

大语言模型赋能强化学习智能体研究综述

周鑫, 曹雷, 陈希亮, 曹育箐

引用本文

周鑫, 曹雷, 陈希亮, 曹育箐. 大语言模型赋能强化学习智能体研究综述[J]. 计算机科学, 2026, 53(10): 254-267.

ZHOU Xin, CAO Lei, CHEN Xiliang, CAO Yuqing. [Review of Research on Large Language Model-empowered Reinforcement Learning Agents](#) [J]. Computer Science, 2026, 53(10): 254-267.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[PacketScope:大语言模型赋能eBPF的协议栈安全防御](#)

PacketScope:LLM-enhanced eBPF Framework for Protocol Stack Security

计算机科学, 2026, 53(10): 440-446. <https://doi.org/10.11896/jsjx.260100048>

[自动化代码重构:基于大语言模型的提示演化方法](#)

Automatic Code Refactoring:LLM-based Prompt Evolution Approach

计算机科学, 2026, 53(10): 396-412. <https://doi.org/10.11896/jsjx.251200012>

[面向安全关键软件单元测试的数据集智能生成方法](#)

Intelligent Dataset Generation Method for Safety-critical Software Unit Testing

计算机科学, 2026, 53(10): 345-355. <https://doi.org/10.11896/jsjx.250900141>

[基于语义驱动执行的代码质量评测框架](#)

Semantic-driven Execution-based Framework for Code Quality Evaluation

计算机科学, 2026, 53(10): 320-335. <https://doi.org/10.11896/jsjx.250900140>

[CLERAG:基于认知层次评估的法律问答检索增强生成方法](#)

CLERAG:Retrieval-augmented Generation Framework for Legal QA with Cognitive-levelEvaluation

计算机科学, 2026, 53(10): 295-300. <https://doi.org/10.11896/jsjx.250900130>

大语言模型赋能强化学习智能体研究综述

周鑫 曹雷 陈希亮 曹育箐

陆军工程大学指挥控制工程学院 南京 210007

(zhou1230x@qq.com)

摘要 随着大语言模型能力的快速发展,利用大语言模型在样本处理、推理生成和任务规划等方面展现的强大能力赋能强化学习技术,正成为机器学习领域的新兴热点,为应对强化学习面临的挑战与提升其性能提供了新思路。首先,从任务分解与规划、状态表征及增强、奖励设计和塑形、策略生成与优化这4个角度系统阐述了大语言模型对强化学习基本要素的优化机制,揭示了如何将大语言模型整合到强化学习中以提高其性能,并通过分析强化学习的局限及大语言模型的能力,对上述优化机制及整合方法的原理和架构进行了解析;然后,介绍了其在游戏AI、机器人技术和自动驾驶等领域的应用情况;最后,总结了当前面临的主要挑战,并对未来探索领域进行了展望。

关键词: 人工智能;强化学习智能体;大语言模型;大语言模型赋能强化学习

中图分类号 TP181

Review of Research on Large Language Model-empowered Reinforcement Learning Agents

ZHOU Xin, CAO Lei, CHEN Xiliang and CAO Yuqing

College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China

Abstract With the rapid development of the capabilities of large language models (LLMs), the technology of LLM-empowered reinforcement learning (RL), which leverages the powerful abilities of LLMs in sample processing, reasoning generation, and task planning, has emerged as a new hotspot in the field of machine learning. This technology provides new ideas for addressing the challenges faced by RL and enhancing its performance. Firstly, this paper systematically elaborates on the optimization mechanisms of LLMs on the basic elements of RL from four perspectives: task decomposition and planning, state representation and enhancement, reward design and shaping, and policy generation and optimization. It reveals how to integrate LLMs into RL to improve its performance and analyzes the principles and architectures of the above optimization mechanisms and integration methods based on the limitations of RL and the capabilities of LLMs. Subsequently, the applications of LLM-assisted RL in the fields of game AI, robotics, and autonomous driving are introduced. Finally, the main challenges currently faced are summarized, and future research directions are prospected.

Keywords Artificial intelligence, Reinforcement learning agent, Large language model, Large language model-empowered reinforcement learning

1 引言

作为人工智能的核心分支之一,强化学习是一种解决控制与决策问题的强大学习范式,其核心是通过不断与环境交互使智能体学习最优行为策略,以最大化长期累积奖励^[1]。强化学习的结构如图1所示。然而,最初的强化学习方法虽然在低维状态空间和简单环境中表现出色,但往往难以应对高维状态空间和复杂环境^[2]。2013年DeepMind团队开发的深度Q网络^[3](Deep Q-Network, DQN),首次将深度学习技术与强化学习结合,使得算法能够在高维感知输入上学习策略。DQN在Atari游戏上的成功应用^[4],也标志着深度强化

学习时代的开始。而后的A3C算法^[5]和PPO算法^[6]进一步提升了深度强化学习的性能,使其在各个领域都取得了突破。

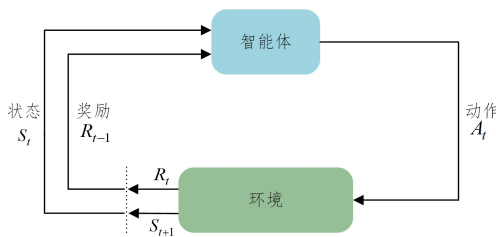


图1 强化学习的结构

Fig. 1 Structure of reinforcement learning

到稿日期:2025-08-26 返修日期:2025-12-11

基金项目:国家自然科学基金(62273356)

This work was supported by the National Natural Science Foundation of China(62273356).

通信作者:曹雷(caolei.nj@163.com)

然而,强化学习仍然面临着诸多挑战:一方面,传统强化学习算法大多缺乏与特定任务相关的先验知识,需要大量采样以逼近环境动态;另一方面,由于当前的主流算法通常依赖神经网络的黑盒特性实现端到端决策,因此强化学习策略普遍存在可解释性不足的问题。此外,强化学习的训练过程稳定性较差,这是因为与传统监督学习丰富的数据集相比,强化学习仅以标量奖励信号作为监督,这种单一维度的监督信号在复杂现实场景中难以充分表征环境动态,而实际环境可以提供文本反馈、视觉信号等多模态感知信息。因此,研究者们正在探索将多模态信息整合到强化学习中,以提供更丰富的反馈和监督,使强化学习模型在动态环境中表现出更强大的泛化能力和解释性。

早在2021年,Tasrin等^[7]就提出了通过自然语言赋能强化学习智能体的框架,这为大语言模型(Large Language Model, LLM)赋能强化学习智能体的研究奠定了基础。同时,有研究发现,让智能体执行下一步行动的时候,加上大模型自己的思考过程,并将思考过程、执行的工具和参数以及执行的结果等信息放到提示词中,就能使得模型对当前和先前的任务完成度有更好的反思能力,从而提升模型解决问题的能力^[8]。

2017年Google提出的Transformer模型^[9],通过注意力机制解决了长距离依赖问题,成为了LLM的通用架构。2019年,Lu等首次将BERT扩展到图文对数据^[10],通过独立编码器和跨模态注意力构建的双流架构实现了视觉-语言的联合学习,该架构目前已经支持文本、图像、音频和视频等数据的混合输入与生成。LLM实现了从单模态基础到跨模态对齐再到多模态交互的飞跃,具备多模态信息的处理能力。同时,DeepSeek^[11]和o1等大推理模型的出现,标志着LLM从直觉式生成迈向深度逻辑推理,实现了推理和规划能力的突破。LLM的发展为人工智能领域带来了巨大变革,随着LLM能力的不断发展,利用其赋能强化学习智能体的潜力越来越大。

为了使LLM能够赋能强化学习智能体的训练过程,需要将自然语言转化为智能体可用的知识。对此,Rodríguez-Sánchez等^[12]引入了一种领域特定语言RLang,通过定义精确的语法和语义,允许使用者使用自然语言描述状态因素、目标、分层动作、策略结构、动作限制、奖励函数、转移函数等马尔可夫决策过程(Markov Decision Process, MDP)元素,然后将自然语言转化为强化学习智能体能够利用的关于MDP各组成部分的部分先验知识。LLM可以将自然语言翻译成RLang,而智能体利用RLang提供的相关知识能提高学习效率。Feng等^[13]受人类语言驱动策略的启发,提出了自然语言强化学习(Natural Language Reinforcement Learning, NLRL),其通过将传统强化学习的数学框架映射到自然语言空间,提高了强化学习决策过程的可解释性。NLRL结合LLM,以自然语言的形式构建了所有元素,将传统的MDP转换为基于文本的MDP,利用文本描述来表示MDP中的状态、动作、奖励和策略等基本概念。从RLang和NLRL的实现可以看到,利用LLM赋能强化学习智能体主要是对强化学习的MDP中的目标、状态、奖励和策略等进行干预。因此,本

文围绕LLM针对强化学习的目标、状态、奖励和策略的赋能方法展开讨论,如图2所示。

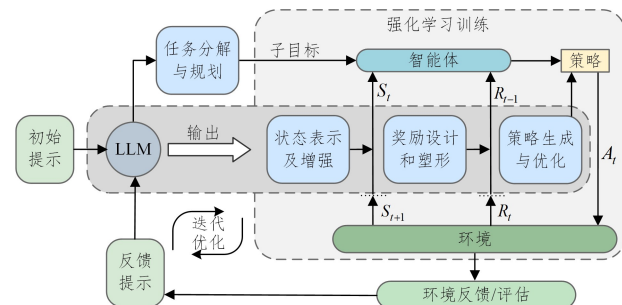


图2 大语言模型赋能强化学习智能体的结构

Fig.2 Architecture of large language model-empowered reinforcement learning agent

当前关于LLM赋能强化学习智能体的研究成果不断涌现,但是缺乏综述性的文章对方法和框架进行详细的分析。为此,本文系统探讨了LLM赋能强化学习智能体的研究进展,基于其对强化学习基本要素的差异化干预方式,如任务分解、状态表征、奖励设计、策略生成等,分类整合了国内外的前沿方法与框架;通过剖析强化学习的固有局限与LLM的赋能能力,阐明了LLM如何针对性地赋能强化学习智能体的核心组成部分,以提升整体性能,并介绍了当前研究在部分领域的应用;最后,对当前挑战与发展趋势进行了系统性总结与展望。

