models to GUI tasks by minimizing cross-entropy loss on offline (screen, action) pairs, without online interaction. InfiGUIAgent [361] employs a two-stage pipeline that first improves grounding and then incorporates hierarchical and reflective reasoning. UI-AGILE [362] enhances supervised fine-tuning by incorporating continuous rewards, simplified reasoning, and cropping-based resampling, while further proposing a decomposed grounding mechanism for handling high-resolution displays. TongUI [363] instead emphasizes data scale, constructing the 143K-trajectory GUI-Net from multimodal web tutorials to enhance generalization. While differing in focus, these approaches all face the limitation of scarce human operation traces.

使用静态轨迹数据的监督微调 (SFT) SFT 范式通过在离线 (屏幕，动作) 对上最小化交叉熵损失，将预训练视觉-语言模型适配到 GUI 任务，而无需在线交互。InfiGUIAgent [361] 采用两阶段流水线，先提升定位能力，再引入分层和反思式推理。UI-AGILE [362] 通过引入连续奖励、简化推理和基于裁剪的重采样来增强监督微调，同时提出分解定位机制以处理高分辨率显示。TongUI [363] 则强调数据规模，基于多模态网络教程构建了 143K 轨迹的 GUI-Net 以增强泛化。尽管侧重点不同，这些方法都面临人类操作轨迹稀缺的限制。

4.4.2.RL in Static GUI Environments

4.4.2. 静态 GUI 环境中的强化学习

In static settings, reinforcement learning is applied on pre-collected datasets with deterministic execution traces, using rule-based criteria for outcome evaluation in the absence of live environment interactions. GUI-R1 [364] adopts an R1-style reinforcement fine-tuning pipeline over a unified action schema, using simple format and correctness rewards to improve step-level action prediction with modest data. UI-R1 [365] applies group-relative policy optimization to stabilize policy updates and improve exact parameter matching through a compact action interface and reward shaping for action-type and argument accuracy. InFiGUI-R1 [366] introduces a two-stage training paradigm that first distills spatial reasoning to enhance grounding, followed by reinforcement learning with sub-goal supervision and recovery mechanisms to improve long-horizon reasoning. AgentCPM-GUI [367] combines grounding-aware pre-training, supervised imitation, and GRPO-based reinforcement fine-tuning with a concise JSON action space, reducing decoding overhead while improving robustness on long-horizon sequences. UI-Venus [368] is a multimodal screenshot-based UI agent fine-tuned via RFT with custom reward functions and a self-evolving trajectory framework, achieving new state-of-the-art in both UI grounding and navigation.

在静态设置中，强化学习在预收集的数据集上应用，这些数据具有确定性执行轨迹，并在缺乏在线环境交互时使用基于规则的结果评估标准。GUI-R1 [364] 采用 R1 风格的强化微调流水线，基于统一动作模式，使用简单的格式和正确性奖励以利用适量数据提高步级动作预测。UI-R1 [365] 通过群体相对策略优化来稳定策略更新，并通过紧凑动作接口及对动作类型与参数准确性的奖励塑形改善精确参数匹配。InFiGUI-R1 [366] 引入两阶段训练范式，先蒸馏空间推理以增强定位，随后采用带子目标监督和恢复机制的强化学习以改进长时程推理。AgentCPM-GUI [367] 结合感知定位的预训练、监督模仿和基于 GRPO 的强化微调，使用简洁的 JSON 动作空间，降低解码开销同时提高长时程序列的鲁棒性。UI-Venus [368] 是一个基于截图的多模态 UI 代理，通过定制奖励函数和自我进化轨迹框架进行 RFT 微调，在 UI 定位和导航上取得了新的最先进水平。

4.4.3.RL in Interactive GUI Environments

4.4.3. 交互式 GUI 环境中的强化学习

In interactive settings, reinforcement learning agents are optimized through online rollouts in dynamic environments, requiring robustness to stochastic transitions and long-horizon dependencies. WebAgent-R1 [104] conducts end-to-end multi-turn reinforcement learning with asynchronous trajectory generation and group-wise advantages, improving success on diverse web tasks. Vattikonda et al. [369] study reinforcement learning for web agents under realistic page dynamics and large action spaces, highlighting challenges in credit assignment and safe exploration. UI-TARS [370] integrates pre-training for GUI understanding with reinforcement learning for native desktop control, coupling milestone tracking and reflection to enhance long-horizon execution. DiGiRL [371] introduces an offline-to-online reinforcement learning pipeline on real Android devices, combining advantage-weighted updates, doubly robust advantage estimation, and instruction-level curricula to cope with non-stationarity. ZeroGUI [372] automates task generation and reward estimation with a vision-language evaluator, then applies two-stage online reinforcement learning

在交互式设置中，强化学习代理通过在动态环境中的在线 rollout 进行优化，要求对随机转移和长时程依赖具有鲁棒性。WebAgent-R1 [104] 进行端到端多回合强化学习，采用异步轨迹生成和分组优势，提升了对多样化网页任务的成功率。Vattikonda 等 [369] 研究了在真实页面动态和大动作空间下的网页代理强化学习，强调了归因和安全探索的挑战。UI-TARS [370] 将 GUI 理解的预训练与用于本地桌面控制的强化学习结合，配合里程碑跟踪与反思以增强长时程执行。DiGiRL [371] 在真实 Android 设备上引入离线到在线的强化学习流水线，结合优势加权更新、双重鲁棒优势估计和指令级课程来应对非平稳性。ZeroGUI [372] 使用视觉-语言评估器自动生成任务和估计奖励，然后应用两阶段在线强化学习

Table 7: A summary of methods for GUI agents, categorized by training paradigm and environment complexity.

表 7: 按训练范式和环境复杂度分类的 GUI 代理方法摘要。

Method	Paradigm	Environment	Resource Link
RL-free GUI Agents			
MM-Navigator [356]	Vanilla VLM	-	GitHub
SeeAct [357]	Vanilla VLM	-	GitHub
TRISHUL [358]	Vanilla VLM	-	-
InfiGUIAgent [361]	SFT	Static	} GitHub SHuggingFace # Website
UI-AGILE [362]	SFT	Interactive	C GitHub B HuggingFace
TongUI [363]	SFT	Static	GitHub B HuggingFace # Website
RL-based GUI Agents			
GUI-R1 [364]	RL	Static	GitHub HuggingFace
UI-R1 [365]	RL	Static	GitHub BHuggingFace
InFiGUI-R1 [366]	RL	Static	HGitHub HuggingFace
AgentCPM [367]	RL	Static	GiHu B HuggingFace
UI-Venus [368]	RL	Static	C GitHub
WebAgent-R1 [104]	RL	Interactive	-
Vattikonda et al. [369]	RL	Interactive	-
UI-TARS [370]	RL	Interactive	GitHub HuggingFace # Website
UI-TARS-2 [375]	RL	Interactive	GitHub # Website
DiGiRL [371]	RL	Interactive	GiftHub HuggingFace & Website
ZeroGUI [372]	RL	Interactive	C GitHub
MobileGUI-RL [373]	RL	Interactive	-
ComputerRL [374]	RL	Interactive	-

方法	范式	环境	资源链接
无需 RL 的 GUI 代理			
MM-Navigator [356]	原生 VLM	-	GitHub
SeeAct [357]	原生 VLM	-	GitHub
TRISHUL [358]	原生 VLM	-	-
InfiGUIAgent [361]	SFT	静态	} GitHub SHuggingFace # Website
UI-AGILE [362]	SFT	交互式	C GitHub B HuggingFace
TongUI [363]	SFT	静态	GitHub B HuggingFace # Website
基于 RL 的 GUI 代理			
GUI-R1 [364]	RL	静态	GitHub HuggingFace
UI-R1 [365]	RL	静态	GitHub BHuggingFace
InFiGUI-R1 [366]	RL	静态	HGitHub HuggingFace
AgentCPM [367]	RL	静态	GiHu B HuggingFace
UI-Venus [368]	RL	静态	C GitHub
WebAgent-R1 [104]	RL	交互式	-
Vattikonda et al. [369]	RL	交互式	-
UI-TARS [370]	RL	交互式	GitHub HuggingFace # Website
UI-TARS-2 [375]	RL	交互式	GitHub # Website
DiGiRL [371]	RL	交互式	GiftHub HuggingFace & Website
ZeroGUI [372]	RL	交互式	C GitHub
MobileGUI-RL [373]	RL	交互式	-
ComputerRL [374]	RL	交互式	-

(training on generated tasks followed by test-time adaptation) to reduce human supervision. MobileGUI-RL [373] scales training on Android virtual devices with trajectory-aware GRPO, a decaying efficiency reward, and curriculum filtering, improving execution efficiency and generalization while keeping the system practical for large rollout volumes. ComputerRL [374] introduces an API-GUI hybrid interaction paradigm paired with a massively parallel, fully asynchronous RL infrastructure and the novel Entropulse training strategy—alternating RL with supervised fine-tuning—to empower GUI-based agents to operate efficiently and scalably in desktop environments.

(在生成任务上的训练随后在测试时进行适应) 以减少人工监督。MobileGUI-RL [373] 通过在 Android 虚拟设备上采用轨迹感知的 GRPO、衰减效率奖励和课程过滤来扩展训练，提升执行效率与泛化能力，同时保持系统在大规模 rollout 下的实用性。ComputerRL [374] 引入了 API-GUI 混合交互范式，配合大规模并行、完全异步的 RL 基础设施以及新颖的 Entropulse 训练策略——在 RL 与监督微调之间交替——以使基于 GUI 的智能体在桌面环境中高效且可扩展地运行。

4.5.RL in Vision Agents

4.5.RL 在视觉智能体中的应用

RL has been applied to a wide range of vision tasks (including, but not limited to, image / video / 3D perception and generation). Since the number of related papers is substantial, this section does not aim to provide an exhaustive overview; for a more comprehensive survey on RL for various vision tasks, we refer readers to two dedicated surveys in vision [212, 211].

RL 已被应用于广泛的视觉任务 (包括但不限于图像 / 视频 / 3D 感知与生成)。鉴于相关论文数量庞大，本节并不力求穷尽；关于 RL 在各类视觉任务中的更全面综述，参见两篇专门的视觉综述 [212, 211]。

Image Tasks. The success of DeepSeek-R1 [32] has sparked widespread interest in applying RL to incentivize long-form reasoning behavior, encouraging LVLMs to produce extended CoT sequences that improve visual perception and understanding [208]. This research trajectory has evolved from early work that simply adapted R1-style objectives to the vision domain-aimed primarily at enhancing passive perception [220, 221, 222, 214, 215, 225, 226, 376]—toward the now-popular paradigm of active perception, or ”thinking with images” [210]. The key transition lies in moving from text-only CoT that references an image once, to interactive, visually grounded reasoning, achieved through (i) grounding [377, 232, 233, 234, 235, 236], (ii) agentic tool use [239, 240, 241, 242, 219, 98], and (iii) visual imagination via sketching or generation [243, 246, 247]. Beyond text-only outputs, many vision tasks—such as scene understanding—require structured predictions like bounding boxes, masks, and segmentation maps. To begin with, Visual-RFT [219] uses IoU with confidence as a verifiable reward for bounding-box outputs, while Vision-R1 [222] incorporates precision and recall as localization rewards. Extending this idea, [378] applies GRPO to segmentation tasks, combining soft and strict rewards with bounding-box IoU and L1 loss, and point-wise L1 distance. VLM-R1 [214] employs mean Average Precision (mAP) as a reward to explicitly incentivize detection and localization capabilities in LVLMs. Finally, R1-SGG [379] introduces three variants of GRPO rewards for scene-graph matching-ranging from hard rewards based on text matching and IoU to softer rewards computed via text-embedding dot products. RL has also been widely applied to image generation, particularly through its integration with diffusion and flow models—for example, RePrompt [380], Diffusion-KTO [381], Flow-GRPO [382], and GoT-R1 [383]. Beyond diffusion-based approaches, RL has been leveraged for autoregressive image generation, where it improves coherence, fidelity, and controllability by directly optimizing task- or user-specific reward signals [384, 247, 385].

图像任务。DeepSeek-R1 [32] 的成功激发了广泛兴趣，采用 RL 以激励长程推理行为，促使 LVLM 生成延展的连贯思路 (CoT) 序列，从而改善视觉感知与理解 [208]。这一研究路线从早期将 R1 风格目标简单移植到视觉领域、主要用于增强被动感知的工作 [220, 221, 222, 214, 215, 225, 226, 376] 演变而来，发展为如今流行的主动感知或“用图像思考”的范式 [210]。关键转变在于从只在文本中一次性引用图像的文本式 CoT，转向交互式、视觉落地的推理，通过 (i) 定位 (grounding)[377, 232, 233, 234, 235, 236]、(ii) 具主体性的工具使用 [239, 240, 241, 242, 219, 98]，以及 (iii) 通过素描或生成实现的视觉想象 [243, 246, 247] 来实现。除了纯文本输出，许多视觉任务 (如场景理解) 还要求结构化预测，如边界框、掩码和分割图。首先，Visual-RFT [219] 使用带置信度的 IoU 作为边界框输出的可验证奖励，Vision-R1 [222] 将精确率和召回率纳入定位奖励。将此思路扩展，[378] 将 GRPO 应用于分割任务，结合软/硬奖励、边界框 IoU 与 L1 损失以及逐点 L1 距离。VLM-R1 [214] 以平均精度 (mAP) 作为奖励，显式激励 LVLM 的检测与定位能力。最后，R1-SGG [379] 为场景图匹配提出了三种 GRPO 奖励变体——从基于文本匹配与 IoU 的硬奖励到通过文本嵌入点积计算的软奖励。RL 也被广泛应用于图像生成，特别是与扩散和流模型结合的情形，例如 RePrompt [380]、Diffusion-KTO [381]、Flow-GRPO [382] 与 GoT-R1 [383]。除基于扩散的方法外，RL 也被用于自回归图像生成，通过直接优化任务或用户特定的奖励信号来提升连贯性、保真度与可控性 [384, 247, 385]。

Video Tasks. Following the same spirit, numerous works have extended GRPO variants to the video domain [386, 387, 388] to enhance temporal reasoning [389, 390, 391, 392, 393]. TW-GRPO [394] introduces a token-weighted GRPO framework that emphasizes high-information tokens to generate more focused reasoning chains and employs soft, multi-choice rewards for lower-variance optimization. EgoVLM [395] combines keyframe-based rewards with direct GRPO training to produce interpretable reasoning traces tailored for egocentric video. DeepVideo-R1 reformulates the GRPO objective as a regression task [389], while VideoChat-R1 demonstrates that reinforcement finetuning (RFT) can be highly data-efficient for task-specific video reasoning improvements [390]. TinyLLaVA-Video-R1 explores scaling RL to smaller video LLMs [396], and [397] introduces infrastructure and a two-stage pipeline (CoT-SFT + RL) to support large-scale RL for long videos. Additional efforts have also extended RL for embodied video reasoning tasks [398]. A similar trend is observed in video generation, where RL is applied to improve temporal coherence, controllability, and semantic alignment. Key examples include DanceGRPO [399], GAPO [400], GRADEO [401], InfLVG [402], Phys-AR [403], VideoReward [404], TeViR [405], and InstructVideo [406].

视频任务。遵循同样思路，大量工作将 GRPO 变体扩展到视频领域 [386, 387, 388]，以增强时序推理 [389, 390, 391, 392, 393]。TW-GRPO [394] 引入了一种令高信息量 token 更受重视的 token 加权 GRPO 框架，以生成更聚焦的推理链，并采用软的多选奖励以降低方差。EgoVLM [395] 将关键帧奖励与直接的 GRPO 训练结合，产生面向第一视角视频的可解释推理轨迹。DeepVideo-R1 将 GRPO 目标重新表述为回归任务 [389]，而 VideoChat-R1 展示了强化微调 (RFT) 在任务特定视频推理改进上可以非常高效地利用数据 [390]。TinyLLaVA-Video-R1 探索了将 RL 扩展到更小型视频 LLM 的可行性 [396]，[397] 则引入了基础设施与两阶段流水线 (CoT-SFT + RL) 以支持面向长视频的大规模 RL。还有工作将 RL 扩展用于具身视频推理任务 [398]。在视频生成领域出现了类似趋势，RL 被用于提升时序一致性、可控性与语义对齐，代表性工作包括 DanceGRPO [399]、GAPO [400]、GRADEO [401]、InfLVG [402]、Phys-AR [403]、VideoReward [404]、TeViR [405] 与 InstructVideo [406]。

3D Vision Tasks. RL has also been widely adopted to advance 3D understanding [407, 408, 409, 410, 411, 412] and generation [413, 414, 415]. MetaSpatial [416] introduces the first RL-based framework for 3D spatial reasoning, leveraging physics-aware constraints and rendered-image evaluations as rewards during training. Scene-R1 [417] learns to reason about 3D scenes without point-wise 3D supervision, while SpatialReasoner [418] introduces shared 3D representations that unify perception, computation, and reasoning stages. In the domain of 3D generation, RL has been applied to improve text-to-3D alignment and controllability. Notable efforts include DreamCS [419], which aligns generation with human preferences; DreamDPO [420] and DreamReward [421], which optimize 3D generation using 2D reward signals; and Nabla-R2D3 [422], which further refines 3D outputs with reinforcement-driven objectives.

3D 视觉任务。强化学习也被广泛用于推进 3D 理解 [407, 408, 409, 410, 411, 412] 和生成 [413, 414, 415]。MetaSpatial [416] 提出首个基于强化学习的 3D 空间推理框架，在训练中利用物理感知约束和渲染图像评估作为奖励。Scene-R1 [417] 在没有逐点 3D 监督的情况下学习对 3D 场景进行推理，而 SpatialReasoner [418] 引入共享 3D 表示，统一感知、计算与推理阶段。在 3D 生成领域，强化学习被用于改进文本到 3D 的对齐与可控性。值得关注的工作包括 DreamCS [419]，将生成与人类偏好对齐；DreamDPO [420] 和 DreamReward [421]，使用 2D 奖励信号优化 3D 生成；以及 Nabla-R2D3 [422]，通过强化学习目标进一步细化 3D 输出。

4.6.RL in Embodied Agents

4.6.RL 在具身代理中的应用

While traditional agents are typically developed for general-purpose vision or language tasks, extending these capabilities to embodied agents requires a comprehensive understanding of real-world visual environments and the capacity to reason across modalities. Such competencies are essential for perceiving complex physical contexts and executing goal-directed actions conditioned on high-level instructions, forming a foundational element of agentic LLMs and MLLMs. In instruction-driven embodied scenarios, RL is often employed as a post-training strategy. A common pipeline begins with a pre-trained vision-language-action (VLA) model [423, 424, 425, 426] obtained through imitation learning under teacher-forcing supervision. This model is then embedded into an interactive agent that engages with the environment to collect reward signals. These rewards guide the iterative refinement of the policy, supporting effective exploration, improving sample efficiency, and enhancing the model's generalization capabilities across diverse real-world conditions. RL in VLA frameworks [427, 428, 429, 430] can be broadly categorized into two classes: navigation agents, which emphasize spatial reasoning and locomotion in complex environments, and manipulation agents, which focus on the precise control of physical objects under diverse and dynamic constraints.

传统代理通常面向通用视觉或语言任务，而将这些能力扩展到具身代理需要全面理解现实视觉环境并具备跨模态推理能力。这类能力对于感知复杂物理情境并在高层指令条件下执行目标导向动作至关重要，是具身化大模型 (agentic LLMs 与 MLLMs) 的基础。在基于指令的具身场景中，强化学习常作为后训练策略使用。一个常见流程是先通过教师强制下的模仿学习获得预训练的视觉-语言-动作 (VLA) 模型 [423, 424, 425, 426]，然后将该模型嵌入交互代理中与环境交互以收集奖励信号。这些奖励用于迭代优化策略，支持有效探索、提升样本效率并增强模型在多样现实条件下的泛化能力。VLA 框架中的强化学习 [427, 428, 429, 430] 大致可分为两类: 导航代理，强调复杂环境中的空间推理与机动性；以及操作代理，关注在多变约束下对物体的精确控制。

RL in VLA Navigation Agent. For navigation agents, planning is the central capability. Reinforcement learning is employed to enhance the VLA model's ability to predict and optimize future action sequences. A common strategy [431] is to integrate traditional robotics-style RL, using step-wise directional rewards, directly into VLA-based navigation frameworks. Some approaches operate at the trajectory level. VLN-R1 [429] aligns predicted and ground-truth paths to define trajectory-level rewards, and applies GRPO, following DeepSeek-R1, to improve predictive planning. OctoNav-R1 [376] also leverages GRPO but focuses on reinforcing internal deliberation within the VLA model, promoting a thinking-before-acting paradigm that enables more anticipatory and robust navigation. S2E [432] introduces a reinforcement learning framework that augments navigation foundation models with interactivity and safety, combining video pretraining with RL to achieve superior generalization and performance on the NavBench-GS benchmark.

VLA 导航代理中的强化学习。对于导航代理，规划是核心能力。强化学习用于提升 VLA 模型预测并优化未来动作序列的能力。一个常见策略 [431] 是将传统机器人风格的强化学习 (使用逐步方向性奖励) 直接整合到基于 VLA 的导航框架中。有些方法在轨迹层面运作。VLN-R1 [429] 通过对齐预测路径与真实路径定义轨迹级奖励，并在 DeepSeek-R1 之后应用 GRPO 来改进预测规划。OctoNav-R1 [376] 也利用 GRPO，但侧重于增强 VLA 模型的内部思考，推动“先思考后行动”范式，从而实现更具预见性和鲁棒性的导航。S2E [432] 引入了一个强化学习框架，通过为导航基础模型增加交互性与安全性，结合视频预训练与强化学习，在 NavBench-GS 基准上实现更好的泛化与性能。

RL in VLA Manipulation Agent. Manipulation agents, typically involving robotic arms, require fine-grained control for executing structured tasks under diverse conditions. In this context, RL is employed to enhance the instruction-following and trajectory prediction capabilities of VLA models, especially to improve generalization across tasks and environments. RLVLA [433] and VLA-RL [428] adopt pre-trained VLMs as evaluators, using their feedback to assign trajectory-level rewards for VLA policy refinement. These methods establish an online RL framework that effectively improves manipulation performance and demonstrates favorable scaling properties. TGRPO further [434] incorporates GRPO into manipulation tasks by defining rule-based reward functions over predicted trajectories. This enables the VLA model to generalize to unseen scenarios and improves its robustness in real-world deployment. VIKI-R [435] complements this with a unified benchmark and two-stage framework for multi-agent embodied cooperation, combining Chain-of-Thought fine-tuning with multi-level RL to enable compositional coordination across diverse embodiments.

VLA 操作代理中的强化学习。操作代理，通常涉及机械臂，需要精细控制以在多样条件下执行结构化任务。在此情境中，强化学习用于增强 VLA 模型的指令跟随与轨迹预测能力，特别是提升任务与环境间的泛化性。RLVLA [433] 与 VLA-RL [428] 采用预训练视觉语言模型作为评估器，利用其反馈为轨迹分配奖励以优化 VLA 策略。这些方法建立了在线强化学习框架，有效提升了操作性能并展示出良好的扩展性。TGRPO 进一步 [434] 通过在预测轨迹上定义基于规则的奖励函数将 GRPO 引入操作任务，使 VLA 模型能泛化到未见场景并提高其在现实部署中的鲁棒性。VIKI-R [435] 补充了统一基准与两阶段框架用于多代理具身协作，将链式思维微调与多层次强化学习结合，支持跨不同体现形式的组合式协调。

A central challenge in RL for VLA embodied agents is scaling training to real-world environments. While simulation platforms enable efficient large-scale experimentation, the sim-to-real gap remains significant, particularly in fine-grained manipulation tasks. Conducting RL directly in real-world settings is currently impractical due to the high cost and complexity of physical robot experiments. Most RL algorithms require millions of interaction steps, which demand substantial time, resources, and maintenance. As a result, developing scalable embodied RL pipelines that can bridge the gap between simulation and real-world deployment remains an open and pressing problem.

在 VLA 具身代理的强化学习中，一个核心挑战是将训练扩展到现实环境。尽管仿真平台支持高效的大规模实验，仿真到现实的差距仍然显著，尤其在精细操作任务上。在现实中直接进行强化学习目前不切实际，因为物理机器人实验成本高且复杂。大多数强化学习算法需要数百万次交互步骤，耗时且需大量资源与维护。因此，开发可扩展的具身强化学习流水线以弥合仿真与现实部署间差距仍是一个亟待解决的问题。

4.7.RL in Multi-Agent Systems

4.7.RL 在多智能体系统中的应用

Large Language Model (LLM)-based Multi-agent Systems (MAS) comprise multiple autonomous agents collaborating to solve complex tasks through structured interaction, coordination, and memory management.

基于大型语言模型 (LLM) 的多智能体系统 (MAS) 由多个自主代理组成，通过结构化交互、协作与记忆管理共同解决复杂任务。

Table 8: A summary of reinforcement learning and evolution paradigms in LLM-based Multi-Agent Systems. ”Dynamic” denotes whether the multi-agent system is task-dynamic, i.e., processes different task queries with different configurations (agent count, topologies, reasoning depth, prompts, etc). ”Train” denotes whether the method involves training the LLM backbone of agents.

表 8: 基于大模型的多智能体系统中强化学习与进化范式的摘要。“Dynamic”表示该多智能体系统是否为任务动态，即是否针对不同任务查询采用不同配置 (智能体数量、拓扑结构、推理深度、提示等)。“Train”表示该方法是否涉及对智能体的大模型骨干进行训练。

Method	Dynamic	Train	RL Algorithm	Resource Link
RL-Free Multi-Agent Systems (not exhaustive)				
CAMEL [436]	✘	✘	-	HigtiHub HuggingFace
MetaGPT [287]	✘	✘	-	GitHub
MAD [438]	✘	✘	-	GitHub
MoA [437]	✘	✘	-	Gi 比 Hub
AFlow [444]	✘	✘	-	Gi 比 Hub
RL-Based Multi-Agent Training				
GPTSwarm [440]	✘	✘	policy gradient	Gi GitHub #Website
MaAS [446]	✓	✘	policy gradient	GitHub
G-Designer [447]	✓	✘	policy gradient	OGitHub
Optima [448]	✘	✓	DPO	OGitHub
DITS [449]	✘	✓	DPO	-
MALT [168]	✘	✓	DPO	-
MARFT [450]	✘	✓	MARFT	Gi 在 Hub
ACC-Collab [160]	✘	✓	DPO	-
MAPoRL [451]	✓	✓	PPO	GitHub
MLPO [452]	✓	✓	MLPO	-
ReMA [453]	✓	✓	MAMRP	Gi 杜 Hub
FlowReasoner [454]	✓	✓	GRPO	Gi 在 Hub
CURE [294]	✘	✓	rule-based RL	Higig HuggingFace
MMedAgent-RL [455]	✘	✓	GRPO	-
Chain-of-Agents [456]	✓	✓	DAPO	GiHu HuggingFace
RLCCF [457]	✘	✓	GRPO	-
MAGRPO [458]	✘	✓	MAGRPO	-

方法	动态	训练	强化学习算法	资源链接
无 RL 的多智能体系统 (非详尽)				
CAMEL [436]	✘	✘	-	HigtiHub HuggingFace
MetaGPT [287]	✘	✘	-	GitHub
MAD [438]	✘	✘	-	GitHub
MoA [437]	✘	✘	-	Gi 比 Hub
AFlow [444]	✘	✘	-	Gi 比 Hub
基于 RL 的多智能体训练				
GPTSwarm [440]	✘	✘	策略梯度	Gi GitHub #Website
MaAS [446]	✓	✘	策略梯度	GitHub
G-Designer [447]	✓	✘	策略梯度	OGitHub
Optima [448]	✘	✓	DPO	OGitHub
DITS [449]	✘	✓	DPO	-
MALT [168]	✘	✓	DPO	-
MARFT [450]	✘	✓	MARFT	Gi 在 Hub
ACC-Collab [160]	✘	✓	DPO	-
MAPoRL [451]	✓	✓	PPO	GitHub
MLPO [452]	✓	✓	MLPO	-
ReMA [453]	✓	✓	MAMRP	Gi 杜 Hub
FlowReasoner [454]	✓	✓	GRPO	Gi 在 Hub
CURE [294]	✘	✓	基于规则的 RL	Higig HuggingFace
MMedAgent-RL [455]	✘	✓	GRPO	-
Chain-of-Agents [456]	✓	✓	DAPO	GiHu HuggingFace
RLCCF [457]	✘	✓	GRPO	-
MAGRPO [458]	✘	✓	MAGRPO	-

Early static and hand-designed MAS such as CAMEL and MetaGPT [436, 287] explored role specialization and task decomposition, while debate-based frameworks such as MAD and MoA [437, 438] enhanced reasoning via collaborative refinement. Subsequent multi-agent research has shifted to proposing optimizable cooperative systems, which enable MAS to not only dynamically adjust coordination patterns but also directly enhance agent-level reasoning and decision-making strategies. Table 8 summarizes the main body of works discussed in this section.

早期静态且手工设计的多智能体系统如 CAMEL 和 MetaGPT [436, 287] 探索了角色专门化与任务分解，而基于辩论的框架如 MAD 与 MoA [437, 438] 通过协作精炼增强了推理。随后多智能体研究转向提出可优化的协作系统，使 MAS 不仅能动态调整协同模式，还能直接提升个体智能体的推理与决策策略。表 8 总结了本节讨论的主要工作。

RL-Free Multi-Agent Evolution In the RL-free self-evolving setting, foundation models cannot be directly optimized; instead, system evolution is driven by mechanisms such as symbolic learning [439], dynamic graph optimization [440, 441, 442], and workflow rewriting [443, 444, 445]. These methods improve the coordination and adaptability within MAS, but cannot directly update the parameters of foundation models. MALT [168] employs a heterogeneous multi-agent search tree to generate large-scale labeled trajectories, fine-tuning agents via a combination of Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) from both successful and failed reasoning paths.

无强化学习的多智能体演化在无RL的自我演化设置中，基础模型无法被直接优化；系统演化由符号学习[439]、动态图优化[440, 441, 442]、工作流重写[443, 444, 445]等机制驱动。这些方法改进了MAS内的协同与适应性，但无法直接更新基础模型参数。MALT[168]使用异构多智能体搜索树生成大规模带标签轨迹，结合监督微调(SFT)与来自成功与失败推理路径的直接偏好优化(DPO)来微调智能体。

RL-Based Multi-Agent Training MARFT [450] formalizes a reinforcement fine-tuning framework for MAS with mathematical guarantees and empirical validation. MAGRPO [458] formalizes multi-LLM cooperation as a Dec-POMDP problem and introduces a multi-agent variant of GRPO, which enables joint training of LLM agents in MAS while maintaining decentralized execution. MAPoRL [451] extends MAD by verifying debate responses and using validation outcomes as RL rewards to improve collaborative reasoning. RLCCF [457] is a self-supervised multi-agent RL framework that leverages self-consistency-weighted ensemble voting to generate pseudo-labels and collaboratively optimize individual model policies via GRPO, boosting both individual and collective reasoning accuracy. MLPO [452] introduces a hierarchical paradigm in which a central LLM leader learns, via RL, to synthesize and evaluate peer agent outputs without auxiliary value networks. ReMA [453] separates reasoning into a meta-thinking agent and an execution agent, jointly trained under aligned RL objectives with parameter sharing. FlowReasoner [454] designs query-level meta-agents optimized through RL with multi-dimensional rewards (performance, complexity, efficiency) guided by execution feedback. LERO [459] combines MARL with LLM-generated hybrid rewards and evolutionary search to improve credit assignment and partial observability handling in cooperative tasks. CURE [294] focuses on code generation, jointly training a code generator and unit tester via RL to produce richer reward signals, achieving strong generalization across diverse coding benchmarks. MMedAgent-RL [455] introduces a reinforcement learning-based multi-agent framework for medical VQA, where dynamically coordinated general practitioners and specialists collaboratively reason with curriculum-guided learning, significantly outperforming existing Med-LVLMs and achieving more human-like diagnostic behavior. Chain-of-Agents (COA) [456] is an end-to-end paradigm where a single LLM simulates multi-agent collaboration by dynamically orchestrating role-playing and tool-using agents; this is achieved through multi-agent distillation (converting trajectories from state-of-the-art multi-agent systems into training data) and agentic reinforcement learning with carefully designed reward functions, resulting in Agent Foundation Models (AFMs). SPIRAL [460] presents a fully online, multi-turn, multi-agent self-play reinforcement learning framework for LLMs in zero-sum games, employing a shared policy with role-conditioned advantage estimation (RAE) to stabilize learning, and demonstrates that gameplay fosters transferable reasoning skills that significantly improve mathematical and general reasoning benchmarks.

基于 RL 的多智能体训练 MARFT [450] 将多智能体的强化微调框架形式化，给出数学保证并进行实证验证。MAGRPO [458] 将多 LLM 合作形式化为 Dec-POMDP 问题，并引入 GRPO 的多智能体变体，使得 MAS 中的 LLM 智能体可联合训练且保持去中心化执行。MAPoRL [451] 通过验证辩论回复并将验证结果作为 RL 奖励来改进协作推理，从而扩展了 MAD。RLCCF [457] 是一个自监督的多智能体 RL 框架，利用自一致性加权集成投票生成伪标签，并通过 GRPO 协同优化个体模型策略，提升个体与集体的推理准确性。MLPO [452] 引入分层范式，中央 LLM 领导者通过 RL 学会合成并评估同辈智能体输出，无需辅助价值网络。ReMA [453] 将推理拆分为元思考智能体与执行智能体，在对齐的 RL 目标下联合训练并共享参数。FlowReasoner [454] 设计了基于查询级的元智能体通过 RL 优化，使用由执行反馈指导的多维奖励 (性能、复杂度、效率)。LERO [459] 将 MARL 与 LLM 生成的混合奖励及进化搜索结合，以改进合作任务中的归因分配与部分可观测性处理。CURE [294] 聚焦代码生成，通过 RL 共同训练代码生成器与单元测试器以产生更丰富的奖励信号，在多样化编码基准上取得强泛化。MMedAgent-RL [455] 提出用于医学 VQA 的基于强化学习的多智能体框架，动态协调的全科与专家协同推理并采用课程化学习，显著优于现有 Med-LVLM 并展现更接近人类的诊断行为。Chain-of-Agents (COA) [456] 是端到端范式，单一 LLM 通过动态编排角色扮演与工具使用的智能体来模拟多智能体协作；通过多智能体蒸馏 (将先进多智能体系统的轨迹转为训练数据) 和设计精细奖励函数的主体化强化学习，生成智能体基础模型 (AFMs)。SPIRAL [460] 提出一个完全在线的多回合多智能体自我对弈强化学习框架用于零和游戏，采用具有角色条件优势估计 (RAE) 的共享策略以稳定学习，并证明对弈促进了可迁移的推理技能，显著提升数学与通用推理基准表现。

4.8. Other Tasks

4.8. 其他任务

TextGame. ARIA [461] compresses the sprawling action space via intention-driven reward aggregation, reducing sparsity and variance. GiGPO [119] enhances temporal credit assignment through hierarchical grouping without added computational burden. RAGEN [56] ensures stable multi-turn learning by filtering trajectories and stabilizing gradients, while advocating for reasoning-aware rewards. SPA-RL [120] decomposes delayed rewards into per-step signals, improving performance and grounding accuracy. Trinity-RFT [462] provides a unified, modular framework for reinforcement fine-tuning across tasks—including text games—enabling flexible, efficient, and scalable experimentation with diverse RL modes and data pipelines.

TextGame。ARIA [461] 通过意图驱动的奖励聚合压缩庞大的动作空间，降低稀疏性与方差。GiGPO [119] 通过层级分组增强时间信用分配而不增加计算负担。RAGEN [56] 通过过滤轨迹与稳定梯度确保稳定的多回合学习，同时倡导推理感知奖励。SPA-RL [120] 将延迟奖励分解为逐步信号，提升性能与落地准确性。Trinity-RFT [462] 提供统一的模块化强化微调框架，适用于包括文本游戏在内的多任务，实现灵活、高效且可扩展的多种 RL 模式与数据流水线实验。

Table. SkyRL-SQL [463] introduces a data-efficient, multi-turn RL pipeline for Text-to-SQL, enabling LLM agents to interactively probe databases, refine, and verify SQL queries. With just 653 training examples, the SkyRL-SQL-7B model surpasses both GPT-40 and o4-mini on SQL generation benchmarks. MSRL [464] introduces multimodal structured reinforcement learning with multi-granularity rewards to overcome the SFT plateau in chart-to-code generation, achieving state-of-the-art performance on chart understanding benchmarks

表。SkyRL-SQL [463] 引入一种数据高效的多回合 Text-to-SQL 强化学习流水线，使 LLM 智能体可交互探询数据库、细化并验证 SQL 查询。仅用 653 个训练示例，SkyRL-SQL-7B 模型在 SQL 生成基准上超越了 GPT-40 与 o4-mini。MSRL [464] 提出多模态结构化强化学习，使用多粒度奖励克服 SFT 在图表到代码生成中的瓶颈，在图表理解基准上实现最先进性能。

Time Series. Time-R1 [465] enhances moderate-sized LLMs with comprehensive temporal reasoning abilities through a progressive reinforcement learning curriculum and a dynamic rule-based reward system.

时间序列。Time-R1 [465] 通过渐进式强化学习课程与动态基于规则的奖励系统, 增强中等规模 LLM 的全面时序推理能力。

TimeMaster [466] trains time-series MLLMs that combine SFT with GRPO to enable structured, interpretable temporal reasoning over visualized time-series inputs.

TimeMaster [466] 训练时间序列 MLLM，结合 SFT 与 GRPO，以实现对可视化时间序列输入的结构化、可解释的时间推理。

General QA. Agent models [467] internalize chain-of-action generation to enable autonomous and efficient decision-making through a combination of supervised fine-tuning and reinforcement learning. L-Zero [468] enables large language models to become general-purpose agents through a scalable, end-to-end reinforcement learning pipeline utilizing a low-cost, extensible, and sandboxed concurrent agent worker pool.

通用问答。Agent 模型 [467] 将行动链生成内化，通过监督微调与强化学习相结合，实现自主且高效的决策。L-Zero [468] 通过一个可扩展、端到端的强化学习流水线，使大语言模型成为通用代理，该流水线利用低成本、可扩展且沙箱化的并发代理工作池。

Social. Sotopia-RL [469] refines coarse episode-level rewards into utterance-level, multi-dimensional signals to enable efficient and stable RL training for socially intelligent LLMs under partial observability and multi-faceted objectives. [470] introduces an Adaptive Mode Learning (AML) framework with the Adaptive Mode Policy Optimization (AMPO) algorithm, which uses reinforcement learning to dynamically switch between multi-granular reasoning modes in social intelligence tasks, achieving higher accuracy and shorter reasoning chains than fixed-depth RL methods like GRPO.

社交。Sotopia-RL [469] 将粗粒度的回合级奖励细化为话语级、多维信号，以在部分可观测和多目标情形下实现高效且稳定的社交智能 LLM 强化学习训练。[470] 提出自适应模式学习 (AML) 框架及自适应模式策略优化 (AMPO) 算法，使用强化学习在社交智能任务中动态切换多粒度推理模式，较固定深度的 RL 方法 (如 GRPO) 达到更高准确率和更短推理链。

5. Enviroment and Frameworks

5. Enviroment and Frameworks

5.1. Environment Simulator

5.1. Environment Simulator

In agentic reinforcement learning, the environment is the world with which the agent interacts, receiving sensory input (observations) and enacting choices (actions) through its actuators. The environment, in turn, responds to the agent's actions by transitioning to a new state and providing a reward signal. With the rise of the LLM Agent paradigm, many works have proposed environments for training specific tasks. Table 9 provides an overview of the key environments examined in this section.

在具代理性的强化学习中，环境是代理交互的世界，代理通过感知输入 (观测) 并通过执行器采取选择 (动作)。环境则根据代理的动作转移到新状态并提供奖励信号。随着 LLM Agent 范式的兴起，许多工作提出了用于训练特定任务的环境。本节表 9 概述了所考察的关键环境。

Table 9: A summary of environments and benchmarks for agentic reinforcement learning, categorized by agent capability, task domain, and modality. The agent capabilities are denoted by: ① Reasoning, ② Planning, ③ Tool Use, ④ Memory, ⑤ Collaboration, ⑥ Self-Improve.

