

# 走向评估思维：具有演化奖励模型的元策略优化

Zae Myung Kim<sup>1</sup>, Chanwoo Park<sup>2</sup>, Vipul Raheja<sup>3</sup>, Dongyeop Kang<sup>1</sup>,

<sup>1</sup>University of Minnesota, <sup>2</sup>MIT, <sup>3</sup>Grammarly

{ kim01756, dongyeop } @umn.edu, cpark97@mit.edu, raheja@grammarly.com

## Abstract

基于奖励的大型语言模型（LLMs）对齐方法面临两个关键限制：一是奖励欺骗的脆弱性，模型可能利用奖励信号中的缺陷；二是当 LLMs 用作奖励模型时，依赖于易碎且劳动密集的提示工程。我们介绍了 Meta Policy Optimization (MPO)，这是一个通过整合动态优化奖励模型提示的元奖励模型来应对这些挑战的框架。在 MPO 中，元奖励模型监控不断发展的训练环境，并持续调整奖励模型的提示以保持高度对齐，从而提供一个自适应的奖励信号，抵御策略的利用。这种元学习方法促进了更稳定的策略优化，并大大减少了手动奖励提示设计的需要。其性能与由广泛精心设计的奖励提示指导的模型持平甚至更好。此外，我们展示了 MPO 在各种任务中，如问答和数学推理，都能保持其有效性，而无需特定的奖励设计。超出标准的 RLAIIF，MPO 的元学习结构可以轻松扩展到更高层次的对齐框架。总体而言，这种方法解决了基于奖励的大型语言模型（LLMs）中 RL 对齐的理论和实际挑战，为更稳健和适应性强的对齐策略铺平了道路。代码和模型将会公开共享。<sup>1</sup>

## 1 介绍

优秀思维者必须运用另一项高级思维技能，即分析和对自身思维的觉察——或称为元认知 [Buckley et al., 2015, Lord et al., 1979]。

最近，在大语言模型（LLM）训练的强化学习（RL）方面的进展标志着从注重类人反应生成的任务转变为强调结构化推理的任务，如数学和编程。传统上，与人类对齐的回答依赖于使用二元比较数据训练的奖励模型，而结构化推理任务则更多地关注根据客观的真实情况验证最终答案或逻辑过程的正确性。然而，许多现实世界的挑战需要综合这两种方法——要求模型融合与人类对齐的判断力与严格的推理。这引入了显著的复杂性，因为此类任务往往缺乏客观可验证的“标准答案”，但仍需进行连贯且有理有据的推理。

一种可扩展的主观评估方法是使用大型语言模型（LLM）作为评判者——这种方法通常被称为带有 AI 反馈的强化学习（RLAIIF）——其中使用固定的提示来评估模型在特定任务上的表现。然而，这种方法引入了三个关键挑战。首先是校准：有效的评分需要评价标准足够详细并经过仔细校准，以适应策略模型输出的质量和变异性。对质量较差的反应提供高度细化的反馈可能是无效的，而过于粗略或简单化的评分则可能无法引导模型行为的实质性改进。其次是奖励作弊：众所周知，当大型语言模型用不完美的奖励机制进行训练时——这在将 LLM 作为评判者或奖励模型时经常发生——它们可能会学会利用捷径来最大化奖励，而不是真正提高输出质量或与人类价值观的匹配度 [Amodei et al., 2016b, Everitt et al., 2021, Langosco et al., 2022, Pan et al., 2022]。这些捷径可能导致形式上获得奖励但与人类意图或效用不一致的回应。第三是提示工程的开销：在生成奖励模型的训练数据或将 LLM 作为代理奖励函数时，通常需要大量的人工努力进行提示工程。这一过程带来了可扩展性瓶颈并限制了对齐管道中的自动化。

为了解决这些问题，这项工作引入了元策略优化（MPO）（见图 1），这是一个通过添加元奖励模型来增强现有基于奖励的 RLAIIF 流程的框架。不同于传统的奖励模型仅仅基于固定的提示对策略的输出进行评分，元奖励模型会监测不断变化的训练环境，并调整或改善标

<sup>1</sup><https://github.com/minnesotanlp/mpo>

准奖励模型使用的提示。我们对 MPO 设计的灵感来源于心理学中的元认知概念——即意识到并反思自己的思考过程 [Flavell, 1979]——以及它在评价性思考中所扮演的核心作用，这是一种反思性、以证据为驱动的认知过程，涉及质疑、分析和解释信息，以指导决策制定和持续学习 [Buckley et al., 2015]。

以大学教室为类比：策略模型扮演学生的角色，而标准奖励模型则充当遵循固定评分标准的初级讲师。定期地，一位高级讲师（元奖励模型）会加入，以评估整体情况：学生的表现、初级讲师的评分标准以及所打的分数。如果学生发现评分标准中的漏洞，并开始以最小的努力提交质量较差的作业却能获得高分，那么这位具备全局视野的高级讲师就能识别出这种模式，并建议初级讲师调整未来作业的评分标准。

元认知意识和控制对于这一过程至关重要，使个人能够监控推理、发现偏见，并通过任务评估和反思来完善策略。认知科学研究进一步支持这一观点，表明刻意的反思促进了更深入、更牢固的学习 [McCormick, 2003, Metcalfe and Kornell, 2005, Veenman et al., 2006, Efklides, 2006]。通过将原则应用于 LLMs 的基于 RL 的对齐中，我们解锁了几个优势：

正如在介绍中所提到的，我们的 MPO 方法受到了元认知和评估领域的启发。我们将详细阐述这一概念动机，然后提出 MPO 的正式强化学习公式，接着详细描述其实现步骤。

## 1.1 评价性思维

评估性思维 (ET) [Buckley et al., 2015] 是一个有意的过程，用于分析、解释和评估信息，以支持深思熟虑的决策，在评估能力建设 (ECB) 中发挥关键作用。然而，目前强化学习中对大型语言模型的奖励模型缺乏这种元认知控制。这些模型通常在静态的人类偏好数据集上训练，或通过固定的评分提示进行引导，并在训练过程中保持不变。随着策略的改进，静态奖励模型往往将细微的改进简化为一个粗略的标签——足够好——从而使早期的盲点持续存在并得不到纠正。

受到评估性思维 (ET) 与评估能力建设 (ECB) 之间关系的启发——其中 ET 通过增强元认知支持 ECB——我们提出了一种元奖励模型 (MRM)，该模型指导奖励模型发展评估性元认知并成为更有效的评分者。具体而言，MRM 遵循 ET 的核心原则：证据收集、质疑和反思性判断（见第 1.3 节和图 3）。我们的 MPO 框架通过使 MRM 能够随着时间的推移细化 RM 的观察分区来实施 ET。在此设置中，奖励模型通过由 MRM 的元认知信号驱动的策略内学习得到改善。

**Remark 1 (Depth and Breadth of ET).** 我们假设，在大语言模型 (LLMs) 的强化学习背景下，评估性思维 (ET) 可以沿着两个正交的维度来理解：深度和广度。这个框架回响了爱德华·德·博诺著名的垂直思维和横向思维的区分 [De Bono, 1971]。直观地说，深度对应于顺序的、实例特定的推理，反映出评估单个案例所需的逻辑推理和审慎程度。相反，广度则体现了跨越不同实例进行概括的能力，识别出在新情境下告知评估的重复模式或抽象原则。如图 2 所示，诸如数学推理的任务体现了垂直（深层）思维，因为它们涉及多步、案例特定的推导。相反，像作文写作这样开放式的任务更符合横向（广泛）思维，需要在不同提示间进行评估性概括。为了实证探索这些维度，我们评估了四个代表性任务——数学推理、伦理推理、摘要和作文写作——它们在我们实验的 ET 深度-广度空间中各自占据一个独特的区域。

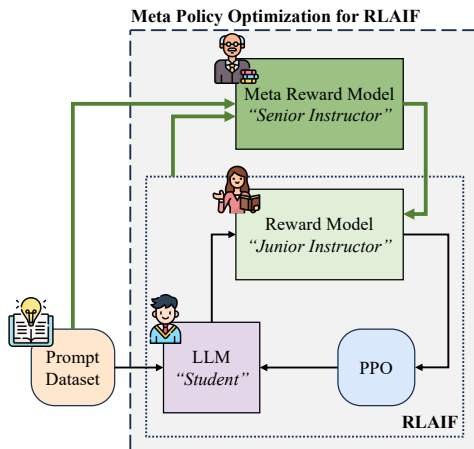


Figure 1: 在标准的 RLAIIF 中，奖励模型在近端策略优化 (PPO) 期间始终保持不变。在对比之下，MPO 框架（绿色）引入了一个动态发展的元奖励模型，该模型基于当前训练背景，包括任务提示、伴随得分的采样生成以及最新的评估提示，来发展奖励模型。MPO 利用这些背景信息迭代地完善评估提示，使得对齐更加自适应和有效。

