

简单规则释放复杂智能

——小模型创意生成的 GRPO 路径研究”

摘要：强化学习（RL）作为 2024 年图灵奖的授予领域，其在现代人工智能，特别是大型语言模型（LLM）训练中的核心作用日益显现。从广泛应用的 RLHF（基于 PPO）到引人注目的 GRPO 算法，RL 为引导 LLM 生成更符合人类偏好或特定目标的输出提供了强大工具。然而，如何将这些先进的 RL 技术，尤其是像 GRPO 这样的新算法，有效应用于参数量较小的语言模型，并激发其在创意文本生成等复杂任务上的潜力，仍然是一个值得探索的问题。本研究针对此问题，基于 GRPO 算法，利用大量高质量的数据，对小参数模型改进了训练范式。该范式首先利用原始创意文本进行初步监督微调（SFT），随后提取文本关键词并进行第二阶段的 SFT 以增强模型对关键信息的捕捉与表达能力。在此基础上，我们运用改进的 GRPO 算法进行强化学习，本研究不仅为 GRPO 算法的应用拓展提供了实证支持，加深了对其工作原理的理解，也为优化小参数量语言模型的训练范式贡献了新的见解和技术路径。

关键词：大语言模型；强化学习；监督微调；GRPO 算法

目录

1 引言	3
1.1 研究介绍	3
1.2 成果介绍	4
2 方法范式	5
2.1 数据处理与构建	5
2.1.1 基本清洗	5
2.1.2 深度加工	6
2.2 Policy Optimization	8
2.2.1 从策略梯度到 PPO	9
2.2.2 Rule Based GRPO	14
2.3 挫折、回退、更新	17
3 讨论	20
3.1 研究局限	20
3.2 展望改进	21
4 结论	22
5 参考文献	23

1 引言

1.1 研究介绍

本研究的核心目标是开发并优化一个能根据用户指定关键词生成高质量创意文本（如歌词）的小参数量语言模型。为实现这一目标，我们系统地开展了数据构建、模型训练与效果评估工作。

在数据层面，我们首先通过网络爬虫及 API 接口收集了约 10 万条歌词等创意文本。原始数据经过严格的清洗与去重处理。随后，利用先进的 DeepSeek V3-0324 模型为每条文本提取核心关键词。值得注意的是，我们还利用该模型根据提取的关键词，生成了 1000 条包含思维链（Chain-of-Thought, CoT）的高质量样本，这些 CoT 数据旨在引导模型学习更具逻辑性和结构化的生成过程，为后续的监督微调（SFT）与强化学习（RL）训练提供了关键支持。

在模型层面，考虑到性能与效率的平衡，经过初步实验筛选，我们选用 Qwen-0.5B 和 DeepSeek-R1-Distill-Qwen-1.5B 这两个参数量相对较小的模型作为研究的基础。基于这两个模型，我们探索了不同的训练策略，最终生成并评估了四个代表性模型：

- Qwen_LoRA: 基于 Qwen-0.5B 进行标准 LoRA 监督微调。
- Qwen_CoT_LoRA: 在 Qwen-0.5B 上，结合 CoT 数据进行 LoRA 监督微调，探索 CoT 对 SFT 阶段的潜在增益。
- DS_LoRA: 基于 DeepSeek-R1-Distill-Qwen-1.5B 进行标准 LoRA 监督微调。
- DS_RL_model: 在 DS_LoRA 模型的基础上，进一步采用强化学习进行优化，旨在提升生成的创意性和用户偏好对齐度。

在评价层面，为全面衡量模型性能，我们采用了人工评估与大语言模型（LLM）辅助评估相结合的方式。邀请人类评估员从创意性、连贯性、与关键词的相关性等多个维度对生成文本进行 1-10 分的评分。同时，利用 LLM 进行辅助打分与分析，力求获得客观、多维度的评价结果。

1.2 成果介绍

对四个代表性模型的评估揭示了模型规模、训练策略与创意文本生成效果之间的重要关系。

在基于 Qwen-0.5B 的小模型实验中，采用标准 LoRA 微调的 Qwen_LoRA 模型在生成短文本方面表现良好，能够产出可读性高、格式清晰的内容，平均评分达到 8 分（满分 10 分）。然而，在 SFT 阶段引入 CoT 数据的 Qwen_CoT_LoRA 模型，其生成文本在格式上存在明显不足，并且有时会包含冗余或不必要的推理步骤——即“思考过度”现象。我们进一步尝试在 Qwen_LoRA 和 Qwen_CoT_LoRA 基础上都应用 GRPO 算法进行强化学习，但未能观察到显著的性能提升。

我们推测，小参数量模型的表达能力和学习复杂策略 GRPO 所优化的策略的能力有限，这可能是导致 RL 效果不彰的主要原因。这一观察与学术界关于模型规模与能力涌现的普遍认识相符，即某些高级能力——如复杂的推理和创意生成可能需要达到一定的参数量阈值才能有效激发。

基于此，我们转向了参数量更大的 DeepSeek-R1-Distill-Qwen-1.5B 模型。我们改进我们的规则奖励函数，针对格式以及文本长度增大了奖励以及惩罚系数。接着，对该模型应用完整的训练范式——包括多阶段 SFT 和后续的改进 GRPO 强化学习，最终得到 DS_RL_model，取得了显著的进步。DS_RL_model 不仅能生成格式规范、创意性强的文本，其输出内容常常伴有清晰、详细的思考过程，类似 CoT 结构，显示出对任务要求和生成逻辑的更深层次理解。

这些结果表明，虽然基础 SFT 在小模型上能取得一定效果，但要充分发挥如 GRPO 等先进强化学习算法在创意生成任务上的潜力，充足的模型参数量是关键因素。结合了恰当模型规模、多阶段 SFT 预热以及 GRPO 优化的训练范式，能够有效提升模型生成创意文本的质量、结构性和可解释性。

