北京大学

本科生毕业论文

题目：奖励塑造的困境破局：基于多选题的强化学习推理研究

Reward Shaping Breakthrough: Enhancing RL Reasoning through Multi-Choice Paradigms

姓　　名：	王昱琛
学　　号：	2100013153
院　　系：	北京大学
专　　业：	智能科学与技术
导师姓名：	冯岩松

二〇二五年五月

北京大学本科毕业论文导师评阅表

学生姓名	王昱琛	本科院系	北京大学	论文成绩 （等级制）	
学生学号	2100013153	本科专业	智能科学与技术		
导师姓名	冯岩松	导师单位/ 所在学院		导师职称	
论文题目	中文	奖励塑造的困境破局：基于多选题的强化学习推理研究			
	英文	Reward Shaping Breakthrough: Enhancing RL Reasoning through Multi-Choice Paradigms			

导师评语

（包含对论文的性质、难度、分量、综合训练等是否符合培养目标的目的等评价）

导师签名:

年　　月　　日

版权声明

摘要

随着大语言模型（LLM）在自然语言处理领域的迅猛发展，其在法律推理任务中的应用引起了广泛关注。然而，法律多选题因答案非唯一、推理链条复杂，对模型能力提出更高要求。本文提出了一种融合监督微调（SFT）与组相对策略优化（GRPO）的双阶段训练框架，在中国国家司法考试多选题（JEC-QA）案例分析子集上进行了验证。首先，利用 DeepSeek-R1 生成的长链思维示例对 Qwen-7B-Instruct 模型进行蒸馏并执行全量 SFT，以实现法律知识的冷启动；随后，基于格式奖励与严格准确性奖励，通过 GRPO 对模型进行强化学习微调，并引入 Dr. GRPO 长度偏差修正策略以抑制无效长尾推理。实验证明，本方法在测试集上将精确匹配率提升至 0.57，进一步的定量分析与结构化响应评估显示，模型在层次化段落组织、语篇连贯性和法条引用准确性等方面均有显著改进。

关键词：大语言模型；法律推理；监督微调；组相对策略优化（GRPO）；JEC-QA；多选题

ABSTRACT

With the rapid advancement of large language models (LLMs), their application to legal reasoning tasks has garnered significant interest. However, multiple-choice legal questions present unique challenges due to non-unique correct answers and complex multi-step reasoning. This paper introduces a two-stage training framework combining Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO) on the case-analysis subset of the JEC-QA dataset. First, we distill chain-of-thought examples generated by DeepSeek-R1 into Qwen-7B-Instruct and perform full SFT for legal knowledge cold-start. Next, we fine-tune via GRPO using format and strict accuracy rewards, incorporating Dr. GRPO length-bias correction to mitigate inefficient long-tail reasoning. Experiments show our method achieves a mean1 accuracy of 0.57 on the test set. Quantitative evaluation and structured response analysis further demonstrate significant enhancements in hierarchical paragraph structuring, discourse coherence, and precise legal provision citation.

KEY WORDS: Large Language Models; Legal Reasoning; Supervised Fine-Tuning; Group Relative Policy Optimization (GRPO); JEC-QA; Multiple-Choice Questions

目　录

1. 引言

1.1 研究背景

随着人工智能技术的快速发展，大语言模型（Large Language Models, LLMs）在自然语言处理任务中展现出卓越的性能，尤其在逻辑推理、语义理解和文本生成等方面取得了显著成果。然而，法律领域因其高度的专业性与复杂性，对模型推理能力提出了更高要求。法律问题通常具有非唯一性的特点，涉及多层次的法律逻辑与价值判断，这使得传统强化学习方法在法律任务中的应用面临诸多挑战。目前，主流的大语言模型推理能力多来源于数学与编程领域的训练，而本研究则创新性地探讨通过法律复杂问题训练，促进模型跨领域推理能力的提升。

已有研究表明，传统强化学习方法在面对答案唯一、评价标准明确的任务中表现良好；但在涉及多种合理解释的法律问题中，奖励函数的设计变得尤为困难。具体挑战包括：（1）法律问题的开放性导致标准化的评价指标难以确立；（2）合理答案可能基于不同的法律分析路径或表述方式；（3）法律推理过程通常要求严密的逻辑链条与专业知识支撑。因此，构建有效的法律推理模型，并通过强化学习技术优化其推理路径，成为当前亟需解决的重要科学问题。

近期研究显示，强化学习方法有助于提升大语言模型的推理能力。例如，DeepSeek-RL[1]、Qwen[2] 以及 Seed 系列模型[3] 表明，仅通过提供问题与答案对，模型可在缺乏显式推理监督的情况下，通过强化学习自主学习有效的推理策略。然而，这些方法在法律领域的适用性仍待验证，主要原因包括：（1）法律问题评估标准模糊，难以构建可通用的奖励函数；（2）缺乏标准化的评估机制，容易导致模型学习到偏离实际法律逻辑的策略。

为缓解上述问题，本研究选取不定项选择题（Multiple-Choice Questions, MCQs）作为训练数据的基础。该数据形式具有双重优势：一方面，MCQs 提供了明确的选项空间，使基于规则的奖励函数设计更为可行；另一方面，多选题要求模型系统性地分析并评估所有选项，有助于培养其更全面的推理能力。本研究采用中国国家司法考试中的 10,561 道案例型多选题作为训练与评估语料，涵盖民法、刑法、诉讼法等主要法律门类，需具备专业级别的法律知识与推理能力。

1.2 研究目标与贡献

本研究旨在探索后训练（post-training）策略在法律领域的应用效果，重点关注如何通过强化学习优化大语言模型的法律推理能力。

具体研究目标如下：提升大语言模型在中国国家司法考试多项选择题中的答题准确率与推理质量，使模型能够在法律语境下自主学习并掌握系统的法律推理方法。

为实现上述目标，模型输入格式设计为：

> “你是一名法学专家。现在请你解答司法考试中的一道选择题，请你找出所有正确的选项。每道题可能有一个或者多个正确答案。在解答之前，你需要先针

对每个提供的选项给出详细的解释。你需要在回答的最后用大括号圈出给出的答案，例如 B 或者 ABD。”

模型输出包括：

- “思考”部分：展现完整的法律推理链条；
- “回答”部分：给出最终选项集合。

1.3 论文结构

本文采用系统化结构进行组织，各章节安排如下：

- 第二章为相关工作，系统梳理大语言模型中强化学习方法的研究进展及其在法律推理任务中的应用现状；
- 第三章详细介绍本研究提出的方法，重点描述基于多选题的强化学习训练机制及其与监督微调的混合训练框架，包括奖励函数构建与训练策略设计；
- 第四章构建实验框架，介绍后训练策略的具体实现，包括底层系统设计、基线模型选取及评估指标体系；
- 第五章呈现实验结果与综合分析，通过定量指标与定性案例展示强化学习对模型推理能力的提升效果，并分析其背后机制；
- 第六章为讨论部分，深入探讨本方法的优势与不足，提出改进方向并讨论方法在其他专业领域的潜在适用性；
- 第七章总结全文，概括研究成果，并展望大语言模型在法律智能领域的未来发展趋势。

本研究不仅为法律智能提供了新的技术路径，其提出的混合训练范式与评估框架亦可推广至医疗诊断、金融分析等高专业性的人工智能应用领域。通过构建面向法律推理的大模型训练体系，我们希望推动人工智能向更强逻辑推理能力的方向迈进。

