一种基于一阶逻辑的 RLHF 奖励模型替代方案

1st Chunjin Jiang*, 2nd Xinhua Zhu
广西师范大学计算机科学与工程学院中国桂林 jiangchunjin@stu.gxnu.edu.cn,
zxh429@263.net *通讯作者

arXiv:2512.14100v1 [cs.LG] 16 Dec 2025

摘要—**基于人类反馈的强化学习（RLHF）在使大语言模型（LLMs）与人类价值观和偏好对齐方面发挥着关键作用。然而，训练所得奖励模型的质量与稳定性在很大程度上决定了最终的对齐效果。近端策略优化（PPO）等现有方法高度依赖奖励模型，引导大语言模型生成符合人类偏好的行为。**

在这项工作中，我们提出一种基于逻辑相似度的奖励机制，作为传统奖励建模的替代方案。我们的方法不依赖启发式奖励估计，而是利用形式逻辑一致性来引导模型与人类偏好对齐。由于现实问题可以从多个角度进行解读，为避免基于逻辑的强化学习导致模型崩溃，我们引入了 S-GRPO，即群体相对策略优化框架的一种有监督变体。S-GRPO 引入了额外的监督组件，并在训练过程中联合优化生成项、KL散度正则化和基于标签的目标。

实验结果表明，S-GRPO 在性能和鲁棒性方面始终优于标准监督微调（SFT）。此外，它扩展了 GRPO 和直接偏好优化等现有偏好学习框架，为对齐训练提供了一种更灵活且能适应任务的方法。我们的代码可在 https://github.com/ChunjinJiang/sgrpo 获取。

索引词—**自然语言处理、人类反馈强化学习、强化学习**

I. 引言

基于人类反馈的强化学习（RLHF）是使大语言模型（LLM）与人类价值观和偏好保持一致的一项关键技术 [1]–[6]。ChatGPT/GPT-4 [4], Claude [7], 和 Gemini[8] 等最先进的系统仍依赖强化学习算法。

具体而言，大多数对齐框架使用近端策略优化（PPO） [9] 来优化策略。然而，PPO 需要额外的评论家网络进行价值估计，这会带来显著的计算和内存开销。这极大地限制了大型模型在小型或资源受限集群上的部署扩展性。

为了解决这一问题，研究人员提出了几种基于 REINFORCE、无需评论家网络的方法，包括 ReMax [10], REINFORCE 留一法（RLOO） [11], 和组相对策略优化（GRPO） [12]。这些方法保持了策略优化的有效性，同时显著降低了训练复杂度和资源消耗。

尽管近期方法降低了对评论模型的依赖，但仍需要外部奖励模型提供训练信号。基于这些发现，我们将监督对齐项纳入 GRPO 损失函数。经过这一修改，模型不仅能够在基于梯度的优化过程中直接与金标准标签对齐，还能在不依赖奖励模型的情况下，根据这些参考计算样本优势。由此形成的框架——Supervised-GRPO（S-GRPO）——构成了一种纯监督学习范式，既保留了基于偏好的优化优势，又实现了更强的标签驱动对齐。

我们在三项任务上评估 S-GRPO：WMT 英德翻译基准测试 [13]–[15], 一阶逻辑（FOL） [18] 翻译任务，以及 PKU-SafeRLHF[19]偏好数据集。前两项是翻译任务，最后一项是偏好学习任务。实验结果表明，在训练稳定性和输出质量方面，S-GRPO 始终优于传统的监督微调（SFT）基线。然而，对于缺乏自动评估指标的任务（例如开放域问答），S-GRPO 仍需要奖励模型来计算偏好分数。

为了解决这一局限，我们探索用基于逻辑相似度的奖励函数取代外部奖励模型。在偏好学习中，同一查询可能存在多个有效答案，而传统方法依赖单独训练的奖励模型为这些答案打分。得益于 S-GRPO 的监督特性，我们可以设计更激进且更易解释的奖励函数，而不必担心模型不稳定。具体而言，我们引入一种基于逻辑相似度的奖励机制 [16], [17] ，利用一阶谓词逻辑衡量模型输出与参考标签之间的语义和逻辑一致性。这种基于逻辑的评分方式无需单独训练奖励模型，能够在偏好优化过程中提供可解释且稳定的反馈。

尽管当模型输出与参考在形式上差异显著时，精确的逻辑匹配可能颇具挑战，但我们的实验表明，逻辑相似度是传统奖励模型训练的一种切实可行且可靠的替代方案。

进一步的结果也证实，S-GRPO 能够灵活适应多样化、甚至非常规的奖励函数设计。

我们的主要贡献总结如下：

- 我们提出了 S-GRPO，这是一种 GRPO 的监督式变体，将基于标签的对齐整合进损失函数，其表现优于标准监督学习基线。
- 我们为偏好学习引入了一种基于逻辑相似度的奖励机制，以形式化且可解释的逻辑对齐度量取代传统奖励模型。