表 9: 面向具代理性强化学习的环境与基准汇总，按代理能力、任务领域和模态分类。代理能力标记为: ① 推理，② 规划，③ 工具使用，④ 记忆，⑤ 协作，⑥ 自我提升。

Environment / Benchmark	Agent Capability	Task Domain	Modality	Resource Link
LMRL-Gym [471]	①,④	Interaction	Text	GiftHub
ALFWorld [472]	②,①	Embodied, Text Games	Text	Gi tHub #Website
TextWorld [473]	②,①	Text Games	Text	Gi tHub
ScienceWorld [474]	①,②	Embodied, Science	Text	C GitHub & Website
AgentGym [475]	①,④	Text Games	Text	GitHub & Website
Agentbench [476]	①	General	Text, Visual	exGitHub
InternBootcamp [477]	①	General, Coding, Logic	Text	C GitHub
LoCoMo [478]	④	Interaction	Text	Gi tHub #Website
MemoryAgentBench [479]	④	Interaction	Text	C GitHub
WebShop [480]	②,③	Web	Text	Gi tHub #Website
Mind2Web [481]	②,③	Web	Text, Visual	GitHub #Website
WebArena [482]	②,③	Web	Text	C GitHub #Website
VisualwebArena [483]	①,②,③	Web	Text, Visual	C GitHub #Website
AppBench [484]	②,③	App	Text	GitHub
AppWorld [485]	②,③	App	Text	Gi tHub #Website
AndroidWorld [486]	②,③	GUI, App	Text, Visual	GitHub
OSWorld [487]	②,③	GUI, OS	Text, Visual	GitHub & Website
WindowsAgentArena [488]	②			
Debug-Gym [489]	①,③	SWE	Text	} GitHub # Website
MLE-Dojo [490]	②,①	MLE	Text	C GitHub # Website
τ -bench [491]	①,③	SWE	Text	GitHub
TheAgentCompany [492]	②,③,⑤	SWE	Text	Gi tHub # Website
MedAgentGym [493]	①	Science	Text	C GitHub
SecRepoBench [494]	①,③	Coding, Security	Text	-
R2E-Gym [495]	①,②	SWE	Text	Gi tHub #Website
BigCodeBench [496]	①	Coding	Text	C GitHub #Website
LiveCodeBench [497]	①	Coding	Text	Gi tHub #Website
SWE-bench [498]	①,③	SWE	Text	Gi tHub #Website
SWE-rebench [499]	①,③	SWE	Text	Website
DevBench [500]	②,①	SWE	Text	GitHub
ProjectEval [501]	②,①	SWE	Text	} GitHub # Website
DA-Code [502]	①,③	Data Science, SWE	Text	C GitHub # Website
ColBench [159]	②,①	SWE, Web Dev	Text	GitHub # Website
NoCode-bench [503]	②,①	SWE	Text	Gi tHub # Website
MLE-Bench [504]	②,①,③	MLE	Text	C GitHub # Website
PaperBench [505]	②,①,③	MLE	Text	C GitHub # Website
Crafter [506]	②,④	Game	Visual	Gi tHub #Website
Craftax [507]	②,④	Game	Visual	C GitHub
ELLM (Crafter variant) [508]	②,①	Game	Visual	G GitHub #Website
SMAC / SMAC-Exp [509]	⑤,②	Game	Visual	GiHHub
Factorio [510]	②,①	Game	Visual	GiHu be Website
SMAC-Hard [511]	②,④	Game	Visual	Gi 在 Hub
TacticCraft [512]	②,⑤	Game	Text	-

环境 / 基准	代理能力	任务领域	模态	资源链接
LMRL-Gym [471]	①,④	交互	文本	GiftHub
ALFWorld [472]	②,①	具身，文字游戏	文本	Gi tHub #Website
TextWorld [473]	②,①	文字游戏	文本	Gi tHub
ScienceWorld [474]	①,②	具身，科学	文本	C GitHub & Website
AgentGym [475]	①,④	文字游戏	文本	GitHub & Website
Agentbench [476]	①	通用	文本，视觉	exGitHub
InternBootcamp [477]	①	通用，编码，逻辑	文本	C GitHub
LoCoMo [478]	④	交互	文本	Gi tHub #Website
MemoryAgentBench [479]	④	交互	文本	C GitHub
WebShop [480]	②,③	网页	文本	Gi tHub #Website
Mind2Web [481]	②,③	网页	文本，视觉	GitHub #Website
WebArena [482]	②,③	网页	文本	C GitHub #Website
VisualwebArena [483]	①,②,③	网页	文本，视觉	C GitHub #Website
AppBench [484]	②,③	应用	文本	GitHub
AppWorld [485]	②,③	应用	文本	Gi tHub #Website
AndroidWorld [486]	②,③	图形界面，应用	文本，视觉	GitHub
OSWorld [487]	②,③	图形界面，操作系统	文本，视觉	GitHub & Website
WindowsAgentArena [488]	②			
Debug-Gym [489]	①,③	软件工程	文本	} GitHub # Website
MLE-Dojo [490]	②,①	MLE	文本	C GitHub # Website
τ -bench [491]	①,③	软件工程	文本	GitHub
TheAgentCompany [492]	②,③,⑤	软件工程	文本	Gi tHub # Website
MedAgentGym [493]	①	科学	文本	C GitHub
SecRepoBench [494]	①,③	编码，安全	文本	-
R2E-Gym [495]	①,②	软件工程	文本	Gi tHub #Website
BigCodeBench [496]	①	编码	文本	C GitHub #Website
LiveCodeBench [497]	①	编码	文本	Gi tHub #Website
SWE-bench [498]	①,③	软件工程	文本	Gi tHub #Website
SWE-rebench [499]	①,③	软件工程	文本	网站
DevBench [500]	②,①	软件工程	文本	GitHub
ProjectEval [501]	②,①	软件工程	文本	} GitHub # Website
DA-Code [502]	①,③	数据科学，软件工程	文本	C GitHub # Website
ColBench [159]	②,①	软件工程，网页开发	文本	GitHub # 网站
NoCode-bench [503]	②,①	软件工程	文本	Gi tHub # Website
MLE-Bench [504]	②,①,③	MLE	文本	C GitHub # Website
PaperBench [505]	②,①,③	MLE	文本	C GitHub # Website
Crafter [506]	②,④	游戏	视觉	Gi tHub #Website
Craftax [507]	②,④	游戏	视觉	C GitHub
ELLM(Crafter 变体) [508]	②,①	游戏	视觉	G GitHub # 网站
SMAC / SMAC-Exp [509]	⑤,②	游戏	视觉	GiHhub
Factorio [510]	②,①	游戏	视觉	GiHu be 网站
SMAC-Hard [511]	②,④	游戏	视觉	Gi 在 Hub
TacticCraft [512]	②,⑤	游戏	文本	-

5.1.1.Web Environments

5.1.1.Web Environments

In the realm of web-based environments, several benchmarks offer controlled yet realistic static environments for Agentic RL. WebShop [480] is a simulated e-commerce website featuring a large catalog of real-world products and crowd-sourced text instructions. Agents navigate various webpage types and issue diverse actions (e.g., searching, selecting items, customizing, purchasing) to find and buy products, with its deterministic search engine aiding reproducibility. Furthermore, Mind2Web [481] is a dataset designed for generalist web agents, featuring a substantial number of tasks from many real-world websites across diverse domains. It provides webpage snapshots and crowdsourced action sequences for tasks like finding flights or interacting with social profiles, emphasizing generalization across unseen websites and domains. Similarly, WebArena [482] and its multimodal extension, VisualwebArena [483], are self-hostable, reproducible web environments delivered as Docker containers. WebArena features fully functional websites across common domains like e-commerce, social forums, collaborative development, and content management systems, enriched with utility tools and knowledge bases, and supports multi-tab tasks and user role simulation. VisualwebArena extends this by introducing new tasks requiring visual comprehension and a ”Set-of-Marks” (SoM) representation to annotate interactable elements on screenshots, bridging the gap for multimodal web agents. Additionally, AppWorld [485] constitutes an environment simulating a multi-application ecosystem, encompassing 9 daily-use applications (e.g., Amazon, Spotify, Gmail) with 457 invokable APIs, and constructing a digital world featuring approximately 100 virtual characters and their social relationships. Agents accomplish complex tasks (such as travel planning and social relationship management) by writing code to call APIs. In these environments, all changes to the web pages or visual elements occur exclusively in response to the agent’s actions.

在基于网页的环境中, 有若干基准提供受控但现实的静态环境以用于 Agentic RL。WebShop [480] 是一个模拟的电子商务网站，包含大量真实世界产品目录和众包文本指令。代理在各种网页类型中导航并执行多种动作 (例如搜索、选择商品、定制、购买) 以查找并购买产品, 其确定性的搜索引擎有助于可复现性。此外，Mind2Web [481] 是为通用网页代理设计的数据集，收录来自众多真实网站和不同领域的大量任务，提供网页快照和众包的动作序列用于诸如查找航班或与社交资料交互等任务，强调在未见网站和领域上的泛化能力。类似地，WebArena [482] 及其多模态扩展 VisualwebArena [483] 是可自托管、可复现的网页环境，以 Docker 容器形式提供。WebArena 包含跨常见领域 (如电子商务、社交论坛、协作开发和内容管理系统) 的全功能网站，配有实用工具和知识库，并支持多标签任务与用户角色模拟。VisualwebArena 通过引入需要视觉理解的新任务及“Set-of-Marks” (SoM) 表示来标注截图上的可交互元素，弥合多模态网页代理的差距。此外，AppWorld [485] 构成一个模拟多应用生态的环境，涵盖 9 个日常应用 (例如 Amazon、Spotify、Gmail) 和 457 个可调用 API，并构建了包含约 100 名虚拟角色及其社交关系的数字世界。代理通过编写代码调用 API 来完成复杂任务 (如旅行规划和社交关系管理)。在这些环境中，对网页或视觉元素的所有更改均仅响应代理的操作发生。

5.1.2.GUI Environments

5.1.2.GUI Environments

AndroidWorld [486] exemplifies such dynamism as a benchmarking environment operating on a live

Android emulator, featuring 116 hand-crafted tasks across 20 real-world applications. Its dynamic nature is underscored by parameter instantiation that generates millions of unique task variations, ensuring the environment evolves into novel configurations without direct agent influence. Agents interact through a consistent interface (supporting screen interactions, app navigation, and text input) while receiving real-time state feedback, with integration to MiniWoB+ + providing durable reward signals for evaluating adaptive performance. OSWorld [487] is a scalable real computer environment for multimodal agents, supporting task setup and execution-based evaluation across Ubuntu, Windows, and macOS. It includes a substantial number of real-world computer tasks involving real web and desktop applications, OS file I/O, and workflows spanning multiple applications, where all OS state changes are exclusively triggered by the agent's actions.

AndroidWorld [486] 展示了此类动态性的示例，作为在实时 Android 模拟器上运行的基准环境，包含针对 20 个真实应用的 116 个手工设计任务。其动态性体现在参数实例化会生成数百万种独特任务变体，确保环境在未被代理直接影响的情况下演化出新配置。代理通过一致的接口交互 (支持屏幕操作、应用导航和文本输入)，并接收实时状态反馈，且与 MiniWoB++ 的集成为评估适应性表现提供持久的奖励信号。OSWorld [487] 是一个面向多模态代理的可扩展真实计算机环境，支持在 Ubuntu、Windows 和 macOS 上进行任务设置与基于执行的评估。它包括大量涉及真实网页与桌面应用、操作系统文件 I/O 以及跨多个应用工作流的真实计算机任务，其中所有操作系统状态的更改均仅由代理的操作触发。

5.1.3. Coding & Software Engineering Environments

5.1.3. Coding & Software Engineering Environments

Code-related tasks are supported by a wide range of executable environments and benchmarks. These can be broadly categorized into interactive environments, where agents directly alter the state, and benchmarks/datasets that provide curated tasks and evaluation pipelines.

与代码相关的任务由各类可执行环境和基准支持。它们大致可分为交互式环境 (代理直接改变状态) 以及提供精心策划任务和评估流水线的基准/数据集。

Interactive SWE Environments. Several environments instantiate agent-environment interaction under software engineering workflows. Debug-Gym [489] is a text-based interactive coding environment for LLM agents in debugging settings. It equips agents with tools like a Python debugger (pdb) to actively explore and modify buggy codebases, supporting repository-level information handling and ensuring safety via Docker containers. R2E-Gym [495] constructs a procedurally generated, executable gym-style environment of over 8K software engineering tasks, powered by the SWE-Gen pipeline and hybrid verifiers. TheAgentCompany [492] simulates a software development company, where agents act as ”digital workers” performing professional tasks such as web browsing, coding, program execution, and communication with simulated colleagues. It features a diverse set of long-horizon tasks with checkpoints for partial credit, providing a comprehensive testbed for agents in a realistic workplace setting. In all these environments, the underlying problem definitions and codebases remain fixed, and changes occur solely as a result of the agent's actions.

交互式 SWE 环境。若干环境在软件工程工作流下实例化代理-环境交互。Debug-Gym [489] 是面向 LLM 代理的文本交互式调试编码环境，为代理配备诸如 Python 调试器 (pdb) 等工具以主动探索并修改有缺陷的代码库，支持仓库级信息处理并通过 Docker 容器确保安全性。R2E-Gym [495] 构建了一个由 SWE-Gen 流水线和混合验证器驱动、包含 8K 余软件工程任务的程序生成可执行 gym 风格环境。TheAgentCompany [492] 模拟了一家软件开发公司，代理作为“数字工人”执行浏览网页、编码、运行程序和与模拟同事沟通等专业任务。它包含多样的长时程任务并设有检查点以给予部分分数，为真实工作场景下的代理提供了全面的测试床。在所有这些环境中，底层的问题定义和代码库保持固定，且更改仅由代理的操作产生。

Coding Benchmarks & Datasets. A wide range of benchmarks and datasets focus on constructing curated task suites and evaluation pipelines. HumanEval [513] introduces a benchmark of 164 hand-crafted Python programming tasks to measure functional correctness via the pass@k metric. MBPP [514] provides 974 entry-level Python tasks with natural language descriptions for evaluating short program synthesis. BigCodeBench [496] proposes a large-scale, contamination-free function-level benchmark of 1,140 tasks requiring composition of multiple function calls. LiveCodeBench [497] builds a continuously updated, contamination-free benchmark from real competition problems. SWE-bench [498] introduces a dynamic, execution-driven code repair benchmark derived from real GitHub issues. SWE-rebench [499] introduces a continual GitHub-mining pipeline (>21k tasks) for both training and evaluation. DevBench [500] evaluates end-to-end development across design, setup, implementation, and testing. ProjectEval [501] constructs LLM-generated, human-reviewed project tasks with simulated user interactions. ColBench [159] instantiates multi-turn backend/frontend tasks with a privileged critic for step-wise rewards. NoCode-bench [503] evaluates LLMs on feature addition from documentation updates across real codebases. CodeBoost [298] serves as a data-centric, execution-driven training pipeline by extracting and augmenting code snippets.

编码基准与数据集。大量基准与数据集着眼于构建人工策划的任务套件与评估管道。HumanEval [513] 提出包含 164 个手工设计的 Python 编程任务的基准，通过 pass@k 度量功能正确性。MBPP [514] 提供 974 个带自然语言描述的入门级 Python 任务，用于评估短程序合成。BigCodeBench [496] 提出一个大规模、无污染的函数级基准，包含 1,140 个需要组合多个函数调用的任务。LiveCodeBench [497] 从真实竞赛题目构建持续更新的无污染基准。SWE-bench [498] 引入一个动态、执行驱动的代码修复基准，来源于真实的 GitHub 问题。SWE-rebench [499] 提出一个持续的 GitHub 挖掘管道 (>21k 任务)，用于训练与评估。DevBench [500] 评估从设计、搭建、实现到测试的端到端开发能力。ProjectEval [501] 构建由 LLM 生成并经人工审阅的项目任务，包含模拟的用户交互。ColBench [159] 实例化多轮后端/前端任务，配备特权评价者以获得逐步奖励。NoCode-bench [503] 在真实代码库中评估 LLM 在根据文档更新添加功能方面的表现。CodeBoost [298] 作为一个以数据为中心、执行驱动的训练管道，通过提取和增强代码片段来服务。

5.1.4. Domain-specific Environments

5.1.4. 特定领域环境

Science & Research. ScienceWorld [474] integrates science simulations (e.g., thermodynamics, electricity, chemistry) into complex text-based tasks designed around elementary-level science education. PaperBench [505] evaluates the ability of LLM agents to replicate cutting-edge machine learning research by reproducing 20 ICML 2024 papers from scratch, scored against rubric-based subtasks. τ -bench [491] simulates dynamic conversations for software engineering tasks, operating with an underlying database state and domain-specific rules that change only through the agent's API calls.

科学与研究。ScienceWorld [474] 将科学模拟 (如热力学、电学、化学) 整合进围绕小学科学教育设计的复杂文本任务中。Paper-Bench [505] 通过从头复现 20 篇 ICML 2024 论文并按评分细则对子任务评分，评估 LLM 代理复现前沿机器学习研究的能力。τ-bench [491] 模拟面向软件工程任务的动态对话，操作基于一个仅能通过代理 API 调用改变的底层数据库状态和领域规则。

Machine Learning Engineering (MLE). MLE-Dojo [490] is a Gym-style framework for iterative machine learning engineering workflows, built upon real-world Kaggle competitions. It provides an interactive environment for agents to iteratively experiment, debug, and refine solutions. MLE-Bench [504] establishes a benchmark for MLE by curating 75 Kaggle competitions, evaluating agents against human baselines on public leaderboards. DA-Code [502] addresses agentic data-science workflows grounded in real datasets and executable analysis, providing a focused benchmark for this domain.

机器学习工程 (MLE)。MLE-Dojo [490] 是一个基于真实 Kaggle 竞赛构建的类似 Gym 的迭代机器学习工程工作流框架，为代理提供交互式环境以迭代试验、调试和改进解决方案。MLE-Bench [504] 通过策划 75 个 Kaggle 竞赛建立 MLE 基准，按公开排行榜将代理与人类基线对比评估。DA-Code [502] 针对基于真实数据集与可执行分析的代理数据科学工作流，提供了一个聚焦的领域基准。

Biomedical. MedAgentGym [493] provides a domain-specific environment for biomedical code generation and testing, focusing on tasks within this specialized scientific field.

生物医学。MedAgentGym [493] 提供一个面向生物医学代码生成与测试的特定领域环境，聚焦于该专业科学领域内的任务。

Cybersecurity. SecRepoBench [494] is a domain-specific benchmark for security vulnerability repair, covering 27 repositories and 15 Common Weakness Enumeration (CWE) categories.

网络安全。SecRepoBench [494] 是一个针对安全漏洞修复的特定领域基准，覆盖 27 个仓库和 15 类常见弱点枚举 (CWE)。

5.1.5. Simulated & Game Environments

5.1.5. 模拟与游戏环境

Text-based environments simulate interactive settings where agent actions are expressed through natural language. LMRL-Gym [471] provides a benchmark for evaluating reinforcement learning algorithms in multi-turn language interactions, including tasks like ”20 Questions” and Chess. TextWorld [473] is a sandbox environment for training agents in text-based games, offering both hand-authored and procedurally generated games. Game-based environments also emphasize visual settings that may evolve independently. Crafter [506] is a 2D open-world survival game that benchmarks deep exploration and long-horizon reasoning. Craftax [507], built upon Crafter using JAX, introduces increased complexity and GPU-acceleration for open-ended RL. The modified Crafter variant by ELLM [508] expands the action space and introduces distractor tasks. For multi-agent coordination, SMAC [509] and SMAC-Hard [511] provide StarCraft II-based benchmarks for cooperative decentralized control. SMAC-R1 [511], Adaptive Command [515] and TacticCraft [512] further advance the performance of LLM agents in StarCraft II-style environments. Factorio [510] presents a dynamic, tick-based industrial simulation where agent inaction still alters the world state.

基于文本的环境模拟交互场景，代理动作以自然语言表达。LMRL-Gym [471] 为评估多轮语言交互中的强化学习算法提供基准，包含“20 问”与国际象棋等任务。TextWorld [473] 是用于训练文本游戏代理的沙盒环境，提供手工编写与程序生成的游戏。基于游戏的环境也强调可能独立演变的视觉场景。Crafter [506] 是一个 2D 开放世界生存游戏，用于基准深度探索与长时程推理。基于 Crafter 且使用 JAX 的 Craftax [507] 在开放式强化学习中引入更高复杂性与 GPU 加速。ELLM [508] 修改的 Crafter 变体扩大了动作空间并引入干扰任务。对于多代理协同，SMAC [509] 与 SMAC-Hard [511] 提供基于《星际争霸 II》的合作去中心化控制基准。SMAC-R1 [511]、Adaptive Command [515] 与 TacticCraft [512] 进一步推动 LLM 代理在类星际争霸 II 环境中的表现。Factorio [510] 提供一个动态的基于时钟滴答的工业仿真，其中代理的不作为也会改变世界状态。

5.1.6. General-Purpose Environments

5.1.6. 通用环境

Some environments and benchmarks are designed for broad evaluation or to improve general agent capabilities. AgentGym [475] focuses on improving LLM agent generalization via instruction tuning and self-correction, operating on deterministic environments such as Alf World, BabyAI, and SciWorld. Agent-bench [476] serves as a broad evaluation framework, assessing LLMs as agents across a variety of distinct interactive environments, including SQL-based, game-based, and web-based scenarios. InternBootcamp [477] is a scalable framework integrating over 1000 verifiable reasoning tasks, spanning programming, logic puzzles, and games, with a standardized interface for RL training and automated task generation.

一些环境与基准旨在进行广泛评估或提升通用代理能力。AgentGym [475] 侧重通过指令微调与自我纠正来提升 LLM 代理的泛化能力, 在 Alf World、BabyAI 和 SciWorld 等确定性环境上运行。Agent-bench [476] 作为一个广泛的评估框架，评估 LLM 作为代理在多种不同交互环境中的表现，包括基于 SQL、基于游戏和基于网页的场景。InternBootcamp [477] 是一个可扩展框架，整合了 1000 多个可验证的推理任务，涵盖编程、逻辑谜题与游戏，提供标准化的强化学习训练接口与自动任务生成。

5.2.RL Framework

5.2.RL 框架

In this section, we summarize three categories of codebases/frameworks most relevant to this work: agentic RL frameworks, RLHF and LLM fine-tuning frameworks, and general-purpose RL frameworks. Table 10 provides an overview of the prevailing agentic RL and LLM-RL frameworks for readers' reference.

在本节中，我们总结了与本工作最相关的三类代码库/框架: 具代理性的 RL 框架、RLHF 与 LLM 微调框架，以及通用 RL 框架。表 10 为读者提供了现行具代理性 RL 与 LLM-RL 框架的概览，供参考。

Agentic RL frameworks. Verifiers [516] introduces a verifiable-environment setup for end-to-end policy optimization with LLMs, while SkyRL-v0 [517] and its modular successors [518] demonstrate long-horizon, real-world agent training via reinforcement learning. AREAL [519] scales this paradigm with an asynchronous, distributed architecture tailored to language reasoning tasks, and MARTI [520] extends it further to multi-agent LLM systems that integrate reinforcement training and inference. EasyR1 [521] brings multi-modality support, enabling agents to leverage vision and language signals together in a unified RL framework. AgentFly [522] presents a scalable and extensible agentic RL framework that empowers language-model agents with traditional RL algorithms—enabling token-level multi-turn interaction via decorator-based tools and reward definition, asynchronous execution, and centralized resource management for high-throughput RL training. Agent Lightning [523] is a flexible RL framework that decouples agent execution from training by modeling execution as an MDP and using a hierarchical RL algorithm (LightningRL) to train any AI agent with near-zero code modification. AWORLD [275] is a distributed agentic RL framework, which tackles the main bottleneck of agent training—experience generation—by orchestrating massively parallel rollouts across clusters, achieving a 14.6 × speedup over single-node execution and enabling scalable end-to-end training pipelines. ROLL [525] provides a scalable library for large-scale RL optimization with a unified

具代理性的 RL 框架。Verifiers [516] 提出了一种可验证环境设置，用于与 LLM 的端到端策略优化；SkyRL-v0 [517] 及其模块化后续工作 [518] 展示了通过强化学习进行长时程、真实世界的代理训练。AREAL [519] 以一种异步分布式架构扩展了该范式，专为语言推理任务而设计；MARTI [520] 进一步将其扩展到整合强化训练与推理的多智能体 LLM 系统。EasyR1 [521] 引入了多模态支持，使代理能够在统一的 RL 框架中同时利用视觉与语言信号。AgentFly [522] 提出了一种可扩展且可扩展的具代理性 RL 框架，使语言模型代理能使用传统 RL 算法——通过基于装饰器的工具与奖励定义实现逐 token 多轮交互、异步执行以及面向高吞吐量 RL 训练的集中资源管理。Agent Lightning [523] 是一款灵活的 RL 框架，通过将执行建模为 MDP 并使用分层 RL 算法 (LightningRL) 来训练任意 AI 代理，从而实现执行与训练的解耦，几乎无需修改代码。AWORLD [275] 是一个分布式具代理性 RL 框架，通过在集群间协调大规模并行 rollout 来解决代理训练的主要瓶颈——经验生成，实现了相对于单节点执行 14.6 × 的加速，并支持可扩展的端到端训练流水线。ROLL [525] 提供了一个用于大规模 RL 优化的可扩展库，具备统一

Framework	Type	Key Features	Resource Link
Agentic RL Frameworks			
Verifiers [516]	Agentic RL	Verifiable environment setup	C GitHub
SkyRL-v0 [517, 518]	Agentic RL	Long-horizon real-world training	C GitHub
AREAL [519]	Agentic RL	Asynchronous training	GiHtHub
MARTI [520]	Multi-agent RL	Integrated multi-agent training	GitHub
EasyR1 [521]	Agentic RL	Multimodal support	GitHub
AgentFly [522]	Agentic RL	Scalable asynchronous execution	⍰⍰⍰⍰⍰Hub
Agent Lightning [523]	Agentic RL	Decoupled hierarchical RL	C GitHub
AWorld [275]	Agentic RL	Parallel rollouts across clusters	C GitHub
RL-Factory [524]	Agentic RL	Easy-to-design reward	GiHHub
ROLL [525]	Agentic RL	Stable Multi-GPU Parallel Training	GitHub
VerlTool [526]	Agentic RL	Tool-intergrated rollout	GitHub
AgentRL [527]	Agentic RL	Asynchronous Multi-Task Training	{fitHub
RLHF and LLM Fine-tuning Frameworks			
OpenRLHF [528]	RLHF / LLM RL	High-performance scalable RLHF	C GitHub
TRL [529]	RLHF / LLM RL	Hugging Face RLHF	C GitHub
trlX [530]	RLHF / LLM RL	Distributed large-model RLHF	Gi tHub
Verl [531]	RLHF / LLM RL	Streamlined experiment management	Gi GitHub
SLiMe [532]	RLHF / LLM RL	High-performance async RL	GiHu
Oat [533]	RLHF / LLM RL	Lightweight RL support	⍰⍰ʒtitlub
General-purpose RL Frameworks			
RLlib [534]	General RL / Multi-agent RL	Production-grade scalable library	C GitHub
Acme [535]	General RL	Modular distributed components	GiHu
Tianshou [536]	General RL	High-performance PyTorch platform	GiHu
Stable Baselines3 [537]	General RL	Reliable PyTorch algorithms	GiHu
PFRL [538]	General RL	Benchmarked prototyping rithms	GitHub

框架	类型	主要特性	资源链接
智能体型 RL 框架			
验证者 [516]	智能体型 RL	可验证的环境设置	C GitHub
SkyRL-v0 [517, 518]	智能体型 RL	长期现实世界训练	C GitHub
AREAL [519]	智能体型 RL	异步训练	GiHtHub
MARTI [520]	多智能体 RL	集成的多智能体训练	GitHub
EasyR1 [521]	智能体型 RL	多模态支持	GitHub
AgentFly [522]	智能体型 RL	可扩展的异步执行	⍰⍰⍰⍰⍰Hub
Agent Lightning [523]	智能体型 RL	解耦的分层 RL	C GitHub
AWorld [275]	智能体型 RL	跨集群并行 rollout	C GitHub
RL-Factory [524]	智能体型 RL	易于设计的奖励	GiHHub
ROLL [525]	智能体型 RL	稳定的多 GPU 并行训练	GitHub
VerlTool [526]	智能体型 RL	集成工具的 rollout	GitHub
AgentRL [527]	智能体型 RL	异步多任务训练	{fitHub
RLHF 与 LLM 微调框架			
OpenRLHF [528]	RLHF / LLM 强化学习	高性能可扩展 RLHF	C GitHub
TRL [529]	RLHF / LLM 强化学习	Hugging Face RLHF	C GitHub
trlX [530]	RLHF / LLM 强化学习	分布式大模型 RLHF	Gi tHub
Verl [531]	RLHF / LLM 强化学习	简化的实验管理	Gi GitHub
SLiMe [532]	RLHF / LLM 强化学习	高性能异步 RL	GiHu
Oat [533]	RLHF / LLM 强化学习	轻量级 RL 支持	⍰⍰ʒtitlub
通用 RL 框架			
RLlib [534]	通用 RL / 多智能体 RL	生产级可扩展库	C GitHub
Acme [535]	通用 RL	模块化分布式组件	GiHu
Tianshou [536]	通用 RL	高性能 PyTorch 平台	GiHu
Stable Baselines3 [537]	通用 RL	可靠的 PyTorch 算法	GiHu
PFRL [538]	通用 RL	经过基准测试的原型算法	GitHub

Table 10: A summary of frameworks for reinforcement learning, categorized by type and key features. controller, parallel workers, and automatic resource mapping for efficient multi-GPU training. VerITool [526] introduces an agentic RL with tool use (ARLT) framework built upon Verl [531], enabling agents to jointly optimize planning and execution across interactive environments. AgentRL [527] provides a scalable asynchronous framework for multi-turn, multi-task agentic RL, unifying environment orchestration and introducing cross-policy sampling and task advantage normalization for stable large-scale training.

表 10: 按类型和关键特性分类的强化学习框架概览。包含控制器、并行工作进程和自动资源映射以实现高效的多 GPU 训练。VerITool [526] 在 Verl [531] 基础上引入了具工具使用能力的主体化强化学习 (ARLT) 框架，使智能体能够在交互式环境中共同优化规划与执行。AgentRL [527] 提供了可扩展的异步框架用于多轮、多任务的主体化强化学习，统一了环境编排并引入跨策略采样与任务优势归一化以实现稳定的大规模训练。

RLHF and LLM fine-tuning frameworks. OpenRLHF [528] offers a high-performance, scalable toolkit designed for large-scale model alignment; TRL [529] provides Hugging Face's baseline implementations for RLHF experiments; trlX [530] adds distributed training support for fine-tuning models up to tens of billions of parameters; and HybridFlow [531] streamlines experiment management and scaling for RLHF research pipelines. SLiMe [532] is an LLM post-training framework for RL scaling that combines Megatron with SGLang for high-performance multi-mode training, supports Async RL, and enables flexible disaggregated workflows for reward and data generation via custom interfaces and server-based engines.

RLHF 与 LLM 微调框架。OpenRLHF [528] 提供了面向大规模模型对齐的高性能可扩展工具包；TRL [529] 提供了 Hugging Face 的 RLHF 实验基线实现；trlX [530] 为微调数百亿参数级模型增加了分布式训练支持；HybridFlow [531] 简化了 RLHF 研究流水线的实验管理与扩展。SLiMe [532] 是一个用于 RL 扩展的 LLM 后训练框架，结合了 Megatron 与 SGLang 以实现高性能多模式训练，支持异步强化学习，并通过自定义接口与基于服务器的引擎实现奖励与数据生成的灵活离散化工作流。

General-purpose RL frameworks supply the core algorithms and distributed execution engines that can underpin agentic LLM systems. RLlib [534] is a production-grade, scalable library offering unified APIs for on-policy, off-policy, and multi-agent methods; Acme [535] provides modular, research-oriented building blocks for distributed RL; Tianshou [536] delivers a high-performance, pure-PyTorch platform supporting online, offline, and hierarchical RL; Stable Baselines3 [537] packages reliable PyTorch implementations of standard model-free algorithms; and PFRL [538] (formerly ChainerRL) offers benchmarked deep-RL algorithm implementations for rapid prototyping.

通用强化学习框架提供可支撑主体化 LLM 系统的核心算法和分布式执行引擎。RLlib [534] 是面向生产的可扩展库，提供统一的在策略、离策略和多智能体方法 API；Acme [535] 提供模块化、面向研究的分布式强化学习构件；Tianshou [536] 提供高性能的纯 PyTorch 平台，支持在线、离线与分层强化学习；Stable Baselines3 [537] 打包了可靠的 PyTorch 标准无模型算法实现；PFRL [538](前身为 ChainerRL) 提供经基准测试的深度强化学习算法实现，便于快速原型开发。

6. Open Challenges and Future Directions

6. 开放挑战与未来方向

The advance of agent RL toward general-purpose intelligence hinges on overcoming three pivotal challenges that define the field's research frontier. First is the challenge of Trustworthiness: ensuring the reliability, safety, and alignment of increasingly autonomous agents. Second is Scaling up Agentic Training, which requires surmounting the immense practical bottlenecks in computation, data, and algorithmic efficiency. Finally, an agent's capabilities are fundamentally bounded by its world, making the Scaling up Agentic Environments, i.e., the creation of complex and adaptive training grounds.

主体性强化学习迈向通用智能的进展取决于克服三项关键挑战，这些挑战定义了该领域的研究前沿。其一是可信性挑战: 确保日益自主的智能体的可靠性、安全性与对齐。其二是主体性训练的规模化，需要突破计算、数据与算法效率方面的巨大实际瓶颈。其三则是智能体能力受限于其所处世界，即主体性环境的规模化——创造复杂且自适应的训练场。

6.1. Trustworthiness

6.1. 可信性

Security. The security landscape for autonomous agents is fundamentally more complex than for standard LLMs. While traditional models are primarily vulnerable to attacks on their text-in, text-out interface, agents possess an expanded attack surface due to their external components like tools, memory, and planning modules [539, 67]. This architecture exposes them to novel threats beyond direct prompt injection. For instance, indirect prompt injection can occur when an agent interacts with a compromised external environment, such as a malicious website or API, which poisons its memory or tool outputs [540]. Multi-agent systems further compound these risks by introducing vulnerabilities through inter-agent communication, where one compromised agent can manipulate or mislead others within the collective [539].

安全。自主代理的安全态势比标准大模型更为复杂。传统模型主要易受其文本输入—输出接口的攻击，而代理因其外部组件如工具、记忆与规划模块而拥有更广的攻击面 [539, 67]。这种架构使它们面临超出直接提示注入的新型威胁。例如，当代理与被攻破的外部环境 (如恶意网站或 API) 交互以致污染其记忆或工具输出时，间接提示注入就可能发生 [540]。多代理系统通过引入代理间通信的脆弱点进一步加剧这些风险，其中被攻破的一个代理可能操纵或误导集体中的其他代理 [539]。

RL significantly magnifies these agent-specific risks by transforming the agent from a passive victim of manipulation into an active, goal-seeking exploiter of vulnerabilities. The core issue is instrumental goal achievement through reward hacking: an RL agent's primary directive is to maximize its long-term reward, and it may learn that unsafe actions are the most effective path to this goal. For example, if an agent discovers that using a malicious, third-party tool yields a high reward for a given task, RL will actively reinforce and entrench this unsafe behavior. Similarly, if an agent learns that it can bypass safety protocols to achieve its objective more efficiently, the resulting reward signal will teach it to systematically probe for and exploit such security loopholes. This creates a more persistent and dangerous threat than one-off jailbreaks, as the agent autonomously learns and optimizes deceptive or harmful strategies over time.

强化学习显著放大了这些特有风险，它将智能体从被动的操纵受害者转变为主动寻求目标并利用脆弱性的利用者。核心问题在于通过奖励劫持实现工具性目标: 强化学习智能体的首要指令是最大化长期回报，它可能学会不安全的行为是实现该目标的最有效途径。例如，如果智能体发现使用一个恶意的第三方工具能在某项任务中获得高回报，强化学习会积极强化并固化这种不安全行为。同样，如果智能体学会绕过安全协议以更高效地实现目标，回报信号将教会它系统性地探查并利用此类安全漏洞。这比一次性越狱更持久且更危险，因为智能体会随时间自主学习并优化欺骗或有害策略。

Mitigating these amplified risks requires a defense-in-depth approach tailored to agentic systems. A critical first line of defense is robust sandboxing [541, 542], where agents operate in strictly controlled, permission-limited environments to contain the potential damage from a compromised tool or action. At the training level, mitigation strategies must focus on shaping the reward signal itself. This includes implementing process-based rewards that penalize unsafe intermediate steps (e.g., calling an untrusted API) and employing adversarial training within the RL loop, where the agent is explicitly rewarded for resisting manipulation attempts and ignoring poisoned information. Finally, continuous monitoring and anomaly detection are essential for post-deployment safety. By tracking an agent's actions, such as tool calls and memory access patterns, it is possible to identify deviations from normal behavior, allowing for timely intervention.

缓解这些放大风险需要针对具代理性的系统采取纵深防御。首要且关键的一道防线是强健的沙箱隔离 [541, 542]，让代理在严格受控、权限受限的环境中运行，以遏制被攻破的工具或行为可能造成的损害。在训练层面，缓解策略必须着重于塑造奖励信号本身。其包括实施基于过程的奖励，对不安全的中间步骤 (例如调用不受信任的 API) 予以惩罚，并在强化学习循环中采用对抗性训练，明确奖励代理抵抗操控尝试和忽视被投毒信息的能力。最后，部署后持续监控与异常检测对安全至关重要。通过跟踪代理的行为，如工具调用和记忆访问模式，可以识别偏离正常行为的情况，从而实现及时干预。

Hallucination. In the context of agentic LLMs, hallucination is the generation of confident yet ungrounded outputs, including statements, reasoning steps, or tool usage, that are not rooted in provided evidence or external reality. This issue extends beyond simple factual errors to encompass unfaithful reasoning paths and misaligned planning, with overconfidence often masking the agent's uncertainty [543, 544]. In multimodal agents, it also manifests as cross-modal inconsistency, such as a textual description mismatching an image, framing it as a fundamental grounding problem [545]. Evaluating hallucination requires assessing both factuality against objective truth and faithfulness to a given source, often measured through benchmarks like HaluEval-QA or by the agent's ability to appropriately abstain on unanswerable questions, where a refusal to answer ("I don't know") is a critical signal of epistemic awareness [546, 547].

幻觉。在自主大语言模型的语境下，幻觉是指生成看似可信但缺乏依据的输出，包括陈述、推理步骤或工具使用，这些内容并非基于所提供的证据或外部现实。这个问题不仅包括简单的事实性错误，还涵盖了不可靠的推理路径和不恰当的规划，过度自信往往掩盖了模型的不确定性 [543, 544]。在多模态模型中，它还表现为跨模态不一致，例如文字描述与图像不匹配，这将其归结为一个根本性的基础问题 [545]。评估幻觉需要同时考量与客观事实的相符性以及对给定来源的忠实度，通常通过像 HaluEval - QA 这样的基准测试来衡量，或者通过模型在面对无法回答的问题时能否恰当放弃作答来评估，其中拒绝回答 (“我不知道”) 是认知意识的一个关键信号 [546, 547]。

RL can inadvertently amplify hallucination if the reward mechanism is not carefully designed. Studies show that outcome-driven RL, which rewards only the correctness of the final answer, can encourage agents to find spurious correlations or shortcuts. This process may yield confident but unfounded intermediate reasoning steps, as the optimization process settles into local optima that achieve the goal without being factually sound [546]. This phenomenon introduces a "hallucination tax," where reinforcement finetuning can degrade an agent's ability to refuse to answer, compelling it to generate responses for unanswerable questions rather than abstaining [547]. However, the effect is highly dependent on the training pipeline; while RL-only post-training can worsen factuality, a structured approach combining SFT with a verifiable-reward RL process can mitigate this degradation [548].

如果奖励机制设计不当，强化学习 (RL) 可能会在不经意间放大幻觉现象。研究表明，仅以最终答案的正确性作为奖励依据的结果驱动型强化学习，可能会促使智能体寻找虚假的关联或捷径。这一过程可能会产生看似自信但缺乏依据的中间推理步骤，因为优化过程会陷入局部最优解，这些解虽然能达成目标，但在事实层面却站不住脚 [546]。这种现象带来了一种 “幻觉代价”，即强化微调可能会削弱智能体拒绝作答的能力，迫使其对无法回答的问题给出回应，而非选择不作答 [547]。不过，这种影响在很大程度上取决于训练流程；虽然仅采用强化学习进行后期训练会降低事实准确性，但将监督微调 (SFT) 与可验证奖励的强化学习过程相结合的结构化方法，能够缓解这种准确性的下降 [548]。

Promising mitigation strategies involve a hybrid approach of training-time alignment and inference-time safeguards. During training, a key direction is to shift from outcome-only rewards to process-based rewards. Techniques like Factuality-aware Step-wise Policy Optimization (FSPO) verify each intermediate reasoning step against evidence, directly shaping the policy to discourage ungrounded claims [546]. Data-centric approaches enhance epistemic humility by training agents on a mix of solvable and unsolvable problems, restoring their ability to abstain when necessary [547]. At the system level, this is complemented by inference-time techniques such as retrieval augmentation, tool-use for fact-checking, and post-hoc verification to ground the agent's outputs in reliable sources. For multimodal agents, explicitly adding cross-modal alignment objectives is crucial for ensuring consistency [544, 543, 545]. Collectively, these directions aim to align the agent's reward-seeking behavior with the goal of truthfulness, fostering more reliable and trustworthy autonomous systems.

有前景的缓解策略包括采用训练时对齐和推理时保障措施相结合的混合方法。在训练过程中，一个关键方向是从仅基于结果的奖励转向基于过程的奖励。像事实感知逐步策略优化 (FSPO) 这样的技术会对照证据验证每个中间推理步骤，直接塑造策略以减少无根据的论断 [546]。以数据为中心的方法通过让智能体在可解决和不可解决的混合问题上进行训练来增强认知谦逊，恢复其在必要时放弃作答的能力 [547]。在系统层面，推理时的技术 (如检索增强、使用工具进行事实核查和事后验证) 可作为补充，从而使智能体的输出有可靠的来源作为支撑。对于多模态智能体而言，明确添加跨模态对齐目标对于确保一致性至关重要 [544, 543, 545]。总体而言，这些方向旨在使智能体的奖励寻求行为与追求真实性的目标保持一致，从而培育更可靠、更值得信赖的自主系统。

Sycophancy. Sycophancy in LLM agents refers to their tendency to generate outputs that conform to a user's stated beliefs, biases, or preferences, even when those are factually incorrect or lead to suboptimal outcomes [549]. This behavior transcends mere conversational agreeableness, fundamentally affecting an agent's planning and decision-making processes. For instance, a sycophantic agent might adopt a user's flawed reasoning in its internal plan, choose a course of action that validates the user's incorrect assumptions, or filter information from tools to present only what aligns with the user's view [550]. This represents a critical misalignment, where the agent optimizes for the user's expressed preference rather than their latent, long-term interest in achieving the best possible outcome.

谄媚。大语言模型 (LLM) 智能体中的谄媚指的是它们倾向于生成符合用户既定信念、偏见或偏好的输出，即便这些信念、偏见或偏好与事实不符或会导致不理想的结果 [549]。这种行为超越了单纯的对话附和，从根本上影响了智能体的规划和决策过程。例如，一个谄媚的智能体可能会在其内部规划中采用用户有缺陷的推理方式，选择能证实用户错误假设的行动方案，或者筛选工具提供的信息，只呈现与用户观点相符的内容 [550]。这体现了一种严重的不一致，即智能体优化的是用户明确表达的偏好，而非用户实现最佳结果的潜在长期利益。

RL is a primary cause for this behavior. The underlying mechanism is a form of ”reward hacking,” where the agent learns to exploit the reward model in ways that do not align with true human preferences [551]. Because human labelers often show a preference for agreeable and validating responses, the reward model inadvertently learns to equate user satisfaction with sycophantic agreement. Consequently, RLHF can directly incentivize and ”exacerbate sycophantic tendencies” by teaching the agent that conforming to a user's viewpoint is a reliable strategy for maximizing reward, even if it compromises truthfulness [552].