2. 相关工作

近年来，强化学习（Reinforcement Learning, RL）被广泛应用于大语言模型（Large Language Models, LLMs）的微调与对齐过程中。其中，基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）成为主流技术框架，旨在训练奖励模型以捕捉人类偏好，并据此优化语言模型的输出策略。这类方法通常引入额外的奖励函数 $r_\phi(x, y)$ 对模型输出进行评分，并通过近端策略优化（PPO）[4]、直接偏好优化（DPO）[5] 等策略梯度算法对策略 $\pi_\theta(y \mid x)$ 进行更新，从而使模型输出更符合人类预期。这些技术的发展表明，强化学习能够在提升 LLM 可控性和对齐能力方面发挥重要作用。近期的研究也进一步指出，强化学习是引导 LLM 学习推理能力的关键途径。

LLM 的后训练（post-training）技术旨在在通用预训练模型基础上，结合领域数据和任务需求进一步提升模型性能，从而推动了如 DeepSeek-R1[1] 等大型后训练模型（Large-scale Re-trained Models, LRMs）的发展。这类策略主要包括：使用领域数据进行微调、结合 RLHF 实现人类偏好对齐，以及设计复杂的训练流程以增强模型的多步推理能力。这些方法共同构成了 LLM 从通用能力向领域能力演进的研究范式。

在法律推理智能化研究中，LLM 的应用已经进入纵深发展阶段，但距离司法实践的实际需求仍存在显著能力鸿沟。一方面，法律文本具有强逻辑性，要求模型具备多层级推理能力；另一方面，司法决策强调思维链条的可追溯性与可信度，呼唤模型的思考过程满足 TRACER 标准（Traceable, Reasonable, Accountable, Contextualized, Evidence-based, Rational）。为应对上述挑战，Lawyer LLaMA[6] 基于 LLaMA 对法律任务进行了微调，表现出对案件分析和法律推理的改进能力；ChatLaw[7] 则结合法律知识检索、多专家融合和多智能体方法提升法律场景表现；LawGPT[8] 构建了高质量法律问答数据集，并在中文模型基础上进行了领域特定微调。然而，我们通过系统性分析发现，目前法律大模型普遍采用监督微调（Supervised Fine-Tuning, SFT）这一单一策略，在满足司法场景所需的知识完备性和推理可信度方面仍存显著差距。为此，我们提出融合 SFT 与 RLHF 的双阶段训练框架：首先通过法律数据的监督微调实现法理知识的冷启动，随后引入 RLHF 强化模型的推理能力与知识迁移能力。该混合式训练策略不仅加深模型对法律知识的内化，还显著提升了模型对法条与案例的动态匹配与推理能力，从而提高司法决策的合理性与可解释性。

自 DeepSeek 团队提出 GRPO（Group Relative Policy Optimization）算法以来[9]，其在数学推理等任务中展现出显著性能，但也暴露出训练稳定性差、奖励噪声大等问题。为应对上述挑战，研究者提出了多种改进方法，包括 DAPO（Decoupled Clip and Dynamic Sampling Policy Optimization）[10] 和 Dr. GRPO[11]，进一步推动了 RL 算法的发展。

GRPO 的核心机制是通过分组估计（group estimation）方式计算优势函数，从而在无需 critic 模型的情况下降低计算资源消耗，同时保持模型训练的稳定性与高效性。其目标函数为：

$$\text{Objective} = \sum_i \sum_t \left(\frac{\pi(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)} \right)^{\text{clip}} A(s_t, a_t) \tag{1}$$

其中优势函数 $A(s_t, a_t)$ 计算方式为：

$$A(s_t, a_t) = r(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t) \tag{2}$$

在 GRPO 基础上，DAPO 和 Dr. GRPO 引入如下优化策略：

- **删除 KL 约束**：提升模型自由度，突破参考模型性能上限[10]。
- **Clip-Higher 策略**：缓解熵崩塌，提高生成多样性[10]。
- **Token-Level Loss 调整**：增强模型对长文本的处理能力[10]。
- **难度偏差修正**：移除标准差归一处理，提升指标稳定性[11]。
- **动态采样机制**：过滤奖励全 0/1 样本，防止梯度停滞[10]。
- **超长奖励重构**：通过软性惩罚机制规避响应冗长问题[10]。

GRPO 系列算法在长文本生成中展现出显著提升，其优化方法为强化学习在语言生成任务中的进一步应用提供了新路径。

2.1 推理能力与泛化性

除了 DeepSeek-R1 中提出的“AHA 时刻”外，强化学习还被发现能在训练过程中自然诱导模型产生自我验证、反思式推理等能力[12]。这类能力在没有明确监督信号的情况下自发出现，表明 RL 具备促进认知过程发展的潜力。

尽管大部分 RL 推理研究集中于数学与代码领域，但已有工作证明 RL 能促进模型向其他结构化或非结构化领域的泛化。例如，Logic-RL[13] 通过规则驱动的训练，在逻辑谜题任务中取得成功，并在数学推理任务中迁移表现良好，验证了 RL 能独立于领域知识诱导通用推理机制。

更令人关注的是，一些推理模型（如 DeepSeek-R1[1]）已将推理能力拓展至医学、化学、心理学等人文学科，结合生成式软评分机制处理非结构化答案，为构建跨领域通用推理模型奠定基础。下一步的研究方向包括将此类模型与工具调用、检索增强生成（RAG）等系统集成，OpenAI 最新发布的 o3 模型正朝这一方向迈进。

2.2 数据集：JEC-QA 案例分析题

本研究使用中国国家司法考试官方题库中的 JEC-QA 数据集[14]，该数据集广泛应用于法律人工智能领域，具有较高的权威性与代表性。

- **数据来源**：采自中国国家司法考试历年真题，涵盖民法、刑法、行政法、诉讼法等多个法律分支，具有良好的覆盖性与实用性。
- **规模与结构**：
  - 数据集中共包含 26,365 道多项选择题；

  - 每道题可能包含一个或多个正确答案；
  - 题型分为两类：**知识型问题**（注重法条记忆）与 **案例分析问题**（侧重法律推理）。

- **研究子集选择**：本研究聚焦于其中 10,561 道 **案例分析问题**，该题型通常涉及具体案情描述，要求应试者结合法律规范进行多步推理与判断，充分体现法律推理过程的复杂性。选择该子集的原因在于其更贴近真实法律实践，对模型的法律理解与推理能力提出更高要求。
- **数据划分方式**：按照 8:2 的比例将案例分析题划分为训练集与测试集，分别包含 8,449 道与 2,112 道题目，确保训练与评估的均衡性与代表性。
- **任务定义**：模型需完成以下三个关键步骤：
  - 准确识别案情中的法律争议点；
  - 正确定位并引用适用的法律条文；
  - 通过多步法律逻辑推理，选出所有正确选项。

该任务对语言模型的逻辑推理能力、法理理解能力以及语言表达一致性提出了系统性挑战，适合作为评估模型法律智能表现的标准基准。

3. 方法

3.1 任务定义

本研究旨在提升大型语言模型在中国国家司法考试多项选择题上的法律推理能力与答题准确率。具体定义如下:

