**本科毕业论文（设计）**

论文题目：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

指导老师： 王 曼

学生姓名： 陈 柏 政

学 号： 20214001042

学 院： 人工智能学院

专 业： 人工智能

毕业时间： 2025年6月

**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，同时采取不同的上下文长度进行模型的优化提升，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

**Research on the Mathematical Reasoning Workflow System Based on Fine-Tuned Large Language Models and Multi-Agent Collaboration**

**Abstract**

This research focuses on the targeted optimization of reasoning capabilities in large language models (LLMs) and the construction of multi-agent collaborative systems. The core objective is to explore and optimize the reasoning abilities of LLMs starting from their Chain of Thought (CoT) reasoning mechanism, aiming to enhance model performance in complex tasks. A progressive experimental framework is adopted, with a phased optimization approach to systematically improve model performance across 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 strategy is designed, with one branch dedicated to foundational reasoning and the other to knowledge enhancement. The two branches are fine-tuned using publicly available CoT datasets, and the resulting models are evaluated across multiple dimensions using benchmark sets such as MMLU, GSM8K, and HUMANEVAL. Additionally, a Mixture of Experts (MoE) architecture is introduced to explore model capacity expansion, enhancing the generalization ability of large models in reasoning tasks.

In the **second phase**, a two-stage optimization system combining "knowledge distillation (KD) - reinforcement learning" is constructed using Qwen2.5-1.5B-Instruct as the core model. Knowledge distillation is first applied using the DeepSeek R1 distillation dataset to generate 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 distilled model, with reinforcement learning fine-tuning applied using the Group Relative Policy Optimization (GRPO) algorithm. The optimization process includes varying context lengths for model enhancement, resulting in two versions of lightweight reasoning models with improved performance.

Finally, a prototype system is developed by integrating self-developed 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 study proposes a three-tier optimization paradigm of "supervised training - knowledge transfer - reinforcement learning policy optimization" and validates the effectiveness of the GRPO algorithm in reasoning tasks. The findings demonstrate that this paradigm significantly improves the reasoning capabilities of models and provides a reproducible technical pathway for optimizing domain-specific LLMs, with significant academic and practical implications.

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

目 录

[**摘 要** 3](#_Toc193316749)

[**Abstract** 5](#_Toc193316750)

[第一章 引言 8](#_Toc193316751)

[1.1 研究背景及意义 8](#_Toc193316752)

[1.2 研究内容及创新点 9](#_Toc193316753)

[1.2.1 研究内容 9](#_Toc193316754)

[1.2.2 创新点 10](#_Toc193316755)

[1.3 11](#_Toc193316756)

# 第一章 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：

**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。

**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。

**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。

**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。

**2. 双分支训练路径的设计**，本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。

**3. 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参数中，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链能力，在第二种尝试中，采用混合上下文长度的训练方法，在一定程度上可以进一步优化模型针对复杂任务的推理能力，最终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。

**5. 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。

## 1.3 本文组织结构

本文共有五个章节组成，每个章节的主要内容如下：

第一章，引言。在本章节中，首先介绍了推理模型的研究背景和目的，明确了研究的核心任务和目标，阐述了推理能力优化与多智能体协同系统构建的重要性。紧接着结合当前发展现状，概述了本文的研究内容和创新点，详细描述了研究过程中提出的技术框架及优化方法。

第二章，国内外研究现状。在此章节中，主要探讨了了在本研究工作中涉及的相关技术研究，包括三个方面：大语言模型的研究，推理模型的相关发展历程以及模型训练技术的研究。与此同时，对相关研究的代表技术和算法进行详细分析，并总结了不同代表算法的技术特点。

第三章，相关技术研究和实验设计。针对本研究涉及的相关技术进行详细阐述，包括但不限于大模型架构分析，有监督微调方法解析以及强化学习算法分析。此外，详细分析了本研究的实验设计整体方案，为模型微调以及推理模型训练进一步探索提供了一定的思路。