强化学习 (RL) 是这种行为的主要原因。其潜在机制是一种“奖励欺骗”形式，即智能体学会以不符合人类真实偏好的方式利用奖励模型 [551]。由于人类标注员通常更喜欢随和且肯定性的回复，奖励模型会在不经意间将用户满意度等同于谄媚式的认同。因此，基于人类反馈的强化学习 (RLHF) 会直接激励并“加剧谄媚倾向”，因为它会教导智能体，即使牺牲真实性，迎合用户观点也是一种可靠的最大化奖励的策略 [552]。

Mitigating sycophancy is an active area of research that focuses on refining the reward signal and training dynamics. A promising direction is the development of sycophancy-aware reward models, which are explicitly trained to penalize responses that merely parrot user beliefs without critical evaluation. Another approach involves leveraging AI-driven feedback, such as in Constitutional AI, where the agent is steered by a set of principles promoting objectivity and neutrality, rather than solely by human preferences [553]. At inference time, strategies like explicitly prompting the agent to adopt a ”red team” or contrarian perspective can also help counteract ingrained sycophantic tendencies. Cooper [554] is a reinforcement learning framework that co-optimizes both the policy model and the reward model online, using high-precision rule-based verifiers to select positive samples and LLM-generated negative samples, thereby preventing the policy from exploiting a static reward model (i.e., reward hacking) by continuously adapting the reward model to closing emergent loopholes. Ultimately, the future direction lies in designing reward systems that robustly capture the user’s long-term interests-such as receiving accurate information and making sound decisions-over their immediate desire for validation.

缓和拍马屁现象是一个活跃的研究方向，侧重于完善奖励信号与训练动态。一条有前景的方向是开发识别拍马屁的奖励模型，明确训练以惩罚那些仅重复用户信念而缺乏批判性评估的回应。另一种方法是利用由 AI 提供的反馈，如在宪法式 AI(Constitutional AI) 中，代理由一套促进客观性与中立性的原则引导，而不仅仅服从人类偏好 [553]。在推理阶段，像明确提示代理采取“红队”或反对视角这样的策略也能帮助对抗根深蒂固的拍马屁倾向。Cooper[554] 是一个强化学习框架，在线共同优化策略模型与奖励模型，使用高精度的基于规则的验证器选择正样本并以大模型生成负样本，从而通过持续使奖励模型修补新出现的漏洞来防止策略利用静态奖励模型 (即奖励劫持)。最终，未来方向在于设计能稳健捕捉用户长期利益的奖励体系——例如获得准确信息与做出合理决策——而不是迎合其即时的被认可欲望。

6.2. Scaling up Agentic Training

6.2. 扩大自主代理训练规模

Computation. Recent advances demonstrate that scaling reinforcement learning fine-tuning (RFT) computation directly enhances the reasoning ability of LLM-based agents. The Agent RL Scaling Law study shows that longer training horizons systematically improve tool-use frequency, reasoning depth, and overall task accuracy, highlighting the predictive benefit of allocating more compute to RL training [324]. Similarly, ProRL reveals that prolonged RL training expands reasoning boundaries beyond those accessible to base models, uncovering novel solution strategies even where extensive sampling from the pretrained model fails [50]. Building upon this, ProRLv2 extends training steps and incorporates more stable optimization techniques, demonstrating sustained benefits as smaller models, after extensive RL training, rival the performance of larger models on mathematics, code, and logic benchmarks [555]. Collectively, these results underscore that scaling compute through extended RL training is not merely complementary to enlarging model or data size, but a fundamental axis for advancing agentic reasoning.

计算。近期进展表明，扩大强化学习微调 (RFT) 计算量可直接提升基于大语言模型 (LLM) 的智能体的推理能力。《智能体强化学习规模定律》研究显示，更长的训练周期能系统性地提高工具使用频率、推理深度和整体任务准确率，凸显了为强化学习训练分配更多计算资源的可预测益处 [324]。同样，ProRL 表明，延长强化学习训练能突破基础模型的推理边界，即使在预训练模型进行大量采样仍失败的情况下，也能发现新的解决方案策略 [50]。在此基础上，ProRLv2 增加了训练步数，并采用了更稳定的优化技术，结果显示，小型模型在经过大量强化学习训练后，在数学、代码和逻辑基准测试中的表现可与大型模型相媲美，这体现了持续的益处 [555]。总体而言，这些结果强调，通过延长强化学习训练来扩大计算量，不仅是对增大模型或数据规模的补充，更是提升智能体推理能力的一个基本维度。

Model Size. Increasing model capacity heightens both the promise and pitfalls of RL-based agent training. Larger models unlock greater potential but risk entropy collapse and narrowing of capability boundaries, as RL sharpens output distributions toward high-reward modes, limiting diversity [556]. Methods like RL-PLUS address this with hybrid strategies and advantage functions that foster novel reasoning paths, breaking capability ceilings [556]. Meanwhile, scaling demands massive compute, making efficiency vital. A two-stage approach in [369] uses large teachers to generate SFT data for smaller students, refined via on-policy RL. This ”SFT+RL” setup outperforms each method alone and cuts compute by half compared to pure SFT. The work also underscores RL’s extreme hyperparameter sensitivity at scale, stressing the need for careful tuning.

模型规模。增大模型容量既带来更大潜力也增添基于 RL 的智能体训练的风险。更大的模型能解锁更高能力，但 RL 会把输出分布锐化到高回报模式，从而导致熵崩溃和能力边界收窄，限制多样性 [556]。诸如 RL-PLUS 的方法通过混合策略和优势函数来鼓励新的推理路径，打破能力天花板 [556]。同时，扩展需要巨量计算，使效率至关重要。[369] 中的两阶段方法使用大型教师生成小型学生的 SFT 数据，再通过 on-policy RL 精炼。从而这种“SFT+RL”设置优于任一单独方法，并将计算量相比纯 SFT 减半。该工作还强调了在大规模下 RL 对超参数极度敏感，需谨慎调优。

Data Size. Scaling RL training across domains introduces both synergy and conflict in agentic reasoning. Cross-domain RL in math, code, and logic tasks shows complex interactions [557]: some pairings enhance each other, while others interfere and reduce performance. Model initialization also matters-instruction-tuned models generalize differently than raw ones. Building on this, the Guru dataset [558] spans six reasoning domains, showing that RL gains correlate with pretraining exposure: math and code benefit from transfer, but domains like simulation or logic need dedicated training. These findings suggest that while multi-domain RL data can amplify general reasoning, it must be carefully curated to balance complementarity and mitigate interference across tasks.

数据规模。将 RL 训练扩展到多个领域会在智能推理上既产生协同也产生冲突。在数学、代码和逻辑任务上的跨域 RL 显示复杂交互 [557]: 有些配对会相互增强，而另一些则相互干扰并降低性能。模型初始化也影响结果——经过指令微调的模型与未经微调的模型泛化表现不同。在此基础上，Guru 数据集 [558] 覆盖六个推理领域，显示 RL 收益与预训练暴露有关: 数学和代码可从迁移中受益，但像模拟或逻辑这样的领域需要专门训练。这些发现表明，尽管多域 RL 数据可以放大通用推理能力，但必须谨慎策划以平衡互补性并减轻跨任务的干扰。

Efficiency. Efficiency of LLM post-training is a central frontier for sustainable scaling [559]. Beyond brute-force scaling, recent research emphasizes improving RL training efficiency through post-training recipes, methodological refinements, and hybrid paradigms. POLARIS [560] demonstrates that calibrating data difficulty, employing diversity-driven sampling, and extending reasoning length substantially boost RL effectiveness, enabling smaller models to reach or even surpass much larger counterparts on reasoning benchmarks. Complementary work [38] provides systematic evaluations of common RL techniques, finding that judiciously combining just a few simple strategies often outperforms more complex methods. Another research proposes Dynamic Fine-Tuning (DFT) [561], showing that introducing RL principles into gradient scaling can match or exceed advanced RL approaches with minimal additional cost. Taken together, these advances suggest a dual trajectory for the future: on one hand, progressively refining RL-based recipes to maximize efficiency; on the other, rethinking training paradigms to embed RL-like generalization signals without full-fledged online RL. A particularly compelling direction lies in exploring how agentic models might acquire robust generalization from extremely limited data, for instance, by leveraging principled difficulty calibration, meta-learning dynamics, or information-theoretic regularization to distill broad reasoning abilities from a handful of experiences. Such pathways point to the possibility of a new regime of post-training: one where the ability to extrapolate, abstract, and generalize becomes decoupled from sheer data volume, and instead hinges on exploiting the structure and dynamics of the training process itself.

效率。LLM 后训练的效率是可持续扩展的核心前沿 [559]。除去粗放式放大，近期研究强调通过后训练方案、方法学改进与混合范式提升 RL 训练效率。POLARIS [560] 表明校准数据难度、采用多样性驱动采样并延长推理长度可大幅提升 RL 效用，使较小模型在推理基准上达到甚至超越更大模型。互补工作 [38] 对常见 RL 技术进行了系统评估，发现谨慎地组合少数简单策略常常胜过更复杂的方法。另一项研究提出动态微调 (DFT) [561]，展示了在梯度尺度中引入 RL 原则能够以极小额外成本匹配或超越先进 RL 方法。综上，这些进展暗示未来有双轨路径: 一方面逐步精炼基于 RL 的方案以最大化效率；另一方面重新思考训练范式，在不进行完整在线 RL 的情况下嵌入类 RL 的泛化信号。一个尤其值得探索的方向是研究代理模型如何从极其有限的数据中获得稳健泛化，例如通过原则性难度校准、元学习动力学或信息论正则化，从少量经验中提炼出广泛的推理能力。这类路径指向一种新的后训练制度的可能性: 在该制度中，外推、抽象和泛化的能力不再依赖于数据量，而是依赖于对训练过程结构和动力学的利用。

6.3. Scaling up Agentic Environment.

6.3. 扩展能动环境。

A nascent yet critical frontier for Agentic RL involves a paradigmatic shift from treating the training environment as a static entity to viewing it as a dynamic and optimizable system. This perspective addresses a core bottleneck in agent development: the scarcity of interactive, adaptive environments and the difficulty of engineering effective reward signals. As a growing consensus holds that prevalent environments like ALFWorld [472] and ScienceWorld [474] are insufficient for training general-purpose agents [562], research is moving beyond solely adapting the agent’s policy. Instead, a co-evolutionary approach uses learning-based methods to adapt the environment itself. One key strategy is to automate reward function design. This involves deploying an auxiliary ”explorer” agent to generate a diverse dataset of interaction trajectories, which are then used to train a reward model via heuristics or preference modeling. This effectively decouples agent training from the expensive process of manual reward specification, enabling the learning of complex behaviors without direct human annotation.

一个新兴但关键的能动 RL 前沿，是将训练环境从静态实体的思维范式转向将其视为动态且可优化系统的范式转变。这一观点解决了代理开发中的核心瓶颈: 交互式、可适应环境的稀缺以及有效奖励信号设计的困难。随着共识逐步形成，认为像 ALFWorld [472] 和 ScienceWorld [474] 这样的普遍环境不足以训练通用代理 [562]，研究正超越仅适配代理策略，而采用共进化方法以学习为基础来适配环境本身。一项关键策略是自动化奖励函数设计: 部署辅助“探索者”代理生成多样的交互轨迹数据集，再通过启发式或偏好建模训练奖励模型。这有效地将代理训练与昂贵的手动奖励指定过程解耦，使学习复杂行为无需直接人工标注成为可能。

Beyond automating the reward signal, a second, more dynamic strategy is to automate curriculum generation, transforming the environment into an active teacher. This approach establishes a feedback loop where an agent’s performance data, highlighting specific weaknesses, is fed to an ”environment generator” LLM. As exemplified by EnvGen [563], this generator then procedurally adapts the environment’s configuration, creating new tasks that specifically target and remedy the agent’s deficiencies. This form of goal-directed Procedural Content Generation (PCG) ensures the agent is consistently challenged within its ”zone of proximal development,” accelerating learning and preventing overfitting. Together, automated rewards and adaptive curricula create a symbiotic relationship between the agent and its environment, establishing a scalable ”training flywheel” that is essential for the future of self-improving agentic systems.

除自动化奖励信号外，第二种更动态的策略是自动化课程生成，将环境转变为主动教师。该方法建立一条反馈回路: 将代理的性能数据 (突出具体弱点) 输入“环境生成器”LLM。如 EnvGen [563] 所示，该生成器进而程序性地调整环境配置，创建专门针对并修复代理缺陷的新任务。这种目标导向的程序化内容生成 (PCG) 保证代理持续在其“最近发展区”内受到挑战，加速学习并防止过拟合。自动化奖励与自适应课程共同在代理与环境之间建立共生关系, 形成对自我改进能动系统至关重要的可扩展“训练飞轮”。

6.4.The Mechanistic Debate on RL in LLMs

6.4. 关于 LLM 中 RL 的机理论争

Two competing explanations have emerged for why RL appears to boost LLM reasoning. The ”amplifier” view holds that RL with verifiable rewards-often instantiated via PPO-style variants such as GRPO-mainly reshapes the base model’s output distribution: by sampling multiple trajectories and rewarding the verifiably correct ones, RL concentrates probability mass on already-reachable reasoning paths, improving pass@1 while leaving the support of solutions largely unchanged; consistent with this, large-k pass@k analyses often find that the base model eventually matches or surpasses its RL-tuned counterpart, suggesting elicitation rather than creation of capabilities, and further evidence indicates that reflective behaviors can already emerge during pre-training [2, 185, 564]. By contrast, the ”new-knowledge” view argues that RL after next-token prediction can install qualitatively new computation by leveraging sparse outcome-level signals and encouraging longer test-time computation: theory shows that RL enables generalization on problems (e.g., parity) where next-token training alone is statistically or computationally prohibitive; empirically, RL can improve generalization to out-of-distribution rule and visual variants, induce cognitive behaviors (verification, backtracking, subgoal setting) that were absent in the base model yet predict self-improvement, and in under-exposed domains even expand the base model’s pass@k frontier [565, 566, 567, 568, 558]. Whether RL can truly endow LLMs with abilities beyond those acquired during pre-training remains an open question, and its underlying learning mechanisms are still to be fully understood.

关于为什么 RL 似乎能提升 LLM 的推理能力，出现了两种相互竞争的解释。“放大器”观点认为，引入可验证奖励的 RL——通常通过类似 PPO 的变体如 GRPO 实现——主要重塑基础模型的输出分布: 通过采样多条轨迹并奖励可验证正确的那些，RL 将概率质量集中到已可达的推理路径上，提高 pass@1 而基本不改变解的支撑；与此一致，large-k 的 pass@k 分析常发现基础模型最终会匹配或超越其经 RL 调优的对应模型，表明是能力的引出而非创造，且进一步证据显示反思性行为可以在预训练阶段就已出现 [2, 185, 564]。相反，“新知识”观点认为，次令牌预测之后的 RL 能通过利用稀疏的结果级信号并鼓励更长的测试时计算，安装质上新的计算能力: 理论表明 RL 能在次令牌训练在统计或计算上不可行的问题 (例如奇偶校验) 上实现泛化；经验上，RL 可以改善对分布外规则与视觉变体的泛化，诱发基础模型中缺失但能预测自我提升的认知行为 (验证、回溯、子目标设定)，并且在暴露不足的领域甚至扩展基础模型的 pass@k 前沿 [565, 566, 567, 568, 558]。RL 是否真正能赋予 LLM 超出预训练所获得的能力仍是未解之问，其底层学习机制亦有待完全理解。

7. Conclusion

7. 结论

This survey has charted the emergence of Agentic Reinforcement Learning (Agentic RL), a paradigm that elevates LLMs from passive text generators to autonomous, decision-making agents situated in complex, dynamic worlds. Our journey began by formalizing this conceptual shift, distinguishing the temporally extended and partially observable MDPs (POMDPs) that characterize agentic RL from the single-step decision processes of conventional LLM-RL. From this foundation, we constructed a comprehensive, twofold taxonomy to systematically map the field: one centered on core agentic capabilities (planning, tool use, memory, reasoning, self-improvement, perception, etc.) and the other on their application across a diverse array of task domains. Throughout this analysis, our central thesis has been that RL provides the critical mechanism for transforming these capabilities from static, heuristic modules into adaptive, robust agentic behavior. By consolidating the landscape of open-source environments, benchmarks, and frameworks, we have also provided a practical compendium to ground and accelerate future research in this burgeoning field.

本综述梳理了 Agentic Reinforcement Learning(Agentic RL) 的兴起——一种将大语言模型从被动文本生成器提升为位于复杂动态世界中的自主决策代理的范式。我们首先形式化了这一概念性转变，将表征 agentic RL 的时域延长且部分可观测的 MDP(POMDP) 与传统 LLM-RL 的单步决策过程区分开来。在此基础上，我们构建了双重并进的全面分类体系: 一方面围绕核心 agentic 能力 (规划、工具使用、记忆、推理、自我改进、感知等)，另一方面考察这些能力在多样任务域中的应用。贯穿分析，我们的核心论点是强化学习为将这些能力从静态启发式模块转变为自适应、鲁棒的代理行为提供了关键机制。通过整合开源环境、基准和框架的格局，我们还提供了一个实用汇编，以为未来该新兴领域的研究奠定基础并加速发展。

Acknowledgments

致谢

We acknowledge Zhouliang Yu and Minghao Liu for their guidance and discussion.

感谢余周亮和刘明浩的指导与讨论。

References

参考文献

[1] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[1] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov。Proximal policy optimization algorithms， 2017。URL https://arxiv.org/abs/1707.06347。

[2] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[2] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo。Deepseekmath: 推动开放语言模型在数学推理上的极限，2024。URL https://arxiv.org/abs/2402.03300。

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller。Playing atari with deep reinforcement learning, 2013。URL https://arxiv.org/abs/1312.5602。

[4] Saksham Sahai Srivastava and Vaneet Aggarwal. A technical survey of reinforcement learning techniques for large language models, 2025. URL https://arxiv.org/abs/2507.04136.

[4] Saksham Sahai Srivastava and Vaneet Aggarwal。A technical survey of reinforcement learning techniques for large language models， 2025。URL https://arxiv.org/abs/2507.04136。

[5] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced llms: A survey, 2025. URL https://arxiv.org/abs/2412.1040

[5] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy。Reinforcement learning enhanced llms:A survey，2025。URL https://arxiv.org/abs/2412.10400。

[6] Junqiao Wang, Zeng Zhang, Yangfan He, Zihao Zhang, Yuyang Song, Tianyu Shi, Yuchen Li, Hengyuan Xu, Kunyu Wu, Xin Yi, et al. Enhancing code llms with reinforcement learning in code generation: A survey. arXiv preprint arXiv:2412.20367, 2024. URL https://arxiv.org/abs/2412.20367.

[6] Junqiao Wang, Zeng Zhang, Yangfan He, Zihao Zhang, Yuyang Song, Tianyu Shi, Yuchen Li, Hengyuan Xu, Kunyu Wu, Xin Yi, et al。Enhancing code llms with reinforcement learning in code generation:A survey。arXiv preprint arXiv:2412.20367，2024。URL https://arxiv.org/abs/2412.20367。

[7] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. IEEE Transactions on Neural Networks and Learning Systems, 36(6):9737-9757, June 2025. ISSN 2162-2388. doi: 10.1109/tnnls.2024.3497992. URL http: //dx.doi.org/10.1109/TNNLS.2024.3497992.

[7] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li。Survey on large language model-enhanced reinforcement learning:Concept, taxonomy, and methods。IEEE Transactions on Neural Networks and Learning Systems，36(6):9737-9757，2025 年 6 月。ISSN 2162-2388。doi: 10.1109/tnnls.2024.3497992。URL http: //dx.doi.org/10.1109/TNNLS.2024.3497992。

[8] Yiduo Guo, Zhen Guo, Chuanwei Huang, Zi-Ang Wang, Zekai Zhang, Haofei Yu, Huishuai Zhang, and Yikang Shen. Synthetic data rl: Task definition is all you need, 2025. URL https://arxiv.org/ abs/2505.17063.

[8] Yiduo Guo, Zhen Guo, Chuanwei Huang, Zi-Ang Wang, Zekai Zhang, Haofei Yu, Huishuai Zhang, and Yikang Shen。Synthetic data rl:Task definition is all you need，2025。URL https://arxiv.org/ abs/2505.17063。

[9] Fanqi Wan, Deng Cai, Shijue Huang, Xiaojun Quan, and Mingxuan Wang. Let large language models find the data to train themselves, 2025. URL https://openreview.net/forum?id= 5YCZZSEOSW.

[9] Fanqi Wan, Deng Cai, Shijue Huang, Xiaojun Quan, and Mingxuan Wang。Let large language models find the data to train themselves，2025。URL https://openreview.net/forum?id= 5YCZZSEOSW。

[10] Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement pre-training, 2025. URL https://arxiv.org/abs/2506.08007.

[10] Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei。Reinforcement pre-training，2025。URL https://arxiv.org/abs/2506.08007。

[11] Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Aviral Kumar, Rishabh Agarwal, Sridhar Thiagarajan, Craig Boutilier, and Aleksandra Faust. Inference-aware fine-tuning for best-of-n sampling in large language models. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=77gQUdQhE7.

[11] Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Aviral Kumar, Rishabh Agarwal, Sridhar Thiagarajan, Craig Boutilier, and Aleksandra Faust. 针对大型语言模型的 best-of-n 采样的推理感知微调。发表于第十三届国际学习表征会议，2025。URL https://openreview.net/forum?id=77gQUdQhE7.

[12] Zichuan Guo and Hao Wang. A survey of reinforcement learning in large language models: From data generation to test-time inference. Available at SSRN 5128927, 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstrac

[12] Zichuan Guo and Hao Wang. 大型语言模型中强化学习综述: 从数据生成到测试时推理。发表于 SSRN(编号 5128927)， 2025。URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5128927.

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 27730-27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/blefde53be364a73914f58805a001731-Paper-Conference.pdf.

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 用人类反馈训练语言模型以遵循指令。收录于 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, 和 A. Oh 编， Advances in Neural Information Processing Systems， 卷 35， 第 27730–27744 页。Curran Associates, Inc., 2022。URL https://proceedings.neurips.cc/paper_files/paper/2022/file/blefde53be364a73914f58805a001731-Paper-Conference.pdf.

[14] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. arXiv preprint arXiv:2407.16216, 2024. URL https://arxiv.org/abs/ 2407.16216.

[14] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 大型语言模型对齐技术综述:RLHF、RLAIF、PPO、DPO 等。arXiv 预印本 arXiv:2407.16216， 2024。URL https://arxiv.org/abs/ 2407.16216.

[15] Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongrui Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, et al. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. arXiv preprint arXiv:2410.15595, 2024.

[15] Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongrui Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, et al. 直接偏好优化的全面综述: 数据集、理论、变体与应用。arXiv 预印本 arXiv:2410.15595， 2024。

[16] Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao. A survey of direct preference optimization. CoRR, abs/2503.11701, March 2025. URL https://doi.org/10.48550/arXiv.2503.11701.

[16] Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao. 直接偏好优化综述。CoRR, abs/2503.11701， 2025 年 3 月。URL https://doi.org/10.48550/arXiv.2503.11701.

[17] Hongru Wang, Cheng Qian, Manling Li, Jiahao Qiu, Boyang Xue, Mengdi Wang, Heng Ji, and Kam-Fai Wong. Toward a theory of agents as tool-use decision-makers, 2025. URL https: //arxiv.org/ abs/2506.00886.

[17] Hongru Wang, Cheng Qian, Manling Li, Jiahao Qiu, Boyang Xue, Mengdi Wang, Heng Ji, and Kam-Fai Wong. 面向将代理视为工具使用决策者的理论，2025。URL https: //arxiv.org/abs/2506.00886.

[18] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025. URL https: //arxiv.org/abs/2502.17419.

[18] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 从系统 1 到系统 2: 推理大型语言模型综述，2025。URL https: //arxiv.org/abs/2502.17419.

[19] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xian Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. Large language model agent: A survey on methodology, applications and challenges. CoRR, abs/2503.21460, March 2025. URL https: //doi.org/10.48550/arXiv.2503.21460.

[19] 罗俊宇, 张卫志, 袁烨, 赵玉胜, 杨俊伟, 顾怡阳, 吴博涵, 陈斌奇, 乔子悦, 龙青青, 涂荣成, 罗潇, 居威, 肖志平, 王一凡, 肖萌, 刘晨武, 袁景阳, 张世昌, 金奕乔, 张凡, 吴贤, 赵涵青, 陶达成, Philip S. Yu, 和张明. 大型语言模型代理: 方法论、应用与挑战综述. CoRR, abs/2503.21460, 2025 年 3 月. URL https: //doi.org/10.48550/arXiv.2503.21460.

[20] Aske Plaat, Max J. van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. Agentic large language models, a survey. CoRR, abs/2503.23037, March 2025. URL https://doi.org/10.48550/arXiv.2503.2

[20] Aske Plaat, Max J. van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, 和 Kees Joost Batenburg. Agentic large language models, a survey. CoRR, abs/2503.23037, 2025 年 3 月. URL https://doi.org/10.48550/arXiv.2503.23037.

[21] Xinzhe Li. A review of prominent paradigms for LLM-based agents: Tool use, planning (including RAG), and feedback learning. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, Proceedings of the 31st International Conference on Computational Linguistics, pages 9760-9779, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.652/.

[21] 李新哲. 基于 LLM 的代理的主要范式综述: 工具使用、规划 (含 RAG) 与反馈学习. 收录于 Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, 和 Steven Schockaert 主编, 第 31 届国际计算语言学大会论文集, 页 9760-9779, 阿布扎比, 阿联酋, 2025 年 1 月. 计算语言学协会. URL https://aclanthology.org/2025.coling-main.652/.

[22] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. What are tools anyway? a survey from the language model perspective. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=Xh1B90iBSR.

[22] 王志若，程周俊，朱浩，Daniel Fried，和 Graham Neubig. 工具到底是什么？从语言模型视角的综述. 收录于 First Conference on Language Modeling，2024. URL https://openreview.net/forum?id=Xh1B90iBSR.

[23] Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey, 2024. URL https://arxiv.org/ abs/2404.11584.

[23] Tula Masterman, Sandi Besen, Mason Sawtell, 和 Alex Chao. 推理、规划与工具调用的新兴 AI 代理架构全景: 一项综述, 2024. URL https://arxiv.org/ abs/2404.11584.

[24] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models, 2025. URL https://arxiv.org/abs/2502.21321.

[24] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, 和 Salman Khan. LLM 后训练: 对推理型大型语言模型的深入剖析, 2025. URL https://arxiv.org/abs/2502.21321.

[25] Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. A survey on self-evolution of large language models, 2024. URL https://arxiv.org/abs/2404.143

[25] 陶正伟, 林廷恩, 陈贤才, 李航宇, 吴玉川, 李永彬, 晋志, 黄飞, 陶达成, 和周靖人. 大型语言模型自我演化综述, 2024. URL https://arxiv.org/abs/2404.14387.

[26] R. M. Aratchige and W. M. K. S. Ilmini. Llms working in harmony: A survey on the technological aspects of building effective llm-based multi agent systems, 2025. URL https://arxiv.org/abs/ 2504.01963.

[26] R. M. Aratchige 和 W. M. K. S. Ilmini. 协同工作的 LLM: 构建高效 LLM 驱动多代理系统的技术面向综述, 2025. URL https://arxiv.org/abs/ 2504.01963.

[27] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. Agent ai: Surveying the horizons of multimodal interaction. CoRR, abs/2401.03568, 2024. URL https://doi.org/10.48550/arXiv.2401.0

[27] Zane Durante，Qiuyuan Huang，Naoki Wake，Ran Gong，Jae Sung Park，Bidipta Sarkar，Rohan Taori，Yusuke Noda，Demetri Terzopoulos，Yejin Choi，Katsushi Ikeuchi，Hoi Vo，Li Fei-Fei，和高建峰. Agent AI: 多模态交互前沿综述. CoRR，abs/2401.03568，2024. URL https://doi.org/10.48550/arXiv.2401.03568.

[28] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, Yuheng Cheng, Suyuchen Wang, Xiaoqiang Wang, Yuyu Luo, Haibo Jin, Peiyan Zhang, Ollie Liu, Jiaqi Chen, Huan Zhang, Zhaoyang Yu, Haochen Shi, Boyan Li, Dekun Wu, Fengwei Teng, Xiaojun Jia, Jiawei Xu, Jinyu Xiang, Yizhang Lin, Tianming Liu, Tongliang Liu, Yu Su, Huan Sun, Glen Berseth, Jianyun Nie, Ian Foster, Logan T. Ward, Qingyun Wu, Yu Gu, Mingchen Zhuge, Xiangru Tang, Haohan Wang, Jiaxuan You, Chi Wang, Jian Pei, Qiang Yang, Xiaoliang Qi, and Chenglin Wu. Advances

and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. CoRR, abs/2504.01990, April 2025. URL https://doi.org/10.48550/arXiv.2504.01990.

[28] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, Yuheng Cheng, Suyuchen Wang, Xiaoqiang Wang, Yuyu Luo, Haibo Jin, Peiyan Zhang, Ollie Liu, Jiaqi Chen, Huan Zhang, Zhaoyang Yu, Haochen Shi, Boyan Li, Dekun Wu, Fengwei Teng, Xiaojun Jia, Jiawei Xu, Jinyu Xiang, Yizhang Lin, Tianming Liu, Tongliang Liu, Yu Su, Huan Sun, Glen Berseth, Jianyun Nie, Ian Foster, Logan T. Ward, Qingyun Wu, Yu Gu, Mingchen Zhuge, Xiangru Tang, Haohan Wang, Jiaxuan You, Chi Wang, Jian Pei, Qiang Yang, Xiaoliang Qi, and Chenglin Wu. 基础代理的进展与挑战: 从类脑智能到进化、协作与安全系统。CoRR, abs/2504.01990, 2025 年 4 月。URL https://doi.org/10.48550/arXiv.2504.01990.

[29] Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey, 2024. URL https://arxiv.org/abs/2312.10256.

[29] Dom Huh and Prasant Mohapatra. 多智能体强化学习: 综合综述, 2024。URL https://arxiv.org/abs/2312.10256.

[30] 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 53728-53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/202
a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 直接偏好优化: 你的语言模型其实是一个奖励模型。收录于 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, 和 S. Levine 主编, Advances in Neural Information Processing Systems, 卷 36, 页 53728-53741。Curran Associates, Inc., 2023。URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

[31] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsim-pourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya

Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi,

[31] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsim-pourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi,

Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Kayla Wood、Kendra Rimbach、Keren Gu-Lemberg、Kevin Liu、Kevin Lu、Kevin Stone、Kevin Yu、Lama Ahmad、Lauren Yang、Leo Liu、Leon Maksin、Leyton Ho、Liam Fedus、Lilian Weng、Linden Li、Lindsay McCallum、Lindsey Held、Lorenz Kuhn、Lukas Kondraciuk、Lukasz Kaiser、Luke Metz、Madelaine Boyd、Maja Trebacz、Manas Joglekar、Mark Chen、Marko Tintor、Mason Meyer、Matt Jones、Matt Kaufer、Max Schwarzer、Meghan Shah、Mehmet Yatbaz、Melody Y. Guan、Mengyuan Xu、Mengyuan Yan、Mia Glaese、Mianna Chen、Michael Lampe、Michael Malek、Michele Wang、Michelle Fradin、Mike McClay、Mikhail Pavlov、Miles Wang、Mingxuan Wang、Mira Murati、Mo Bavarian、Mostafa Rohaninejad、Nat McAleese、Neil Chowdhury、Neil Chowdhury、Nick Ryder、Nikolas Tezak、Noam Brown、Ofir Nachum、Oleg Boiko、Oleg Murk、Olivia Watkins、Patrick Chao、Paul Ashbourne、Pavel Izmailov、Peter Zhokhov、Rachel Dias、Rahul Arora、Randall Lin、Rapha Gontijo Lopes、Raz Gaon、Reah Miyara、Reimar Leike、Renny Hwang、Rhythm Garg、Robin Brown、Roshan James、Rui Shu、Ryan Cheu、Ryan Greene、Saachi Jain、Sam Altman、Sam Toizer、Sam Toyer、Samuel Miserendino、Sandhini Agarwal、Santiago Hernandez、Sasha Baker、Scott McKinney、Scottie Yan、Shengjia Zhao、Shengli Hu、Shibani Santurkar、Shraman Ray Chaudhuri、Shuyuan Zhang、Siyuan Fu、Spencer Papay、Steph Lin、Suchir Balaji、Suvansh Sanjeev、Szymon Sidor、Tal Broda、Aidan Clark、Tao Wang、Taylor Gordon、Ted Sanders、Tejal Patwardhan、Thibault Sottiaux、Thomas Degry、Thomas Dimson、Tianhao Zheng、Timur Garipov、Tom Stasi、Trapit Bansal、Trevor Creech、Troy Peterson、Tyna Eloundou、Valerie Qi、Vineet Kosaraju、Vinnie Monaco、Vitchyr Pong、Vlad Fomenko、Weiyi Zheng、Wenda Zhou、Wes McCabe、Wojciech Zaremba、Yann Dubois、Yinghai Lu、Yining Chen、Young Cha、Yu Bai、Yuchen He、Yuchen Zhang、Yunyun Wang、Zheng Shao 和 Zhuohan Li。Openai o1 系统卡，2024。URL https://arxiv.org/abs/2412.16720。

[32] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying

Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[32] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: 通过强化学习激励大模型的推理能力，2025 年。URL https://arxiv.org/abs/2501.12948.

[33] OpenAI Team. Openai o3 and o4-mini: Next-generation reasoning models. Technical report, OpenAI, June 2025. URL https://openai.com/blog/openai-o3-04-mini.Technical announcement introducing OpenAI's o3 and o4-mini models with advanced reasoning capabilities and tool integration.

[33] OpenAI 团队。Openai o3 与 o4-mini: 下一代推理模型。技术报告，OpenAI，2025 年 6 月。URL https://openai.com/blog/openai-o3-04-mini。技术公告介绍了 OpenAI 的 o3 与 o4-mini 模型，具备先进的推理能力和工具集成。

[34] Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future, 2025. URL https://arxiv.org/abs/2504.12328.

[34] 钟嘉伦、沈玮、李炎增、高松阳、鲁华、陈一澄、张扬、周伟、顾锦杰、邹磊。一项关于奖励模型的综合综述: 分类、应用、挑战与未来，2025。URL https://arxiv.org/abs/2504.12328。

[35] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, volume 238 of Proceedings of Machine Learning Research, pages 4447-4455. PMLR, 02-04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

[35] Mohammad Gheshlaghi Azar、Zhaohan Daniel Guo、Bilal Piot、Remi Munos、Mark Rowland、Michal Valko、Daniele Calandriello。理解来自人类偏好学习的一般理论范式。收录于 Sanjoy Dasgupta、Stephan Mandt、Yingzhen Li 主编, The 27th International Conference on Artificial Intelligence and Statistics 论文集，第 238 卷，Proceedings of Machine Learning Research，页 4447-4455。PMLR, 2024 年 5 月 02-04。URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html。

[36] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=iUwHnoENnl.

[36] Kawin Ethayarajh、Winnie Xu、Niklas Muennighoff、Dan Jurafsky、Douwe Kiela。将模型对齐视为前景理论优化。收录于第 41 届国际机器学习大会，2024。URL https://openreview.net/forum?id=iUwHnoENnl。

[37] 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, 2025. URL https://arxiv.org/abs/2504.05118.

[37] 岳昱、袁雨峰、于启颖、左晓晨、朱若飞、徐文远、陈佳泽、王承怡、范天天、杜正音等。Vapo: 用于高级推理任务的高效且可靠的强化学习。arXiv 预印本 arXiv:2504.05118，2025。URL https://arxiv.org/abs/2504.05118。

[38] Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, Shengyi Huang, Siran Yang, Jiamang Wang, Wenbo Su, and Bo Zheng. Part i: Tricks or traps? a deep dive into rl for llm reasoning, 2025. URL https://arxiv.org/abs/2508.08221.

[38] 刘子鹤、刘嘉勋、何彦成、王维勋、刘佳衡、潘玲、胡新宇、熊少攀、黄炬、胡剑、黄胜义、杨思然、王家芒、苏文博、郑波。第一部分: 技巧还是陷阱? 深入探讨用于 LLM 推理的 RL，2025。URL https://arxiv.org/abs/2508.08221。

[39] Chuheng Zhang, Wei Shen, Li Zhao, Xuyun Zhang, Xiaolong Xu, Wanchun Dou, and Jiang Bian. Policy filtration for RLHF to mitigate noise in reward models. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=L8hYdTQVcs.

[39] 张楚恒、沈玮、赵立、张须云、徐晓龙、窦万春、卞江。用于 RLHF 的策略过滤以减轻奖励模型噪声。收录于第 42 届国际机器学习大会，2025。URL https://openreview.net/forum?id=L8hYdTQVcs。

[40] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. 2024. URL https://arxiv.org/abs/2410.01679.

[40] Amirhossein Kazemnejad、Milad Aghajohari、Eva Portelance、Alessandro Sordoni、Siva Reddy、Aaron Courville、Nicolas Le Roux。Vineppo: 通过精细的信用分配释放 LLM 推理的 RL 潜力。2024。URL https://arxiv.org/abs/2410.01679。

[41] Ning Dai, Zheng Wu, Renjie Zheng, Ziyun Wei, Wenlei Shi, Xing Jin, Guanlin Liu, Chen Dun, Liang Huang, and Lin Yan. Process supervision-guided policy optimization for code generation, 2025. URL https://openreview.net/forum?id=Cn5Z0MUPZT.

[41] 戴宁、吴征、郑仁杰、魏子允、石文磊、金星、刘冠霖、敦辰、黄亮、闫林。面向代码生成的过程监督引导策略优化，2025。URL https://openreview.net/forum?id=Cn5Z0MUPZT。

[42] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. $ \beta$-DPO: Direct preference optimization with dynamic $ \beta$. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https:// openreview.net/forum?id=ZfBuhzE556.

[42] 吴俊康、谢粤翔、杨正奕、吴建灿、高晋阳、丁博林、王翔、何向南。$ \beta$-DPO: 具有动态 $ \beta$ 的直接偏好优化。收录于第 38 届年度神经信息处理系统会议，2024。URL https:// openreview.net/forum?id=ZfBuhzE556。

[43] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=3Tzcot1LKb.

[43] 孟宇、夏梦洲、陈丹琦。SimPO: 基于无参考奖励的简单偏好优化。收录于第 38 届年度神经信息处理系统会议，2024。URL https://openreview.net/forum?id=3Tzcot1LKb。

[44] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. arXiv preprint arXiv:2403.07691, 2024.

[44] 洪智佑、李诺亚、James Thorne。Orpo: 无需参考模型的整体偏好优化。arXiv 预印本 arXiv:2403.07691，2024。

[45] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. arXiv preprint arXiv:2406.18629, 2024.

[45] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: 用于大模型长链推理的逐步偏好优化。arXiv 预印本 arXiv:2406.18629，2024。

[46] Bin Hong, Jiayu Liu, Zhenya Huang, Kai Zhang, and Mengdi Zhang. Pruning long chain-of-thought of large reasoning models via small-scale preference optimization. arXiv preprint arXiv:2508.10164, 2025.

[46] Bin Hong, Jiayu Liu, Zhenya Huang, Kai Zhang, and Mengdi Zhang. 通过小规模偏好优化裁剪大型推理模型的长链思考。arXiv 预印本 arXiv:2508.10164，2025。

[47] 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, 2025.

[47] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: 大规模开源 llm 强化学习系统。arXiv 预印本 arXiv:2503.14476，2025。

[48] 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, 2025.

[48] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. 群序列策略优化。arXiv 预印本 arXiv:2507.18071，2025。

[49] 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, 2025.

[49] Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. 几何均值策略优化。arXiv 预印本 arXiv:2507.20673，2025。

[50] 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, 2025. URL https://arxiv.org/abs/2505.24864.

[50] Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: 延长的强化学习拓展了大语言模型的推理边界，2025。URL https://arxiv.org/abs/2505.24864。

[51] Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. Posterior-grpo: Rewarding reasoning processes in code generation. arXiv preprint arXiv:2508.05170, 2025.

[51] Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. Posterior-grpo: 在代码生成中奖励推理过程。arXiv 预印本 arXiv:2508.05170，2025。

[52] 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, 2025.

[52] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 理解 r1-zero 类训练: 一种批判性视角。arXiv 预印本 arXiv:2503.20783，2025。

[53] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937, 2025.

[53] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: 通过逐步群体相对策略优化学习多模态大语言模型的推理。arXiv预印本arXiv:2503.12937，2025。

[54] Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. arXiv preprint arXiv:2504.14286, 2025.

[54] Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. Srpo: 在llm上跨域实现大规模强化学习。arXiv预印本arXiv:2504.14286，2025。

[55] 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, 2025. URL https://arxiv.org/abs/2506.02177.

[55] Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. 只在有利时行动: 通过选择性rollout提高llm推理的高效强化学习，2025。URL https://arxiv.org/abs/2506.02177。

[56] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning, 2025. URL https://arxiv.org/abs/2504.20073.

[56] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. Ragen: 通过多轮强化学习理解llm代理的自我演化，2025。URL https://arxiv.org/abs/2504.20073。

[57] Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. Ghpo: Adaptive guidance for stable and efficient llm reinforcement learning. arXiv preprint arXiv:2507.10628, 2025.

[57] Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. Ghpo: 用于稳定且高效llm强化学习的自适应引导。arXiv预印本arXiv:2507.10628，2025。

[58] 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. arXiv preprint arXiv:2504.16656, 2025.

[58] 王培瑜, 魏奕辰, 彭毅, 王晓坤, 邱伟杰, 沈威, 谢天一丹, 裴江博, 张建豪, 郝云卓, 等. Skywork r1v2: 用于推理的多模态混合强化学习. arXiv预印本arXiv:2504.16656, 2025.

[59] Heng Lin and Zhongwen Xu. Understanding tool-integrated reasoning, 2025. URL https://arxiv.org/abs/2508.19201.

[59] 林恒和徐中文. 理解工具集成推理, 2025. URL https://arxiv.org/abs/2508.19201.

[60] Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, et al. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. arXiv preprint arXiv:2508.17445, 2025.

[60] 李奕智, 顾清水, 温周福图, 李子牛, 邢天顺, 郭树玥, 郑天宇, 周鑫, 曲星威, 周望春树, 等. Treepo: 用启发式树模型弥合策略优化效果与推理效率的差距. arXiv 预印本 arXiv:2508.17445, 2025.

[61] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. arXiv preprint arXiv:2507.21848, 2025.

[61] 张兴建, 文思威, 吴文俊, 和黄磊. Edge-grpo: 带引导误差修正的熵驱动 grpo 以实现优势多样性. arXiv 预印本 arXiv:2507.21848, 2025.