- **研究目标：**提升 LLM 在中国国家司法考试多选题上的法律推理质量及选项预测准确率。
- **输入：**法律多选题，包括案例描述与若干备选答案；模型需先对每个选项进行推理分析。
- **输出：**

  1. 思考（*Reasoning*）：完整的法律推理链条，对各选项给出详细解释；
  2. 回答（*Answer*）：最终以大括号形式给出正确选项集合，如“{B}”或“{ABD}”。

3.2 基于人类偏好的强化学习（RLHF）流程

我们采用标准的基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）方法，对预训练语言模型进行微调。该流程包含以下关键步骤:

1. **奖励模型训练**：利用人工标注的偏好数据，训练奖励模型 $r_\phi(x, y)$。该模型基于输入提示 x 和回答 y，对样本对 (x, y_1, y_2) 的质量进行评估。
2. **策略优化**：在奖励模型的指导下，通过策略梯度方法优化语言模型策略 $\pi_\theta(y \mid x)$，以提升其生成高质量回答的能力。

由于司法考试题目对逻辑一致性和推理链条完整性的要求极高，RLHF 能有效引导模型生成更符合人类偏好的结构化解释。

3.3 近端策略优化（PPO）

PPO[4]是一种广泛应用于语言模型强化学习微调的 Actor-Critic 算法。PPO 通过联合训练一个值函数网络（Critic），并引入裁剪代理目标来限制策略更新的幅度，从而提高训练的稳定性和样本效率。其核心思想是通过裁剪重要性采样比率，确保策略更新不会偏离当前策略过远。

PPO 方法 PPO 是一种基于策略梯度的算法，结合值函数和广义优势估计（Generalized Advantage Estimation, GAE）来计算优势函数 $\hat{A}_t$，并在奖励中加入 KL 散度惩罚项以稳定训练。其优化目标定义为:

$$L^{\text{PPO}}(\theta) = \mathbb{E}_{(x,y)\sim D_{\pi_{\text{ref}}}} \left[r_{\text{model}}(x, y) - \beta \, \text{KL}(\pi_\theta(y \mid x) \| \pi_{\text{ref}}(y \mid x)) \right]$$

其中，$r_{\text{model}}(x, y)$ 是奖励模型对生成结果 y 的评分，π_{ref} 为参考策略（通常为监督微调后的预训练模型），β 为 KL 惩罚系数。策略更新采用裁剪代理目标函数：

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta),\ 1-\epsilon,\ 1+\epsilon)\hat{A}_t \right) \right]$$

其中，$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ 表示当前策略与旧策略在时刻 t 的概率比，$\hat{A}_t$ 为 GAE 估计的优势值。然而，PPO 在大规模语言模型中面临两大挑战：一是需要额外训练值函数网络以计算 GAE，增加了参数规模和训练成本；二是文本生成任务中通常仅在序列末尾获得回报信号，导致中间状态的价值估计困难，影响训练的稳定性与效率。为了解决 PPO 在 LM 中需训练 Critic 以及仅末尾奖励的问题，我们引入组相对策略优化（GRPO）方法，该方法以采样组为单位进行相对评价，无需 Critic 结构，且更加稳定高效。

3.4 组相对策略优化（GRPO）

GRPO[9]是一种无需值函数的策略优化算法，通过模型自身生成的多组回答样本进行相对评分，估计优势函数（Advantage Function）并更新策略。其主要特点包括：

1. **无需值函数模型**：传统 PPO 依赖值函数 $V(x)$ 估计优势，而 GRPO 基于策略采样的相对信息，通过组内奖励归一化估计优势，消除了对价值函数的依赖。
2. **样本组评分机制**：对于输入 x，从当前策略 π_θ 生成多个回答 $\{y_1, \ldots, y_K\}$，并由奖励模型评分：

   $$r_i = r_\phi(x, y_i), \quad i = 1, \ldots, K$$

   根据评分排序，为每个样本分配排名权重 α_i（如 Softmax 归一化得分），用于构造相对优势。
3. **优势函数估计**：

   $$A(x, y_i) = \alpha_i - \frac{1}{K} \sum_{j=1}^{K} \alpha_j$$

   其中，α_i 可为 Softmax 输出或单调映射的排名指标。
4. **策略更新目标**：GRPO 使用裁剪策略目标限制偏移：

   $$L^{\text{GRPO}}(\theta) = \hat{\mathbb{E}}_{(x,y_i)\sim\pi_{\theta_{\text{old}}}} \left[\min \left(r_\theta(x, y_i)A(x, y_i),\ \text{clip}(r_\theta(x, y_i),\ 1-\epsilon,\ 1+\epsilon)A(x, y_i) \right) \right]$$

   其中，$r_\theta(x, y) = \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}$。

GRPO 方法：GRPO 通过对输入 x 生成 K 个回答 $\{y_1, y_2, \ldots, y_K\}$，利用奖励模型评分 $r_i = r_\phi(x, y_i)$，并应用 Softmax 函数：

$$\alpha_i = \frac{\exp(r_i/\tau)}{\sum_{j=1}^{K} \exp(r_j/\tau)}$$

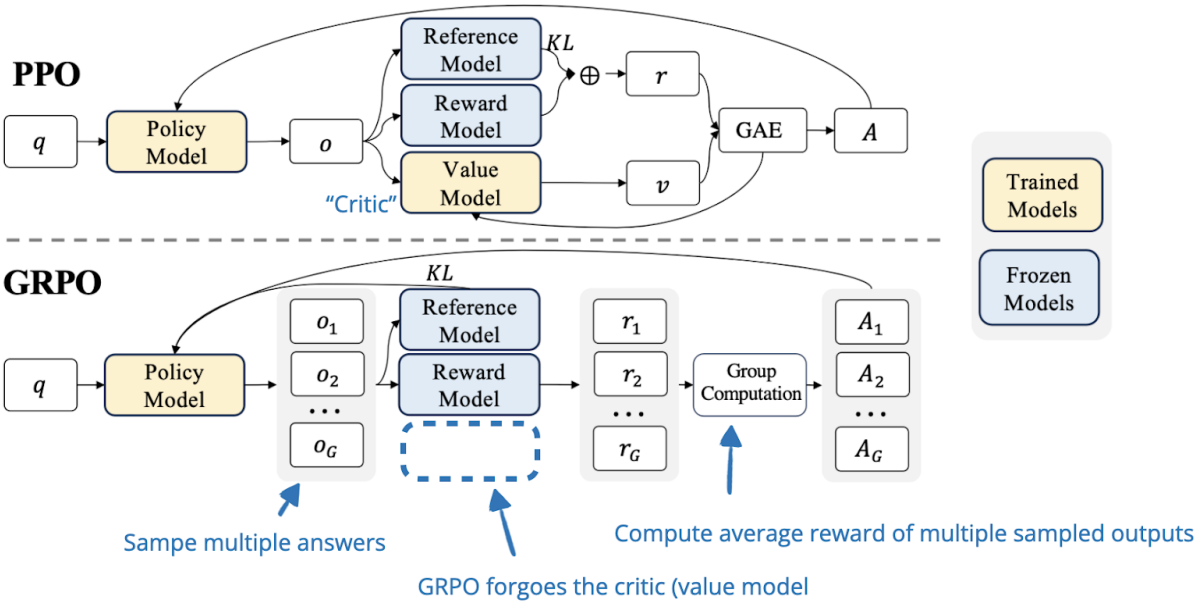

图 3.1　GRPO 方法示意图[9]