2 方法范式

2.1 数据处理与构建

高质量的数据是驱动语言模型性能的关键因素，尤其是在需要模型掌握特定风格或能力的创意文本生成任务中。本研究采用了多阶段的数据处理与构建流程，旨在为后续模型微调（SFT）和强化学习（RL）提供定制化、高质量的数据支持。整个流程包括基础数据处理和深度数据增强两个主要部分。

2.1.1 基本清洗

本研究的原始数据来源于公开网络资源。我们通过定制化的网络爬虫及公开 API 接口，收集了约 10 万条以歌词为主的创意文本样本。这些原始数据包含了丰富的语言风格和创作模式，但也存在格式不一、噪声信息、非目标语言内容以及重复内容等问题。为确保数据质量并聚焦于中文创意文本生成，我们执行了以下基础预处理步骤：

数据清洗 (Cleaning)：移除了与文本内容无关的 HTML 标签、特殊字符、元数据——如来源注释、时间戳等)。对文本格式进行了初步统一，例如规范化标点符号、处理异常换行等，以减少对模型训练的干扰。

语言过滤 (Language Filtering)：为了确保模型专注于中文创意文本的生成，我们对数据进行了语言检测。利用基于字符集判断识别并移除了主要内容非中文的文本条目。此步骤保证了后续训练数据的语言一致性。

数据去重 (Deduplication)：采用了基于哈希或语义相似度的方法，识别并移除了完全相同或高度相似的文本条目。这有助于减少数据冗余，提高训练效率，并避免模型对特定重复样本产生过拟合。

经过上述基础预处理，我们得到了一个相对干净、多样化且以中文为主的创意文本语料库，为后续深度加工和模型训练奠定了坚实的基础。

其中一条数据如下：

有多少次我尝试写首诗留下你那婀娜多姿的样子/有时候娇横伤人有时候娇柔依人/我只能用最奢侈的玻璃为你筑起不惧风雨的天地/只要能够陪着你我都愿

意我就是你忠实的园丁/只要能够想着你我就欢喜/即使你宁可自由自在呼吸那一窗星星/名字背过又忘记符号充满了神秘/原在梦里忽而又中世纪我像圣堂下为你禁欲的僧侣

2.1.2 深度加工

为了更好地引导模型根据用户意图进行创作，并提升生成内容的结构性和逻辑性，我们对预处理后的中文创意文本语料库进行了深度增强。此阶段核心目标是生成带有关键词标签的数据以及包含思维链（Chain-of-Thought, CoT）的高质量样本。这两个环节均系统性地利用了先进的大语言模型 DeepSeek V3-0324 的 API 接口。

关键词提取 (Keyword Extraction):

为每条创意文本标注核心主题或元素，使模型能够学习关键词与文本内容之间的映射关系，从而实现基于关键词的条件生成。

方法：我们利用 DeepSeek V3-0324 API，通过多线程处理，为基础预处理阶段得到的全部约 10 万条中文创意文本样本逐一提取了关键词。针对每条文本，模型被指示输出能代表其核心内容的关键词集合，通常为 3-5 个。

此过程为整个语料库生成了对应的关键词标签。这批带有关键词标注的数据——约 10 万条，构成了后续进行第二阶段监督微调 Keyword-SFT 的关键数据集，并可直接用于推理阶段作为条件输入。

数据展示如下：

温柔，轮廓，洒脱：在我眼前竟是这般米罗/你的温柔你的轮廓你的洒脱/在我心底都是飘逸水墨/不知怎么样开口有太多的感动/因为太难以形容只能对着自己说/躲在梦里喃喃念出来/藏在手心迟迟打不开/密麻错落的情怀/你的言语你的欢声你的呼吸/播放一种巴洛克的旋律/你的步伐你的转身你的沉静/框进画里印象派的意境

Chain-of-Thought (CoT) 数据生成：

探索通过显式地展示创作思考过程来提升模型生成创意文本的逻辑性、结构化和创意深度。CoT 数据旨在引导模型学习“如何思考”来创作，而不仅仅是模仿最终文本。

样本选择：基于“少即是多” (Less is More) 的原则，特别是受到相关研究 LIMO: Less is More for Reasoning 的启发，即少量高质量的推理导向数据即能显著提升模型相关能力，我们并未对所有数据生成 CoT。相反，我们从 10 万条预处理后的语料中精心筛选出 1000 条被认为是最高质量、最具代表性的样本的关键词。针对这 1000 条精选样本，我们设计了特定的 Prompt，并调用 DeepSeek V3-0324 API 来生成具有 Chain-of-Thought 结构的创作过程或解释。这些 CoT 样本通常详细展示了从理解输入、构思主线、组织段落/歌词节、到具体遣词造句的逐步思考流程。

Prompt 示例：例如：我们设计的 Prompt 可能要求模型：“请分析以下歌词的核心主题和情感。然后，构思一个创作过程，包括如何开头，如何发展情感。最后，基于这个思考过程重写或解释这首歌词，并输出完整的思考步骤和最终文本。”

通过此方法，我们获得了 1000 条高质量的 CoT 数据样本。这些样本具有清晰的思考脉络和结构化的表达。尽管数量不多，但基于 LIMO 原则，我们预期这批独特的 CoT 数据能有效用于后续的特定 SFT 阶段如训练 Qwen_CoT_LoRA 模型，以及作为强化学习 (RL) 阶段的重要输入或高质量轨迹范例，用以鼓励模型生成更具深度和逻辑性的创意内容。