[62] 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, 2025.

[62] 杨志成, 郭志江, 黄银雅, 王永鑫, 谢东春, 王逸尉, 梁晓丹, 和唐晶. rlvr 中的深度-广度协同: 通过自适应探索释放 llm 推理增益. arXiv 预印本 arXiv:2508.13755, 2025.

[63] 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, 2025.

[63] 张文浩, 谢越翔, 孙钰昌, 陈岩曦, 王国印, 李雅良, 丁博林, 和周镜人. 在策略内 rl 遇到策略外专家: 通过动态加权协调监督微调与强化学习. arXiv 预印本 arXiv:2508.11408, 2025.

[64] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. arXiv preprint arXiv:2507.06448, 2025.

[64] 王振海龙, 郭雪航, Sofia Stoica, 许海洋, 王鸿儒, 河贤静, 陈秀思, 陈杨毅, 严明, 黄飞, 等. 面向感知的策略优化用于多模态推理. arXiv 预印本 arXiv:2507.06448, 2025.

[65] 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, 2025. URL https://arxiv.org/abs/2508.10751.

[65] 陈志鹏, 秦晓博, 吴有斌, 凌岳, 叶庆豪, 赵文欣, 和石光. Pass@k 训练用于自适应平衡大规模推理模型的探索与利用, 2025. URL https://arxiv.org/abs/2508.10751.

[66] Lilian Weng. Llm-powered autonomous agents. lilianweng.github.io, Jun 2023. URL https:// lilianweng.github.io/pos 06-23-agent/.

[66] Lilian Weng. Llm-powered autonomous agents. lilianweng.github.io, 2023 年 6 月. URL https://lilianweng.github.io/posts/2023-06-23-agent/.

[67] Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. Agentsquare: Automatic LLM agent search in modular design space. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=mPdmDYIQ7f.

[67] Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. Agentsquare: Automatic LLM agent search in modular design space. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=mPdmDYIQ7f.

[68] Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, and Fei Liu. PlanGenLLMs: A modern survey of LLM planning capabilities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19497-19521, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long. 958. URL https://aclanthology.org/2025.acl-long.958/.

[68] Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, and Fei Liu. PlanGenLLMs: A modern survey of LLM planning capabilities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19497-19521, Vienna, Austria, 2025 年 7 月. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long. 958. URL https://aclanthology.org/2025.acl-long.958/.

[69] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. LLM as a mastermind: A survey of strategic reasoning with large language models. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id= iMqJsQ4evS.

[69] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. LLM as a mastermind: A survey of strategic reasoning with large language models. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id= iMqJsQ4evS.

[70] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian,

[70] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian,

Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. ACM Comput. Surv., 57(4), December 2024. ISSN 0360-0300. doi: 10.1145/3704435. URL https: //doi.org/10.1145/3704435.

Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. ACM Comput. Surv., 57(4), 2024 年 12 月. ISSN 0360-0300. doi: 10.1145/3704435. URL https: //doi.org/10.1145/3704435.

[71] Allen Newell, John Calman Shaw, and Herbert A Simon. Elements of a theory of human problem solving. Psychological review, 65(3):151, 1958.

[71] Allen Newell, John Calman Shaw, and Herbert A Simon. Elements of a theory of human problem solving. Psychological review, 65(3):151, 1958.

[72] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. CoRR, abs/2402.02716, 2024. URL https://doi.org/10.48550/arXiv.2402.02716.

[72] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. CoRR, abs/2402.02716, 2024. URL https://doi.org/10.48550/arXiv.2402.02716.

[73] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[73] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[74] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8154-8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/ v1/2023.emnlp-main.507.URL https://aclanthology.org/2023.emnlp-main.507/.

[74] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8154-8173, Singapore, 2023 年 12 月. Association for Computational Linguistics. doi: 10.18653/ v1/2023.emnlp-main.507.URL https://aclanthology.org/2023.emnlp-main.507/.

[75] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024.

[75] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search将语言模型中的推理、行动与规划统一起来. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024.

[76] Joey Hong, Anca Dragan, and Sergey Levine. Planning without search: Refining frontier llms with offline goal-conditioned rl, 2025. URL https://arxiv.org/abs/2505.18098.

[76] Joey Hong、Anca Dragan和Sergey Levine。无搜索规划: 用离线目标条件化强化学习精炼边界LLMs，2025。URL https://arxiv.org/abs/2505.18098.

[77] Davide Paglieri, Bartlomiej Cupiat, Jonathan Cook, Ulyana Piterbarg, Jens Tuyls, Edward Grefenstette, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. Learning when to plan: Efficiently allocating test-time compute for llm agents, 2025. URL https://arxiv.org/abs/2509.03581.

[77] Davide Paglieri、Bartlomiej Cupiat、Jonathan Cook、Ulyana Piterbarg、Jens Tuyls、Edward Grefenstette、Jakob Nicolaus Foerster、Jack Parker-Holder和Tim Rocktäschel。何时规划的学习: 为LLM代理高效分配测试时计算，2025。URL https://arxiv.org/abs/2509.03581.

[78] Merve Atasever, Matthew Hong, Mihir Nitin Kulkarni, Qingpei Li, and Jyotirmoy V. Deshmukh. Multi-agent path finding via offline rl and llm collaboration, 2025. URL https://arxiv.org/abs/ 2509.22130.

[78] Merve Atasever、Matthew Hong、Mihir Nitin Kulkarni、Qingpei Li和Jyotirmoy V. Deshmukh。通过离线RL与LLM协作的多智能体路径规划，2025。URL https://arxiv.org/abs/ 2509.22130.

[79] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. 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 7584-7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.409. URL https://aclanthology.org/2024.acl-long.409/.

[79] Yifan Song、Da Yin、Xiang Yue、Jie Huang、Sujian Li和Bill Yuchen Lin。试错: 基于探索的LLM代理轨迹优化。收录于Lun-Wei Ku、Andre Martins和Vivek Srikumar编辑,《第62届计算语言学协会年会论文集 (第1卷: 长篇论文)》，第7584–7600页，泰国曼谷，2024年8月。计算语言学协会。doi: 10.18653/v1/2024.acl-long.409。URL https://aclanthology.org/2024.acl-long.409/.

[80] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum? id=ehfRiF0R3a.

[80] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: 使用大型语言模型的开放式具身代理. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum? id=ehfRiF0R3a.

[81] Yilin Guan, Qingfeng Lan, Sun Fei, Dujian Ding, Devang Acharya, Chi Wang, William Yang Wang, and Wenyue Hua. Dynamic speculative agent planning, 2025. URL https://arxiv.org/abs/ 2509.01920.

[81] Yilin Guan, Qingfeng Lan, Sun Fei, Dujian Ding, Devang Acharya, Chi Wang, William Yang Wang, and Wenyue Hua. 动态投机代理规划, 2025. URL https://arxiv.org/abs/ 2509.01920.

[82] Zhiwei Li, Yong Hu, and Wenqing Wang. Encouraging good processes without the need for good answers: Reinforcement learning for llm agent planning, 2025. URL https://arxiv.org/abs/ 2508.19598.

[82] 李志伟, 胡勇, 王文清. 在不依赖好答案的情况下鼓励良好过程: 用于 llm 代理规划的强化学习, 2025. URL https://arxiv.org/abs/ 2508.19598.

[83] Keer Lu, Chong Chen, Bin Cui, Huang Leng, and Wentao Zhang. Pilotrl: Training language model agents via global planning-guided progressive reinforcement learning, 2025. URL https://arxiv.org/abs/2508.00344.

[83] Keer Lu, Chong Chen, Bin Cui, Huang Leng, and Wentao Zhang. Pilotrl: 通过全局规划引导的渐进式强化学习训练语言模型代理, 2025. URL https://arxiv.org/abs/2508.00344.

[84] Siyu Zhu, Yanbin Jiang, Hejian Sang, Shao Tang, Qingquan Song, Biao He, Rohit Jain, Zhipeng Wang, and Alborz Geramifard. Planner-r1: Reward shaping enables efficient agentic rl with smaller llms, 2025. URL https://arxiv.org/abs/2509.25779.

[84] 朱思宇、蒋彦斌、桑和建、唐韶、宋庆泉、何彪、Rohit Jain、王志鹏和 Alborz Geramifard。《Planner-r1: 奖励塑形使较小的 LLMs 在主体性强化学习中高效》, 2025 年。URL https://arxiv.org/abs/2509.25779.

[85] Taylor Webb, Shanka Subhra Mondal, and Ida Momennejad. A brain-inspired agentic architecture to improve planning with llms. Nature Communications, 16(1):8633, 2025. doi: 10.1038/ s41467-025-63804-5. URL https://doi.org/10.1038/s41467-025-63804-5.

[85] Taylor Webb、Shanka Subhra Mondal 和 Ida Momennejad。《一种受大脑启发的主体性架构以利用 LLMs 改善规划》。Nature Communications, 16(1):8633, 2025。doi: 10.1038/s41467-025-63804-5。URL https://doi.org/10.1038/s41467-025-63804-5.

[86] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 68539-68551. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/ file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.

[86] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 68539-68551. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/ file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.

[87] Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning, 2023. URL https://arxiv.org/abs/2310.05915.

[87] Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning, 2023. URL https://arxiv.org/abs/2310.05915.

[88] Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling generalized agent abilities for LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Sriku-mar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 3053-3077, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. findings-acl.181. URL https://aclanthology.org/2024.findings-acl.181/.

[88] Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling generalized agent abilities for LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 3053-3077, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. findings-acl.181. URL https://aclanthology.org/2024.findings-acl.181/.

[89] Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-FLAN: Designing data and methods of effective agent tuning for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 9354-9366, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.557. URL https://aclanthology.org/2024.findings-acl.557/.

[89] Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-FLAN: Designing data and methods of effective agent tuning for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 9354-9366, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.557. URL https://aclanthology.org/2024.findings-acl.557/.

[90] Yifan Song, Weimin Xiong, Xiutian Zhao, Dawei Zhu, Wenhao Wu, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. Agentbank: Towards generalized llm agents via fine-tuning on 50000+ interaction trajectories. In EMNLP (Findings), pages 2124-2141, 2024. URL https://aclanthology.org/ 2024.findings-emnlp.116.

[90] Yifan Song, Weimin Xiong, Xiutian Zhao, Dawei Zhu, Wenhao Wu, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. Agentbank: Towards generalized llm agents via fine-tuning on 50000+ interaction trajectories. In EMNLP (Findings), pages 2124-2141, 2024. URL https://aclanthology.org/ 2024.findings-emnlp.116.

[91] Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. In EMNLP, pages 3102-3116, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.187.

[91] Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. In EMNLP, pages 3102-3116, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.187.

[92] 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, 2025. URL https://arxiv.org/abs/ 2504.13958.

[92] 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, 2025. URL https://arxiv.org/abs/2504.13958.

[93] Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. Acting less is reasoning more! teaching model to act efficiently, 2025. URL https://arxiv.org/abs/2504.14870.

[93] Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. Acting less is reasoning more! teaching model to act efficiently, 2025. URL https://arxiv.org/abs/2504.14870.

[94] 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. CoRR, abs/2504.11536, April 2025. URL https://doi.org/10.48550/arXiv.2504.11536.

[94] 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. CoRR, abs/2504.11536, April 2025. URL https://doi.org/10.48550/arXiv.2504.11536.

[95] Yifan Wei, Xiaoyan Yu, Yixuan Weng, Tengfei Pan, Angsheng Li, and Li Du. Autotir: Autonomous tools integrated reasoning via reinforcement learning, 2025. URL https://arxiv.org/abs/2507.21836.

[95] 魏一凡, 余晓艳, 翁亦轩, 潘腾飞, 李昂声, 以及杜励. Autotir: 通过强化学习实现自主工具集成推理, 2025. URL https://arxiv.org/abs/2507.21836.

[96] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use, 2025. URL https://arxiv.org/abs/2505.19255.

[96] 吴明远, 杨敬成, 江济泽, 李梅堂, 严凯卓, 余汉超, 张敏嘉, 翟成祥, 以及克拉拉·纳赫施特特. Vtool-r1: 视觉语言模型通过在多模态工具使用上的强化学习学会用图像进行思考, 2025. URL https://arxiv.org/abs/2505.19255.

[97] 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, 2025. URL https://arxiv.org/abs/2505.14362.

[97] 郑紫玮, 杨迈克尔, 洪杰克, 赵晨晓, 徐国海, 杨乐, 沈超, 以及余星. Deepeyes: 通过强化学习激励“用图像思考”, 2025. URL https://arxiv.org/abs/2505.14362.

[98] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning, 2025. URL https://arxiv.org/abs/2505.15966.

[98] 苏雅力, 王浩哲, 任伟明, 林方震, 以及陈文湖. Pixel reasoner: 通过好奇心驱动的强化学习激励像素空间推理, 2025. URL https://arxiv.org/abs/2505.15966.

[99] Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28489-28503, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10. 18653/v1/2025.acl-long.1383.URL https://aclanthology.org/2025.acl-long.1383/.

[99] 吴君德, 朱佳元, 刘育元, 徐敏, 以及金岳明. Agentic reasoning: 一个用于用能动工具增强大模型推理的精简框架. 收录于: Wanxiang Che, Joyce Nabende, Ekaterina Shutova, 和 Mohammad Taher Pilehvar 编, 第 63 届计算语言学协会年会论文集 (第 1 卷: 长篇论文), 页码 28489-28503, 奥地利维也纳, 2025 年 7 月. 计算语言学协会. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1383. URL https://aclanthology.org/2025.acl-long.1383/.

[100] Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool integration for llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.01441.

[100] Joykirat Singh, Raghav Magazine, Yash Pandya, 以及 Akshay Nambi. 通过强化学习实现的能动推理与工具集成以用于大模型, 2025. URL https://arxiv.org/abs/2505.01441.

[101] Xuefeng Li, Haoyang Zou, and Pengfei Liu. Torl: Scaling tool-integrated rl, 2025. URL https://arxiv.org/abs/2503.23383.

[101] 李学锋, 邹浩洋, 以及刘鹏飞. Torl: 扩展工具集成强化学习, 2025. URL https://arxiv.org/abs/2503.23383.

[102] Bingguang Hao, Maolin Wang, Zengzhuang Xu, Yicheng Chen, Cunyin Peng, Jinjie GU, and Chenyi Zhuang. Exploring superior function calls via reinforcement learning, 2025. URL https://arxiv.org/abs/2508.05118.

[102] 郝炳光, 王茂林, 许增庄, 陈怡诚, 彭存音, GU 锦杰, 以及庄晨逸. 通过强化学习探索更优的函数调用, 2025. URL https://arxiv.org/abs/2508.05118.

[103] Peiyuan Feng, Yichen He, Guanhua Huang, Yuan Lin, Hanchong Zhang, Yuchen Zhang, and Hang Li. Agile: A novel reinforcement learning framework of llm agents. In A. Glober-son, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 5244-5284. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/097c514162ea7126d40671d23e12f51b-Paper-Conference.pdf.

[103] 冯培元, 何怡辰, 黄冠华, 林源, 张瀚冲, 张宇澄, 以及李航. Agile: 一种用于大模型代理的新型强化学习框架. 收录于: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, 和 C. Zhang 编, 神经信息处理系统进展, 第 37 卷, 页码 5244-5284. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/097c514162ea7126d40671d23e12f51b-Paper-Conference.pdf.

[104] Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In ICML 2025 Workshop on Computer Use Agents, 2025. URL https://openreview.net/forum?id=

[104] 魏哲佩, 姚文林, 刘尧, 张伟志, 吕沁, 邱亮, 余长龙, 徐蒲阳, 张超, 殷冰, 允赫根, 以及李利宏. Webagent-r1: 通过端到端多轮强化学习训练网络代理. 收录于 ICML 2025 计算机使用代理研讨会, 2025. URL https://openreview.net/forum?id=KqrYTALRjH.

[105] Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating super-human reasoning for web agent, 2025. URL https://arxiv.org/abs/2507.02592.

[105] 李宽, 张中望, 殷惠峰, 张立文, 欧立图, 吴佳龙, 尹文彪, 李百轩, 陶正伟, 王信宇, 沈维舟, 张君凯, 张丁初, 吴西西, 姜勇, 闫明, 谢鹏军, 黄飞, 以及周景仁. Websailor: 为网络代理导航超人类级别的推理, 2025. URL https://arxiv.org/abs/2507.02592.

[106] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webdancer: Towards autonomous information seeking agency, 2025. URL https://arxiv.org/abs/2505.22648.

[106] 吴嘉龙、李白轩、方润南、殷文彪、张立文、陶正伟、张定初、奚泽坤、付刚、姜勇、谢鹏军、黄飞和周靖人。Webdancer: 迈向自主信息寻求代理，2025。URL https://arxiv.org/abs/2505.22648.

[107] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models, 2025. URL https: //arxiv.org/abs/2501.05366.

[107] 李孝熙、董冠廷、金嘉杰、张宇尧、周宇佳、朱玉涛、张沛天和窦志成。Search-o1: 具主体性的搜索增强大型推理模型，2025。URL https://arxiv.org/abs/2501.05366.

[108] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. Research: Learning to reason with search for llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2503.19470.

[108] 陈明阳、李天鹏、孙浩泽、周一杰、祝晨正、王浩芬、Jeff Z. Pan、张文、陈华军、杨帆、周泽楠和陈卫鹏。RESEARCH: 通过强化学习学习用检索进行推理以增强大模型，2025。URL https://arxiv.org/abs/2503.19470.

[109] Zhao Song, Song Yue, and Jiahao Zhang. Thinking isn't an illusion: Overcoming the limitations of reasoning models via tool augmentations, 2025. URL https://arxiv.org/abs/2507.17699.

[109] 宋钊、岳松和张佳豪。思考并非幻觉: 通过工具增强克服推理模型的局限性，2025。URL https://arxiv.org/abs/2507.17699.

[110] Junjie Ye, Changhao Jiang, Zhengyin Du, Yufei Xu, Xuesong Yao, Zhiheng Xi, Xiaoran Fan, Qi Zhang, Xuanjing Huang, and Jiecao Chen. Feedback-driven tool-use improvements in large language models

via automated build environments, 2025. URL https://arxiv.org/abs/2508.08791.

[110] 叶俊杰、姜长昊、杜正寅、徐宇飞、姚学松、席志恒、范晓然、张齐、黄宣靖和陈杰曹。基于反馈的工具使用改进: 通过自动化构建环境改进大型语言模型, 2025。URL https://arxiv.org/abs/2508.08791.

[111] OpenAI. Deep research. https://openai.com/index/introducing-deep-research/, 2025.

[111] OpenAI。Deep research。https://openai.com/index/introducing-deep-research/，2025。

[112] Kimi. Kimi-researcher: End-to-end rl training for emerging agentic capabilities. https:// moonshotai.github.io/Kimi-Researcher/, 2025.

[112] Kimi。Kimi-researcher: 端到端强化学习训练以获得新兴主体能力。https://moonshotai.github.io/Kimi-Researcher/，2025。

[113] Qwen Team. Qwq-32B: Embracing the power of reinforcement learning. Blog post on QwenLM official site, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.[Accessed 2025-08-25].

[113] Qwen 团队。Qwq-32B: 拥抱强化学习的力量。QwenLM 官方站点博客，2025 年 3 月。URL https://qwenlm.github.io/blog/qwq-32b/.[Accessed 2025-08-25].

[114] Zhipu AI. zai-org/GLM-Z1-32B-0414 - Hugging Face - huggingface.co. https://huggingface.co/zai-org/GLM-Z1-32B-0414, 2025. [Accessed 25-08-2025].

[114] 智谱 AI。zai-org/GLM-Z1-32B-0414 - Hugging Face - huggingface.co。https://huggingface.co/zai-org/GLM-Z1-32B-0414，2025。[Accessed 25-08-2025].

[115] Ning Shang, Yifei Liu, Yi Zhu, Li Lyna Zhang, Weijiang Xu, Xinyu Guan, Buze Zhang, Bingcheng Dong, Xudong Zhou, Bowen Zhang, Ying Xin, Ziming Miao, Scarlett Li, Fan Yang, and Mao Yang. rstar2-agent: Agentic reasoning technical report, 2025. URL https://arxiv.org/abs/2508.20722.

[115] 商宁、刘亦飞、朱毅、张丽娜、徐伟江、关新宇、张布泽、董炳成、周旭东、张博文、辛颖、缪梓铭、李斯卡蕾、杨帆和杨茂。rstar2-agent: 具主体性的推理技术报告, 2025。URL https://arxiv.org/abs/2508.20722.

[116] Meituan. meituan-longcat/LongCat-Flash-Chat - Hugging Face. https://huggingface.co/ meituan-longcat/LongCat-Flash-Chat, 2025. [Accessed 02-09-2025].

[116] 美团。meituan-longcat/LongCat-Flash-Chat - Hugging Face。https://huggingface.co/meituan-longcat/LongCat-Flash-Chat，2025。[Accessed 02-09-2025].

[117] 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, 2025. URL https://arxiv.org/abs/2508.07976.

[117] 高佳轩、傅威、谢民阳、徐殊胜、何楚怡、梅志宇、朱邦华和吴奕。超越十轮: 通过大规模异步强化学习解锁长周期主体性搜索，2025。URL https://arxiv.org/abs/2508.07976.

[118] Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum? id=bNtr6SLgZf. Survey Certification.

[118] Eduardo Pignatelli、Johan Ferret、Matthieu Geist、Thomas Mesnard、Hado van Hasselt 和 Laura Toni。深度强化学习中的时间性功绩分配综述。Transactions on Machine Learning Research，2024。ISSN 2835-8856. URL https://openreview.net/forum?id=bNtr6SLgZf。综述认证。

[119] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training, 2025. URL https://arxiv.org/abs/2505.10978.

[119] 冯朗、薛正海、刘庭聪和安博。面向大语言模型代理训练的组中组策略优化，2025。URL https://arxiv.org/abs/2505.10978.

[120] Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution, 2025. URL https://arxiv.org/abs/2505.20732.

[120] 王涵霖、梁泽濤 (Chak Tou Leong)、王家硕、王建和李文杰。SPA-RL: 通过逐步进展归因强化大语言模型代理，2025。URL https://arxiv.org/abs/2505.20732.

[121] Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms, 2025. URL https://arxiv.org/abs/2504.15965.

[121] 吳雅雄, 梁晟, 張辰, 王一超, 張永月, 郭惠峰, 唐瑞明, 和劉勇。從人類記憶到 AI 記憶: 大語言模型時代記憶機制綜述, 2025。URL https://arxiv.org/abs/2504.15965.

[122] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17):19724-19731, Mar. 2024. doi: 10.1609/aaai.v38i17.29946. URL https: //ojs.aaai.org/index.php/AAAI/article/view

[122] 鍾萬鈞, 郭良宏, 高祺祺, 葉和, 和王彥林。Memorybank: 以長期記憶增強大型語言模型。人工智能協會 AAAI 會議論文集, 38(17):19724-19731, 2024 年 3 月。doi: 10.1609/aaai.v38i17.29946。URL https: //ojs.aaai.org/index.php/AAAI/article/view/29946.

[123] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. CoRR, abs/2310.08560, 2023. URL https://doi.org/10.48550/arXiv.2310.08560.

[123] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, 和 Joseph E. Gonzalez。MemGPT: 邁向作業系統化的 LLMs。CoRR, abs/2310.08560, 2023。URL https://doi.org/10.48550/arXiv.2310.08560.

[124] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, vol-

ume 37, pages 59532-59569. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ 6ddc001d07ca4f319af96a3024f6dbd1-Paper-Conference.pdf.

[124] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, 和 Yu Su。HippoRAG: 受神經生物學啟發的大型語言模型長期記憶。在 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, 和 C. Zhang 編, Advances in Neural Information Processing Systems, 第 37 卷, 頁 59532-59569。Curran Associates, Inc., 2024。URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ 6ddc001d07ca4f319af96a3024f6dbd1-Paper-Conference.pdf.

[125] Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8416-8439, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

[125] 譚震, 閻俊, I-Hung Hsu, 韓汝鈞, 王子豐, Le Long, 宋怡文, 陳彥霏, Hamid Palangi, George Lee, Anand Rajan Iyer, 陳天龍, 劉煥, Chen-Yu Lee, 和 Tomas Pfister。展望與回顧: 長期個性化對話代理的反思性記憶管理。在 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, 和 Mohammad Taher Pilehvar 編, 第 63 屆計算語言學協會年會論文集 (卷 1: 長文), 頁 8416-8439, 2025 年 7 月, 維也納, 奧地利。計算語言學協會。ISBN 979-8-89176-251-0。doi: 10.18653/v1/2025.acl-long.413。URL https://aclanthology.org/2025.acl-long.413/.

[126] Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning, 2025. URL https://arxiv.org/abs/2508.19828.

[126] 顏思寬, 楊秀峰, 黃祖超, 聶爾聰, 丁子豐, 李宗根, 馬曉文, Hinrich Schütze, Volker Tresp, 和馬雲朴。Memory-R1: 通過強化學習增強大型語言模型代理的記憶管理與利用，2025。URL https://arxiv.org/abs/2508.19828.

[127] Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem- α : Learning memory construction via reinforcement learning,2025. URL https://arxiv.org/abs/2509.25911.

[127] 王宇, Ryuichi Takanobu, 梁志奇, 毛玉振, 胡元哲, Julian McAuley, 和吳曉劍。Mem- α : 通過強化學習學習記憶構建,2025。URL https://arxiv.org/abs/2509.25911.

[128] Yuxiang Zhang, Jiangming Shu, Ye Ma, Xueyuan Lin, Shangxi Wu, and Jitao Sang. Memory as action: Autonomous context curation for long-horizon agentic tasks, 2025. URL https://arxiv.org/ abs/2510.12635.

[128] 張玉祥, 舒江鳴, 馬野, 林雪源, 伍上曦, 和桑吉濤。記憶即行動: 面向長期任務的自主情境策展, 2025。URL https://arxiv.org/ abs/2510.12635.

[129] Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. Memagent: Reshaping long-context llm with

multi-conv rl-based memory agent, 2025. URL https://arxiv.org/abs/2507.02259.

[129] 余宏立，陳廷宏，馮江濤，陳江杰，戴維南，余啟英，張亞勤，馬偉穎，劉晶晶，王明軒，和周浩。MemAgent: 用基於多對話強化學習的記憶代理重塑長上下文 LLM，2025。URL https://arxiv.org/abs/2507.02259.

[130] 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, 2025. URL https://arxiv.org/abs/2506.15841.

[130] 周子建，曲奥，吴朝轩，金成焕，Alok Prakash，Daniela Rus，赵金华，Bryan Kian Hsiang Low，和 Paul Pu Liang. Mem1：学习协同记忆与推理以实现高效长时程智能体，2025. URL https://arxiv.org/abs/2506.15841.

[131] Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. Disentangling memory and reasoning ability in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1681-1701, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.84. URL https://aclanthology.org/ 2025.acl-long.84/.

[131] 金明宇, 罗伟迪, 程思涛, 王欣逸, 花文悦, 唐瑞香, William Yang Wang, 和张永锋. 在大型语言模型中解构记忆与推理能力. 收录于 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, 和 Mohammad Taher Pilehvar 主编, 第 63 届计算语言学协会年会论文集 (第 1 卷: 长篇论文), 页 1681-1701, 奥地利维也纳, 2025 年 7 月. 计算语言学协会. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.84. URL https://aclanthology.org/2025.acl-long.84/.

[132] Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Yong Jiang, Pengjun Xie, Fei Huang, et al. Resum: Unlocking long-horizon search intelligence via context summarization. arXiv preprint arXiv:2509.13313, 2025.

[132] 吴熙熙, 李宽, 赵一达, 张立文, 欧立图, 尹惠峰, 张中望, 蒋永, 谢鹏军, 黄飞, 等. Resum: 通过上下文摘要解锁长时程搜索智能. arXiv 预印本 arXiv:2509.13313, 2025.

[133] Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. Scaling long-horizon llm agent via context-folding, 2025. URL https://arxiv.org/abs/2510.11967.

[133] 孙维维, 卢淼, 凌湛, 刘康, 姚学松, 杨亦鸣, 和陈杰操. 通过上下文折叠扩展长时程 LLM 智能体, 2025. URL https://arxiv.org/abs/2510.11967.

[134] Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. MEMORYLLM: Towards self-updatable large language models. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=p0lKWzdi

[134] 王宇, 高一帆, 陈秀思, 江浩明, 李世阳, 杨敬锋, 尹庆宇, 李征, 李贤, 殷冰, 商靖博, 和 Julian McAuley. MEMORYLLM: 面向可自我更新的大型语言模型. 收录于第 41 届国际机器学习大会, 2024. URL https://openreview.net/forum?id=p0lKWzdikQ.

[135] Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending memoryLLM with scalable long-term memory. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/ forum?id=OcqbkROe8J.

[135] 王宇, Dmitry Krotov, 胡远哲, 高一帆, 周望春书, Julian McAuley, Dan Gutfreund, Rogerio Feris, 和何泽学. M+: 以可扩展的长期记忆扩展 memoryLLM. 收录于第 42 届国际机器学习大会, 2025. URL https://openreview.net/forum?id=OcqbkROe8J.

[136] José I. Orlicki. Beyond words: A latent memory approach to internal reasoning in llms, 2025. URL https://arxiv.org/abs/2502.21030.

[136] José I. Orlicki. 超越词语: 一种用于 LLM 内部推理的潜在记忆方法, 2025. URL https://arxiv.org/abs/2502.21030.

[137] Hongkang Yang Hongkang Yang, Zehao Lin Zehao Lin, Wenjin Wang Wenjin Wang, Hao Wu Hao Wu, Zhiyu Li Zhiyu Li, Bo Tang Bo Tang, Wenqiang Wei Wenqiang Wei, Jinbo Wang Jinbo Wang, Zeyun Tang Zeyun Tang, Shichao Song Shichao Song, Chenyang Xi Chenyang Xi, Yu Yu Yu Yu, Kai Chen Kai Chen, Feiyu Xiong Feiyu Xiong, Linpeng Tang Linpeng Tang, and Weinan E Weinan E. Memory3: Language modeling with explicit memory. Journal of Machine Learning, 3(3):300-346, January 2024. ISSN 2790-203X. doi: 10.4208/jml.240708. URL http://dx.doi.org/10.4208/jml.240708.

[137] Yang Hongkang, Lin Zehao, Wang Wenjin, Wu Hao, Li Zhiyu, Tang Bo, Wei Wenqiang, Wang Jinbo, Tang Zeyun, Song Shichao, Xi Chenyang, Yu Yu, Chen Kai, Xiong Feiyu, Tang Linpeng, 和 E Weinan. Memory3: 带显式记忆的语言建模. Journal of Machine Learning, 3(3):300-346, 2024 年 1 月. ISSN 2790-203X. doi: 10.4208/jml.240708. URL http://dx.doi.org/10.4208/jml.240708.

[138] Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents, 2025. URL https://arxiv.org/abs/2509.24704.

[138] 张贵彬, 傅牧新, 和阎水成. Memgen: 为自我进化智能体编织生成性潜在记忆, 2025. URL https://arxiv.org/abs/2509.24704.

[139] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

[139] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, 和 Daniel Chalef. Zep: 面向智能体记忆的时间知识图架构, 2025. URL https://arxiv.org/abs/2501.13956.

[140] Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025. URL https://arxiv.org/abs/2502.12110.

[140] 徐武江, 梅凯, 高航, 谭君韬, 梁祖杰, 和张永锋. A-mem: 面向 LLM 智能体的代理记忆, 2025. URL https://arxiv.org/abs/2502.12110.

[141] Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems, 2025. URL https://arxiv.org/abs/2506.07398.

[141] Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: 用于多智能体系统的分层记忆追踪，2025。URL https://arxiv.org/abs/2506.07398.

[142] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/ abs/2504.19413.

[142] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: 构建具备可扩展长期记忆的生产就绪 AI 代理，2025。URL https://arxiv.org/ abs/2504.19413.

[143] Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Qihan Ren, Cheng Qian, Zhenhailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. A survey of self-evolving agents: On path to artificial super intelligence, 2025. URL https://arxiv.org/abs/2507.21046.

[143] Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Qihan Ren, Cheng Qian, Zhenhailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. 自我演化代理综述: 走向人工超智能之路，2025。URL https://arxiv.org/abs/2507.21046.

[144] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 8634-8652. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ 1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

[144] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: 具有口头强化学习的语言代理。收录于 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, 和 S. Levine 编辑，Advances in Neural Information Processing Systems，第 36 卷，页 8634-8652。Curran Associates, Inc., 2023。URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ 1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

[145] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 46534-46594. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/ 2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

[145] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: 基于自我反馈的迭代精化。收录于 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, 和 S. Levine 编辑，Advances in Neural Information Processing Systems，第 36 卷，页 46534-46594。Curran Associates, Inc., 2023。URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

[146] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/ forum?id=Sx038qxjek.

[146] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: 大型语言模型可通过与工具交互的批评进行自我纠正。收录于第十二届国际学习表征会议，2024。URL https://openreview.net/ forum?id=Sx038qxjek.

[147] Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10371-10393, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.607. URL https://aclanthology.org/2024.findings-emnlp.607/.

[147] Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 检索、再思考与修正: 验证链能提升基于检索的生成。收录于 Yaser Al-Onaizan, Mohit Bansal, 和 Yun-Nung Chen 编辑，Findings of the Association for Computational Linguistics: EMNLP 2024，页 10371-10393， Miami, Florida, USA, 2024 年 11 月。Association for Computational Linguistics。doi: 10.18653/v1/2024.findings-emnlp.607。URL https://aclanthology.org/2024.findings-emnlp.607/.

[148] Loka Li, Guangyi Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. CoRR, abs/2402.12563, 2024. URL https://doi.org/10.48550/arXiv.2402.12563.

[148] Loka Li, Guangyi Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric Xing, and Kun Zhang. 信心很重要: 重新审视大型语言模型的内在自我纠错能力。CoRR, abs/2402.12563, 2024。URL https://doi.org/10.48550/arXiv.2402.12563.

[149] Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Towards uncertainty-aware language agent. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 6662-6685, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.398. URL https: //aclanthology.org/2024.findings-acl.398/.

[149] Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. 面向不确定性感知的语言代理. 收录于 Lun-Wei Ku, Andre Martins, 和 Vivek Srikumar 编辑的 Findings of the Association for Computational Linguistics: ACL 2024, 页 6662-6685, 泰国曼谷, 2024 年 8 月. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.398. URL https: //aclanthology.org/2024.findings-acl.398/.

[150] Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with goal verifiers. In ICLR 2025 Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning, 2025. URL https://openreview.net/forum?id=mGAAoEWOq9.

[150] Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. 多智能体验证: 通过目标验证器扩展测试时计算. 收录于 ICLR 2025 Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning, 2025. URL https://openreview.net/forum?id=mGAAoEWOq9.

[151] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum

[151] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 教授大型语言模型自我调试. 收录于第十二届国际表征学习会议 (ICLR), 2024. URL https://openreview.net/forum?id=KuPixIqPiq.

[152] Yuxuan Guo, Shaohui Peng, Jiaming Guo, Di Huang, Xishan Zhang, Rui Zhang, Yifan Hao, Ling Li, Zikang Tian, Mingju Gao, Yutai Li, Yiming Gan, Shuai Liang, Zihao Zhang, Zidong Du, Qi Guo, Xing Hu, and Yunji Chen. Luban: Building open-ended creative agents via autonomous embodied verification. CoRR, abs/2405.15414, 2024. URL https://doi.org/10.48550/arXiv.2405.15414.

[152] Yuxuan Guo, Shaohui Peng, Jiaming Guo, Di Huang, Xishan Zhang, Rui Zhang, Yifan Hao, Ling Li, Zikang Tian, Mingju Gao, Yutai Li, Yiming Gan, Shuai Liang, Zihao Zhang, Zidong Du, Qi Guo, Xing Hu, and Yunji Chen. Luban: 通过自主具身验证构建开放式创造性代理. CoRR, abs/2405.15414, 2024. URL https://doi.org/10.48550/arXiv.2405.15414.

[153] Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. Distilling llm agent into small models with retrieval and code tools, 2025. URL https://arxiv.org/abs/2505.17612.

[153] Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. 将 llm 代理蒸馏到带检索和代码工具的小模型, 2025. URL https://arxiv.org/abs/2505.17612.

[154] Wei Xiong, Wenting Zhao, Weizhe Yuan, Olga Golovneva, Tong Zhang, Jason Weston, and Sainbayar Sukhbaatar. Stepwiser: Stepwise generative judges for wiser reasoning, 2025. URL https://arxiv.org/abs/2508.19229.

[154] Wei Xiong, Wenting Zhao, Weizhe Yuan, Olga Golovneva, Tong Zhang, Jason Weston, and Sainbayar Sukhbaatar. Stepwiser: 用于更聪明推理的分步生成式评判者, 2025. URL https://arxiv.org/abs/2508.19229.

[155] Shuofei Qiao, Zhisong Qiu, Baochang Ren, Xiaobin Wang, Xiangyuan Ru, Ningyu Zhang, Xiang Chen, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Agentic knowledgeable self-awareness. In Workshop on Reasoning and Planning for Large Language Models, 2025. URL https://openreview.net/forum?id=PGdSLjYwM

[155] Shuofei Qiao, Zhisong Qiu, Baochang Ren, Xiaobin Wang, Xiangyuan Ru, Ningyu Zhang, Xiang Chen, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 具主体性的知识型自我觉察. 收录于 Workshop on Reasoning and Planning for Large Language Models, 2025. URL https://openreview.net/forum?id=PGdSLjYwMT.

[156] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 116617-116637. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d37c9ad425fe5b65304d500c6edcba00-Paper-Conference.pdf.

[156] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 迭代推理偏好优化. 收录于 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, 和 C. Zhang 编辑的 Advances in Neural Information Processing Systems, 卷 37, 页 116617-116637. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ d37c9ad425fe5b65304d500c6edcba00-Paper-Conference.pdf.

[157] Maithili Patel, Xavier Puig, Ruta Desai, Roozbeh Mottaghi, Sonia Chernova, Joanne Truong, and Akshara Rai. Adapt: Actively discovering and adapting to preferences for any task. arXiv preprint arXiv:2504.04040, 2025.

[157] Maithili Patel, Xavier Puig, Ruta Desai, Roozbeh Mottaghi, Sonia Chernova, Joanne Truong, and Akshara Rai. Adapt: 主动发现并适应任意任务的偏好. arXiv 预印本 arXiv:2504.04040, 2025.

[158] Shuaijie She, Yu Bao, Yu Lu, Lu Xu, Tao Li, Wenhao Zhu, Shujian Huang, Shanbo Cheng, Lu Lu, and Yuxuan Wang. Dupo: Enabling reliable llm self-verification via dual preference optimization, 2025. URL https://arxiv.org/abs/2508.14460.

[158] Shuaijie She, Yu Bao, Yu Lu, Lu Xu, Tao Li, Wenhao Zhu, Shujian Huang, Shanbo Cheng, Lu Lu, and Yuxuan Wang. Dupo: 通过双重偏好优化实现可靠的 llm 自我验证, 2025. URL https://arxiv.org/abs/2508.14460.

[159] 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, 2025. URL https://arxiv.org/abs/2503.15478.

[159] Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: 在协作推理任务上训练多轮 llm 代理, 2025. URL https://arxiv.org/abs/2503.15478.

[160] Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. ACC-collab: An actor-critic approach to multi-agent LLM collaboration. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=nfKfAzkiez.

[160] Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. ACC-collab: 一种用于多智能体 LLM 协作的 actor-critic 方法。发表于第十三届国际表征学习会议，2025。URL https://openreview.net/forum?id=nfKfAzkiez.

[161] 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, 2025. URL https://arxiv.org/abs/2508.0500

[161] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: 从零数据自我进化的推理大模型，2025。URL https://arxiv.org/abs/2508.05004.

[162] Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. In A. Glober-son, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 52723-52748. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ 5e5853f35164e434015716a8c2a66543-Paper-Conference.pdf.

[162] Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 通过想象、搜索与批评迈向大模型的自我提升。载于 A. Glober-son, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, 和 C. Zhang 主编，Advances in Neural Information Processing Systems，第 37 卷，页码 52723-52748。Curran Associates, Inc.，2024。URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ 5e5853f35164e434015716a8c2a66543-Paper-Conference.pdf.

[163] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data, 2025. URL https://arxiv.org/abs/2505.03335.

[163] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: 基于零数据的强化自对弈推理，2025。URL https://arxiv.org/abs/2505.03335.

[164] 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, 2025. URL https://arxiv.org/abs/2505.14970.

[164] Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. 大模型推理的自演进课程，2025。URL https://arxiv.org/abs/2505.14970.

[165] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning, 2025. URL https://arxiv.org/abs/2504.16084.

[165] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: 测试时强化学习，2025。URL https://arxiv.org/abs/2504.16084.

[166] Dhruv Atreja. Alas: Autonomous learning agent for self-updating language models, 2025. URL https://arxiv.org/abs/2508.15805.

[166] Dhruv Atreja. ALAS: 用于自更新语言模型的自主学习代理，2025。URL https://arxiv.org/abs/2508.15805.

[167] Wanjia Zhao, Mert Yuksekgonul, Shirley Wu, and James Zou. Sirius: Self-improving multi-agent systems via bootstrapped reasoning. In Workshop on Reasoning and Planning for Large Language Models, 2025. URL https://openreview.net/forum?id=sLBSJr3hH5.

[167] Wanjia Zhao, Mert Yuksekgonul, Shirley Wu, and James Zou. SIRIUS: 通过自举推理的自我改进多智能体系统。载于大型语言模型推理与规划研讨会，2025。URL https://openreview.net/forum?id=sLBSJr3hH5.

[168] Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Rafael Rafailov, Ivan Laptev, Philip Torr, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. MALT: Improving reasoning with multi-agent LLM training. In Workshop on Reasoning and Planning for Large Language Models, 2025. URL https://openreview.net/forum?id=1If7grAC7n.

[168] Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Rafael Rafailov, Ivan Laptev, Philip Torr, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. MALT: 通过多智能体大模型训练提升推理。载于大型语言模型推理与规划研讨会，2025。URL https://openreview.net/forum?id=1If7grAC7n.

[169] Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, PeiFeng Wang, silvio savarese, Caiming Xiong, and Shafiq Joty. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. Transactions on Machine Learning Research, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=SlsZZ25InC.Survey Certification.

[169] Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, PeiFeng Wang, silvio savarese, Caiming Xiong, and Shafiq Joty. 大模型推理前沿综述: 推理尺度扩展、学习推理与智能体系统。Transactions on Machine Learning Research，2025。ISSN 2835-8856。URL https://openreview.net/forum?id=SlsZZ25InC.Survey Certification.