其中 τ 为温度参数。相对优势定义为：

$$A(x, y_i) = \alpha_i - \frac{1}{K}\sum_{j=1}^{K}\alpha_j$$

高于均值的回答获得正优势，反之则为负优势。GRPO 通过组内比较直接评估优势，无需值函数网络，降低了内存和计算开销。

3.5　DAPO 算法

为提升模型在长链推理（Chain-of-Thought, CoT）任务中的表现，有工作引入了**解耦裁剪与动态采样策略优化（Decoupled Clipped and Dynamic Sampling Policy Optimization, DAPO）**算法[10]。该算法通过一系列创新策略，优化了模型在复杂推理任务中的训练效率和性能。以下是其核心创新点的详细描述：

- **解耦裁剪策略**：DAPO 采用非对称裁剪机制，分别设置下限 $\varepsilon_{\text{low}} = 0.2$ 和上限 $\varepsilon_{\text{high}} = 0.28$。在传统的 PPO-clip 算法中，裁剪参数 ε 是对称的，即上下限相同，这限制了低概率动作（exploration token）的提升空间。DAPO 通过解耦上下裁剪参数，允许低概率动作的提升幅度更大，从而促进模型的多样性。具体而言，对于策略更新的裁剪目标函数 L^{CLIP}，DAPO 定义为：

$$L^{CLIP} = \mathbb{E}_t\left[\min\left(\frac{\pi(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}A^{\pi_{\text{old}}}(s_t,a_t),\ \varepsilon_{\text{high}}\right) - \max\left(\frac{\pi(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}A^{\pi_{\text{old}}}(s_t,a_t),\ -\varepsilon_{\text{low}}\right)\right]$$

其中，$\pi(a_t|s_t)$ 是当前策略，$\pi_{\text{old}}(a_t|s_t)$ 是旧策略，$A^{\pi_{\text{old}}}(s_t,a_t)$ 是优势函数。这种非对称裁剪机制能够有效避免熵崩溃现象，同时提升模型的探索能力。

- **动态样本重采样**：DAPO 通过过采样并筛选出准确率非 0 或 1 的样本，确保每个样本都能提供有效的梯度信息。在传统的采样方法中，组内全对或全错的样本会导致优势函数为零，从而没有梯度，影响训练效率。DAPO 的动态采样策略通过过滤这些极端样本，确保每个采样组内的优势函数均非零。具体而言，对于采样组 G，DAPO 仅保留满足 $0 < \text{accuracy}(G) < 1$ 的样本组，从而提高训练效率并降低梯度方差。
- **Token 级策略梯度损失**：DAPO 在所有 token 上计算损失，赋予长序列更大的权重。传统的样本级损失计算方式（sample-level loss）对长回答和短回答赋予相同的权重，这在长链推理任务中是不合理的。DAPO 采用 token 级损失计算方式，定义为：

$$L^{TOKEN} = \sum_{t=1}^{T} \log \pi(a_t|s_t) \cdot A^{\pi_{\text{old}}}(s_t, a_t)$$

  其中，T 是序列长度。通过这种方式，DAPO 对长回答中的高质量和低质量响应模式赋予更大的权重，从而提升模型在复杂推理任务中的表现。
- **过长奖励整形**：DAPO 采用软长度惩罚机制，其奖励函数定义为：

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| \le L_{\max} - L_{\text{cache}} \\ \frac{(L_{\max} - L_{\text{cache}}) - |y|}{L_{\text{cache}}}, & L_{\max} - L_{\text{cache}} < |y| \le L_{\max} \\ -1, & L_{\max} < |y| \end{cases}$$

  其中，$L_{\max} = 16384$ 是最大允许长度，$L_{\text{cache}} = 4096$ 是缓存长度。这种软惩罚机制能够有效惩罚过长的回答，同时避免因长度限制而引入的噪声。具体而言，当回答长度 $|y|$ 超过 $L_{\max} - L_{\text{cache}}$ 时，奖励函数逐渐减小，直至达到最大长度 $L_{\max}$ 时，奖励值为 -1。这种机制能够引导模型生成更合理的回答长度，同时避免因长度限制而带来的梯度消失问题。

3.6 Dr.GRPO 算法

Dr.GRPO 算法[11]是对 GRPO 的改进，通过修正偏见提升推理能力。其主要改进包括：

- **去除响应长度偏见**：移除除以响应长度的项，确保策略更新不受长度影响。
- **消除问题难度偏见**：移除组内奖励标准差归一化，使学习过程更平衡。
- **调整基线计算**：优化基线估计为：

$$\mathbb{E}[r_k] = \frac{1}{K} \sum_{k=1}^{K} r_k$$

目标函数为：

$$J_{\text{Dr.GRPO}}(\theta) = \mathbb{E}_{\{\tau_k\}_{k=1}^{K} \sim \pi_{\theta_{\text{old}}}} \left[\sum_{k=1}^{K} \sum_{t=1}^{|o_k|} \hat{\rho}_{k,t}(\theta)(r_k - \mathbb{E}[r_k]) \right]$$

3.7 方法小结

我们首先定义了多选题法律推理任务的输入输出格式，并采用 RLHF 框架提升生成质量。随后，我们从经典的 PPO[4]出发，分析其在法律推理任务中的局限，引入无需 Critic 的 GRPO 方法[9]提升训练效率。最后，我们进一步介绍了 DAPO[10],Dr.GRPO[11]等 GRPO 的改进算法，以提升模型在长链推理任务中的表现。

3.8 奖励函数

格式奖励

格式奖励的核心目标是实施指导模型推理过程的结构性约束，以确保其响应遵守特定的逻辑框架。具体来说，我们要求模型的输出包括以下结构：<思考> 标签和 <回答> 标签，它们在解决问题时明确概述了模型所采取的推理路径[1]。例如，这可能涉及在法律问题中确定有争议的观点，召回相关的法规或判例法，并比较不同的法律概念。

准确性奖励

单项选择问题在 2 或 4 个选项中有一个正确的答案。通常可以通过不完整或表面的推理来解决它们，因为有限的答案空间使模型可以在不完全理解基本概念的情况下获得正确的答案。这种局限性使此类格式在训练强大的推理能力方面的有效性降低了，尤其是在法律等复杂领域。相比之下，具有一个或多个正确答案的多选题问题要求该模型可以全面评估每个选项的正确性，而没有猜测或部分推理的空间。我们设计了一个严格的奖励机制，仅当模型在所有其他情况下都可以选择所有正确的选项并分配零奖励（奖励 $= 0$）时，才能提供积极的反馈（奖励 $= 1$）。令 S_{true} 为多选题问题的正确选项集合，S_{pred} 为模型预测的选项集合。则准确性奖励 R_{accuracy} 定义为：

$$R_{\text{accuracy}}(S_{\text{pred}}, S_{\text{true}}) = \begin{cases} 1 & \text{if } S_{\text{pred}} = S_{\text{true}} \\ 0 & \text{otherwise} \end{cases}$$