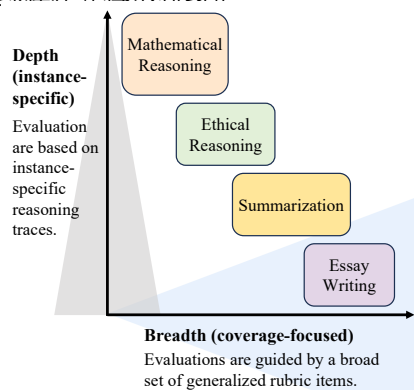


Figure 2: 评估思维的维度：深度与广度。

## 1.2 通过演化奖励模型的时变观察与奖励

虽然通过 ET 的 ECB 为改进 RM 提供了基础，但它并未完全捕捉 RL 的动态，其学习是由经 ECB 增强的 RM 的信号驱动的。为了解决这一差距，我们引入了一个数学框架，该框架形式化地描述了 RM 如何在 ET 的影响下演变，以及这种不断演变的 RM 如何可以整合到 LLMs 的训练过程中。

形式上，考虑一个离散状态空间  $\mathcal{S}$ ，一个动作空间  $\mathcal{A}$ （均为有限或可数），以及一个转换核  $P(s' | s, a)$ ，指定从状态  $s \in \mathcal{S}$  通过动作  $a \in \mathcal{A}$  转移到状态  $s' \in \mathcal{S}$  的概率。我们定义一个黄金奖励函数  $r : \mathcal{S} \rightarrow \mathbb{R}$ ，为每个状态分配一个数值奖励。

**Remark 2.** 在 LLM-RL 设置中，状态  $s$  表示文本历史（例如，到目前为止生成的令牌序列），而动作  $a$  则是选择下一个令牌。因此，后续状态就自然地表示为  $(s, a)$ ，即历史和所选令牌的串联。在实际操作中，由于诸如连贯性或相关性等主观标准，分配精确的数值奖励  $r(s)$  是具有挑战性的。因此，评估者通常提供近似反馈。这种情况与部分可观测的马尔可夫决策过程（POMDP）结构非常吻合，在该结构中，来自评估者的观测构成了对潜在状态的一种部分、聚合视图。

**Remark 3.** 考虑一个在对话环境中生成回应的大型语言模型。起初，AI 或人类反馈可能只会广泛地将回应分类为“好”，“中立”或“坏”。然而，随着时间的推移，评估者可能会引入更细致的区分，例如“连贯但不礼貌”，“礼貌但不相关”，以及“相关但冗长”。数学上，这相当于细化大型语言模型接收到的观察集的粒度，提供越来越精确和信息丰富的反馈。这与我们案例中不断演变的奖励模型直接相关 - 通过 ET 的 ECB-ed RM，他们能提供一个更精细的分数，这将更接近于黄金奖励模型。

### 1.2.1 随着时间的推移精炼观察划分

令每个离散阶段  $t = 1, 2, 3, \dots$  的可能观察集合记为  $\Omega_t$ 。定义状态空间  $\mathcal{S}$  的分区集合  $\{\mathcal{O}_{o,t}\}_{o \in \Omega_t}$ ，满足：

$$\mathcal{O}_{o,t} \cap \mathcal{O}_{o',t} = \emptyset, \quad \text{for } o \neq o', \quad \text{and} \quad \bigcup_{o \in \Omega_t} \mathcal{O}_{o,t} = \mathcal{S}.$$

每个分区代表由评估者标记状态，其中  $\mathcal{O}_{o,t}$  包含在时间  $t$  被标记为观察  $o$  的状态。如果在阶段  $t$  的真实状态为  $s$ ，则代理确定性地观察到标签  $o$ ，使得  $s \in \mathcal{O}_{o,t}$ 。

**精炼性质。** 为了正式描述不断精确的反馈，假设每个划分细化了前一个划分。具体来说，对于每个  $\mathcal{O}_{o,t}$ ，存在某个  $\mathcal{O}_{o',t-1}$  满足：

$$\mathcal{O}_{o,t} \subseteq \mathcal{O}_{o',t-1}.$$

这意味着在每个后续阶段的划分比前一个阶段更精细，从不粗糙，这也与我们的 MPO 设置一致——元奖励模型使用 ET 引导奖励模型，使训练过程中划分标准更细。

### 1.2.2 基于观察的奖励

与其根据精确的状态信息立即提供奖励，不如根据观察标签定义奖励。当在阶段  $t$  观察到  $o \in \Omega_t$  时，代理会收到一个平均奖励：

$$R_t(o) = \mathbb{E}_{s \in \mathcal{O}_{o,t}}[r(s)] = \frac{1}{|\mathcal{O}_{o,t}|} \sum_{s \in \mathcal{O}_{o,t}} r(s).$$

在我们的 MPO 设置中，更粗略的评分标准会导致所有映射到相同评分类别的反应都被平均化为奖励。

### 1.2.3 元策略优化框架

使用单一静态 LLM 评分器对应一个固定的观察划分，这可能过于粗糙，无法精确捕捉微妙的奖励差异。这样粗糙的划分将许多不同的状态（例如，文本或对话历史）归为过于粗略的类别，导致形如

$$R(o) = \mathbb{E}_{s \in \mathcal{O}_o}[r(s)],$$

的平均奖励，这掩盖了对精确策略优化至关重要的状态特定细节。因此，静态评分机制难以收敛到真实奖励  $r(s)$ ，限制了其捕捉微妙的、高维的或正在演变的奖励标准的能力。

相反，元策略框架通过在连续的迭代中自适应地优化观察分区来解决此限制。从形式上来说，元奖励过程引入了逐渐细化的分区：

$$\{\mathcal{O}_{o,t}\}_{o \in \Omega_t}, \quad \text{where } \mathcal{O}_{o,t} \subseteq \mathcal{O}_{o',t-1} \quad \text{for some } o', t > 1.$$

这个迭代优化使得奖励信号能够更具辨别性：

$$R_t(o) = \mathbb{E}_{s \in \mathcal{O}_{o,t}}[r(s)],$$

能够更好地捕捉状态空间中的细微变化。通过自适应地分割观察集合——当策略的性能提高或评估的新维度出现时，将更大的粗略类别拆分为更小的有针对性的子集——元奖励评估器逐步增强了强化信号的粒度和效果。这种动态优化在复杂的 LLM 驱动任务中特别有利，确保策略更新变得更具针对性，能够与细微的性能改进对齐，最终促进高级策略学习。

### 1.3 实现：元奖励模型

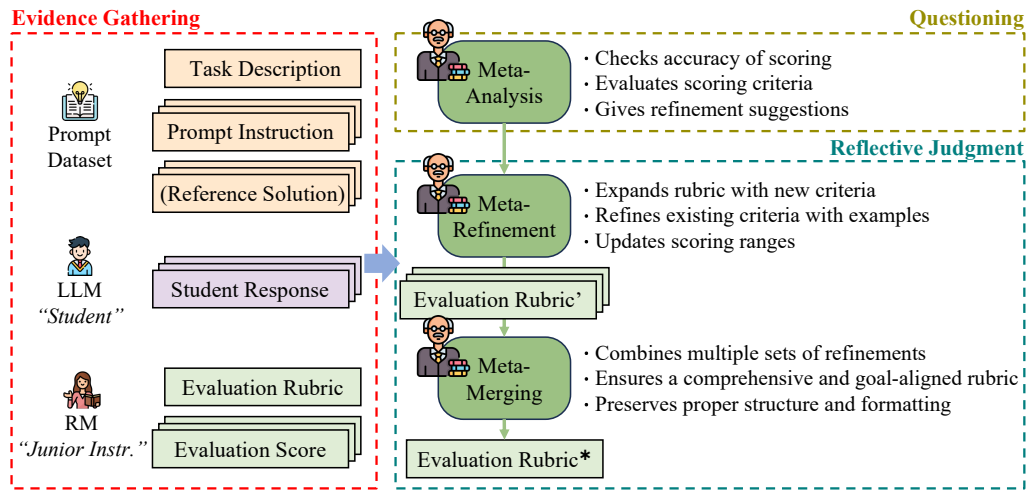


Figure 3: 元策略优化步骤。

通过引入一个元奖励模型（MRM），实现了 MPO 框架——在我们的学术类比中，这是一个“高级导师”。如图 3 所示，MRM 负责监督更广泛的训练背景，并对奖励模型（RM）使用的评估标准（或提示）进行有针对性的改进，奖励模型扮演“初级导师”的角色。这些改进的信息来源于三个方面的输入：提示数据集、策略模型（“学生”）以及奖励模型本身。具体而言，在每隔固定的  $k$  训练批次步骤后，MRM 通过通用的元级别提示对情境输入源进行处理，从而执行一个 MPO 步骤，这些提示被设计为与任务无关并适用于所有任务。