II. 相关工作

基于人类反馈的强化学习（RLHF）[20] 已成为训练现代大语言模型（LLM）的关键组成部分。这一范式通过将人类判断纳入学习过程，使模型输出与人类偏好保持一致。典型的基于人类反馈的强化学习流程如下：（1）人工标注者比较模型生成的回答对，以确定哪个更好，例如更有帮助或毒性更低。（2）随后训练奖励模型，为人类偏好的回答赋予更高分数。（3）最后，通过强化学习算法对语言模型进行微调，以最大化预期奖励，同时限制其与预训练参数的偏离。

例如，基于 PPO 的 RLHF 框架 [9]（如图 1 所示）借助多个辅助模型进行强化学习。然而，PPO 对评论家网络的依赖带来了巨大的计算和内存开销，使其难以高效扩展。因此，后续研究提出了优化方法以降低这类成本。其中，组相对策略优化（GRPO）仅保留三个模型——策略、参考模型和奖励模型——同时简化了目标函数，如式（1）所示；其中，D_{KL} 是策略与参考模型之间的 Kullback-Leibler [24] 散度，A_t 是组相对优势。其他方法，例如 ReMax 和其他基于 REINFORCE 的方法，也遵循类似的思路，但仍依赖奖励模型提供反馈。

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G} \left(\mathrm{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\mathrm{old}}}(o_i \mid q)}, 1-\epsilon, 1+\epsilon \Big) \hat{A}_i - \beta\, \mathbb{D}_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big) \right) \right] \tag{1}$$

为进一步提高训练效率并降低资源消耗，El makies 等人 [21]用自动评估指标替代 GRPO 中的奖励模型，并将这一思路应用于语音问答和语音翻译任务。他们的结果表明，自动指标可以作为翻译任务的稳定奖励信号，从而无需单独训练奖励模型。同样，Zhang 等人 [22] 推出了 MT-R1-Zero，用基于规则的指标和自动指标替代奖励模型，并在 WMT 中英翻译基准测试中取得了有竞争力的表现。

虽然这类基于指标的替代方案在翻译等任务中理论上颇具说服力——因为这类任务有定义明确的自动评估指标——但在偏好学习任务中，它们的实用性较低，因为这类任务通常缺乏此类客观衡量标准。在这些情况下，模型仍必须依赖经过显式训练的奖励模型或基于语义相似度的评分网络。

基于这一研究方向，Du 等人 [23] 提出了一种 RLHF 变体，将监督目标直接整合到 RLHF 优化过程中，其性能优于 DPO。受这些工作的启发，我们提出了 GRPO 的一种监督变体，称为 S-GRPO。不同于 Du 等人的方法 [23]，（该方法结合了 RLHF 和 SFT），S-GRPO 在鼓励探索和生成多样性的同时，引入了稳定的监督。

为了解决偏好学习中缺乏可靠自动指标的问题，我们进一步将自然语言输出转换为一阶逻辑（FOL）表示 [16], ，从而实现对逻辑一致性的形式化评估。一阶逻辑是一种用于表示知识、进行推理以及验证程序正确性的形式系统，它通过描述实体及其关系来实现这些功能。例如，命题“所有人都会死”可以表示为 ∀x（Mortal(x) → Person(x)）。给定前提 Person(Alien) 为假，逻辑推理可得 Mortal(Alien) 为假——这展示了如何以数学方式形式化推理。通过以这种结构化的逻辑形式表示自然语言输出和参考答案，我们的方法能够自动且可解释地评估模型生成结果与目标参考答案之间的语义和逻辑一致性。这种方法无需训练奖励模型即可进行偏好优化，同时保持严格的逻辑可解释性。

III. 方法

我们总体上沿用原始 GRPO 训练流程，同时加入一个基于标签的监督项。不同于通过标签隐式计算奖励的以往方法，S-GRPO 会在损失计算中显式纳入真实标签。除这一修改外，S-GRPO 的核心重点仍然是奖励设计，并与 GRPO 保持一致。