2 相关概念

2.1 强化学习问题建模

强化学习是一种通过奖励机制驱动的学习方式。在强化学习中,智能体与环境的交互过程常用MDP进行数学建模,由一个五元组 (S, A, P, R, γ) 来定义。其中, S 为状态集; A 为动作集; P 为状态转移概率; R 为奖励函数; γ 为折扣因子,描述了智能体如何在完全可观测的环境中进行决策。MDP定义了一个动态系统,在该系统中,智能体根据当前策略 π 选择一个动作 $a_t \in A$,环境根据当前的状态 S_t 和智能体选择的动作 a_t ,使用状态转移函数 P 转移到下一个状态 $S_{t+1} \in S$ 并给出奖励 $r_t \in R$,智能体接收到奖励 r_t 并观察新状态 S_{t+1} 更新策略。该过程的核心就是智能体通过学习最优策略来最大化累积奖励。

而当智能体无法直接观测到全局的环境状态信息时,通常会将问题建模为部分可观测马尔可夫决策过程(Partially Observable Markov Decision Processes, POMDP)。POMDP扩展了标准马尔可夫决策过程,通常由一个七元组 $(S, A, T, R, O, Z, \gamma)$ 来定义,通过引入观测空间 O 和观测函数 Z 解决智能体在信息受限场景下的决策问题。在POMDP中,智能体无法直接观测环境的真实状态,因此只能通过不断收集环境信息来更新自己所处状态的可信度,并且结合观测和动作的历史序列来做决策。

2.2 大语言模型及其能力演进

LLM是一种基于深度学习技术的自然语言处理模型,其基础架构是Transformer。随着大模型技术的不断发展,当前

的 LLM 已经远超最初的文本生成模型,其核心能力经历了深度演进,主要可以概括为 3 个阶段。

1)语言理解与生成。理解和生成能力是 LLM 核心竞争力的基础;能够处理多形式语言输入,根据上下文信息动态调整对当前信息的理解并感知隐含的信息^[14],使其能够更加准确地理解情况,从而得到符合需求的输出;能够根据输入的提示生成对话、文章、代码等多种形式的文本内容,且输出内容逻辑连贯,语义通顺。此外,LLM 还能根据不同的提示词调整输出,生成符合特定主题、情感色彩甚至文化背景的内容。

2)深度推理与规划。通过思维链^[15]、思维树^[16]以及思维图^[17]等先进的提示工程技术,LLM 在回答问题时可以将问题的解决过程分解为能够执行且可验证的步骤,并构建清晰且易于理解的推理逻辑。这种思维过程使其在处理复杂任务时表现出色^[18]。同时,强化学习通过奖励函数鼓励 LLM 拆解复杂问题,增强其复杂推理和问题解决的能力^[11]。这种强大的、内生的推理与规划能力,为 LLM 解决强化学习中的长期任务规划与分解问题提供了根本保证。

3)多模态融合交互。当前的前沿模型已发展为多模态大模型。它们能够直接处理并关联图像、视频、音频等多模态信息与文本,同时精准提取核心语义并实现多模态信息与文本之间的语义对齐^[19],通过统一编码器或跨模态注意力机制,将图像和视频等非文本模态与文本模态映射到同一语义空间。这使得模型能够直接理解像素级信息背后的高层语义,实现视觉问答、图像描述、基于文本指令的图像编辑等复杂任务。

2.3 大语言模型赋能强化学习智能体

传统的强化学习问题是通过 MDP 进行建模,依靠数值化的状态、动作和奖励来进行求解。这种类型的数据往往难以准确地反映实际问题,可能会导致最后训练得到的模型无法准确地解决问题。而 LLM 凭借其强大的理解、生成和推理能力以及丰富的先验知识,可以很好地弥补这方面的缺陷;同时,LLM 不仅可以处理自然语言数据,在处理视觉数据方面也有强大的能力^[20]。因此,LLM 具备了处理和解释复杂的多模态信息的能力,可以更加准确地反映实际问题,为 LLM 赋能强化学习智能体提供了有力支撑。LLM 的理解和生成能力能够将多种信息输入与先验知识结合,为强化学习智能体提供结构化先验知识^[21],使智能体在训练初期具备方向性探索能力,避免完全随机探索导致资源浪费,显著减少试错成本。LLM 的推理能力使其能够对强化学习任务进行解析,对复杂任务进行分解与规划^[22],或者在奖励稀疏的环境中找到潜在的中间奖励^[23],提高强化学习的效率。LLM 的泛化性赋能强化学习智能体,使其能够适应不同的环境或任务^[24],增强了其环境适应性。

3 任务分解与规划

在强化学习中,任务分解与规划是解决复杂问题的重要方式。任务分解是指将一个复杂的任务拆解为多个简单的子任务,并为每个子任务设计明确的奖励和目标;任务规划是指智能体根据当前的环境状态和任务目标,通过一系列的决策来制定最优的行动策略,以完成任务。当前的任务分解与规

划方法主要是基于规则或约束算法,大多依赖于符号规划器^[25]。然而,基于符号的方法往往缺乏良好的鲁棒性,难以合理地将约束条件分配给子任务,导致某些子任务须同时处理过多的约束,从而大幅增加学习难度;并且在面对大规模任务时,往往会出现规划效率低下或规划结果不准确的问题^[26]。最新的研究表明,LLM 在任务规划和分解方面具有强大的能力^[8]。如图 3 所示,LLM 凭借其强大的推理能力和语义理解能力,能够对复杂任务进行结构化拆解,将模糊的高层目标转化为明确、可操作的子目标序列,从而引导智能体在面对复杂任务时高效地进行探索,并且通过环境反馈来动态调整输出,得到更加合理的目标序列。

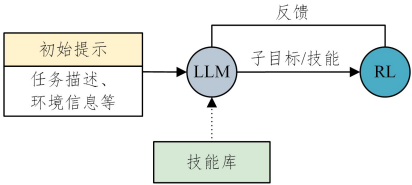


图 3 任务分解与规划结构

Fig. 3 Task decomposition and planning architecture

3.1 子目标分解

在强化学习中,目标是引导强化学习智能体探索的重要因素^[27],它不仅为智能体提供了明确的学习方向,还直接影响智能体的动作策略和学习效率。从实践来看,合理的目标设定可以提高样本效率并减少训练过程中的环境交互次数。对于复杂的长期任务,智能体需要处理长期依赖关系,且奖励信号通常非常稀疏,这使得智能体难以学习到有用的策略。任务分解可以把复杂任务分解为多个简单的子目标,从而降低任务的复杂性。传统的子目标分解方法在处理复杂任务时,难以有效地进行分解。LLM 凭借其强大的推理能力,能够将复杂的长期任务目标拆解为一系列具有逻辑顺序的子目标。Dalal 等^[22]使用 LLM,结合任务描述将任务拆解为多个阶段性子目标,并定义各个子目标的终止条件,然后通过运动规划使智能体到达目标区域,强化学习模块则基于局部观察在线学习局部控制策略。此外,还可以利用 LLM 直接生成动作序列。Bhambri 等^[28]使用 LLM 将上层子任务分解为动作序列 π_g ,再使用基于环境模型的规则检查器验证动作的可执行性。

对于多智能体强化学习任务,智能体数量的增加将导致动作状态空间呈指数级增长,这给策略空间探索带来了巨大的挑战。Qu 等^[24]提出的 LEMAE 框架,利用 LLM 生成关键状态鉴别函数 f_i ,从轨迹中自动定位对任务完成至关重要的关键状态,并基于关键状态构建了一个树形结构,引导智能体以关键状态为目标进行探索。LEMAE 在多个任务上实现了最高达 10 倍的加速。该方法不但提高了智能体的探索效率,而且通过提示工程使智能体获得了跨任务的泛化能力。

在上述方法中,LLM 可以将复杂的任务拆解为多个子目标。但是,由于环境在动态变化,一开始生成的一系列任务子目标不一定适用于后续的任务。为了让分解后的目标能够适应环境的动态变化,Du 等^[23]利用状态描述函数:

$$C_{obs}:\Omega\rightarrow\Sigma^* \tag{1}$$

将当前状态空间 Ω 映射为自然语言文本 Σ^* , 然后使用当前状态描述与任务指令提示 LLM 生成可能的探索目标。Huang 等^[29] 提出可以利用 LLM 生成机器人的操作序列, 并且根据成功检测反馈、场景描述反馈、人类交互反馈等来自环境的各种反馈信息对 LLM 初始生成的操作序列进行调整, 以提高操作序列对环境的适应性。Dasgupta 等^[30] 将强化学习任务划分为规划器、执行器和报告器。规划器由 LLM 充当, 可以根据报告信息调整发出的指令来指导执行器的行动目标; 执行器是负责在环境中执行规划器生成的简单指令的强化学习智能体; 报告器负责观测环境和执行器的动作, 并向规划器报告信息。通过结合环境的变化信息, LLM 能实时调整子目标, 以适应动态变化的环境, 提高强化学习智能体在复杂环境的探索效率和成功率。

3.2 分层技能规划

由于 LLM 可能缺乏关于特定领域的相关知识, 尽管它能够对复杂任务进行分解并生成任务目标, 但其生成的目标可能不适用于特定任务中的智能体。而预训练技能将一系列低层次的动作封装为高层次行为单元, 使得 LLM 能够基于当前任务的技能库对任务进行分解, 提升了生成目标的可靠性。对此, Yuan 等^[31] 针对开放世界游戏 Minecraft 中进行多任务学习和长期任务规划的挑战, 提出利用 LLM 构建基本技能之间的关系, 然后在任务执行时通过该图生成合适的技能序列。类似地, Shek 等^[32] 利用 LLM, 根据任务指令 l_i 和初始状态 s_0 将任务分解为子目标序列 G_n :

$$G_n : \{g_{1i}, g_{2i}, \dots, g_{ji}\} = LLM(l_i, s_{i_0}) \tag{2}$$

使用 DQN 来学习状态-子目标值函数, 得到最优子目标 g_i :

$$g_i = \arg \max_{g_i \in G} Q_{\psi}(s_i, g_i, l) \tag{3}$$

然后根据当前状态 s_i 和子目标 g , 选择使最终收益最大的选项 o_i :

$$o_i = \arg \max_{o_i \in O(s_i)} Q_{\psi}(s_i, o_i, g) \tag{4}$$

o_i 是一个时延拓展动作, 定义为三元组 (I, π, β) , 其中 I 是启动条件, π 是内部策略, π 执行时生成一个动作序列, β 是终止条件。

LLM 的训练数据大多来自于互联网规模的未过滤文本, 这些噪声数据可能导致带有偏差, 进而被模型继承, 导致模型输出出现偏差。直接利用 LLM 构建技能图, 可能会因为 LLM 的知识偏差导致任务失败。LLM 通过结合实时的环境变化给出接下来的目标, 能够降低其固有缺陷对输出的影响; 并且在任务失败后, 可以利用 LLM 进行目标纠正。对此, Brohan 等^[33] 通过组合 LLM 基于当前环境生成下一步技能的概率分布 $p(\ell_{\pi} | i)$ 和当前状态 s 下技能 ℓ_{π} 的成功概率 $p(c_{\pi} | s, \ell_{\pi})$, 得到候选技能的概率, 然后智能体基于概率分布进行定向探索。概率分布的计算式如下:

$$p(c_i | i, s, \ell_{\pi}) \propto p(c_{\pi} | s, \ell_{\pi}) p(\ell_{\pi} | i) \tag{5}$$