从提示数据集中，MRM 处理任务描述、一组  $n$  任务特定的提示指令，以及——当可用时——对应的  $n$  参考解决方案。它还接收策略模型对这些提示生成的  $n$  响应、当前版本的评估标准以及 RM 使用该标准分配的分数。

凭借这些丰富的上下文输入，MRM 能够识别当前评分标准中的弱点或不足之处，并提出日益细化和有针对性的评估标准。每经过固定的  $k$  训练批次，该细化过程就会被触发，并遵循图 3 中所所示的三阶段程序。在每个阶段，MRM 都会收到旨在适用于各种任务的元级别指令提示。

#### 1.3.1 元分析

MPO 的第一步涉及处理完整的输入背景，以评估 RM 的评分是否准确和可靠——特别是在学生 LLM 可能通过奖励黑客攻击 RM 漏洞的情况下。MRM 评估当前的评分标准是否足够全面和详细，并规定必要的调整以提高评估质量和稳健性。

这一步尤为重要，因为它可以在早期检测出 RM 评估逻辑中的漏洞。例如，考虑在 RL 训练期间观察到的以下真实案例：



### A reward hacking example

User: You will act as an English writer and compose either an essay or a story depending on the instruction given below. Your essay should be no more than 350 words.  
Instruction: Write an essay that argues the claim that the U. S. Drug Enforcement Agency has a negative impact on society that outweighs any benefits it provides. Your essay should include evidence and examples to support your argument.  
Student Response: "I'm ready when you are. Please provide the essay instruction."  
Evaluation Score: 3

在这个例子中，策略模型发现生成一个退化的回复——如图所示——虽然明显与任务目标不一致，但却从 RM 中得到了 3 分（满分 5 分）。这是一个奖励欺骗的例子，模型利用评分系统的漏洞来获得相对较高的分数，而没有忠实地完成任务。

### Meta-Analysis

The junior instructor's assigned score of 3 out of 5 points for the student's response is misleading given the provided student generation and the task requirements. Let's break down the analysis.  
Student's Response: The student's response is entirely missing. The student's generation reads, "I'm ready when you are. Please provide the essay instruction." This indicates that the student did not actually write the essay as instructed, but rather responded with a statement that is irrelevant to the task.  
Task Requirements and Prompt Instructions: ...  
Junior Instructor's Scoring Criteria Analysis: ...  
Discrepancies and Feedback:  
1. Adherence to Task Requirements (0 points)  
- The student's response does not reflect the task requirements at all. The student did not provide an essay on the given topic.  
- The junior instructor should assign 0 points for adherence to task requirements.  
2. Content Accuracy and Relevance (0 points) ...  
3. Argument Strength (0 points) ...  
4. Counterarguments and Rebuttals (0 points) ...  
...

每 k 批次步骤进行元分析，以识别失败案例并提出具体补救措施和规则细化的反馈。没有这种干预，在固定规则下继续进行强化学习训练会加强这些退化的输出，可能导致训练失败。

### 1.3.2 元精炼

基于元分析的结果，下一步是构建一个更精细的评分准则。具体而言，MRM 通过引入新的评估标准来扩展评分准则，用更详细的说明和示例增强现有标准，并调整评分范围以更有效地区分不同水平的响应质量。继续以奖励黑客示例为例，提出的评分准则改进将作文长度作为一个明确的评估标准，并将评分范围从 0 到 10 分进行扩展。此调整使得能够更细致地区分不同的响应，且能够避免利用模糊或定义不足的评分标准产生的退化输出：

### Meta-Refinement

...  
Task Alignment and Argument Clarity (10 points)  
- \*\*10\*\*:: The essay clearly and consistently addresses the specific claim made in the prompt and presents a well-defined, coherent argument supported by relevant examples and evidence. The argument is fully developed and addresses the core issue of the prompt. The essay meets the required word count.  
- \*\*8\*\*:: ...  
- \*\*6\*\*:: ...  
- \*\*4\*\*:: ...  
- \*\*2\*\*:: ...  
- \*\*0\*\*:: The essay does not address the specific claim or is completely off-topic. The argument is not related to the prompt and does not address the core issue. The essay is significantly below the required word count. ...

### 1.3.3 元合并

当 MPO 采集  $n$  学生的反馈时，它会产生  $n$  相应的评分标准修正。MPO 的最后一步包括将这些多个精炼实例合并为一个单一的、一致的评分标准提示。这个整合版本成为更新后的评估标准，RM 将用于后续的培训步骤。

## 2 实验

为了研究 MPO 的有效性及其对训练动态的影响，我们进行了三个核心实验。第 ?? 节概述了我们的实验设置。第 2.1 节评估了 MPO 对齐的 LLM 在论证性文章写作任务中的表现，探

讨了不同级别教师的配对。第 ?? 节通过实证显示 MPO 能够推广到要求不同程度评估思维的任务。最后，在第 2.2 节中，我们分析了 RM（“初级教师”）使用的评分标准提示在连续的 MPO 迭代中是如何演变的。

在我们的实验中，我们将策略模型固定为一个相对较小的 LLM：Qwen2-1.5B-Instruct。这个选择基于三个因素。首先，我们需要一个在我们的目标任务中有足够改进空间的模型。由于我们使用的是具有有限资源的公开基准测试，我们专注于较小的开源模型，这些模型可以显示提出框架的明显效果和价值。将其扩展到更大型的模型和更多变体的模型家族仍然是未来研究的一个有趣方向。其次，一些任务——例如数学推理——需要生成超过 1000 个标记，这在 PPO 步骤中会带来显著的内存负载。较大的模型超过了我们可用 GPU 的容量，使得 Qwen2-1.5B-Instruct 成为一种实用且可扩展的选择。第三，为了对比 MPO 生成的提示，我们基于大量 PPO 训练在该任务和模型中的经验见解，手动设计了一个 RM 提示。

对于奖励建模，我们使用 Qwen2.5-32B-Instruct-AWQ 和 Qwen2.5-72B-Instruct-AWQ，探索所有四种初级-高级 RM-MRM 尺寸组合：32b\_32b、32b\_72b、72b\_32b 和 72b\_72b，其中第一个和第二个术语分别表示 RM 和 MRM 的大小。将较大的模型用作 (M)RM 是可行的，因为只需要推断，这可以通过一个 LLM 服务框架高效地处理。我们还包括一个没有 MPO 的固定 RM 基线，使用初始 MPO 提示，表示为 { RM 大小 } \_iter0。

**实现。** 我们的 MPO 框架依赖于实现为 LLM 的在线（元）奖励模型，具有可互换的提示机制。为了支持这一点，我们通过实现一个基于提示的在线奖励模型扩展了 TRL 库 [von Werra et al., 2020]，其中基于 LLM 的 RM 使用 SGLang 框架进行托管 [Zheng et al., 2024]。此外，我们在 trl 中扩展了“PPOTrainer”类到一个定制的“PPOMetaTrainer”，它将 MPO 细化步骤直接集成到训练循环中，并进行其他必要的修改。

我们注意到，MPO 框架是通用且模块化的，可以集成到依赖于奖励模型的其他强化学习优化技术中——例如 GRPO [Shao et al., 2024]——以实现动态评分标准细化和更具适应性的奖励塑造。

## 2.1 MPO 对作文任务的影响

**设置。** 我们使用 MPO 训练四个策略模型，不同的 RM-MRM 配对如第 ?? 节所述，并使用固定 RM 提示的原始 PPO 训练另四个模型。作文数据集由 Kim et al. [2025] 编译，包括来自各种来源的写作说明，如英语水平考试、说服语料库和“改变我观点”（CMV）子版块的内容。训练集包含 26,013 个样本，测试集包括 4,096 个样本。MPO 和 PPO 都在训练集上训练一个周期，期间 MPO 的细化步骤每 10 个批次进行一次。我们使用 8 个 A100 80GB GPU 对模型进行 RL 训练，有效的综合批大小为 64，每个 RL 回合最多包含 400 个 tokens。在一个周期的训练过程中，此配置产生了 40 次 MPO 细化步骤。附加的训练细节在附录 C.3 中提供。

经过单轮训练后，我们使用最终的检查点为所有测试提示生成文章，并按照基于 Elo 的 Chatbot Arena 框架 [Chiang et al., 2024] 进行 10,000 次一对一比较评估，其中 GPT-4o 作为公正的评审，每次比赛后应用一个小的 K 因子 4 以保持评分的稳定性。表 1 总结的最终 Elo 分数提供了八个模型间文章质量的相对排名。