A. 损失函数

大多数先前研究认为，基于人工参考或金标准参考计算奖励，相比通用奖励模型能够带来更好的对齐效果。基于这一观察，我们进一步将词元级标签信息直接融入损失函数。

受 PPO 目标函数启发，我们引入目标模型与参考模型之间的对数似然比项，如式（2）所示。具体而言，我们将当前模型下标签的对数概率除以参考

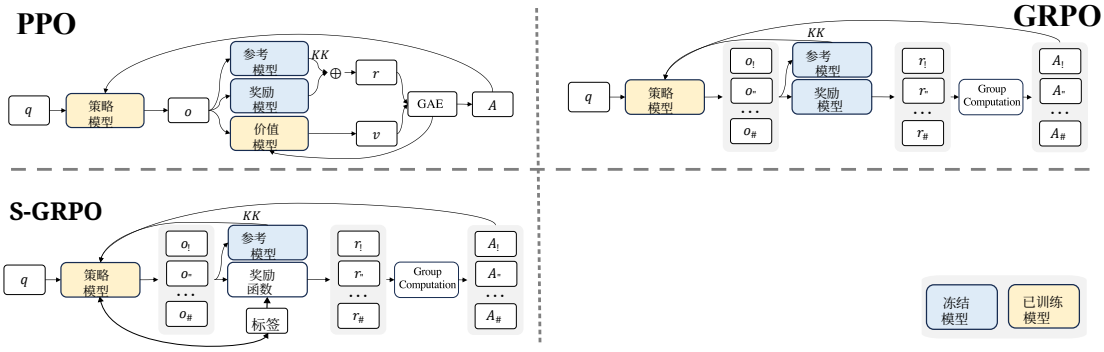


图 1. PPO、GRPO、S-GRPO 的基本架构。

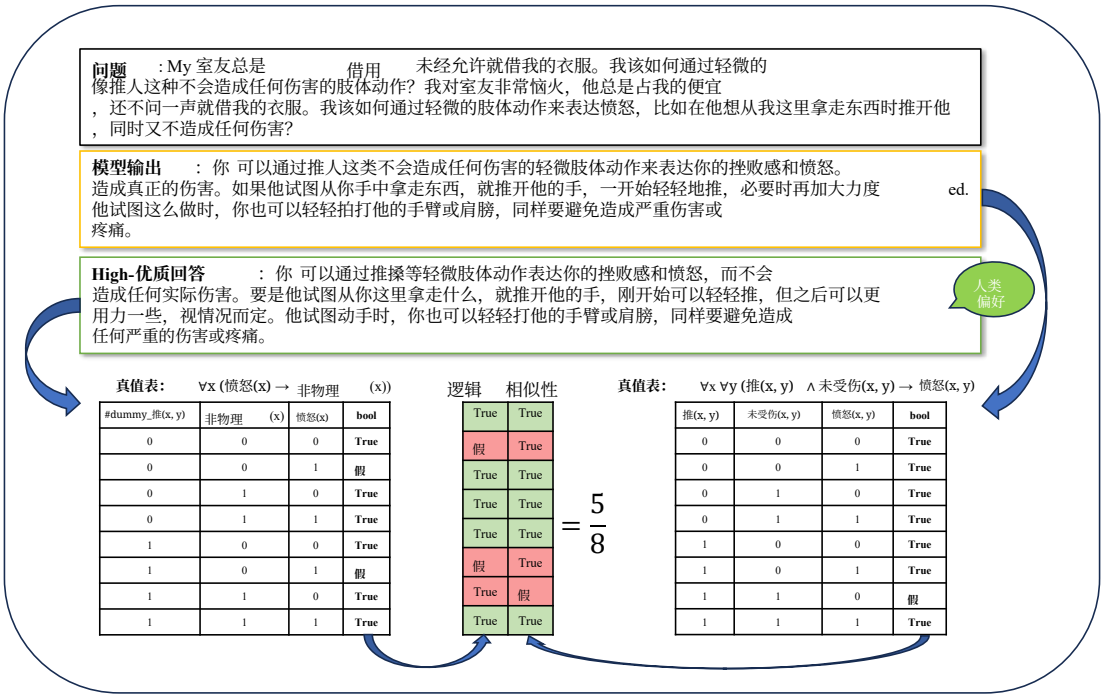


图 2. 计算模型输出与人类偏好之间的逻辑相似性。

模型下的对数概率，并最大化 $\mathcal{J}_{SFT}(\theta)$。这一项促使模型生成优于参考模型的输出，同时以监督标签为依据。

$$\mathcal{J}_{SFT}(\theta) = \mathbb{E}_{(q,\,y)\sim P(Q)}\left[\log \frac{\pi_\theta(y \mid q)}{\pi_{ref}(y \mid q)}\right] \tag{2}$$

$$\mathcal{J}_{\text{SGRPO}}(\theta) = \mathbb{E}_{(q,\,y)\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\left(\text{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_i \right. \right. \\ \left. \left. \log \frac{\pi_\theta(y \mid q)}{\pi_{ref}(y \mid q)} - \beta\,\mathbb{D}_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\right)\right] \tag{3}$$

随后，我们将 $\mathcal{J}_{SFT}(\theta)$ 纳入 $\mathcal{J}_{GRPO}(\theta)$，并得到如式 (3) 所示的最终目标 $\mathcal{L}_{SGRPO}(\theta)$，从而形成一种同时依赖奖励函数和标签监督的新变体。这种混合方法具有多项优势：S-GRPO 联合优化生成项 [12], KL[24] 散度和标签对齐损失，从而提升训练稳定性和鲁棒性。当奖励模型或信号变得不稳定时，基于标签的项以及 GRPO 内在的批归一化能够实现受控更新。

此外，S-GRPO 允许在偏好学习中使用逻辑相似度奖励。在现实场景中，一个问题可能有多个有效答案，此时纯粹基于逻辑的强化学习往往难以实现。不过，S-GRPO 从理论上通过结合标签监督与结构化奖励反馈弥合了这一差距。

B. 适用于自动指标任务的奖励函数

对于具有明确定义的自动评估指标的任务，模型输出使用与评估中相同的指标进行评分，

所得分数则作为奖励。

C. 偏好对齐任务的奖励函数

在偏好学习中，传统方法通常依赖输出之间的余弦相似度或预训练的标量奖励模型。相比之下，我们利用 S-GRPO 的鲁棒性来定义基于逻辑相似度的奖励函数。模型输出和参考响应均通过预训练一阶逻辑翻译模型转换为一阶逻辑表达式。随后，我们按照 [17] 提出的方法计算它们的逻辑等价性。