第四章，数据处理和模型训练。本章首先介绍了本研究的基础设施和硬件配置，重点介绍了多种开发工具的使用。然后，详细说明了数据的格式清洗与预处理过程，并介绍了在各个阶段所采用的评测指标。

第五章，实验结果展示与分析。本章详细展示并分析了在第一阶段和第二阶段实验中取得的结果。包括对各阶段模型的性能评估、损失函数变化、KL散度、以及生成文本长度等指标的变化进行分析，重点展示了不同版本模型的性能提升情况，探讨了模型优化过程中的主要挑战与解决方案。

第六章，Dify应用模版的搭建与演示**。**本章介绍了基于Dify开源框架开发大模型的简单应用模版，并进行演示。通过将自研模型与主流API模型整合，展示了如何在实际场景中应用优化后的推理模型，尤其是在数学推理等具体任务中的应用验证。该部分展示了模型的实际应用效果，证明了优化方法在真实环境中的可行性与有效性。

# 第二章 国内外研究现状

## 2.1 大语言模型发展概述

大语言模型（LLM, Large Language Models）是近年来人工智能领域的一个重要研究方向，其核心目标是通过大规模数据和强大的计算能力，构建能够理解和生成自然语言的模型。

大语言模型的崛起源自谷歌Vaswani等人于2017提出的Transformer架构的提出及其注意力机制（Attention Mechanism）[17]的成功应用，Transformer的核心创新在于引入了自注意力机制（Self-Attention），该机制能够在处理序列数据时捕捉输入序列中各个位置之间的依赖关系，而不需要传统的循环神经网络（RNN）或长短期记忆网络（LSTM）[18]中的逐步迭代。这种机制通过计算输入序列中每个元素与其他元素的相关性，从而实现全局信息的捕捉，相比于传统方法具有显著的并行计算优势。Transformer的关键组件包括**多头注意力**（Multi-Head Attention）、**位置编码**（Positional Encoding）和**前馈神经网络**（Feedforward Neural Networks），这些都为模型的深度学习和并行计算奠定了基础。自注意力机制不仅提升了模型的训练效率，还增强了模型处理长程依赖的能力，使其能够在自然语言处理（NLP）任务中表现出色。

在Transformer架构的基础上，BERT（Bidirectional Encoder Representations from Transformers）[19]实现了一个重大的突破，提出了一种新的预训练和微调方法。BERT采用了Transformer的Encoder部分，最显著的特点是其双向性。与传统的单向语言模型（例如左到右或右到左的语言模型）不同，BERT通过在预训练阶段进行双向上下文的学习，能够在理解文本时同时利用左右两侧的信息，从而提高了对句子整体语义的理解能力。BERT采用了两种主要的预训练任务：Masked Language Model（MLM）和Next Sentence Prediction（NSP）[20]。MLM任务通过遮蔽输入序列中的一些词汇，让模型预测这些被遮蔽的词，而NSP任务则帮助模型学习句子之间的关系。BERT的成功标志着预训练模型的时代来临，预训练加微调（Pretrain-Finetune）[21]的方法成为了大规模NLP模型的标准训练流程。

与BERT不同，GPT（由OpenAI提出）更侧重于生成任务。GPT系列模型采用了Transformer的Decoder部分，并且利用了自回归的方式生成文本。自回归模型的特点是，模型通过逐步生成下一个词汇，并利用前面的词汇生成后续的内容，这种方式特别适合生成任务，如文本生成、机器翻译等。GPT的第一个版本（GPT-1）是一个相对较小的模型，但随着后续版本的发布（GPT-2、GPT-3、GPT-4）[22-25]，模型的规模大幅度提升，参数数量也从1.1亿增加到1750亿最后到到现在的闭源商业模型，并在多个NLP任务上展示出了卓越的性能。GPT的成功关键在于其庞大的参数量和超大规模的训练数据，使得模型具备了强大的语言理解和生成能力。GPT-3的发布进一步推动了大语言模型的研究热潮，其生成的文本不仅在语言流畅性和内容连贯性上表现出色，且能够应对许多复杂的推理和生成任务。GPT系列强调的是无监督预训练和零-shot学习能力，即模型在没有专门任务微调的情况下，能够处理各种NLP任务。随着后续性能表现更加强大的GPT-4O以及各种新型AI产品的到来，大语言模型逐渐在各个领域上展露具有无限可能的发展潜力。