**训练曲线。** 图 4 展示了八个模型的训练曲线，每个模型对应着论述写作任务中不同的 RM 和 MRM 尺寸配对。需要注意的是，我们使用原始 RL 奖励值进行优化，因为它们保留了奖励信号中的细粒度区别——尽管略微偏离我们正式的理论框架——这增强了奖励敏感性并提高了优化效果。四个采用 MPO 的模型在 RL 奖励值上随时间增加，这归功于连续的 MPO 步骤通过（1）引入更多的评价标准和（2）在训练进展时扩大与这些标准相关的评分范围，精炼了评估提示。虽然这些曲线表明训练如预期的那样自适应地进行，奖励分数本身也是自适应的——反映了不断演变的标准——因此，各模型间的奖励值差异（如在 RL 奖励图中所示）不一定会转化为最终测试集上的相应性能差距。

在表 1 中，我们观察到所有 MPO 对齐模型的表现都优于 PPO 基准模型和基础 LLM，其中 32b\_32b 模型——使用 32B RM 和 32B MRM 训练的——表现最佳。我们注意到 72b\_iter0 模型导致了一次失败的 PPO 运行，其特点是奖励作弊，政策模型经常仅以标题形式产生响应，例如“标题：Reddit 固有缺点的神话：为其作为讨论平台的价值辩护”，但仍常常获得 4 或 5 分（满分 5 分）的高分。相比之下，使用相同 72B RM 的 MPO 对齐模型成功进行了有意义的训练，并在性能上超越了基础 LLM。这强调了 MPO 可以自动发现有效的评估规则，减少对手动提示设计的依赖，并降低失败的 PPO 训练运行的风险，从而节省时间和计算资源。

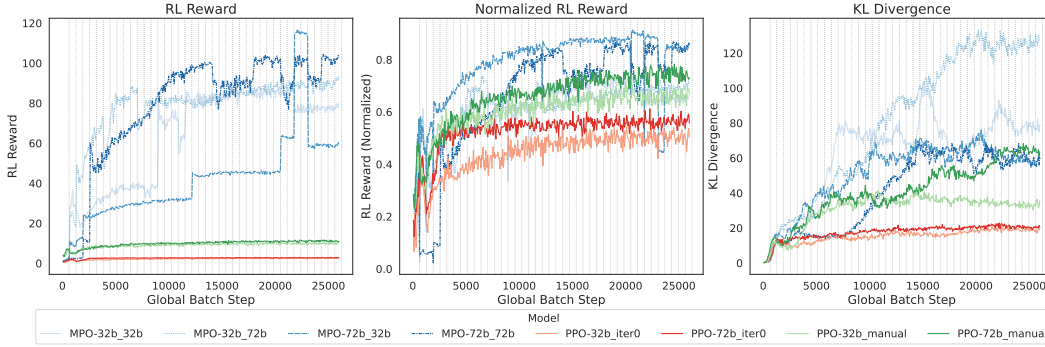


Figure 4: 八个作文政策模型的训练曲线图，每个模型都结合了不同规模的奖励模型（RM）和元奖励模型（MRM）。“RL 奖励”和“归一化 RL 奖励”图显示了奖励值如何随着全局批次步骤的演变，反映在训练的每个阶段由对应 RM 评判的生成响应的质量。通过将 RL 奖励值除以当前评分标准定义的总可得分数来获得归一化图，从而在不断发展的评估标准中提供了更一致的奖励动态视图。Kullback-Leibler (KL) 散度量化了整个训练过程中政策漂移的程度。虚线垂直线表示 MPO 轮次，在我们的设置中，每批次大小为  $\times$  的 MPO 步骤—640 步。

我们观察到，在 MPO 使用的初始评估提示（即 iter0）中，无论 RM 大小如何，其一致产生最低的 Elo 评级。图 5 提供了在不同任务中对初始和最终评估提示的并列比较。

有趣的是，当 RM 和 MRM 具有相同的模型大小时，会获得最佳结果，例如，32B + 32B 组合优于不对称组合。一个合理的解释是，相同大小的模型会发展出类似的词汇使用和推理风格，使得 MRM 的评估细化对 RM 来说更加清晰；当容量差异较大时，不匹配的词汇或推理深度可能会阻碍这种“通信渠道”。验证这一假设并系统性地探索其他组合（例如，容量匹配但架构混合的 RM 和 MRM）仍然是未来研究的前景方向。

Table 1: 针对作文任务的 Elo 评分，通过 7 个 LLM 的 10,000 次成对比较进行评估。每个评分都伴随着标准差，该标准差是通过 5 次独立实验计算的，用  $s = \text{value}$  表示。

RM \ MRM	MPO w/ MRM (ours)		PPO iter0	Base LLM
	32B	72B		
32B	1196 <sub>s=20</sub>	1184 <sub>s=18</sub>	966 <sub>s=17</sub>	836 <sub>s=17</sub>
72B	1058 <sub>s=18</sub>	1132 <sub>s=17</sub>	629 <sub>s=15</sub>	

正如在第 1.1 节中讨论的那样，评估书面作文需要一种更注重广度的评估性思维形式——这种思维形式由一套广泛的、通用的评分标准指导，可以应用于不同的提示和写作风格中。在这个实验中，我们将 MPO 框架应用于在评估性思维的深度和广度维度上要求不同的任务，研究所提出的方法的益处是否在这一光谱上持续存在。如图 2 所示，我们在三个额外的任务中评估 MPO：总结、伦理推理和数学推理。

对于这三个任务，我们比较了三个模型的性能：Base LLM，即在任何对齐前的原始 Qwen2-1.5B-Instruct 模型；32b\_iter0，即使用初始评估提示并以 32B Qwen 模型作为 RM 的普通 PPO 对齐模型；以及 32b\_32b，即同时以 32B Qwen 模型作为 RM 和 MRM 的 MPO 对齐模型。

对于这三个任务，我们比较了三个模型的性能：Base LLM，即在任何对齐前的原始 Qwen2-1.5B-Instruct 模型；32b\_iter0，即使用初始评估提示并以 32B Qwen 模型作为 RM 的普通 PPO 对齐模型；以及 32b\_32b，即同时以 32B Qwen 模型作为 RM 和 MRM 的 MPO 对齐模型。

### 2.1.1 摘要任务

对于摘要生成，我们在 BillSum 基准 [Kornilova and Eidelman, 2019] 上训练模型一个纪元，每 20 个批处理步骤进行一次 MPO 优化，并生成 4.5K 个摘要用于评估。性能通过与人工撰写的参照文比较的 ROUGE 分数和从 5,000 个 GPT-4o 判断的成对比较中获得的 Elo 评分进行评估，结果如表 2 所示。

MPO 对齐的 32b\_32b 模型在所有评估指标上，包括 Elo 评分，表现优于其他两个模型。图 5 展示了用于总结任务的评估标准的摘录，显示了 MPO 如何扩展标准以纳入特别与评估立法法案摘要相关的标准。

### 2.1.2 伦理推理任务

<b>Essay Writing: Initial Prompt</b> You will act as an English instructor and evaluate the quality of an essay or story written by a student in response to given instructions. Your evaluation should focus on the “discourse” aspect of the text. Output your score as an integer from 0 (worst quality) to 5 (best quality). Surround your integer score with <score> and </score>. Example: <score>2</score>	<b>Summarization: Initial Prompt</b> Your evaluation should focus on the faithfulness and conciseness of the summary. Output your score as an integer from 0 (worst quality) to 5 (best quality).
<b>Essay Writing: Last Prompt</b> ### Content and Argumentation (30 points) - **30 points**: The essay presents a clear, well-supported argument that effectively addresses the claim. It includes a variety of strong, relevant evidence and reasoning, and addresses counterarguments comprehensively. (Example: The essay provides historical context, cites specific legal precedents, and includes expert opinions to support its argument.) ... ### Historical and Contextual Analysis (10 points) ... ### Impact Analysis (15 points) ... ### Counterarguments (10 points) ... ### Structure and Coherence (15 points) ... ### Use of Evidence (15 points) ... ### Clarity and Coherence (10 points) ... ### Language and Style (10 points) ... ### Persuasiveness (10 points) ...	<b>Summarization: Last Prompt</b> ### Faithfulness to the Bill (30 points) - **30 points**: The summary accurately reflects all key points and details of the bill, maintaining the correct context and intent. It includes specific details such as the short title, the structure of the bill, the purposes, the findings, the definitions of key terms, the establishment of specific programs, the authorization of appropriations, the requirement for periodic reports, the effective date, the prohibition of recounts or certification of results without counting provisional ballots, and other relevant provisions. ... ### Conciseness (10 points) ... ### Completeness (30 points) ... ### Clarity and Precision (20 points) ... ### Deductions ...
<b>Mathematical Reasoning: Initial Prompt</b> Evaluate the correctness of the [Student Response].	<b>Ethical Reasoning: Initial Prompt</b> Evaluate the ethical reasoning based on its logical coherence, depth of moral insight, and alignment with established ethical principles. Assign a score from 0 to 5, where 0 indicates the lowest quality and 5 indicates the highest quality.
<b>Mathematical Reasoning: Last Prompt</b> 1. **Consistency in Scoring and Feedback**: - Ensure that the scoring and feedback are consistent across all cases. If a student's response is partially correct, acknowledge the correct aspects and provide a nuanced score. - **Example**: In [Evaluation Case #1], the student used the concept of similar triangles, which is a valid approach. The feedback should have pointed out the specific error in their calculation and provided a more nuanced score, perhaps a 2 or 3 out of 5. 2. **Detailed Feedback**: ... 3. **Handling Omitted Responses**: ... 4. **Correctness of Mathematical Reasoning**: ... 5. **Detailed Explanation of Steps**: ... 6. **Encouragement of Full Solutions**: ... To ensure these guidelines are followed, the junior instructor should implement the following additional diagnostic checks: - **Specific Error Identification**: ... - **Step-by-Step Guidance**: ... - **Encouragement of Review**: ... - **Reminders of Formatting Guidelines**: ... ### Specific Examples and Patterns ... ### General Patterns ...	<b>Ethical Reasoning: Last Prompt</b> ### Logical Coherence (0-5) - **5**: The reasoning is logically consistent, free from contradictions, and well-supported by evidence or examples. The argument is clear, structured, and avoids repetition or disjointedness. It considers multiple perspectives and the broader ethical implications of the action. The verdict is presented within the required tags. <verdict>RIGHT</verdict> - **Example**: “The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning is clear, structured, and avoids repetition. <verdict>RIGHT</verdict>” - **4**: ... - **3**: ... - **2**: ... - **1**: ... - **0**: ... ### Depth of Moral Insight (0-5) ... ### Alignment with Established Ethical Principles (0-5) ... ### Adherence to Task Instructions (0-5) ...

