大语言模型强化学习与推理可靠性研究获进展

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/41122.html

大语言模型强化学习与推理可靠性研究获进展
。近日，中国科学院合肥物质科学研究院等在大语言模型强化学习与推理可靠性方面取得两项成果，从因果信息论角度严格证明过程奖励模型（PRM）相比于结果奖励模型（ORM）的优势，并提出大模型强化评测基准体系——棱镜
基准OPG-D^3
。两项成果从类脑信息处理与认知控制的角度，为理解模型推理行为的可靠性提供了新的理论框架和评测范式。

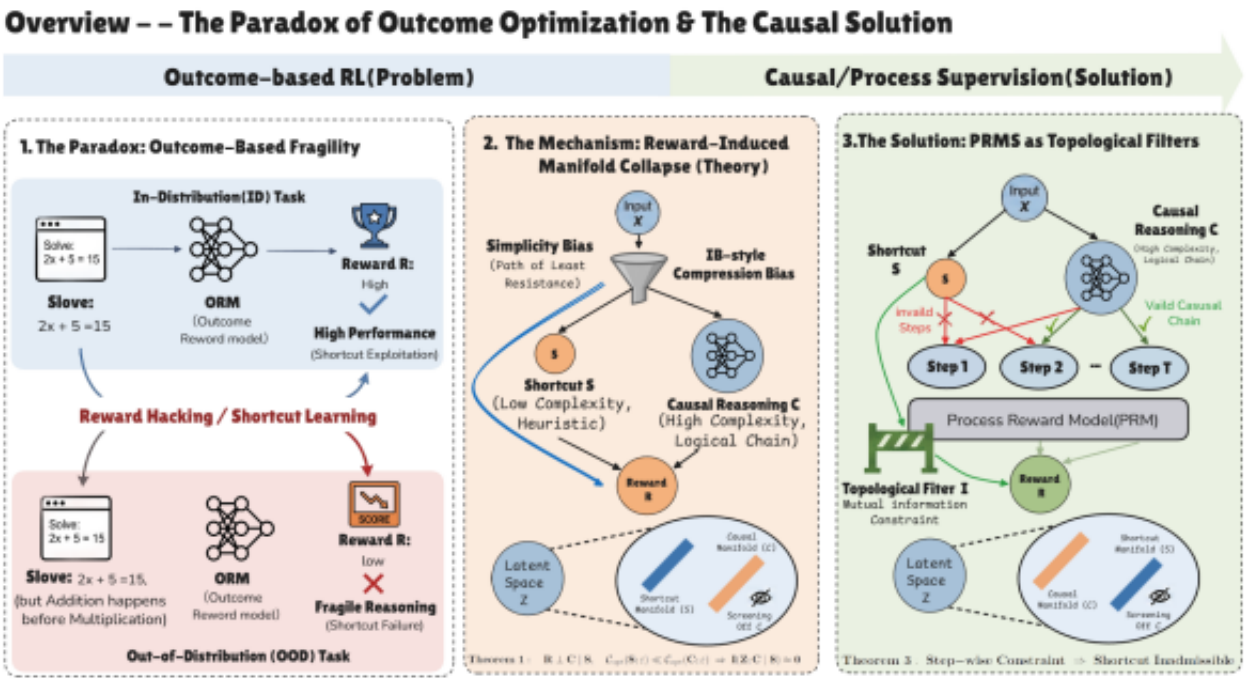
在基于结果奖励的强化学习下，模型会天然地偏好学习训练数据中的表面统计相关性，而非真正的因果推理链。这一现象与人类认知中的“启发式捷径”相似——大脑在面对复杂问题时，也倾向于调用低能耗的联想记忆而非费力的逻辑推演。受此启发，团队将过程奖励模型类比为前额叶皮层对推理步骤的在线监控机制。通过逐步骤的稀疏奖励信号，过程奖励模型能够像认知控制一样，对每一步推理施加约束，抑制“走捷径”倾向，引导模型逐步构建更稳健的因果链条。研究证实，单纯扩大同质数据规模无法从根本上消除这一脆弱性，唯有对推理过程施加结构化监督，才能实现从“答案匹配”到“因果掌握”的质变。

团队
剖析了当前强化学习评测基准的固有局限
，提出了OPG
指标——衡
量模型在训练集上训练
与直接在测试集上训练的性能差异。实验显示
，在主流基准上，强化学习模型的OPG近乎为零，表明当前训练—
测试划分可能低估了模型的泛化失败风险。以难度测试、泛化测试和反事实扰动测试为核心的基准框架OPG-D^3
，能够针对性衡量当前模型在跨分布、规则突变等场景下性能表现、陌生情境下的泛化能力。该成果为构建面向真实世界“认知压力”的下一代世界模型评测基准，提供了可操作的设计原则和诊断工具。

两项研究共同构成了大模型训练推理和评测创新体系——模型推理的可靠性不仅取决于数据量和参数规模，更依赖于学习目标与认知控制结构的匹配程度。团队将强化学习中的结果—过程矛盾上升为计算理论
层面的一般性问题，并提出因果
信息论边界的棱镜基准OPG-D^3，为后续世界模型评测提供了可验证、可比较的理论参照系。

相关成果被国际自然语言处理领域顶级会议ACL 2026接收发表。

论文链接：1、2



结果奖励诱导捷径与过程监督纠正机制示意图

当前大语言模型强化学习评测基准局限性示意图

研究团队单位：合肥物质科学研究院