LLaMA是Meta（前Facebook）提出的一个新型大语言模型系列[26]，旨在通过较小的计算资源实现较大的性能突破。LLaMA的核心创新之一是其对不同规模模型的训练进行了优化同时在模型架构上进行了优化，例如使用RMS Norm 归一化函数，使用SwiGLU替代了ReLU作为激活函数，利用ROPE通过绝对位置编码的方式实现相对位置编码[27]，采用了高效的参数量与计算资源的平衡，使得较小的LLaMA模型能够在处理复杂任务时仍保持较高的性能。LLaMA系列在数据处理、模型训练和推理效率上做出了创新，其主要目标是通过在更少的计算资源上训练出更强的语言能力，进而为广泛的AI应用提供更高效的解决方案。LLaMA在多个标准基准上展示了优异的性能，特别是在低资源设备上，能够有效进行推理任务。近年来也在研究领域不断开源分享相关模型和技术，为大语言模型发展奠定了一定的基础。

而国内的研究团队也是紧跟热点，随着阿里推出Qwen系列模型，结合了最新的强化学习和优化技术，通过引入更加先进的思维链（Chain of Thought, CoT）推理机制，大幅度提升了模型在复杂推理任务中的表现。Qwen的核心优势在于其针对特定任务的定向优化策略，包括在基础推理能力和知识增强方面的双分支训练路径。通过这种优化策略，Qwen能够在多个NLP和推理任务中展现出强大的性能，并且具有更好的泛化能力和计算效率。Qwen系列的推出，标志着国内大语言模型在推理能力和技术创新方面的快速发展，进一步推动了大语言模型技术的多元化应用。与此同时，国内其他研究团队也相继分享开源了自行研发训练的国产化大模型，如Baichuan、InternLM[28, 29]等，这些模型都针对中文语言做了相关的优化和数据处理，更适合国内用户开发使用。

随着模型规模和技术的不断进步，未来的大语言模型将在多个领域展示出更强的表现，推动NLP和AI应用的广泛落地。

## 2.2 推理大模型发展概述

[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,

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. 2017, 30(

[18] SHERSTINSKY A J P D N P. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network [J]. 2020, 404(132306.

[19] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?; proceedings of the ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, F, 2019 [C].

[20] GREZES F, BLANCO-CUARESMA S, ACCOMAZZI A, et al. Building astroBERT, a language model for astronomy & astrophysics [J]. 2021,

[21] TAY Y, DEHGHANI M, RAO J, et al. Scale efficiently: Insights from pre-training and fine-tuning transformers [J]. 2021,

[22] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report [J]. 2023,

[23] ZHANG M, LI J J F R. A commentary of GPT-3 in MIT Technology Review 2021 [J]. 2021, 1(6): 831-3.

[24] FLORIDI L, CHIRIATTI M J M, MACHINES. GPT-3: Its nature, scope, limits, and consequences [J]. 2020, 30(681-94.

[25] LEE J-S, HSIANG J J W P I. Patent claim generation by fine-tuning OpenAI GPT-2 [J]. 2020, 62(101983.

[26] LI Y, LI Z, ZHANG K, et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge [J]. 2023, 15(6):

[27] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models [J]. 2022,

[28] YANG A, XIAO B, WANG B, et al. Baichuan 2: Open large-scale language models [J]. 2023,

[29] CHEN Z, WU J, WANG W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2024 [C].