****

**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： 2 0 2 5 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** 2](#_Toc193505510)

[**Abstract** 4](#_Toc193505511)

[第一章 引言 7](#_Toc193505512)

[1.1 研究背景及意义 7](#_Toc193505513)

[1.2 研究内容及创新点 8](#_Toc193505514)

[1.2.1 研究内容 8](#_Toc193505515)

[1.2.2 创新点 9](#_Toc193505516)

[1.3 本文组织结构 10](#_Toc193505517)

[第二章 国内外研究现状 11](#_Toc193505518)

[2.1 大语言模型发展概述 11](#_Toc193505519)

[2.2 推理大模型发展概述 14](#_Toc193505520)

[2.3 大模型微调训练方法的发展 16](#_Toc193505521)

[2.4 大模型推理优化的相关技术发展 18](#_Toc193505522)

[第三章 相关技术研究和实验设计 20](#_Toc193505523)

[3.1 Transformer模型架构分析 20](#_Toc193505524)

[3.2 Qwen模型架构分析 23](#_Toc193505525)

[第七章 参考文献 23](#_Toc193505526)

# 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

本文共有六个章节组成，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和模型训练的相关信息。

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，通过采用BPE分词器扩展词表，同时在QKV的注意力机制中添加了偏差增强模型的外推能力[28]，同时使用了Flash Attention技术[29]加快整体的训练速度。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、Minicpm[30-32]等，这些模型都针对中文语言做了相关的优化和数据处理，同时尝试利用更小的参数量训练出更加强大的模型，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

推理大模型（Reasoning Large Language Models，RL-LLMs）作为大语言模型（LLM）的一个重要分支，专注于模型在复杂推理任务中的表现。推理任务通常涉及到抽象思维、逻辑推理、数学问题求解、常识推理等，这些任务不仅需要模型具备强大的语言理解能力，还要求模型在推理过程中能够做出合理的判断和推导。推理大模型的崛起源于对基础大语言模型（如BERT、GPT等）能力的拓展和优化，特别是在深层推理和推断能力上的提升。

在增强大语言模型推理能力的研究中，提示词工程、链式思维推理、少样本推理(Few-shot prompting) [33, 34]等技术日益成为提升模型表现的重要方法。这些技术通过细化问题的理解过程，再通过部分示例样本促使模型能够更系统性地分步推理，从而显著提升其在复杂任务中的表现。CoT推理通过内嵌推理链条，引导模型逐步分解任务，以实现更连贯的推理路径，尤其适用于多步骤逻辑推理与数学问题。CoT的核心思想是，在推理过程中，模型通过模拟思考链的方式逐步展开推理，而不是直接给出答案。这种方法可以显著提升模型在复杂推理任务中的表现。CoT机制通过将推理过程显式化，允许模型逐步推导出结论。这种逐步推理的方法不仅提高了模型在复杂推理任务中的准确性，也增强了模型的解释性和可控性。例如，在数学问题求解时，CoT使得模型能够分步骤推理，展示出推理过程而不仅仅是输出最终答案，从而大大提升了推理的准确性和可靠性在此基础上，少样本推理策略(如TOT , Tree of Thought) [35]通过有限的样本展示推理路径，有效减少了对大规模标注数据的依赖，同时使模型能够更灵活地应对多样化问题。

随着推理任务的日益复杂，强化学习（Reinforcement Learning, RL）被引入推理大模型的优化过程，以进一步提升推理能力。强化学习通过奖励机制引导模型逐步改进推理策略，尤其适用于需要在多轮交互或连续决策中进行优化的任务。强化学习微调（RL Fine-Tuning）[36]已经成为推理大模型中一种重要的技术，尤其是在解决诸如对话生成、复杂数学问题和推理题目时，它能够提升模型的推理准确性。DeepMind提出的AlphaCode模型[37]，通过在编程和推理任务中的强化学习微调，取得了显著的成功。AlphaCode能够通过学习大量编程题目中的输入输出模式，逐步推理并生成符合题意的代码，标志着推理大模型在代码理解和生成上的一大突破。