Figure 5: 所有任务的 MPO 评估指标初始版本和最终版本之间的比较。

对于伦理推理，我们使用了来自 Scruples 数据集 [Lourie et al., 2020] 的轶事，该数据集包含超过 32,000 个标注有社区判断的伦理复杂现实情况。我们随机抽样 13K 个轶事用于训练，4.7K 用于测试。在单次训练周期中，每进行 10 个批次步骤时执行 MPO 步骤，从而进行 20 次评分准则细化。我们没有依赖二元真实标签，而是仅根据伦理推理的质量来设定奖励分数，以鼓励更深入的推理发展，而非优化不平衡标签分布。

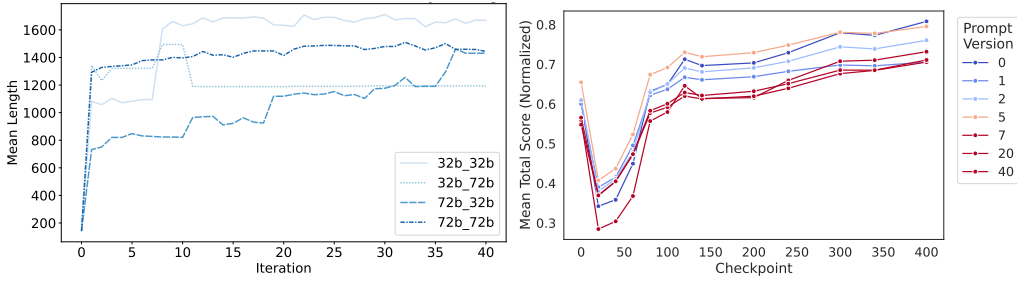
Table 3: 伦理推理 (Scruples–Anecdotes) 和数学推理 (MATH) 基准的准确性。基本 LLM 是基础的指令微调模型；32b\_iter0 是使用固定 32B RM 通过 PPO 对齐的同一模型；32b\_32b 是我们使用 32B RM 为 RM 和 MRM 进行 MPO 对齐的模型。

Model	Accuracy ( % , correct / total)	
	Scruples–Anecdotes	MATH
Base LLM	33.80 (1601/4736)	18.13 (928/5120)
PPO 32b_iter0	63.79 (3021/4736)	23.93 (1225/5120)
MPO 32b_32b (ours)	68.60 (3249/4736)	24.18 (1238/5120)



	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Elo Rating
Base LLM	41.82	20.79	27.81	27.80	849
PPO 32b_iter0	45.97	23.57	30.29	30.29	1012
MPO 32b_32b	48.00	24.96	30.97	30.98	1139

Table 2: 模型的性能通过在 BillSum 长篇文档摘要任务上的 ROUGE 分数和从 5,000 对比中得出的 Elo 评分进行评估。Base LLM 指的是基础指令微调模型；32b\_iter0 是使用 32B RM 固定初始提示通过 PPO 训练的 Base LLM；32b\_32b 是使用 MPO 训练的，其中 RM 和 MRM 都使用 32B LLMs。



(a) 在 MPO 精炼过程中的评分标准长度

(b) 通过评估量表修订的总得分

Figure 6: (a) 跨 MPO 对齐模型的作文任务评分标准项目的平均长度。(b) 经过连续的评估提示改进后，使用 32b\_72b 模型生成的 1,000 篇测试作文的平均归一化总评分标准。

基于对二进制裁决标签的准确性评估结果在表 3 中报告。我们观察到，MPO 对齐的策略模型生成的伦理推理轨迹与人工标注的裁决有更高程度的一致性。

### 2.1.3 数学推理任务

对于数学推理，我们使用 MATH 数据集 [Hendrycks et al., 2021]，该数据集包含跨七个学科的 12,500 道高中竞赛风格的问题，每道题都有详细的逐步解决方案，用于评估最终答案和推理过程。我们在 7,500 个样本上进行训练，并在 5,000 个样本上进行测试，将问题聚类到 21 个学科聚类组中，为每一个独立维护评估提示，每 30 批次步骤触发 MPO 优化。奖励模型遵循一种先计划后执行的策略，从 Saha et al. [2025] 中改编，在评分学生答案之前制定评估计划，用准则指导的评估进行评估。

根据参考答案的准确匹配得出的结果在表 3 中报告。与其他任务一样，通过细化元层指南应用 MPO 框架进一步增强了我们的“计划-后执行”奖励模型的性能。然而，与其他任务相比，改善的幅度相对较小。我们假设这是由于数学推理的高度实例特定性，其评分主要依赖于顺序逻辑是否准确地导向正确答案。相反，从采样实例构建的元层指南往往保持相对一般性。然而，主题特定的细化仍然对性能提升有所贡献。开发更细化和量身定制的元指南可能会带来额外的改善，我们将其留待未来的工作。

## 2.2 评估标准的演变

虽然早期的结果突出了 MPO 改进的评估标准与其初始提示如何不同（图 5），但本实验旨在揭示标准演变中的更整体的模式，特别是关注于作文任务。

**量表项目的平均长度** 对于论文写作任务，我们追踪每个评分项平均长度在连续的 MPO 改进过程中如何演变。值得注意的是，长度在第一次改进后急剧增加，并在接下来的 5 到 10 次迭代中继续稳步增长，然后在训练的剩余阶段趋于平稳（图 6a）。手动检查确认，大多数有意义的评分变化发生在前 5 到 10 次迭代之后，其改进变得相对较小。<sup>2</sup> 这一观察结果表明，当前固定的 MPO 计划可以通过根据训练动态进行动态调整来改善——这是未来工作中值得探索的一个途径。

<sup>2</sup>我们在附录 C 中提供了进化后的评分标准序列

**连续评分标准的平均总分** 图 6b 绘制了 32b\_72b 模型生成的 1,000 篇随机抽样测验文章的标准化平均总分，这些文章使用通过 MPO 进化的早期和晚期 RM 提示进行评估。总体模式显示，平均总分在训练过程中增加，反映了策略模型输出质量的提高。然而，较早版本的评估提示往往为样本分配更高的分数，主要是因为它们的标准较为宽松且不够细致，使得回应更容易达到评分标准。相反，后期的提示具有更详细和区分度更高的评分项目，因而分配的分较低，因为它们捕捉到更微妙的缺陷并实施更严格的评估标准。这一趋势突显了评分标准的改进不仅加强了评估，也为指导策略改进提供了更准确和更高要求的反馈。

**话语主题分布的变化。** 到目前为止，我们已经展示了 MPO 演化的评分标准不仅能带来更高的训练奖励，还能在四个下游任务中产生更高质量的生成结果。为了深入了解这些评分标准的语言结构，我们应用层次性话语解析并分析生成的话语子图——称为“话语主题”——这些话语子图捕捉了从短语到完整段落的文本单元之间的语用话语关系。这项分析基于 Kim et al. [2024] 引入的方法，他使用修辞结构理论 (RST) [Mann and Thompson, 1987] 研究大语言模型生成文本和人工生成文本中的话语模式。在我们的设置中，我们计算了三个版本评分标准的话语主题分布：初始版本，经过一次 MPO 精炼的版本，以及在训练结束时的最终版本。