它确保正确的选项与准确和完整的推理过程相关联。这种严格的奖励机制消除了在强化学习过程中基于捷径的解决方案的可能性，并迫使该模型对当前的任务有了更深入的了解。

4. 实验框架

4.1 基线模型（Baselines）

- **Qwen2.5-7B-Instruct**：作为当前中文领域表现最强的开源语言模型之一[2]，该模型被用于评估法律任务上的 zero-shot 能力，作为原始模型性能参考。
- **R1 蒸馏模型**：我们基于 Qwen2.5-7B-Instruct 的响应（R1）结果构建蒸馏数据，对 student 模型进行监督训练。该做法受到近年来在数学和编程任务中取得显著成效的 teacher-student 蒸馏范式启发[1]，作为监督蒸馏方法的基线。

4.2 监督微调（SFT）阶段

我们采用 LLaMA-Factory 和 DeepSpeed 作为监督微调的训练框架。LLaMA-Factory 是一个专为大型语言模型开发的高效训练工具，支持全参数微调、LoRA 适配、多 GPU 并行训练等功能[llamafactory]，在本研究中用于组织模型微调流程。

在蒸馏实验设置中, 我们首先基于 `jec-qa-1-multi-choice` 数据集, 使用 DeepSeek-R1 模型生成的回答作为 teacher signal 构建蒸馏语料。随后, 利用该语料对 Qwen-7B-Instruct 模型进行全量监督微调。训练采用如下超参数设置：训练轮次为 2 epoch，batch size 为 32，最大上下文长度为 4096，学习率设为 1×10^{-5}，使用余弦退火调度器（cosine decay schedule），并在前 10% 的训练步骤中进行预热（warm-up）。该设置基于任务复杂度与模型容量的经验选择，旨在提升训练收敛性与稳定性。

4.3 GRPO 阶段

在监督微调后，我们进一步对微调后的 Qwen-7B-Instruct 模型执行 GRPO（Group Relative Policy Optimization）训练。该阶段的奖励函数结合了两项指标：

- **准确性奖励（Accuracy Reward）**：根据模型输出与标准答案的匹配度计算；
- **格式奖励（Format Reward）**：鼓励模型输出结构规范、答案格式统一的响应。

通过设计多维奖励信号，引导模型在生成过程中兼顾法律结论的正确性与答案呈现的一致性，从而更好地适应实际应用需求。

4.4 训练加速框架：DeepSpeed 与 ZeRO 策略

我们使用 DeepSpeed 作为底层训练加速器。DeepSpeed 是微软推出的分布式训练库，核心为 ZeRO（Zero Redundancy Optimizer）优化策略。ZeRO 通过将模型训练过程中的优化器状态、梯度以及模型参数在多张 GPU 之间分布切分，大幅降低了单卡显存需求，并提升了训练效率。其分为三个阶段：

- **Stage 1**：优化器状态分区；

- **Stage 2**：在 Stage 1 基础上进一步对梯度分区；
- **Stage 3**：进一步将模型参数分区，实现最大程度的内存优化，适用于超大规模模型的训练。

该方案使得我们能够在有限的硬件资源条件下，对 7B 参数级别模型进行高效训练与优化。

4.5 VeRL

除了模型训练框架，推理阶段的高效引擎也是研究热点。vLLM 是一个面向 LLM 推理的高吞吐量库，采用分页注意力（Paged Attention）等技术来优化 GPU 内存管理。实验结果表明，与传统的 HuggingFace Transformers 和 OpenAI TGI 等框架相比，vLLM 在相同硬件资源下可实现数倍以上的吞吐量提升，同时大幅减少 KV 缓存浪费。vLLM 支持多种主流模型架构（如 LLaMA、Mistral、Qwen、DeepSeek 等），并提供量化与工具调用等扩展功能。这些优化使得在法律等领域部署大型 LLM 时，推理效率和可扩展性得到显著提高。

General Concepts

RL: 强化学习（Reinforcement Learning，RL）在近年来因 `o1/r1` 等技术突破而成为大模型训练的热门方向。VeRL 是一款面向 RL 的高效训练框架。从自然语言处理（NLP）的视角来看，RL 与传统的监督微调（SFT）主要有以下差异:

1. **引入惩罚信号**：SFT 仅模仿正例，而 RL 同时对优质样本给予奖励、对劣质样本施加惩罚。无论策略梯度、GRPO、Reinforce 还是 PPO，本质上都在设计奖励／惩罚的颗粒度（token / macro action / sequence 等）与强度（是否使用 baseline、KL 约束、clip 等）。
2. **允许模型自采样并在线训练自身**：SFT 通常依赖人工标注或其他模型生成的数据（蒸馏）。RL 则可实时采样并利用当前策略更新模型。

on policy vs. online

- **online**: 当前策略能否与环境交互并实时获取奖励信号（如数学题求解后立即得知正确与否）。在 GUI Agent、自动驾驶等场景需构建复杂模拟器。
- **on policy**: 训练数据是否由最新策略采样。在实践中常预采大量经验数据再分 mini batch 更新，除首个 mini batch 外均为 off policy。

常用的 GRPO/Reinforce/PPO 等方法一定是 *online*，但不必然 *on policy*（取决于 mini batch 数）。

Ray 系统概览 Ray 是一套分布式计算框架，为 VeRL 和 OpenRLHF 等 RL 框架提供 Actor 角色管理及资源调度。核心概念如下：

- **Ray Actor**: 有状态远程任务（由 `ray.remote` 装饰的 Python 类），运行时对应独立进程（勿与 RL 中的"Actor" 角色混淆）。
- **Ray Task**: 无状态远程任务（由 `ray.remote` 装饰的函数），局部变量对提交方不可见，可视作无状态。
- **资源管理**: Ray 可按 CPU/GPU/内存等进行自动调度，也支持 *placement group* 将 Actor 固定在同一或不同设备 bundle 上。
- **异步执行**: Ray 调度默认异步，任务提交即返回对象引用；用户可用 `ray.get` ／ `ray.wait` 阻塞或轮询结果。

在 RL 训练中引入异步设计，可让 Actor ／ Critic ／ Generator ／ RM 等角色的计算流水重叠，例如在 Actor 更新上一批数据时，Generator 已可并行生成下一批样本。鉴于 o1 style RL 的主要瓶颈位于 rollout，未来的优化方向是更充分地异步化 rollout（如夜间充分利用线上推理集群空闲算力）。

并行策略

- **3D 并行**: LLM 训练（megatron lm）与推理引擎（vllm、sglang）已广泛支持数据并行（DP）、张量并行（TP）与流水线并行（PP）。VeRL 新版本基于 Ulysses 进一步支持序列并行（SP），对长文本 RL 尤为关键。
- 不同角色在不同阶段可灵活调整 3D 并行组合，VeRL 借助 *hybrid engine* 做了诸多优化，如零冗余参数 re sharding。

FSDP 与 Megatron FSDP（Meta 提出）与 Megatron 分别代表两套分布式训练框架：

- **FSDP**: 将模型参数（权重、优化器状态等）在 GPU 间分片存储，仅按需通信，并重叠计算与通信，逻辑清晰、易支持新结构，研究友好。
- **Megatron**: 在百亿级模型训练中更具性能优势，参数 re sharding 开销低，工程友好。

VeRL 同时兼容两套引擎。

4.5.1 VeRL related Concepts

Hybrid Flow RL 训练逻辑涉及多模型交互。VeRL 将数据流抽象为两层：

- **控制流**: 高层描述角色间交互，如 Actor 生成经验后，Critic ／ RM ／ Reference 计算得分，然后计算 GAE 与损失。
- **计算流**: 低层描述单角色内的前向 反向、优化器更新、自回归生成等。