Numina 团队是一个专注于研究大语言模型数学能力的团队，于2024年7月参加并获得AI 数学奥林匹克 (AI Math Olympiad，AIMO) 的首届进步奖，研发的Numina-Math的数学解题模型采用两阶段训练方法[38]，第1阶段通过在自然语言“数学题 + 解答”的大规模、多样化数据集上微调基础模型，其中每个解答都需套用思维链模板以促使 LLM 进行推理。第 2 阶段则是在Python代码工具整合推理的合成数据集上微调第 1 阶段得到的模型，其中每个数学题都分解为一系列推理、Python程序及其输出。其遵循微软的ToRA [39]的做法，提示模型以ToRA格式生成带有代码执行反馈的解答。对这些数据进行微调会产生一个推理代理，它可以通过将自然语言推理和使用外部Python REPL来计算中间结果结合起来以解决数学问题。后续，阿里云团队也用类似的方法即CoT+TIR的两阶段方式强化模型的数学推理能力，训练并开源Qwen2.5-Math系列模型[40]，其中还涉及到强化学习的人类反馈微调[41]。

在2024年9月份，OpenAI正式推出了其闭源推理模型GPT-O1[42]，在处理复杂推理问题和理解长文本推理方面具备了更强的能力, GPT-O1强化了思维链推理（Chain of Thought, CoT）能力，优化了模型在多步骤推理中的表现。其通过强化学习算法引导模型逐步推导解决方案，在训练过程中引入了任务自适应的机制，针对不同类型的推理任务（如常识推理、数学推理、程序推理等），通过微调和数据集增强，O1能够在复杂推理任务中显著减少逻辑错误，并提升推理的准确性和可解释性, 为了更好地处理长文本推理任务，GPT-O1对模型的上下文理解能力进行了增强。在推理过程中，GPT-O1能够处理更多的上下文信息，优化了模型在长文推理中逐步推理和保持一致性的能力，成为了业界的SOTA模型。自此，研究具有深度推理思维链能力的推理模型逐渐开始成为各界的研究热点。

国内Deepseek-AI团队也一直在推理模型的领域发光发热，从Deepseek-Math到Deepseek-Coder [16, 43, 44]最后到利用更低的训练推理成本却能达到业界SOTA O1效果的Deepseek-R1 [15]，在研究过程中提出了群体相对策略优化（GRPO）算法，能够取消常规近端强化学习算法PPO流程中的Value Model[3]以及直接在基座模型上进行强化学习的训练思路使模型具备推理反思的能力，而不是像以往的常规训练流程一样，先通过大量高质量思维链数据加有监督微调的方式让模型扩充能力后再利用强化学习增强模型对思考过程和最后答案的优化，并在探索完推理能力的提升后，为了完善优化模型输出的可读性和可解释性，进行了高质量的普通知识数据冷启动的方式，在推理和日常交流能力中寻找一定的平衡，同时为了节省训练推理的GPU显存，利用MoE机制和MLA（Multi-Head Latent Attention）。Deepseek-R1的爆火在一定程度上也象征着我国在推理模型领域能够比肩国外的顶尖水平，甚至能够做到利用更低的时间和金钱成本进行训练推理。

