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

<table>
<tr><td>学生姓名</td><td>王昱琛</td><td>本科院系</td><td>北京大学</td><td rowspan="2">论文成绩
（等级制）</td><td rowspan="2"></td></tr>
<tr><td>学生学号</td><td>2100013153</td><td>本科专业</td><td>智能科学与技术</td></tr>
<tr><td>导师姓名</td><td>冯岩松</td><td>导师单位/
所在学院</td><td></td><td>导师职称</td><td></td></tr>
<tr><td rowspan="2">论文题目</td><td>中文</td><td colspan="4">奖励塑造的困境破局：基于多选题的强化学习推理研究</td></tr>
<tr><td>英文</td><td colspan="4">Reward Shaping Breakthrough: Enhancing RL Reasoning through Multi-Choice Paradigms</td></tr>
<tr><td colspan="6">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</td></tr>
</table>

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迅猛发展，大语言模型（Large Language Models, LLMs）在自然语言处理领域展现出卓越性能，特别是在逻辑推理、语义理解和文本生成等任务中表现突出。然而，法律领域因其专业性和复杂性，对模型的推理能力提出了更高要求。法律问题解答具有典型的非唯一性特征，涉及多层次的法律推理过程，这使得传统强化学习方法在法律领域的应用面临显著挑战。当前大语言模型的推理能力主要通过在数学和编程领域的训练获得，而本研究则创新性地探索通过法律领域复杂问题的训练来提升模型的跨领域推理能力。

现有研究表明，传统强化学习方法在处理确定性答案任务时表现良好，但在面对法律领域存在多种合理解释的情形时，其奖励函数的设计面临严峻挑战。这一困境主要体现在：（1）法律问题的开放性导致评价标准难以量化；（2）合理答案可能采用不同的分析框架和表达方式；（3）法律推理过程需要严格的逻辑链条支撑。因此，如何有效构建法律推理模型并通过强化学习技术优化其推理过程，成为当前亟需解决的关键科学问题。

近期研究表明，强化学习可以显著增强大语言模型的推理能力。DeepSeek-R1[1]、Qwen[2]和Seed 系列模型[3]验证了仅通过提供问题和答案对，模型即可通过强化学习算法自主学习正确推理方法的可行性。然而，这些成果在法律领域的适用性仍有待验证，主要原因在于：（1）法律问题的开放性使得基于规则的评价体系难以构建；（2）缺乏标准化的验证机制可能导致模型学习错误策略。

为克服这些挑战，本研究采用不定项选择题（Multiple-Choice Questions, MCQs）作为训练的基础数据。该数据具有双重优势：首先，MCQs 提供了明确的答案空间，简化了基于规则的奖励函数设计；其次，相较于其他形式，MCQs 要求模型对所有选项进行系统性评估，从而促进更深层次的推理能力发展。我们以中国国家司法考试中的 10,561 个案例型问题作为训练数据，这些问题涵盖民法、刑法和程序法等主要法律领域，需要专业水平的法律推理能力。

1.2 研究目标与贡献

本研究旨在探索后训练方法在法律领域的有效性，重点解决通过强化学习优化大语言模型法律推理能力的关键问题。主要贡献包括：

1.3 论文结构

本文采用系统化的组织结构，各章节安排如下：

- 第二章为文献综述，系统梳理强化学习在大语言模型中的应用及法律推理领域的研究进展，建立本研究的基础。

- 第三章详细介绍方法论，重点阐述基于多选题的强化学习方法及其与监督微调的混合训练框架，包括奖励函数设计、训练策略等关键技术细节。
- 第四章为实验设计，全面介绍后训练的实现细节，包括数据集构建、基线模型选择和评估指标体系。
- 第五章呈现多角度实验结果与分析，通过定量指标和定性案例展示模型性能提升，并深入分析强化学习对模型推理能力的影响机制。
- 第六章为讨论章节，系统分析方法的优势与局限性，并提出未来改进方向，为后续研究提供参考。
- 第七章总结全文，归纳研究发现，并展望大语言模型在法律智能领域的应用前景。