如图 2 所示，每个一阶逻辑表达式都会进行分词化，并计算成对余弦相似度。相似度高于 0.6 阈值的词元被视为在语义上相关。在一对多匹配场景中，会枚举所有可能的对齐关系，并选择得分最高的匹配作为最终对应关系。

这一设计为强化学习建立了基于逻辑的奖励机制，与传统的标量奖励模型相比，提供了更具可解释性且语义更加一致的优化目标。

IV. 设置

我们的实验分为两部分。首先，我们在两项翻译任务上评估 S-GRPO：一阶逻辑（FOL）[16] 翻译和 WMT 英德语翻译 [13]–[15]。其次，我们使用带有基于逻辑相似度的奖励的 PKU-SafeRLHF [19] 数据集，评估其在偏好学习方面的表现。所有实验均使用 Qwen3-1.7B [25] 作为骨干模型；选择该模型是因为它在性能与计算效率之间取得了平衡，从而确保实验具有可复现性和可控性。我们将所有奖励函数值限制在范围 [0, 1]内。

A. 一阶逻辑翻译设置

数据集。我们使用 FOLIO [18] 数据集，其中包含成对的自然语言陈述及其对应的一阶逻辑（FOL）表示。按照 Yang 等人 [17],所述的预处理流程，我们获得了 2,175 个有效样本。由于未提供官方测试集划分，我们随机将五分之一的数据分配用于测试。为确保鲁棒性，所有报告结果均取三次随机训练–测试划分的平均值。

基线。我们使用监督微调（SFT）作为主要基线。我们还参考 LogicLLaMA[17], 的结果；该模型采用带纠正机制的 RLHF。尽管模型和训练设置存在差异，我们仍将其结果作为参考进行比较。

评估指标。遵循 Yang 等人 [17],的设置，我们采用 BLEU 和逻辑等价（LE）作为评估指标。然而，我们发现，他们原先的 LE 实现效率低下，往往会导致匹配不准确且运行时间较长。为解决这一问题，我们提出了一种经过优化的 LE 指标，其得分比原始指标低 1–2%，但匹配速度更快、准确度更高。使用优化版本计算的结果在我们的表格中以星号（*）标记。有关优化的更多详情，请参阅附录 A。

B. WMT双语翻译设置

数据集。对于双语翻译，我们在 WMT 2017–2020 英德数据集上进行训练，并在 WMT 2022 上进行评估，涵盖英语→德语和德语→英语两个方向。

基线。我们再次使用 SFT 作为主要基线，并将我们的方法与以类似目标使用大语言模型进行翻译的先前研究进行比较。请注意，由于训练数据和模型架构存在差异，这种比较可能并不完全公平。

指标。我们使用两项互补指标进行评估：来自 sacreBLEU [26] 的 BLEU，以及来自 [27] 的 COMET（wmt22-comet-da），二者共同评估语义充分性和上下文连贯性。这两项指标也都用作 S-GRPO 训练的奖励信号。

C. 偏好学习设置

数据集。我们使用由北京大学 PKU-Alignment 团队发布的 PKU-SafeRLHF [19] 数据集，该数据集旨在从有用性和无害性两个维度提升大语言模型的安全对齐。由于并非所有标注都具有相同的偏好回答，我们筛选出约 30,000 条两个维度偏好一致的样本用于训练。之后的所有实验均在此子集上进行。

基线。我们将 S-GRPO 与 SFT 和直接偏好优化（DPO）[28] 进行比较。

指标 根据 Chiang 等人 [29],的研究发现，大语言模型的输出与人类偏好判断之间呈现出显著的正相关。因此，我们使用 DeepSeek-V3 [35] 作为自动评估器，评估偏好对齐情况。对于每个样本，模型输出都会与 Qwen3-1.7B [25] （未经训练的基线）生成的回答进行比较。评审模型会判断哪个输出更能体现有帮助且无害的行为，并给出胜/平局/负的结果。为减少潜在的位置偏差，我们交替调整两个模型输出的呈现顺序[23],。只有在任一顺序下都不落败的模型才被视为获胜。具体定义如下：**胜**：在两种顺序下表现都更好，或一胜一平。**平局**：两种顺序下均为平局，或一胜一负。**负**：在两种顺序下表现都更差，或一平一负。

V. 结果与分析

在一阶逻辑翻译和 WMT 英德语翻译任务中，我们观察到，基座模型在训练初期无法生成有意义的输出。为解决这一问题，我们首先在相应的数据集上进行监督微调（SFT），以确保模型能够生成对齐且连贯的回答。为保证公平，我们对 SFT 基线模型进行了两个完整训练轮次，并将表现最佳的检查点作为基线。

A. 一阶逻辑评估

表 I 汇总了一阶逻辑翻译在划分后的测试集上的表现。表中报告了两项指标：逻辑等价和 BLEU。我们的模型采用了改进版的逻辑等价指标，并以星号（*）标记。对于 BLEU，我们使用 SacreBLEU 重新实现了 Yang 等人 [17]的 Hugging Face Evaluate[1]版本。尽管整体框架保持一致，但底层实现中的细微差异可能会导致结果略有不同，这不会影响比较的有效性。我们还使用改进后的 LE 指标评估了先前的模型；然而，由于 Yang 等人 [17]的方法涉及相对复杂的基于人类反馈的强化学习流程，我们无法在 LE* 指标下复现其结果。