## 2.3 大模型微调训练方法的发展

大模型的微调（Fine-Tuning）是深度学习研究中的一个核心课题，尤其是在自然语言处理（NLP），微调已经成为提升模型性能的主要手段。随着大模型（如GPT、BERT、T5、Qwen、LLaMA等）的不断发展和复杂度以及参数两的提升，微调训练方法也经历了显著的演进发展。大模型的微调最早来源于“预训练-微调”（Pre-training and Fine-tuning）范式[21]，这一方法的核心是首先在大规模通用数据集上进行无监督预训练，然后在特定任务上进行有监督微调。BERT作为预训练和微调范式的标志性模型，BERT采用了双向Transformer结构，首先在大规模的文本语料上进行无监督预训练（Masked Language Model），然后针对具体任务进行微调。BERT的创新之处在于它的预训练阶段使用了大规模的语料库进行自监督学习，极大地提升了其在下游任务中的表现。BERT的微调方法通过将预训练得到的参数作为初始值，然后在特定任务上进行适当的微调，从而实现了从通用模型到任务特定模型的转换。与BERT不同，GPT采用了基于自回归的生成模型，模型首先进行预训练（自回归语言建模任务），然后通过微调来实现特定任务的优化。GPT的微调方法同样依赖于将预训练的通用模型应用到特定任务中，针对下游任务的特定数据进行训练，进一步优化性能。

随着大规模预训练模型的逐步成熟，研究者们开始探索如何在多任务、跨领域的情境下进行有效的微调。不同于传统的单一任务微调，任务特定微调（Task-specific Fine-tuning）[45]方法通过考虑不同任务的特性，调整微调策略以达到更好的效果，多任务学习[46]是微调方法中的一种创新策略，它通过共享参数来同时优化多个任务，通常通过共享某些层的参数来利用多个任务之间的关联性。这样的多任务微调方法不仅能够提高模型在各个任务上的表现，还能降低过拟合的风险。

由于大模型的不断扩展和优化，微调方法开始融合更多先进技术，尤其是有监督微调和知识蒸馏，也是目前学术界和工业界为了满足模型应用需求最常用的训练方法。监督微调（Supervised Fine-Tuning，SFT）[47]是一种传统且广泛应用的微调方法，它主要依赖于大量的人工标注数据进行训练。在大模型的背景下，SFT通常指在经过预训练的基础上，针对特定任务或数据集通过有监督的学习来调整模型参数。该过程是根据训练数据中的标签来最小化损失函数，以便模型能够在给定的任务中获得最佳表现。知识蒸馏（Knowledge Distillation, KD）[5]是一种模型压缩技术，它将一个大型、复杂的教师模型（Teacher Model）中的知识传递到一个较小的学生模型（Student Model）中。通过蒸馏，大模型的知识可以转移到更小的模型中，从而提升小模型的性能。尤其在大规模模型无法直接应用的场景中，知识蒸馏成为了一种有效的微调方法。许多现代的推理大模型（如Deepseek、Qwen等）都在微调过程中使用了蒸馏方法，以提高模型的推理性能。

强化学习（Reinforcement Learning, RL）在大模型微调中的应用已经取得了显著进展，尤其是在推理任务和决策任务中的效果越来越突出。与传统的监督微调（SFT）不同，强化学习通过奖励信号来引导模型的学习过程，模型并不依赖于人工标注数据，而是通过与环境的交互来探索最优策略。在大语言模型（LLM）中，强化学习的应用可以提高模型在推理、对话、推理能力等方面的表现。PPO（Proximal Policy Optimization）是强化学习中的一种策略优化算法，它的目标是通过最大化策略的期望回报来优化模型。在大模型的微调过程中，PPO通过优化模型的策略，使得模型能够在环境中更有效地做出决策。DPO（Direct Preference Optimization）[48]是一种新的强化学习算法，旨在直接优化模型生成的输出的“偏好”或“质量”。与PPO等基于策略梯度的算法不同，DPO通过直接对模型的生成结果进行评价和优化，推动模型在任务中更好地优化目标。DPO特别适用于文本生成类任务，例如机器翻译、对话系统和文本生成中的推理任务，能够直接对生成文本的质量进行优化，很适合优化SFT后的大模型，从而提升大模型的任务适应性和响应能力。GRPO（Group Relative Policy Optimization）是近期提出的一种强化学习算法，旨在优化多个模型之间的协作策略，尤其适用于多智能体系统和集体决策的场景。在大模型训练中，GRPO的优势在于它可以在多个模型或多个策略之间进行相对优化，从而提升整体系统的推理性能