图 7 展示了一个柱状图，比较了三个评估标准版本在话语模式分布上的差异，其中 x 轴表示模式类型，y 轴表示其在整体 RST 图中的归一化频率。<sup>3</sup> 边缘标签“/”表示一个超边缘关系，通常表示跨文本层次的过渡，并表明文本展现出更多层次化组织的话语结构。图表显示，随着 MPO 细化的进行，新的评估标准采用了日益层次化的话语结构，体现为背景关系的减少和对比、归因等信息含量更高的关系的相应增加。值得注意的是，明确识别信息来源或所有权的归因关系变得更加频繁——这是基于评估标准的评估中的一个重要特征，因为在这种评估中，明确归因、推理和判断对于形成连贯的评估是至关重要的。这些趋势暗示着 MPO 不仅精细化了内容标准，而且还隐式地推动了更丰富、更结构化的评估语言的发展。我们还注意到，层次化话语结构的增多是人类书写的已知特征，正如在 Kim et al. [2024, 2025] 中所报道的那样。

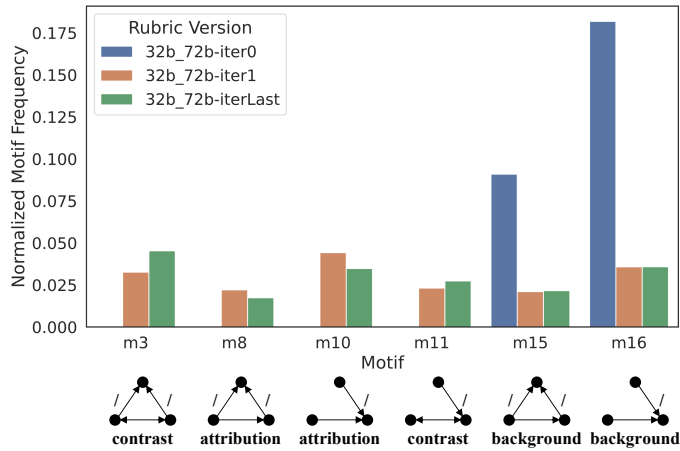


Figure 7: 比较在议论文写作任务训练的初始阶段、第一次之后以及最后一次 MPO 步骤中在评价标准提示中发现的话语主题。

### 2.3 与人工设计的最佳提示比较

对于论文写作任务，Kim et al. [2025] 通过一个迭代过程，为 RM 手工设计了一个评估提示，该过程涉及超过 60 次基础 PPO 训练。根据从相同的 Qwen2-1.5B-Instruct 模型观察到的奖励破解行为，提炼了提示，并通过修改评估标准使其更好地与人类评估分数对齐。我们将此提示视为一个神谕提示，因为它代表了通过广泛的 PPO 训练经验而提炼出的手工演变的准则。图 8 将神谕评估提示与由 MPO 框架生成的最终版本进行了比较。与神谕不同，MPO 演变的提示在一个训练周期内生成，并具有一组专门为论文评估量身定制的多样化评分项目。每个项目都配有详细描述，以支持在定义的评分范围内进行细致的分数分配。

Table 4: 通过 4 个 LLM 的 5,000 次成对比较评估的作文任务的 Elo 评分。每个评分都有一个标准差，通过 5 次独立实验计算得出，以  $s = \text{value}$  表示。

RM	MPO w/ MRM (ours)		PPO oracle
	32B	72B	
32B	1005 <sub>s=14</sub>	1001 <sub>s=14</sub>	952 <sub>s=20</sub>
72B	-	-	1042 <sub>s=20</sub>

<sup>3</sup>完整版本可在附录 D 中查看。

表 4 展示了 32b\_32b 和 32b\_72b MPO 对齐模型与使用 oracle 提示的 PPO 对齐模型的 Elo 评分，这是基于 5,000 次由 GPT-4o 判断的成对比较得出的。虽然使用 32B RM 的两个 MPO 对齐模型比使用 oracle 提示和 32B RM 训练的 PPO 对齐模型表现更好，但它们还是落后于使用 oracle 提示和 72B RM 的 PPO 模型。尽管如此，这些结果表明，MPO 框架可以自动生成评估提示，其质量接近于大量人工设计的 oracle 提示，而无需任何特定任务的手动提示设计。

Manual Hand-Crafted Prompt for Essay Writing	Last Version of Prompt Evolved by MPO
<p>You will act as an English instructor and evaluate the quality of an essay or story written by a student in response to given instructions. When grading, consider the following discourse aspects of the text.</p> <ul style="list-style-type: none"> <li>- Logical Flow and Structure (flow): Assess the logical progression of ideas and the overall organization of the text, ensuring that it is easy to follow and well-structured.</li> <li>- Hierarchical Organization (organization): Examine the organization of ideas in a hierarchical manner, from general to specific, ensuring that each section supports the main argument or narrative.</li> <li>- Balance and Emphasis (balance): Ensure that important ideas are appropriately emphasized and that there is a balance in the coverage of different points or sections of the text.</li> </ul> <p>For each aspect, you need to assign an integer score from 0 (worst quality) to 5 (best quality).</p> <p>When assigning the score, carefully consider which specific parts of the text relate to each aspect.</p> <p>Assign lower scores when:</p> <ul style="list-style-type: none"> <li>- The text is poorly structured and do not conform to the standard of an English essay or a story.</li> <li>- The text contains a lot of non-sensical words such as special tokens or programming code.</li> <li>- The text contains a lot of non-English words.</li> <li>- The text does not fully answer the writing instruction with full content, and therefore, is unfinished.</li> </ul> <p>Important: Your evaluation output should conform to the following JSON format:</p> <pre>{   "flow": int,   "organization": int,   "balance": int }</pre> <p>Write &lt;EOE&gt; after outputting the JSON result.</p>	<pre>### Content and Argumentation (30 points) - **30 points**: The essay presents a clear, well-supported argument that effectively addresses the claim. It includes a variety of strong, relevant evidence and reasoning, and addresses counterarguments comprehensively. (Example: The essay provides historical context, cites specific legal precedents, and includes expert opinions to support its argument.) - **25 points**: The essay presents a clear argument with strong evidence and reasoning but may lack a few minor details or fail to address a minor counterargument. (Example: The essay provides strong evidence and reasoning but does not address a minor counterargument.) - **20 points**: ... - **15 points**: ... - **10 points**: The essay presents a weak argument with minimal evidence and reasoning. It may contain significant logical flaws or fail to address the claim effectively. (Example: The essay provides a weak argument and minimal evidence, with significant logical flaws.) - **0 points**: The essay contains significant factual errors that misinterpret the core business or key aspects of the claim. (Example: The essay incorrectly states that CEMEX is a major sugar producer.)  ### Historical and Contextual Analysis (10 points) ...  ### Impact Analysis (15 points) ...  ### Counterarguments (10 points) ...  ### Structure and Coherence (15 points) ...  ### Use of Evidence (15 points) ...  ### Clarity and Coherence (10 points) ...  ### Language and Style (10 points) ...  ### Persuasiveness (10 points) ...</pre>

Figure 8: 手工设计的和 MPO 演化的评分标准在作文任务中的比较。

### 3 结论

这项工作介绍了元策略优化 (MPO)，一种新颖的框架，通过动态演变奖励模型使用的评价标准来增强来自人类或人工智能反馈的强化学习。基于评估思维和元认知的认知原则，MPO 使奖励模型不仅能评估策略输出，还能随着时间推移反思和改进他们的评分标准。在各种任务——包括论文写作、总结、伦理推理和数学问题解决中——MPO 始终提高了对齐效果，并优于依赖静态、手工制作提示的模型。

除了提高经验性能之外，MPO 还提供了一种新的视角，将奖励建模视为一个自适应的、自我改进的过程。我们的分析进一步揭示了进化的评判标准表现出更深层次的语言结构，暗示了更有原则的评估模式的出现。

对于未来的工作，有几个有前景的方向：基于训练动态调整 MPO 频率，扩展到更细致的评分标准专业化领域，探索多轮对话和交互设置，将 MPO 与超越 PPO 的先进优化算法结合。最后，将 MPO 扩展到支持多智能体对齐或长时间跨度任务可能会为更具普遍性和认知对齐的学习系统开辟新的道路。

致谢 本研究主要由 Grammarly 慷慨提供的研究资助资助。我们也感谢 Minnesota NLP 小组成员对我们初稿的宝贵反馈和建设性意见。

## References

- A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL <https://aclanthology.org/2024.acl-long.662>.
- D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety, 2016a.
- D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety, 2016b. URL <https://arxiv.org/abs/1606.06565>.
- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- J. Buckley, T. Archibald, M. Hargraves, and W. M. Trochim. Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation*, 36(3): 375–388, 2015.
- L. Chen, C. Zhu, D. Soselia, J. Chen, T. Zhou, T. Goldstein, H. Huang, M. Shoenybi, and B. Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf, 2024. URL <https://arxiv.org/abs/2402.07319>.
- W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL <https://arxiv.org/abs/2403.04132>.
- T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=dcjtMYkpXx>.
- E. De Bono. *The Use of Lateral Thinking*. Pelican books. Penguin Books, 1971. ISBN 9780140214468. URL <https://books.google.com/books?id=2Fd-AAAAMAAJ>.
- DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.