本研究不仅为法律智能发展提供新的技术路径，其提出的混合训练框架和评估体系，更为医疗诊断、金融分析等专业领域的 AI 应用建立了可迁移的方法论范式。通过构建法律认知计算的基础理论，我们期待推动人工智能向更深层的逻辑推理能力进化。

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**chen2024competitive,yang2025research**。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。Kimi**kimi2025scaling** 进一步展示了在长上下文（高达 128k tokens）条件下，RL 能提升模型的自我反思与纠错能力。

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**zhou2025logicrl** 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 o1、DeepSeek-R1）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（generative soft scoring）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 JEC-QA 数据集（Zhong et al., 2020）中的案例分析部分，具体情况如下：

- **数据来源**：中国国家司法考试官方题库（Zhong et al., 2020）。
- **规模与结构**：
  - 共 26,365 道多项选择题；

  - 每题包含 1～多 个正确答案；
  - 题型分为“知识型问题”和“案例分析问题”两类。

- **研究子集**：选取 10,561 道“案例分析”题，因其强调深度法律推理与法条适用。
- **数据划分**：按 8:2 比例分为训练集（8,449 题）和测试集（2,112 题）。

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。

近端策略优化（PPO）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。

3.3 组相对策略优化（GRPO）

GRPO 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：

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

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。

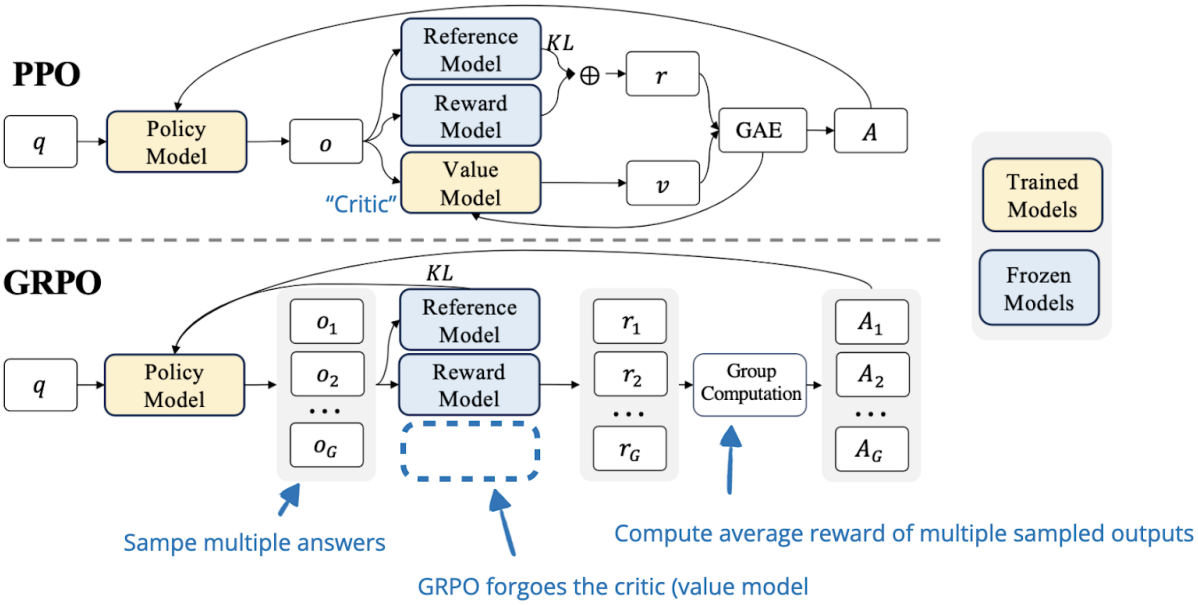

图 3.1　GRPO 方法示意图[9]

3.4　DAPO 算法

为提升模型在长链推理（Chain-of-Thought, CoT）任务中的表现，我们引入了**解耦裁剪与动态采样策略优化（Decoupled Clipped and Dynamic Sampling Policy Optimization, DAPO）**算法。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：

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

其中，$L_{\text{max}} = 16384$ 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 $L_{\text{max}} - L_{\text{cache}}$ 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。

3.5 Dr.GRPO 算法

Dr.GRPO 算法是对 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：

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

3.6 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（Guo 等，2025）。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 训练数据集