特别地, Nottingham 等^[34] 通过 LLM 构建了一个抽象世界模型 (Abstract World Models, AWM), 它是由各个任务节点和任务之间的依赖关系构成的有向无环图。任务节点以合成的物品为目标, 被抽象为可复用技能, 通过 AWM 引导智能体进行探索, 同时在与环境的交互过程中对 AWM 进行验证和修改。AWM 的引导显著提高了探索效率, 并且通过与环境交互进行修正, 增强了对 LLM 输出错误的鲁棒性; 同时, 其样本效率比无 LLM 指导的版本提高了 12 倍。

3.3 小结

利用 LLM 将复杂的高层目标分解为结构化子目标, 能够有效降低对高维、离散动作空间进行直接探索和规划的难度, 为解决离散决策问题提供了高效的语义引导, 提高了智能体的学习效率。如表 1 所列, 与传统的基于规则的方法或约束算法相比, LLM 的赋能革新了任务分解的范式, 重塑了规划的流程, 使其具备了语义理解和逻辑推理的全新能力。然而, 在长期规划场景中, LLM 可能出现逻辑断裂的问题, 导致无法生成有效的规划, 因此需要将其与其他方法结合使用, 以提高规划的合理性。

表 1 任务分解与规划方法的对比
Table 1 Comparison of task decomposition and planning methods

维度	传统方法	LLM 赋能	对比分析
任务分解能力	基于规则或优化算法, 难以处理复杂逻辑和动态约束	利用 LLM 推理生成逻辑子目标, 并能动态调整	LLM 的方法更灵活, 适应动态环境; 传统方法在简单任务中更稳定
任务规划效率	在大规模任务中效率低下	通过技能图快速生成规划并实时优化序列	LLM 方法显著提升了速度和效率; 传统方法在低维空间更可预测
局限	难以处理高维状态和稀疏奖励	LLM 存在知识偏差和长期规划局限	

4 状态表示及增强

对于高维状态的强化学习任务, 需要通过状态表示将原始的高维状态向量映射到低维表示。状态表示方法主要分为基于特征的状态表示^[35] 和基于自动编码的状态表示^[36]。基于特征的状态表示需要通过特征工程来选择能够充分表示状态的特征。此种表示方法依赖于设计者的领域知识和经验, 不同人可能会选择不同的特征, 而这可能会导致状态表示的质量参差不齐。基于自动编码的方式需要大量的数据来学习有效的低维表示, 如果数据量不足或数据质量不高, 可能导致学习到的表示不够准确。

凭借强大的多模态信息处理能力, LLM 能够处理语言和图像等非结构化数据, 提取其中的重要信息并对原始状态空间进行优化, 从只依靠数据构建的状态空间优化为知识与数据共同驱动的新型状态空间。优化过程如图 4 所示。

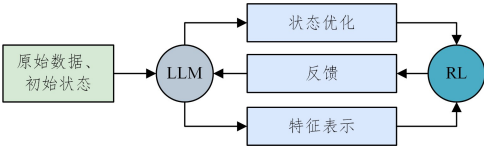


图 4 特征表示和状态优化过程
Fig. 4 Process of feature representation and state optimization

4.1 多模态信息语义转换

自然语言能够为强化学习智能体提供有价值的信息,特别是在状态信号建模时。虽然越来越多的论文将语言融入强化学习,但大多数研究集中在简单的强化学习任务 and 合成语言上。这些语言具有高度的结构化和指导性。而对于非结构化和描述性文本,强化学习在理解自然语言的复杂性和多样性方面还存在挑战^[37]。交互式虚构游戏是提升语言型强化学习智能体泛化能力的重要应用场景,但是文本游戏领域的生成非常耗时。为此,Basavatia 等^[38]提出可以利用基于自然语言的指令生成包含对象、动作、奖励条件、任务序列的游戏元数据,然后将游戏元数据转换为可执行的文本游戏供强化学习智能体学习所需的技能。在工具操作任务中,由于需要理解不同工具的属性和功能,且学习新工具时需要进行多次交互,因此强化学习的训练效率低下。对此,Ren 等^[39]利用 LLM 从自然语言中获得的各类工具的描述生成它们的特征表示,然后通过元学习来执行策略,并使用 Soft Actor-Critic (SAC)算法对策略进行优化,减少了与工具的交互,从而帮助控制策略更快地适应给定任务的新工具。LLM 的语义理解能力可以把非结构化的描述文本转换成智能体可以理解的信息,利用这些信息可以大大提高强化学习的效率。

除了语义信息外,LLM 凭借其多模态信息处理能力,能够将图像和视频等非结构化的信息转化为语言化状态标签。Rocamonde 等^[40]通过映射函数 $\psi(s)$ 将当前状态转化为视觉模型 VLM 可处理的图像,将该图像与任务描述 l 和上下文信息 c 结合并转化为强化学习奖励信号:

$$R_{VLM}(s)=VLM(l,\psi(s),c) \tag{6}$$

相较于人工奖励工程,该方法无需任何针对特定任务的微调或训练,极大地降低了将强化学习应用于新任务的门槛。LLM 能够建立原始数据与高层语义的映射,使智能体可以更准确地理解这些信息,弥补了智能体无法利用这些信息的缺陷,从而提高了强化学习的效率。

4.2 状态空间优化

强化学习在现实世界应用中面临的挑战之一是高维状态和动作空间^[41]。在高维状态下,智能体需要处理大量的输入信息,且训练过程中存在大量的与任务无关的状态,这将大大增加智能体的学习复杂度。为了解决这一问题,Jeong 等^[42]利用 LLM 直接对特征重要性进行评分或排序,如式(7)所示,其评分表示为输入特征 c_j 关于目标概念 c_y 的重要性分数 s_j 。

$$s_j=M(prompt_{score}(c_j,c_y)),\forall j\in\{1,\cdots,d\} \tag{7}$$

排序表示为全部输入特征集合 $\mathbf{c}$ 和排序后得到的与特征概念 c_y 相关的排序表 r 。

$$r=M(prompt_{rank}(\mathbf{c},c_y)) \tag{8}$$

在包含多个真实世界数据集的广泛实验中,该方法在选择最好的 30% 的特征时,其下游模型的平均测试性能与最好的数据驱动基线方法不相上下。

此外,传统强化学习的状态表示缺乏任务相关性,这将导致价值网络在建立从状态到任务奖励的映射时会出现映射不连续和样本效率低的情况^[43]。对此,Wang 等^[44]提出利用

LLM 生成与任务相关的状态表示函数。

$$F:S\rightarrow S^* \tag{9}$$

该函数通过将原始状态 S 扩展为任务相关的高维表示 S^* 来增强与任务相关的状态表示,以促进有效训练。最终,智能体在标准强化学习环境 Mujoco 任务中的平均累积奖励提升 29%,在 Gym-Robotics 任务中的平均胜率提升 30%。利用 LLM 的先验知识和推理生成能力可以减小或拓展原始状态空间,对原始状态空间的优化,能提高强化学习的效率。

4.3 上下文关联推理

在 POMDP 中,智能体只能观测到环境的部分状态,可能会由于观测到的信息太少导致智能体出现决策偏差。而自然语言天然地结合了来自环境的上下文和语义线索。如图 5 所示,LLM 凭借其强大的推理能力,通过自然语言结合环境上下文信息推断隐含的状态。Wei 等^[45]提出,可以利用 LLM,根据提示,通过推断全局环境信息生成能够补充局部观测不足的观测增强函数。

$$o_i'=OE(o_i,h_i) \tag{10}$$

该函数通过智能体的观测值 o_i 和观测历史 h_i 得到增强后的观测值 o_i' ,能够为 QMIX 智能体提供明确的地标距离和伙伴智能体位置估计等空间特征,使其性能提升了 41.7%。Shek 等^[46]提出的 LANCAR 框架,在每个时间步将人类观察者或视觉语言模型对环境上下文的描述输出给 LLM,由其提取环境属性并生成上下文嵌入,然后将上下文嵌入与观测结果结合后作为强化学习智能体的输入。强化学习智能体根据具有上下文感知的输入,使用控制策略产生动作并在环境中执行。通过 LLM 推断环境中隐含的信息,能够帮助智能体更准确地做出决策。

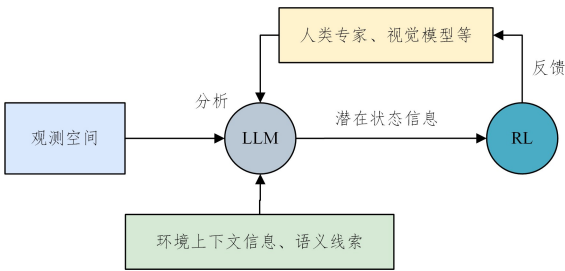


图 5 上下文关联推理过程

Fig. 5 Process of contextual correlation inference

4.4 小结

如表 2 所列,与传统的状态表征方法相比,LLM 通过语义转换、空间优化和上下文推理,将原始的、符号化的离散状态空间转化为富含语义的、连续的低维表示,不仅缓解了维数灾难,还能够减少无关特征对智能体决策的干扰。在只能观测到部分环境状态的 POMDP 任务中,LLM 能够推断出隐含的环境信息,提高智能体决策的可靠性。但是,LLM 的推理延迟问题使其难以在实时决策问题中得到很好的应用,且由于幻觉现象,其生成的状态可能会包含一定的错误语义,尤其是在领域知识不足的情况下,这可能会给智能体传输错误的信息,从而导致错误的决策。

表 2 状态表征方法的对比

Table 2 Comparison of state representation methods

维度	传统方法	LLM 赋能
信息处理能力	依赖人工设计的低层特征	能够处理文本、图像等多模态信息
状态空间效率	特征冗余度高,存在大量无关特征	动态筛选相关特征
环境观测能力	在 POMDP 中,只能观测到环境的部分状态	能够推断出环境隐含的环境信息
数据依赖性	优秀表征往往须依赖专家知识进行特征工程,或依赖大量任务标注数据以有监督方式训练	利用 LLM 的先验知识可以缓解数据稀缺问题
计算开销	自动化编码器轻量化、低延时	在实时场景中,LLM 推理延迟比较高