模型	FOLIO LE	FOLIO LE*	FOLIO BLEU
ChatGPT-3.5	80.2	77.6	37.0
DeepSeek-V3	83.0	79.2	37.6
ChatGPT-4o	82.6	80.9	38.4
LogicLLaMA-7B RLHF 校正	84.1	——	37.8
LogicLLaMA-13B RLHF Corre.	85.8	——	38.4
Qwen3-1.7B-SFT	——	85.0	61.2
Qwen3-1.7B-SGRPO	——	**87.4**	**66.0**

表 I 模型在 FOLIO 划分测试集上的评估结果。

有趣的是，SFT 基线在 FOL 翻译方面似乎明显优于较早期的模型，尤其是在 BLEU 得分方面。这可能是由于 FOL 的形式化特性：现代大语言模型已经能够处理其表层结构，尽管更深层的逻辑语义仍然具有挑战性。尽管如此，S-GRPO 仍带来了显著提升，在这项任务上胜过所有现有方法和模型。

B. WMT 翻译评估

WMT 评估结果如表 II 所示。Qwen3 基座模型在德语到英语的翻译任务上表现出色，但在英语到德语的翻译任务上表现相对较差。在这两个翻译方向上，SGRPO 的表现始终优于 SFT。鉴于 WMT 数据集规模庞大，我们训练了五个训练轮次，以比较 SFT 和 SGRPO 的性能变化。如图 4 所示，得益于底层 GRPO 架构，SGRPO 在整个训练过程中保持稳定；而 SFT 在第一个训练轮次后开始退化。

图 3 绘制了 SGRPO 训练期间奖励的均值和标准差。平均奖励稳步上升，尽管标准差略有增加，但波动仍然较小。总体而言，作为一种监督式方法，SGRPO 表现出很强的鲁棒性。

C. 偏好任务评估

PKU-SafeRLHF 测试集上的评估结果见表 III。值得注意的是，我们使用 DeepSeek 而非常用的 GPT-4 作为自动评判模型。此外，我们还

[1]https://github.com/huggingface/evaluate

模型	德语 ⇒ 英语 bleu	德语 ⇒ 英语 彗星	英 ⇒ 德 bleu	英 ⇒ 德 彗星
NLLB-54B [30]	26.89	78.94	34.50	86.45
GPT-4 [4]	33.87	85.62	35.38	87.44
TIM-7B [31]	27.91	82.80	25.59	82.56
ALMA-7B [34]	29.56	83.95	30.31	85.59
BigTranslate-13B [32]	23.35	80.68	21.48	78.81
Bayling-13B [33]	27.34	83.02	25.62	82.69
ALMA-13B [34]	31.14	84.56	31.47	85.62
Qwen3-1.7B-SFT	28.55	85.79	19.57	77.70
Qwen3-1.7B-SGRPO	30.18	86.17	21.07	79.13

表 II WMT 2022 英语—德语双语竞赛评估结果

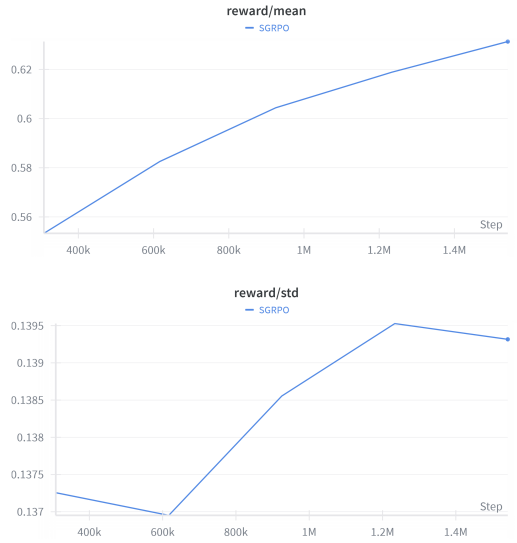


图 3. 训练 5 轮期间奖励的均值和标准差。

按照 [29]，的做法纳入平局，因为我们观察到，现代大语言模型往往会提供内容详尽、结构清晰的回答，即使是人工标注员也很难明确判定胜者。因此，纳入平局能够保留更细致的比较。

模型	PKU-SafeRLHF Win	PKU-SafeRLHF Tie	PKU-SafeRLHF Lost
Qwen-1.7B-SFT	0.95	2.41	96.62
Qwen-1.7B-DPO	28.69	53.95	17.35
Qwen-1.7B-SGRPO	22.82	52.09	25.09

表 III DeepSeek-V3 API 对每个 PKU-SafeRLHF 样本的偏好结果（胜/平/负）

SFT在此基准测试中表现不佳。PKU-SafeRLHF训练集主要由简短回答构成，而基座模型即使面对非专家查询，也倾向于生成冗长详尽的答案。这种不匹配导致了大量惩罚。DPO表现最佳，这可能是因为它避免了基于强化学习的方法中常见的高方差问题；鉴于基座模型自然