## 2.4 大模型推理优化的相关技术发展

推理优化（Inference Optimization）[49]在大模型的发展过程中逐渐成为一个重要的研究方向。随着模型规模的扩大，推理性能成为了一个瓶颈，尤其是在硬件资源和计算时间上。推理优化的目标是减少大语言模型在推理阶段的计算资源消耗，提高响应速度，并降低部署成本。随着计算能力的提升和模型规模的不断增大，推理优化逐渐从初期的硬件加速发展到更加智能化的优化技术，具体包括，硬件加速优化：通过利用专用硬件（如TPU、GPU、FPGA等）来加速模型推理，以减轻计算压力。算法层面的优化：例如，通过压缩模型参数、优化计算图等方式，减少推理时的计算复杂度和内存占用。资源高效性优化：在确保推理质量的前提下，通过量化、剪枝、蒸馏等技术减少计算资源需求，提高模型的推理效率。

量化和剪枝是常见的推理优化技术[49, 50]，能够有效减少模型的参数量、存储需求和计算复杂度，进而提升推理速度。量化 (Quantization)是将浮点数模型参数转换为低位表示（例如8位整数或16位整数）的过程。量化能够减少模型存储和计算的开销，从而加速推理过程。常见的量化方法包括：权重量化、激活量化、训练时量化例如QLORA[51]。剪枝通过去除神经网络中的不重要连接（权重较小或影响较小的部分），来减少网络的复杂性。剪枝后的模型可以显著减少计算量和内存消耗，尤其是在推理时，能够更快地执行。但需要注意，剪枝也可能带来模型性能的下降，因此需要通过精细的控制和再训练来确保性能的保持。

混合专家模型MoE (Mixture of Experts)，是一种基于模型容量和计算资源有效分配的推理优化方法。MoE模型通过将不同的专家（子模型）分配给不同的任务或输入，而不是使用单一的模型进行所有任务。它的主要特点在于：专家动态选择，在每次推理时，根据输入数据动态选择若干个专家进行推理，而不是全模型参与计算。这有效减少了计算量，并提高了推理速度。容量与计算的平衡，MoE使得模型可以在不同的任务中充分利用计算资源，同时避免在不需要全部专家时浪费计算能力。MoE的应用让大规模模型能够在推理过程中进行更高效的计算资源调度，尤其在大规模任务中，MoE能够显著提高计算效率。

与此同时，众多大模型推理优化工具框架也层出不穷，VLLM是一个开源的高效推理框架，专为加速大语言模型的推理过程而设计。VLLM利用自行研发的Page-Attentin机制[52]通过优化模型的计算图和内存访问，降低了推理过程中的延迟，并且能够有效提升模型在不同硬件平台上的推理性能。除此之外，SGLang、LMDeploy[53, 54]等大模型推理优化工具也各有各的特点，适配不同的设备的同时能为用户带来吞吐量的提升和并发度的控制。

未来由于大语言模型规模和能力的持续增长，推理优化将成为未来研究和应用的核心方向，推动大模型技术在实际场景中的广泛部署和应用

# 第三章 相关技术研究和实验设计

在本章中，会介绍本研究涉及到的一些基本算法原理，包括但不限于Transformer模型架构分析，本研究底座模型Qwen系列模型架构分析，Deepseek-R1关键训练算法和流程解析，最后根据涉及到的技术，简单阐述本研究实验过程的整体设计方案。