其中一个数据样例：

信，天涯，晚风<think>嗯，用户让我根据“信、天涯、晚风”这三个关键词写一首诗。首先，我得确定这三个元素如何有机地结合起来。信通常与思念、沟通有关；天涯暗示距离或遥远的地方；晚风则带来一种黄昏或夜晚的轻柔氛围。可能需要营造一种在黄昏时分，借着晚风传递书信到远方的意境。接下来，考虑诗的结构。中文诗常用五言或七言律诗，或者现代诗的形式。用户没

有指定，可能希望有古典韵味，所以五言或七言比较合适。但也可以尝试现代诗，更自由地表达。然后，思考如何将这三个词串联起来。比如，晚风吹动信纸，或者信的内容随风飘向天涯。或者，写信的人在晚风中思念远在天涯的人。需要营造一种静谧而略带忧伤的感觉。可能需要加入一些自然意象，比如雁群、暮色、星辰，来增强画面感。例如，用“雁群”象征传递信息，但这里已经有“信”了，可能不需要重复。或者用“折痕”来表现信纸的褶皱，暗示反复阅读或保存已久。还要注意押韵和节奏，使诗句流畅。比如，每句的结尾字押韵，或者采用对仗结构。例如，前两句描绘场景，中间两句表达情感，后两句深化主题。可能会想到这样的句子：“信纸的折痕里，住着迟到的春天”，这里将信与季节联系起来，表达等待的漫长。然后“暮色垂钓星辰”描绘黄昏的景象，同时引出天涯的概念。“晚风解开行囊”暗示风的到来，带来或带走某种情感，最后“字句在掌心结冰”表现无法传递的遗憾。检查是否有重复或冲突的意象，确保每个词都有其独特的作用。可能需要调整用词，使整体更连贯。例如，确保“天涯”与“暮色垂钓星辰”之间的联系是否自然，或者是否需要更直接的表达。最后，确定诗的情感基调，是忧伤、希望，还是混合？根据关键词，可能偏向于思念与距离带来的无奈，所以结尾的“结冰”可以强化这种无法传达的感觉。同时，加入“候鸟”和“指纹”来增加层次，表现时间的流逝和个人的痕迹。整体结构可能需要分段，每段集中一个意象，逐步展开主题。确保过渡自然，避免突兀。完成初稿后，再润色用词，使语言更凝练，意境更深远。

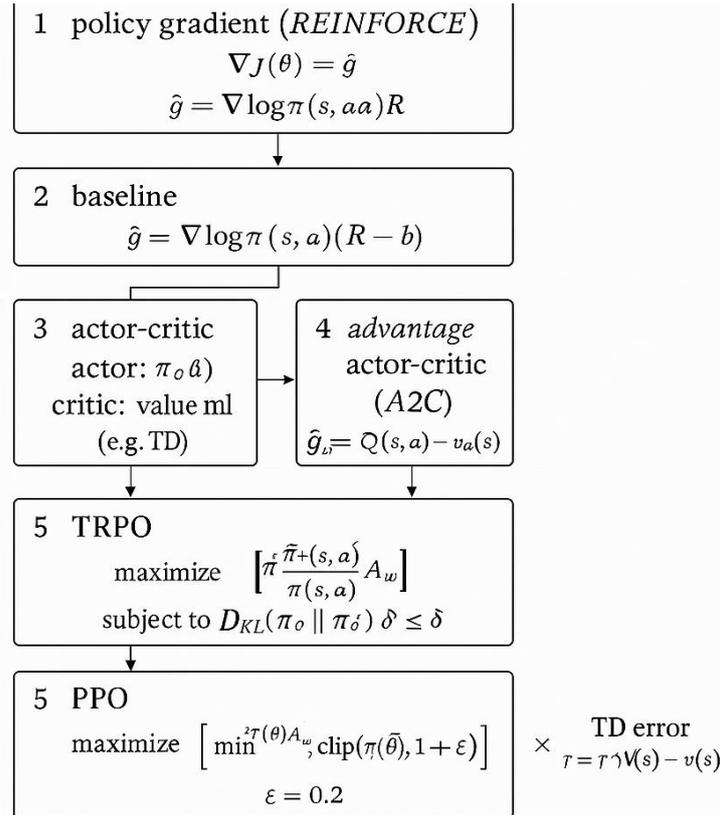
</think>: 《未寄出的褶皱》/信纸的折痕里/住着迟到的春天/暮色垂钓星辰时/天涯正被揉成碎屑/晚风解开行囊/抖落一地候鸟的指纹/那些横平竖直的笔画/在掌心结冰/成为地质年轮中/未曾启封的琥珀。

2.2 Policy Optimization

强化学习 (RL) 的目标是训练一个智能体 (Agent)，使其通过与环境交互来学习一个最优策略 (Policy)，以最大化累积奖励 (Cumulative Reward)。在大型语言模型 (LLM) 的微调中，RL，特别是策略优化 (Policy Optimization) 方法，已被证明能有效引导模型生成更符合人类偏好或特定目标的文本。本节

将回顾策略优化方法的演进，从基础的策略梯度方法到广泛应用的 PPO 算法，最终引出本研究重点关注的 Rule-Based GRPO 算法。

2.2.1 从策略梯度到 PPO



策略优化方法的核心思想是直接参数化策略 $\pi_\theta(a|s)$ ，并通过梯度下降。来更新参数 θ ，以最大化期望累积奖励。

$$\theta \leftarrow \theta + \alpha \times \nabla \theta \times \text{reward}$$

和经典传统的梯度下降法不同，参数更新中加入了奖励值。然而，基础的方法存在高方差的问题，即梯度的估计值在不同采样批次间波动很大，导致训练过程不稳定且收敛缓慢。这主要是因为奖励信号本身可能带有噪声，且单个轨迹的奖励可能与策略的长期表现关联不大。

