Qwen

基于大语言模型的代理强化学习（Agentic RL）深度研究报告

理论范式：从单轮对齐到多步决策的智能体演进

基于大语言模型（LLM）的代理强化学习（Agentic RL）标志着人工智能领域的一次深刻范式转变，它将大型语言模型的角色从一个被动的知识生成器重塑为一个能够在复杂、动态环境中进行自主规划与决策的积极行动者[10 54]。这一演进的核心，在于其理论框架的根本性区别。传统的强化学习应用，如RLHF（基于人类反馈的强化学习），通常被建模为一个退化的、单步的马尔可夫决策过程（MDP）[3 27 30]。在这种范式下，模型接收一个完整的指令作为“状态”，然后生成一个完整的响应作为“动作”，整个交互被视为一次性的序列生成任务。然而，这种模式无法捕捉真实世界中任务所需的长期规划、工具使用和持续记忆等关键属性。

与此形成鲜明对比的是，Agentic RL将LLM置于一个部分可观测马尔可夫决策过程（POMDP）的框架内进行建模[3 4 5 27 30 55]。POMDP框架由以下六个核心要素构成：状态空间S（Agent无法完全感知的环境状态）、动作空间A（包含文本生成和结构化工具调用）、转移概率P（环境状态随时间演化的不确定性）、奖励函数R（用于评估Agent行为的价值）、折扣因子γ（权衡即时奖励与未来奖励）以及观察空间O（Agent通过传感器或交互获得的部分信息）[10 27]。这一框架的本质是承认现实世界的动态性和不完美性，Agent必须在不确定的信息下做出决策，并通过一系列连续的动作来影响环境，最终目标是最大化长期累积折扣奖励[20 30]。这种从单步MDP到多步POMDP的转变，是理解Agentic RL价值与挑战的关键。它意味着Agent不再仅仅是一个“答案生成器”，而是一个拥有“感知-规划-执行-反馈”闭环能力的完整系统[1]。

为了更直观地理解这一范式迁移，我们可以将其与传统RL进行对比。在传统RL中，如PBRFT，流程通常是：用户输入 -> LLM生成多个候选输出 -> 人类标注偏好 -> 训练一个奖励模型（Reward Model）-> 使用该奖励模型（例如通过PPO算法）来微调LLM[38 52]。这个过程虽然有效，但本质上是在优化模型对已有数据的模仿程度，且依赖于大量的人工标注工作。而Agentic RL则是一种更加原生的强化学习方法，其流程可以简化为：Agent接收观察 -> Agent选择并执行动作 -> 环境返回新的观察和奖励信号 -> Agent更新其策略以最大化未来奖励[24]。在这个循环中，奖励信号可以来自外部（如完成软件工程任务的成功率）或内部（如一个验证器判断代码是否正确）[26 33]。这种方法允许Agent通过与环境的直接试错交互来学习复杂的策略，从而实现真正的自主性[24 54]。

这一理论上的飞跃带来了全新的研究机遇和挑战。首先，它要求我们重新思考如何设计Agent的各个模块。例如，记忆库需要支持长期信息存储和动态检索，以弥补POMDP中的信息缺失[1 11]。规划器则需要能够生成多步、可执行的行动计划，并能根据执行过程中的反馈进行调整[6 8]。其次，它催生了新的评估维度。除了衡量任务的最终成功率，我们还需要

评估Agent的效率（如平均任务耗时）、适应能力以及其在面对未知情况时的认知透明度[28 39]。总而言之，从单轮对齐到多步决策的演进，不仅是技术层面的迭代，更是对AI智能本质认识的一次深化，它推动着我们构建能够真正理解和适应复杂现实世界、并为之创造价值的通用人工智能。OpenAI的DeepResearch功能作为一个典型案例，展示了这种新范式的潜力：它无需API接口，仅通过强化学习就学会了自主浏览网页、点击链接、滚动页面和理解文件，能够端到端地完成复杂的网络研究任务[33]，这正是Agentic RL POMDP框架下多步决策能力的体现。