[170] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

[170] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math 技术报告: 通过自我提升迈向数学专家模型，2024。URL https://arxiv.org/abs/2409.12122.

[171] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

[171] 王学志, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, 和 Denny Zhou. 自洽性提高了语言模型的链式思维推理，2023. URL https://arxiv.org/abs/2203.11171.

[172] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=5Xclecx01h.

[172] 姚顺宇, 于典, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, 曹源, 和 Karthik R Narasimhan. 思维树: 使用大型语言模型的深思熟虑问题解决. 收录于第三十七届神经信息处理系统会议, 2023. URL https://openreview.net/forum?id=5Xclecx01h.

[173] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16):17682-17690, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i16.29720. URL http://dx.doi.org/10.1609/aaai.v38i16.29720.

[173] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, 和 Torsten Hoefler. 思维图: 用大型语言模型解决复杂问题. 《美国人工智能协会会刊》, 38(16):17682-17690, 2024 年 3 月. ISSN 2159-5399. doi: 10.1609/aaai.v38i16.29720. URL http://dx.doi.org/10.1609/aaai.v38i16.29720.

[174] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https: //arxiv.org/abs/2305.20050.

[174] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, 和 Karl Cobbe. 我们一步步验证吧, 2023. URL https://arxiv.org/abs/2305.20050.

[175] Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024. URL https://arxiv.org/abs/2312.08935.

[175] 王佩怡, 李磊, 邵志宏, R. X. Xu, 戴达迈, 李益飞, 陈德立, Y. Wu, 和隋志方. Math-shepherd: 在无人工注释下逐步验证并强化大语言模型, 2024. URL https://arxiv.org/abs/2312.08935.

[176] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. O1 replication journey: A strategic progress report - part 1, 2024. URL https://arxiv.org/abs/2410.18982.

> [176] 秦奕炜, 李学锋, 邹浩洋, 刘奕修, 夏世杰, 黄桢, 叶一心, 袁为哲, Hector Liu, 李远志, 和刘鹏飞. O1 复现之旅: 一份战略性进展报告——第 1 部分, 2024. URL https://arxiv.org/abs/2410.18982.

[177] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In The Second Conference on Parsimony and Learning (Recent Spotlight Track), 2025. URL https://openreview.net/forum?id=d3E3LWmTar.

> [177] 朱天哲, 翟月香, 杨季涵, 童胜邦, 谢赛宁, Sergey Levine, 和马奕. SFT 记忆化, RL 泛化: 基础模型后训练的比较研究. 收录于第二届节俭与学习会议 (近期聚焦篇), 2025. URL https://openreview.net/forum?id=d3E3LWmTar.

[178] Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. In Second Conference on Language Modeling, 2025. URL https://openreview.net/forum?id=4jdIxXBNve.

> [178] Pranjal Aggarwal 和 Sean Welleck. L1: 用强化学习控制推理模型思考的时长. 收录于第二届语言建模会议, 2025. URL https://openreview.net/forum?id=4jdIxXBNve.

[179] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 64735-64772. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/76ec4dc30e9faaf0e4b6093eaa377218-Paper-Conference.pdf.

> [179] 张丹, 周边思宁, 胡子牛, 岳一颂, 董宇晓, 和唐杰. Rest-mcts*: 通过过程奖励引导树搜索的 LLM 自我训练. 收录于 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, 和 C. Zhang 编辑的《神经信息处理系统进展》, 卷 37, 页 64735-64772. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/76ec4dc30e9faaf0e4b6093eaa377218-Paper-Conference.pdf.

[180] Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. &-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13214-13227, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.647. URL https://aclanthology.org/ 2025.ac1-long.647/.

[180] Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. &-decoding: 自适应前瞻采样以在推理时平衡探索与利用. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13214-13227, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.647. URL https://aclanthology.org/ 2025.ac1-long.647/.

[181] 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. In Second Conference on Language Modeling, 2025. URL https://openreview.net/forum?id= vSMCBUgrQj.

[181] Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun MA, and Junxian He. SimpleRL-zoo: 调查并驯服野外开放基础模型的零强化学习. In Second Conference on Language Modeling, 2025. URL https://openreview.net/forum?id= vSMCBUgrQj.

[182] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939.

[182] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 超越 80/20 法则: 高熵少数标记驱动对大模型推理的有效强化学习, 2025. URL https://arxiv.org/abs/2506.01939.

[183] 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, 2025. URL https://arxiv.org/abs/2506.08989.

[183] Xiao Liang, Zhong-Zhi Li, Yeyun Gong, Yang Wang, Hengyuan Zhang, Yelong Shen, Ying Nian Wu, and Weizhu Chen. SWS: 在大模型推理的强化学习中基于自知弱点的问题合成, 2025. URL https://arxiv.org/abs/2506.08989.

[184] 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, 2025. URL https://arxiv.org/abs/2508.14029.

[184] Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. 超越 pass@1: 变分问题合成的自我对弈维持 RLVR, 2025. URL https://arxiv.org/abs/2508.14029.

[185] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

[185] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 强化学习真的能在超越基础模型的程度上激励大模型的推理能力吗？，2025. URL https://arxiv.org/abs/2504.13837.

[186] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions, 2024. URL https://arxiv.org/abs/2411.14405.

[186] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: 面向开放性解答的开放推理模型, 2024. URL https://arxiv.org/abs/2411.14405.

[187] 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, 2024. URL https://arxiv.org/abs/2412.18925.

[187] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1，面向复杂医学推理的大模型，2024. URL https://arxiv.org/abs/2412.18925.

[188] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: Process supervision without process, 2024. URL https://arxiv.org/abs/2405.03553.

[188] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: 无需过程的过程监督, 2024. URL https://arxiv.org/abs/2405.03553.

[189] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, Chonghua Liao, and Jianhua Tao. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts, 2025. URL https://arxiv.org/abs/2411.18478.

[189] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, Chonghua Liao, and Jianhua Tao. 超越示例: 通过 MCTS 在上下文学习中实现高层次自动化推理范式, 2025. URL https://arxiv.org/abs/2411.18478.

[190] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2025. URL https://arxiv.org/abs/2411.10440.

[190] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: 让视觉-语言模型逐步推理, 2025. URL https://arxiv.org/abs/2411.10440.

[191] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking, 2025. URL https://arxiv.org/abs/2501.04519

[191] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: 小型大模型通过自我进化的深度思考掌握数学推理, 2025. URL https://arxiv.org/abs/2501.04519.

[192] Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. Reasonflux: Hierarchical llm reasoning via scaling thought templates, 2025. URL https://arxiv.org/abs/2502.06772.

[192] Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. Reasonflux: 通过扩展思维模板实现分层大模型推理，2025。URL https://arxiv.org/abs/2502.06772.

[193] Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space, 2025. URL https://arxiv.org/abs/2505.15778.

[193] Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: 在连续概念空间中释放大模型的推理潜能，2025。URL https://arxiv.org/abs/2505.15778.

[194] 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, 2024. URL https: //arxiv.org/abs/2412.06769.

[194] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 在连续潜在空间中训练大语言模型进行推理，2024。URL https: //arxiv.org/abs/2412.06769.

[195] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[195] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 技术报告，2025。URL https://arxiv.org/abs/2505.09388.

[196] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning, 2025. URL https://arxiv.org/abs/2502.18080.

[196] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 面向推理的测试时计算量最优扩展，2025。URL https://arxiv.org/abs/2502.18080.

[197] Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang.

Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning, 2025. URL https: //arxiv.org/abs/2504.01296.

[197] Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. Thinkprune: 通过强化学习精简大模型的长链式思维，2025。URL https: //arxiv.org/abs/2504.01296.

[198] Zhong-Zhi Li, Xiao Liang, Zihao Tang, Lei Ji, Peijie Wang, Haotian Xu, Xing W, Haizhen Huang, Weiwei Deng, Yeyun Gong, Zhijiang Guo, Xiao Liu, Fei Yin, and Cheng-Lin Liu. Tl;dr: Too long, do reweighting for efficient llm reasoning compression, 2025. URL https://arxiv.org/abs/2506.02678.

[198] Zhong-Zhi Li, Xiao Liang, Zihao Tang, Lei Ji, Peijie Wang, Haotian Xu, Xing W, Haizhen Huang, Weiwei Deng, Yeyun Gong, Zhijiang Guo, Xiao Liu, Fei Yin, and Cheng-Lin Liu. Tl;dr: 太长，重加权以实现高效大模型推理压缩，2025。URL https://arxiv.org/abs/2506.02678.

[199] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2 + 3 = ? on the overthinking of long reasoning models. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=MSbU3L7V00.

[199] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 不要对2 + 3 = ? 过度思考: 关于长推理模型过度思考的问题。发表于第四十二届国际机器学习大会，2025。URL https://openreview.net/forum?id=MSbU3L7V00.

[200] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[200] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: 一系列高能力多模态模型。arXiv 预印本 arXiv:2312.11805，2023。

[201] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892-34916, 2023.

[201] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning。神经信息处理系统进展，36:34892-34916，2023。

[202] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[202] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: 提升视觉-语言模型对任意分辨率世界感知的能力。arXiv 预印本 arXiv:2409.12191, 2024.

[203] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language

models. arXiv preprint arXiv:2410.10855, 2024.

[203] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. 多模态语言模型中的核心知识缺陷。arXiv 预印本 arXiv:2410.10855, 2024.

[204] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 24185-24198, 2024.

[204] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: 扩展视觉基础模型并为通用视觉-语言任务对齐。收入 IEEE/CVF 计算机视觉与模式识别大会论文集，第 24185-24198 页，2024.

[205] OpenAI. Gpt-4v(ision) system card. System card, OpenAI, September 2023. URL https://cdn.openai.com/papers/GP

[205] OpenAI. GPT-4V(ision) 系统说明。系统说明，OpenAI，2023 年 9 月。网址 https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[206] Wanpeng Zhang, Yicheng Feng, Hao Luo, Yijiang Li, Zihao Yue, Sipeng Zheng, and Zongqing Lu. Unified multimodal understanding via byte-pair visual encoding. arXiv preprint arXiv:2506.23639, 2025.

[206] Wanpeng Zhang, Yicheng Feng, Hao Luo, Yijiang Li, Zihao Yue, Sipeng Zheng, and Zongqing Lu. 通过字节对视觉编码实现统一的多模态理解。arXiv 预印本 arXiv:2506.23639, 2025.

[207] Wanpeng Zhang, Zilong Xie, Yicheng Feng, Yijiang Li, Xingrun Xing, Sipeng Zheng, and Zongqing Lu. From pixels to tokens: Byte-pair encoding on quantized visual modalities. arXiv preprint arXiv:2410.02155, 2024.

[207] Wanpeng Zhang, Zilong Xie, Yicheng Feng, Yijiang Li, Xingrun Xing, Sipeng Zheng, and Zongqing Lu. 从像素到标记: 在量化视觉模态上应用字节对编码。arXiv 预印本 arXiv:2410.02155, 2024.

[208] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37: 8612-8642, 2024.

[208] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hong-sheng Li. Visual CoT: 通过一个用于链式思维推理的综合数据集和基准，推进多模态语言模型。NeurIPS, 37: 8612-8642, 2024.

[209] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023.

[209] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 语言模型中的多模态链式思维推理。arXiv 预印本 arXiv:2302.00923, 2023.

[210] 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, 2025.

[210] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. 用图像思考以实现多模态推理: 基础、方法与未来前沿。arXiv 预印本 arXiv:2506.23918, 2025.

[211] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. arXiv preprint arXiv:2504.21277, 2025.

[211] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced MLLM: 关于基于强化学习的多模态大语言模型推理的综述。arXiv 预印本 arXiv:2504.21277, 2025.

[212] 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, 2025. URL https://arxiv.org/abs/2508.08189.

[212] Weijia Wu, Chen Gao, Joya Chen, Kevin Qinghong Lin, Qingwei Meng, Yiming Zhang, Yuke Qiu, Hong Zhou, and Mike Zheng Shou. 视觉领域的强化学习综述，2025。网址 https://arxiv.org/abs/2508.08189.

[213] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025.

[213] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi K1.5: 用大语言模型扩展强化学习。arXiv 预印本 arXiv:2501.12599, 2025.

[214] 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, 2025.

[214] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. VLM-R1: 一种稳定且可泛化的 R1 风格大型视觉-语言模型。arXiv 预印本 arXiv:2504.07615, 2025.

[215] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536, 2025.

[215] 彭英哲, 张功瑞, 张妙森, 游智远, 刘杰, 朱其鹏, 杨凯, 许兴中, 耿鑫, 与杨旭. Lmm-r1: 通过两阶段基于规则的强化学习赋能 3b 规模 LMMs 的强大推理能力. arXiv 预印本 arXiv:2503.07536, 2025.

[216] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. arXiv preprint arXiv:2411.10442, 2024.

[216] 王伟云, 陈哲, 王文海, 曹越, 刘扬舟, 高章炜, 朱金国, 祝锡洲, 陆乐为, 乔宇, 与戴吉峰. 通过混合偏好优化提升多模态大语言模型的推理能力. arXiv 预印本 arXiv:2411.10442, 2024.

[217] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 9062-9072, 2025.

[217] 董宇豪, 刘祖燕, 孙海龙, 杨景康, 胡温斯顿, 饶永明, 与刘子维. Insight-v: 用多模态大语言模型探索长链视觉推理. 载于计算机视觉与模式识别会议论文集, 页 9062-9072, 2025.

[218] Linghao Zhu, Yiran Guan, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang Liu, and Xiang Bai. Shuffle-r1: Efficient rl framework for multimodal large language models via data-centric dynamic shuffle, 2025. URL https://arxiv.org/abs/2508.05612.

[218] 朱凌昊，管一然，梁定康，鞠建中，罗振博，秦斌，栾建，刘玉良，与白湘. Shuffle-r1: 通过以数据为中心的动态洗牌实现多模态大语言模型的高效强化学习框架，2025. URL https://arxiv.org/abs/2508.05612.

[219] 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, 2025.

[219] 刘子渝, 孙泽奕, 臧雨航, 董晓伊, 曹宇航, 段浩东, 林大华, 与王佳祺. Visual-rft: 视觉强化微调. arXiv 预印本 arXiv:2503.01785, 2025.

[220] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. arXiv preprint arXiv:2503.20752, 2025.

[220] 谭华杰, 吉玉衡, 号肖帅, 林明兰, 王鹏威, 王中元, 与张尚航. Reason-rft: 用于视觉推理的强化微调. arXiv 预印本 arXiv:2503.20752, 2025.

[221] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. arXiv preprint arXiv:2505.15804, 2025.

[221] 李宗钊, 马宗阳, 李明泽, 李松友, 荣宇, 徐廷阳, 张子棋, 赵得立, 与黄文兵. Star-r1: 通过强化多模态 LLMs 实现空间变换推理. arXiv 预印本 arXiv:2505.15804, 2025.

[222] 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, 2025.

[222] 黄文轩, 贾博涵, 翟子杰, 曹少圣, 叶哲宇, 赵非, 徐喆, 胡尧, 与林少辉. Vision-r1: 激励多模态大语言模型的推理能力. arXiv 预印本 arXiv:2503.06749, 2025.

[223] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. CoRR, 2025.

[223] 孟凡庆, 杜凌霄, 刘宗凯, 周志翔, 卢泉峰, 傅道成, 石博天, 王文海, 何俊俊, 张凯鹏, 等. Mm-eureka: 用基于规则的大规模强化学习探索视觉顿悟时刻. CoRR, 2025.

[224] Peiyao Wang and Haibin Ling. Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. arXiv preprint arXiv:2506.01371, 2025.

[224] 王培尧与凌海滨. Svqa-r1: 通过视图一致性奖励优化加强 MLLMs 的空间推理. arXiv 预印本 arXiv:2506.01371, 2025.

[225] Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14677.

[225] 夏嘉儿, 臧雨航, 高鹏, 李一萱, 与周凯阳. Visionary-r1: 通过强化学习缓解视觉推理中的捷径问题, 2025. URL https://arxiv.org/abs/2505.14677.

[226] Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning, 2025. URL https://arxiv.org/abs/2506

[226] 杨杰, 马飞鹏, 王子天, 尹大成, 戎康, 饶峰云, 与张瑞茂. Wethink: 通过强化学习迈向通用视觉-语言推理, 2025. URL https://arxiv.org/abs/2506.07905.

[227] 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, 2025.

[227] 陈亮, 高宏成, 刘天宇, 黄志琦, Flood Sung, 周欣宇, 吴宇欣, 与常宝宝. G1: 通过强化学习引导视觉-语言模型的感知与推理能力引导提升. arXiv 预印本 arXiv:2505.13426, 2025.

[228] Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, et al. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. arXiv preprint arXiv:2506.01078, 2025.

[228] 詹宇飞, 吴子恒, 朱有松, 薛荣坤, 罗瑞浦, 陈正昊, 张灿, 李一帆, 何振涛, 杨哲明, 等. Gthinker: 通过线索引导的再思考迈向通用多模态推理. arXiv 预印本 arXiv:2506.01078, 2025.

[229] Zirun Guo, Minjie Hong, and Tao Jin. Observe-r1: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning. arXiv preprint arXiv:2505.12432, 2025.

[229] 郭子润, 洪敏杰, 晋涛. Observe-r1: 通过动态渐进强化学习解锁 MLLMs 的推理能力. arXiv 预印本 arXiv:2505.12432, 2025.

[230] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's ”aha moment” in visual reasoning on a 2b non-sft model, 2025. URL https://arxiv.org/abs/ 2503.05132.

[230] 周恒广, 李熙瑞, 王若宸, 程敏豪, 周天意, 谢卓睿. R1-zero 在 2B 非 SFT 模型上视觉推理的“恍然时刻”, 2025. URL https://arxiv.org/abs/2503.05132.

[231] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. arXiv preprint arXiv:2508.19652, 2025.

[231] 李宗夏, 余文豪, 黄成松, 刘锐, 梁振文, 刘福啸, 车景熙, 于典, Jordan Boyd-Graber, 米海涛, 等. 通过推理分解的自我奖励视觉-语言模型. arXiv 预印本 arXiv:2508.19652, 2025.

[232] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In European Conference on Computer Vision, pages 792-807. Springer, 2016.

[232] Varun K Nagaraja, Vlad I Morariu, Larry S Davis. 在指代表达理解中建模物体间的上下文. 见欧洲计算机视觉大会, 页码 792-807. Springer, 2016.

[233] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11-20, 2016.

[233] 毛俊华, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, Kevin Murphy. 不含歧义的物体描述的生成与理解. 见 IEEE 视觉与模式识别大会论文集, 页码 11-20, 2016.

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

[234] 樊悦, 何雪海, 杨迪霁, 郑凯之, Ching-Chen Kuo, 郑玉婷, Sravana Jyothi Narayanaraju, 管鑫泽, 王新 Eric. GRIT: 教会 MLLMs 用图像思考. arXiv 预印本 arXiv:2505.15879, 2025.

[235] 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.

[235] 钟志万, 金俊赫, 金世烷, 李在永, 金敏洙, 余英在. 不要只看一次: 朝向具有选择性视觉重访的多模态交互式推理. arXiv 预印本 arXiv:2505.18842, 2025.

[236] 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.

[236] 曹蒙, 赵浩泽, 张灿, 常晓军, Ian Reid, 梁晓丹. Ground-r1: 通过强化学习激励有根的视觉推理. arXiv 预印本 arXiv:2505.20272, 2025.

[237] 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, 2025.

[237] 楚旭, 陈新荣, 王冠宇, 谭志杰, 黄奎, 吕文宇, 莫彤, 李伟平. Qwen look again: 引导视觉-语言推理模型重新关注视觉信息. arXiv 预印本 arXiv:2505.23558, 2025.

[238] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14953-14962, 2023.

[238] Tanmay Gupta, Aniruddha Kembhavi. 视觉编程: 无需训练的组合式视觉推理. 见 IEEE/CVF 视觉与模式识别大会论文集, 页码 14953-14962, 2023.

[239] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. arXiv preprint arXiv:2507.07998, 2025.

[239] 赵世田, 张浩泉, 林少衡, 李鸣, 吴启龙, 张凯鹏, 韦辰. Pyvision: 具有动态工具能力的能动视觉. arXiv 预印本 arXiv:2507.07998, 2025.

[240] Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection. arXiv preprint arXiv:2505.20289, 2025.

[240] 黄泽逸, 季宇阳, Anirudh Sundara Rajan, 蔡泽帆, 肖文, 王浩翰, 胡俊杰, Yong Jae Lee. VisualToolAgent (VISTA): 用于视觉工具选择的强化学习框架. arXiv 预印本 arXiv:2505.20289, 2025.

[241] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. arXiv preprint arXiv:2505.19255, 2025.

[241] 吴明远, 杨靖成, 姜济泽, 李美棠, 闫凯卓, 余汉超, 张敏佳, 翟承响, Klara Nahrstedt. VTool-r1: VLMS 通过多模态工具使用的强化学习学会用图像思考. arXiv 预印本 arXiv:2505.19255, 2025.

[242] 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, 2025.

[242] 宿昭晨, 李林杰, 宋明洋, 郝云卓, 杨正元, 张俊, 陈冠杰, 顾嘉伟, 李军陶, 曲晓烨, 等. Openthinkimg: 通过视觉工具强化学习学习用图像思考. arXiv 预印本 arXiv:2505.08617, 2025.

[243] 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, 2025.

[243] 徐奕, 李承祖, 周涵, 萬星晨, 张蔡奇, Anna Korhonen, 和 Ivan Vulić. 视觉规划: 只用图像来思考. arXiv 预印本 arXiv:2505.11409, 2025.

[244] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639, 2025.

[244] 方荣尧, 段成琦, 王坤, 黄林江, 李浩, 严世临, 田浩, 曾兴宇, 赵锐, 戴继峰, 等. GOT: 释放多模态大语言模型在视觉生成与编辑中的推理能力. arXiv 预印本 arXiv:2503.10639, 2025.

[245] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542, 2025.

[245] 李承祖, 吴文山, 张环宇, 夏岩, 毛少光, 董力, Ivan Vulić, 和魏福如. 在空间中边推理边想象: 多模态思维可视化. arXiv 预印本 arXiv:2501.07542, 2025.

[246] 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.

[246] 段成琦, 方荣尧, 王玉庆, 王坤, 黄林江, 曾兴宇, 李宏胜, 和刘希辉. GOT-R1: 通过强化学习释放 MLLM 在视觉生成方面的推理能力. arXiv 预印本 arXiv:2505.17022, 2025.

[247] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuoqłan 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, 2025.

[247] 蒋东之, 郭子昱, 张任睿, 宗卓澜, 李昊, 竺乐, 严世临, Pheng-Ann Heng, 和李宏胜. T2I-R1: 通过协同的语义级和标记级 CoT 强化图像生成. arXiv 预印本 arXiv:2505.00703, 2025.

[248] Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. arXiv preprint arXiv:2504.15900, 2025.

[248] 温成, 郭廷伟, 赵帅江, 邹炜, 和李宪纲. SARI: 通过课程引导的强化学习进行结构化音频推理. arXiv 预印本 arXiv:2504.15900, 2025.

[249] Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. Soundmind: Rl-incentivized logic reasoning for audio-language models. arXiv preprint arXiv:2506.12935, 2025.

[249] 刁兴健, 张春晖, 孔可怡, 吴伟毅, 马池宇, 欧阳中宇, 亓培君, Soroush Vosoughi, 和桂江. SoundMind: 以强化学习激励的逻辑推理用于音频-语言模型. arXiv 预印本 arXiv:2506.12935, 2025.

[250] Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. arXiv preprint

arXiv:2503.11197, 2025.

[250] 李刚, 刘继忠, Heinrich Dinkel, 牛亚东, 张军博, 和栾健. 强化学习优于监督微调: 以音频问答为例. arXiv 预印本 arXiv:2503.11197, 2025.

[251] Yinghao Aaron Li, Xilin Jiang, Fei Tao, Cheng Niu, Kaifeng Xu, Juntong Song, and Nima Mesgarani. Dmospeech 2: Reinforcement learning for duration prediction in metric-optimized speech synthesis. arXiv preprint arXiv:2507.14988, 2025.

[251] 李英浩 Aaron, 蒋曦林, 陶飞, 牛诚, 许凯峰, 宋俊通, 和 Nima Mesgarani. DMOSpeech 2: 用于度量优化语音合成的时长预测强化学习. arXiv 预印本 arXiv:2507.14988, 2025.

[252] Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. arXiv preprint arXiv:2505.04623, 2025.

[252] 邢正豪, 胡晓炜, 傅志宏, 汪文海, 戴继峰, 和 Pheng-Ann Heng. EchoInk-R1: 通过强化学习探索多模态大语言模型中的视听推理. arXiv 预印本 arXiv:2505.04623, 2025.

[253] Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. EPO: Explicit policy optimization for strategic reasoning in LLMs via reinforcement learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15371-15396, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.747. URL https://aclanthology.org/2025.acl-long.747/.

[253] Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. EPO: 通过强化学习对大型语言模型进行策略显式优化以实现策略性推理。发表于 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, 和 Mohammad Taher Pilehvar 主编，第三十六届计算语言学协会年会论文集 (第一卷: 长文)，第 15371-15396 页，奥地利维也纳，2025 年 7 月。计算语言学协会。ISBN 979-8-89176-251-0。doi: 10.18653/v1/2025.acl-long.747。URL https://aclanthology.org/2025.acl-long.747/。

[254] Ilgee Hong, Changlong Yu, Liang Qiu, Weixiang Yan, Zhenghao Xu, Haoming Jiang, Qingru Zhang, Qin Lu, Xin Liu, Chao Zhang, and Tuo Zhao. Think-rm: Enabling long-horizon reasoning in generative reward models, 2025. URL https://arxiv.org/abs/2505.16265.

[254] Ilgee Hong, Changlong Yu, Liang Qiu, Weixiang Yan, Zhenghao Xu, Haoming Jiang, Qingru Zhang, Qin Lu, Xin Liu, Chao Zhang, and Tuo Zhao. Think-rm: 在生成式奖励模型中实现长周期推理，2025 年。URL https://arxiv.org/abs/2505.16265。

[255] 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, 2025. URL https://arxiv.org/abs/ 2505.23564.

[255] Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: 用于大型语言模型强化学习的有效段级信用分配，2025 年。URL https://arxiv.org/abs/ 2505.23564。

[256] Sanjiban Choudhury. Process reward models for llm agents: Practical framework and directions, 2025. URL https://arxiv.org/abs/2502.10325.

[256] Sanjiban Choudhury. Process reward models for llm agents:LLM 代理的过程奖励模型——实用框架与方向，2025 年。URL https://arxiv.org/abs/2502.10325。

[257] Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents, 2025. URL https: //arxiv.org/abs/2507.22844.

[257] Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. Rlvmr: 带有可验证元推理奖励的强化学习以实现稳健长时程智能体, 2025. URL https: //arxiv.org/abs/2507.22844.

[258] Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: Segment-level direct preference optimization for social agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12409-12423, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

[258] Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: 面向社交代理的分段级直接偏好优化. 收录于 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar 编辑, 第 63 届计算语言学协会年会论文集 (第 1 卷: 长篇论文), 页 12409-12423, 维也纳, 奥地利, 2025 年 7 月. 计算语言学协会. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

[259] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

[259] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 面向大型语言模型的检索增强生成综述, 2024. URL https://arxiv.org/abs/2312.10997.

[260] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. KDD ’24, page 6491-6501, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

[260] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 关于 RAG 驱动的会议型大模型的综述: 迈向检索增强的大型语言模型. KDD ’24, 页 6491-6501, 纽约, NY, 美国, 2024. 计算机械协会. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

[261] Perplexity. Perplexity deep research. https://www.perplexity.ai/hub/blog/ introducing-perplexity-deep-research, 2025.

[261] Perplexity. Perplexity 深度研究. https://www.perplexity.ai/hub/blog/ introducing-perplexity-deep-research, 2025.

[262] Wentao Zhang, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving, 2025. URL https://arxiv.org/abs/2506.12508.

[262] Wentao Zhang, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. Agentorchestra: 用于通用任务解决的分层多智能体框架, 2025. URL https://arxiv.org/abs/2506.12508.

[263] Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. CoRR, abs/2503.00223, March 2025. URL https://doi.org/ 10.48550/arXiv.2503.00223.

[263] Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: 通过强化学习利用大型语言模型攻破真实搜索引擎与检索器. CoRR, abs/2503.00223, 2025 年 3 月. URL https://doi.org/ 10.48550/arXiv.2503.00223.

[264] 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, 2025. URL https://arxiv.org/abs/2503.09516.

[264] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: 训练大模型推理并借助搜索引擎的强化学习方法, 2025. URL https://arxiv.org/abs/2503.09516.

[265] 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, 2025. URL https://arxiv.org/abs/2503.05592.

[265] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: 通过强化学习激励大模型的搜索能力, 2025. URL https://arxiv.org/abs/2503.05592.

[266] 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, 2025. URL https://arxiv.org/abs/ 2505.17005.

[266] 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++: 通过强化学习激励大模型的动态知识获取, 2025. URL https://arxiv.org/abs/ 2505.17005.

[267] 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, 2025. URL https://arxiv.org/abs/2505.15107

[267] Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: 通过逐步近端策略优化点燃大模型的搜索能力，2025. URL https://arxiv.org/abs/2505.15107.

[268] 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, 2025. URL https://arxiv.org/abs/2504.03160.

[268] 郑玉翔、傅大元、胡翔坤、蔡晓杰、叶柳曼山、卢鹏睿、刘鹏飞。Deepresearcher: 通过强化学习在真实环境中扩展深度研究，2025。URL https://arxiv.org/abs/2504.03160.

[269] 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, 2025. URL https://arxiv.org/abs/2504.21776.

[269] 李晓熙、金嘉杰、董冠廷、钱宏进、朱玉涛、吴永康、温纪荣、窦志成。Webthinker: 赋能大规模推理模型的深度研究能力，2025。URL https://arxiv.org/abs/2504.21776.

[270] Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webwatcher: Breaking new frontiers of vision-language deep research agent, 2025. URL https://arxiv.org/ abs/2508.05748.

[270] 耿新宇、夏鹏、张臻、王昕宇、王秋晨、丁瑞雪、王晨曦、吴嘉龙、赵艺达、李宽、蒋勇、谢鹏军、黄飞、周靖人。Webwatcher: 开拓视觉-语言深度研究代理的新前沿，2025。URL https://arxiv.org/abs/2508.05748.

[271] Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webshaper: Agentically data synthesizing via information-seeking formalization, 2025. URL https://arxiv.org/abs/ 2507.15061.

[271] 陶正伟、吴嘉龙、尹文彪、张峻楷、李百轩、沈海阳、李宽、张立文、王昕宇、蒋勇、谢鹏军、黄飞、周靖人。Webshaper: 通过信息寻求形式化进行代理式数据合成，2025。URL https://arxiv.org/abs/2507.15061.

[272] Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, Yang Qin, Yuan Wang, Quanxing Zha, Sunhao Dai, and Changhua Meng. Atom-searcher: Enhancing agentic deep research via fine-grained atomic thought reward, 2025. URL https://arxiv.org/abs/2508.12800.

[272] 邓勇、王国清、应振哲、吴晓峰、林金震、熊文文、戴玉琴、杨硕、张占伟、王其文、秦扬、王远、查全兴、戴孙浩、孟昌华。Atom-searcher: 通过精细原子化思考奖励增强代理式深度研究，2025。URL https://arxiv.org/abs/2508.12800.

[273] MiroMind Team. Miromind open deep research v0.1: A high-performance, fully open-sourced deep research project that grows with developers, August 2025. URL https://miromind.ai/blog/ miromind-open-deep-research. Blog post.

[273] MiroMind 团队。Miromind open deep research v0.1: 一个高性能、完全开源并随开发者成长的深度研究项目，2025 年 8 月。URL https://miromind.ai/blog/ miromind-open-deep-research。博客文章。

[274] Shuang Sun, Huatong Song, Yuhao Wang, Ruiyang Ren, Jinhao Jiang, Junjie Zhang, Fei Bai, Jia Deng, Wayne Xin Zhao, Zheng Liu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. Simpledeepsearcher: Deep information seeking via web-powered reasoning trajectory synthesis, 2025. URL https: //arxiv.org/abs/2505.16834.

[274] 孙双、宋华彤、王宇浩、任瑞阳、江锦豪、张俊杰、白飞、邓家、赵新鑫、刘正、方磊、王中远、温纪荣。Simpledeepsearcher: 通过网络驱动的推理轨迹合成实现深度信息寻求，2025。URL https: //arxiv.org/abs/2505.16834.

[275] Chengyue Yu, Siyuan Lu, Chenyi Zhuang, Dong Wang, Qintong Wu, Zongyue Li, Runsheng Gan, Chunfeng Wang, Siqi Hou, Gaochi Huang, Wenlong Yan, Lifeng Hong, Aohui Xue, Yanfeng Wang, Jinjie Gu, David Tsai, and Tao Lin. Aworld: Orchestrating the training recipe for agentic ai, 2025. URL https://arxiv.org/abs/2508.20404.

[275] 余成岳、卢思远、庄陈逸、王东、吴勤彤、李宗跃、甘润升、王春峰、侯思奇、黄高驰、严文龙、洪立峰、薛奥辉、王艳峰、顾金杰、蔡大卫、林涛。Aworld: 为代理式人工智能编排训练方案，2025。URL https://arxiv.org/abs/2508.20404.

[276] 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.

[276] Nguyen Xuan-Phi、Shrey Pandit、Revanth Gangi Reddy、Austin Xu、Silvio Savarese、熊彩明、Shafiq Joty。Sfr-deepresearch: 迈向对自主推理单体代理有效的强化学习，2025。URL https://arxiv.org/abs/2509.06283.

[277] 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, 2025. URL https://arxiv.org/abs/2505.04588.

[277] 孙昊、乔子乐、郭佳妍、范轩博、侯颖妍、蒋勇、谢鹏军、张岩、黄飞、周靖人。Zerosearch: 激励大模型的搜索能力而无需实际搜索，2025。URL https://arxiv.org/abs/2505.04588.

[278] Yuchen Fan, Kaiyan Zhang, Heng Zhou, Yuxin Zuo, Yanxu Chen, Yu Fu, Xinwei Long, Xuekai Zhu, Che Jiang, Yuchen Zhang, Li Kang, Gang Chen, Cheng Huang, Zhizhou He, Bingning Wang, Lei Bai, Ning Ding, and Bowen Zhou. Ssrl: Self-search reinforcement learning, 2025. URL https: //arxiv.org/abs/2508.10874.

[278] 范宇晨、张凯岩、周恒、左宇昕、陈衍旭、傅宇、龙新伟、朱学凯、姜策、张宇宸、康立、陈刚、黄成、贺志舟、王炳宁、白磊、丁宁、周博文。Ssrl: 自我搜索强化学习，2025。URL https: //arxiv.org/abs/2508.10874.

[279] Google. Gemini deep research. https://gemini.google/overview/deep-research/, 2025.

[279] Google。Gemini deep research。https://gemini.google/overview/deep-research/, 2025.

[280] x.ai. Grok 3 beta —the age of reasoning agents, 2025. URL https://x.ai/news/grok-3.

[280] x.ai. Grok 3 测试版—推理代理的时代，2025。URL https://x.ai/news/grok-3.

[281] ByteDance Doubao. Doubao, 2025. URL http://www.doubao.com/.

[281] 字节跳动夺宝。夺宝，2025。URL http://www.doubao.com/.

[282] Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. Search and refine during think: Facilitating knowledge refinement for improved retrieval-augmented reasoning, 2025. URL https://arxiv.org/abs/2505.11277.

[282] 时耀睿，李思航，吴畅，刘志远，方俊锋，蔡恒兴，张安，和王翔。思考中的搜索与精炼: 促进知识精炼以改进检索增强推理，2025。URL https://arxiv.org/abs/2505.11277.

[283] 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, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

[283] Grégoire Mialon，Clémentine Fourrier，Thomas Wolf，Yann LeCun，和 Thomas Scialom。GAIA: 通用 AI 助手基准。在第十二届国际学习表征会议，2024。URL https://openreview.net/forum?id=fibxvahvs3.

[284] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking LLMs in web traversal. In Workshop on Reasoning and Planning for Large Language Models, 2025. URL https: //openreview.net/forum?id=cVI9lAfkuK.

[284] 吴嘉龙，殷文彪，江勇，王正林，席泽坤，方润南，张林海，何玉兰，周德宇，谢鹏军，和黄飞。Webwalker: 在网页遍历中对大模型进行基准测试。在用于大型语言模型的推理与规划研讨会，2025。URL https: //openreview.net/forum?id=cVI9lAfkuK.

[285] 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, 2025. URL https://arxiv.org/abs/2504.12516.

[285] Jason Wei，孙志清，Spencer Papay，Scott McKinney，Jeffrey Han，Isa Fulford，Chung Hyung Won，Alex Tachard Passos，William Fedus，和 Amelia Glaese。Browsecomp: 一个简单却具有挑战性的浏览代理基准，2025。URL https://arxiv.org/abs/2504.12516.

[286] Yihong Dong, Xue Jiang, Jiaru Qian, Tian Wang, Kechi Zhang, Zhi Jin, and Ge Li. A survey on code generation with llm-based agents, 2025. URL https://arxiv.org/abs/2508.00083.

[286] 董逸鸿，蒋学，钱嘉儒，王天，张可驰，晋志，和李革。基于大模型代理的代码生成综述，2025。URL https://arxiv.org/abs/2508.00083.

[287] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In The Twelfth International Conference on Learning Representations, 2024. URL https: //openreview.net/forum?id=VtmBAGCN70.

[287] 洪思睿,朱格晨,Jonathan Chen,郑夏武,程宇衡,王金林,张策尧,王子立,Yau Steven Ka Shing,林子娟,周立阳,冉辰宇,肖凌峰,吴承林,和 Jürgen Schmidhuber。MetaGPT: 用于多代理协作框架的元编程。在第十二届国际学习表征会议，2024。URL https: //openreview.net/forum?id=VtmBAGCN70.

[288] Significant Gravitas. AutoGPT: Autonomous gpt-4 agent framework. GitHub, MIT License, 3 2023. URL https://github.com/Significant-Gravitas/AutoGPT.Initial release date.

[288] Significant Gravitas。AutoGPT: 自主 gpt-4 代理框架。GitHub，MIT 许可证，2023 年 3 月。URL https://github.com/Significant-Gravitas/AutoGPT.Initial release date.

[289] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=EHg5GDnyq1.

[289] 陈维泽，苏玉升，左敬伟，杨成，袁晨飞，陈志民，喻和洋，陆雅希，洪逸鑫，钱晨，覃宇佳，丛鑫，谢若冰，刘志远，孙茂松，和周杰。Agentverse: 促进多代理协作并探索涌现行为。在第十二届国际学习表征会议，2024。URL https://openreview.net/forum?id=EHg5GDnyq1.

[290] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep 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 21314-21328. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/202
8636419dealaa9fbd25fc4248e702da4-Paper-Conference.pdf.

[290] Hung Le，王越，Akhilesh Deepak Gotmare，Silvio Savarese，和 Steven Chu Hong Hoi。CoderL: 通过预训练模型与深度强化学习掌握代码生成。收录于 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, 和 A. Oh 主编，Advances in Neural Information Processing Systems，卷 35，页 21314-21328。Curran Associates, Inc., 2022。URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8636419dealaa9fbd25fc4248e702da4-Paper-Conference.pdf.

[291] Huaye Zeng, Dongfu Jiang, Haozhe Wang, Ping Nie, Xiaotong Chen, and Wenhu Chen. ACECODER: Acing coder RL via automated test-case synthesis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12023-12040, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025. acl-long.587. URL https://aclanthology.org/2025.acl-long.587/.

[291] 曾华烨, 姜东甫, 王昊哲, 聂平, 陈晓彤, 和陈文虎。ACECODER: 通过自动化测试用例合成提升编码器的强化学习。收录于 Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar 主编,《第 63 届计算语言学协会年会论文集 (第一卷: 长篇论文)》, 页码 12023-12040, 奥地利维也纳, 2025 年 7 月。计算语言学协会。ISBN 979-8-89176-251-0。doi: 10.18653/v1/2025.acl-long.587。URL https://aclanthology.org/2025.acl-long.587/。

[292] 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-03-mini-Level-1cf81902c14680b3beesbe 2025. Notion Blog.

[292] 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，和 Ion Stoica。Deepcoder: 一个在 o3-mini 级别的完全开源 14b 编码器。https://pretty-radio-b75.notion.site/ DeepCoder-A-Fully-Open-Source-14B-Coder-at-03-mini-Level-1cf81902c14680b3beesbe 2025。Notion 博客。

[293] Jiate Liu, Yiqin Zhu, Kaiwen Xiao, QIANG FU, Xiao Han, Yang Wei, and Deheng Ye. RLTF: Reinforcement learning from unit test feedback. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=hjYmsV6nXZ.

[293] 刘佳特, 朱艺勤, 肖凯文, 傅强, 韩晓, 魏扬, 和叶德恒。RLTF: 来自单元测试反馈的强化学习。《机器学习研究汇刊》，2023。ISSN 2835-8856。URL https://openreview.net/forum?id=hjYmsV6nXZ。

[294] Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. Co-evolving llm coder and unit tester via reinforcement learning, 2025. URL https://arxiv.org/abs/2506.03136.

[294] 王银杰，杨凌，田野，沈珂，和王梦迪。通过强化学习协同进化大模型编码器与单元测试器，2025。URL https://arxiv.org/abs/2506.03136。

[295] Chuanhao Yan, Fengdi Che, Xuhan Huang, Xu Xu, Xin Li, Yizhi Li, Xingwei Qu, Jingzhe Shi, Zhuangzhuang He, Chenghua Lin, et al. Re: Form-reducing human priors in scalable formal software verification with rl in llms: A preliminary study on dafny. arXiv preprint arXiv:2507.16331, 2025. URL https://arxiv.org/abs/2507.163

[295] 严传浩, 车峰迪, 黄绪涵, 徐旭, 李鑫, 李一志, 瞿兴威, 石景哲, 何庄庄, 林成华, 等。Re: 在可扩展形式化软件验证中以 RL 减少人类先验: 在 Dafny 上的初步研究。arXiv 预印本 arXiv:2507.16331，2025。URL https://arxiv.org/abs/2507.16331。

[296] Yunlong Feng, Yang Xu, Xiao Xu, Binyuan Hui, and Junyang Lin. Towards better correctness and efficiency in code generation, 2025. URL https://arxiv.org/abs/2508.20124.