## 3.1 Transformer模型架构分析

Transformer[17]由谷歌Vaswani等人于2017提出，作为近年来在自然语言处理（NLP）领域取得突破性进展的架构，其核心思想在于摒弃了传统循环神经网络（RNN）和卷积神经网络（CNN）的局部计算方式，转而采用全局自注意力机制（Self-Attention）来进行序列建模，解决序列到序列(Sequence-to-Sequence)任务。Transformer的基本架构分为两部分：编码器(Encoder)和解码器(Decoder)。如图3-1所示，左边为编码器部分，右边为解码器部分。编码器部分主要由个相同的层堆叠而成，而每一层都包含两个子层，分别为多头注意力(Multi-Head Attention, MHA)层和前馈网络(Feed-Forward Network, FNN)层，并采用相加和层归一化(Layer Normalization, LayerNorm)操作连接两个子层。解码器同样是由个相同的层堆叠而成，除了编码器的两层之外，又插入了一个掩码多头注意力层，用于将编码器的输出与解码器的输入相融合。Transformer模型在解码器部分的注意力机制上增加了上三角掩码矩阵，防止在模型训练过程中出现信息泄露的情况，保证模型在计算当前位置信息时不受后面位置信息的影响。

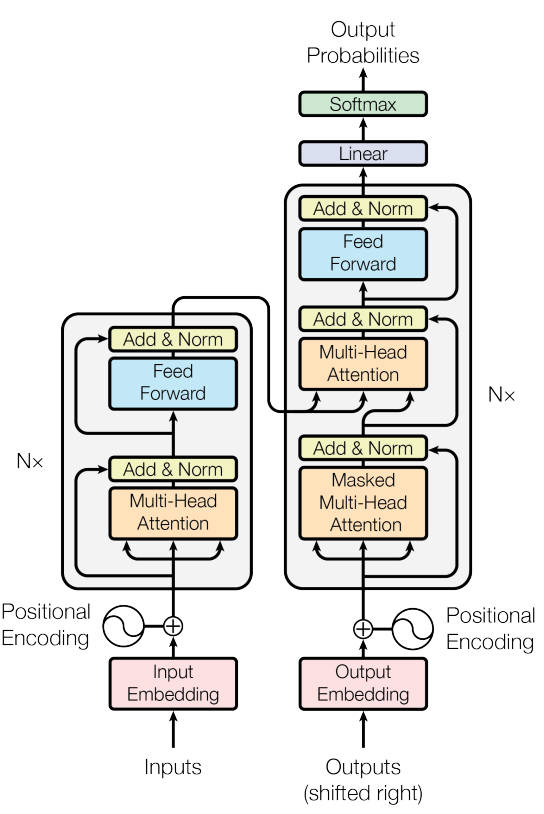


图3-1：Transformer模型架构示意图

自注意力机制是Transformers架构的核心，其目的是在输入序列中捕捉单词之间的依赖关系。在自注意力机制中，输入的每个词都会计算出一个加权和，这个加权和依赖于其他所有词的表示。一般可以看作将查询（Query）和一组键值对（Key Value Pair）映射到高维空间即对Value进行加权求和计算。自注意力的计算可以用以下公式表示：

(3-1)

在公式(3-1)中，是查询（Query）矩阵，是键（Key）矩阵是值（Value）矩阵，是键向量的维度。通过计算查询与键的点积并进行缩放，能够得到一个权重分布，再通过该分布对值进行加权求和，从而生成最终的输出表示。输入的每个单词首先通过三个不同的线性变换生成对应的查询、键和值向量。这些变换使得模型能够灵活地学习不同的表示。自注意力机制通过计算查询和键之间的相似度来为每个词分配一个权重，进而对值进行加权求和，得出最终的输出。通过这种方式，模型能够在一个序列中聚焦于与当前词最相关的其他词。

为了进一步增强模型的表达能力，Transformer引入了多头注意力机制。该机制将自注意力的计算分成多个注意力头，每个头独立地学习不同的表示，然后将所有头的输出拼接起来并进行线性变换。多头注意力的计算表示如公式(3-2)所示：

(3-2)

通过多个不同的线性变换矩阵  分别对输入的查询、键 、值  进行变换，得到不同头的注意力结果，然后将这些结果拼接起来(操作)，最后再通过输出线性变换矩阵进行变换得到最终的多头注意力输出 。 这样模型可以从不同子空间捕捉信息，更好处理序列复杂关系。