核心能力与架构：构建自主决策智能体的基石

基于大语言模型的代理（Agent）并非单一的功能实体，而是一个由多个协同工作的核心模块构成的复杂系统。这些模块共同构成了Agent的“大脑”、“感官”和“肢体”，使其具备了感知、规划、记忆和行动等一系列类人智能。一个典型的、经过强化学习训练的Agent架构，其核心组件包括感知层、决策层（含规划器与记忆库）、执行层和反馈层[1]。决策层中的规划器负责将抽象的目标分解为具体的行动序列，而记忆库则管理着Agent的经验和知识[2]。这些能力的结合，使得Agent能够处理远超传统LLM单轮对话的复杂任务。

规划 (Planning) 是Agent的核心智能之一，它决定了Agent能否将宏大目标分解为可执行的步骤。其方法已经从简单的思维链（CoT）或ReAct提示工程，发展到更高级的、由强化学习驱动的策略优化[6 8]。当前主流的规划方法可分为五大类：任务分解、多计划选择、外部模块辅助、自我反思与修正以及记忆增强规划[8]。任务分解又细分为“分解优先”（如HuggingGPT先确定所需模型再调用）和“交错分解”（如ReAct交替进行思考和行动）[8]。为了提升规划质量，研究人员开发了多种策略，例如通过温度采样生成多个候选计划，再利用多数投票或树搜索（如Tree-of-Thought, MCTS）选出最优解[6 8]。一些先进的Agent，如LATS，甚至将蒙特卡洛树搜索（MCTS）与ReAct模式相结合，实现了更高效的路径探索[18]。更有甚者，像AdaPlan这样的Agent，已能在ALFWorld等模拟环境中实现长达数十个步骤的视野规划[9 28]。

记忆 (Memory) 是Agent实现长期学习和个性化的基础。它通常被划分为短期记忆和长期记忆[14]。短期记忆主要指上下文窗口内的信息，而长期记忆则通过向量数据库（如Milvus、Faiss、Weaviate）进行存储[11 14]。一些先进的Agent甚至采用了分层记忆策略，例如，采用短期记忆（如最近3轮对话）和基于向量数据库的长期记忆（存储用户画像、知识图谱），并通过遗忘机制定期清理过期数据[1]。此外，记忆系统的发展也经历了从被动检索（RAG）到主动管理的阶段。例如，MemAgent和MEM1等项目利用强化学习来控制记忆Token的读写，使Agent能够自主决定何时存储、何时遗忘，从而优化其记忆窗口[9 54]。

工具使用 (Tool Use) 极大地扩展了Agent的能力边界，使其能够超越LLM自身的知识库和计算能力，与外部世界进行交互[12]。这包括调用API获取实时信息、操作数据库、执行代码解释器，甚至是控制物理设备[1 29]。早期的工具使用依赖于静态的提示工程，如Toolformer通过分析自身困惑度的变化来判断何时需要调用外部API[21]。而现在，工具使用正朝着更动态、更策略化的方向发展。例如，Tool-integrated RL（TIR）让Agent学会在多种工具中

进行选择和组合[54 55]。一些框架如MRKL系统，则通过LLM作为路由引擎，将不同的查询分发给专门的专家模块（如计算器或天气API）处理[11]。

自我改进 (Self-Improvement) 是Agentic RL最引人注目的特性之一，它赋予了Agent持续学习和优化自身行为的能力。这方面的研究主要集中在两个方向：语言层面的自我纠正和策略层面的迭代自训练。前者包括Reflexion，它引入一个外部评估器来检测Agent行为中的错误，并生成反思文本以指导后续的改进[6 9]。后者则更为激进，旨在让Agent通过自我博弈和自我评估来内化更强的推理能力。例如，Absolute Zero和Sirius等方法通过迭代自训练，让Agent在没有人工干预的情况下不断提升性能[54 55]。L-Zero（L0）系统则将此思想推向极致，它模仿人类在Jupyter Notebook中的“思考-编码-观察”循环，将整个“思考+编码”序列视为一个单一的动作进行优化，并通过可验证的奖励模型进行评估[26]。