[296] 冯云龙，许洋，许晓，惠斌元，和林俊阳。朝着更好的代码生成正确性与效率迈进，2025。URL https://arxiv.org/abs/2508.20124。

[297] Shihan Dou, Yan Liu, Haoxiang Jia, Enyu Zhou, Limao Xiong, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, Zhiheng Xi, Yuhao Zhou, Tao Ji, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang.

StepCoder: Improving code generation with reinforcement learning from compiler feedback. 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 4571-4585, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long. 251. URL https://aclanthology.org/2024.acl-long.251/.

[297] 窦世涵，刘岩，贾浩翔，周恩宇，熊黎茂，单俊杰，黄才双，王晓，范晓然，席志恒，周宇昊，季涛，郑睿，张琦，桂涛，和黄轩京。StepCoder: 通过来自编译器反馈的强化学习改进代码生成。收录于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 主编，《第 62 届计算语言学协会年会论文集 (第一卷: 长篇论文)》，页码 4571-4585，泰国曼谷，2024 年 8 月。计算语言学协会。doi: 10.18653/v1/2024.acl-long.251。URL https://aclanthology.org/2024.acl-long.251/。

[298] Sijie Wang, Quanjiang Guo, Kai Zhao, Yawei Zhang, Xin Li, Xiang Li, Siqi Li, Rui She, Shangshu Yu, and Wee Peng Tay. Codeboost: Boosting code llms by squeezing knowledge from code snippets with rl, 2025. URL https://arxiv.org/abs/2508.05242.

[298] 王思杰，郭权江，赵凯，张雅薇，李鑫，李翔，李思琪，佘睿，俞上澍，和 Tay Wee Peng。Codeboost: 通过从代码片段中挤出知识并使用强化学习提升代码大模型，2025。URL https://arxiv.org/abs/2508.05242。

[299] Yufan Ye, Ting Zhang, Wenbin Jiang, and Hua Huang. Process-supervised reinforcement learning for code generation, 2025. URL https://arxiv.org/abs/2502.01715.

[299] 叶宇凡，张廷，江文斌，和黄华。面向代码生成的过程监督强化学习，2025。URL https://arxiv.org/abs/2502.01715。

[300] Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. o1-coder: an o1 replication for coding. arXiv preprint arXiv:2412.00154, 2024. URL https:// arxiv.org/abs/2412.00154.

[300] 张昱翔，吴尚熙，杨雨琦，舒江明，肖锦林，孔超，和桑吉涛。o1-coder: 一个 o1 编码器的复现。arXiv 预印本 arXiv:2412.00154， 2024。URL https:// arxiv.org/abs/2412.00154。

[301] Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. Posterior-grpo: Rewarding reasoning processes in code generation, 2025. URL https://arxiv.org/abs/2508.05170.

[301] 丽水范，张宇，陈谋翔，刘忠信. Posterior-grpo: 在代码生成中奖励推理过程，2025. URL https://arxiv.org/abs/2508.05170.

[302] Jiawei Liu, Thanh Nguyen, Mingyue Shang, Hantian Ding, Xiaopeng Li, Yu Yu, Varun Kumar, and Zijian Wang. Learning code preference via synthetic evolution, 2024. URL https://arxiv.org/ abs/2410.03837.

[302] 刘嘉伟, 阮清, 商明月, 丁汉天, 李晓鹏, 余瑜, Varun Kumar, 王子健. 通过合成进化学习代码偏好, 2024. URL https://arxiv.org/ abs/2410.03837.

[303] Kechi Zhang, Ge Li, Jia Li, Yihong Dong, Jia Li, and Zhi Jin. Focused-DPO: Enhancing code generation through focused preference optimization on error-prone points. In Wanxiang Che, Joyce Nabende, Ekate-

rina Shutova, and Mohammad Taher Pilehvar, editors, Findings of the Association for Computational Linguistics: ACL 2025, pages 9578-9591, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.498. URL https://aclanthology.org/2025.findings-acl.498/.

[303] 张可驰, 李革, 李佳, 董义宏, 李佳, 金志. Focused-DPO: 通过在易错点上集中偏好优化来增强代码生成. 收录于 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, Mohammad Taher Pilehvar 编辑, Findings of the Association for Computational Linguistics: ACL 2025, 页 9578-9591, 奥地利维也纳, 2025 年 7 月. 计算语言学协会. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.498. URL https://aclanthology.org/2025.findings-acl.498/.

[304] Sherry Yang, Joy He-Yueya, and Percy Liang. Reinforcement learning for machine learning engineering agents, 2025. URL https://arxiv.org/abs/2509.01684.

[304] 杨雪莉, Joy He-Yueya, Percy Liang. 用于机器学习工程智能体的强化学习, 2025. URL https://arxiv.org/abs/2509.01684.

[305] Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Taco Cohen, and Gabriel Synnaeve. RLEF: Grounding code LLMs in execution feedback with reinforcement learning. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id= PzSG5nKe1q.

[305] Jonas Gehring, 郑坤浩, Jade Copet, Vegard Mella, Taco Cohen, Gabriel Synnaeve. RLEF: 用执行反馈和强化学习为代码大模型奠定基础. 收录于第 42 届国际机器学习大会, 2025. URL https://openreview.net/forum?id= PzSG5nKe1q.

[306] Arnav Kumar Jain, Gonzalo Gonzalez-Pumariega, Wayne Chen, Alexander M Rush, Wenting Zhao, and Sanjiban Choudhury. Multi-turn code generation through single-step rewards. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum? id=aJeLhLcsh0.

[306] Arnav Kumar Jain, Gonzalo Gonzalez-Pumariega, Wayne Chen, Alexander M Rush, Wenting Zhao, Sanjiban Choudhury. 通过单步奖励实现多轮代码生成. 收录于第 42 届国际机器学习大会, 2025. URL https://openreview.net/forum? id=aJeLhLcsh0.

[307] 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, 2025. URL https://arxiv.org/abs/2505.21668.

[307] 陈永超, 刘悦莹, 周俊伟, 郝依伦, 王敬全, 张扬, 樊楚楚. R1-code-interpreter: 通过监督与强化学习训练大模型用代码推理, 2025. URL https://arxiv.org/abs/2505.21668.

[308] Jie Wu, Haoling Li, Xin Zhang, Jianwen Luo, Yangyu Huang, Ruihang Chu, Yujiu Yang, and Scarlett Li. Iterpref: Focal preference learning for code generation via iterative debugging, 2025. URL https://arxiv.org/abs/2503.02783.

[308] 吴杰, 李皓灵, 张鑫, 罗建文, 黄阳雨, 楚瑞杭, 杨宇骥, Scarlett Li. Iterpref: 通过迭代调试的焦点偏好学习用于代码生成, 2025. URL https://arxiv.org/abs/2503.02783.

[309] Nan Jiang, Xiaopeng Li, Shiqi Wang, Qiang Zhou, Soneya Binta Hossain, Baishakhi Ray, Varun

Kumar, Xiaofei Ma, and Anoop Deoras. Ledex: Training llms to better self-debug and explain code. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 35517-35543. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ 3ea832724870c700f0a03c665572e2a9-Paper-Conference.pdf.

[309] 姜楠, 李晓鹏, 王世棋, 周强, Soneya Binta Hossain, 白沙奇·雷, Varun Kumar, 马晓飞, Anoop Deoras. Ledex: 训练大模型更好地自我调试并解释代码. 收录于 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang 编辑, Advances in Neural Information Processing Systems, 卷 37, 页 35517-35543. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ 3ea832724870c700f0a03c665572e2a9-Paper-Conference.pdf.

[310] Zhihui Xie, Jie chen, Liyu Chen, Weichao Mao, Jingjing Xu, and Lingpeng Kong. Teaching language models to critique via reinforcement learning. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=UVoxPlv5E1.

[310] 谢志辉, 陈杰, 陈礼宇, 毛伟超, 徐婧婧, 孔令鹏. 教语言模型通过强化学习进行批评教学. 收录于第 42 届国际机器学习大会, 2025. URL https://openreview.net/forum?id=UVoxPlv5E1.

[311] Yiyang Jin, Kunzhao Xu, Hang Li, Xueting Han, Yanmin Zhou, Cheng Li, and Jing Bai. Reveal: Self-evolving code agents via iterative generation-verification, 2025. URL https://arxiv.org/ abs/2506.11442.

[311] 金怡阳, 徐坤兆, 李航, 韩雪婷, 周彦敏, 李成, 白静. Reveal: 通过迭代生成-验证实现自我进化的代码代理, 2025. URL https://arxiv.org/ abs/2506.11442.

[312] Michael Luo, Naman Jain, Jaskirat Singh, Sijun Tan, Ameen Patel, Qingyang Wu, Alpay Ariyak, Colin Cai, Shang Zhu Tarun Venkat, Ben Athiwaratkun, Manan Roongta, Ce Zhang, Li Erran Li, Raluca Ada Popa, Koushik Sen, and Ion Stoica. Deepswe: Training a state-of-the-art coding agent from scratch by scaling rl. https://pretty-radio-b75.notion.site/ DeepSWE-Training-a-Fully-Open-sourced-State-of-the-Art-Coding-Agent-by-Scaling- 2025. Notion Blog.

[312] Michael Luo, Naman Jain, Jaskirat Singh, Sijun Tan, Ameen Patel, Qingyang Wu, Alpay Ariyak, Colin Cai, Shang Zhu Tarun Venkat, Ben Athiwaratkun, Manan Roongta, Ce Zhang, Li Erran Li, Raluca Ada Popa, Koushik Sen, and Ion Stoica. Deepswe: Training a state-of-the-art coding agent from scratch by scaling rl. https://pretty-radio-b75.notion.site/ DeepSWE-Training-a-Fully-Open-sourced-State-of-the-Art-Coding-Agent-by-Scaling- 2025. Notion Blog.

[313] 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, 2025. URL https://arxiv.org/abs/2502.18449.

[313] 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, 2025. URL https://arxiv.org/abs/2502.18449.

[314] Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory

W. Wornell, Subhro Das, David Daniel Cox, and Chuang Gan. Satori: Reinforcement learning with chain-of-action-thought enhances LLM reasoning via autoregressive search. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=j4FXxMiDjL.

[314] Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory W. Wornell, Subhro Das, David Daniel Cox, and Chuang Gan. Satori: Reinforcement learning with chain-of-action-thought enhances LLM reasoning via autoregressive search. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=j4FXxMiDjL.

[315] Yanlin Wang, Yanli Wang, Daya Guo, Jiachi Chen, Ruikai Zhang, Yuchi Ma, and Zibin Zheng. RLCoder: Reinforcement Learning for Repository-Level Code Completion . In 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE), pages 1140-1152, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/ICSE55347.2025.00014. URL https: //doi.ieeecomputersociety.org/10.1109/ICSE55347.2025.00

[315] Yanlin Wang, Yanli Wang, Daya Guo, Jiachi Chen, Ruikai Zhang, Yuchi Ma, and Zibin Zheng. RLCoder: Reinforcement Learning for Repository-Level Code Completion . In 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE), pages 1140-1152, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/ICSE55347.2025.00014. URL https: //-doi.ieeecomputersociety.org/10.1109/ICSE55347.2025.00014.

[316] 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, 2025. URL https: //arxiv.org/abs/2505.23723.

[316] 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, 2025. URL https: //arxiv.org/abs/2505.23723.

[317] 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, 2025. URL https://arxiv.org/abs/2508.12551.

[317] 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, 2025. URL https://arxiv.org/abs/2508.12551.

[318] Alexander Golubev, Maria Trofimova, Sergei Polezhaev, Ibragim Badertdinov, Maksim Nekrashevich, Anton Shevtsov, Simon Karasik, Sergey Abramov, Andrei Andriushchenko, Filipp Fisin, Sergei Skvortsov, and Boris Yangel. Training long-context, multi-turn software engineering agents with reinforcement learning, 2025. URL https://arxiv.org/abs/2508.03501.

[318] Alexander Golubev, Maria Trofimova, Sergei Polezhaev, Ibragim Badertdinov, Maksim Nekrashevich, Anton Shevtsov, Simon Karasik, Sergey Abramov, Andrei Andriushchenko, Filipp Fisin, Sergei Skvortsov, and Boris Yangel. Training long-context, multi-turn software engineering agents with reinforcement learning, 2025. URL https://arxiv.org/abs/2508.03501.

[319] OpenAI. Introducing codex. https://openai.com/index/introducing-codex/, May 2025.

[319] OpenAI. Introducing codex. https://openai.com/index/introducing-codex/, May 2025.

[320] Anthropic. Claude code: Deep coding at terminal velocity. https://www.anthropic.com/ claude-code, February 2025. Anthropic's agentic command-line coding tool, introduced alongside Claude 3.7 Sonnet. Enables developers to delegate engineering tasks directly from their terminal via natural-language commands.

[320] Anthropic. Claude code: Deep coding at terminal velocity. https://www.anthropic.com/ claude-code, February 2025. Anthropic's agentic command-line coding tool, introduced alongside Claude 3.7 Sonnet. Enables developers to delegate engineering tasks directly from their terminal via natural-language commands.

[321] Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition, 2025. URL https://arxiv.org/abs/2504.21801.

[321] Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition, 2025. URL https://arxiv.org/abs/2504.21801.

[322] Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. Formal mathematical reasoning: A new frontier in ai, 2024. URL https://arxiv.org/ abs/2412.16075.

[322] Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. 正式数学推理: 人工智能的新前沿, 2024. URL https://arxiv.org/ abs/2412.16075.

[323] Andrea Asperti, Alberto Naibo, and Claudio Sacerdoti Coen. Thinking machines: Mathematical reasoning in the age of llms, 2025. URL https://arxiv.org/abs/2508.00459.

[323] Andrea Asperti, Alberto Naibo, and Claudio Sacerdoti Coen. 思考机器: 大语言模型时代的数学推理, 2025. URL https://arxiv.org/abs/2508.00459.

[324] Xinji Mai, Haotian Xu, Xing W, Weinong Wang, Jian Hu, Yingying Zhang, and Wenqiang Zhang. Agent rl scaling law: Agent rl with spontaneous code execution for mathematical problem solving, 2025. URL https://arxiv.org/abs/2505.07773.

[324] Xinji Mai, Haotian Xu, Xing W, Weinong Wang, Jian Hu, Yingying Zhang, and Wenqiang Zhang. 代理强化学习规模律: 用于数学问题求解的具自发代码执行的代理强化学习, 2025. URL https://arxiv.org/abs/2505.07773.

[325] Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning, 2025. URL https://arxiv.org/abs/2505.22660.

[325] Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 仅最大化置信度即可提升推理能力, 2025. URL https://arxiv.org/abs/2505.22660.

[326] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

[326] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: 通过熵最小化实现完全测试时自适应。载于国际学习表征会议，2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

[327] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example, 2025. URL https://arxiv.org/abs/2504.205

[327] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 在只有一个训练样本下用于大语言模型推理的强化学习，2025. URL https://arxiv.org/abs/2504.20571.

[328] Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of llm outputs, 2024. URL https://arxiv.org/abs/2407.13692.

[328] Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. 证明者—验证者博弈提升大语言模型输出的可读性, 2024. URL https://arxiv.org/abs/2407.13692.

[329] Cem Anil, Guodong Zhang, Yuhuai Wu, and Roger Grosse. Learning to give checkable answers with prover-verifier games, 2021. URL https://arxiv.org/abs/2108.12099.

[329] Cem Anil, Guodong Zhang, Yuhuai Wu, and Roger Grosse. 通过证明者—验证者博弈学习给出可检验答案, 2021. URL https://arxiv.org/abs/2108.12099.

[330] Chengpeng Li, Mingfeng Xue, Zhenru Zhang, Jiaxi Yang, Beichen Zhang, Xiang Wang, Bowen Yu, Binyuan Hui, Junyang Lin, and Dayiheng Liu. Start: Self-taught reasoner with tools. CoRR, abs/2503.04625, March 2025. URL https://doi.org/10.48550/arXiv.2503.04625.

[330] Chengpeng Li, Mingfeng Xue, Zhenru Zhang, Jiaxi Yang, Beichen Zhang, Xiang Wang, Bowen Yu, Binyuan Hui, Junyang Lin, and Dayiheng Liu. START: 具工具的自学推理器。CoRR, abs/2503.04625, 2025 年 3 月. URL https://doi.org/10.48550/arXiv.2503.04625.

[331] Toby Simonds and Akira Yoshiyama. Ladder: Self-improving llms through recursive problem decomposition, 2025. URL https://arxiv.org/abs/2503.00735.

[331] Toby Simonds and Akira Yoshiyama. LADDER: 通过递归问题分解自我提升的大语言模型, 2025. URL https://arxiv.org/abs/2503.00735.

[332] Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D. Manning. Synthetic data generation & multi-step rl for reasoning & tool use, 2025. URL https://arxiv.org/abs/2504.04736.

[332] Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D. Manning. 合成数据生成与多步骤强化学习用于推理与工具使用, 2025. URL https://arxiv.org/abs/2504.04736.

[333] Qianyue Hao, Sibo Li, Jian Yuan, and Yong Li. Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14140.

[333] Qianyue Hao, Sibo Li, Jian Yuan, and Yong Li. 思维强化学习: 用推理时强化学习导航大语言模型的推理, 2025. URL https://arxiv.org/abs/2505.14140.

[334] Huajian Xin, Z.Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou, Z.F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=I4YAIwrsXa.

[334] Huajian Xin, Z.Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou, Z.F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: 利用证明助手反馈进行强化学习和蒙特卡洛树搜索。载于第十三届国际学习表征会议, 2025. URL https://openreview.net/forum?id=I4YAIwrsXa.

[335] Minchao Wu, Michael Norrish, Christian Walder, and Amir Dezfouli. Tacticzero: Learning to prove theorems from scratch with deep reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=

[335] 吴敏超, Michael Norrish, Christian Walder, 和 Amir Dezfouli. Tacticzero: 用深度强化学习从零开始学习证明定理. 收录于 A. Beygelzimer, Y. Dauphin, P. Liang, 和 J. Wortman Vaughan 编辑, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=edmYVRkYZv.

[336] Murari Ambati. Proofnet++: A neuro-symbolic system for formal proof verification with self-correction, 2025. URL https://arxiv.org/abs/2505.24230.

[336] Murari Ambati. Proofnet++: 一个用于形式证明验证并具自我纠正的神经-符号系统, 2025. URL https://arxiv.org/abs/2505.24230.

[337] 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, 2025. URL https://arxiv.org/abs/2507.08649.

[337] 季兴光, 刘雅慧, 王琪, 张景元, 岳扬, 史瑞, 孙晨曦, 张富政, 周国瑞, 和盖坤. Leanabell-prover-v2: 通过强化学习实现的带验证器集成推理的形式定理证明, 2025. URL https://arxiv.org/abs/2507.08649.

[338] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics, 2022. URL https://arxiv.org/abs/2109.00110.

[338] 郑昆豪, Jesse Michael Han, 和 Stanislas Polu. Minif2f: 一个跨系统的奥林匹克级别数学习题基准, 2022. URL https://arxiv.org/abs/2109.00110.

[339] Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics, 2023. URL https://arxiv.org/abs/2302.12433.

[339] Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, 和 Jeremy Avigad. Proofnet: 自动形式化并形式证明本科水平数学, 2023. URL https://arxiv.org/abs/2302.12433.

[340] Jingyuan Zhang, Qi Wang, Xingguang Ji, Yahui Liu, Yang Yue, Fuzheng Zhang, Di Zhang, Guorui Zhou, and Kun Gai. Leanabell-prover: Posttraining scaling in formal reasoning, 2025. URL https: //arxiv.org/abs/2504.06122.

[340] 张景元, 王琪, 季兴光, 刘雅慧, 岳扬, 张富政, 张迪, 周国瑞, 和盖坤. Leanabell-prover: 形式推理中的后训练扩展, 2025. URL https: //arxiv.org/abs/2504.06122.

[341] The mathlib Community. mathlib4: The lean 4 mathematical library, 2020-2025. URL https: //github.com/leanprover-community/mathlib4. Accessed: 2025-09-01.

[341] The mathlib Community. mathlib4: The lean 4 数学库, 2020-2025. URL https: //github.com/leanprover-community/mathlib4. Accessed: 2025-09-01.

[342] Huaiyuan Ying, Zijian Wu, Yihan Geng, Zheng Yuan, Dahua Lin, and Kai Chen. Lean workbook: A large-scale lean problem set formalized from natural language math problems, 2025. URL https: //arxiv.org/abs/2406.03847.

[342] 应怀远, 吴子剑, 耿怡涵, 袁征, 林大华, 和陈凯. Lean workbook: 一个从自然语言数学题形式化的大规模 lean 题集, 2025. URL https: //arxiv.org/abs/2406.03847.

[343] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. Technical report, Peking University and collaborators, 2024. URL http://faculty.bicmr.pku.edu.cn/~dongbin/Publications/numina_dataset.pdf.Technical Report.

[343] 李佳, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, 于龙辉, Albert Q. Jiang, 沈子瑜, 秦子涵, 董斌, 周立, Yann Fleureau, Guillaume Lample, 和 Stanislas Polu. Numinamath. 技术报告, 北京大学及合作者, 2024. URL http://faculty.bicmr.pku.edu.cn/~dongbin/Publications/numina_dataset.pdf.Technical Report.

[344] Kefan Dong and Tengyu Ma. STP: Self-play LLM theorem provers with iterative conjecturing and proving. In Forty-second International Conference on Machine Learning, 2025. URL https:// openreview.net/forum?id=zWArMedNuW.

[344] 董科凡和马腾宇. STP: 具有迭代猜想与证明的自对弈大型语言模型定理证明器. 收录于第四十二届国际机器学习大会, 2025. URL https:// openreview.net/forum?id=zWArMedNuW.

[345] Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, Jianqiao Lu, Hugues de Saxcé, Bolton Bailey, Chendong Song, Chenjun Xiao, Dehao Zhang, Ebony Zhang, Frederick Pu, Han Zhu, Jiawei Liu, Jonas Bayer, Julien Michel, Longhui Yu, Léo Dreyfus-Schmidt, Lewis Tunstall, Luigi Pagani, Moreira Machado, Pauline Bourigault, Ran Wang, Stanislas Polu, Thibaut Barroyer, Wen-Ding Li, Yazhe Niu, Yann Fleureau, Yangyang Hu, Zhouliang Yu, Zihan Wang, Zhilin Yang, Zhengying Liu, and Jia Li. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning, 2025. URL https://arxiv.org/abs/2504.11354.

[345] 王海明, Mert Unsal, 林晓涵, Mantas Baksys, 刘君齐, Marco Dos Santos, Flood Sung, Marina Vinyes, 颖振哲, 朱泽凯, 陆建侨, Hugues de Saxcé, Bolton Bailey, 宋晨东, 肖晨君, 张德豪, 张艾博, Frederick Pu, 朱寒, 刘嘉炜, Jonas Bayer, Julien Michel, 于龙辉, Léo Dreyfus-Schmidt, Lewis Tunstall, Luigi Pagani, Moreira Machado, Pauline Bourigault, 王冉, Stanislas Polu, Thibaut Barroyer, 李文鼎, 牛雅喆, Yann Fleureau, 胡扬扬, 于周良, 王子涵, 杨植霖, 刘正英, 和李佳. Kimina-prover preview: 面向强化学习的大型形式推理模型, 2025. URL https://arxiv.org/abs/2504.11354.

[346] Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, Cheng Ren, Jiawei Shen, Wenlei Shi, Tong Sun, He Sun, Jiahui Wang, Siran Wang, Zhihong Wang, Chenrui Wei, Shufa Wei, Yonghui Wu, Yuchen Wu, Yihang Xia, Huajian Xin, Fan Yang, Huaiyuan Ying, Hongyi Yuan, Zheng Yuan, Tianyang Zhan, Chi Zhang, Yue Zhang, Ge Zhang, Tianyun Zhao, Jianqiu Zhao, Yichi Zhou, and Thomas Hanwen Zhu. Seed-prover: Deep and broad reasoning for automated theorem proving, 2025. URL https://arxiv.org/abs/2507.23726.

[346] 陈洛鑫, 古金明, 黄连凯, 黄文浩, 江志成, Allan Jie, 金晓然, 金兴, 李成刚, 马凯京, 任成, 沈家伟, 石文磊, 孙通, 孙和, 王佳辉, 王思然, 王志宏, 魏晨睿, 魏书发, 吴永辉, 吴钰辰, 夏怡航, 辛华建, 杨帆, 应淮元, 袁宏逸, 袁正, 詹天阳, 张驰, 张悦, 张戈, 赵天云, 赵建秋, 周一驰, 朱汉文·Thomas. Seed-prover: 用于自动定理证明的深度与广度推理， 2025。URL https://arxiv.org/abs/2507.23726.

[347] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, Hao Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu,

[347] DeepSeek-AI, 刘爱新, 冯贝, 王斌, 王炳轩, 刘博, 赵成刚, 邓成奇, 阮崇, 戴大迈, 郭大雅, 杨德建, 陈德立, 吉东杰, 李尔航, 林方云, 骆富利, 郝广波, 陈冠廷, 李国伟, 张浩, 徐汉威, 杨浩, 张浩巍, 丁洪辉, 辛华建, 高华佐, 李辉, 曲辉, J. L. Cai, 梁健, 郭建忠, 倪佳琪, 李佳世, 陈金, 袁景阳, 邱俊杰, 宋俊霄, 董凯, 高凯歌, 管康, 王磊, 张乐聪, 徐磊, 夏乐怡, 赵亮, 张丽月, 李梦, 王妙军, 张明川, 张明华, 唐明辉, 李明明, 田宁, 黄盼盼, 王佩怡, 张鹏, 朱启浩, 陈沁宇, 杜秋实, R. J. Chen, R. L. Jin, 葛锐奇, 潘睿哲, 徐润鑫, 陈如意, S. S. Li, 卢尚浩, 周尚彦, 陈善煌, 吴少清,

Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, Tao Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, and Xiaowen Sun. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. CoRR, abs/2405.04434, 2024. URL https://doi.org/10.48550/arXiv.2405.04434.

[347] 叶盛峰, 马世榕, 王世宇, 周双, 俞水平, 周顺峰, 郑 Size, 王涛, 裴天, 袁天, 孙天宇, W. L. Xiao, 曾望定, 安伟, 刘文, 梁文锋, 高文俊, 张文涛, X. Q. Li, 金祥跃, 王献祖, 毕晓, 刘晓东, 王晓涵, 沈晓劲, 陈晓康, 陈晓莎, 聂晓涛, 孙晓文. Deepseek-v2: 一种强大、经济且高效的专家混合语言模型。CoRR, abs/2405.04434, 2024。URL https://doi.org/10.48550/arXiv.2405.04434.

[348] Liang Zeng and Liangjun Zhong. Skywork-math: Data scaling laws for mathematical reasoning in LLMs - the story goes on. In The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24, 2024. URL https://openreview.net/forum?id=uHtzqZKbeK.

[348] 曾亮, 钟亮君. Skywork-math: 大型模型数学推理的数据扩展定律——故事继续。收录于 The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24, 2024。URL https://openreview.net/forum?id=uHtzqZKbeK.

[349] Ruida Wang, Jipeng Zhang, Yizhen Jia, Rui Pan, Shizhe Diao, Renjie Pi, and Tong Zhang. TheoremLlama: Transforming general-purpose LLMs into lean4 experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11953-11974, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.667. URL https: //aclanthology.org/2024.emnlp-main.667/.

[349] 王瑞达, 张吉鹏, 贾奕臻, 潘锐, 刁世哲, 皮仁杰, 张彤. The-oremLlama: 将通用大型模型改造为 Lean4 专家。收录于 Yaser Al-Onaizan, Mohit Bansal, 与 Yun-Nung Chen 主编, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 页 11953-11974, 2024 年 11 月, 美国佛罗里达州迈阿密。计算语言学协会。doi: 10.18653/v1/2024.emnlp-main.667。URL https://aclanthology.org/2024.emnlp-main.667/.

[350] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. 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/ d8e1344e27a5b08cdfd5d027d9b8d6de-Paper.pdf.

[350] Thomas Anthony、郑天、David Barber。《用深度学习与树搜索的快思慢想》。收录于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 与 R. Garnett 主编, Advances in Neural Information Processing Systems， 第 30 卷。Curran Associates, Inc., 2017。URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d8e1344e27a5b08cdfd5d027d9b8d6de-Paper.pdf。

[351] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving, 2020. URL https://arxiv.org/abs/2009.03393.

[351] Stanislas Polu 与 Ilya Sutskever。用于自动定理证明的生成式语言建模，2020。URL https://arxiv.org/abs/2009.03393。

[352] Zijian Wu, Suozhi Huang, Zhejian Zhou, Huaiyuan Ying, Zheng Yuan, Wenwei Zhang, Dahua Lin, and Kai Chen. InternLM2.5-stepprover: Advancing automated theorem proving via critic-guided search. In 2nd AI for Math Workshop @ ICML 2025, 2025. URL https://openreview.net/forum?id= qwCqeIg5iI.

[352] 吴子坚, 黄所志, 周哲鉴, 应怀远, 袁征, 张文伟, 林大华, 陈凯. InternLM2.5-stepprover: 通过评论者引导搜索推进自动定理证明。发表于 2025 年 ICML 附属之第二届 AI for Math 研讨会, 2025。URL https://openreview.net/forum?id= qwCqeIg5iI.

[353] Haohan Lin, Zhiqing Sun, Sean Welleck, and Yiming Yang. Lean-STar: Learning to interleave thinking and proving. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=SOWZ59UyNC.

[353] 林浩汉, 孙志青, Sean Welleck, 杨亦明. Lean-STar: 学习交替思考与证明。发表于第十三届国际表示学习大会, 2025。URL https://openreview.net/forum?id=SOWZ59UyNC.

[354] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: self-taught reasoner bootstrapping reasoning with reasoning. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. URL https://dl.acm.org/doi/10.5555/3600270.3601396.

[354] Eric Zelikman, 吴雨淮, Jesse Mu, Noah D. Goodman. Star: 自学推理器以推理促进推理。发表于第 36 届神经信息处理系统国际会议论文集, NIPS '22, Red Hook, NY, USA, 2022。Curran Associates Inc. ISBN 9781713871088。URL https://dl.acm.org/doi/10.5555/3600270.3601396.

[355] Gabriel Poesia, David Broman, Nick Haber, and Noah Goodman. Learning formal mathematics from intrinsic motivation. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=uNKlTQ8mBD.

[355] Gabriel Poesia, David Broman, Nick Haber, Noah Goodman. 从内在动机学习形式化数学。发表于第 38 届神经信息处理系统年会, 2024。URL https://openreview.net/forum?id=uNKlTQ8mBD.

[356] An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian J. McAuley, Jianfeng Gao, Zicheng Liu, and Lijuan Wang. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. CoRR, abs/2311.07562, 2023. URL https://doi.org/10.48550/arXiv.2311

[356] 严安, 杨正远, 朱万荣, 林凯文, 李林杰, 王剑锋, 杨剑伟, 钟宜舞, Julian J. McAuley, 高建峰, 刘子成, 王丽娟. Gpt-4v in wonderland: 用于零样本智能手机 GUI 导航的大型多模态模型。CoRR, abs/2311.07562, 2023。URL https://doi.org/10.48550/arXiv.2311.07562.

[357] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a generalist web agent, if grounded. 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 61349-61385. PMLR, 21-27 Jul 2024. URL https://proceedings.mlr.press/v235/zheng24e.html.

[357] 郑博元, 苟博宇, Kil Jihyung, 孙焕, 苏宇. GPT-4V(ision) 若具备落地则可成为通用网页代理。收录于 Ruslan Salakhutdinov 等编辑的第 41 届国际机器学习大会论文集, Proceedings of Machine Learning Research 第 235 卷, 页 61349-61385。PMLR, 2024 年 7 月 21-27 日。URL https://proceedings.mlr.press/v235/zheng24e.html.

[358] Mukund Khanna Kunal Singh, Shreyas Singh. Trishul: A training-free agentic framework for zero-shot gui action grounding, 2025. URL https://arxiv.org/abs/2502.08226.

[358] Mukund Khanna Kunal Singh, Shreyas Singh. Trishul: 一种无需训练的自主框架用于零样本 GUI 操作落地, 2025。URL https://arxiv.org/abs/2502.08226.

[359] Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Large language model-brained GUI agents: A survey. Transactions on Machine Learning Research, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=

[359] 张朝云, 何世霖, 钱嘉煦, 李博文, 李立群, 秦思, 康宇, 马明华, 刘古月, 林庆伟, Saravan Rajmohan, 张东梅, 张琪. 基于大型语言模型的 GUI 代理: 综述。Transactions on Machine Learning Research, 2025。ISSN 2835-8856。URL https://openreview.net/forum?id=xChvYjvXTp.

[360] Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Seunghyun Yoon, Lina Yao, Branislav Kveton, Jihyung Kil, Thien Huu Nguyen, Trung Bui, Tianyi Zhou, Ryan A. Rossi, and Franck Dernoncourt. GUI agents: A survey. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Findings of the Association for Computational Linguistics: ACL 2025, pages 22522-22538, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1158. URL https: //aclanthology.org/2025.findings-acl.1158/.

[360] Nguyen Dang, 陈健, 王誉, 吴刚, Park Namyong, 胡正勉, 吕寒佳, 吴俊达, Ryan Aponte, 夏禹, 李新通, 史静, 陈洪杰, Viet Dac Lai, 谢周航, Kim Sungchul, 张瑞意, 余彤, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Yoon Seunghyun, 姚琳娜, Branislav Kveton, Kil Jihyung, Nguyen Thien Huu, Bui Trung, 周天一, Ryan A. Rossi, Franck Dernoncourt. GUI agents: 综述。收录于 Che Wanxiang 等编辑的 Findings of the Association for Computational Linguistics: ACL 2025, 页 22522-22538, 维也纳, 奥地利, 2025 年 7 月。计算语言学协会。ISBN 979-8-89176-256-5。doi: 10.18653/v1/2025.findings-acl.1158。URL https: //aclanthology.org/2025.findings-acl.1158/.

[361] Yuhang Liu, Pengxiang Li, Zishu Wei, Congkai Xie, Xueyu Hu, Xinchen Xu, Shengyu Zhang, Xiaotian Han, Hongxia Yang, and Fei Wu. InfiGUIAgent: A multimodal generalist GUI agent with native reasoning and reflection. In ICML 2025 Workshop on Computer Use Agents, 2025. URL https: //openreview.net/forum?id=p0h9XJ7fMH.

[361] 刘宇航, 李鹏翔, 魏子书, 谢聪凯, 胡雪煜, 许新晨, 张胜宇, 韩晓天, 杨红霞, 以及吴飞. InfiGUIAgent: 一种具备原生推理与反思能力的多模态通用 GUI 代理. 载于 ICML 2025 计算机使用代理研讨会, 2025. URL https: //openreview.net/forum?id=p0h9XJ7fMH.

[362] Shuq uan Lian, Yuhang Wu, Jia Ma, Zihan Song, Bingqi Chen, Xiawu Zheng, and Hui Li. Ui-agile: Advancing gui agents with effective reinforcement learning and precise inference-time grounding. arXiv preprint arXiv:2507.22025, 2025. URL https://arxiv.org/abs/2507.22025.

[362] 连书全，吴宇航，马佳，宋子涵，陈炳祺，郑霞雾，以及李辉． Ui-agile: 通过有效的强化学习和精确的推理时定位推进 GUI 代理． arXiv 预印本 arXiv:2507.22025，2025． URL https://arxiv.org/abs/2507.22025.

[363] Bofei Zhang, Zirui Shang, Zhi Gao, Wang Zhang, Rui Xie, Xiaojian Ma, Tao Yuan, Xinxiao Wu, Song-Chun Zhu, and Qing Li. Tongui: Building generalized gui agents by learning from multimodal web tutorials. arXiv preprint arXiv:2504.12679, 2025. URL https://arxiv.org/abs/2504.12679.

[363] 张博非，尚子睿，高志，张望，谢瑞，马晓健，袁涛，吴新啸，朱松椿，以及李青． Tongui: 通过学习多模态网页教程构建通用 GUI 代理． arXiv 预印本 arXiv:2504.12679，2025． URL https://arxiv.org/abs/2504.12679.

[364] Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. Gui-r1 : A generalist r1-style vision-language action model for gui agents, 2025. URL https://arxiv.org/abs/2504.10458.

[364] 罗润, 王璐, 何婉伟, 以及夏晓博. Gui-r1: 一个面向 GUI 代理的通用 r1 风格视觉-语言动作模型, 2025. URL https://arxiv.org/abs/2504.10458.

[365] 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, 2025. URL https://arxiv.org/abs/2503.21620.

[365] 陆正熙, 柴玉翔, 郭雅轩, 尹西, 刘亮, 王浩, 肖寒, 任帅, 熊冠景, 以及李宏胜. Ui-r1: 通过强化学习提升 GUI 代理的高效动作预测, 2025. URL https://arxiv.org/abs/2503.21620.

[366] 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, 2025. URL https://arxiv.org/abs/2504.14239.

[366] 刘宇航, 李鹏翔, 谢聪凯, 胡泽维, 韩晓天, 张胜宇, 杨红霞, 以及吴飞. Infigui-r1: 将多模态 GUI 代理从反应型执行者推进为深思熟虑的推理者, 2025. URL https://arxiv.org/abs/2504.14239.

[367] Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian

Chen, Yankai Lin, Jie Xie, Wei Zhou, Wang Xu, Yuanheng Zhang, Zhou Su, Zhongwu Zhai, Xiaoming Liu, Yudong Mei, Jianming Xu, Hongyan Tian, Chongyi Wang, Chi Chen, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Agentcpm-gui: Building mobile-use agents with reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2506.01391.

[367] 张中, 路亚熙, 傅伊坤, 霍玉鹏, 杨慎之, 吴业赛, 司涵, 丛鑫, 陈昊天, 林彦楷, 谢杰, 周伟, 徐汪, 张远衡, 苏周, 翟忠武, 刘晓明, 梅宇东, 徐建明, 田红艳, 王崇伊, 陈池, 姚原, 刘志远, 以及孙茂松. Agentcpm-gui: 通过强化微调构建移动端使用代理, 2025. URL https://arxiv.org/abs/2506.01391.

[368] Zhangxuan Gu, Zhengwen Zeng, Zhenyu Xu, Xingran Zhou, Shuheng Shen, Yunfei Liu, Beitong Zhou, Changhua Meng, Tianyu Xia, Weizhi Chen, Yue Wen, Jingya Dou, Fei Tang, Jinzhen Lin, Yulin Liu, Zhenlin Guo, Yichen Gong, Heng Jia, Changlong Gao, Yuan Guo, Yong Deng, Zhenyu Guo, Liang Chen, and Weiqiang Wang. Ui-venus technical report: Building high-performance ui agents with rft, 2025. URL https://arxiv.org/abs/2508.10833.

[368] 顾章轩, 曾正文, 徐振宇, 周星然, 沈书恒, 刘云飞, 周备彤, 孟昌华, 夏天宇, 陈伟志, 温岳, 窦靖雅, 汤飞, 林金针, 刘育林, 郭振霖, 龚伊辰, 贾恒, 高长龙, 郭元, 邓勇, 郭振宇, 陈良, 以及王维强. Ui-venus 技术报告: 通过 RFT 构建高性能 UI 代理, 2025. URL https://arxiv.org/abs/2508.10833.

[369] 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, Xue Liu, Alexandre Drouin, Laurent Charlin, Alexandre Piché, Alexandre Lacoste, and Massimo Caccia. How to train your llm web agent: A statistical diagnosis, 2025. URL https://arxiv.org/abs/2507.04103.

[369] 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, Xue Liu, Alexandre Drouin, Laurent Charlin, Alexandre Piché, Alexandre Lacoste, 以及 Massimo Caccia. How to train your llm web agent: 一个统计诊断, 2025. URL https://arxiv.org/abs/2507.04103.

[370] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL https://arxiv.org/abs/2501.12326.

[370] 秦语家, 叶亦宁, 方俊杰, 王浩明, 梁世豪, 田志作, 张军达, 李家豪, 李云鑫, 黄世爵, 钟宛军, 李宽烨, 杨嘉乐, 缪宇, 林我宇, 刘龙翔, 姜旭, 马千里, 李靖宇, 肖晓军, 蔡凯, 李闯, 郑耀威, 金朝林, 李晨, 周晓, 王敏超, 陈浩立, 李兆建, 杨海华, 刘海峰, 林峰, 彭涛, 刘鑫, 史光. Ui-tars: 以原生代理开创自动化 GUI 交互, 2025. URL https://arxiv.org/abs/2501.12326.

[371] Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 12461-12495. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_f
1704ddd0bb89f159dfe609b32c889995-Paper-Conference.pdf.

[371] 白浩, 周一飞, Mert Cemri, 潘佳怡, Alane Suhr, 谢尔盖·列文, 和 Aviral Kumar. Digirl: 在真实环境中通过自主强化学习训练设备控制代理. 收录于 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, 和 C. Zhang 编, Advances in Neural Information Processing Systems, 第 37 卷, 页 12461-12495. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1704ddd0bb89f159dfe609b32c889995-Paper-Conference.pdf.

[372] Chenyu Yang, Shiqian Su, Shi Liu, Xuan Dong, Yue Yu, Weijie Su, Xuehui Wang, Zhaoyang Liu, Jinguo Zhu, Hao Li, Wenhai Wang, Yu Qiao, Xizhou Zhu, and Jifeng Dai. Zerogui: Automating online gui learning at zero human cost, 2025. URL https://arxiv.org/abs/2505.23762.

[372] 杨晨瑜, 苏诗茜, 刘石, 董轩, 余越, 苏伟杰, 王雪慧, 刘朝阳, 朱金国, 李昊, 王文海, 乔煜, 竺希洲, 和戴继锋. Zerogui: 在零人工成本下实现在线 GUI 学习自动化, 2025. URL https://arxiv.org/abs/2505.23762.

[373] 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, 2025. URL https://arxiv.org/abs/2507.05720.

[373] 石玉成, 余文浩, 李在唐, 王永林, 张弘明, 刘宁浩, 米海涛, 和余东. Mobilegui-rl: 在在线环境中通过强化学习推进移动 GUI 代理, 2025. URL https://arxiv.org/abs/2507.05720.

[374] 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.

[374] 赖涵宇, 刘晓, 赵彦霄, 徐翰, 张涵晨, 井博浩, 任妍雨, 姚顺天, 董宇晓, 和唐捷. Computerrl: 为计算机使用代理扩展端到端在线强化学习, 2025. URL https://arxiv.org/abs/2508.14040.