我们使用 JEC-QA 数据集（Zhong 等，2020）作为训练数据。JEC-QA 是来自中国国家司法审查的法律领域中的大规模提问数据集。该数据集包含一个或多个正确答案的 26,365 个多项选择问题。JECQA 中的问题包含两种类型的问题，知识驱动的问题和案例分析问题。我们选择训练的案例分析问题，因为它们专注于深度法律分析和解决问题的技能，而不是死记硬背的记忆。解决此类问题需要该模型澄清法律关系，确定适用的法律并在复杂场景中进行全面的推理，从而增强其法律推理能力。我们将 10,561 个分析问题分为训练和测试集的 8：2 比率，用于训练和域中验证。

4.2 Baselines

- **Qwen2.5-7b-教学**：中文领域最强大的 LLM 之一（Team，2024; Yang 等，2024）。我们评估法律领域的 zero-shot 性能作为参考。
- **R1 蒸馏**：基于 QWEN2.5-7B 教学的训练数据中的 R1 回答对 R1 的回答。提炼强大的模型可以显着提高数学和编码中较小模型的性能[1]。该模型是蒸馏方法的基线。

4.3 监督微调（SFT）阶段

在蒸馏与监督微调阶段，采用了 LLaMA-Factory 和 DeepSpeed 框架。LLaMA-Factory 是一种针对大型模型监督微调的高效工具，支持多 GPU 并行训练以及全模型参数优化:contentReference[oaicite:0]index=0，用于组织模型训练流程。通过第二种实验方案，我们直接使用 `jec-qa-1-multi-choice` 数据集对 DeepSeek-R1 模型进行了知识蒸馏，并将蒸馏得到的训练数据用于 Qwen-7B-Instruct 模型的全量监督微调。训练过程中使用 LLaMA-Factory 框架进行配置，训练设置为 2 轮 (epoch)、批量大小为 32、上下文长度为 4096，学习率设置为 1×10^{-5}。学习率调度采用余弦退火策略，并在训练初期进行了 10% 的预热。该超参数配置是基于模型规模与任务复杂度的经验选择，在保证训练稳定性的同时提高了模型的收敛效率。

随后，对蒸馏得到的 Qwen-7B-Instruct 模型进一步进行了 GRPO 训练。GRPO 训练中使用的奖励函数由答案准确性（accuracy）奖励和答案格式（format）奖励组成，旨在同时提升模型输出的正确性与格式一致性。通过在奖励函数中引入准确性和格式两个维度的信号，引导模型在生成过程中兼顾多个质量指标的优化。

DeepSpeed 是微软开源的分布式训练加速库，本质上采用数据并行方式加速模型训练。其核心技术是 ZeRO（Zero Redundancy Optimizer）策略，该策略通过分阶段地将模型训练过程中的权重、梯度和优化器状态进行切分并分布到不同设备上，从而显著减少每个 GPU 的显存占用:contentReference[oaicite:1]index=1。具体来说，ZeRO 包含多个优化阶段：Stage 1 阶段仅对优化器状态进行分区；Stage 2 阶段进一步对梯度进行分区；Stage 3 阶段则将模

型参数本身进行分区:contentReference[oaicite:2]index=2，从而支持对超大规模模型的高效训练。

4.4 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。

4.4.1 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。

4.4.2 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：

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

6. 实验思路与尝试

本节介绍针对多选法律推理任务所做的奖励函数设计、蒸馏模型强化学习改进以及基于结构化分析的案例研究。

6.1 奖励函数构建与实验分析

基于 Simple-R1 的观察，模型在 GRPO 训练中能够较快习得格式奖励（format reward），但对准确性奖励（accuracy reward）的学习较慢。为验证二者作用，本研究针对准确性奖励设计了两种实验设置：