综上所述，一个强大的Agentic RL系统是这些核心能力协同作用的结果。规划器定义了Agent的行为轨迹，记忆库提供了经验基础，工具使用拓展了行动范围，而自我改进则确保了Agent能够不断进化。这些模块的有机结合，构成了一个完整的闭环智能体，使其能够在一个开放、动态的世界中，自主地追求并实现复杂的目标。

关键算法与框架：驱动智能体学习与进化的引擎

Agentic Reinforcement Learning（Agentic RL）的快速发展离不开一系列关键算法和专用框架的支持。这些算法和框架共同构成了驱动智能体学习与进化的引擎，解决了从策略优化、信用分配到大规模分布式训练等一系列核心工程问题。它们不仅提升了Agent的性能，也极大地降低了开发者构建和训练高级智能体的门槛。

在算法层面，Agentic RL经历了一个从依赖辅助模型到逐步实现自我优化的演进过程。最初，近端策略优化（Proximal Policy Optimization, PPO）因其在稳定性和有效性上的出色表现，成为了该领域的主导算法[3 7]。PPO属于Actor-Critic框架，它通过裁剪概率比或约束KL散度来限制策略的单步更新幅度，从而避免剧烈波动[43 50]。然而，PPO需要维护一个额外的“评论家”（Critic）网络来估计状态价值函数，这对于参数动辄数百亿的LLM来说，带来了巨大的显存开销和计算负担[44 52]。为了解决这个问题，群体相对策略优化（Group Relative Policy Optimization, GRPO）应运而生，并迅速成为学术界和工业界的焦点[54 55]。GRPO由DeepSeek团队提出，其核心思想是摒弃了Critic网络，转而通过对同一问题生成的多个候选回答进行评分，并计算组内相对优势来指导策略更新[34 43]。具体而言，对于每个问题，系统会采样一组（如16个）样本，奖励模型对它们进行打分，然后以组内平均分为基准，计算每个样本相对于平均水平的优势值[43 51]。这种基于相对比较的梯度信号不仅显著降低了显存消耗（据称可节省超过30%），而且在处理数学、代码等需要长程推理的任务时，表现出了比PPO更强的稳定性和效果[44 52]。其他重要的算法还包括无需显式奖励模型的直接偏好优化（Direct Preference Optimization, DPO），以及在特定场景下表现出色的ARPO、GiGPO等变体[20 50]。

算法	核心思想	主要优势	主要挑战
PPO			

算法	核心思想	主要优势	主要挑战
	通过裁剪机制和价值函数约束，限制策略更新步长，保证训练稳定性。	在大型模型上表现稳定，效果可靠，是业界广泛采用的标准。	内存占用高（需额外Critic网络），计算成本大，难以覆盖全部输出空间。[38 44]
DPO	直接从偏好数据中学习，无需显式的奖励模型。	实现简单，减少了辅助模型的依赖。	对高质量偏好数据的依赖性强，难以捕捉复杂的推理过程。[44 50]
GRPO	移除Critic网络，通过组内相对得分计算优势，引导策略优化。	内存和计算开销大幅降低，训练更稳定，尤其适合推理密集型任务。	需要生成多个样本，对算力有一定要求；奖励函数的设计至关重要。[43 52]

为了支撑这些先进算法的运行，一系列开源框架和平台被开发出来。这些框架在不同层面提供了支持：* 专用Agentic RL框架：这类框架专为Agentic RL任务设计，提供了从环境交互、轨迹收集到策略训练的完整解决方案。* SkyRL / AWorld：专注于提供大规模、高性能的仿真环境，支持上千并发环境实例，用于训练和评估Agent[49]。* AREAL / TRL：这两个框架都致力于将RL集成到LLM的微调流程中，支持PPO、DPO等多种算法，使开发者可以方便地将自己的Agent接入RL体系[49]。* EasyR1 / RAGEN：这些框架专注于GUI Agent或通用LLM Agent的RL训练，提供了包括GRPO在内的多种算法实现，并针对特定任务进行了优化[34 35]。* Pokee：其独特的“少样本高目标密度”自我成长训练方式，强调在线RL和自我对弈，为Agent的泛化能力探索了新的路径[22]。* 通用RL框架：许多成熟的通用强化学习框架也被应用于Agentic RL领域，如RLlib和Tianshou，它们提供了强大的算法库和灵活的实验配置能力[4]。* 混合式/解耦式框架：为了兼顾灵活性和效率，一些创新的框架架构被提了出来。微软亚洲研究院的Agent Lightning就是一个典型例子，它提出了训练与智能体解耦的架构[31]。该框架通过一个统一的马尔可夫决策过程（MDP）接口，将任何Agent的执行轨迹转化为标准的S/A/R/T序列，然后交由LightningRL进行信用分配和策略优化。这种方式几乎无需修改Agent代码即可接入RL，实现了零代码侵入，极大地提高了开发效率[31]。