5 奖励设计和塑形

手工设计奖励函数和专家示范设计奖励函数,是传统强化学习奖励设计的主要方法。前者根据任务需求和目标人工定义状态特征与奖励的映射关系,但需要大量的领域知识,并且在复杂任务中难以设计出合适的奖励函数^[47];后者利用专家的行为数据来推断奖励函数,但依赖于高质量的专家示范数据,这类数据的获取难度较大。而 LLM 凭借其丰富的先验知识和强大的语言理解和生成能力,可以通过结合任务描述或自然语言指令的提示直接生成奖励函数^[24],如图 6 所示。其过程是将高级别的语言指令转化为具体的奖励结构,并且 LLM 可以根据不同的任务需求调整生成的奖励函数,不需要人工设计或调整奖励参数,使得智能体能够动态适应不同的任务。此外,LLM 还能够根据对任务的理解及先验知识,或通过评估当前的动作和状态,为智能体的探索直接生成奖励信号。

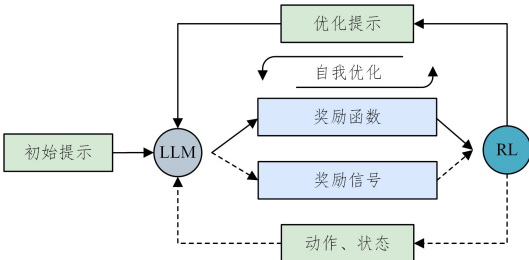


图 6 大语言模型生成奖励函数的过程

Fig. 6 Reward function/signal generation process via LLM

5.1 奖励函数生成与优化

奖励函数是强化学习中的一个基本组成部分,用于评估当前环境状态下特定行为的效果,需要通过体现任务的关键性能指标来帮助智能体在多个行为中做出合理的决策^[48]。但自然语言任务描述通常蕴含多维度的语义信息,人工设计须将其拆解为数学表达式,这会因人类认知偏差与语义信息丢失而难以找到足够的关键性能指标^[49]。Turcato 等^[50]提出通过自然语言任务描述将用户需求输入 LLM,根据特定的提示,LLM 能够生成奖励函数 $r^k(s_t, a_t)$ 和任务终止条件,结合终端奖励 $R_F(s_t, a_t)$ 和二元分类器 $\Phi(s_{t+1})$,最终奖励函数设计为:

$$r(s_t, a_t, s_{t+1}) = \sum_{k=1}^K r^k(s_t, a_t) + R_F(s_t, a_t) \cdot \Phi(s_{t+1}) \quad (11)$$

其中,二元分类器定义为:

$$\Phi(s_t) = \begin{cases} 1, & \text{If in } s_t \text{ the task is solved} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

在推箱子、抓取、插入、双操作臂协作等 10 个机器人操作

任务中,该方法生成的奖励函数均优于人工设计的基线奖励。

为了增强强化学习决策问题的可解释性,Xu 等^[51]利用 LLM 在训练开始前预先生成所有可能状态的语义特征并存储为查找表,然后对于每个状态,在训练过程中通过查找表来对比当前状态下目标网络 $\hat{g}(h_{t+1})$ 和预测网络 $g_\varphi(h_{t+1})$ 的语义特征,并根据二者的误差生成语义分支奖励。

$$\bar{b}_t^{\text{md}} = g_\varphi(h_{t+1}) - \hat{g}(h_{t+1})^2 \quad (13)$$

为了增强生成奖励和动作间的逻辑性,Park 等^[52]利用 LLM 在训练前生成动作树和每个动作的奖励函数,通过动作-奖励的树状结构分解来提升训练的稳定性 and 效率。

为了缓解 LLM 自身固有局限对奖励函数的影响,可以结合环境反馈等信息调整输出,使得智能体能够在交互过程中与环境状态对齐并实时调整其行为,生成更加符合任务需求的奖励函数。LLM 可以根据不同的信息反馈来优化奖励函数。Li 等^[53]通过分析错误轨迹来优化 LLM 生成的奖励函数。Sun 等^[54]采用 LLM 驱动的编码器-评估器架构,其中编码器生成奖励函数代码,评估器通过过程反馈、轨迹反馈和轨迹偏好评估迭代优化奖励,无需人类干预或重复强化学习训练。Ma 等^[55]提出的 EUREKA 方法,能够利用进化搜索和奖励反思,基于策略训练统计的文本总结迭代地改进 LLM 生成的奖励函数,实现了零样本奖励生成和上下文改进。在 EUREKA 的基础上,Masadome 等^[56]提出了一个分层架构,在高层利用 LLM 生成预定义的最优动作目标的混合列表,在每个状态下将当前选择的动作与混合列表对比并生成奖励;低层根据高层提供的动作目标和当前状态选择动作,并使用 EUREKA 生成的奖励函数评估动作质量来更新策略。高层代理能够通过环境反馈实时更新目标策略。

为了让奖励能够准确地反映人类的期望,使智能体学习到与人类目标一致的策略,Nazir 等^[57]使用 LLM 处理带有任务描述或提示的结构化输入来生成反馈以代替人类反馈。该过程定义为:

$$P(yx, t) = f_\sigma(x, t) \quad (14)$$

其表示给定输入 x 和任务描述 t 的情况下预测输出 y 的概率分布。将大型语言模型反馈 $\zeta_{\text{LLM-D}}(s, a; \sigma)$ 添加到来自环境的内在奖励中,以增强每个时间步骤的奖励。

$$\tilde{r}_t = r_t + \zeta_{\text{LLM-D}}(s, a; \sigma) \quad (15)$$

Hazra 等^[58]将人类反馈整合到 LLM 提示中,LLM 根据人类反馈输出相应的奖励信号来训练智能体,使智能体能够学习到符合人类偏好的行为策略。Verma 等^[59]将 LLM 与社会理论结合,确保了奖励函数既符合用户语言提示的优先级,又平衡了多目标的资源分配。上述研究说明,LLM 可以对人类提供的自然语言反馈或预设规则进行解释和扩展,将其转

化为更具体、更详细的奖励调整建议,生成符合人类期望的奖励函数。

5.2 稀疏奖励转化

在一些任务中,智能体可能只有在完成整个任务时才能获得奖励,而任务过程中没有或者很少有中间奖励。由于缺乏中间奖励这种明确的指导信号来指示其行动,在这种奖励稀疏的任务中智能体很难进行有效的学习^[60]。稀疏奖励转化主要关注如何将稀疏的奖励信号转化为更密集、更具有指导性的信号,以便强化学习算法能够更加有效地学习。面对这一挑战,LLM 基于其对任务的理解和先验知识,可以为智能体提供中间奖励信号,从而引导智能体有效地探索环境。LLM 将复杂任务拆解为具有逻辑关联的子目标,并结合反馈信息,能够为每个子目标赋予即时奖励,这是一类实现稀疏奖励转化的具体方法。为了给每个子目标赋予即时奖励,可以通过语义相似度^[61]来生成奖励,其计算式为:

$$\theta(\alpha, \beta) = \frac{\alpha \cdot \beta}{\|\alpha\| \|\beta\|} \quad (16)$$

式(16)用两个词向量 α 和 β 的点积除以它们模长的乘积来表示它们的近似程度,将其运用在强化学习中,可以衡量智能体的当前状态或行为与目标任务描述之间的匹配程度。LLM 可以通过其强大的语义理解能力将环境状态或智能体行为转化为可解释的语义表示,并通过计算其与目标的语义相似度生成奖励。Du 等^[23]提出先利用 LLM 生成当前探索目标,然后使用转移描述生成器 $C_{\text{transition}}()$ 将初始观测状态 o 、执行动作 a 、动作执行后的新观测状态 o' 转化为文本形式的自然语言描述 $C_{\text{transition}}(o, a, o')$, 计算该自然语言描述与目标 g 的语义相似度作为奖励。语义相似度的计算式为:

$$\Delta(C_{\text{transition}}(o, a, o'), g) = \frac{E(C_{\text{transition}}(o, a, o')) \cdot E(g)}{\|E(C_{\text{transition}}(o, a, o'))\| \|E(g)\|} \quad (17)$$

类似地, Rocamonde 等^[40]计算当前环境状态的图像嵌入与文本嵌入的余弦相似度作为即时奖励反馈给智能体进行训练。

除了使用语义相似度外,LLM 还可以根据环境信息直接评估动作的好坏并生成奖励。Chu 等^[62]提出的 Lafite-RL 框架,根据场景信息推断当前任务阶段,并评估当前动作的好坏,将其映射为奖励 r_{LLM} 。其映射规则为:

$$r_{\text{LLM}} = \begin{cases} 1, & \text{if goodmove} \\ -1, & \text{if badmove} \\ 0, & \text{else} \end{cases} \quad (18)$$

然后,结合 LLM 的反馈和环境奖励 r_{env} 来优化奖励函数。Place 等^[63]提出了一个双层优化框架,内层使用 LLM 根据当前状态和操作的描述生成指示,外层根据 LLM 的输入和输出生成一个奖励值 r_{LLM} 反馈给智能体。为了修正 LLM 的错误建议,还引入了可学习的补偿奖励项 $f_{\eta}(s, a)$,再融合原始任务奖励 $r_p(s, a)$ 得到最终奖励,计算式为:

$$r_{\eta}(s, a) = r_p(s, a) + z_{\eta}(s, a) r_{\text{LLM}}(s, a) + f_{\eta}(s, a) \quad (19)$$

该方法在有害奖励和随机奖励设置下均优于或匹配基线方法 BIPARS,展现了强大的适应能力和鲁棒性,并且在 Crafter 成就解锁率和任务完成速度上均表现出优势。此外,

Cao 等^[64]提出可以利用 LLM 直接对策略模型的生成结果进行评估并将其转化为强化学习训练期间的中间步骤的内在奖励 r_t^{in} ,将最终奖励定义为环境奖励 r^{ex} 和内在奖励的加权和:

$$r(s, a) = \alpha_1 r^{\text{ex}}(s, a) + \alpha_2 r^{\text{in}}(s, a) \quad (20)$$