为了降低方差，引入了基线 $b(s)$ 的概念。更新时不再使用原始奖励，而是使用奖励与基线的差值 $\text{Reward} - b(s)$ 。一个常用的基线是状态价值函数 $V(s)$ ，即在状态 s 下遵循当前策略所能获得的期望累积奖励。直观地说，如果获得的奖励高于预期（基线），则增加相应动作的概率；反之则减少。这有效降低了梯度估计的方差，提升了训练稳定性。

$$\theta \leftarrow \theta + \alpha \times \nabla \theta \times (\text{reward} - b(s))$$

为了克服基础策略梯度方法中高方差的问题，并提高学习效率，Actor-Critic, AC 架构被广泛采用。AC 架构明确地分离了智能体的两个核心功能

Actor: 通常指策略网络 (π_θ , 即 Policy Model), 负责根据当前状态 s 选择动作 a ; 例如, 在 LLM 中生成下一个词元或整个文本序列。其目标是学习一个能最大化累积奖励的策略。

Critic: 通常指价值网络, 负责评估 Actor: 所选择的动作或所处状态的好坏。它可以学习状态价值函数 $V(s)$ ——表示从状态 s 出发遵循当前策略的期望回报或状态-动作价值函数 $Q(s, a)$ ——表示在状态 s 执行动作 a 后遵循当前策略的期望回报。

Critic 提供的价值评估为 Actor 的策略更新提供了关键的指导信号。相比于直接使用可能延迟、稀疏或充满噪声的原始奖励信号, 基于价值评估的反馈通常更稳定、信息更密集。

优势函数 (Advantage Function) 在 AC 框架中扮演着核心角色。它旨在量化在特定状态 s 下, 采取某个动作 a 相较于该状态下平均动作表现的相对优势。其标准定义为:

$$A(s, a) = Q(s, a) - V(s)$$

使用优势函数 $A(s, a)$ 而非简单的奖励或 $Q(s, a)$ 来指导策略更新，具有显著优点：它天然地引入了一个状态相关的基线 $V(s)$ ，有效降低了梯度估计的方差，使得策略更新更稳定、更精确。直观地说，策略更新的目标是增加具有正优势值的动作被选中的概率，同时减少具有负优势值的动作被选中的概率。因此，策略参数 θ 的更新方向通常与优势加权的策略梯度相关：

$$\nabla_{\theta} J(\theta) \propto E [\nabla_{\theta} \log \pi_{\theta}(a|s) * A(s, a)]$$

(其中 $J(\theta)$ 是策略目标函数)

RLHF 训练范式

OpenAI 等机构提出的 RLHF (Reinforcement Learning from Human Feedback) 范式巧妙地将 AC 架构和优势函数的思想应用于大型语言模型的对齐任务。在典型的 RLHF 流程中：

数据收集与奖励建模：首先，让待优化的 LLM (Policy Model) 对一系列提示生成多个不同的回答。然后，由人类标注者对这些回答进行比较和排序，选出更符合期望——如更有帮助、更无害、更诚实的回答。利用这些大规模的人类偏好数据，训练一个独立的奖励模型 (Reward Model, RM)。这个 RM 的作用类似于 AC 架构中的评论家，但它学习的目标是预测一个标量分数，该分数反映了给定回答在多大程度上符合人类偏好，而不是传统的 $V(s)$ 或 $Q(s, a)$ 。

RL 微调：在此阶段，Policy Model (Actor) 生成回答，而预训练好的 Reward Model (Critic) 为这些回答打分，提供奖励信号。这个奖励信号随后被用来计算优势估计值。例如，优势可以被估计为 RM 对当前回答的打分减去一个基线（可能是 RM 对一个参考回答的打分，或者是该状态下 RM 得分的期望值，通常也由一个价值网络学习得到）。最后，利用计算出的优势估计值和 PPO 等策略优化算法来更新 Policy Model 的参数 θ ，使其倾向于生成能获得更高 RM 分数的回答。

通过这种方式，RLHF 将 AC 架构和优势函数的原理与人类的隐式知识，体现在偏好数据中结合，有效地将 LLM 的行为引导至更符合人类期望的方向。

TRPO 和 PPO

尽管基于优势函数的 Actor-Critic 方法显著提升了策略优化的稳定性，但如何确定策略更新的步长仍然是一个核心挑战。过大的更新步长可能导致策略性能急剧下降，破坏学习过程；而过小的步长则可能导致收敛速度过慢。为了解决这一问题，研究者们提出了更先进的算法，旨在更直接地控制策略更新的幅度，确保学习过程的稳定性和单调性。其中，TRPO 和 PPO 是两个里程碑式的工作。

信任区域策略优化 (Trust Region Policy Optimization - TRPO)

TRPO 的核心思想是限制每次策略更新的幅度，确保新的策略 π_{new} 与旧的策略 π_{old} 保持在一定的“距离”之内，这个距离定义了一个“信任区域”。其理论基础是，在信任区域内进行的策略更新更有可能带来单调的性能提升。TRPO 通过引入 KL 散度 (Kullback - Leibler divergence) 约束来实现这一点。KL 散度是衡量两个概率分布之间差异的一种度量。TRPO 的目标是最大化一个替代目标函数，该函数估计了新策略相对于旧策略的期望优势，但同时强制要求新旧策略之间的平均 KL 散度不超过一个预设的小阈值 δ ：

$$\text{maximize } \theta E_s [E_{\{a \sim \pi_{\text{old}}\}} [r(\theta) * A_{\text{old}}(s, a)]]$$

$$\text{subject to } E_s [KL(\pi_{\text{old}}(\cdot | s), \pi_{\theta}(\cdot | s))] \leq \delta$$