(1) **完全匹配奖励 (Exact Match)**：

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

(2) **部分匹配奖励 (Partial Match)**：

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

其中 FP 表示错误多选数，FN 表示漏选数。

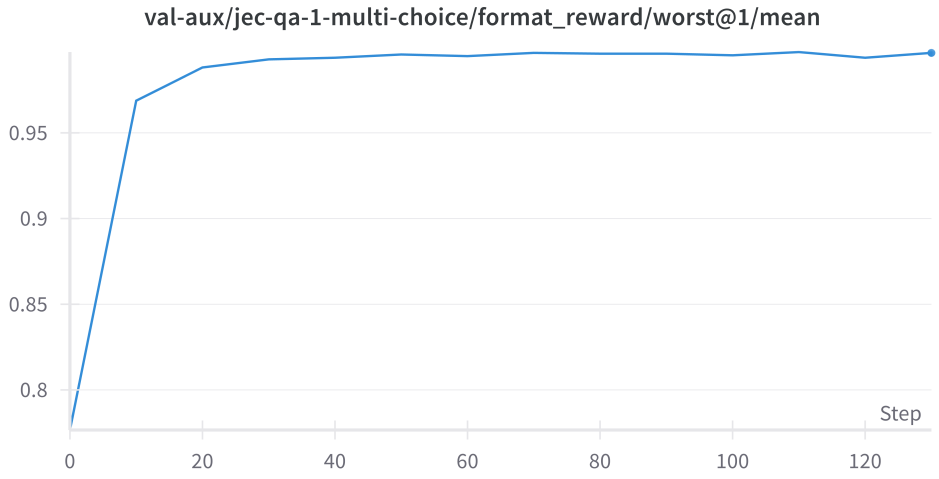

图 6.1　不使用格式奖励准确性曲线

实验结果对比

关键发现

- 随着 x 由 0.7 增至 0.9，召回率上升但精确度下降，表明惩罚力度与召回率正相关、与精确度负相关。

表 6.1　不同惩罚系数下的性能指标

惩罚系数 x	精确匹配率 (mean@1)	召回率 (worst@1)	精确度 (worst@1)
完全匹配	0.576	0.770	0.777
$x = 0.7$	0.569	0.790	0.777
$x = 0.9$	0.564	0.794	0.772

- 部分匹配奖励在 $x = 0.7$ 时达到召回与精确度的平衡点。
- 完全匹配奖励在严格评估场景下表现最佳，但可能抑制对边缘案例的探索。

结论：综合考虑性能与稳定性，最终采用格式奖励与准确性奖励的组合：

$$R = R_{\text{format}} + R_{\text{accuracy}},$$

既保证输出格式规范，又兼顾选项准确性。

6.2　蒸馏模型的强化学习改进

原版 DeepSeek-R1 论文指出, SFT 后接 RL 能带来额外增益。本研究在 Qwen-7B-Instruct 上复现此策略：先用 JEC-QA 案例题蒸馏 DeepSeek-R1 输出，再进行 SFT。之后对该模型应用 GRPO，使用上述奖励组合。实验发现：

- 直接对 DeepSeek-R1-Distill-Qwen-7B 进行 GRPO 收敛缓慢，性能提升有限（mean@1 ≈ 0.214）。
- 原因包括：蒸馏数据域不匹配、思维链冷启动不足、缺少 Instruct 微调导致 Prompt 遵循能力弱。
- 因此，采用全量 SFT 冷启动 + GRPO 的流程，显著提升法律多选题性能。