5.3 信度分配

在强化学习中,一个长期的复杂任务可以把整个任务分解为多个子任务或阶段,每个阶段都有其特定的目标和动作。在智能体完成整个强化学习任务获得环境反馈的最终奖励或者惩罚后,需要确定每个阶段的动作或者决策对结果的影响程度才能够合理地评估和优化每个阶段的学习策略。在这类时序信度分配问题中面临着延迟效率、低行为影响等挑战^[65]。对此,Qu 等^[66]提出 LaRe 框架,利用 LLM 生成一个奖励函数,该函数将环境状态编码为潜在奖励,通过引入潜在变量 $z_{r,t}$,将最终奖励 R 的分布分解为一个依赖观测序列的编码器-解码器生成模型,如式(21)所示。通过结合多个维度信息的潜在奖励,可以更好地衡量各阶段动作或决策的价值,解决强化学习中的时序信用分配问题。在 MuJoCo 任务中,LaRe 表现出更高的样本效率和更快的收敛速度。与基线方法相比,LaRe 在训练步数较少时就能达到更高的回报值;同时,在多智能体粒子环境(MPE)中,LaRe 能有效分配智能体间的贡献。

$$\begin{aligned} p(R | s_{1:T}, a_{1:T}) \\ &= \int p(R, r_{1:T}, z_{r,1:T} | s_{1:T}, a_{1:T}) dz dr \\ &= \int \left[\prod_{t=1}^T \underbrace{p(r_t | z_{r,t})}_{\text{decoder}} \underbrace{p(z_{r,t} | s_t, a_t)}_{\text{encoder}} \right] p(R | r_{1:T}) dz dr \end{aligned} \quad (21)$$

强化学习中除了时序信度分配外,还有结构信度分配问题。在多智能体任务下,评估每个智能体的行为对团队整体成功或失败的贡献这一任务也面临着巨大的挑战。多智能体强化学习任务的核心目标是优化联合策略参数 θ ,最大化团队累积期望奖励。

$$\max_{\theta} \mathbb{E}_{s_0 \sim \rho, a_k \sim \pi_{\theta}(o_k)} \left(\sum_{k=0}^{\infty} R(o_k, a_k) \right) \quad (22)$$

Nagpal 等^[67]通过解析函数 F_{MCA} 将 LLM 的文本输出 C_{LLM} 提取为个体信用值 c_k^i ,其数学表达式为:

$$F_{\text{MCA}}(C_{\text{LLM}}(p_k, R(o_k, a_k), o_{k+1}, a_k)) = (c_k^1, \dots, c_k^N) \quad (23)$$

然后,将智能体目标重构为最大化累积个体信用之和。

$$\max_{\theta} \sum_{s_0 \sim \rho} \left(\sum_{k=0}^{\infty} \sum_{i=1}^N F_{\text{MCA}}(C_{\text{LLM}}(p_k, R(o_k, a_k), o_{k+1}, a_k)) \right) \quad (24)$$

此外,Wei 等^[45]提出利用 LLM,根据提示生成用于分解全局团队奖励 $R_{\text{global}}(S, a)$ 和个人贡献 $R_{\text{local}}(o_i, S, a_i)$ 的混合奖励函数,利用二者构造最终奖励 $R_i(o_i, S, a_i)$ 。

$$R_i(o_i, S, a_i) = \alpha_i R_{\text{local}}(o_i, S, a_i) + (1 - \alpha_i) R_{\text{global}}(S, a) \quad (25)$$

同时,通过进化算法迭代优化混合奖励函数用于指导 LLM 生成下一代函数,以解决多智能体任务的信度分配问题。

5.4 小结

传统奖励设计的方法依赖于领域知识和专家示范,并且在稀疏奖励处理和信度分配方面都面临着巨大的挑战。如表 3 所列,通过将 LLM 整合到强化学习中,优化了奖励设计和塑形的方

稀疏和信度分配问题提供了新思路。不过,LLM 的输出时延使其在复杂任务中生成奖励的耗时较长,且高度依赖于任务描述、环境反馈和提示工程的质量。此外,LLM 中

潜在幻觉的存在可能导致其生成与真实任务目标语义不一致但表面合理的奖励函数,引导智能体学习出无效甚至危险的行为策略。

表 3 奖励设计和塑形方法的对比
Table 3 Comparison of reward design and shaping methods

维度	传统方法	LLM 赋能	突破
设计方式	依赖于领域知识或专家示范	根据任务描述自动生成	降低设计门槛,支持跨任务泛化
迭代优化方式	调整需要重新进行建模	结合环境或人类反馈实时调整	适应动态环境,提升奖励与目标的匹配度
稀疏奖励处理	依赖于课程学习、奖励塑形等补充技术	直接生成密集中间奖励	显著增加探索效率
信度分配	基于人工规则分解	通过潜在奖励、多目标权重分配等自动分解	解决时序和结构层面的信度分配问题
局限	泛化性差,复杂任务设计困难	依赖于提示工程,在实时交互时存在一定的延时	计算成本较高,需要平衡实时性和精确性

6 策略生成与优化

强化学习的最终目的是学习一个最优策略,以最大化智能体在与环境交互过程中获得的长期累积奖励。在传统强化学习中,策略初始化方法通常是随机初始化或专家演示初始化。随机初始化是让策略参数被随机赋值,这种方法不依赖于任何先验知识,适用于没有专家数据或先验知识的情况,但是可能导致智能体在训练初期需要大量时间来探索有效的策略,从而减慢收敛速度;专家演示初始化利用专家演示数据来初始化策略,这种方法虽然可以显著提高学习效率,但依赖于大量高质量的专家演示数据^[68]。LLM 凭借其强大的推理生成能力和先验知识,可以在没有专家数据的情况下生成先验策略,还能结合状态、动作等信息优化策略性能,并加速策略的收敛。其过程如图 7 所示。

结合多步观测和当前策略给出下一步策略。Karine 等^[73]利用 LLM 解析用户偏好语义,剔除不合适的强化学习动作并注入环境约束,而后根据新的状态,利用贝叶斯推理更新动作分布。其均值和协方差矩阵的更新方程如下:

Σ_{(t+1)a} = σ_{Ya}² (v_t^T v_t + σ_{Ya}² Σ_{ta}⁻¹)⁻¹ (27)

μ_{(t+1)a} = Σ_{(t+1)a} ((σ_{Ya}²)⁻¹ r_t v_t + Σ_{ta}⁻¹ μ_{ta}) (28)

Zhang 等^[74]利用 LLM 接收状态 *s* 并输出动作分布 π_{*h*}^{LLM}(*a* | *s*),通过正则化值函数,借助 KL 散度约束,使 RL 策略 π 向 LLM 输出策略的高概率动作区域选靠拢,以此完成策略优化。正则化值函数如下:

V_{*h*}[†](*s*) = max_{π ∈ Δ_A} { Σ_{*a*} π(*a* | *s*) Q_{*h*}[†](*s*) - λ KL(π(· | *s*) || π_{*h*}^{LLM}(· | *s*)) } (29)

在多个基准测试中,该方法实现了最高的成功率,超越了纯 RL 和纯 LLM 方法,并且在包括 embodied 环境(ALF-World)、交互式编码(InterCode)和经典规划(BlocksWorld)的不同环境类型中均表现优异,展示了强大的泛化能力。

为了提高强化学习策略的可解释性,Wang 等^[75]将专家演示轨迹和环境规则输入 LLM,生成文本化的策略列表,并通过与环境的相互作用不断改进其策略列表及其策略模型,通过文本化的策略列表,提升了策略的可解释性。

6.2 安全约束

安全性是强化学习现实应用面临的一个重要挑战。在自动驾驶、机器人控制等领域,需要规避物理世界的风险,防止智能体执行危险的动作。考虑到安全因素,要对智能体的行为施加一定的安全约束,以限制其随机探索^[76]。但是,目前的方法需要以数学形式指定约束条件^[77]。对此,Yang 等^[78]将自然语言约束和观测状态输入约束解释器,输出约束违反概率预测 $\hat{y}$ 给策略网络进行训练,训练约束解释器的损失函数。损失函数如式(30)所示。训练得到的智能体将约束违反次数,并且在多种任务环境中实现了更高的累积奖励;同时,其能无缝迁移到新任务和未见过的文本约束中,无需重新训练约束解释器。

L(θ₁) = - E_{(o₁, x) ~ D_{train}} [1 / |M_C| Σ_{i,j=1}ⁿ y log y + (1-y) log (1- $\hat{y}$)] (30)

随着 LLM 技术的成熟,约束解释器可以用 LLM 来代替。LLM 可以根据环境观测,结合安全知识生成带有安全约束的策略,使智能体能够在动态且复杂的环境中做出安全决策,并且可以提高策略的可解释性。对此,Gross 等^[79]在策略

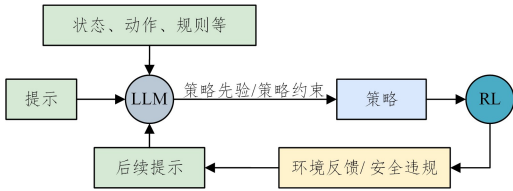


图 7 策略生成与优化过程

Fig. 7 Process of policy generation and optimization

6.1 策略优化

受人类如何利用现有知识做出决策并从新任务中学习的启发,结合 LLM 的先验知识和在处理决策任务中展现的能力^[69],Yan 等^[21]将 LLM 视为动作先验分布。其过程被描述为:

p_{LLM}(*a* | *s_t*) = Σ_{*o*} p(*a* | *o*) LLM(*o* | *s_t*) (26)

LLM 根据当前状态 *s_t* 输出文本建议 *o*,并通过规则映射函数 *p*(*a* | *o*) 将其转化为动作 *a* 的概率,然后利用强化学习在其生成的候选动作范围内进行优化。为了减少对样本的依赖,Kheirandish 等^[70]使用 LLM,结合子任务目标为每个子任务生成规则,然后通过强化学习智能体将规则调整为可解释策略,以解决子任务。

除了直接生成强化学习策略,LLM 还能结合环境、状态等信息对现有的策略进行优化,以提升策略的性能。为了解决高维状态-动作空间的低效问题,Ma 等^[71]利用 LLM,根据示例图像和任务描述等探索提示生成候选策略代码,然后随机选择 LLM 生成的探索策略或当前强化学习智能体的策略来进行行为决策。Peng 等^[72]利用 LLM 进行高层决策,通过

训练完成后,通过形式化验证定位安全风险状态,然后利用 LLM 进行反事实推理,以解释原策略生成动作的故障并提出有效的替代行动,并强制策略选择 LLM 生成的替代行为来提高训练后的强化学习策略的安全性能。在 EUREKA^[55]的基础上, Ma 等^[80]在奖励生成提示中强制加入安全约束 l_{safety} :

$$R_{\text{DrEureka}}, \pi_{\text{initial}} = \text{Eureka}(M, l_{\text{task}} + l_{\text{safety}})$$

(31)

