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和多智能体协同系统的构建。研究的核心目标是从大语言模型的思维链（Chain of Thought, CoT）推理机制入手，探索并优化其推理能力，进而提升模型在复杂任务中的表现。我们采用了递进式的实验框架，分阶段进行系统化优化探索，以确保在多个层面上对模型性能进行全面的提升。

第一阶段，以 Qwen2-1.5B-Instruct 作为基础模型，实施监督微调（Supervised Fine-Tuning, SFT）。在此基础上，我们设计了双分支训练路径，其中一分支专注于基础推理能力的训练，另一分支则专注于知识增强。通过构建并利用公开的 CoT 数据集，分别对这两条分支进行了微调，并通过 MMLU、GSM8K 和 HUMANEVAL 等标准基准集对产出的模型进行了多维度评估。此外，为探究如何提升模型推理性能，研究引入了混合专家模型（Mixture of Experts, MoE）架构，探索了模型容量的扩展方案，以提升大模型在推理任务中的泛化能力。

第二阶段以 Qwen2.5-1.5B-Instruct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 蒸馏数据集进行知识蒸馏（Knowledge Distillation, KD），以此数据为基础微调产出了两版 R1-Distlled-1.5B 基础模型；后续再基于Qwen2.5基座模型以及Deepseek官方蒸馏模型尝试复现R1-Zero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，最终得到两版的性能表现均有一定提升的轻量级推理模型。

最终通过整合自研模型与主流 API 模型，基于 Dify 开源框架开发一个简单的演示框架进行应用层的封装，实现数学推理能力的场景化验证。

本研究不仅提出了一个基于 "监督训练 - 知识迁移 - 强化学习策略优化" 的三级优化范式，还在技术层面验证了 GRPO 算法在推理任务中的有效性。研究结果表明，该范式不仅能显著提升模型的推理能力，还为领域特定的大语言模型的优化提供了一个具有可复现性的技术路径，具有重要的学术价值和实践意义。

**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 development of multi-agent collaborative systems. The primary objective is to explore and optimize the reasoning abilities of LLMs through the Chain of Thought (CoT) reasoning mechanism, thereby enhancing model performance on complex tasks. A progressive experimental framework is adopted, involving phased optimization to ensure comprehensive performance improvement at multiple levels.

In the first phase, Qwen2-1.5B-Instruct is used as the base model for supervised fine-tuning (SFT). A dual-branch training approach is designed, with one branch focused on basic reasoning and the other on knowledge enhancement. The two branches are fine-tuned using publicly available CoT datasets, and the resulting models are evaluated across multiple dimensions using benchmarks such as MMLU, GSM8K, and HUMANEVAL. To further explore ways to improve reasoning performance, a Mixture of Experts (MoE) architecture is introduced to investigate model capacity expansion and improve the generalization ability of large models on reasoning tasks.

In the second phase, a two-stage optimization system of "knowledge distillation (KD) – reinforcement learning" is constructed using Qwen2.5-1.5B-Instruct as the core. Knowledge distillation is first performed using the DeepSeek R1 distillation dataset, resulting in two versions of the R1-Distilled-1.5B base model. Subsequently, training routes from R1-Zero are replicated using the Qwen2.5 base model and the DeepSeek official distillation model, with reinforcement learning fine-tuning conducted via the Group Relative Policy Optimization (GRPO) algorithm. The final models show performance improvements in lightweight reasoning tasks.

A prototype system is developed by integrating in-house models with mainstream API models, using the Dify open-source framework to create an application layer for the scene-based validation of mathematical reasoning capabilities.

This research not only proposes a three-tier optimization paradigm of "supervised training – knowledge transfer – reinforcement learning policy optimization" but also validates the effectiveness of the GRPO algorithm in reasoning tasks. The findings demonstrate that this paradigm significantly enhances the reasoning abilities of LLMs and provides a reproducible technical pathway for optimizing domain-specific models, with notable academic and practical implications.