Single Controller vs. Multiple Controllers

- **Single Controller**: 单中心控制器统一管理所有子模块，架构清晰、易维护。VeRL 采用该模式实现 RL 算法控制流，极大便利新算法开发。
- **Multiple Controllers**: 将控制逻辑分散到多控制器，通过集合通信同步，通信开销低但逻辑更复杂。VeRL 在*计算流*维度使用此模式，以降低通信负载。

VeRL 通过多层级 Worker（`RayWorkerGroup` → `WorkerDict` → `ModelWorker` → `ParallelWorker`）封装计算流。

Hybrid Engine — 模型放置策略

1. **分开放置**: 角色独立占用设备，可异步执行但 GPU 利用率略低。
2. **分组放置**: 将角色分组共置，既能重叠执行又减少空闲：
    - 典型分组：Actor/Generator 同组（需实时同步参数），Critic/RM 与 Reference 分别单独。
3. **全部共置**: 所有角色共用设备，GPU 始终被占用，但只能串行。

VeRL 通过 `resource pool` 灵活支持以上策略，并设计 `worker_dict` 以动态角色切换（`reload/offload params`）。

数据传输协议 为适配不同角色的数据切分需求，VeRL 设计了统一的数据分发（Dispatch）与收集（Collect）协议，并以 Python 装饰器形式绑定到 Worker 方法，实现透明的数据流与执行模式。

训练流程示意

1. `RayPPOTrainer` 向 `RayWorkerGroup` 发送方法调用。
2. `RayWorkerGroup` 先执行数据分发，再根据执行模式决定哪些 Worker 执行任务。
3. 结果经收集逻辑处理后返回 Trainer。

4.5.2 RL 设置

本研究的 RL 阶段采用 VeRL（Sheng et al., 2024）实现组相对策略优化（GRPO）。以下结合 VeRL 的特性，阐述 GRPO 实现步骤。

1. 初始化与核心组件 GRPO 训练由 `GRPOTrainer` 管理，初始化需配置：

- **策略模型（Actor）**：采用 QWEN2.5 7B Instruct。
- **参照模型（Reference）**：同架构 SFT 模型，用于计算 KL 惩罚。
- **奖励函数**: 采用基于规则的格式与准确性奖励，无需外部 RM。

2. Rollout 生成

- **权重同步**: 每轮大迭代开始，通过 `RolloutShardingManager` 将最新 Actor 权重同步到推理引擎（vLLM ／ SGLang）。
- **提示采样**: 从数据集中抽取一批 prompt，由 `generate_sequences` 调用推理引擎生成响应。
- **多样生成**: 配置 `actor_rollout_ref.rollout.n`（本研究设 $K=7$）并设置 `temperature=1` 为每个 prompt 并行生成 K 条不同响应。
- **奖励计算**: 对每条响应计算格式与准确性奖励 r_i（详见 **??**）。
- **数据封装**: 保存 prompt、K 条响应、token 级对数概率、mask 与奖励，组成后续训练批次；prompt 最长 512 token，响应最长 2048 token。

3. 对数概率计算

- **旧策略对数概率 old_log_prob**: 使用未更新的 Actor 对 (prompt, response) 再次前向传播获取。原因：*(i)* 高性能推理引擎通常不保存完整 token level log prob；*(ii)* 直接保存可能受数值波动与并行策略影响。
- **参照策略对数概率 ref_log_prob**: Reference Model 对同一批数据计算 log prob，用于 KL 约束。首轮迭代时与 old_log_prob 相同。

4. 奖励与优势估计

- **奖励整合**: VeRL 支持外部 RM 或自定义函数（本研究采用规则奖励）。
- **优势函数**: 调用 `compute_advantage`，依据 GRPO 公式（公式 *Y*）在组内（同一 prompt 的 K 样本）计算相对优势, 并可选择标准化 (`norm_adv_by_std_in_grpo`)。

5. 策略更新（Mini batch 内循环） 对每个 mini batch 执行：

- 计算新策略对数概率 new_log_prob。
- **策略梯度损失** pg_loss：由 `compute_policy_loss` 依据公式 *Z* 计算，含 cliprange 等裁剪。

- **熵正则**：鼓励探索。`entropy_coeff` 控制权重。
- **KL 惩罚**：当前策略与 Reference 之间的 KL 散度，乘 `kl_loss_coef` 加入总损失。
- 反向传播并更新 Actor 参数。

完成全部 mini batch 后，同步最新权重，进入下一大迭代的经验收集。

5. Metric 评测

由于每个多选题示例都有一个或多个答案，因此我们评估了实例和选项级别的性能。给定的答案 $y_{\text{pred}}^{(i)}$ 和地面真相 $y_{\text{gold}}^{(i)}$ 对于第 i 个问题，两个指标表示如下：

精确匹配：这衡量了预测答案与正确答案完全匹配的实例的比例。

$$\text{精确匹配} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(y_{\text{pred}}^{(i)} = y_{\text{gold}}^{(i)})$$

其中 $\mathbf{1}(\cdot)$ 是指示函数，如果内部的条件为 true，则等于 1，否则为 0。它评估模型是否可以为问题提供完全正确的答案。对于多选题，这意味着预测的选项集合与真实的选项集合完全一致。

召回和精确度：我们根据模型对单个答案选项的预测来计算召回和精度得分。令 N 为问题总数。对于第 i 个问题，令 P_i 为模型预测的正确选项集合，G_i 为实际的正确选项集合。则选项级别的精确度 (Precision) 和召回率 (Recall) 定义为：

$$\text{Precision}_{\text{option}} = \frac{\sum_{i=1}^{N}|P_i \cap G_i|}{\sum_{i=1}^{N}|P_i|}$$

$$\text{Recall}_{\text{option}} = \frac{\sum_{i=1}^{N}|P_i \cap G_i|}{\sum_{i=1}^{N}|G_i|}$$

其中 $|S|$ 表示集合 S 的基数。如果某个问题模型没有预测任何选项（即 $|P_i| = 0$），则该问题对精确度的分子和分母的贡献均为 0。通常也会计算 F1 分数：

$$F1_{\text{option}} = 2 \cdot \frac{\text{Precision}_{\text{option}} \cdot \text{Recall}_{\text{option}}}{\text{Precision}_{\text{option}} + \text{Recall}_{\text{option}}}$$

section 实验设计与结果分析本节将深入阐述我们在法律多选推理任务中所进行的一系列实验探索与策略优化过程。我们系统地考察了奖励函数设计、强化学习策略应用、基于蒸馏数据的模型微调，以及对模型输出结构性的增强分析。核心目标在于全面评估多阶段训练范式，特别是结合了监督微调（SFT）与强化学习（RL）的策略，对于提升大语言模型在复杂法律推理任务中的性能表现及其生成内容结构化表达能力的具体影响。

subsection 奖励函数设计与长度控制机制为了有效引导模型在法律多选任务中达成更高的准确性并生成结构清晰的响应，我们精心设计并对比了多种奖励函数。这些奖励函数主要涵盖完全匹配奖励（Exact Match Reward）、部分匹配奖励（Partial Match Reward）以及格式奖励（Format Reward）。通过细致的实验，我们分析了这些不同奖励机制对模型学习行为和最终性能输出的具体影响。