其中： $A_{\text{old}}(s, a)$ 是基于旧策略 π_{old} 计算的优势函数估计值。

$r(\theta) = \pi_{\theta}(a|s) / \pi_{\text{old}}(a|s)$ 是重要性采样 (Importance Sampling) 因子或比率 (ratio)。它允许我们使用从旧策略 π_{old} 采样的数据来评估新策略 π_{θ} 的性能。这个比率反映了在新旧策略下采取相同动作 a 的概率变化。

TRPO 通过求解上述约束优化问题来找到更新后的策略参数。虽然 TRPO 在理论上具有良好的单调改进保证，并且在实践中表现稳定，但其实现通常需要复

杂的二阶优化方法来处理 KL 散度约束，计算成本较高且不易集成到现代深度学习框架中。

近端策略优化 (Proximal Policy Optimization - PPO)

为了克服 TRPO 的复杂性，PPO 被提出，旨在通过更简单的一阶优化方法实现类似 TRPO 的稳定性能。PPO 已成为 RLHF 等领域中最受欢迎和广泛应用的策略优化算法之一。PPO 的核心创新在于引入了一种裁剪的替代目标函数 (Clipped Surrogate Objective)，它通过直接修改目标函数本身来限制策略更新的幅度，而非使用显式的 KL 约束。其目标函数通常形式如下：

$$L^{\text{CLIP}}(\theta) = E_t [\min(r_{t(\theta)} * A_t, \text{clip}(r_{t(\theta)}, 1 - \epsilon, 1 + \epsilon) * A_t)]$$

$r_{t(\theta)} = \pi_{\theta}(a_t|s_t) / \pi_{\{\theta_{old}\}}(a_t|s_t)$ 是在时间步 t 的重要性采样比率。

A_t 是在时间步 t 的优势函数估计值。

ϵ 是一个小的超参数，定义了裁剪区间 $[1 - \epsilon, 1 + \epsilon]$ 。

$\text{clip}(r_{t(\theta)}, 1 - \epsilon, 1 + \epsilon)$ 函数将比率 $r_{t(\theta)}$ 限制在 $[1 - \epsilon, 1 + \epsilon]$ 区间内。

这个目标函数的直观理解是：

当优势 A_t 为正时（即动作比平均预期要好），目标函数倾向于增加 $r_{t(\theta)}$ ，但增长被 $(1 + \epsilon) * A_t$ 限制。这防止了策略因为一个好的动作而更新得过于激进。当优势 A_t 为负时（即动作比平均预期要差），目标函数倾向于减小 $r_{t(\theta)}$ ，但减小被 $(1 - \epsilon) * A_t$ 限制。这防止了策略因为一个坏的动作而更新得过于保守。

通过取 min 操作，PPO 有效地创建了一个悲观下界 (Pessimistic Bound)，惩罚那些使得比率 $r_{t(\theta)}$ 远离 1 的更新。这种方式在实践中被证明非常有效，它既能保持策略更新的稳定性，又易于使用标准的梯度下降优化器实现，并且具有良好的样本效率。总结来说，从 AC 到 TRPO 再到 PPO，策略优化的发展体

现了在保证学习有效性的同时，不断追求算法稳定性、实现简易性和计算效率的趋势。

2.2.2 Rule Based GRPO

尽管 PPO 算法及其在 RLHF 等框架中的应用已取得显著成功，推动了大型语言模型对齐技术的发展，但在实践中，其标准范式仍面临若干挑战：

奖励模型 (Reward Model) 的构建与维护成本：在典型的 RLHF 流程中，PPO 依赖于一个独立的奖励模型来评估生成内容的质量或符合人类偏好的程度。训练一个高质量、泛化能力强的 RM 通常需要大量的人工标注数据（，这不仅耗费巨大的财力与人力资源，而且标注过程本身可能引入主观偏差。

计算资源要求：在强化学习微调阶段，通常需要将 Policy Model 和 Reward Model 同时加载到计算设备的显存中。对于参数量巨大的现代 LLM，这会产生显著的计算开销和高昂的显存需求，给资源相对有限的中小型研究团队或企业带来了较高的准入门槛。

潜在的奖励作弊 (Reward Hacking) 风险：这是一个更为根本性的问题。由于 RM 是基于有限数据学习到的对人类偏好的近似模拟，LLM 在优化过程中可能发现并利用 RM 评分机制中的“漏洞”或捷径，生成表面上能获得高分、但实际上并不符合预期目标的输出。这种模型“投机取巧”以最大化代理奖励信号而非真正实现预期目标的行为，即为奖励作弊，是基于学习奖励模型的 RL 方法普遍面临的固有挑战。

针对上述挑战，GRPO (Group Reward Policy Optimization) 算法并提出。GRPO 提供了一种不同的范式，旨在简化奖励生成过程，降低对复杂奖励模型的依赖，并可能缓解奖励作弊问题。其核心思路是采用基于规则的启发式方法——Rule-Based Heuristics) 直接从模型的生成内容 (中计算奖励信号。这种方法有望通过更直接、可解释的奖励机制，规避训练和部署复杂 RM 带来的高成本，并减少模型利用奖励模型漏洞的机会，

基于规则的 GRPO 并不试图从大规模比较数据中隐式学习人类偏好，而是倡导通过一系列明确的、预定义的规则或启发式方法来界定生成文本所期望的行

为和属性。这些规则直接作为评估策略模型 (Actor) 输出的机制。规则的设计可以高度灵活并依赖于具体任务，可能包含的标准有：

语法正确性与流畅度：评估语言的语法完整性和自然程度。

语义相关性：检查输出是否与输入提示或给定关键词保持一致。

格式遵循度：核实输出是否符合特定的结构要求（诗歌的韵律结构）。

内容约束：确保包含期望的元素或排除有害/不期望的内容（安全对齐）。

结构与逻辑（尤其结合 CoT）：评估生成的连贯性、一致性和逻辑流程，特别是在利用思维链 (Chain-of-Thought) 输出时。规则可以评估特定推理步骤的质量或存在与否。其核心机制涉及将这些定义的规则应用于每个生成的样本，以计算出一个标量的奖励信号。与查询大型奖励模型相比，这种计算通常是确定性的，并且计算开销较小。

