研究提出基于信息论的大模型强化学习微调框架

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/36204.html

研究提出基于信息论的大模型强化学习微调框架
。近日，中国科学院软件研究所研究团队聚焦大语言模型（LLMs）在复杂推理任务中的优化问题，提出了基于信息论的强化微调框架Learning to Think（L2T），旨在平衡模型的推理效果和效率，为大语言模型在实际应用中的推理优化提供了新的技术路径。

随着LLMs能力提升，其应用场景已从基础自然语言处理任务，扩展到需要多步逻辑推理的复杂问题。分析发现，对于复杂推理任务，现有LLMs多以推理计算的最终结果为奖励信号，缺乏对中间推理步骤的及时反馈，使模型产生冗余计算，造成资源浪费，甚至可能降低推理效果。

针对上述问题，L2T框架进行了问题重构，将推理过程建模为多回合层次化对话，同时引入基于信息论的稠密过程奖励机制。该机制通过评估每一推理回合带来的信息增益，并采用改进的GRPO算法策略对大语言模型进行优化，鼓励有理推理步骤、抑制冗余生成，从而实现对推理路径的精细化调控，提升推理质量和效率。

通过AIME、AMC和HumanEval等推理基准测试，L2T在不同规模的基础模型如DeepScaleR-1.5B-Preview、DeepSeek-R1-Distill-Qwen-1.5B上，均表现出稳定的性能提升。结果显示，与基于结果奖励的方法相比，L2T在准确率上提升超过3.2%，同时token效率翻倍；与基于过程奖励的基线相比，L2T在准确率上仍有约2%的提升，效率提高约1.2倍。同时，在多任务评估中，L2T在不同难度任务上实现了平均近3%的准确率提升，并在不同token预算下均保持稳定的性能优势。

相关论文发表在人工智能领域顶级会议NeurIPS 2025上。

论文链接

研究团队单位：软件研究所