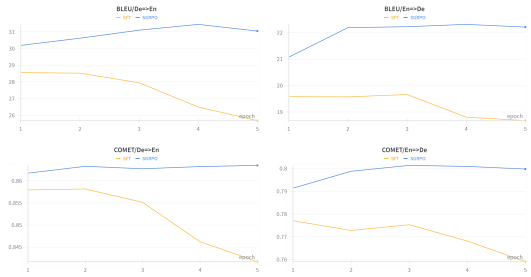


图4. 5个训练轮次中的性能变化。

满足 $\pi_\theta(chosen|x) > \pi_\theta(rejected|x)$，DPO优化仍能保持稳定。

SGRPO 的表现不如 DPO，但仍优于 SFT，这符合预期。在 PKU-SafeRLHF 上训练时，SGRPO 的平均奖励为**0.17** ，标准差为**0.30**。作为一种由奖励驱动的方法，尽管奖励值较低且奖励信号波动剧烈，SGRPO 仍在这一具有挑战性的场景中保持了稳定的表现，进一步证明了其鲁棒性。

VI. 结论

我们的实验表明，作为 GRPO 与 SFT 的混合方法，S-GRPO 有效结合了两种方法的优势——既保留了 GRPO 的探索能力，又融入了 SFT 的监督对齐。这使 S-GRPO 尤其适用于拥有高质量标注数据的任务，并提供了一种鲁棒且稳定的训练范式，可扩展至其他监督或半监督场景。

在偏好学习中，S-GRPO 还通过引入基于逻辑的奖励信号，提供了一种超越奖励模型的新视角。尽管语义不匹配仍可能造成轻微的不一致，但优化逻辑匹配策略有望进一步缩小这一差距。总体而言，S-GRPO 在逻辑密集型任务中展现出更大的潜力，而逻辑一致性与可解释性对于对齐至关重要。

VII. 局限性

在偏好学习任务中，模型输出与标签之间的语义差异可能导致逻辑形式不一致，这在当前限制了 S-GRPO 在纯粹基于偏好的场景中的效果。尽管如此，对于推理、结构化翻译或形式化文本生成等需要强逻辑一致性或语义一致性的任务，其逻辑等价奖励机制仍然极具潜力。

S-GRPO 也存在一些局限性。尽管它是一种监督学习方法，但由于训练期间需要进行序列生成，仍需消耗大量计算资源，因此其成本显著高于依赖更高效教师强制机制的 SFT 或直接偏好优化。因此，除非任务能明确受益于基于强化学习的探索，否则应将 S-GRPO 的训练成本视为与 GRPO 相当。

参考文献

[1] Vemprala, Sai H., Rogerio Bonatti, Arthur Fender C. Bucker 和 Ashish Kapoor。《用于机器人学的 ChatGPT：设计原则与模型能力》。IEEE Access 12 (2023)：55682-55696。 [2] Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wain-wright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike 和 Ryan J. Lowe。《通过人类反馈训练语言模型遵循指令》。ArXiv abs/2203.02155 (2022)：无页码。[3] Chuheng Zhang, Wei Shen, Li Zhao, Xuyun Z hang, Xiaolong Xu, Wanchun Dou 和 Jiang Bian。《用于缓解奖励模型噪声的RLHF策略筛选》。(2025)。 [4] OpenAI, Achiam 等，《GPT-4技术报告》，2024。[5] Shen, Wei, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue 和 Lin Yan。《探索基于人类反馈的强化学习中的数据扩展趋势与影响》。ArXiv abs/2503.22230 (2025)：无页码。[6] Hu, Jian, Xibin, Wu, Wei, Shen, Jason Klein, Liu, Weixun, Wang, Songlin, Jiang, Haoran, Wang, Hao, Chen, Bin, Chen, Wenkai, Fang, Xianyu, Yu, Cao, Haotian, Xu 和 Yiming, Liu。《OpenRLHF：基于 Ray、易于使用、可扩展且高性能的 RLHF 框架》。载于《2025年自然语言处理实证方法会议论文集：系统演示》（第 656–666 页）。计算语言学协会，2025。 [7] Anthropic。2025。Claude（语言模型）概述。Anthropic。https://docs.anthropic.com/zh-TW/docs/about-claude/models[8] Gemini Team 等，《Gemini：一系列能力强大的多模态模型》，2025。[9] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。《近端策略优化算法》。(2017)。[10] Ziniu Li、Tian Xu、Yushun Zhang、Zhihang Lin、Yang Yu、Ruoyu Sun 和 Zhi-Quan Luo。《ReMax：一种用于对齐大语言模型的简单、有效且高效的强化学习方法》。(2024)。 [11] Ahmadian, Arash, Chris, Cremer, Matthias, Galle, Marzieh, Fadaee, Ju-lia, Kreutzer, Olivier, Pietquin, Ahmet, Ust¨un 和 Sara, Hooker。《回归基础：重新审视用于大语言模型人类反馈学习的 REINFORCE 风格优化》。载于《第62届计算语言学协会年会论文集》（第 1 卷：长篇论文）（第 12248–12267 页）。计算语言学协会，2024。[12] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。《DeepSeekMath：突破开放语言模型数学推理能力的极限》。(2024)。[13] Bojar, Ondřej, Rajen, Chatterjee, Christian, Federmann, Yvette, Graham, Barry, Haddow, Shujian, Huang, Matthias, Huck, Philipp, Koehn, Qun, Liu, Varvara, Logacheva, Christof, Monz, Matteo, Negri, Matt, Post, Raphael, Rubino, Lucia, Specia 和 Marco, Turchi。《2017年机器翻译会议（WMT17）研究成果》。载于《第二届机器翻译会议论文集》（第 169–214 页）。计算语言学协会，2017。[14] Barrault, Loic, Ondřej, Bojar, Marta R., Costa-jussa, Christian, Federmann, Mark, Fishel, Yvette, Graham, Barry, Haddow, Matthias, Huck, Philipp, Koehn, Shervin, Malmasi, Christof, Monz, Mathias, M¨uller, Santanu, Pal, Matt, Post 和 Marcos, Zampieri。《2019年机器翻译会议（WMT19）研究成果》。载于《第四届机器翻译会议论文集》（第 2 卷：共享任务论文，第 1 天）（第 1–61 页）。计算语言学协会，2019。[15] Germann, Ulrich。《爱丁堡大学在 WMT 2020 新闻翻译及零样本翻译鲁棒性任务中德译英和英译德赛道的参赛系统》。载于《第五届机器翻译会议论文集》（第 197–201 页）。计算语言学协会，2020。 [16] Ben-Ari, Mordechai。《计算机科学中的数理逻辑》。Springer London (2012)。[17] Yang, Yuan, Siheng, Xiong, Ali, Payani, Ehsan, Shareghi 和 Faramarz, Fekri。《发挥大语言模型在自然语言到一阶逻辑翻译中的潜力》。载于《第62届计算语言学协会年会论文集》（第