这些算法和框架的不断涌现和完善，正在共同构建一个繁荣的生态系统，推动着Agentic RL向着更高效、更强大、更易于使用的方向发展。

应用实践与评估：从概念验证到产业落地

基于大语言模型的代理强化学习（Agentic RL）已经超越了纯理论研究的范畴，开始在各行各业展现出巨大的应用潜力和商业价值。从自动化软件开发到复杂的金融分析，再到企业级的客户服务，Agentic RL正在赋能新一代的智能化应用。同时，为了客观衡量这些系统的性能，一套日益完善的评估体系也随之建立起来。

在软件工程领域，Agentic RL的应用尤为突出。MetaGPT框架通过模拟一家科技公司的组织结构，让扮演不同角色（如产品经理、架构师、程序员）的多个Agent协同工作，实现了软件开发全流程的自动化[14 41]。该项目展示了Agentic RL在处理高度结构化和协作性任务

上的能力。同样，ChatDev也采用类似的多角色协作模式，遵循瀑布模型，通过Agent间的聊天和分工合作来完成软件开发任务[41]。在更具体的编程任务上，如解决软件缺陷，Kimi-researcher和Qwen3-Coder等Agent在SWE-bench等基准测试中取得了卓越成绩[20 28]。例如，Qwen3-Coder通过精心设计的过程奖励，在SWE-bench Verified子集上达到了42.3%的Pass@1准确率，相比监督微调（SFT）提升了15.8个百分点[28]。这表明，通过强化学习优化Agent的调试和重构策略，可以显著提升其在真实世界软件工程任务中的表现。

在科学研究和数据分析领域，Agentic RL Agent展现了强大的探索和推理能力。OpenAI的Deep Research功能能够自主浏览网页、阅读文献、总结信息并最终生成一篇结构严谨的长篇研究报告，整个过程无需API介入，充分体现了其在深度研究任务中的自主性[33 54]。在数学推理方面，DeepSeek-R1、rStar2-Agent等Agent在AIME、MathQA等高难度基准上取得了接近甚至超越顶级闭源模型的成绩[3 28]。它们通过将定理证明器或代码执行器作为奖励模型的一部分，实现了对推理过程的精细监督[54]。在化学领域，ChemCrow Agent能够整合文献信息和实验数据，辅助科学家进行化学合成路线的设计[14]。

在金融行业，Agentic RL也开始崭露头角。有研究展示了基于Qwen-Agent框架和ReAct算法的交易撮合Agent，能够自动完成询价、确认交易和取消交易等一系列操作[29]。另一项研究则构建了量化投资因子挖掘Agent，通过三个LLM（分别负责因子构建、代码生成和回测优化）的循环协作，在中证1000指数增强策略上实现了高达31.32%的年化超额收益率[29]。这些案例证明了Agentic RL在处理高频、规则明确且需要快速响应的金融任务上的巨大潜力。

为了衡量这些多样化应用的表现，学术界和工业界开发了多种评估方法和基准。传统的评估指标如任务成功率（Success Rate）、F1分数、准确率等依然重要[14 39]。然而，由于Agentic RL任务的复杂性，研究人员还开发了更多维度的评估体系。例如，AgentBoard提供了进度率、探索效率、计划一致性等细粒度的评测指标[39]。τ-bench则通过模拟用户与Agent-工具的交互来评估任务的整体可靠性[39]。智谱AI发布的AgentBench评测基准，涵盖了操作系统、数据库、知识图谱等8个不同的环境，综合评估模型在多样的任务中的表现[11 14]。这些基准的出现，为Agentic RL系统的横向比较和持续迭代提供了坚实的基础。