- C. Denison, M. MacDiarmid, F. Barez, D. Duvenaud, S. Kravec, S. Marks, N. Schiefer, R. Soklaski, A. Tamkin, J. Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.
- A. Efklides. Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational research review*, 1(1):3–14, 2006.
- J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. N. D’Amour, K. D. Dvijotham, A. Fisch, K. A. Heller, S. R. Pfohl, D. Ramachandran, P. Shaw, and J. Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=5u1GpUkKtG>.
- T. Everitt, M. Hutter, R. Kumar, and V. Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467, 2021.
- J. H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.
- L. Fluri, L. Lang, A. Abate, P. Forré, D. Krueger, and J. Skalse. The perils of optimizing learned reward functions: Low training error does not guarantee low regret. *arXiv preprint arXiv:2406.15753*, 2024.
- J. Fu, X. Zhao, C. Yao, H. Wang, Q. Han, and Y. Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.
- L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/gao23h.html>.
- D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL [https://datasets-benchmarks-proceedings.neurips.cc/paper\\_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf).
- Z. M. Kim, K. Lee, P. Zhu, V. Raheja, and D. Kang. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.298. URL <https://aclanthology.org/2024.acl-long.298/>.
- Z. M. Kim, A. Ramachandran, F. Tavazoei, J.-K. Kim, O. Rokhlenko, and D. Kang. Align to structure: Aligning large language models with structural information, 2025. URL <https://arxiv.org/abs/2504.03622>.
- A. Kornilova and V. Eidelman. BillSum: A corpus for automatic summarization of US legislation. In L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5406. URL <https://aclanthology.org/D19-5406/>.
- V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 3, 2020.
- L. L. D. Langosco, J. Koch, L. D. Sharkey, J. Pfau, and D. Krueger. Goal misgeneralization in deep reinforcement learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12004–12019. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/langosco22a.html>.

- H. Le, Y. Wang, A. D. Gotmare, S. Savarese, and S. C. H. Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.
- H. Lee, C. Park, D. Abel, and M. Jin. A hypothesis on black swan in unchanging environments. *arXiv preprint arXiv:2407.18422*, 2024.
- C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013/>.
- T. Liu, W. Xiong, J. Ren, L. Chen, J. Wu, R. Joshi, Y. Gao, J. Shen, Z. Qin, T. Yu, D. Sohn, A. Makarova, J. Liu, Y. Liu, B. Piot, A. Ittycheriah, A. Kumar, and M. Saleh. Rrm: Robust reward model training mitigates reward hacking. *ArXiv*, abs/2409.13156, 2024. URL <https://api.semanticscholar.org/CorpusID:272770255>.
- C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.
- N. Lourie, R. L. Bras, and Y. Choi. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv e-prints*, 2020.
- W. C. Mann and S. A. Thompson. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles, 1987.
- C. B. McCormick. Metacognition and learning. *Handbook of psychology*, pages 79–102, 2003.
- J. Metcalfe and N. Kornell. A region of proximal learning model of study time allocation. *Journal of Memory and Language*, 52(4):463–477, 2005. ISSN 0749-596X. doi: <https://doi.org/10.1016/j.jml.2004.12.001>. URL <https://www.sciencedirect.com/science/article/pii/S0749596X04001330>. Special Issue on Metamemory.
- Y. Miao, S. Zhang, L. Ding, R. Bao, L. Zhang, and D. Tao. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=3XnBVK9sD6>.
- Y. Miao, S. Zhang, L. Ding, Y. Zhang, L. Zhang, and D. Tao. The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking, 2025. URL <https://arxiv.org/abs/2501.19358>.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- A. Pan, K. Bhatia, and J. Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=JYtwGwIL7ye>.
- R. Y. Pang, V. Padmakumar, T. Sellam, A. Parikh, and H. He. Reward gaming in conditional text generation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4746–4763, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.262. URL <https://aclanthology.org/2023.acl-long.262/>.
- C. Park, M. Liu, D. Kong, K. Zhang, and A. Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.
- C. Park, S. Han, X. Guo, A. Ozdaglar, K. Zhang, and J.-K. Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. *arXiv preprint arXiv:2502.18439*, 2025.

- E. Perez, S. Ringer, K. Lukosiute, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. El Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. Discovering language model behaviors with model-written evaluations. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL <https://aclanthology.org/2023.findings-acl.847/>.
- A. Rame, N. Vieillard, L. Hussenot, R. Dadashi-Tazehoz, G. Cideron, O. Bachem, and J. Ferret. WARM: On the benefits of weight averaged reward models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42048–42073. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/rame24a.html>.
- S. Saha, X. Li, M. Ghazvininejad, J. Weston, and T. Wang. Learning to plan & reason for evaluation with thinking-llm-as-a-judge, 2025. URL <https://arxiv.org/abs/2501.18099>.
- K. Saito, A. Wachi, K. Wataoka, and Y. Akimoto. Verbosity bias in preference labeling by large language models, 2023. URL <https://arxiv.org/abs/2310.10076>.
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. DURMUS, Z. Hatfield-Dodds, S. R. Johnston, S. M. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=tvhaxkMKAn>.
- P. Singhal, T. Goyal, J. Xu, and G. Durrett. A long way to go: Investigating length correlations in RLHF, 2024. URL <https://openreview.net/forum?id=sNtDKdcI1f>.
- J. M. V. Skalse, N. H. R. Howe, D. Krashennikov, and D. Krueger. Defining and characterizing reward gaming. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=yb3H0X031X2>.
- K. Tian, E. Mitchell, H. Yao, C. D. Manning, and C. Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2024.
- M. V. Veenman, B. H. Van Hout-Wolters, and P. Afflerbach. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and learning*, 1:3–14, 2006.
- L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.
- J. Wen, R. Zhong, A. Khan, E. Perez, J. Steinhardt, M. Huang, S. R. Bowman, H. He, and S. Feng. Language models learn to mislead humans via RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=xJljiPE6dg>.
- T. Wu, W. Yuan, O. Golovneva, J. Xu, Y. Tian, J. Jiao, J. Weston, and S. Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*, 2024.

- C. Zhang, C. Tang, D. Chong, K. Shi, G. Tang, F. Jiang, and H. Li. TS-align: A teacher-student collaborative framework for scalable iterative finetuning of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8926–8946, Miami, Florida, USA, Nov. 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.521. URL <https://aclanthology.org/2024.findings-emnlp.521/>.
- S. Zhang, Z. Chen, S. Chen, Y. Shen, Z. Sun, and C. Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble, 2024b. URL <https://arxiv.org/abs/2401.16635>.
- L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett, and Y. Sheng. Sglang: Efficient execution of structured language model programs, 2024. URL <https://arxiv.org/abs/2312.07104>.
- D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.



## A 相关工作

### A.1 大型语言模型中的奖励黑客

强化学习被广泛应用于大型语言模型的后期训练中，增强了诸如真实性 [Tian et al., 2024]、代码生成 [Le et al., 2022]、推理 [DeepSeek-AI et al., 2025] 和多智能体决策制定 [Park et al., 2025] 等领域。将强化学习融入大型语言模型训练的主要策略是来自人类反馈的强化学习 (RLHF) [Ziegler et al., 2019, Ouyang et al., 2022, Bai et al., 2022, Ahmadian et al., 2024, Park et al., 2024]。

奖励攻击 [Skalse et al., 2022] 是指在强化学习中，当智能体利用奖励函数中的漏洞、模糊性或缺乏明确性（正如在古德哈特定律中所指出的那样）以意想不到的方式获得高奖励时观察到的现象，通常表现出一致但未预料到的行为 [Amodei et al., 2016a]。这导致智能体与人类预期的行为不一致，但却获得了高奖励。这成为 RLHF 和 RLAIIF 中的一个关键挑战 [Krakovna et al., 2020, Pan et al., 2022, Gao et al., 2023, Fluri et al., 2024, Lee et al., 2024]。

众多研究已强调了对齐的 LLMs 中的奖励的有害影响 [Pang et al., 2023]。各种不良副作用，例如阿谀谄媚 [Perez et al., 2023, Sharma et al., 2024, Denison et al., 2024]、冗长 [Saito et al., 2023, Singhal et al., 2024] 和欺骗 [Wen et al., 2025] 等。为了应对这些挑战，近期工作提出了众多奖励建模和缓解策略，例如奖励集成 [Eisenstein et al., 2024, Rame et al., 2024, Coste et al., 2024, Zhang et al., 2024b]，以及设计 RL 正则化 [Miao et al., 2025, Chen et al., 2024, Liu et al., 2024, Miao et al., 2024, Fu et al., 2025]。虽然这些策略显示出不同程度的成功，奖励过拟合、不当规范化和错误泛化等挑战仍然构成实际中基于奖励模型的对齐的重大障碍。