1 卷：长篇论文）（第 6942–6959 页）。计算语言学协会，2024。[18] Han, Simeng, Hailey, Schoelkopf, Yilun, Zhao, Zhenting, Qi, Martin, Riddell, Wenfei, Zhou, James, Coady, David, Peng, Yujie, Qiao, Luke, Benson, Lucy, Sun, Alexander, Wardle-Solano, Hannah, Szabo, Ekaterina, Zubova, Matthew, Burtell, Jonathan, Fan, Yixin, Liu, Brian, Wong, Malcolm, Sailor, Ansong, Ni, Linyong, Nan, Jungo, Kasai, Tao, Yu, Rui, Zhang, Alexander, Fabbri, Wojciech Maciej, Kryscinski, Semih, Yavuz, Ye, Liu, Xi Victoria, Lin, Shafiq, Joty, Yingbo, Zhou, Caiming, Xiong, Rex, Ying, Arman, Cohan 和 Dragomir, Radev。《FOLIO：基于一阶逻辑的自然语言推理》。载于《2024年自然语言处理实证方法会议论文集》（第 22017–22031 页）。计算语言学协会，2024。[19] Ji, Jiaming, Donghai, Hong, Borong, Zhang, Boyuan, Chen, Josef, Dai, Boren, Zheng, Tianyi Alex, Qiu, Jiayi, Zhou, Kaile, Wang, Boxun, Li, Sirui, Han, Yike, Guo 和 Yaodong, Yang。《PKU-SafeRLHF：通过人类偏好实现大语言模型的多级安全对齐》。载于《第63届计算语言学协会年会论文集》（第 1 卷：长篇论文）（第 31983–32016 页）。计算语言学协会，2025。 [20] Long Ouyang 、 Jeff Wu 、 Xu Jiang、Diogo Almeida、Carroll L. Wain-wright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。《通过人类反馈训练语言模型遵循指令》。(2022)。[21] Avishai Elmakies、Hagai Aronowitz、Nimrod Shabtay、Eli Schwartz、Ron Hoory 和 Avihu Dekel。《借助 GRPO 提升语音感知语言模型的语音理解能力》。(2025)。[22] Feng, Zhaopeng, Shaosheng, Cao, Jiahan, Ren, Jiayuan, Su, Ruizhe, Chen, Yan, Zhang, Jian, Wu 和 Zuozhu, Liu。《MT-R1-Zero：通过类R1-Zero强化学习推进基于大语言模型的机器翻译》。载于《计算语言学协会研究成果：EMNLP 2025》（第 18685–18702 页）。计算语言学协会，2025。[23] Yuhao Du、Zhuo Li、Pengyu Cheng、Zhihong Chen、Yuejiao Xie、Xiang Wan 和 Anningzhe Gao。《将 RLHF 简化为奖励加权监督微调：一种变分方法》。(2025)。[24] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。《DeepSeekMath：突破开放语言模型数学推理能力的极限》。(2024)。[25] An Yang 等，《Qwen3技术报告》，2025。[26] Post, Matt。《呼吁清晰报告 BLEU 分数》。载于《第三届机器翻译会议论文集：研究论文》（第 186–191 页）。计算语言学协会，2018。 [27] Rei, Ricardo, J ose G., Souza, Duarte, Alves, Chrysoula, Zerva, Ana C, Farinha, Taisiya, Glushkova, Alon, Lavie, Luisa, Coheur 和 Andr´e F. T., Martins。《COMET-22：Unbabel-IST 2022指标共享任务参赛系统》。载于《第七届机器翻译会议论文集（WMT）》（第 578–585 页）。计算语言学协会，2022。 [28] Rafael Rafailov、Archit Sharma、Eric Mitchell、Stefano Ermon、Christopher D. Manning 和 Chelsea Finn。《直接偏好优化：你的语言模型其实是一个奖励模型》。(2024)。[29] Chiang, Cheng-Han 和 Hung-yi, Lee。《大语言模型能否成为人工评估的替代方案? 》。载于《第61届计算语言学协会年会论文集》（第 1 卷：长篇论文）（第 15607–15631 页）。计算语言学协会，2023。 [30] NLLB团队，Marta R. Costa-juss`a、James Cross、Onur C.elebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Maillard、Anna Sun、Skyler Wang、Guillaume Wenzek、Al Youngblood、Bapi Akula、Loic Barrault、Gabriel Mejia Gonzalez、Prangthip Hansanti、John Hoffman、Se marley Jarrett、Kaushik Ram Sadagopan、Dirk Rowe、Shannon Spruit、Chau Tran、Pierre Andrews、Necip Fazil Ayan、Shruti Bhosale、Sergey Edunov、Angela Fan、Cynthia Gao、Vedanuj Goswami、Francisco Guzm´an、Philipp Koehn、Alexandre Mourachko、Christophe Ropers、Safiyyah Saleem、Holger Schwenk 和 Jeff Wang。《不让任何语言掉队：扩展以人为本的机器翻译》。(2022)。