加入安全约束的提示 LLM 能够将安全约束编译为可执行的代码。

6.3 离线策略增强

强化学习是通过智能体与环境交互进行学习的。但是,在安全性要求高或者交互成本昂贵的场景,在线探索的风险很大,在这种情况下通常使用离线强化学习通过监督学习来拟合由状态-动作-奖励轨迹组成的数据集,将强化学习的控制问题作为序列建模问题来处理。但离线强化学习受限于数据集的多样性,难以泛化到新任务。而 LLM 可以用于构建知识库^[81],基于这样的能力, Pang 等^[82]提出使用离线数据集对 LLM 进行监督学习,使 LLM 具备为新技能生成想象推演数据的能力;然后将 LLM 生成的数据与原始数据混合,使用离线强化学习算法来训练最终策略模型 π ,使学到的策略能够泛化到前所未见的目标和任务上。实验表明,该方法在未

见目标任务上的成功率从基线的 26% 提升至 46%,实现了近一倍的性能改进。Bhambri 等^[83]通过基于规则的模型对 LLM 的输出进行批判,使其最终输出一个完整的、能够到达目标的序列。对序列中的每个状态-动作对赋予一个均匀递增的内在奖励:

$$F(s, a) = \gamma \phi(s') - \phi(s)$$

(32)

其中, $\phi(s)$ 为势函数,越靠近目标的状态势能值越高。将内在奖励与原始奖励相加,以增强离线强化学习经验回放缓冲区上的奖励信号,从而更高效地学习到目标策略。

6.4 小结

如表 4 所列, LLM 的先验知识降低了强化学习策略初始化时对专家数据的依赖并提高了初始策略的质量,同时通过结合环境反馈动态调整策略,提高了策略的环境适应性并加快了收敛速度。尽管其自然语言解释还提高了策略的可解释性,但其实际有效性高度依赖领域知识的完整性、提示工程的质量以及环境反馈的验证机制,需注意 LLM 解释可能存在幻觉、逻辑跳跃或过度简化等问题,要结合多模态状态验证或人工评估来保障其可靠性。同时, LLM 的推理延迟可能影响决策效率,需要结合轻量化模型和高效的协作框架以平衡性能与实时性的需求。

表 4 策略生成与优化的对比

Table 4 Comparison of strategy generation and optimization

维度	传统方法	LLM 赋能	突破
初始化依赖	随机初始化或专家演示数据	LLM 生成先验策略	降低专家数据依赖,提升初始策略的质量
优化机制	基于奖励函数的梯度更新	LLM 结合环境反馈生成优化建议	增强环境适应性并加速策略收敛
安全性保障	需人工设计数学约束条件	自然语言规则自动编码为安全约束	降低安全策略的设计门槛
可解释性	黑盒策略,可解释性弱	文本策略描述,增强可解释性	提供策略过程语义解释

7 应用研究

强化学习作为一种强大的人工智能技术,已经在游戏 AI、自动驾驶、生物医学、LLM 和机器人等多个领域得到了广泛的应用^[84]。随着 LLM 赋能强化学习智能体相关研究的不断发展,其已经在部分场景有了应用,当前的应用主要集中在游戏 AI、机器人技术和自动驾驶领域。

1) 游戏 AI: 在类似于“Minecraft”这样的探索空间巨大、奖励稀疏且任务长期依赖的开放自由游戏中,传统强化学习容易陷入“探索困境”,其随机探索策略在解决需要多步逻辑推理的长期任务时效率低下。LLM 根据从海量数据中得到的任务范式,将抽象的长期目标分解为具有逻辑顺序的子目标序列^[34]。该序列为智能体构建了探索路径,使得强化学习算法可以为每一个子目标的达成设计密集的内在奖励,从根本上缓解了长期信度分配难题;并且 LLM 的理解生成能力使其可以结合环境的动态反馈指导智能体实时调整探索目标^[23],增强规划对环境的适应性。但是,这种规划高度依赖 LLM 输出的合理性。然而,由于预训练数据与游戏世界规则不完全对齐, LLM 可能生成不符合特定游戏机制或物理规则的目标,若不对其加以验证,将引导智能体执行无效甚至导致任务失败的动作序列。对此,需引入基于环境模型的规则校验器来对 LLM 的规划输出进行验证。

2) 机器人技术: 在机器人控制任务中,环境给出的往往是

回合奖励,只有在任务成功时才得到回报,这导致其训练效率低下。LLM 能够基于对场景的语义理解,评估智能体的动作并将其转化为与任务目标相关的具体奖励数值^[62]。这种基于语义的奖励塑形提供了比回合奖励更加密集的奖励信号,显著提升了训练效率。同时,对于长期机器人任务, LLM 的规划能力使其能够将任务进行高层次的规划^[85],通过阶段性的目标指导机器人高效训练。对于场景迁移困难这一痛点,为缓解仿真到现实的迁移差距, LLM 可以充当“域随机化引擎”^[80],利用其内在物理常识推理出哪些物理参数是影响任务的关键,并据此生成参数扰动范围。这使得在仿真环境训练的强化学习策略能够覆盖更加合理的物理场景。然而,机器人技术中 LLM 的应用仍面临独特的挑战:一是计算开销与性能的权衡,使用高参数数量的 LLM 需要大量的算力,但使用低参数数量的 LLM 会导致语义理解精度下降,从而出现误判;二是在动态环境中,由于 LLM 的延迟推理可能导致响应不及时,从而出现碰撞的风险,需通过优化模型推理链路或通过异步的方式减少时延。

3) 自动驾驶: 传统强化学习在复杂驾驶环境中依赖于原始传感器数据,导致智能体难以真正理解场景含义从而做出合理的决策;而 LLM 蕴含了海量的常识知识和推理能力。通过提示工程,将环境感知信息作为提示输入 LLM,令其输出高层指令或潜在风险的语义描述。这些语义输出可以作为强化学习智能体状态空间的增强^[86],使其能够产生符合人类

预期、可解释的驾驶行为。此外,安全约束难以量化是自动驾驶领域的一大挑战。LLM 凭借其推理能力,基于强化学习输出的动作预测自车的未来轨迹^[87],从而做出安全的预测。此外,LLM 的理解能力和推理能力使其能够理解用户自然语言指令,并结合场景描述和历史记忆,生成既符合用户意图又考虑安全性的高级策略^[88],实现满足用户个性化需求的驾驶体验。然而,LLM 的固有幻觉问题导致其可能生成不安全或违规的危险指令,这对于自动驾驶来说是致命的,需要通过安全验证机制或对抗性测试修正 LLM 的输出;并且 LLM 的推理延迟高,难以满足驾驶任务这种毫米级实时控制的需求,需要通过异步架构或语义缓存^[86]等方式缓解时延。

除了游戏 AI、机器人技术和自动驾驶领域,LLM 赋能强化学习智能体在医疗保健^[73]和无人机网络^[89]等领域也有相关研究。在自适应干预等医疗场景,传统强化学习策略难以实时解析并响应动态变化以及以文本形式表达的用户偏好,这导致干预策略与用户需求脱节。LLM 利用其自然语言理解能力,能够将文本偏好解析为策略约束,并立即影响动作选择,使策略动态适应变化,让强化学习生成符合用户偏好的医疗干预措施^[73]。但 LLM 的幻觉可能产生有害的医疗建议,并且对 LLM 进行领域微调或知识增强需要在保护患者数据隐私的前提下进行。在无人机网络领域中的多无人机协同任务中,传统多智能体强化学习依赖全局奖励,导致信度分配困难,难以准确评估单个无人机在复杂任务中的贡献,导致策略学习不均。LLM 的链式思维推理可分析 UE(User Equipment)分布和网络拓扑等全局信息,生成角色区分化的部署建议,使策略更加均衡^[89]。但 LLM 的集中式推理与无人机的分布执行架构冲突,将其深度集成到系统中将引发巨大的通信开销;并且在动态对抗环境中,LLM 的决策延迟难以有效指导无人机群的实时协作。

8 挑战与展望

强化学习的一个特征在于其试错学习范式,智能体必须通过与环境的交互来收集数据和更新策略。然而,传统的探索方法在复杂、稀疏奖励的任务中会导致极高的样本复杂性和训练成本。尽管引入 LLM 确实增加了单次决策的计算开销,但其关键优势在于通过提升探索的有效性,从根本上减少无效试错,从而显著减少达到最优策略所需的环境交互轮次,实现加速收敛。尽管利用 LLM 提升探索效率有望降低智能体的试错代价,但当前该技术路径的发展主要受限于 LLM 自身的固有局限性以及 LLM 与强化学习智能体协同机制的不成熟。这两大挑战直接决定了 LLM 生成的先验知识、规划与奖励信号等能否被可靠、高效地转化为对 RL 智能体的有效赋能。具体来说,LLM 赋能强化学习智能体主要面临着以下几个挑战。

1)潜在幻觉现象。LLM 固有的幻觉会使其输出看似合理但事实上不正确、无根据或与输入上下文相矛盾的信息^[90]。其主要源于内在知识的缺陷、对上下文的误解,以及在长序列规划中注意力分散导致的推理链断裂。在强化学习中,幻觉的影响是系统的。在奖励设计中,它可能生成与目标语义不一致的奖励函数;在状态表征中,可能注入错误的语义

信息,导致对环境认知扭曲;在任务规划中,可能生成与实际环境不符的规划目标;在策略生成中,可能输出危险或无效的动作指令,此类错误会被智能体学习并放大,导致策略偏离甚至引发现实世界的安全风险。缓解这种现象可能的方法有:通过 RAG^[91]技术减少其对内部可能错误知识的依赖;通过推理约束限制 LLM 的输出逻辑;通过环境模型或其他验证器进行协同验证,或开发能够对 LLM 输出不确定性量化的方式,确保 LLM 输出的合理性和安全性。

2)重现性不稳定。模型固有的随机性和对提示工程的敏感性,导致不同 LLM 输出可能是可变的,并且通常很难通过提示来重现相同的结果;其次,对于相同的 LLM,它在被输入小的事实错误后,很难重现相同的输出结果,并且自然语言提示的细微差异可能会导致 LLM 的输出显著不同。这种不稳定性使得实验和部署难以复现,增加了研究和应用的不确定性,严重阻碍了技术的比较、验证与迭代。对此,需要对提示工程进行标准化和版本化,通过构建规范的提示词提高 LLM 输出的可重现性。

3)缺乏领域知识。训练数据的固有偏差、知识更新的滞后性以及领域知识的不足,导致 LLM 在某些非常专业或者罕见的领域,即使有精心设计的提示,也无法生成符合期望的输出,这会导致最终的策略的偏差和决策失误。在强化学习中,这种问题会沿着决策链传导并放大。在奖励设计中,LLM 可能因不了解领域内安全规范而生成鼓励危险行为的奖励函数;在状态表征中,可能无法识别专业特性,导致状态信息缺失;在任务规划中,可能提出违背实际的动作序列。对此,可能需要使用外部知识图谱和 RAG^[91]等知识注入技术将领域知识注入 LLM,或者使用特定任务的数据对 LLM 进行微调,从而提升其在特定任务中的表现。