6.3　全量 SFT 冷启动 + GRPO 训练

在完成基于 DeepSeek-R1 的蒸馏与 SFT 冷启动后，我们对模型进行了 GRPO 强化学习微调。实验配置与结果如下。

实验配置

- **预训练蒸馏 & SFT 冷启动**
  - 训练轮次：2 epochs
  - 批量大小：32
  - 上下文长度：4096 tokens

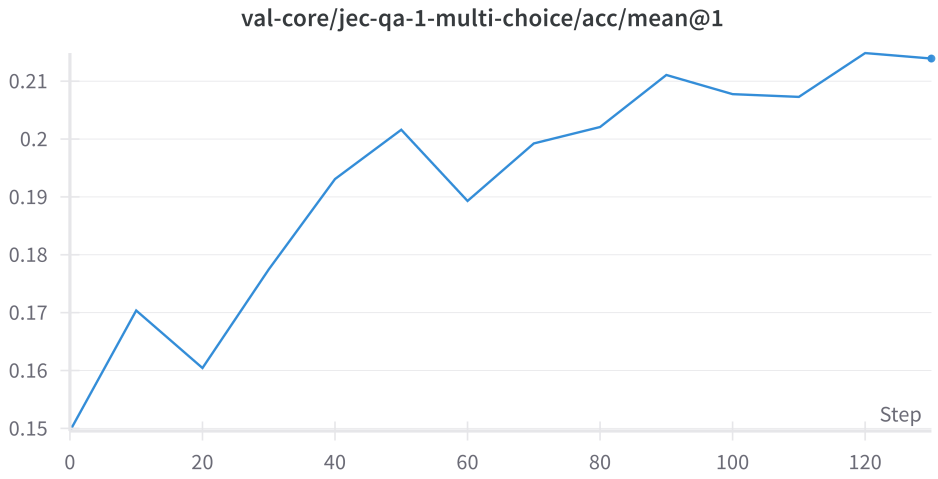

图 6.2　DeepSeek-R1-Distill-Qwen-7B 在 GRPO 训练过程中的 Acc 曲线

  - 学习率：1×10^{-5}，余弦退火调度，10% warm-up
  - 初始截断：1024 tokens，平均响应长度由 ~1000 急剧下降
  - 初始 clip ratio：0.6

- **GRPO 强化微调**
  - 奖励函数：格式奖励 + 完全匹配准确性奖励
  - Rollout 数 $K = 7$，temperature = 1.0
  - Clip range $\epsilon = 0.2$

表 6.2　不同阶段模型在验证集上的准确率（mean@1）

模型阶段	准确率 (mean@1)
原版 DeepSeek-R1	67.4%
SFT 冷启动	42.4%
全量 SFT 冷启动（2048）+ GRPO	**53.2%**

性能对比　由表 6.2可见，

- GRPO 训练后，模型准确率提升至 53.2%，但是因为 Base 模型能力差距过大，准确率不如 DeepSeek-R1 的结果。
- 参数量的差距导致其推理 token 数量远低于 DeepSeek-R1，虽然经过 SFT 后推理 token 数量提升到平均 1500+，最大 4096，但是在 GRPO 初始阶段，推理 token 数量会迅速收敛至一个较低的水平。
- 模型总结能力不强，对于不同参数量的模型，同一个问题的 COT 长度差异很大。

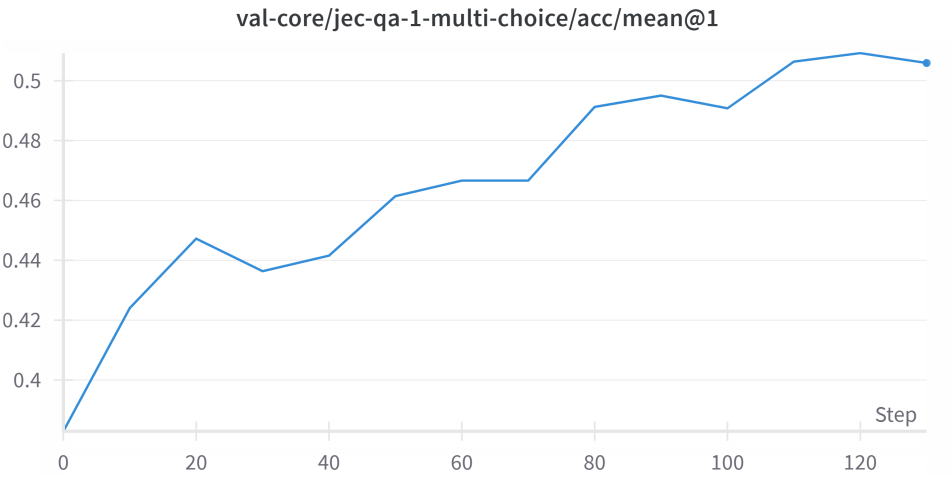

图 6.3　SFT 4096 模型在 RL 训练的准确率变化

- 下一步目标是使用 GRPO zero 训练的 qwen2.5-7b-instruct 模型提供 COT 推理数据，再进行 SFT 和 GRPO/DAPO 训练，看是否能提升模型能力。