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

**目录**

# 引言

## 1.1 研究背景及意义

近年来，大语言模型（LLM）[1]在自然语言处理（NLP）任务中的表现取得了显著突破，尤其在问答、推理、代码生成等任务中展现出强大的泛化能力。然而，尽管 LLM 具备一定的推理能力，其在复杂逻辑推理、数学计算、多步骤问题求解等方面仍存在明显局限性。推理能力不足不仅影响模型的可解释性和可靠性，也限制了其在专业领域的应用。因此，如何针对性地优化 LLM 的推理能力，提升其在复杂任务中的表现，已成为当前学术界和工业界广泛关注的研究热点。

为提升 LLM 的推理能力，近年来涌现出诸多方法，如**思维链推理（Chain of Thought, CoT）**、**自回归推理增强**、**检索增强生成（RAG）** 以及**基于强化学习的策略优化[2-4]**等。其中，CoT 通过引导模型进行逐步推理，使其在多步骤推理任务中具备更好的逻辑一致性，已被广泛应用于数学推理、代码生成等任务。然而，当前主流的 CoT 方法多依赖于大规模预训练和少量的指令微调，缺乏对推理能力的定向优化。此外，由于 LLM 训练过程中计算资源需求巨大，研究如何利用知识蒸馏（Knowledge Distillation, KD）和强化学习（Reinforcement Learning, RL）[5, 6]等技术，以更轻量级的方式提升模型推理能力，具有重要的研究价值。

本研究针对 LLM 的推理能力优化问题，结合**监督微调（Supervised Fine-Tuning, SFT）[7]**、**知识蒸馏（KD）** 和 **强化学习（RL）**，提出了一种“监督训练 - 知识迁移 - 强化学习策略优化”的三级优化范式。通过构建多智能体协同优化体系，本研究探索了如何在不同参数量级的模型上增强推理能力，并采用混合专家模型（MoE）架构[8]扩展模型容量，以进一步提升其在多任务环境下的泛化能力。此外，研究还基于 Dify [9] 开源框架开发数学推理能力验证系统，探索 LLM 在应用场景中的可行性。

本研究的意义体现在以下几个方面：

1. **优化 LLM 的推理能力**：针对 LLM 在复杂逻辑推理任务中的局限性，提出了一种系统化的优化方法，通过指令微调嵌入思维模型和同构强化学习令模型自发的拥有推理能力，提升模型在数学推理、多步骤推理等任务上的表现。
2. **提出可复现的优化范式**：结合 SFT、KD 和 RL 等技术，构建了一种可复现的优化框架，为不同参数规模的 LLM 训练提供技术路径参考。
3. **提升轻量级模型的推理性能**：通过知识蒸馏与强化学习微调，使小型模型在推理能力上接近甚至超越更大规模的模型，从而降低计算成本，提高部署效率。
4. **验证 GRPO 算法的有效性**：在强化学习优化环节，引入群体相对策略优化（GRPO）算法，并在推理型LLM 的微调任务中进行了实验验证，进一步拓展了该算法在大模型训练中的应用价值。

本研究不仅尝试为 LLM 的推理能力优化提供了一条高效、可复现的技术路径，同时也为轻量级 LLM 的推理增强、强化学习在大模型训练中的应用等方向提供了相关思路，在一定程度上具有重要的学术价值和应用前景。

## 1.2 研究内容及创新点

### 1.2.1 研究内容

本研究围绕大语言模型（LLM）推理能力的定向优化和多智能体协同系统的构建展开，具体研究内容可归纳为以下三个阶段：

**第一阶段：基础推理能力的优化与模型扩展**

在此阶段，研究以 Qwen2-1.5B-Instruct [10] 为基础模型，实施监督微调（Supervised Fine-Tuning, SFT）。通过设计双分支训练路径，分别优化基础推理能力和知识增强能力。第一条分支专注于基础推理能力的训练，利用公开的 CoT 数据集进行微调；另一条分支则针对知识增强，通过相应的优化策略提高模型的知识整合能力。在此基础上，研究通过 MMLU、GSM8K 和 HUMANEVAL等基准数据集[11-13]对优化后的模型进行多维度评估，确保模型的推理能力在不同任务上的有效性。进一步地，为探索如何提升模型推理性能，研究引入了混合专家模型（Mixture of Experts, MoE）架构，以扩展模型容量，并提升模型在多任务推理中的泛化能力。