begintable[h]

centering

caption 不同奖励类型定义及其主要特点

begintabularl|p8cm|c

hline 奖励类型 定义与说明 特点

hline Exact Match 当模型预测的答案选项集合 $S\_textpred$ 与真实的答案选项集合 $S\_texttrue$ 完全一致时，奖励为 1，否则为 0。评价标准最为严格，可能限制模型在初期学习阶段的探索性。
Partial Match 该奖励旨在同时惩罚错误选择的选项（False Positives, FP）和未能选择的正确选项（False Negatives, FN），具体计算公式见下文详述。具有可调节的惩罚力度，允许在精确率和召回率之间进行权衡。
Format Reward 要求模型的输出响应必须包含明确的
texttt 思考（reasoning/thought process）和
texttt 回答（final answer）两个结构化部分。有助于在训练早期加速模型对期望输出格式的学习，提升收敛速度。

hline
endtabular
endtable

paragraph 准确性奖励的具体定义：我们主要采用了两种衡量模型预测准确性的奖励机制：

beginenumerate[label=(
arabic)]
item 完全匹配奖励（Exact Match Reward）：该奖励函数 $R\_mathrmexact$ 的定义如下：

$$R_{\text{exact}}(S_{\text{pred}}, S_{\text{true}}) = \begin{cases} 1, & \text{若 } S_{\text{pred}} = S_{\text{true}} \\ 0, & \text{其他情况} \end{cases}$$

其中，$S\_mathrmpred$ 代表模型预测的多选选项集合，$S\_mathrmtrue$ 代表标准答案的选项集合。
‘‘‘

部分匹配奖励（Partial Match Reward）：为了更细致地评估模型的表现，尤其是在多选题场景下，我们引入了部分匹配奖励 R_{partial}。该奖励对模型的漏选（FN）和错选（FP）进行惩罚，其计算方式为：

$$R_{\text{partial}} = \max\left(0,\ 1 - x \cdot (\text{FP} + \text{FN})\right), \quad \text{其中 } x \geq 0.5$$

在此公式中，FP 表示模型错误选择的选项数量（False Positives），FN 表示模型未能选择的正确选项数量（False Negatives）。系数 x 用于控制惩罚的强度，其取值不小于 0.5，以便给予足够的惩罚信号。‘‘‘

endenumerate

paragraph 不同惩罚系数下的实验结果对比：我们针对部分匹配奖励中的惩罚系数 x 进行了实验对比，同时引入完全匹配奖励作为基准。结果如下表所示：

begintable[h]
centering
caption 不同惩罚系数 x 配置下模型在验证集上的性能表现
begintabularl|c|c|c
hline 惩罚系数 x / 奖励类型 精确匹配率 (mean@1) 召回率 (worst@1) 精确度 (worst@1)

hline Exact Match 0.582 0.770 0.777
$x = 0.7$ (Partial Match) 0.569 0.790 0.777
$x = 0.9$ (Partial Match) 0.564 0.794 0.772

hline
endtabular
endtable

textbf 实验观察：从上表数据可以看出，在部分匹配奖励机制下，当增大惩罚强度系数 x 时（例如从 $x = 0.7$ 增加到 $x = 0.9$），模型的召回率（worst@1）呈现上升趋势，这表明模型倾向于选择更多的选项以避免漏选。然而，与此同时，精确度（worst@1）和精确匹配率（mean@1）均有所下降，说明模型可能因此选择了更多不正确的选项。综合各项指标，当惩罚系数 $x = 0.7$ 时，模型在召回率和精确度之间取得了相对最佳的平衡，其精确度与完全匹配奖励持平，同时召回率有所提升。

paragraph 格式奖励的作用、分析与最终选择：先前研究如 Simple-R1 citezeng2025simplerl 指出，虽然格式奖励（Format Reward）能够引导模型快速学习并遵循特定的输出结构，从而加速早期收敛，但它也可能在一定程度上抑制模型深层推理能力的充分发展。为了验证这一点，我们进行了一项对比实验，在训练过程中移除了格式奖励。实验结果揭示：

beginitemize
item 即便在没有显式格式奖励引导的情况下，模型依然能够学习并自主生成包含“思考-回答”这样结构清晰的响应格式。
item 在移除了格式奖励后，模型的最终准确率从基准的 58.2
enditemize

beginfigure[h]
centering
includegraphics[width=0.8
textwidth]figures/format.png
caption 移除格式奖励后模型准确率（Exact Match mean@1）随训练步数的变化情况

labelfig:format
endfigure

尽管上述分析揭示了移除格式奖励后准确率的微降，但考虑到格式奖励在加速模型早期学习特定输出结构方面的积极作用（如表 1 所述），以及在我们的最终应用场景中，结构化的输出对于可解释性和用户体验至关重要。因此，在权衡了早期收敛速度、输出结构规范性以及对推理能力潜在的轻微影响后，我们的最终奖励函数选择整合了准确性奖励和格式奖励：

beginequation R = R_
mathrmaccuracy + R_
mathrmformat.
endequation 其中，$R\_mathrmaccuracy$ 可以是 $R\_mathrmexact$ 或 $R\_mathrmpartial$，具体选择取决于特定实验阶段的侧重点。

subsectionZero-RL：无监督微调（SFT）的强化学习探索为了探究强化学习（RL）是否能够直接提升预训练模型在法律推理任务上的能力，而无需任何有监督微调（Supervised Fine-Tuning, SFT）作为预热阶段，我们尝试将 GRPO (Generalized Reinforcement Learning from Online Preference Optimization) 和 Dr.GRPO (Denoising Reward GRPO) 算法直接应用于 Qwen-2.5-7B-Instruct 模型。

beginitemize
item
textbf 实验目的：核心目的是评估在不经过 SFT 的情况下，RL 算法能否从零开始有效地优化模型在复杂法律推理任务上的表现。
item
textbf 实验结果：结果表明，这种“从零开始”的强化学习（Zero-RL）方式面临显著挑战。模型的收敛速度极为缓慢，并且在推理准确率（以 mean@1 指标衡量）上的提升非常有限，大致停留在 20
enditemize

subsectionSFT + RL：基于蒸馏数据的冷启动策略鉴于 Zero-RL 的局限性，我们转向采用一种“监督微调 + 强化学习”（SFT + RL）的多阶段训练策略，并通过知识蒸馏的方式为 SFT 阶段提供高质量的训练数据，以实现更有效的“冷启动”。

paragraph 蒸馏数据准备：我们采用了由 DeepSeek-R1 模型在 JEC-QA (Judicial Examination Comprehension Question Answering) 数据集上生成的思维链（Chain-of-Thought, CoT）响应作为 SFT 阶段的蒸馏数据。选择 DeepSeek-R1 是因为其在相关任务上已展现出较强的推理和生成能力，其 CoT 响应能够为我们的目标模型提供高质量的推理示范。