GRPO 中的“组” (Group) 概念通常涉及将多个输出样本组成一个小批次 (group of completions)，并在批次内对原始的基于规则的奖励分数进行标准化。最常见的做法是对每个样本

$$\hat{r}_i = \frac{r_i - \mu_{group}}{\sigma_{group}}$$

其中 r_i 是样本的原始规则奖励， μ_{group} ， σ_{group} 分别是该组内所有样本奖励的均值和标准差。

这一归一化步骤有两个关键目的：

增强学习信号的稳定性：通过消除奖励信号的尺度波动，使得策略优化更稳定；引入相对排名优势 (Group-wise Advantage)：这种机制类似于优势函数 (Advantage Function)，但不依赖价值函数的估计，而是通过组内比较来动态衡量“好于平均”的程度。可以理解为，这种组间优势的思想使得模型更新不仅仅基于绝对奖励得分，而是强化那些在一组候选输出中相对更优的生成，从而提高优化效率，减少对复杂价值函数或回报折算的依赖。

这个基于规则的归一化奖励信号 \hat{r}_i 可进一步用于构造策略更新所需的优势估计值 A_i^{rule} 从而策略优化目标函数类似于 PPO 中裁剪的目标，可表达为：

$$L_{GRPO}(\theta) = E_{(s,a) \sim \pi_\theta} [\min(r_t(\theta) * A_i^{rule}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) * A_i^{rule})]$$

通过采用这种基于规则的框架，GRPO 为先前确定的挑战提供了潜在的解决方案：

降低成本与复杂性：它绕过了大规模人类偏好数据收集以及训练和部署独立奖励模型所需的大量资源。

提高透明度与可控性：奖励机制是明确且可解释的。研究人员和开发人员可以直接理解为何某个输出获得高分或低分，并可以通过修改规则来微调模型行为。

缓解奖励作弊：虽然并非完全免疫（模型可能找到满足规则的表面方式），但明确的规则相较于从数据中学到的不透明近似模型，可能不易受到微妙的利用。奖励的标准是清晰固定的。

提供高效且稳健的优化信号：通过组间标准化与优势机制，能够有效利用局部对比信息，使模型更快收敛于“相对更好”的策略。

然而，基于规则的 GRPO 的有效性在很大程度上取决于手工制定的规则的质量和全面性。设计一套能够准确捕捉像创意文本生成这类复杂任务所有期望细微差别的规则可能具有挑战性，并且可能需要大量的领域专业知识和迭代优化。

在本研究的背景下，我们正是利用了基于规则的 GRPO 的原理，对其规则设计进行了调整和可能的扩展，以专门针对提升小型语言模型在创意文本生成方面的能力，特别是利用我们生成的 CoT 数据所提供的结构信息。这种方法使我们能够将期望的创意属性和逻辑连贯性直接编码到奖励信号中，与传统的 RLHF 方法相比，可能提供一条资源效率更高、更易于控制的微调路径。

2.3 挫折、回退、更新

尽管理论层面以及 DeepSeek 团队的实证研究表明，GRPO (Group Reward Policy Optimization) 算法具备极高的性价比——以显著更低的训练资源开销，实现与传统 RLHF 方法相当甚至更优的微调效果，体现出高度的第一性原理驱动特性。但在我们针对小规模语言模型（例如 500M 量级）以及生成创意文本这类输出空间高度灵活、评估标准主观性较强的任务中，GRPO 的直接应用仍面临显著困难。

初始尝试：Qwen-0.5B 模型与逐步训练策略

我们选取了 Qwen-0.5B 作为初始的基础模型 (backbone)，并遵循以下两个阶段进行训练：

第一阶段：纯监督微调

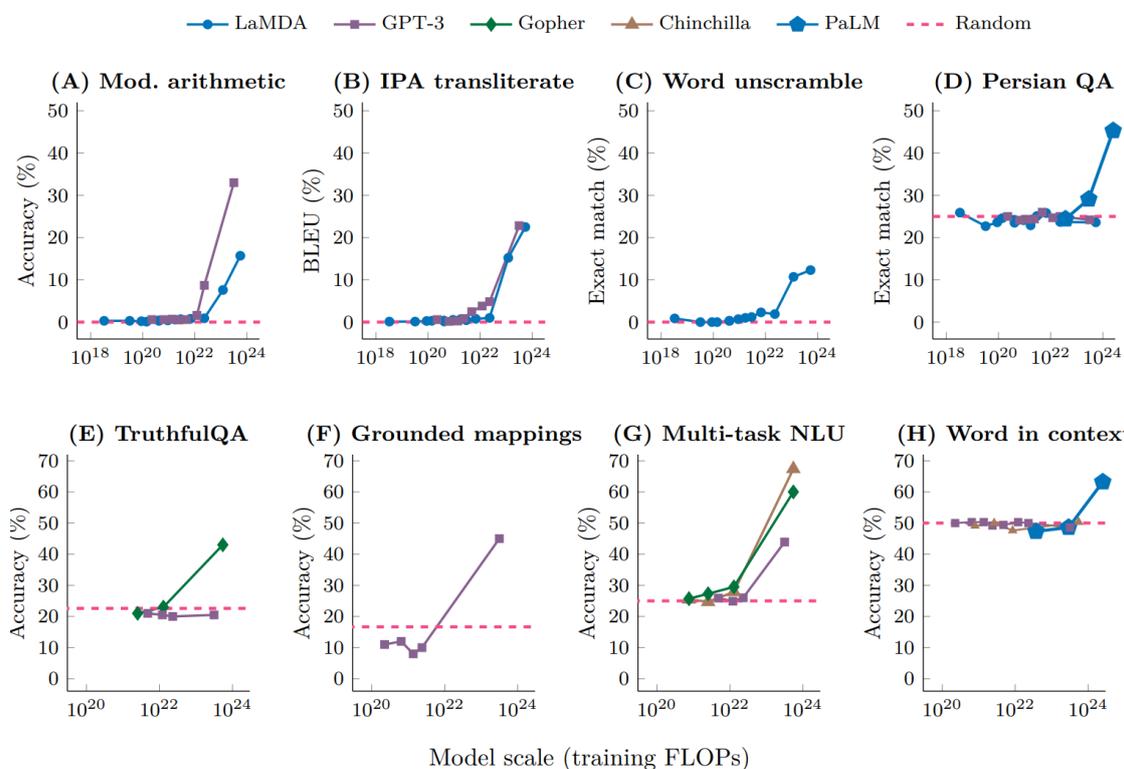
在该阶段，我们未引入任何关键词提示或思维链 (Chain-of-Thought, CoT) 结构，而是直接使用原始的人类创意写作数据对模型进行监督微调，目标是使模型学习基本的语言风格、文本格式以及创作语境下的语言节奏与表达方式。