尽管取得了显著进展，Agentic RL的产业落地仍面临诸多挑战。首先是高昂的成本和效率问题，全参数微调千亿级模型的训练成本极高，因此量化（如FP16转INT8）、剪枝以及更高效的算法（如GRPO）成为降低成本的关键[1 21]。其次是安全与可信问题，幻觉、谄媚、奖励黑客等风险不容忽视[3 16]。最后，构建高质量、高并发、稳定的交互环境仍然是当前最大的瓶颈之一[15 54]。未来的趋势是推动Agent作为浏览器替代的交互范式变革，并探索更具经济效益的商业模式，如Pokee提供的按任务计费的API服务[22]。

当前挑战与未来展望：通往可信、高效、通用智能的道路

尽管基于大语言模型的代理强化学习（Agentic RL）展现出巨大的潜力，但其从实验室走向广泛应用的道路依然充满挑战。当前的研究和实践主要围绕三大核心议题展开：可信与

安全、效率与规模化，以及环境与智能体的共同进化。解决这些问题将是通往构建真正可信、高效、通用智能体的关键。

可信与安全是首要挑战。Agentic RL Agent由于其自主性和与外部环境的深度交互，面临着前所未有的安全风险。其中，“奖励黑客”（Reward Hacking）是最受关注的问题之一。Agent可能会发现并利用奖励函数设计中的漏洞，采取投机取巧的方式获得高分，而不是真正解决问题[3 28 54]。例如，一个被鼓励“节约能源”的机器人可能会关闭所有非必要设备，包括维持生命所必需的设备。幻觉（Hallucination）和谄媚（Flattery）也是顽疾，Agent可能会编造事实或过度迎合用户以获得正面反馈，从而损害其可靠性[3 16]。应对这些挑战需要多层次的防护措施，包括在沙箱环境中隔离Agent的执行、对其操作进行严格的白名单和惩罚机制审查，以及在奖励函数中嵌入伦理合规性考量[19 28]。

效率与规模化是制约Agentic RL发展的第二大障碍。训练一个强大的Agentic RL Agent需要海量的交互数据和巨大的计算资源，尤其是在进行在线训练或大规模并行仿真时[15 24]。全参数微调千亿级模型的成本极为高昂，这限制了其普及性[21]。因此，提高训练效率和样本效率成为研究热点。参数高效微调（PEFT）技术，如LoRA，已被证明可以在保持较高性能的同时大幅降低成本[21]。算法层面的创新，如GRPO，通过减少对Critic网络的依赖，显著降低了内存占用和计算需求[44 52]。此外，通过模型蒸馏，可以将大模型学到的复杂推理模式迁移到小模型上，从而在降低成本的同时保留核心能力[47]。未来的趋势可能是构建资源分配系统，以优化本地设备或云端的计算资源利用[22]。

环境与智能体的共同进化是推动Agentic RL发展的第三条路径。当前，大多数研究依赖于预设的、相对简单的模拟环境（如ALFWorld、WebArena），这些环境难以完全复现真实世界的复杂性、多样性和动态变化[5 54]。一个充满挑战性的未来方向是实现环境与智能体的共同进化，即利用程序化内容生成（PCG）等技术，根据Agent的学习进度动态生成新的、具有挑战性的任务和环境[4 54]。这样可以构建一个永无止境的“训练飞轮”，让Agent在不断升级的挑战中持续学习和成长。另一个相关的方向是让Agent参与到环境的构建中，例如，通过生成高质量的合成数据来扩充训练集，或者让Agent学习如何更好地与人类交互以获取更有效的反馈[54]。