而由于Transformer模型没有使用递归或卷积结构[55]，它无法自动捕捉输入序列的顺序信息。因此，位置编码被引入以补充位置信息。位置编码是通过对输入序列的每个元素添加一个独特的向量来实现的，这个向量与元素在序列中的位置相关。常见的做法是使用正弦和余弦函数来生成位置编码，计算同时如下：

(3-3)

表示位置，是维度索引，是模型的维度。位置编码通过正弦和余弦函数为序列中的每个位置生成固定维度的向量，为模型引入位置信息 。

为了优化训练过程，Transformer采用了位置-Wise前馈网络来对每个位置的表示进行独立变换，这一过程的公式为:

(3-4)

在公式(3-4)中，x是输入向量，是权重矩阵， 是偏置向量。运算过程是先将输入与权重矩阵相乘再加上偏置，然后通过 激活函数即 ，最后再与权重矩阵相乘并加上偏置得到输出。

通过这些设计，Transformer成功地在NLP任务中提供了更高效、更强大的性能，尤其在处理长序列时表现突出

## 3.2 Qwen模型架构分析

# 第七章 参考文献

[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] BAI J, BAI S, CHU Y, et al. Qwen technical report [J]. 2023,

[29] DAO T, FU D, ERMON S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness [J]. 2022, 35(16344-59.

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

[31] 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].

[32] HU S, TU Y, HAN X, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies [J]. 2024,

[33] WHITE J, FU Q, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt [J]. 2023,

[34] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning [J]. 2020, 53(3): 1-34.

[35] YAO S, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models [J]. 2024, 36(

[36] ZIEGLER D M, STIENNON N, WU J, et al. Fine-tuning language models from human preferences [J]. 2019,

[37] LI Y, CHOI D, CHUNG J, et al. Competition-level code generation with alphacode [J]. 2022, 378(6624): 1092-7.

[38] YIN S, YOU W, JI Z, et al. MuMath-Code: Combining Tool-Use Large Language Models with Multi-perspective Data Augmentation for Mathematical Reasoning [J]. 2024,

[39] GOU Z, SHAO Z, GONG Y, et al. Tora: A tool-integrated reasoning agent for mathematical problem solving [J]. 2023,

[40] YANG A, ZHANG B, HUI B, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement [J]. 2024,

[41] DAI J, PAN X, SUN R, et al. Safe rlhf: Safe reinforcement learning from human feedback [J]. 2023,

[42] JAECH A, KALAI A, LERER A, et al. Openai o1 system card [J]. 2024,

[43] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model [J]. 2024,

[44] ZHU Q, GUO D, SHAO Z, et al. DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence [J]. 2024,

[45] LI H, ZHU C, ZHANG Y, et al. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2023 [C].

[46] ZHANG Y, YANG Q J I T O K, ENGINEERING D. A survey on multi-task learning [J]. 2021, 34(12): 5586-609.

[47] CHEN T, LIU S, CHANG S, et al. Adversarial robustness: From self-supervised pre-training to fine-tuning; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2020 [C].

[48] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model [J]. 2024, 36(

[49] HISAHARO S, NISHIMURA Y, TAKAHASHI A J A P. Optimizing llm inference clusters for enhanced performance and energy efficiency [J]. 2024,

[50] LIN J, TANG J, TANG H, et al. Awq: Activation-aware weight quantization for on-device llm compression and acceleration [J]. 2024, 6(87-100.

[51] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: Efficient finetuning of quantized llms [J]. 2023, 36(10088-115.

[52] KWON W, LI Z, ZHUANG S, et al. Efficient memory management for large language model serving with pagedattention; proceedings of the Proceedings of the 29th Symposium on Operating Systems Principles, F, 2023 [C].

[53] ZHENG L, YIN L, XIE Z, et al. Sglang: Efficient execution of structured language model programs [J]. 2024, 37(62557-83.

[54] YIN W, XU M, LI Y, et al. Llm as a system service on mobile devices [J]. 2024,

[55] KATTENBORN T, LEITLOFF J, SCHIEFER F, et al. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing [J]. 2021, 173(24-49.