基于微调大语言模型与多Agent协作的数学推理工作流系统研究

摘 要

本研究聚焦于大语言模型（LLM）推理能力的定向优化与多智能体协同系统构建，研究从 LLM 的思维链（CoT）推理机制切入，重点探索模型推理能力的定向优化策略。研究采用递进式实验框架，分阶段探索模型优化策略：第一阶段以 Qwen2-1.5B-Instruct 为基础模型，实施监督微调（SFT）训练。通过构建双分支训练路径 —— 基础推理模型分支与知识增强分支，分别利用公开 CoT 数据集完成参数优化，并采用 MMLU、GSM8K、HUMANEVAL 基准集进行多维度能力评估。在此基础上，引入混合专家模型（MoE）架构开展模型容量扩展实验。第二阶段以 Qwen2.5 轻量级模型为主体，构建 "知识蒸馏 - 强化学习" 二级优化研究体系：首先通过 DeepSeek R1 蒸馏数据集实施知识蒸馏（KD），生成 R1-1.5B 基础模型；后续再基于基座模型尝试复现R1的训练路线，采用群体相对策略优化算法（GRPO）开展强化学习微调，构建多参数量级的推理模型。最终通过整合自研模型与主流 API 模型，基于 Dify 开源框架开发原型系统，实现数学推理能力的场景化验证。本研究在方法论层面提出了 "监督训练 - 知识迁移 – 强化学习策略优化" 的三级优化范式，在技术层面验证了 GRPO 算法在数学推理任务中的有效性，为领域特定 LLM 的优化提供了可复现的技术路径。

**关键词：大语言模型，推理能力，有监督微调，知识蒸馏，强化学习**

Abstract

This research focuses on the targeted optimization of reasoning capabilities in large language models (LLMs) and the construction of multi-agent collaborative systems. Starting from the chain-of-thought (CoT) reasoning mechanism of LLMs, it explores optimization strategies for model reasoning capabilities. The study adopts a progressive experimental framework with phased optimization approaches:

**Phase I** utilizes Qwen2-1.5B-Instruct as the base model for supervised fine-tuning (SFT) training. By establishing dual-branch training paths—a foundational reasoning model branch and a knowledge-enhanced branch—the research optimizes parameters using public CoT datasets and conducts multi-dimensional capability evaluations via benchmarks such as MMLU, GSM8K, and HUMANEVAL. Additionally, experiments on model capacity expansion are performed through the introduction of a mixture-of-experts (MoE) architecture.

**Phase II** constructs a two-stage optimization system of "knowledge distillation (KD)–reinforcement learning" using Qwen2.5 lightweight models. First, knowledge distillation is implemented with the DeepSeek R1 distillation dataset to generate the R1-1.5B base model. Subsequently, group relative policy optimization (GRPO) is applied to the base model for reinforcement learning fine-tuning, resulting in multi-scale reasoning models. Finally, an integrated prototype system is developed based on the Dify open-source framework, combining self-developed models and mainstream API models to validate mathematical reasoning capabilities in practical scenarios.

Methodologically, this research proposes a three-tier optimization paradigm of "supervised training–knowledge transfer–reinforcement learning policy optimization." Technically, it validates the effectiveness of the GRPO algorithm in mathematical reasoning tasks and provides a reproducible technical path for domain-specific LLM optimization.

**Keywords： Large Language Model，Reasoning Capabilities，Supervised Fine-Tuning， Knowledge Distillation，Reinforcement Learning**

**目录**

# 研究目的和意义

近年来，随着人工智能技术的快速演进，大语言模型（LLM）在自然语言处理领域展现出强大潜力，但在涉及复杂逻辑推理任务（如数学推理）中仍存在逻辑连贯性和可解释性不足、领域适应性有限等问题，而近期推理模型的研究逐渐成为了该领域的热门课题。在现实场景中，一个合格的推理模型应该能够解决结合具体上下文语境的复杂问题以及部分抽象概念的数学证明，并能够生成一定规模和可解释性的思维链，这要求模型具备极强的推理深度和跨领域的知识储备。本研究立足当前技术发展需求，针对 LLM 推理能力的核心瓶颈，通过构建 “监督训练 - 知识迁移 – 强化学习策略优化”的三级技术框架，探究如何让普通模型通过递进式训练体系提升模型的推理深度与跨领域知识整合能力。

本研究以数学推理场景为切入点，通过递进式实验设计：首先构建双分支监督训练体系，结合混合专家模型（MoE）实现模型容量动态扩展；继而采用知识蒸馏（KD）技术压缩模型规模，引入群体相对策略优化算法（GRPO）提升推理逻辑的合理性；最终通过多智能体系统整合不同模型优势，实现复杂数学问题的分步推理与协同决策。研究旨在突破现有模型在特定领域的推理性能瓶颈，为构建可解释、高效能的领域专用推理LLM 提供一定程度上的方法论参考，同时为多智能体协同解决复杂问题提供示例的简单技术范式。