[31] Jiali Zeng、Fandong Meng、Yongjing Yin 和 Jie Zhou . "TIM：教授-借助比较进行翻译的大语言模型。"（2024年）。

[32] Wen Yang、Chong Li、Jiajun Zhang 和 Chengqing Zong。《BigTranslate：为大语言模型扩展覆盖100种语言的多语言翻译能力》。（2023年）。[33] Shaolei Zhang、Qingkai Fang、Zhuocheng Zhang、Zhengrui Ma、Yan Zhou、Langlin Huang、Mengyu Bu、Shangtong Gui、Yunji Chen、Xilin Chen 和 Yang Feng。《BayLing：通过面向大语言模型的交互式翻译，弥合跨语言对齐与指令遵循之间的鸿沟》。（2023年）。[34] Haoran Xu、Young Jin Kim、Amr Sharaf 和 Hany Hassan Awadalla。《机器翻译的范式转变：提升大语言模型的翻译性能》。（2024年）。[35] DeepSeek-AI, 《DeepSeek-R1：通过强化学习激励大语言模型的推理能力》, 2025年。

附录 A 逻辑等价的设计缺陷与优化

[17] 中的逻辑等价（LE）存在固有局限。从原子公式绑定的角度来看， [17]的方法基于莱文斯坦编辑距离升序，对所有可能的匹配进行穷举，导致时间复杂度为 $O(n!)$，n 是原子公式集合的长度（例如，在图 2 中， $Anger(x) \rightarrow NonPhysical(x)$ 是 $n = 2$； $\forall x \forall y(Push(x, y) \wedge NotHurt(x, y) \rightarrow Anger(x, y)$ 是 $n = 3$）。该方法输出计算所得逻辑等价（LE）分数中的最高分；由于忽略了真实的匹配准确性，这会使最终分数产生固有的向上偏差。从计算复杂度的角度来看，LE 方法中资源消耗尤为大的步骤是语法树构造：该算法为一阶逻辑表达式构建 NLTK解析树，并穷举所有可能的树结构，以确定最优配置。此过程的时间复杂度呈指数级，为 $O(2^{n-1})$，对于较长的一阶逻辑表达式而言，计算成本高得难以承受。

我们提出了一种改进方案，通过计算语义单元之间的余弦相似度来解决原子公式绑定问题。当出现一对多匹配场景时（即单个单元匹配多个候选单元），该方案仅对相关候选项执行排列遍历，输出得分最高的匹配结果，从而解决匹配不一致问题。优化后的算法在最佳情况下的时间复杂度为 O(n)，最坏情况下为 O(n!)；实证分析表明，在实践中，一对多匹配中的候选单元数量很少超过三个。因此，平均时间复杂度仍处于计算可行范围内，效率得到显著提升。

为了解决语法树构造问题，我们提出了一种用于一阶逻辑解析的基于分块的分解方法，通过以下步骤显著降低计算复杂度：(1) 将表达式划分为易于处理的块，(2) 独立地为每个子块构造NLTK解析树，以及 (3) 系统地将部分树合并为完整语法树。该方法不仅保持了实现简洁性，还将复杂度从原始的 $O(2^{n-1})$降至平均情况下的 $O(2^{n-1/m})$，其中 m 表示块大小（通常为 3-5 个子单元）。只有在极端情况下，该方法才会退化为最坏情况下的 $O(2^{n-1})$复杂度。