**第二阶段：轻量级模型优化与强化学习微调**

第二阶段的研究以 Qwen2.5-1.5B-Instruct [14] 为核心，构建了一个“知识蒸馏 - 强化学习”二级优化体系。首先，研究通过 DeepSeek R1[15] 蒸馏数据集进行知识蒸馏（Knowledge Distillation, KD），产生了两个 R1-Distilled-1.5B 基础模型。随后，基于 Qwen2.5 基座模型和 DeepSeek 官方蒸馏模型，研究复现了 R1-Zero 的训练路线，并采用群体相对策略优化算法（Group Relative Policy Optimization, GRPO）[16]进行强化学习微调。此阶段的目标是进一步优化轻量级推理模型，在保证模型推理性能的同时，通过少量训练资源达到更高效的优化效果。

**第三阶段：自研与主流 API 模型整合与应用验证**

在最终的系统验证阶段，研究将自研的优化模型与主流 API 模型进行整合，并基于 Dify 开源框架开发了一个简单的演示框架。该框架用于封装推理能力模块，并实现数学推理任务的场景化验证。通过该验证系统，评估模型在实际应用中的表现，进一步证明了所提优化方法在真实场景中的可行性和有效性。

### 1.2.2 创新点

在本研究工作中，研究创新方面可以总结为以下五点：

1. 推理能力定向优化框架

本研究提出了一种基于“监督训练 - 知识迁移 - 强化学习策略优化”的三级优化范式，通过精细化的模型微调和知识增强策略，显著提升大语言模型（LLM）的推理能力。这一优化框架系统性地结合了监督微调、知识蒸馏与强化学习等多种技术，为模型推理性能的提升提供了新的思路。

1. 双分支训练路径的设计

研究创新性地提出了双分支训练路径，其中一分支专注于基础推理能力的提升，另一分支则专注于知识增强。通过对 CoT 数据集的利用，双分支路径能在不同层次上强化模型的推理能力，为复杂任务提供更加精细的优化手段。

1. 混合专家模型（MoE）架构的引入

在推理能力提升的过程中，研究引入了混合专家模型（Mixture of Experts, MoE）架构。这一架构能够在模型容量扩展的同时提高计算效率，进而优化大模型的推理性能，提升其在多任务环境下的泛化能力。

4. 轻量级推理模型的优化与强化学习微调优化的探索

本研究通过知识蒸馏（KD）与群体相对策略优化（GRPO）算法，进行了两种高效的轻量级推理模型优化方案的尝试。通过蒸馏技术将思维链能力通过指令微调的方式嵌入模型生成权重中，以及进行探索通过强化学习微调让模型自发的拥有长推理思维俩能力，两种方法都成功提升了小模型在推理任务中的表现，实现了推理能力和计算效率的平衡。

1. GRPO算法在推理模型训练中的应用验证

本研究通过在LLM 训练任务中验证了群体相对策略优化（GRPO）算法的有效性。通过强化学习优化模型推理策略，GRPO 在提升模型表现的同时，还能有效降低训练资源的消耗，拓展了该算法在大规模模型训练中的应用场景。

[1] 文森, 钱力, 胡懋地, et al. 基于大语言模型的问答技术研究进展综述 [J]. 2024, 1.

[2] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks [J]. 2020, 33(9459-74.

[3] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [J]. 2017,

[4] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. 2022, 35(24824-37.

[5] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: A survey [J]. 2021, 129(6): 1789-819.

[6] KAELBLING L P, LITTMAN M L, MOORE A W J J O A I R. Reinforcement learning: A survey [J]. 1996, 4(237-85.

[7] CHEN T, KORNBLITH S, SWERSKY K, et al. Big self-supervised models are strong semi-supervised learners [J]. 2020, 33(22243-55.

[8] LI J, WANG X, ZHU S, et al. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts [J]. 2024, 37(131224-46.

[9] ARAI K J I J O A C S, APPLICATIONS. Design of On-Premises Version of RAG with AI Agent for Framework Selection Together with Dify and DSL as Well as Ollama for LLM [J]. 2024, 15(12):

[10] YANG A, YANG B, HUI B, et al. Qwen2 technical report [J]. 2024,

[11] LIU B, BUBECK S, ELDAN R, et al. Tinygsm: achieving> 80% on gsm8k with small language models [J]. 2023,

[12] SINGH S, ROMANOU A, FOURRIER C, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation [J]. 2024,

[13] ZHENG Q, XIA X, ZOU X, et al. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x; proceedings of the Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, F, 2023 [C].

[14] YANG A, YANG B, ZHANG B, et al. Qwen2. 5 technical report [J]. 2024,

[15] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning [J]. 2025,

[16] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models [J]. 2024,