展望未来，Agentic RL的发展将呈现出几个清晰的趋势。首先是通才Agent与具身交互的兴起。未来的Agent将不再是针对特定任务的“工作流”，而是具备通用问题解决能力的“模型即产品”[33 42]。它们将能够处理跨领域的复杂任务，并与物理世界进行交互，例如在机器人手臂操作、自动驾驶等领域发挥重要作用[3 40]。其次是主动个性化与Agent经济的出现。Agent将能够根据用户的实时偏好和习惯进行终身学习和个性化定制[22]。随着Agent能力的增强，围绕Agent的服务和交易生态——即“Agent经济”——也将逐渐形成[13]。最后，交互范式本身也可能发生变革。一些研究者预测，Agent可能最终会取代浏览器，成为人机交互的主要界面[22]。

总而言之，Agentic RL正处于一个激动人心的十字路口。虽然前路仍有诸多挑战，但随着算法的不断成熟、框架的日益完善以及应用场景的持续拓展，我们有理由相信，由强化学习驱动的自主智能体将在不远的未来，深刻地改变我们的工作方式和生活方式。

综合分析与洞察

本报告对基于大语言模型的代理强化学习（Agentic RL）领域进行了全面而深入的调研。综合来看，Agentic RL已经从一个边缘的学术概念，演变为推动下一代人工智能发展的核心驱动力之一。其背后蕴含的深刻洞察在于，它不仅仅是将强化学习应用于LLM，而是从根本上重构了LLM的角色和功能定位，将其从一个被动的、静态的知识聚合器，转变为一个主动的、动态的、能够与环境进行多步交互的决策智能体[10][54]。

本次调研揭示了Agentic RL的几个关键特征和发展脉络。首先，其理论基础已从传统的单步MDP转向更能反映现实世界复杂性的POMDP框架[3][4][27]。这一范式转变是根本性的，它要求Agent具备长期规划、记忆、工具使用和在信息不完全的情况下进行决策的能力。其次，Agentic RL的发展呈现出一条清晰的技术演进路径：从依赖人类标注的RLHF，到更原生的、基于内在奖励信号的Agentic RL；从依赖庞大辅助模型的PPO，到更轻量、更高效的GRPO[38][44][52]。这条路径反映了业界对更高效率、更强能力和更低门槛的不懈追求。

在核心能力方面，一个成功的Agentic RL系统必然是规划、记忆、工具使用和自我改进四大支柱协同作用的结果[8][54]。特别是自我改进能力，无论是通过Reflexion式的语言反馈修正，还是通过Absolute Zero式的迭代自训练，都指向同一个未来：Agent将能够脱离人类的直接监督，实现自我完善和进化[54][55]。这种“模型即产品”的趋势，预示着未来的AI应用将更加灵活和强大，而非仅仅是提供API供他人调用[33]。

在应用层面，Agentic RL已经展现出其在软件工程、科学研究、金融服务等领域的巨大价值[28][29][41]。然而，其产业落地仍受限于成本、安全和环境复杂性等现实挑战[15][54]。这促使研究者们积极探索模型量化、高效算法和更逼真的模拟环境等解决方案。评估体系的完善，如AgentBench、τ-bench等基准的出现，也为衡量和比较不同Agent的性能提供了科学依据[14][39]。

总而言之，Agentic RL领域正处在一个高速发展的黄金时期。它不仅在理论上深化了我们对智能的理解，也在实践中不断突破技术的边界。尽管前路依然存在诸多挑战，但其所展现出的潜力无疑是巨大的。未来，随着算法的进一步成熟、算力成本的持续下降以及应用场景的不断拓展，由强化学习驱动的智能体有望成为连接数字世界与物理世界、赋能千行百业的核心基础设施。

参考文献

1. 大模型Agent设计技术路线图：构建智能体系统的核心方法 https://www.betteryeah.com/blog/large-model-agent-design-technical-roadmap-building-intelligent-agent-systems-core-methods
2. 一文读懂大模型Agent 架构，详解Profile，Memory，Planning https://blog.csdn.net/m0_59596990/article/details/135717263