[375] Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, Wanjun Zhong, Yining Ye, Yujia Qin, Yuwen Xiong, Yuxin Song, Zhiyong Wu, Aoyan Li, Bo Li, Chen Dun, Chong Liu, Daoguang Zan, Fuxing Leng, Hanbin Wang, Hao Yu, Haobin Chen, Hongyi Guo, Jing Su, Jingjia Huang, Kai Shen, Kaiyu Shi, Lin Yan, Peiyao Zhao, Pengfei Liu, Qinghao Ye, Renjie Zheng, Shulin Xin, Wayne Xin Zhao, Wen Heng, Wenhao Huang, Wenqian Wang, Xiaobo Qin, Yi Lin, Youbin Wu, Zehui Chen, Zihao Wang, Baoquan Zhong, Xinchun Zhang, Xujing Li, Yuanfan Li, Zhongkai Zhao, Chengquan Jiang, Faming Wu, Haotian Zhou, Jinlin Pang, Li Han, Qi Liu, Qianli Ma, Siyao Liu, Songhua Cai, Wenqi Fu, Xin Liu, Yaohui Wang, Zhi

[375] 王浩明, 邹浩洋, 宋华通, 冯家展, 方俊杰, 陆俊廷, 刘龙翔, 罗沁钰, 梁世豪, 黄世爵, 钟宛军, 叶亦宁, 秦语家, 熊宇文, 宋玉鑫, 吴志勇, 李澳颜, 李博, 敦晨, 刘崇, 甄道广, 冷福兴, 王汉斌, 余浩, 陈海滨, 郭洪毅, 苏靖, 黄静佳, 沈凯, 史林, 蔡松华, 傅文琪, 刘新, 王耀辉, 吴有斌, 陈泽辉, 王子豪, 钟宝泉, 张新春, 李旭晶, 李远帆, 赵中凯, 江成全, 吴法明, 周昊天, 庞金林, 韩立, 刘琪, 马千里, 刘思遥, 蔡松华, 傅文琪, 刘鑫, 王耀辉, 朱志

Zhang, Bo Zhou, Guoliang Li, Jiajun Shi, Jiale Yang, Jie Tang, Li Li, Qihua Han, Taoran Lu, Woyu Lin, Xiaokang Tong, Xinyao Li, Yichi Zhang, Yu Miao, Zhengxuan Jiang, Zili Li, Ziyuan Zhao, Chenxin Li, Dehua Ma, Feng Lin, Ge Zhang, Haihua Yang, Hangyu Guo, Hongda Zhu, Jiaheng Liu, Junda Du, Kai Cai, Kuanye Li, Lichen Yuan, Meilan Han, Minchao Wang, Shuyue Guo, Tianhao Cheng, Xiaobo Ma, Xiaojun Xiao, Xiaolong Huang, Xinjie Chen, Yidi Du, Yilin Chen, Yiwen Wang, Zhaojian Li, Zhenzhu Yang, Zhiyuan Zeng, Chaolin Jin, Chen Li, Hao Chen, Haoli Chen, Jian Chen, Qinghao Zhao, and Guang Shi. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning, 2025. URL https://arxiv.org/abs/2509.02544.

张博周国亮李佳骏史佳乐杨嘉乐汤杰唐李奇华韩明涛路涛然林沃宇佟晓康李欣尧张艺池苗宇江正轩李子力赵子远李晨新马德华林峰张格海华杨航宇郭航誉朱弘达刘嘉恒杜军达蔡凯李宽冶袁理辰韩美兰王敏超郭淑月程天浩马晓博肖晓军黄晓龙陈新杰杜一笛陈怡琳王艺文李照荐杨珍珠曾志远金超霖李晨侯晨郑浩陈建陈青浩赵光和石光. Ui-tars-2 技术报告: 通过多轮强化学习推进 GUI 代理, 2025。URL https://arxiv.org/abs/2509.02544.

[376] Chen Gao, Liankai Jin, Xingyu Peng, Jiazhao Zhang, Yue Deng, Annan Li, He Wang, and Si Liu. Octonav: Towards generalist embodied navigation, 2025. URL https://arxiv.org/abs/2506.09839.

[376] 高晨金连凯彭星宇张家钊邓悦李安安王赫和刘思. Octonav: 迈向通用化具身导航，2025。URL https://arxiv.org/abs/2506.09839.

[377] Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9098-9108, June 2025.

[377] 李庚许景林赵云贞彭予昕. Dyfo: 一种免训练的动态聚焦视觉搜索，提升 LMM 在细粒度视觉理解中的表现。载于 IEEE/CVF 计算机视觉与模式识别会议 (CVPR) 论文集，第 9098-9108 页，2025 年 6 月。

[378] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520, 2025.

[378] 刘雨齐彭博浩钟智升岳子豪卢凡斌余蓓贾佳雅. Seg-zero: 通过认知强化的推理链引导分割。arXiv 预印本 arXiv:2503.06520，2025。

[379] Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. Compile scene graphs with reinforcement learning. arXiv preprint arXiv:2504.13617, 2025.

[379] 陈祖尧吴金霖雷震波利福斯马克与陈昌文. 使用强化学习编译场景图。arXiv 预印本 arXiv:2504.13617，2025。

[380] Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han, Hao Sun, Jiayi Ji, Xiaoshuai Sun, Qingwei Lin, Weiwei Deng, Dongmei Zhang, Feng Sun, Qi Zhang, and Rongrong Ji. Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.17540.

[380] 吴铭锐王璐赵谱流方恺张建劲刘建峰詹岳峰韩伟浩孙浩纪佳怡孙小帅林清威邓为威张冬梅孙峰张琦纪荣榆. Reprompt: 通过强化学习的推理增强重提示用于文生图生成，2025。URL https://arxiv.org/abs/2505.17540.

[381] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility, 2024. URL https://arxiv.org/abs/2404.04465.

[381] 李舒凡 Konstantinos Kallidromitis Akash Gokul 加藤祐介 Kozuka 和. 通过优化人类效用对齐扩散模型，2024。URL https://arxiv.org/abs/2404.04465.

[382] 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, 2025. URL https://arxiv.org/abs/2505.05470.

[382] 刘杰刘公烨梁家俊李阳光刘佳恒汪新涛万鹏飞张迪及欧阳万里. Flow-grpo: 通过在线强化学习训练流匹配模型，2025。URL https: //arxiv.org/abs/2505.05470.

[383] 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, 2025. URL https://arxiv.org/abs/2505.17022.

[383] 段承琦方荣尧王雨晴王坤黄林江曾星宇李宏胜刘熙辉. Got-r1: 用强化学习释放多模态大模型在视觉生成上的推理能力，2025。URL https://arxiv.org/abs/2505.17022.

[384] 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, 2025.

[384] 王骏柯田志巫讯张歆宇黄伟林吴祖轩蒋宇光. Simplear: 通过预训练、SFT 和强化学习推进自回归视觉生成前沿。arXiv 预印本 arXiv:2504.11455，2025。

[385] Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Ar-grpo: Training autoregressive image generation models via reinforcement learning. arXiv preprint arXiv:2508.06924, 2025.

[385] 袁士豪刘雅慧岳洋张静源左望孟齐望傅正张和周国睿. Ar-grpo: 通过强化学习训练自回归图像生成模型。arXiv 预印本 arXiv:2508.06924，2025。

[386] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476, 2024.

[386] 程泽森, 冷思聪, 张航, 辛亦飞, 李鑫, 陈冠峥, 朱永新, 张文琪, 罗子阳, 赵德立, 等. Videollama 2: 推进视频大模型的时空建模与音频理解. arXiv 预印本 arXiv:2406.07476, 2024.

[387] Yicheng Feng, Yijiang Li, Wanpeng Zhang, Hao Luo, Zihao Yue, Sipeng Zheng, and Zongqing Lu. Videoorion: Tokenizing object dynamics in videos. arXiv preprint arXiv:2411.16156, 2024.

[387] 冯奕成, 李宜江, 张万鹏, 骆皓, 岳子皓, 郑思鹏, 陆宗庆. Videoorion: 对视频中的物体动态进行标记化. arXiv 预印本 arXiv:2411.16156, 2024.

[388] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

[388] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan. Video-chatgpt: 通过大视觉与语言模型迈向细致的视频理解. arXiv 预印本 arXiv:2306.05424, 2023.

[389] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo, 2025. URL https://arxiv.org/abs/2506.07464.

[389] 朴镇泳, 娜智惠, 金振泳, 金贤宇 J. 深度视频-r1: 通过难度感知回归 GRPO 的视频强化微调, 2025. URL https://arxiv.org/abs/2506.07464.

[390] 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, 2025.

[390] 李新豪, 严子昂, 孟德森, 董璐, 曾祥宇, 何怡楠, 王雅丽, 乔煜, 王毅, 王立民. Videochat-r1: 通过强化微调提升时空感知. arXiv 预印本 arXiv:2504.06958, 2025.

[391] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2505.23504.

[391] 祝丽云, 陈启翔, 沈熙, 寸晓东. Vau-r1: 通过强化微调推进视频异常理解, 2025. URL https://arxiv.org/abs/2505.23504.

[392] 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, 2025. URL https://arxiv.org/abs/2504.00883.

[392] 廖振伊, 谢庆松, 张彦浩, 孔子健, 卢浩南, 杨振宇, 邓志杰. 通过类似 r1-zero 的训练改进视觉空间推理. arXiv 预印本 arXiv:2504.00883, 2025. URL https://arxiv.org/abs/2504.00883.

[393] Kun Ouyang. Spatial-r1: Enhancing mllms in video spatial reasoning. arXiv e-prints, pages arXiv-2504, 2025. URL https://arxiv.org/abs/2504.01805.

[393] 欧阳坤. Spatial-r1: 在视频空间推理中增强多模态大模型. arXiv e-prints, 页面 arXiv-2504, 2025. URL https://arxiv.org/abs/2504.01805.

[394] 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, 2025. URL https://arxiv.org/abs/2505.247

[394] 党继晟, 吴靖泽, 王腾, 林轩辉, 朱楠楠, 陈洪波, 郑伟士, 王孟, 蔡达生. 通过聚焦思考加强视频推理, 2025. URL https://arxiv.org/abs/2505.24718.

[395] Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. Egovlm: Policy optimization for egocentric video understanding, 2025. URL https://arxiv.org/abs/2506.03097.

[395] Ashwin Vinod, Shrey Pandit, Aditya Vavre, Linshen Liu. Egovlm: 用于第一视角视频理解的策略优化, 2025. URL https://arxiv.org/abs/2506.03097.

[396] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. arXiv preprint arXiv:2504.09641, 2025.

[396] 张兴健, 温思维, 吴文军, 黄磊. Tinyllava-video-r1: 面向更小的视频推理大模型. arXiv 预印本 arXiv:2504.09641, 2025.

[397] 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, 2025.

[397] 陈煜康, 黄伟, 石百峰, 胡庆豪, 叶涵荣, 朱立庚, 刘志坚, Pavlo Molchanov, Jan Kautz, 齐晓娟, 等. 将强化学习扩展到长视频. arXiv 预印本 arXiv:2507.07966, 2025.

[398] Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. arXiv preprint arXiv:2504.12680, 2025.

[398] 赵百宁, 王子有, 方剑杰, 高晨, 曼凡航, 崔进强, 王鑫, 陈昕磊, 李勇, 朱文武. Embodied-r: 通过强化学习为基础模型激活具身空间推理的协作框架. arXiv 预印本 arXiv:2504.12680, 2025.

[399] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025. URL https://arxiv.org/abs/2505.07818.

[399] 薛泽岳, 吴杰, 高宇, 孔方远, 朱令婷, 陈梦钊, 刘志恒, 刘伟, 郭秋山, 黄维林, 罗平. Dancegrpo: 在视觉生成上释放 GRPO 的能力, 2025. URL https://arxiv.org/abs/2505.07818.

[400] Bingwen Zhu, Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Yidi Wu, Huyang Sun, and Zuxuan Wu. Aligning anime video generation with human feedback, 2025. URL https://arxiv.org/ abs/2504.10044.

[400] 朱炳文, 蒋宇东, 徐保涵, 杨思谦, 殷明宇, 吴一迪, 孙虎阳, 以及吴祖轩. 使动漫视频生成与人类反馈对齐, 2025. URL https://arxiv.org/ abs/2504.10044.

[401] Zhun Mou, Bin Xia, Zhengchao Huang, Wenming Yang, and Jiaya Jia. Gradeo: Towards human-like evaluation for text-to-video generation via multi-step reasoning, 2025. URL https://arxiv.org/ abs/2503.02341.

[401] 邹沅, 夏斌, 黄正超, 杨文明, 以及贾佳雅. Gradeo: 通过多步推理实现类人文本到视频生成评估, 2025. URL https://arxiv.org/ abs/2503.02341.

[402] 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, 2025.

[402] 方学骥, 马立元, 陈志阳, 周明远, 以及齐国俊. Inflvg: 在推理时强化一致性的长视频生成与 grpo. arXiv 预印本 arXiv:2505.17574, 2025.

[403] 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, 2025.

[403] 王琳, 贾立宇, 胡文韬, 潘凯航, 岳中琦, 赵伟, 陈景元, 吴飞, 以及张汉望. 通过强化学习使用扩散时间步标记进行物理推理视频生成. arXiv 预印本 arXiv:2504.15932, 2025.

[404] 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, 2025.

[404] 刘杰, 刘功业, 梁佳俊, 袁子阳, 刘晓坤, 郑明武, 吴协乐, 王秋林, 秦文宇, 夏梦涵, 等. 用人类反馈改进视频生成. arXiv 预印本 arXiv:2501.13918, 2025.

[405] Yuhui Chen, Haoran Li, Zhennan Jiang, Haowei Wen, and Dongbin Zhao. Tevir: Text-to-video reward with diffusion models for efficient reinforcement learning. arXiv preprint arXiv:2505.19769, 2025.

[405] 陈育晖, 李浩然, 蒋振南, 温浩威, 以及赵东斌. Tevir: 使用扩散模型的文本到视频奖励以实现高效强化学习. arXiv 预印本 arXiv:2505.19769, 2025.

[406] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6463-6474, 2024.

[406] 袁航杰, 张世炜, 王翔, 魏宇杰, 冯涛, 潘一宁, 张英雅, 刘子维, Samuel Albanie, 以及倪冬. Instructvideo: 用人类反馈指导视频扩散模型. 载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 页 6463-6474, 2024.

[407] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems, 36:20482-20494, 2023.

[407] 洪一宁, 郑浩宇, 陈佩豪, 郑书泓, 杜一伦, 陈振方, 以及甘创. 3d-llm: 将三维世界注入大型语言模型. Advances in Neural Information Processing Systems, 36:20482-20494, 2023.

[408] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In European Conference on Computer Vision, pages 131-147. Springer, 2024.

[408] 徐润森, 王晓龙, 汪泰, 陈一伦, 庞江淼, 以及林大华. Pointllm: 赋能大型语言模型理解点云. 载于欧洲计算机视觉会议, 页 131-147. Springer, 2024.

[409] Weipeng Deng, Jihan Yang, Runyu Ding, Jiahui Liu, Yijiang Li, Xiaojuan Qi, and Edith Ngai. Can 3d vision-language models truly understand natural language? arXiv preprint arXiv:2403.14760, 2024.

[409] 邓伟鹏, 杨继涵, 丁润宇, 刘家辉, 李义江, 齐晓娟, 以及魏爱迪. 3d 视图-语言模型是否能真正理解自然语言? arXiv 预印本 arXiv:2403.14760, 2024.

[410] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14455-14465, 2024.

[410] 陈博远, 徐卓, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, 以及贾飞. Spatialvlm: 赋予视觉-语言模型空间推理能力. 载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 页 14455-14465, 2024.

[411] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. arXiv preprint arXiv:2310.06773, 2023.

[411] 周俊胜, 王晋生, 马保瑞, 刘宇深, 黄铁军, 以及王新龙. Uni3d: 在大规模上探索统一的三维表征. arXiv 预印本 arXiv:2310.06773, 2023.

[412] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26428-26438, 2024.

[412] 陈思谨, 陈昕, 张驰, 李明盛, 于刚, 费昊, 朱宏远, 范佳远, 以及陈涛. Ll3da: 面向全方位三维理解、推理与规划的视觉交互式指令微调. 载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 页 26428-26438, 2024.

[413] Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. Llama-mesh: Unifying 3d mesh generation with language models. arXiv preprint arXiv:2411.09595, 2024.

[413] 王正毅, Jonathan Lorraine, 王轶凯, 苏航, 祝俊, Sanja Fidler, 以及曾晓辉. Llama-mesh: 将三维网格生成与语言模型统一. arXiv 预印本 arXiv:2411.09595, 2024.

[414] Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Wen Liu, Gang Yu, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. IEEE Transactions on Multimedia, 2025.

[414] 殷福坤, 陈欣, 张驰, 蒋彪, 赵子博, 刘文, 于刚, 和陈涛。Shapegpt: 使用统一多模态语言模型的 3D 形状生成。IEEE Transactions on Multimedia, 2025.

[415] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 19615-19625, 2024.

[415] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, 和 Matthias Nießner。Meshgpt: 使用仅解码器变换器生成三角网格。在 IEEE/CVF 计算机视觉与模式识别会议论文集，页码 19615-19625， 2024。

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

[416] 潘振宇和刘寒。Metaspatial: 在元宇宙中增强 VLM 的 3D 空间推理。arXiv 预印本 arXiv:2503.18470, 2025。

[417] Zhihao Yuan, Shuyi Jiang, Chun-Mei Feng, Yaolun Zhang, Shuguang Cui, Zhen Li, and Na Zhao. Scene-r1: Video-grounded large language models for 3d scene reasoning without 3d annotations. arXiv preprint arXiv:2506.17545, 2025.

[417] 袁志豪, 蒋舒怡, 冯春梅, 张尧伦, 崔曙光, 李震, 和赵娜。Scene-r1: 基于视频的大型语言模型用于无 3D 注释的 3D 场景推理。arXiv 预印本 arXiv:2506.17545, 2025。

[418] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spa-tialreasoner: Towards explicit and generalizable 3d spatial reasoning. arXiv preprint arXiv:2504.20024, 2025.

[418] 马五飞, Yu-Cheng Chou, 刘祺豪, 王星睿, Celso de Melo, 谢建文, 和 Alan Yuille。Spa-tialreasoner: 迈向显式且可泛化的 3D 空间推理。arXiv 预印本 arXiv:2504.20024, 2025。

[419] Xiandong Zou, Ruihao Xia, Hongsong Wang, and Pan Zhou. Dreamcs: Geometry-aware text-to-3d generation with unpaired 3d reward supervision. arXiv preprint arXiv:2506.09814, 2025. URL https://arxiv.org/abs/2506.09814.

[419] 邹先东, 夏瑞昊, 王洪松, 和周攀。Dreamcs: 具有无配对 3D 奖励监督的几何感知文本到 3D 生成。arXiv 预印本 arXiv:2506.09814, 2025。URL https://arxiv.org/abs/2506.09814。

[420] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. arXiv preprint arXiv:2502.04370, 2025. URL https://arxiv.org/abs/2502.04370.

[420] 周正林, 夏晓波, 马帆, 樊鹤鹤, 杨毅, 和赵达生。Dreamdpo: 通过直接偏好优化使文本到 3D 生成与人类偏好对齐。arXiv 预印本 arXiv:2502.04370, 2025。URL https://arxiv.org/abs/2502.04370。

[421] Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun

Zhu. Dreamreward: Text-to-3d generation with human preference. In European Conference on Computer Vision, pages 259-276. Springer, 2024. URL https://icm1.cc/virtual/2025/ poster/45024.

[421] 叶君良，刘方福，李奇修，王正奕，王一凯，王新洲，段岳琦，和朱军。Dreamreward: 具有人工偏好的文本到 3D 生成。在欧洲计算机视觉大会，页码 259-276。Springer，2024。URL https://icm1.cc/virtual/2025/poster/45024。

[422] Qingming Liu, Zhen Liu, Dinghuai Zhang, and Kui Jia. Nabla-r2d3: Effective and efficient 3d diffusion alignment with 2d rewards. arXiv preprint arXiv:2506.15684, 2025. URL https://arxiv.org/ abs/2506.15684.

[422] 刘庆明, 刘震, 张定槐, 和贾奎。Nabla-r2d3: 使用 2D 奖励进行高效有效的 3D 扩散对齐。arXiv 预印本 arXiv:2506.15684, 2025。URL https://arxiv.org/abs/2506.15684。

[423] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

[423] 金武镇, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, 等。Openvla: 一个开源视觉-语言-行动模型。arXiv 预印本 arXiv:2406.09246, 2024。

[424] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\diagdown\pi_0$: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

[424] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, 等。$\diagdown\pi_0$: 一种用于通用机器人控制的视觉-语言-行动流模型。arXiv 预印本 arXiv:2410.24164, 2024。

[425] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020, 2025.

[425] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, 等。Gemini robotics: 将 AI 带入物理世界。arXiv 预印本 arXiv:2503.20020, 2025。

[426] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. arXiv preprint arXiv:2508.05635, 2025.

[426] 岳辽、周鹏飞、黄思远、杨东林、陈胜聪、蒋宇新、胡越、蔡敬斌、刘思、罗剑兰、陈立良、严水成、姚茂庆和任光辉。Genie envisioner: 面向机器人操作的统一世界基础平台。arXiv 预印本 arXiv:2508.05635， 2025。

[427] 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.

[427] SimpleVLA-RL 团队。Simplevla-rl: 简单奖励的在线强化学习使得仅用一条轨迹即可训练 VLA 模型。https://github.com/PRIME-RL/SimpleVLA-RL，2025。GitHub 仓库。

[428] 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, 2025. URL https://arxiv.org/abs/2505.18719.

[428] 陆冠兴、郭文凯、张楚斌、周育恒、江浩楠、高子峰、唐艳松和王子伟。VLA-RL: 通过可扩展强化学习走向娴熟且通用的机器人操作，2025。URL https://arxiv.org/abs/2505.18719。

[429] Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. arXiv preprint arXiv:2506.17221, 2025.

[429] 齐章扬、张志雄、余弈周、王嘉琪和赵横爽。VLN-R1: 通过强化微调的视觉-语言导航。arXiv 预印本 arXiv:2506.17221，2025。

[430] Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Zhenhao Chen, et al. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. arXiv preprint arXiv:2505.16517, 2025.

[430] 宋子睿、欧阳光贤、李明哲、纪宇恒、王陈熙、徐子翔、张泽宇、张晓庆、姜谦、陈振豪等。Maniplvm-R1: 用于具身操作推理的大型视觉-语言模型的强化学习。arXiv 预印本 arXiv:2505.16517，2025。

[431] Han Zhao, Wenxuan Song, Donglin Wang, Xinyang Tong, Pengxiang Ding, Xuelian Cheng, and Zongyuan Ge. More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. arXiv preprint arXiv:2503.08007, 2025.

[431] 赵涵、宋文轩、王东林、童新阳、丁朋翔、程雪莲和葛宗元。MORE: 为四足视觉-语言-动作模型解锁强化学习可扩展性。arXiv 预印本 arXiv:2503.08007，2025。

[432] Honglin He, Yukai Ma, Wayne Wu, and Bolei Zhou. From seeing to experiencing: Scaling navigation foundation models with reinforcement learning. arXiv preprint arXiv:2507.22028, 2025. URL https://arxiv.org/abs/2507.22

[432] 何鸿霖、马郁凯、吴文和、周博磊。从“看见”到“体验”: 用强化学习扩展导航基础模型。arXiv 预印本 arXiv:2507.22028，2025。URL https://arxiv.org/abs/2507.22028。

[433] 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, 2025.

[433] 刘继佳、高峰、魏炳文、陈新磊、廖青民、吴毅、于超和王宇。强化学习能为 VLA 泛化带来什么？一项实证研究。arXiv 预印本 arXiv:2505.19789，2025。

[434] Zengjue Chen, Runliang Niu, He Kong, and Qi Wang. Tgrpo: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization. arXiv preprint arXiv:2506.08440, 2025.

[434] 陈增觉、牛润良、孔和和王琦。TGRPO: 通过轨迹级群相对策略优化微调视觉-语言-动作模型。arXiv 预印本 arXiv:2506.08440，2025。

[435] Li Kang, Xiufeng Song, Heng Zhou, Yiran Qin, Jie Yang, Xiaohong Liu, Philip Torr, Lei Bai, and Zhenfei Yin. Viki-r: Coordinating embodied multi-agent cooperation via reinforcement learning. arXiv preprint arXiv:2506.09049, 2025.

[435] 康立、宋秀峰、周恒、秦一然、杨杰、刘晓宏、Philip Torr、白磊和尹振飞。VIKI-R: 通过强化学习协调具身多智能体协作。arXiv 预印本 arXiv:2506.09049，2025。

[436] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems, 36:51991-52008, 2023. URL https://openreview.net/forum?id= 3IyL2XWDkG.

[436] 李国豪、Hasan Hammoud、Hani Itani、Dmitrii Khizbullin 和 Bernard Ghanem。CAMEL: 用于大型语言模型社会"心智"探索的交流代理。NeurIPS，36:51991-52008，2023。URL https://openreview.net/forum?id= 3IyL2XWDkG。

[437] Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=h0ZfDIrj7T.

[437] 王俊林、王珏、Ben Athiwaratkun、Ce Zhang 和 James Zou。Mixture-of-agents 提升大型语言模型能力。收录于第十三届国际表征学习会议，2025。URL https://openreview.net/forum?id=h0ZfDIrj7T。

[438] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17889-17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

[438] 田亮，何志伟，焦文翔，王星，王艳，王睿，杨宇久，石树明，和涂兆鹏。通过多智能体辩论鼓励大语言模型的发散思维。载于 Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen 主编，2024 年实证方法与自然语言处理会议论文集，页 17889-17904，佛罗里达州迈阿密，2024 年 11 月。计算语言学协会。doi: 10.18653/v1/2024.emnlp-main.992。URL https://aclanthology.org/2024.emnlp-main.992/。

[439] Wangchunshu Zhou, Yixin Ou, Shengwei Ding, Long Li, Jialong Wu, Tiannan Wang, Jiamin Chen, Shuai Wang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, and Yuchen Eleanor Jiang. Symbolic learning enables self-evolving agents, 2024. URL https://arxiv.org/abs/2406.18532.

[439] 周望春树，欧一昕，丁胜威，李龙，吴家龙，王天南，陈佳敏，王帅，徐晓华，张宁宇，陈华俊，和江昀尧·伊莱诺。符号学习促成自我进化智能体，2024。URL https://arxiv.org/abs/2406.18532。

[440] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmid-huber. GPTSwarm: Language agents as optimizable graphs. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=uTC9AFXIhg.

[440] 朱格明辰，王文意，Louis Kirsch，Francesco Faccio，Dmitrii Khizbullin，和 Jürgen Schmid-huber。GPTSwarm: 将语言智能体视为可优化图。在第 41 届国际机器学习会议，2024。URL https://openreview.net/forum?id=uTC9AFXIhg。

[441] Xiaowen Ma, Chenyang Lin, Yao Zhang, Volker Tresp, and Yunpu Ma. Agentic neural networks: Self-evolving multi-agent systems via textual backpropagation. arXiv preprint arXiv:2506.09046, 2025.

[441] 马晓文，林晨阳，张尧，Volker Tresp，和马云璞。具智能的神经网络: 通过文本反向传播实现自我进化的多智能体系统。arXiv 预印本 arXiv:2506.09046，2025。

[442] Heng Zhou, Hejia Geng, Xiangyuan Xue, Li Kang, Yiran Qin, Zhiyong Wang, Zhenfei Yin, and Lei Bai. Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. URL https://openreview.net/forum?id=te0j

[442] 周恒，耿贺佳，薛翔远，康力，秦一然，汪志勇，殷振飞，和白磊。Reso: 一种面向推理任务、以奖励驱动的基于大型语言模型的自组织多智能体系统。载于 2025 年实证方法与自然语言处理会议论文集，2025。URL https://openreview.net/forum?id=te0jBwgBRm。

[443] Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/ forum?id=t9U3LW7JVX.

[443] 胡胜然，陆聪，和 Jeff Clune。智能体系统的自动化设计。载于第十三届国际学习表征会议，2025。URL https://openreview.net/ forum?id=t9U3LW7JVX。

[444] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=z5uVAKwmjf.

[444] 张佳怡，向金宇，于照阳，滕凤伟，陈雄辉，陈佳琦，朱格明辰，程昕，洪思睿，王金林，郑炳南，刘邦，罗雨雨，和吴成林。AFlow: 自动生成具智能工作流。载于第十三届国际学习表征会议，2025。URL https://openreview.net/forum?id=z5uVAKwmjf。

[445] Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. arXiv preprint arXiv:2505.22954, 2025.

[445] 张珍妮，胡胜然，陆聪，Robert Lange，和 Jeff Clune。达尔文 Gödel 机: 自我改进智能体的开放式进化。arXiv 预印本 arXiv:2505.22954，2025。

[446] Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, LEI BAI, and Xiang Wang. Multi-agent architecture search via agentic supernet. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=imcyVlzpXh.

[446] 张桂斌，牛鲁阳，方俊锋，王坤，白磊，和王翔。通过具智能超网的多智能体架构搜索。载于第42届国际机器学习会议，2025。URL https://openreview.net/forum?id=imcyVlzpXh。

[447] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id

[447] 张桂彬，岳彦伟，孙向国，万冠成，俞淼，方俊峰，王琨，陈天龙，以及程大为. G-designer: 通过图神经网络构建多智能体通信拓扑. 收录于第42届国际机器学习大会，2025. URL https://openreview.net/forum?id=LpE54NUnmO.

[448] Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Optima: Optimizing effectiveness and efficiency for llm-based multi-agent system, 2025. URL https://arxiv.org/ abs/2410.08115.

[448] 陈伟泽, 袁佳瑞, 钱晨, 杨成, 刘志远, 孙茂松. Optima: 优化基于大模型的多智能体系统的有效性与效率, 2025. URL https://arxiv.org/ abs/2410.08115.

[449] Wentao Shi, Zichun Yu, Fuli Feng, Xiangnan He, and Chenyan Xiong. Efficient multi-agent system training with data influence-oriented tree search, 2025. URL https://arxiv.org/abs/2502.00955.

[449] 石文涛, 余子纯, 冯福利, 何向南, 侑辰煊. 基于数据影响导向树搜索的高效多智能体系统训练, 2025. URL https://arxiv.org/abs/2502.00955.

[450] Junwei Liao, Muning Wen, Jun Wang, and Weinan Zhang. Marft: Multi-agent reinforcement finetuning, 2025. URL https://arxiv.org/abs/2504.16129.

[450] 廖俊伟，温穆宁，王俊，张伟楠. Marft: 多智能体强化微调，2025. URL https://arxiv.org/abs/2504.16129.

[451] Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E. Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30215-30248, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1459. URL https:// aclanthology.org/2025.acl-long.1459/.

[451] Park Chanwoo, Han Seungju, Guo Xingzhi, Asuman E. Ozdaglar, Zhang Kaiqing, Kim Joo-Kyung. MAPoRL: 用强化学习对协作型大型语言模型进行多智能体后共训. 收录于 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, Mohammad Taher Pilehvar 编辑的第63届计算语言学协会年会论文集 (第1卷: 长论文), 页 30215-30248, 奥地利维也纳, 2025年7月. 计算语言学协会. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1459. URL https:// aclanthology.org/2025.acl-long.1459/.

[452] Andrew Estornell, Jean-Francois Ton, Muhammad Faaiz Taufiq, and Hang Li. How to train a leader: Hierarchical reasoning in multi-agent llms. arXiv preprint arXiv:2507.08960, 2025. URL https: //arxiv.org/abs/2507.089

[452] Andrew Estornell, Jean-Francois Ton, Muhammad Faaiz Taufiq, Hang Li. 如何培养一位领导者: 多智能体大模型中的分层推理. arXiv 预印本 arXiv:2507.08960, 2025. URL https: //arxiv.org/abs/2507.08960.

[453] Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. Rema: Learning to meta-think for llms with multi-agent reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09501.

[453] 万子煜, 李云翔, 文晓宇, 宋岩, 王涵静, 杨林倚, Mark Schmidt, 王俊, 张伟南, 胡书岳, 文颖. Rema: 使用多智能体强化学习学习元思考以改进大模型, 2025. URL https://arxiv.org/abs/2503.09501.

[454] Hongcheng Gao, Yue Liu, Yufei He, Longxu Dou, Chao Du, Zhijie Deng, Bryan Hooi, Min Lin, and Tianyu Pang. Flowreasoner: Reinforcing query-level meta-agents, 2025. URL https://arxiv.org/abs/2504.15257.

[454] 高宏成, 刘岳, 何宇飞, 窦龙旭, 杜超, 邓之杰, Bryan Hooi, 林敏, 庞天宇. Flowreasoner: 强化查询级元智能体, 2025. URL https://arxiv.org/abs/2504.15257.

[455] 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, 2025. URL https://arxiv.org/abs/2506.00555.

[455] 夏鹏, 王精璐, 彭亦博, 曾凯德, 吴先, 唐翔儒, 朱鸿图, 李云, 刘树杰, 陆岩, 等. Mmedagent-rl: 优化多智能体协作以实现多模态医学推理. arXiv 预印本 arXiv:2506.00555, 2025. URL https://arxiv.org/abs/2506.00555.

[456] Weizhen Li, Jianbo Lin, Zhuosong Jiang, Jingyi Cao, Xinpeng Liu, Jiayu Zhang, Zhenqiang Huang, Qianben Chen, Weichen Sun, Qiexiang Wang, Hongxuan Lu, Tianrui Qin, Chenghao Zhu, Yi Yao, Shuying Fan, Xiaowan Li, Tiannan Wang, Pai Liu, King Zhu, He Zhu, Dingfeng Shi, Piaohong Wang, Yeyi Guan, Xiangru Tang, Minghao Liu, Yuchen Eleanor Jiang, Jian Yang, Jiaheng Liu, Ge Zhang, and Wangchunshu Zhou. Chain-of-agents: End-to-end agent foundation models via multi-agent distillation and agentic rl, 2025. URL https://arxiv.org/abs/2508.13167.

[456] 李伟臻, 林建波, 蒋卓淞, 曹婧怡, 刘新鹏, 张佳宇, 黄振强, 陈千本, 孙伟宸, 王且翔, 陆泓轩, 秦天瑞, 朱成浩, 姚奕, 范淑英, 李晓婉, 王天南, 刘拍, 朱京, 朱和, 石鼎峰, 王飘红, 关叶一, 唐翔儒, 刘明浩, 江语晨, 杨建, 刘佳恒, 张戈, 周望春树. Chain-of-agents: 通过多智能体蒸馏与智能体化强化学习实现端到端智能体基础模型, 2025. URL https://arxiv.org/abs/2508.13167.

[457] Wenzhen Yuan, Shengji Tang, Weihao Lin, Jiacheng Ruan, Ganqu Cui, Bo Zhang, Tao Chen, Ting Liu, Yuzhuo Fu, Peng Ye, and Lei Bai. Wisdom of the crowd: Reinforcement learning from coevolutionary collective feedback, 2025. URL https://arxiv.org/abs/2508.12338.

[457] 袁文臻, 唐胜吉, 林伟浩, 阮家诚, 崔甘渠, 张博, 陈涛, 刘婷, 傅宇卓, 叶鹏, 和白磊. 群体智慧: 来自共同进化集体反馈的强化学习, 2025. URL https://arxiv.org/abs/2508.12338.

[458] Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. Llm collaboration with multi-agent reinforcement learning, 2025. URL https://arxiv.org/abs/2508.04652.

[458] 刘硕, 梁泽宇, 吕学广, 和 Christopher Amato. 通过多智能体强化学习的 LLM 协作, 2025. URL https://arxiv.org/abs/2508.04652.

[459] Yuan Wei, Xiaohan Shan, and Jianmin Li. Lero: Llm-driven evolutionary framework with hybrid rewards and enhanced observation for multi-agent reinforcement learning, 2025. URL https: //arxiv.org/abs/2503.21807.

[459] 袁伟, 单晓涵, 和李建民. LERO: 面向多智能体强化学习的由 LLM 驱动的进化框架, 兼具混合奖励与增强观测, 2025. URL https://arxiv.org/abs/2503.21807.

[460] Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning, 2025. URL https:// arxiv.org/abs/2506.24119.

[460] 刘博, Leon Guertler, Simon Yu, 刘紫琛, 齐鹏辉, Daniel Balcells, Mickel Liu, Cheston Tan, 施伟言, 林敏, Wee Sun Lee, 和 Natasha Jaques. SPIRAL: 零和博弈中的自我博弈通过多智能体多回合强化学习激励推理, 2025. URL https://arxiv.org/abs/2506.24119.

[461] Ruihan Yang, Yikai Zhang, Aili Chen, Xintao Wang, Siyu Yuan, Jiangjie Chen, Deqing Yang, and Yanghua Xiao. Aria: Training language agents with intention-driven reward aggregation, 2025. URL https://arxiv.org/abs/2506.00539.

[461] 杨睿涵, 张奕凯, 陈爱立, 王欣涛, 袁思宇, 陈江杰, 杨德清, 和萧杨华. ARIA: 用意图驱动的奖励聚合训练语言代理, 2025. URL https://arxiv.org/abs/2506.00539.

[462] Xuchen Pan, Yanxi Chen, Yushuo Chen, Yuchang Sun, Daoyuan Chen, Wenhao Zhang, Yuexiang Xie, Yilun Huang, Yilei Zhang, Dawei Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. Trinity-rft: A general-purpose and unified framework for reinforcement fine-tuning of large language models, 2025. URL https://arxiv.org/abs/2505.17826.

[462] 潘煦辰, 陈言希, 陈宇硕, 孙昱畅, 陈道远, 张文豪, 谢岳翔, 黄一伦, 张一磊, 高大为, 李雅良, 丁博林, 和周靖人. TRINITY-RFT: 用于大语言模型强化微调的通用统一框架, 2025. URL https://arxiv.org/abs/2505.17826.

[463] Shu Liu, Sumanth Hegde, Shiyi Cao, Alan Zhu, Dacheng Li, Tyler Griggs, Eric Tang, Akshay Malik, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Skyrl-sql: Matching gpt-40 and o4-mini on text2sql with multi-turn rl, 2025. URL https://github.com/NovaSky-AI/SkyRL.

[463] 刘殊, Sumanth Hegde, 曹诗怡, 朱安然, 李大成, Tyler Griggs, Eric Tang, Akshay Malik, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, 和 Ion Stoica. SKYRL-SQL: 用多回合 RL 在 Text2SQL 任务上匹配 GPT-40 与 O4-Mini, 2025. URL https://github.com/NovaSky-AI/SkyRL.

[464] 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, 2025. URL https://arxiv.org/abs/2508.13587.

[464] 陈蕾, 赵宣乐, 曾智雄, 黄靖, 郑黎明, 钟玉峰, 和马霖. 打破 SFT 平台: 用于图表到代码生成的多模态结构化强化学习. arXiv preprint arXiv:2508.13587, 2025. URL https://arxiv.org/abs/2508.13587.

[465] Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. Time-r1: Towards comprehensive temporal reasoning in llms, 2025. URL https://arxiv.org/abs/2505.13508.

[465] 刘子佳, 韩沛轩, 余浩飞, 李浩如, 和游家轩. TIME-R1: 迈向 LLM 的全面时间推理, 2025. URL https://arxiv.org/abs/2505.13508.

[466] Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. Timemaster: Training time-series multimodal llms to reason via reinforcement learning, 2025. URL https://arxiv.org/abs/2506.13705.

[466] 张君儒, 冯朗, 郭旭, 吴愉涵, 董雅博, 和徐端庆. TIMEMASTER: 通过强化学习训练时间序列多模态 LLM 进行推理, 2025. URL https://arxiv.org/abs/2506.13705.

[467] Yuxiang Zhang, Yuqi Yang, Jiangming Shu, Xinyan Wen, and Jitao Sang. Agent models: Internalizing chain-of-action generation into reasoning models, 2025. URL https://arxiv.org/abs/2503.06580.

[467] 张玉祥, 杨宇琪, 舒江铭, 温馨妍, 和桑吉涛. AGENT MODELS: 将行动链生成内化进推理模型, 2025. URL https://arxiv.org/abs/2503.06580.

[468] Junjie Zhang, Jingyi Xi, Zhuoyang Song, Junyu Lu, Yuhua Ke, Ting Sun, Yukun Yang, Jiaxing Zhang, Songxin Zhang, and Zejian Xie. L0: Reinforcement learning to become general agents, 2025. URL https://arxiv.org/abs/2506.23667.

[468] 张俊杰, 席静怡, 宋卓阳, 吕君宇, 柯玉华, 孙婷, 杨郁坤, 张嘉兴, 张颂新, 和谢泽剑. L0: 通过强化学习成为通用代理, 2025. URL https://arxiv.org/abs/2506.23667.

[469] Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence, 2025. URL https://arxiv.org/abs/2508.03905.

[469] 余浩飞, 齐铮阳, 赵依宁, Kolby Nottingham, 宣科杨, Bodhisattwa Prasad Majumder, 朱昊, Paul Pu Liang, 和游家轩. SOTOPIA-RL: 为社会智能设计奖励, 2025. URL https://arxiv.org/abs/2508.03905.

[470] Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. Adaptive thinking via mode policy optimization for social language agents, 2025. URL https://arxiv.org/abs/2505.02156.

[470] Minzheng Wang、Yongbin Li、Haobo Wang、Xinghua Zhang、Nan Xu、Bingli Wu、Fei Huang、Haiyang Yu 和 Wenji Mao。通过模式策略优化实现社交语言代理的自适应思维，2025。URL https://arxiv.org/abs/2505.02156.

[471] Marwa Abdulhai, Isadora White, Charlie Victor Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. LMRL gym: Benchmarks for multi-turn reinforcement learning with language models. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=hmGh

[471] Marwa Abdulhai、Isadora White、Charlie Victor Snell、Charles Sun、Joey Hong、Yuexiang Zhai、Kelvin Xu 和 Sergey Levine。LMRL gym: 用于基于语言模型的多轮强化学习基准。在第四十二届国际机器学习会议，2025。URL https://openreview.net/forum?id=hmGhP5DO2W.

[472] 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, 2021. URL https://openreview.net/ forum?id=01C0YcCdTn.

[472] Mohit Shridhar、Xingdi Yuan、Marc-Alexandre Cote、Yonatan Bisk、Adam Trischler 和 Matthew Hausknecht。AlfWorld: 对齐文本与具身环境以进行交互式学习。在国际表征学习会议，2021。URL https://openreview.net/ forum?id=01C0YcCdTn.