4)计算需求制约。由于 LLM 参数规模巨大,单次输出需要巨大的计算资源和内存带宽。这种计算需求对强化学习构成了训练与部署的双重压力。在训练阶段,频繁调用 LLM 会使单次环境交互的时间成本飙升,严重拖慢策略迭代速度。在部署阶段,LLM 的推理延迟,使得其在机器人控制、自动驾驶等需要实时控制的场景中无法满足高频响应的需求,导致智能体无法及时响应环境变化。因此,需要利用微调或蒸馏等技术实现模型轻量化,使中小模型获得领域专家能力,从而降低 LLM 推理计算的成本,同时设计高效的协作框架,如异步决策框架等,减小 LLM 输出时延对强化学习的影响。

5)语言知识与符号任务不匹配。LLM 虽然蕴含丰富的语义知识,但存在一个重大挑战:其与强化学习任务所需的精确、结构化且可执行的符号表示之间不匹配。由于语义的模糊性和符号的精确性之间存在矛盾,因此 LLM 输出可能无法直接映射到环境动作或状态,导致无法对智能体进行有效赋能。为实现 LLM 与强化学习的有效对齐,需要通过设计与领域相关的中间表示(如技能库或领域特定语言),将 LLM 的任务限定为在这些结构化表示上进行规划和组合。同时,将 LLM 的决策过程与环境的实时反馈紧密耦合。通过递归提示将环境状态持续反馈给 LLM,使其能够基于环境反馈进行动态重规划和错误纠正,突破静态知识的局限。

6)多模态融合与对齐。大模型已经全面进入多模态时

代,其与强化学习结合存在巨大潜力,充分利用文本、视觉等多模态信息是构建全面感知并理解开放世界的通用智能体的必经之路,但如何将视觉、语言、音频等异质信息进行有效融合与语义对齐已成为关键挑战。首先,不同模态之间存在语义鸿沟,信息的偏差将影响策略精度;其次,简单的特征拼接易引发维度灾难,缺乏能动态筛选关键跨模态信息的高效方式;此外,跨模态的时序信度分配极为复杂,难以追溯决策失败是源于对视觉的误判还是对语言信息的误判。为解决上述难题,可以在模型内部学习一个统一的语义表征空间^[92],使得不同模态的信息可以在此空间进行交互。未来需要探索多模态大模型和强化学习的对齐机制,使其能够有效且高效赋能强化学习智能体。

LLM 的固有缺陷会影响智能体最终的性能,但是规模定律^[93]指出模型能力密度会随时间呈指数级增强。随着 LLM 技术的不断发展和 LLM 赋能强化学习智能体研究的不断深入,相信这些难题会不断被克服。虽然目前 LLM 赋能强化学习智能体在几个方面已经展现出了良好的性能,但是未来还有很多有待探索的领域。

1) 高效协同框架。当前研究在需要实时控制的场景下,受到计算消耗的严重制约。未来需要研究高效的协同框架,通过分层决策、缓存机制、模型轻量化等方法,使其能够满足实时控制的需求。

2) 多模态融合学习。现有方法多依赖于文本接口,少部分使用视觉信息。未来可以开发多模态学习框架,充分利用多模态信息,提升在开发世界任务中的感知与决策能力。

3) 专用 LLM 微调。当前研究大多都是直接调用现有的通用 LLM。未来可以探索通过强化学习反馈数据对中小型 LLM 进行微调,培养其成为专家模型,在提升效果的同时降低延迟和成本。

4) 现实世界部署。当前研究主要聚焦于高度结构化的仿真环境。未来研究需解决从受控仿真环境向高风险物理场景迁移的核心挑战,探索在高风险的现实世界中如何可靠且安全地进行部署。

结束语 研究表明,LLM 赋能强化学习智能体是人工智能迈向更通用、可解释决策系统的重要一步。LLM 凭借其强大的语言理解和推理生成能力为强化学习智能体提供了丰富的先验知识与语义理解支持,弥补了传统强化学习在复杂环境中的缺陷。这一研究在游戏 AI、机器人技术和自动驾驶等领域的成功应用,展现了其巨大的潜力。虽然目前相关研究已经取得了一定的进展,但是整体仍处于探索阶段。未来随着 LLM 技术的不断发展以及该领域研究的不断深入,LLM 将在安全强化学习、可解释强化学习和迁移强化学习等多个领域发挥更重要的作用,推动构建强化学习新范式,为解决人工智能领域决策问题提供新思路。

参 考 文 献

[1] SUTTON R S,BARTO A G. Reinforcement Learning: An Introduction[M]. Cambridge, USA: MIT Press, 1998.

[2] PÉREZ-DATTARI R,CELEMIN C,RUIZ-DEL-SOLAR J, et al. Continuous control for high-dimensional state spaces: An

interactive learning approach[C] // 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 7611-7617.

[3] MNIH V,KAVUKCUOGLU K,SILVER D,et al. Playing atari with deep reinforcement learning[J]. arXiv:1312.5602,2013.

[4] MNIH V,KAVUKCUOGLU K,SILVER D,et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.

[5] MNIH V,BADIA A P,MIRZA M,et al. Asynchronous methods for deep reinforcement learning[C] // International Conference on Machine Learning. PMLR, 2016:1928-1937.

[6] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal policy optimization algorithms[J]. arXiv:1707.06347,2017.

[7] TASRIN T,AL NAHIAN M S,PERERA H,et al. Influencing reinforcement learning through natural language guidance [C] // Proceedings of the 34th International Florida Artificial Intelligence Research Society Conference (FLAIRS-34). AAAI Press, 2021.

[8] YAO S,ZHAO J,YU D,et al. React: Synergizing reasoning and acting in language models [C] // International Conference on Learning Representations (ICLR). 2023.

[9] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[10] LU J,BATRA D,PARIKH D,et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C] // NeurIPS. 2019.

[11] GUO D Y,YANG D,ZHANG H R,et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[J]. Nature, 2025, 638: 1-9.

[12] RODRIGUEZ-SANCHEZ R,SPIEGEL B A,WANG J,et al. RLang: a declarative language for describing partial world knowledge to reinforcement learning agents[C] // International Conference on Machine Learning. PMLR, 2023: 29161-29178.

[13] FENG X,LIU B,SONG Y,et al. Natural language reinforcement learning[J]. arXiv:2411.14251,2024.

[14] BROWN T,MANN B,RYDER N,et al. Language models are few-shot learners[C] // Advances in Neural Information Processing Systems 33. 2020: 1877-1901.

[15] WEI J,WANG X,SCHUURMANS D,et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[16] YAO S,YU D,ZHAO J,et al. Tree of thoughts: Deliberate problem solving with large language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 11809-11822.

[17] BESTA M,BLACH N,KUBICEK A,et al. Graph of thoughts: Solving elaborate problems with large language models[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 17682-17690.

[18] XU F,LIN Q,HAN J,et al. Are large language models really good logical reasoners? a comprehensive evaluation and beyond [J]. IEEE Transactions on Knowledge and Data Engineering,

- 2025,37(4):1620-1634.
- [19] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models[C]// International Conference on Machine Learning(ICML). PMLR, 2023: 19730-19742.
 - [20] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision[C]// International Conference on Machine Learning(ICML). PMLR, 2021: 8748-8763.
 - [21] YAN X, SONG Y, FENG X, et al. Efficient reinforcement learning with large language model priors[J]. arXiv: 2410. 07927, 2024.
 - [22] DALAL M, CHIRUVOLU T, CHAPLOT D, et al. Plan-seq-learn: Language model guided rl for solving long horizon robotics tasks[J]. arXiv: 2405. 01534, 2024.
 - [23] DU Y, WATKINS O, WANG Z, et al. Guiding pretraining in reinforcement learning with large language models[C]// International Conference on Machine Learning. PMLR, 2023: 8657-8677.
 - [24] QU Y, WANG B, JIANG Y, et al. Choices are more important than efforts: Llm enables efficient multi-agent exploration[J]. arXiv: 2410. 02511, 2024.
 - [25] HASLUM P, LIPOVETZKY N, MAGAZZENI D, et al. An introduction to the planning domain definition language[M]. Morgan & Claypool Publishers, 2019.
 - [26] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: Language agents with verbal reinforcement learning[C]// Advances in Neural Information Processing Systems (NeurIPS). 2023: 8634-8652.
 - [27] CHENTANEZ N, BARTO A, SINGH S. Intrinsically motivated reinforcement learning[C]// NIPS. 2004.
 - [28] BHAMBRI S, BHATTACHARJEE A, KALWAR D, et al. Extracting Heuristics from Large Language Models for Reward Shaping in Reinforcement Learning[J]. arXiv: 2405. 15194, 2024.
 - [29] HUANG W, ABBEEL P, PATHAK D, et al. Inner Monologue: Embodied Reasoning through Planning with Language Models [C] // Proceedings of the Conference on Robot Learning (CoRL). 2022.
 - [30] DASGUPTA I, KAESER-CHEN C, MARINO K, et al. Collaborating with language models for embodied reasoning[J]. arXiv: 2302. 00763, 2023.
 - [31] YUAN H, ZHANG C, WANG H, et al. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks [J]. arXiv: 2303. 16563, 2023.
 - [32] SHEK C L, TOKEKAR P. Option Discovery Using LLM-guided Semantic Hierarchical Reinforcement Learning[J]. arXiv: 2503. 19007, 2025.
 - [33] BROHAN A, CHEBOTAR Y, FINN C, et al. Do as i can, not as i say: Grounding language in robotic affordances[C]// Conference on Robot Learning. PMLR, 2023: 287-318.
 - [34] NOTTINGHAM K, AMMANABROLU P, SUHR A, et al. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling[C]// International Conference on Machine Learning. PMLR, 2023: 26311-26325.
 - [35] BERTSEKAS D P. Feature-based aggregation and deep reinforcement learning: A survey and some new implementations [J]. IEEE/CAA Journal of Automatica Sinica, 2018, 6(1): 1-31.
 - [36] FALZARI M. Autoencoder-based Deep Reinforcement Learning for ground-level walking of Human Musculoskeletal models [D]. Groningen Provincie; University of Groningen, 2022.
 - [37] LUKETINA J, NARDELLI N, FARQUHAR G, et al. A survey of reinforcement learning informed by natural language [C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence(IJCAI). 2019: 6309-6317.
 - [38] BASAVATIA S, MURUGESAN K, RATNAKAR S. Starling: Self-supervised training of text-based reinforcement learning agent with large language models [C]// Findings of the Association for Computational Linguistics: ACL 2024. Bangkok; Association for Computational Linguistics, 2024: 15804-15819.
 - [39] REN A Z, GOVIL B, YANG T Y, et al. Leveraging language for accelerated learning of tool manipulation [C] // Conference on Robot Learning. PMLR, 2023: 1531-1541.
 - [40] ROCAMONDE J, MONTESINOS V, NAVA E, et al. Vision-language models are zero-shot reward models for reinforcement learning [C]// Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). New Orleans, USA: Curran Associates, Inc., 2023.
 - [41] ZHU J, WU F, ZHAO J. An overview of the action space for deep reinforcement learning[C]// Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence. 2021: 1-10.
 - [42] JEONG D P, LIPTON Z C, RAVIKUMAR P. Llm-select: Feature selection with large language models [J]. arXiv: 2407. 02694, 2024.
 - [43] ASADI K, MISRA D, LITTMAN M. Lipschitz continuity in model-based reinforcement learning[C]// International Conference on Machine Learning. PMLR, 2018: 264-273.
 - [44] WANG B, QU Y, JIANG Y, et al. LLM-Empowered State Representation for Reinforcement Learning [C]// Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria; PMLR, 2024: 51348-51375.
 - [45] WEI Y, SHAN X, MIAO R, et al. LERO: LLM-driven Evolutionary framework with Hybrid Rewards and Enhanced Observation for Multi-Agent Reinforcement Learning[C]// International Conference on Intelligent Computing. Singapore; Springer Nature Singapore, 2025: 15-26.
 - [46] SHEK C L, WU X, SUTTLE W A, et al. Lancar: Leveraging language for context-aware robot locomotion in unstructured environments[C]// 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2024: 9612-9619.
 - [47] IBRAHIM S, MOSTAFA M, JNADI A, et al. Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications [J]. IEEE Access, 2024, 12: 175473-17550.
 - [48] ESCHMANN J. Reward function design in reinforcement learning[J]. Reinforcement learning algorithms: Analysis and Appli-