微调后模型在输出风格仿写方面表现合理，但缺乏对内容的控制能力，即生成内容未能充分围绕输入主题展开，缺乏逻辑性与指向性。

第二阶段：带关键词数据上的强化学习

随后，我们在包含关键词与目标文本对的数据上，对上述微调后的模型执行基于 GRPO 的强化学习训练。然而该阶段结果非常不理想：模型在训练后频繁生成高度重复的内容，缺乏基本的语义拓展能力，语言表现极度僵化。我们初步推测这一问题可能源于奖励函数设计存在偏差，或训练数据本身无法支撑强化学习阶段所需的探索空间。为了进一步探查问题本质，我们对训练流程进行了改进

在模型参数量超过特定阈值后才会表现出来。因此，GRPO 等强化学习范式若要在创意写作等高复杂度生成任务中奏效，模型基础能力的容量门槛至关重要。



策略转移：采用更大模型并融合训练流程

基于上述分析，我们将基础模型更换为参数量更大的 DeepSeek-R1-Distill-Qwen-1.5B，并对训练流程进行了融合式优化。

直接使用思维链数据进行监督微调

鉴于上一阶段的分步训练策略在初始阶段缺乏对写作逻辑的建模，本轮我们跳过单独的关键词训练阶段，直接在结构化的 CoT 数据上对模型进行监督微调。该策略使模型在初期即开始接触多步推理结构与写作逻辑链条，从而在后续强化学习阶段减少“冷启动”问题，加速收敛。

强化学习阶段：规则驱动的 GRPO 优化

随后，我们利用设计良好的基于规则的奖励函数——涵盖语言长度、逻辑、关键词对齐、结构完整性等维度——执行了 GRPO 强化学习训练。训练在

1 万步内完成，期间奖励收敛良好，模型行为逐步稳定，最终得到强化后的模型 DS_RL_model。该训练流程不仅显著节省了策略优化阶段的探索代价，也有效缓解了早期强化学习训练中模型输出无序、收敛困难等问题，最终产出的模型在多项主观和客观评估指标上均表现出更强的控制能力、逻辑一致性与创作灵活性。



3 讨论

3.1 研究局限

尽管本研究在小模型场景下验证了基于规则的 GRPO 范式的可行性，并通过逐步构造的训练流程取得了较为显著的提升效果，但我们仍需正视当前工作的若干局限性：

首先，在模型训练过程中，我们未能充分保留和可视化关键的中间指标，尤其是在策略优化阶段——GRPO 的 loss 曲线、KL 散度指标以及奖励分布等方面，缺乏系统的收敛性与稳定性展示。这一缺口限制了我们对训练动态的深入分析，也影响了后续模型调优的可解释性。

其次，在创意文本生成任务的效果评估方面，我们尚未构建起一套体系化、自动化的评估机制。尽管采用了 DeepSeek 官方 API 进行辅助评分，并结合了少量人工评审结果，但整体评价过程缺乏一致性与覆盖面，主观性较强。这种评估方式在一定程度上削弱了对模型性能的全面判断，也不利于在更大规模或跨任务设置中进行推广比较。

此外，由于计算资源和时间的限制，我们在规则设计上采用了相对通用的启发式方法，尚未针对创意文本生成任务中某些更为微妙的语言特性——情绪色彩、比喻创新、节奏韵——进行更细粒度的设计与调整。规则质量的提升空间仍然显著。

3.2 展望改进

针对上述不足，我们提出如下的未来改进方向：

(1) 构建更系统的训练可视化与分析机制。

后续工作将引入更完整的训练日志追踪与可视化手段，覆盖 reward 变化趋势、policy loss、value loss、KL divergence、梯度范数等指标，并结合 tensorboard / wandb 等工具，实现训练全程的细粒度监控，从而更有效地评估模型训练是否收敛、是否出现过拟合或震荡等问题。

(2) 发展更可靠的评价体系。

我们计划结合自动化指标与专家标注，构建适用于创意文本生成的混合型评价框架。一方面，可探索引入如 MAUVE、BLEURT、Fréchet Distance for Text 等用于衡量生成质量与多样性的度量；另一方面，也可以借鉴大语言模型对生成文本进行结构性点评的新方法；如 GPT-4 Judge、LLM-as-a-Judge Paradigm，建立规则一致、偏差较小的模型辅助评分机制，以替代纯人工评审或 API 黑盒调用。