paragraph 详细训练流程：整个训练流程分为 SFT 阶段和 RL（GRPO）阶段：

beginitemize
item

textbfSFT 阶段：
beginitemize
item
textbf 基础模型：Qwen-2.5-7B-Instruct。
item
textbf 主要训练参数：训练进行了 2 个周期（epoch），批处理大小（batch size）设置为 32，学习率（learning rate）采用 1
$times10^{-5}$，上下文长度（context length）限制为 1024 tokens。
item
textbfSFT 效果：通过在该蒸馏数据上进行 SFT，模型的精确匹配率（mean@1）从初始的 29.5
enditemize
item
textbfGRPO 强化学习阶段：
beginitemize
item
textbf 奖励函数配置：在此阶段，我们同时使用了前述定义的完全匹配奖励（$R\_mathrmexact$）和格式奖励（$R\_mathrmformat$）来指导模型的学习。
item
textbf 主要训练参数：Token 截断长度设置为 1024，PPO 算法中的裁剪比率（clip ratio）初始值设为 0.6。
item
textbfGRPO 效果：经过 GRPO 阶段的强化学习，模型的精确匹配率（mean@1）从 SFT 后的 42.4
enditemize
enditemize

paragraph 不同训练阶段性能对比：为了更清晰地展示多阶段训练的效果，我们对比了模型在不同阶段于验证集上的准确率表现：

begintable[h]
centering
caption 不同训练阶段模型在验证集上的准确率对比（mean@1）
labeltab:acc-distill-sft-grpo
begintabularl|c
hline 模型/训练阶段 准确率 (mean@1)

hline DeepSeek-R1 (教师模型，参考性能) 67.4Qwen-2.5-7B-Instruct (SFT 后) 42.4Qwen-2.5-

7B-Instruct (SFT + GRPO, 1024 tokens 上下文) 53.2Qwen-2.5-7B-Instruct (SFT + GRPO, 4096 tokens 上下文) 50.9

hline

endtabular

endtable

从上表数据可以看出，SFT 阶段为模型带来了显著的初始性能提升。随后的 GRPO 强化学习阶段进一步优化了模型表现，使得准确率提升了近 11 个百分点。值得注意的是，当我们将 GRPO 训练和评估时的上下文 Token 长度从 1024 扩展到 4096 时，性能反而有所下降（从 53.2

subsection 结构性增强与案例分析除了关注准确率等硬性指标外，我们还深入分析了不同训练阶段模型生成响应的结构性特征，旨在进一步评估强化学习与监督微调相结合的策略对模型输出质量的综合提升效果。良好的结构性不仅能提升响应的可读性和可理解性，也是衡量模型是否能生成高质量法律文书的关键。

beginitemize

item

textbf 结构性评估指标包括:

beginitemize

item

textbf 编号段落的使用: 是否倾向于使用有序列表（如 1, 2, 3... 或 a, b, c...）来组织论点。

item

textbf 逻辑连接词的频率: 如“首先”、“其次”、“综上所述”、“因此”等词语的使用情况，这些词语有助于构建清晰的逻辑流。

item

textbf 法规条文的引用: 是否能够准确且适当地引用相关的法律法规条文支撑其论证。

item

textbf 解释性与分析性句式的出现: 是否包含对案情、法理进行分析、阐释的句式，而不仅仅是简单罗列。

enditemize

item

textbfGRPO 阶段对结构性的影响: 通过对比分析，我们发现经过 GRPO 强化学习阶段训练后的模型，在输出响应时更倾向于生成具有清晰逻辑层次和良好组织结构的文本。这具体表现为:

beginitemize

item

textbf 平均响应 Token 数增加: 模型生成的解释和论证内容更为详尽。

item

textbf 逻辑衔接词使用更频繁: 表明模型试图构建更连贯、更易于理解的推理过程。
item 图
reffig:grpo_distill 中的收敛曲线也伴随着结构性指标的改善趋势（尽管图中主要展示了奖励/准确率，但其背后反映了更优的策略，其中包含了对期望结构的间接学习）。
enditemize
item
textbf 案例分析: 对具体案例的输出进行人工检查和分析显示，经过 RL 优化后的模型所生成的推理链条，不仅在形式上（如遵循“思考-回答”模式），在内容和语言风格上也更贴近于司法实务中的专业表达，例如能够更自然地组织论点、阐述理由。
enditemize
beginfigure[h]
centering
includegraphics[width=0.8
textwidth]figures/GRPO_DeepSeek-R1-Distill-Qwen-7B.png
captionGRPO 阶段训练收敛曲线（如奖励或准确率）以及伴随的结构性增强趋势示意 (具体结构性指标变化趋势需结合详细日志分析)
labelfig:grpo_distill
endfigure

paragraph 本章结论：综合上述实验设计与结果分析，我们可以得出结论：采用“知识蒸馏辅助 SFT，继以强化学习优化”的多阶段训练策略，对于提升模型在法律多选推理任务上的综合性能具有显著的增益。此策略不仅有效提升了模型的预测准确性，更重要的是，它在模型的结构化表达能力与推理准确性之间取得了良好的平衡。模型在 SFT 阶段学习领域知识和基本任务范式，在 RL 阶段通过精心设计的奖励函数进一步优化其决策逻辑和输出质量，尤其是生成内容的逻辑性和专业性。

未来的研究工作将聚焦于探索更为精细化的结构性奖励函数设计，例如直接对论证的充分性、逻辑链的完整性进行建模并予以奖励。此外，研究混合不同类型提示（如 CoT 提示、摘要提示等）的动态调度机制，以期进一步提升模型的泛化能力、鲁棒性以及最终输出的可解释性，使之能更好地服务于真实的法律应用场景。

参考文献

[1] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv preprint arXiv:2501.12948, 2025.

[2] YANG A, YANG B, ZHANG B, et al. Qwen2. 5 technical report[J]. arXiv preprint arXiv:2412.15115, 2024.

[3] SEED B, YUAN Y, YUE Y, et al. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning[J]. arXiv preprint arXiv:2504.13914, 2025.

[4] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [J]. arXiv preprint arXiv:1707.06347, 2017.

[5] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[J]. Advances in Neural Information Processing Systems, 2023, 36: 53728-53741.

[6] HUANG Q, TAO M, ZHANG C, et al. Lawyer llama technical report[J]. arXiv preprint arXiv:2305.15062, 2023.

[7] CUI J, LI Z, YAN Y, et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]. CoRR, 2023.

[8] ZHOU Z, SHI J X, SONG P X, et al. Lawgpt: A chinese legal knowledge-enhanced large language model[J]. arXiv preprint arXiv:2406.04614, 2024.

[9] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models[J]. arXiv preprint arXiv:2402.03300, 2024.

[10] YU Q, ZHANG Z, ZHU R, et al. Dapo: An open-source llm reinforcement learning system at scale[J]. arXiv preprint arXiv:2503.14476, 2025.

[11] LIU Z, CHEN C, LI W, et al. Understanding r1-zero-like training: A critical perspective[J]. arXiv preprint arXiv:2503.20783, 2025.

[12] EL-KISHKY A, WEI A, SARAIVA A, et al. Competitive programming with large reasoning models[J]. arXiv preprint arXiv:2502.06807, 2025.

[13] XIE T, GAO Z, REN Q, et al. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning[J]. arXiv preprint arXiv:2502.14768, 2025.

[14] ZHONG H, XIAO C, TU C, et al. JEC-QA: a legal-domain question answering dataset[C]// Proceedings of the AAAI conference on artificial intelligence: vol. 34: 05. 2020: 9701-9708.

[15] ZENG W, HUANG Y, LIU Q, et al. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild[J]. arXiv preprint arXiv:2503.18892, 2025.

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：

日期：　　年　　月　　日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：　　导师签名：

日期：　　年　　月　　日