cations,2021,883;25-33.

[49] WU Z,HU Y,SHI W,et al. Fine-grained human feedback gives better rewards for language model training [J]. Advances in Neural Information Processing Systems,2023,36:59008-59033.

[50] TURCATO N,IOVINO M,SYNODINOS A,et al. Towards autonomous reinforcement learning for real-world robotic manipulation with large language models[J]. arXiv:2503. 04280,2025.

[51] XU P,HUANG K Q. An efficient reinforcement learning method based on large language model[J]. Journal of Graphics,2024, 45(6):1165-1177.

[52] PARK S G,KIM H B,LEE Y J,et al. TARG: Tree of Action-reward Generation With Large Language Model for Cabinet Opening Using Manipulator[J]. International Journal of Control, Automation and Systems,2025,23(2):449-458.

[53] LI H,YANG X,WANG Z,et al. Auto mc-reward: Automated dense reward design with large language models for minecraft [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024;16426-16435.

[54] SUN S,LIU R,LYU J,et al. A large language model-driven reward design framework via dynamic feedback for reinforcement learning[J]. Knowledge-Based Systems,2025,326:114065.

[55] MA Y J,LIANG W,WANG G,et al. Eureka: Human-level reward design via coding large language models [C]// The Twelfth International Conference on Learning Representations (ICLR 2024). Vienna,Austria:OpenReview. net,2024.

[56] MASADOME S,HARADA T. Reward design using large language models for natural language explanation of reinforcement learning agent actions[J]. IEEJ Transactions on Electrical and Electronic Engineering,2025,20(8):1203-1211.

[57] NAZIR M S,BANERJEE C. Zero-Shot LLMs in Human-in-the-Loop RL:Replacing Human Feedback for Reward Shaping[J]. arXiv:2503. 22723,2025.

[58] SYGKOUNAS A,HAZRA R,PERSSON A,et al. REvolve:Reward Evolution with Large Language Models using Human Feedback[C]//Greeks in AI Symposium 2025. 2025.

[59] VERMA S,BOEHMER N,KONG L,TAMBE M. Balancing act:Prioritization strategies for LLM-designed restless bandit rewards [C]//Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence(UAI 2024). Rotterdam,Netherlands:AUAI Press,2024.

[60] YANG R,YAN J P,LI X. Survey of sparse reward algorithms in reinforcement learning _ theory and experiment [J]. CAAI Transactions on Intelligent Systems,2020,15(5):888-899.

[61] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe,Nevada,USA: Curran Associates Inc. ,2013;3111-3119.

[62] CHU K,ZHAO X,WEBER C,et al. Accelerating reinforcement learning of robotic manipulations via feedback from large language models[J]. arXiv:2311. 02379,2023.

[63] PLACE A. Adaptive reinforcement learning with llm-augmented reward functions[J]. Authorea Preprints,2023.

[64] CAO M,SHU L,YU L,et al. Beyond sparse rewards:Enhancing reinforcement learning with language model critique in text generation [C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing(EMNLP 2024). Miami,Florida,USA: Association for Computational Linguistics, 2024;9119-9138.

[65] SUTTON R S,MODAYIL J,DELP M,et al. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction[C]//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS). 2011.

[66] QU Y,JIANG Y,WANG B,et al. Latent reward: Llm-empowered credit assignment in episodic reinforcement learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2025;20095-20103.

[67] NAGPAL K,DONG D,MEHR N. Leveraging Large Language Models for Effective and Explainable Multi-Agent Credit Assignment[C]//Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems(AAMAS'25). Richland,SC:International Foundation for Autonomous Agents and Multiagent Systems,2025:1501-1510.

[68] KADAMALA K,CHAMBERS D,BARRETT E. Enhancing HVAC control efficiency:a hybrid approach using Imitation and reinforcement learning[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland,2024;256-270.

[69] ZHAO A,HUANG D,XU Q,et al. Expel: Llm agents are experiential learners[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024;19632-19642.

[70] KHEIRANDISH A,XU D,FEKRI F. LLM-Augmented Symbolic Reinforcement Learning with Landmark-Based Task Decomposition [C]// ICASSP 2025 — 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. New Delhi,India:IEEE,2025.

[71] MA R,LUIJKX J,AJANOVIĆ Z,et al. Explorllm: Guiding exploration in reinforcement learning with large language models [C]//2025 IEEE International Conference on Robotics and Automation(ICRA). IEEE,2025;9011-9017.

[72] PENG H J,ZHANG W Y,CHEN R H,et al. Large-language-model-driven training optimization for air-combat self-defense jamming strategies[J]. Journal of Detection & Control,2025, 47(2):71-81.

[73] KARINE K,MARLIN B M. Combining LLM decision and RL action selection to improve RL policy for adaptive interventions [J]. arXiv:2501. 06980,2025.

[74] ZHANG S,ZHENG S,KE S,et al. How can llm guide rl? a value-based approach[J]. arXiv:2402. 16181,2024.

[75] WANG B,MCKEOWN K,YING R. DYSTIL: Dynamic Strategy Induction with Large Language Models for Reinforcement Learning[J]. arXiv:2505. 03209,2025.

[76] WANG X S,WANG R R,CHENG Y H. Safe reinforcement learning: A survey[J]. Acta Automatica Sinica, 2023, 49(9): 1813-1835.

[77] BRIJEN T ,ASHWIN B,SURAJ N,et al. Recovery rl:Safe reinforcement learning with learned recovery zones[J]. IEEE Robotics and Automation Letters,2021,6(3):4915-4922.

[78] YANG T Y,HU M Y,CHOW Y,et al. Safe reinforcement learning with natural language constraints[J]. Advances in Neural Information Processing Systems,2021,34:13794-13808.

[79] GROSS D,SPIEKER H. Enhancing rl safety with counterfactual llm reasoning[C]// IFIP International Conference on Testing Software and Systems. Cham: Springer Nature Switzerland, 2024:23-29.

[80] MA J,LIANG W,WANG H J,et al. Dreureka:Language model guided sim-to-real transfer[J]. arXiv:2406.01967,2024.

[81] PETRONI F,ROCKTÄSCHEL T,RIEDEL S,et al. Language models as knowledge bases? [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. EMNLP,2019.

[82] PANG J C,YANG S H,LI K,et al. Kalm: Knowledgeable agents by offline reinforcement learning from large language model rollouts[J]. Advances in Neural Information Processing Systems,2024,37:126620-126652.

[83] BHAMBR I,SIDDHAN T,AMRITA BHATTACHARJE E,et al. Efficient reinforcement learning via large language model-based search [C]// NeurIPS 2024 Workshop on Open-World Agents. 2024.

[84] FRANÇOIS-LAVET V,HENDERSON P,ISLAM R,et al. An introduction to deep reinforcement learning[J]. Foundations and Trends © in Machine Learning,2018,11(3/4):219-354.

[85] WANG J,WANG T,CAI W,et al. Boosting Efficient Reinforcement Learning for Vision-and-Language Navigation With Open-Sourced LLM [J]. IEEE Robotics and Automation Letters, 2025,10(1):612-619.

[86] CHEN Z,LENG B,LI Z,et al. HCRMP: A LLM-Hinted Contextual Reinforcement Learning Framework for Autonomous Driving[J]. arXiv:2505.15793,2025.

[87] WANG X,REN H,TAN G Z,et al. Autonomous Driving Decision-making Method Based on Cooperative Reinforcement Learning of Large Language Model[J]. Journal of Transportation Systems Engineering and Information Technology, 2025, 25(4):137-146.

[88] XU C,LIU J,GUO Y,et al. Towards human-centric autonomous driving:a fast-slow architecture integrating large language model guidance with reinforcement learning [J]. arXiv: 2505.06875,2025.

[89] XU Y,HONG W,ZHA J,et al. Scalable UAV Multi-Hop Networking via Multi-Agent Reinforcement Learning with Large Language Models[J]. arXiv:2505.08448. 2025.

[90] HUANG L,YU W,MA W,et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions [J]. ACM Transactions on Information Systems,2023,43:1-55.

[91] LEWIS P,PEREZ E,PIKTUS A,et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems,2020,33:9459-9474.

[92] HUANG W,LI H,QIN F,et al. MFA-NRM: A novel framework for multimodal fusion and semantic alignment in visual neural decoding[J]. Information Fusion, 2025, 127 (Part A): 103717.

[93] KAPLAN J,MCCANDLISH S,HENIGHAN T,et al. Scaling laws for neural language models[J]. arXiv:2001.08361,2020.



ZHOU Xin, born in 2002, postgraduate. His main research interest is deep reinforcement learning.



CAO Lei, born in 1965, professor, Ph.D supervisor. His main research interests include command information system, machine learning and intelligent decision making, etc.

(责任编辑:柯颖)