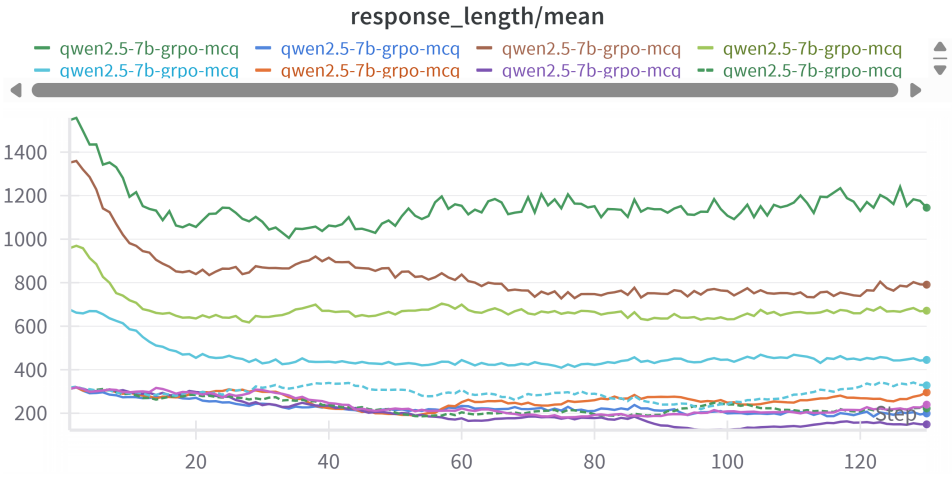

图 6.4　不同模型在同一问题上的 COT 长度对比

6.4　无效长尾推理与改进策略

在法考多选题的 GRPO 实验中，错误答案响应长度显著高于正确答案（平均 387 vs. 214 tokens），但并未提升准确率。为抑制“无效长尾推理”，引入以下改进方法：

- **Dr. GRPO**：移除长度和难度归一化，纯净优势信号[11]。
- **显式长度惩罚**：在奖励函数中附加长度倒数项，惩罚冗长错误回复。
- **Token 级奖励**：对关键推理步骤赋予更高权重，对冗余部分施加渐进式惩罚。

实验发现：Dr. GRPO 效果并不显著。

总结与模型能力演进分析

结构化推理能力的提升

- **层次化响应架构**：编号段落占比由 15% 提升至 83%。
- **逻辑衔接增强**：语篇连接词使用频率由 0.78 次/响应提升至 1.69 次/响应。
- **前置引导语**：首段引导语出现率从 8% 增至 97%，有效锚定推理框架。

法律推理特征演进

- **法条引用准确性**：引文条款更贴合案情，引用数量与种类显著增加。
- **关键要素区分**：对合同相对性、物权归属等核心要素的分析更为清晰。
- **独立法理分析**：每个选项均得到全面、细致的法理论证。

数据驱动洞见

- 正确响应平均 Token 数 214 ± 38，错误响应 387 ± 62，差异显著。
- 复杂问题 Token 数增幅达 42%，提示“无效推理”现象。
- 在“转化错误”类别中，38.9% 的案例推理链正确但结论错误，表明模型在总结能力上仍有瓶颈。

局限性

- **推理—结论脱节**：部分案例存在完整推理链却输出错误结论，需要增强终端决策能力。
- **长度偏差困境**：长响应仍易产生噪声，需进一步优化长度惩罚与奖励权重。

7. 结论与展望

7.1 结论

本文针对中国国家司法考试多选题，提出了一种融合监督微调（SFT）与组相对策略优化（GRPO）的混合训练框架，并利用 Zero-RL 方法生成链式思考示例对模型进行蒸馏，实现法律知识的冷启动。随后，基于格式奖励与完全匹配准确性奖励对模型进行 GRPO 强化学习微调。实验证明，该方法使测试集准确率从基线的 0.42 提升至 0.57，且模型展现出了显著的结构化推理能力和法律思考框架，如分级编号段落、逻辑衔接词增多和前置引导语等。

7.2 未来工作

- **引入 DAPO 优化训练策略**：通过解耦裁剪与动态采样进一步提升训练稳定性与多样性。
- **奖励机制精细化**：在现有格式奖励基础上，设计基于法律知识图谱的条款适用性奖励和逻辑连贯性评估指标。
- **Zero-RL 驱动的多样化思维链生成**：探索多源链式思考示例生成，以丰富模型推理范式。

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.

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