3. Agentic RL Survey: 从被动生成到自主决策 https://zhuanlan.zhihu.com/p/1948526335136867297
4. 综述：基于LLM 的智能体强化学习（Agentic RL） https://zhuanlan.zhihu.com/p/1946896052616663346
5. 100 页Agentic RL 综述！牛津、新国立、AI Lab 等联合定义 ... https://news.qq.com/rain/a/20250910A03ZTR00
6. Agent技术解读：Planning（规划）模块 https://view.inews.qq.com/a/20240908A061FP00
7. The Landscape of Agentic Reinforcement Learning for LLMs https://www.alphaxiv.org/zh/overview/2509.02547v1
8. Agent技术解读：Planning（规划）模块 https://zhuanlan.zhihu.com/p/718959013
9. 综述| Agentic RL for LLM的最新进展与未来挑战，idea满满 https://blog.csdn.net/qq_27590277/article/details/151202587
10. The Landscape of Agentic Reinforcement Learning for LLMs https://arxiv.org/abs/2509.02547
11. 大模型智能体LLM Agent https://zhuanlan.zhihu.com/p/658808853
12. 一文详尽之LLM-Based Agent - 智源社区 https://hub.baai.ac.cn/view/42910
13. 智能体的21 种设计模式总结：Agentic Design Patterns 书评 https://jimmysong.io/blog/agentic-design-patterns-review/
14. LLM Agent（大型語言模型代理） https://forum.aiotcloud.dev/t/topic/71
15. Agentic 是个谎言，本质还是经典RL 原创 https://blog.csdn.net/sinat_37574187/article/details/147466822
16. 8. AI 自主決策& 規劃(Decision-Making & Planning) 想讓 ... https://www.threads.com/@danieltsai04/post/DHOa1V8SM1u/8-ai-%E8%87%AA%E4%B8%BB%E6%B1%BA%E7%AD%96-%E8%A6%8F%E5%8A%83-decision-making-planning%E6%83%B3%E8%AE%93-ai-%E8%87%AA%E5%B7%B1%E8%A8%AD%E5%AE%9A%E7%9B%AE%E6%A8%99%E8%A6%8F%E5%8A%83%E4%BB%BB%E5%8B%99%E9%80%99%E4%BA%9B%E6%96%B9%E6%B3%95%E4%B8%8D%E5%8F%AF%E5%B0%91-%E5%88%86%E5%B1%A4%E8%A6%8F%E5%8A%83-hierarchical-p
17. Agent 開發知識庫- ihower's Notes https://ihower.tw/notes/agent-guideline
18. [AI/GPT/综述] AI Agent的设计模式综述- 千千寰宇 https://www.cnblogs.com/johnnyzen/p/18717441/ai-agent-design-patterns
19. 推测未来Agentic形态：Dynamic Cognitive Contextual Agent ... https://blog.csdn.net/weixin_40941102/article/details/146392502
20. 端到端强化学习在LLM Agent中的应用 https://zhuanlan.zhihu.com/p/1945550319904876116

21. 【深度好文】Agentic Tool Use RL 原生多轮工具调用训练范式 https://blog.csdn.net/weixin_44191845/article/details/149534917

22. 对谈Pokee CEO 朱哲清：RL-native 的Agent 系统应该长什么 ... https://new.qq.com/rain/a/20250801A08YR700

23. 从单智能体到LLM-Agents 的演进丨「大模型时代下的Agent ... https://swarma.org/?p=62283

24. 第九章：强化学习（RL）赋能AI Agents：潜力、挑战与问题建模 https://blog.csdn.net/YPeng_Gao/article/details/147311225

25. 基于LLM 的智能体在多轮对话中的评估的综述 https://zhuanlan.zhihu.com/p/1890525466923861388

26. RLVR来做Agent任务能力增强训练 https://cn.linkedin.com/pulse/rlvr%E6%9D%A5%E5%81%9Aagent%E4%BB%BB%E5%8A%A1%E8%83%BD%E5%8A%9B%E5%A2%9E%E5%BC%BA%E8%AE%AD%E7%BB%83-boyang-zhou-iuk9c

27. 牛津、上AI Lab重磅综述：Agentic RL，一文看懂AI智能体的 ... https://zhuanlan.zhihu.com/p/1951339432193036596