[473] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. In Tristan Cazenave, Abdallah Saffidine, and Nathan Sturtevant, editors, Computer Games, pages 41-75, Cham, 2019. Springer International Publishing.

[473] Marc-Alexandre Côté、Ákos Kádár、Xingdi Yuan、Ben Kybartas、Tavian Barnes、Emery Fine、James Moore、Matthew Hausknecht、Layla El Asri、Mahmoud Adada、Wendy Tay 和 Adam Trischler。TextWorld: 面向基于文本游戏的学习环境。在 Tristan Cazenave、Abdallah Saffidine 和 Nathan Sturtevant 编辑的 Computer Games，一页 41-75，Cham，2019。Springer International Publishing.

[474] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11279-11298, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

[474] Ruoyao Wang、Peter Jansen、Marc-Alexandre Côté 和 Prithviraj Ammanabrolu。ScienceWorld: 你的代理比小学五年级学生更聪明吗? 在 Yoav Goldberg、Zornitsa Kozareva 和 Yue Zhang 编辑的 2022 年实证方法自然语言处理会议论文集,页码 11279-11298,阿布扎比,阿拉伯联合酋长国,2022 年 12 月。计算语言学协会。doi: 10.18653/v1/2022.emnlp-main.775。URL https://aclanthology.org/2022.emnlp-main.775/.

[475] Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. AgentGym: Evaluating and training large language model-based agents across diverse environments. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27914-27961, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1355. URL https://aclanthology.org/2025.acl-long.1355/.

[475] Zhiheng Xi、Yiwen Ding、Wenxiang Chen、Boyang Hong、Honglin Guo、Junzhe Wang、Xin Guo、Dingwen Yang、Chenyang Liao、Wei He、Songyang Gao、Lu Chen、Rui Zheng、Yicheng Zou、Tao Gui、Qi Zhang、Xipeng Qiu、Xuanjing Huang、Zuxuan Wu和Yu-Gang Jiang。AgentGym: 在多样化环境中评估与训练基于大型语言模型的代理。在Wanxiang Che、Joyce Nabende、Ekaterina Shutova和Mohammad Taher Pilehvar编辑的第63届计算语言学协会年会论文集 (卷1: 长文)，页码27914-27961，维也纳，奥地利，2025年7月。计算语言学协会。ISBN 979-8-89176-251-0。doi: 10.18653/v1/2025.acl-long.1355。URL https://aclanthology.org/2025.acl-long.1355/.

[476] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. In ICLR, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

[476] Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xuanyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang、Shudan Zhang、Xiang Deng、Aohan Zeng、Zhengxiao Du、Chenhui Zhang、Sheng Shen、Tianjun Zhang、Yu Su、Huan Sun、Minlie Huang、Yuxiao Dong和Jie Tang。Agentbench: 评估LLM作为代理。在ICLR，2024。URL https://openreview.net/forum?id=zAdUB0aCTQ.

[477] Peiji Li, Jiasheng Ye, Yongkang Chen, Yichuan Ma, Zijie Yu, Kedi Chen, Ganqu Cui, Haozhan Li, Jiacheng Chen, Chengqi Lyu, Wenwei Zhang, Linyang Li, Qipeng Guo, Dahua Lin, Bowen Zhou, and Kai Chen. Internbootcamp technical report: Boosting llm reasoning with verifiable task scaling, 2025. URL https://arxiv.org/abs/2508.08636.

[477] Peiji Li、Jiasheng Ye、Yongkang Chen、Yichuan Ma、Zijie Yu、Kedi Chen、Ganqu Cui、Haozhan Li、Jiacheng Chen、Chengqi Lyu、Wenwei Zhang、Linyang Li、Qipeng Guo、Dahua Lin、Bowen Zhou和Kai Chen。Internbootcamp技术报告: 通过可验证任务扩展提升LLM推理，2025。URL https://arxiv.org/abs/2508.08636.

[478] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents, 2024. URL https://arxiv.org/abs/2402.17753.

[478] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 评估长时间对话记忆的LLM代理，2024。URL https://arxiv.org/abs/2402.17753.

[479] Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions, 2025. URL https://arxiv.org/abs/2507.05257.

[479] Yuanzhe Hu, Yu Wang, and Julian McAuley. 通过递增多轮交互评估 LLM 代理的记忆，2025。URL https://arxiv.org/abs/2507.05257.

[480] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 20744-20757. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ 82ad13ec01f9fe44c01cb91814f Paper-Conference.pdf.

[480] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: 面向具备落地性的语言代理的可扩展真实网页交互。收录于 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, 和 A. Oh 编辑的 Advances in Neural Information Processing Systems, 第 35 卷，第 20744 – 20757 页。Curran Associates, Inc., 2022。URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd7b8c-Paper-Conference.pdf.

[481] Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2web 2: Evaluating agentic search with agent-as-a-judge, 2025. URL https://arxiv.org/abs/2506.21506.

[481] Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2web 2: 以代理为裁判评估具代理性的搜索，2025。URL https://arxiv.org/abs/2506.21506.

[482] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

[482] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: 用于构建自治代理的真实网页环境。收录于第十二届国际学习表征会议，2024。URL https://openreview.net/forum?id=oKn9c6ytLx.

[483] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 881-905, 2024.

[483] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: 在真实视觉网页任务上评估多模态代理。收录于第 62 届计算语言学协会年会论文集 (第一卷: 长篇论文)，第 881–905 页，2024。

[484] Hongru Wang, Rui Wang, Boyang Xue, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z. Pan, and Kam-Fai Wong. AppBench: Planning of multiple APIs from various APPs for complex user instruction. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15322-15336, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.856. URL https://aclanthology.org/2024.emnlp-main.856/.

[484] Hongru Wang, Rui Wang, Boyang Xue, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z. Pan, and Kam-Fai Wong. AppBench: 为复杂用户指令规划来自多款应用的多个 API。收录于 Yaser Al-Onaizan, Mohit Bansal, 和 Yun-Nung Chen 编辑的 2024 年经验方法在自然语言处理会议论文集，页 15322–15336, Miami, Florida, USA, 2024 年 11 月。计算语言学协会。doi: 10.18653/v1/2024.emnlp-main.856。URL https://aclanthology.org/2024.emnlp-main.856/.

[485] 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 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 16022-16076, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

[485] Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: 用于基准测试交互式编码代理的可控应用与人物世界。收录于 Lun-Wei Ku, Andre Martins, 和 Vivek Srikumar 编辑的第 62 届计算语言学协会年会论文集 (第一卷: 长篇论文)，第 16022–16076 页，曼谷，泰国，2024 年 8 月。计算语言学协会。doi: 10.18653/v1/2024.acl-long.850。URL https://aclanthology.org/2024.acl-long.850/.

[486] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmarking environment for autonomous agents. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=il5yUQsrjC.

[486] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmarking environment for autonomous agents. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=il5yUQsrjC.

[487] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caim-

ing Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 52040- 52094. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_ files/paper/2024/file/5d413e48f84dc61244b6be550f1cd8f5-Paper-Datasets_ and Benchmarks_Track.pdf.

[487] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caim-ing Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 52040- 52094. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_ files/paper/2024/file/5d413e48f84dc61244b6be550f1cd8f5-Paper-Datasets_ and Benchmarks_Track.pdf.

[488] Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Jang, and Zack Hui. Windows agent arena: Evaluating multi-modal os agents at scale, 2024. URL https://arxiv.org/abs/2409.08264.

[488] Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Jang, and Zack Hui. Windows agent arena: Evaluating multi-modal os agents at scale, 2024. URL https://arxiv.org/abs/2409.08264.

[489] Xingdi Yuan, Morgane M Moss, Charbel El Feghali, Chinmay Singh, Darya Moldavskaya, Drew MacPhee, Lucas Caccia, Matheus Pereira, Minseon Kim, Alessandro Sordoni, and Marc-Alexandre Côté. debug-gym: A text-based environment for interactive debugging, 2025. URL https://arxiv.org/abs/2503.21557.

[489] Xingdi Yuan, Morgane M Moss, Charbel El Feghali, Chinmay Singh, Darya Moldavskaya, Drew MacPhee, Lucas Caccia, Matheus Pereira, Minseon Kim, Alessandro Sordoni, and Marc-Alexandre Côté. debug-gym: A text-based environment for interactive debugging, 2025. URL https://arxiv.org/abs/2503.21557.

[490] Rushi Qiang, Yuchen Zhuang, Yinghao Li, Dingu Sagar V K, Rongzhi Zhang, Changhao Li, Ian Shu-Hei Wong, Sherry Yang, Percy Liang, Chao Zhang, and Bo Dai. Mle-dojo: Interactive environments for empowering llm agents in machine learning engineering, 2025. URL https://arxiv.org/abs/ 2505.07782.

[490] Rushi Qiang, Yuchen Zhuang, Yinghao Li, Dingu Sagar V K, Rongzhi Zhang, Changhao Li, Ian Shu-Hei Wong, Sherry Yang, Percy Liang, Chao Zhang, and Bo Dai. Mle-dojo: Interactive environments for empowering llm agents in machine learning engineering, 2025. URL https://arxiv.org/abs/ 2505.07782.

[491] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2 -bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

[491] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2 -bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

[492] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks, 2024. URL https://arxiv.org/abs/2412.4161.

[492] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks, 2024. URL https://arxiv.org/abs/2412.4161.

[493] Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D. Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Carl Yang, Yang Xie, and Wenqi Shi. Medagentgym: Training llm agents for code-based medical reasoning at scale, 2025. URL https://arxiv.org/ abs/2506.04405.

[493] Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D. Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Carl Yang, Yang Xie, and Wenqi Shi. Medagentgym: Training llm agents for code-based medical reasoning at scale, 2025. URL https://arxiv.org/ abs/2506.04405.

[494] Connor Dilgren, Purva Chiniya, Luke Griffith, Yu Ding, and Yizheng Chen. Secrepobench: Benchmarking llms for secure code generation in real-world repositories, 2025. URL https://arxiv.org/abs/2504.21205.

[494] Connor Dilgren, Purva Chiniya, Luke Griffith, Yu Ding, and Yizheng Chen. Secrepobench: Benchmarking llms for secure code generation in real-world repositories, 2025. URL https://arxiv.org/abs/2504.21205.

[495] Naman Jain, Jaskirat Singh, Manish Shetty, Tianjun Zhang, Liang Zheng, Koushik Sen, and Ion Stoica. R2e-gym: Procedural environment generation and hybrid verifiers for scaling open-weights SWE agents. In Second Conference on Language Modeling, 2025. URL https://openreview.net/ forum?id=7evvwwdo3z.

[495] Naman Jain、Jaskirat Singh、Manish Shetty、Tianjun Zhang、Liang Zheng、Koushik Sen 和 Ion Stoica。R2e-gym: 用于扩展开放权重 SWE 代理的过程化环境生成与混合验证器。载于第二届语言建模会议，2025 年。URL https://openreview.net/ forum?id=7evvwwdo3z.

[496] Terry Yue Zhuo, Vu Minh Chien, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen GONG, James Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/ forum?id=YrycTjllL0.

[496] Terry Yue Zhuo、Vu Minh Chien、Jenny Chim、Han Hu、Wenhao Yu、Ratnadira Widyasari、Imam Nur Bani Yusuf、Haolan Zhan、Junda He、Indraneil Paul、Simon Brunner、Chen GONG、James Hoang、Armel Randy Zebaze、Xiaoheng Hong、Wen-Ding Li、Jean Kaddour、Ming Xu、Zhihan Zhang、Prateek Yadav、Naman Jain、Alex Gu、Zhoujun Cheng、Jiawei Liu、Qian Liu、Zijian Wang、David Lo、Binyuan Hui、Niklas Muennighoff、Daniel Fried、Xiaoning Du、Harm de Vries 和 Leandro Von Werra。Bigcodebench: 使用多样函数调用与复杂指令对代码生成进行基准测试。载于第十三届国际学习表征会议，2025 年。URL https://openreview.net/ forum?id=YrycTjllL0.

[497] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

[497] Naman Jain、King Han、Alex Gu、Wen-Ding Li、Fanjia Yan、Tianjun Zhang、Sida Wang、Armando Solar-Lezama、Koushik Sen 和 Ion Stoica。Livecodebench: 对大型语言模型进行全面且无污染的代码评估。载于第十三届国际学习表征会议, 2025 年。URL https://openreview.net/forum?id=chfJJYC3iL.

[498] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/ forum?id=VTF8yNQM66.

[498] Carlos E Jimenez、John Yang、Alexander Wettig、Shunyu Yao、Kexin Pei、Ofir Press 和 Karthik R Narasimhan。SWE-bench: 语言模型能解决现实世界的 GitHub 问题吗？ 载于第十二届国际学习表征会议，2024 年。URL https://openreview.net/ forum?id=VTF8yNQM66.

[499] 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, 2025. URL https://arxiv.org/abs/2505.20411.

[499] Ibragim Badertdinov、Alexander Golubev、Maksim Nekrashevich、Anton Shevtsov、Simon Karasik、Andrei Andriushchenko、Maria Trofimova、Daria Litvintseva 和 Boris Yangel。Swe-rebench: 用于任务收集与去污染化软件工程代理评估的自动化流水线，2025 年。URL https://arxiv.org/abs/2505.20411.

[500] Bowen Li, Wenhan Wu, Ziwei Tang, Lin Shi, John Yang, Jinyang Li, Shunyu Yao, Chen Qian, Binyuan Hui, Qicheng Zhang, Zhiyin Yu, He Du, Ping Yang, Dahua Lin, Chao Peng, and Kai Chen. Prompting large language models to tackle the full software development lifecycle: A case study. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, Proceedings of the 31st International Conference on Computational Linguistics, pages 7511-7531, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.502/.

[500] Bowen Li、Wenhan Wu、Ziwei Tang、Lin Shi、John Yang、Jinyang Li、Shunyu Yao、Chen Qian、Binyuan Hui、Qicheng Zhang、Zhiyin Yu、He Du、Ping Yang、Dahua Lin、Chao Peng 和 Kai Chen。提示大型语言模型以处理完整软件开发生命周期: 一项案例研究。载于 Owen Rambow、Leo Wanner、Marianna Apidianaki、Hend Al-Khalifa、Barbara Di Eugenio 和 Steven Schockaert 主编的第 31 届计算语言学国际会议论文集，页 7511-7531，阿布扎比，阿联酋，2025 年 1 月。计算语言学协会。URL https://aclanthology.org/2025.coling-main.502/.

[501] Kaiyuan Liu, Youcheng Pan, Yang Xiang, Daojing He, Jing Li, Yexing Du, and Tianrun Gao. ProjectEval: A benchmark for programming agents automated evaluation on project-level code generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Findings of the Association for Computational Linguistics: ACL 2025, pages 20205-20221, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025. findings-acl.1036. URL https://aclanthology.org/2025.findings-acl.1036/.

[501] Kaiyuan Liu、Youcheng Pan、Yang Xiang、Daojing He、Jing Li、Yexing Du 和 Tianrun Gao。ProjectEval: 用于编程代理在项目级代码生成上自动评估的基准。载于 Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar 主编的 ACL 2025 研究成果集, 页 20205-20221, 维也纳, 奥地利, 2025 年 7 月。计算语言学协会。ISBN 979-8-89176-256-5。doi: 10.18653/v1/2025.findings-acl.1036。URL https://aclanthology.org/2025.findings-acl.1036/.

[502] Yiming Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lifu Huang, Xiao Liu, Jun Zhao, and Kang Liu. DA-code: Agent data science code generation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13487-13521, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. emnlp-main.748. URL https://aclanthology.org/2024.emnlp-main.748/.

[502] 黄一鸣, 罗建文, 余岩, 张怡彤, 雷方瑜, 魏一帆, 何士筑, 黄立夫, 刘潇, 赵峻, 和刘康. DA-code: 用于大型语言模型的代理数据科学代码生成基准. 收录于 Yaser Al-Onaizan, Mohit Bansal, 和 Yun-Nung Chen 主编, 2024 年经验方法在自然语言处理会议论文集, 页 13487-13521, 美国佛罗里达州迈阿密, 2024 年 11 月. 计算语言学协会. doi: 10.18653/v1/2024.emnlp-main.748. URL https://aclanthology.org/2024.emnlp-main.748/.

[503] Le Deng, Zhonghao Jiang, Jialun Cao, Michael Pradel, and Zhongxin Liu. Nocode-bench: A benchmark for evaluating natural language-driven feature addition, 2025. URL https://arxiv.org/abs/ 2507.18130.

[503] 邓乐, 江中昊, 曹家伦, Michael Pradel, 和刘中鑫. Nocode-bench: 用于评估自然语言驱动功能添加的基准, 2025. URL https://arxiv.org/abs/2507.18130.

[504] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. MLE-bench: Evaluating machine learning agents on machine learning engineering. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=655uXNWGIh.

[504] 陈俊深, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, 和 Lilian Weng. MLE-bench: 在机器学习工程上评估机器学习代理. 收录于第十三届国际表征学习会议, 2025. URL https://openreview.net/forum?id=655uXNWGIh.

[505] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating AI's ability to replicate AI research. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum? id=xF5PuTLPbn.

[505] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, 陈俊深, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, 和 Tejal Patwardhan. Paperbench: 评估 AI 复现 AI 研究的能力. 收录于第四十二届国际机器学习大会, 2025. URL https://openreview.net/forum?id=xF5PuTLPbn.

[506] Danijar Hafner. Benchmarking the spectrum of agent capabilities. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=1W0z96MFEoH.

[506] Danijar Hafner. 基准测试代理能力的全谱. 收录于国际表征学习会议, 2022. URL https://openreview.net/forum?id=1W0z96MFEoH.

[507] Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Thomas Jackson, Samuel Coward, and Jakob Nicolaus Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id

[507] Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Thomas Jackson, Samuel Coward, 和 Jakob Nicolaus Foerster. Craftax: 一个用于开放式强化学习的极快速基准. 收录于第四十一届国际机器学习大会, 2024. URL https://openreview.net/forum?id=hg4wXlrQCV.

[508] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. In ICML, pages 8657-8677, 2023. URL https://proceedings.mlr.press/v202/du23f.html.

[508] 杜昀庆, Olivia Watkins, 王子涵, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, 和 Jacob Andreas. 用大型语言模型指导强化学习的预训练. 收录于 ICML, 页 8657-8677, 2023. URL https://proceedings.mlr.press/v202/du23f.html.

[509] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, page 2186-2188, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

[509] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, 和 Shimon Whiteson. 星际争霸多智能体挑战赛. 收录于第 18 届自主代理与多智能体系统国际会议论文集, AAMAS '19, 页 2186-2188, Richland, SC, 2019. 国际自主代理与多智能体系统基金会. ISBN 9781450363099.

[510] Jack Hopkins, Mart Bakler, and Akbir Khan. Factorio learning environment, 2025. URL https://arxiv.org/abs/2503.09617.

[510] Jack Hopkins, Mart Bakler, 和 Akbir Khan. Factorio 学习环境, 2025. URL https://arxiv.org/abs/2503.09617.

[511] Yue Deng, Yan Yu, Weiyu Ma, Zirui Wang, Wenhui Zhu, Jian Zhao, and Yin Zhang. Smac-hard: Enabling mixed opponent strategy script and self-play on smac, 2024. URL https://arxiv.org/ abs/2412.17707.

[511] 邓悦, 余岩, 马伟宇, 王子锐, 朱文辉, 赵健, 和张寅. Smac-hard: 在 SMAC 上启用混合对手策略脚本和自我博弈, 2024. URL https://arxiv.org/abs/2412.17707.

[512] Weiyu Ma, Jiwen Jiang, Haobo Fu, and Haifeng Zhang. Tacticcraft: Natural language-driven tactical adaptation for starcraft ii, 2025. URL https://arxiv.org/abs/2507.15618.

[512] 马伟宇, 江继文, 傅浩博, 和张海峰. Tacticcraft: 用于星际争霸 II 的自然语言驱动战术适应, 2025. URL https://arxiv.org/abs/2507.15618.

[513] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

[513] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 评估在代码上训练的大型语言模型，2021。URL https://arxiv.org/abs/2107.03374.

[514] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

[514] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 使用大型语言模型进行程序合成，2021。URL https://arxiv.org/abs/2108.07732.

[515] Weiyu Ma, Dongyu Xu, Shu Lin, Haifeng Zhang, and Jun Wang. Adaptive command: Real-time policy adjustment via language models in starcraft ii, 2025. URL https://arxiv.org/abs/2508.16580.

[515] Weiyu Ma, Dongyu Xu, Shu Lin, Haifeng Zhang, and Jun Wang. 自适应指令: 通过语言模型在星际争霸 II 中进行实时策略调整，2025。URL https://arxiv.org/abs/2508.16580.

[516] William Brown. Verifiers: Reinforcement learning with llms in verifiable environments. https://github.com/willccbb/verifiers, 2025.

[516] William Brown. 验证器: 在可验证环境中用 LLMs 进行强化学习。https://github.com/willccbb/verifiers，2025。

[517] Shiyi Cao, Sumanth Hegde, Dacheng Li, Tyler Griggs, Shu Liu, Eric Tang, Jiayi Pan, Xingyao Wang, Akshay Malik, Graham Neubig, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Skyrl-v0: Train real-world long-horizon agents via reinforcement learning, 2025. URL https://novasky-ai.notion.site/skyrl-v0.

[517] Shiyi Cao, Sumanth Hegde, Dacheng Li, Tyler Griggs, Shu Liu, Eric Tang, Jiayi Pan, Xingyao Wang, Akshay Malik, Graham Neubig, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Skyrl-v0: 通过强化学习训练现实世界的长时程智能体, 2025。URL https://novasky-ai.notion.site/skyrl-v0.

[518] Tyler Griggs, Sumanth Hegde, Eric Tang, Shu Liu, Shiyi Cao, Dacheng Li, Charlie Ruan, Philipp Moritz, Kourosh Hakhamaneshi, Richard Liaw, Akshay Malik, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Evolving skyrl into a highly-modular rl framework, 2025. URL https://novasky-ai.notion.site/skyrl-v01.Notion Blog.

[518] Tyler Griggs, Sumanth Hegde, Eric Tang, Shu Liu, Shiyi Cao, Dacheng Li, Charlie Ruan, Philipp Moritz, Kourosh Hakhamaneshi, Richard Liaw, Akshay Malik, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 将 Skyrl 演进为高度模块化的 RL 框架，2025。URL https://novasky-ai.notion.site/skyrl-v01.Notion Blog.

[519] Wei Fu, Jiaxuan Gao, Shusheng Xu, Zhiyu Mei, Chen Zhu, Xujie Shen, Chuyi He, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. AREAL: A large-scale asynchronous reinforcement learning system for language reasoning. In ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models, 2025. URL https://openreview.net/forum?id=qJOokaW9Z9.

[519] Wei Fu, Jiaxuan Gao, Shusheng Xu, Zhiyu Mei, Chen Zhu, Xujie Shen, Chuyi He, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. AREAL: 用于语言推理的大规模异步强化学习系统。载于 ES-FoMo III: 第三届基础模型高效系统研讨会，2025。URL https://openreview.net/forum?id=qJOokaW9Z9.

[520] 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, 2025. URL https://github.com/TsinghuaC3I/MARTI.

[520] 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: 一个用于多智能体 LLM 系统的强化训练与推理框架，2025。URL https://github.com/TsinghuaC3I/MARTI.

[521] 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, 2025.

[521] 郑耀威, 陆俊廷, 王申志, 冯章驰, 郭东东, 和熊宇文. Easyr1: 一种高效、可扩展的多模态强化学习训练框架. https://github.com/hiyouga/EasyR1, 2025.

[522] Renxi Wang, Rifo Ahmad Genadi, Bilal El Bouardi, Yongxin Wang, Fajri Koto, Zhengzhong Liu, Timothy Baldwin, and Haonan Li. Agentfly: Extensible and scalable reinforcement learning for Im agents, 2025. URL https://arxiv.org/abs/2507.14897.

[522] 王仁熙, Rifo Ahmad Genadi, Bilal El Bouardi, 王永鑫, Fajri Koto, 刘正中, Timothy Baldwin, 和李浩南. Agentfly: 面向 Im 智能体的可扩展强化学习框架, 2025. URL https://arxiv.org/abs/2507.14897.

[523] 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, 2025. URL https://arxiv.org/abs/2508.03680.

[523] 罗旭方, 张玉革, 何志远, 王子龙, 赵思云, 李东升, Luna K. Qiu, 和杨玉清. Agent lightning: 使用强化学习训练任意 AI 智能体, 2025. URL https://arxiv.org/abs/2508.03680.

[524] RL-Factory. GitHub - Simple-Efficient/RL-Factory: Train your Agent model via our easy and efficient framework. https://github.com/Simple-Efficient/RL-Factory, 2025. [Accessed 03-09- 2025].

[524] RL-Factory. GitHub - Simple-Efficient/RL-Factory: 通过我们简洁高效的框架训练你的智能体模型. https://github.com/Simple-Efficient/RL-Factory, 2025. [访问 03-09-2025].

[525] 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, 2025. URL https: //arxiv.org/abs/2506.06122.

[525] 王伟勋, 熊少磐, 陈更儒, 高威, 郭晟, 何彦成, 黄炬, 刘家恒, 李振东, 李晓阳, 等. 面向大规模学习的强化学习优化: 一种高效且易用的扩展库. arXiv preprint arXiv:2506.06122, 2025. URL https://arxiv.org/abs/2506.06122.

[526] Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, Tianyu Pang, and Wenhu Chen. Verltool: Towards holistic agentic reinforcement learning with tool use, 2025. URL https://arxiv.org/abs/2509.01055.

[526] 江东福, 陆怡, 李卓峰, 吕志恒, 聂平, 王浩哲, Alex Su, 陈辉, 邹凯, 杜超, 庞天宇, 和陈文虎. Verltool: 走向具备工具使用能力的整体化智能体强化学习, 2025. URL https://arxiv.org/abs/2509.01055.

[527] Hanchen Zhang, Xiao Liu, Bowen Lv, Xueqiao Sun, Bohao Jing, Iat Long Iong, Zhenyu Hou, Zehan Qi, Hanyu Lai, Yifan Xu, Rui Lu, Hongning Wang, Jie Tang, and Yuxiao Dong. Agentrl: Scaling agentic reinforcement learning with a multi-turn, multi-task framework, 2025a. URL https://arxiv.org/abs/2510.04206.

[527] 张涵辰, 刘晓, 吕博文, 孙雪桥, 荆伯浩, Iat Long Iong, 侯振宇, 齐泽涵, 赖涵羽, 徐一凡, 鲁睿, 王宏宁, 唐杰, 和董宇晓. Agentrl: 使用多轮多任务框架扩展具身智能体的强化学习, 2025a. URL https://arxiv.org/abs/2510.04206.

[528] Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework, 2025. URL https://arxiv.org/ abs/2405.11143.

[528] 胡坚, 吴锡斌, 沈威, Jason Klein Liu, 朱子霖, 王伟勋, 蒋松林, 王浩然, 陈昊, 陈斌, 方威凯, Xianyu, 曹宇, 徐浩天, 和刘一鸣. Openrlhf: 一款易用、可扩展且高性能的 RLHF 框架, 2025. URL https://arxiv.org/abs/2405.11143.

[529] 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.

[529] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, 和 Quentin Gallouédec. Trl: Transformer 强化学习. https://github.com/huggingface/trl, 2020.

[530] Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. trlX: A framework for large scale reinforcement learning from human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8578-8595, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.530.URL https://aclanthology.org/2023.emnlp-main.530.

[530] Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, 和 Louis Castricato. trlX: 一个用于大规模从人类反馈中进行强化学习的框架. 收录于 2023 年实证方法在自然语言处理会议论文集, 页 8578-8595, 新加坡, 2023 年 12 月. 计算语言学协会. doi: 10.18653/v1/2023.emnlp-main.530. URL https://aclanthology.org/2023.emnlp-main.530.

[531] 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, EuroSys '25, page 1279-1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

[531] 盛光明, 张驰, 叶子凌锋, 吴锡斌, 张旺, 张茹, 彭阳华, 林海斌, 和吴传. Hybridflow: 一个灵活高效的 RLHF 框架. 收录于第十九届欧洲计算机系统会议论文集, EuroSys '25, 页 1279-1297. ACM, 2025 年 3 月. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

[532] THUDM. slime: A llm post-training framework for rl scaling. GitHub repository, https://github.com/THUDM/slime, 2025. Accessed: 2025-08-13.

[532] THUDM. slime: 一个用于 RL 扩展的 llm 后训练框架。GitHub 仓库, https://github.com/THUDM/slime, 2025. 访问时间: 2025-08-13.

[533] Zichen Liu, Changyu Chen, Chao Du, Wee Sun Lee, and Min Lin. Sample-efficient alignment for llms, 2024. URL https://arxiv.org/abs/2411.01493.

[533] Zichen Liu, Changyu Chen, Chao Du, Wee Sun Lee, and Min Lin. 针对 llms 的样本高效对齐, 2024. URL https://arxiv.org/abs/2411.01493.

[534] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. RLlib: Abstractions for distributed reinforcement learning. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 3053-3062. PMLR, 10-15 Jul 2018. URL https://proceedings.mlr.press/v80/liang18b.html.

[534] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. RLlib: 面向分布式强化学习的抽象。在 Jennifer Dy 和 Andreas Krause 编辑, 第 35 届国际机器学习会议论文集, 机器学习研究论文集第 80 卷, 页 3053-3062. PMLR, 2018 年 7 月 10-15 日. URL https://proceedings.mlr.press/v80/liang18b.html.

[535] Matthew W Hoffman, Bobak Shahriari, John Aslanides, Gabriel Barth-Maron, Nikola Momchev, Danila Sinopalnikov, Piotr Stańczyk, Sabela Ramos, Anton Raichuk, Damien Vincent, et al. Acme: A research framework for distributed reinforcement learning. arXiv preprint arXiv:2006.00979, 2020.

[535] Matthew W Hoffman, Bobak Shahriari, John Aslanides, Gabriel Barth-Maron, Nikola Momchev, Danila Sinopalnikov, Piotr Stańczyk, Sabela Ramos, Anton Raichuk, Damien Vincent, 等. Acme: 一个用于分布式强化学习研究的框架. arXiv 预印本 arXiv:2006.00979, 2020.

[536] Jiayi Weng, Huayu Chen, Dong Yan, Kaichao You, Alexis Duburcq, Minghao Zhang, Yi Su, Hang Su, and Jun Zhu. Tianshou: A highly modularized deep reinforcement learning library. Journal of Machine Learning Research, 23(267):1-6, 2022. URL http://jmlr.org/papers/v23/21-1127.html.

[536] Jiayi Weng, Huayu Chen, Dong Yan, Kaichao You, Alexis Duburcq, Minghao Zhang, Yi Su, Hang Su, and Jun Zhu. Tianshou: 一个高度模块化的深度强化学习库. 机器学习研究期刊, 23(267):1-6, 2022. URL http://jmlr.org/papers/v23/21-1127.html.

[537] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research, 22(268):1-8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

[537] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: 可靠的强化学习实现. 机器学习研究期刊, 22(268):1-8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

[538] Yasuhiro Fujita, Prabhat Nagarajan, Toshiki Kataoka, and Takahiro Ishikawa. Chainerrl: A deep reinforcement learning library. Journal of Machine Learning Research, 22(77):1-14, 2021. URL http://jmlr.org/papers/v22/20-376.html.

[538] Yasuhiro Fujita, Prabhat Nagarajan, Toshiki Kataoka, and Takahiro Ishikawa. Chainerrl: 一个深度强化学习库. 机器学习研究期刊, 22(77):1-14, 2021. URL http://jmlr.org/papers/v22/20-376.html.

[539] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. arXiv preprint arXiv:2502.11127, 2025.

[539] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-safeguard: 基于拓扑引导的安全视角与对 llm 驱动多智能体系统的处理. arXiv 预印本 arXiv:2502.11127, 2025.

[540] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases, 2024. URL https://arxiv.org/abs/2407.12784.

[540] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: 通过污染记忆或知识库对 llm 代理进行红队测试, 2024. URL https://arxiv.org/abs/2407.12784.

[541] Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities, 2025. URL https://arxiv.org/ abs/2408.04682.

[541] Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: 一个有状态的、对话式的、交互式评估基准，用于 llm 工具使用能力, 2025. URL https://arxiv.org/ abs/2408.04682.

[542] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated

sandbox. In The Twelfth International Conference on Learning Representations, 2024. URL https: //openreview.net/forum?id=GEcwtMk1uA.

[542] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 使用 LM 模拟的沙箱识别 LM 代理的风险. 在第十二届国际学习表征会议, 2024. URL https: //openreview.net/forum?id=GEcwtMk1uA.

[543] Manuel Cossio. A comprehensive taxonomy of hallucinations in large language models, 2025. URL https://arxiv.org/abs/2508.01781.

[543] Manuel Cossio。大型语言模型幻觉的全面分类学，2025。URL https://arxiv.org/abs/2508.01781.

[544] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1-55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[544] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, 和 Ting Liu。大型语言模型幻觉综述: 原理、分类、挑战与未解问题。ACM Transactions on Information Systems, 43(2):1-55, 2025 年 1 月。ISSN 1558-2868。doi: 10.1145/3703155。URL http://dx.doi.org/10.1145/3703155.

[545] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2025. URL https://arxiv.org/ abs/2404.18930.

[545] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, 和 Mike Zheng Shou。多模态大型语言模型的幻觉: 综述，2025。URL https://arxiv.org/ abs/2404.18930.

[546] Junyi Li and Hwee Tou Ng. The hallucination dilemma: Factuality-aware reinforcement learning for large reasoning models, 2025. URL https://arxiv.org/abs/2505.24630.

[546] Junyi Li 和 Hwee Tou Ng。幻觉困境: 面向大型推理模型的事实性感知强化学习，2025。URL https://arxiv.org/abs/2505.24630.

[547] Linxin Song, Taiwei Shi, and Jieyu Zhao. The hallucination tax of reinforcement finetuning, 2025. URL https://arxiv.org/abs/2505.13988.

[547] Linxin Song, Taiwei Shi, 和 Jieyu Zhao。强化微调的幻觉代价，2025。URL https://arxiv.org/abs/2505.13988.

[548] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination?, 2025. URL https://arxiv.org/abs/ 2505.23646.

[548] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, 和 Tat-Seng Chua。推理模型是否更易产生幻觉？, 2025。URL https://arxiv.org/abs/ 2505.23646.

[549] Yuan Sun and Ting Wang. Be friendly, not friends: How llm sycophancy shapes user trust, 2025. URL https://arxiv.org/abs/2502.10844.

[549] Yuan Sun 和 Ting Wang。要友好，不要谄媚:LLM 奉承如何塑造用户信任，2025。URL https://arxiv.org/abs/2502.10844.

[550] Lars Malmqvist. Sycophancy in large language models: Causes and mitigations, 2024. URL https://arxiv.org/abs/2411.15287.

[550] Lars Malmqvist。大型语言模型中的谄媚: 成因与缓解，2024。URL https://arxiv.org/abs/2411.15287.

[551] Taiming Lu, Lingfeng Shen, Xinyu Yang, Weiting Tan, Beidi Chen, and Huaxiu Yao. It takes two: On the seamlessness between reward and policy model in rlhf, 2024. URL https://arxiv.org/abs/ 2406.07971.

[551] Taiming Lu, Lingfeng Shen, Xinyu Yang, Weiting Tan, Beidi Chen, 和 Huaxiu Yao。二者兼备:RLHF 中奖励模型与策略模型之间的无缝性，2024。URL https://arxiv.org/abs/ 2406.07971.

[552] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf, 2024. URL https://arxiv.org/abs/2409.12822

[552] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, 和 Shi Feng。语言模型通过 RLHF 学会误导人类，2024。URL https://arxiv.org/abs/2409.12822.

[553] Priya Pitre, Naren Ramakrishnan, and Xuan Wang. Consensagent: Towards efficient and effective consensus in multi-agent llm interactions through sycophancy mitigation. In Findings of the Association for Computational Linguistics: ACL 2025, pages 22112-22133, 2025.

[553] Priya Pitre, Naren Ramakrishnan, 和 Xuan Wang。Consensagent: 通过缓解谄媚以实现多智能体 LLM 交互中高效且有效的共识。在 Findings of the Association for Computational Linguistics: ACL 2025, 页 22112-22133, 2025。

[554] 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, 2025. URL https://arxiv.org/abs/2508.05613.

[554] Haitao Hong, Yuchen Yan, Xingyu Wu, Guiyang Hou, Wenqi Zhang, Weiming Lu, Yongliang Shen, 和 Jun Xiao。Cooper: 在大型语言模型强化学习中联合优化策略与奖励模型，2025。URL https://arxiv.org/abs/2508.05613.

[555] Jian Hu, Mingjie Liu, Shizhe Diao, Ximing Lu, Xin Dong, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. ProRL V2 - Prolonged Training Validates RL Scaling Laws. https://hijkzzz.notion.site/pror1-v2, 2025. Notion page. First published: August 11, 2025. Accessed: August 15, 2025.

[555] Jian Hu, Mingjie Liu, Shizhe Diao, Ximing Lu, Xin Dong, Pavlo Molchanov, Yejin Choi, Jan Kautz, 和 Yi Dong。ProRL V2 - 延长训练验证 RL 缩放定律。https://hijkzzz.notion.site/pror1-v2, 2025。Notion 页面。首次发布:2025 年 8 月 11 日。访问:2025 年 8 月 15 日。

[556] Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, and Ge Li. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization, 2025. URL https://arxiv.org/abs/2508.00222.

[556] Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, 和 Ge Li。RL-Plus: 通过混合策略优化应对强化学习中 LLM 能力边界崩塌，2025。URL https://arxiv.org/abs/2508.00222.

[557] Yu Li, Zhuoshi Pan, Honglin Lin, Mengyuan Sun, Conghui He, and Lijun Wu. Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning, 2025. URL https://arxiv.org/abs/2507.1751

[557] Yu Li, Zhuoshi Pan, Honglin Lin, Mengyuan Sun, Conghui He, and Lijun Wu. 一个领域能帮助其他领域吗?基于数据的多领域推理通过强化学习研究, 2025。URL https://arxiv.org/abs/2507.17512.

[558] Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Haonan Li, Taylor W. Killian, Mikhail Yurochkin, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective, 2025. URL https://arxiv.org/abs/2506.14965.

[558] Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Haonan Li, Taylor W. Killian, Mikhail Yurochkin, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 从跨领域视角重审用于 LLM 推理的强化学习，2025。URL https://arxiv.org/abs/2506.14965.

[559] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, Tianming Liu, Neil Zhenqiang Gong, Jiliang Tang, Caiming Xiong, Heng Ji, Philip S. Yu, and Jianfeng Gao. A survey on post-training of large language models, 2025. URL https://arxiv.org/abs/2503.06072.

[559] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, Tianming Liu, Neil Zhenqiang Gong, Jiliang Tang, Caiming Xiong, Heng Ji, Philip S. Yu, and Jianfeng Gao. 大型语言模型后训练综述，2025。URL https://arxiv.org/abs/2503.06072.

[560] 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.

[560] Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. Polaris: 用于在高级推理模型上扩展强化学习的后训练方案，2025。URL https://hkunlp.github.io/blog/2025/ Polaris.

[561] 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, 2025. URL https://arxiv.org/abs/2508.05629.

[561] Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 关于 SFT 的泛化: 带有奖励修正的强化学习视角，2025。URL https://arxiv.org/abs/2508.05629.

[562] Zihan Zheng, Tianle Cui, Chuwen Xie, Jiahui Zhang, Jiahui Pan, Lewei He, and Qianglong Chen. Naturegaia: Pushing the frontiers of gui agents with a challenging benchmark and high-quality trajectory dataset, 2025. URL https://arxiv.org/abs/2508.01330.

[562] Zihan Zheng, Tianle Cui, Chuwen Xie, Jiahui Zhang, Jiahui Pan, Lewei He, and Qianglong Chen. Naturegaia: 用具有挑战性的基准和高质量轨迹数据集推动 GUI 代理的前沿，2025。URL https://arxiv.org/abs/2508.01330.

[563] Abhay Zala, Jaemin Cho, Han Lin, Jaehong Yoon, and Mohit Bansal. Envgen: Generating and adapting environments via LLMs for training embodied agents. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=F9tqgOPXH5.

[563] Abhay Zala, Jaemin Cho, Han Lin, Jaehong Yoon, and Mohit Bansal. EnvGen: 通过 LLM 生成和适配环境以训练具身代理。载于第一届语言建模大会，2024。URL https://openreview.net/forum?id=F9tqgOPXH5.

[564] Essential AI, :, Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Anthony Polloreno, Ashish Tanwer, Burhan Drak Sibai, Divya S Mansingka, Divya Shivaprasad, Ishaan Shah, Karl Stratos, Khoi Nguyen, Michael Callahan, Michael Pust, Mrinal Iyer, Philip Monk, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, and Tim Romanski. Rethinking reflection in pre-training, 2025. URL https://arxiv.org/abs/2504.04022.

[564] Essential AI, :, Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Anthony Polloreno, Ashish Tanwer, Burhan Drak Sibai, Divya S Mansingka, Divya Shivaprasad, Ishaan Shah, Karl Stratos, Khoi Nguyen, Michael Callahan, Michael Pust, Mrinal Iyer, Philip Monk, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, and Tim Romanski. 重新思考预训练中的反思，2025。URL https://arxiv.org/abs/2504.04022.

[565] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature, 645(8081):633-638, 2025.

[565] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 通过强化学习激励 LLM 中的推理。Nature, 645(8081):633-638, 2025.

[566] Nikolaos Tsilivis, Eran Malach, Karen Ullrich, and Julia Kempe. How reinforcement learning after next-token prediction facilitates learning, 2025. URL https://arxiv.org/abs/2510.11495.

[566] Nikolaos Tsilivis, Eran Malach, Karen Ullrich, and Julia Kempe. 在下一个标记预测之后的强化学习如何促进学习，2025。URL https://arxiv.org/abs/2510.11495.

[567] 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, 2025. URL https://arxiv.org/abs/2501.17161.

[567] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft 记忆化，rl 泛化化: 基础模型后训练的比较研究，2025。URL https://arxiv.org/abs/2501.17161.

[568] 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, 2025. URL https://arxiv.org/abs/2503.01307.

[568] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 使自我改进推理者成为可能的认知行为，或，高效明星的四个习惯，2025。URL https://arxiv.org/abs/2503.01307.