(3) 扩大模型与任务规模，验证方法适应性。

本研究集中在小型语言模型及单一文本创意任务的验证，下一阶段我们将引入参数量更大、能力更强的基础模型——DeepSeek-Coder、Mixtral 系列等

——并扩展到多任务或跨模态生成任务上，以系统评估 GRPO 在更复杂场景下的稳定性、可扩展性与收益边界。

通过以上改进方向的推进，我们希望能够进一步释放 GRPO 范式在低资源、高可控场景中的潜力，同时也为提升小模型的能力边界提供更具普适性的方法探索。

4 结论

正如强化学习之父 Richard S. Sutton 教授在其博文《The Bitter Lesson》中所强调的那样，历史经验不断印证：在构建智能系统的过程中，人类设计的启发式规则和精细结构，最终往往难以与“以算力为基础的通用方法”抗衡。以暴力计算、海量数据以及通用优化算法为核心的范式，尽管看似“简单粗暴”，却反复在多个技术浪潮中取得了突破性进展。这种源于第一性原理的理念也正是我们本研究所依赖的逻辑基础。

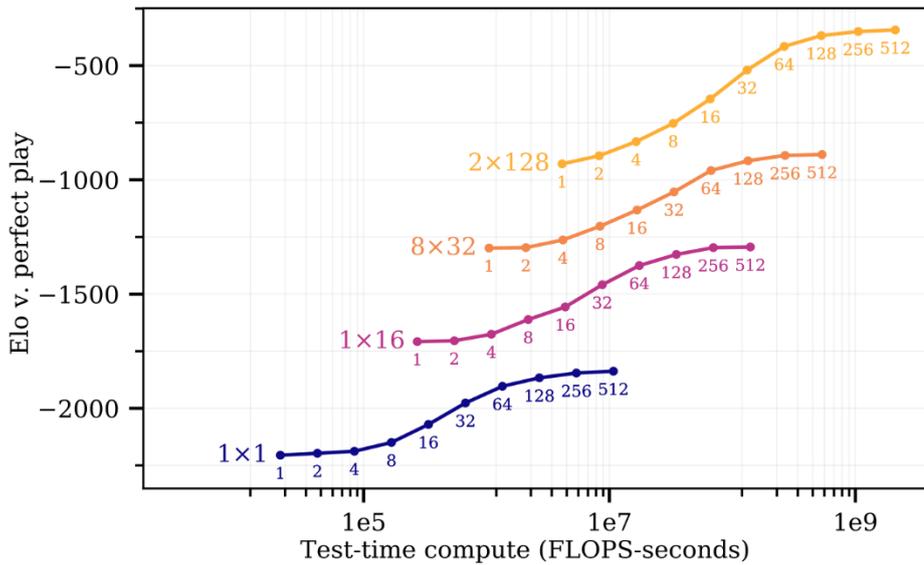
从早期基于显式值函数和策略迭代的强化学习方法，到近年来以策略梯度为核心、强调模型自主优化和“自由思考”的方法；如 PPO 及其衍生，我们可以观察到一种逐步简化、趋向通用的演化趋势。这种趋势不仅体现为算法设计上的简化，更体现为对智能本质的重新理解：智能或许并不依赖复杂手工规则，而在于环境中自主交互中形成的表达与推理能力。

本研究以一种轻量、小规模、低资源的方式对基于规则的 GRPO 范式进行了实证探索。在缺乏强大奖励模型与大规模人类偏好数据的前提下，我们尝试借助一系列透明可控的规则，配合相对简单的策略优化结构，引导小语言模型获得具备逻辑性和创造力的文本生成能力。这种结合了“人类先验规则”与“策略自我优化”的方式，也许正是在小模型受限条件下实现智能提升的一种有效路径。

更进一步地，我们认为这一研究验证了一个值得深思的假设：在通往通用人工智能（AGI）的道路上，简单的结构、强大的算力以及让模型自由探索与表达的机制——即一种从经验驱动转向能力驱动的模式，可能比人类设计更加复

杂的启发式系统走得更远。如何在这种“结构简化 + 自我组织”的路径上持续前行，是我们接下来研究工作的核心关注点之一。

当模型能在极简规则中自洽地优化表达与逻辑，我们或许正在接近人工智能真正的本质。



5 参考文献

- [1] J. Wei et al., “Emergent abilities of large language models,” Trans. Mach. Learn. Res., Aug. 2022.
- [2] DeepSeek-AI, “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,” arXiv preprint arXiv:2501.12948, Jan. 25, 2025.
- [3] N. Muennighoff et al., “s1: Simple test-time scaling,” arXiv preprint arXiv:2501.19393, Jan. 29, 2025.
- [4] Y. Wang et al., “Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs,” arXiv preprint arXiv:2501.18585, Jan. 28, 2025.
- [5] DeepSeek-AI, “DeepSeek-V3 Technical Report,” DeepSeek-AI, 2024. [Technical Report].
- [6] J. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv preprint arXiv:2201.11903, 2022.
- [7] A. Lewkowycz et al., “Solving Quantitative Reasoning Problems with Language Models,” arXiv preprint arXiv:2206.14858, 2022.
- [8] H. Lightman et al., “Let’s Verify Step by Step,” arXiv preprint arXiv:2305.20050, May 2023.
- [9] S. Yao et al., “Tree of Thoughts: Deliberate Problem Solving with Large Language Models,” arXiv preprint arXiv:2305.10601, 2023.
- [10] Z. Yuan et al., “Scaling Relationship on Learning Mathematical Reasoning with Large Language Models,” arXiv preprint arXiv:2308.01825, Aug. 2023.
- [11] X. Wang et al., “Self-Consistency Improves Chain of Thought Reasoning in Language Models,” arXiv preprint arXiv:2203.11171, 2023.