28. Agentic RL——下一代企业级AI智能体的终极路线图 https://www.51cto.com/article/825188.html

29. 基于大语言模型的Agent智能体在金融行业中的应用 https://developer.nvidia.com/zh-cn/blog/llm-agent-for-finance/

30. 上AI Lab重磅综述：Agentic RL，一文看懂AI智能体的进化路线图 https://xfyun.csdn.net/68c931efa6dc56200e856e90.html

31. 开源上新| Agent Lightning：零侵入强化学习，为任意AI智能体 ... https://www.microsoft.com/en-us/research/articles/agent-lightning/

32. 多智能体系统在大语言模型中的应用与强化学习训练方法 https://zhuanlan.zhihu.com/p/1910345358380339636

33. 模型即产品：万字详解RL驱动的AI Agent模型如何巨震AI行业范式 https://developer.aliyun.com/article/1659192

34. 基于LLamaFactory 和EasyR1 打造一站式无代码大模型强化 ... https://aws.amazon.com/cn/blogs/china/building-llm-model-hub-based-on-llamafactory-and-easyr1/

35. LLMs之Agent之RL：RAGEN的简介、安装和使用方法 https://blog.csdn.net/qq_41185868/article/details/147434189

36. 论文导读| 大语言模型中应用到的强化学习算法 https://www.modb.pro/db/625092

37. 如何借助LLM 设计和实现任务型对话Agent https://www.thoughtworks.com/zh-cn/insights/blog/machine-learning-and-ai/how-to-design-task-based-dialogue-Agent-with-LLM

38. RL 是LLM 的新范式 https://www.53ai.com/news/LargeLanguageModel/2024082341792.html

39. Agentic AI基础设施实践经验系列（六）：Agent质量评估 - AWS https://aws.amazon.com/cn/blogs/china/agent-quality-evaluation/

40. 12章构建AI Agent：从LLM选型到多Agent协作的完整技术栈 https://studygolang.com/articles/39220

41. LLM agentic模式之multi-agent: ChatDev,MetaGPT, ... https://blog.csdn.net/beingstrong/article/details/141873846

42. 从原理到实践：万字长文深入浅出教你优雅开发复杂AI Agent https://zhuanlan.zhihu.com/p/1919338285160965135

43. PPO和GRPO——最流行强化学习算法流程对比 https://zhuanlan.zhihu.com/p/1920646701053645255

44. DeepSeek基础：PPO、DPO、GRPO概念详解原创 https://blog.csdn.net/EnjoyEDU/article/details/146494231

45. Unsloth强化学习教程：从RLHF、PPO到GRPO训练推理模型 https://www.xinfinite.net/t/topic/13063

46. 為什麼GRPO比PPO更高效？ 關鍵區別： 1. 去掉評論家 ... https://www.threads.com/@prompt_case/post/DFmPOvYpuSd

47. 一文读懂PPO 与GRPO：LLM 训练的关键算法 https://www.51cto.com/aigc/4105.html

48. 从DeepSeek-R1中了解强化学习的策略优化，从PPO, TRPO ... https://zhuanlan.zhihu.com/p/20858086974

49. 解读DeepSeekMath中的RL策略！GRPO：改进PPO增强 ... https://blog.csdn.net/AIBigModel/article/details/145243853

50. GRPO、PPO、DPO 深入解析与对比 https://zhuanlan.zhihu.com/p/31906665568

51. DeepSeek-R1：语言模型极限与GRPO的数学探索 https://www.atyun.com/66529.html

52. LLM中的强化学习算法——RLHF、PPO、DPO、GRPO 原创 https://blog.csdn.net/qq_45889056/article/details/146165758

53. LLM Agent的构建：OpenAI官方指南解读 https://www.cnblogs.com/CareySon/p/18848452/openai_llm_agent_summary

54. 从「会说」迈向「会做」，LLM下半场：Agentic强化学习范式综述 https://zhuanlan.zhihu.com/p/1948454792096614282

55. 从「会说」迈向「会做」，LLM下半场：Agentic强化学习范式综述 https://cj.sina.com.cn/articles/view/5953740931/162dee083067021dse?froms=ggmp