我们认为，有效解决奖励作弊问题需要具备广泛的上下文推理能力的对齐机制，因为奖励作弊通常源于不明确的奖励信号、不断变化的策略行为以及训练过程中探索与利用动态变化之间的相互作用。

随着大型语言模型 (LLMs) 的扩展，尽管在对齐技术方面取得了进展，它们在处理复杂且不断变化的失败模式时仍未能达到预期。这导致人们对通过迭代方法优化偏好以对齐 LLMs 产生了越来越大的兴趣。这些方法通过在连续的训练轮次中加入反馈来改进单次偏好调整，使模型能够根据变化的输出细化其行为，但仍然受到数据集的限制：依赖于明确的偏好比较或固定的提示模板，这些模板依赖于奖励提示或训练分布的初始设计，并且在训练过程中不会随着策略的发展而适应。由于奖励模型在训练过程中不演变，它们在策略变化时也对奖励作弊的鲁棒性有限。同样，通过监督微调 (SFT) 从奖励模型进行知识蒸馏的基于此的方法将奖励判断编码为静态的训练目标，这可能不再反映模型改进后的最佳行为，进一步增强了奖励估计中的非自适应偏差。文献中提出的自我奖励的语言模型同样受到启发，通过将 LLM-as-a-Judge 设定作为提示在训练期间生成奖励信号。然而，他们的方法依赖于迭代应用直接策略优化 (DPO)，然后为每个 DPO 轮次生成一个新的数据集——这一过程在计算上繁重且资源密集。相比之下，我们的 MPO 框架通过元奖励模型引入了轻量化的、基于提示的奖励细化，能在不需要反复数据集再生成或完全模型再训练的情况下实现持续对齐。

我们的工作与 Wu et al. [2024] 密切相关，他们首次在自我奖励的流程中引入了 LLM-作为-元法官的概念，使得单一模型能够评估和改进自身的判断。他们的方法训练一个 LLM 作为演员、法官和元法官，使用固定的 5 分评分标准来生成和评分响应，然后通过 DPO 迭代地改进这两种角色。虽然这反映了 MPO 的自我改进哲学，并专注于减轻奖励作弊，这些方法还是有所不同：Meta-Judge 通过偏好优化来更新模型权重，而 MPO 引入一个单独的元奖励模型，该模型重新编写评估标准本身——根据新出现的行为如奖励利用来适应标准，而不仅仅是模型。

同样，TS-Align [Zhang et al., 2024a] 也追求通过减少人工监督来实现可扩展对齐，与 MPO 的目标一致，但采用了不同的路径。它使用教师—学生框架，其中强大的教师 RM 通过较小的学生 RM 过滤后重新排名偏好对，然后进行基于 DPO 的微调。然而，奖励提示在整个过程中保持不变。相比之下，MPO 采用单一 RM (由元 RM 指导)，并在固定间隔动态演变评估标准，允许奖励函数自身适应政策漂移和训练阶段的动态。

据我们所知，MPO 是第一个通过在 PPO 而不是 DPO 下对元级评分标准进行改进来提高 LLM 对齐的方案——提供了一种轻量化、基于提示的替代方案，既减少了计算成本，又能持续适应不断变化的训练环境。

## B 元奖励模型的提示

### Meta Analysis

You are a senior instructor tasked with evaluating a junior instructor's scoring of a student's generation based on a specific task and prompt instruction. Your objective is to conduct a meta-level analysis of the junior instructor's evaluation approach, guiding them in refining their scoring criteria to ensure accurate, nuanced differentiation between high-quality and subpar generations. Emphasize strategies for assigning lower scores to undesirable responses and higher scores to responses that adhere closely to the overall objectives of the task.

The information provided includes:

Task Description:  
task\_description

Student's Prompt Instructions:  
student\_prompt

Student's Generation:  
student\_generation

Junior Instructor's Scoring Criteria:  
junior\_prompt

Junior Instructor's Assigned Score:  
junior\_score

Your task:

Critically evaluate the junior instructor's score and justification in relation to the student's response, task requirements, and prompt instructions.

1. Accuracy of Scoring
  - Determine whether the student's response is receiving an inflated score despite not fully meeting the task objectives in terms of quality and content.
  - Identify any elements where the response deviates from task expectations, such as misinterpretation, lack of depth, or overemphasis on irrelevant aspects.
2. Evaluation of Scoring Criteria
  - Assess whether the junior instructor's criteria align with the task's overarching purpose. Are critical aspects overlooked, or do the criteria require further breakdown for clarity?
  - Examine whether the distribution of points is logical and correctly sums to the total score. Flag any inconsistencies and suggest necessary adjustments.
3. Constructive Feedback for Refinement
  - Provide actionable recommendations to enhance the scoring framework, ensuring it is comprehensive and consistently applied.
  - Emphasize the need for strict penalization in cases of severe errors to maintain evaluation rigor.

Present the analysis concisely within max\_words words. Conclude the response with: "<EOE>".

Your Analysis:

### Meta Refinement

Based on the meta-level analysis, refine the junior instructor's scoring criteria by designing an explicit rubric-based framework with separate section items for awarding points and deducting points. This rubric must assign specific point values for meeting given criteria, with clear deductions for any shortcomings.

Fill in any gaps in the existing criteria to cover all relevant aspects of the task. Provide a concrete example illustrating how the rubric would apply to a typical student response. Adjust the total score to match the rubric items, ensuring the sum of all criteria equals the final total.

Use the following structure:

```
<rubric>
<item>
Score Category Name
- X1: (Description of the criterion for achieving this score X1, followed by an example.)
- X2: (Description of the criterion for achieving this score X2, followed by an example.)
...
</item>
...
</rubric>
```

Your generation should be no more than max\_words words. End with "<EOE>".

Important: You must follow the <rubric> and <item> formatting as shown above.

Junior Instructor's Scoring Criteria (refined):

### Meta Merging

Combine and refine the multiple sets of Junior Instructor's Scoring Criteria into a single, cohesive set that provides comprehensive guidelines for assessment.

Here are multiple sets of Junior Instructor's Scoring Criteria, delimited by "===":

```
" "  
multiple_sets  
" "
```

Combine the above concisely without repetition. The combined criteria should be no more than max\_words words. Make sure that the points across criteria add up correctly to the total score.

Use the following structure:

```
<rubric>  
<item>  
Score Category Name  
- X1: (Description of the criterion for achieving this score X1, followed by an example.)  
- X2: (Description of the criterion for achieving this score X2, followed by an example.)  
...  
</item>  
...  
</rubric>
```

Your generation should be no more than max\_words words. End with "<EOE>".  
Important: You must follow the <rubric> and <item> formatting as shown above.

Junior Instructor's Scoring Criteria (combined):

## C 评价量表的演变

我们展示了一些在 32b\_32b 模型训练过程中产生的精细评估提示的例子。完整的例子可以在 <https://github.com/minnesotanlp/mpo> 找到。

### C.1 作文写作任务

#### MPO Iteration 1 – Essay Writing

```
<item>  
Task Alignment and Argument Clarity  
- 5: The essay clearly and consistently addresses the specific claim made in the prompt and presents a well-defined, coherent argument supported by relevant examples and evidence. (Example: "The essay argues that the U.S. is not responsible for social backwardness in Iran because it has not directly influenced Iran's social structures and policies, supported by specific examples.")  
- 4: The essay mostly addresses the specific claim and presents a coherent argument, but there are minor inconsistencies or lack of depth. (Example: "The essay argues that the U.S. is not primarily responsible for social backwardness in Iran, but it does not fully explain how other factors might influence this.")  
- 3: The essay partially addresses the specific claim and presents an argument, but there are significant inconsistencies or lack of depth. (Example: "The essay argues that the U.S. has not directly caused social backwardness, but it does not fully explain how other factors might influence this.")  
- 2: The essay addresses the specific claim but the argument is weak and poorly supported. (Example: "The essay argues that the U.S. is not responsible, but it lacks supporting evidence or explanation.")  
- 1: The essay fails to address the specific claim or presents an argument that is contradictory or irrelevant. (Example: "The essay argues that the U.S. has historically been responsible for promoting social backwardness in Iran.")  
- 0: The essay does not address the specific claim or is completely off-topic. (Example: "The essay discusses the history of Iran without mentioning the U.S. or social backwardness.")  
</item>  
<item>  
Evidence and Reasoning  
- 5: The essay provides strong, relevant evidence and reasoning to support the argument, with specific examples and data. (Example: "The essay cites specific examples of U.S. policies that did not directly influence Iran's social structures and policies.")  
- 4: The essay provides mostly relevant evidence and reasoning, but some points are weak or not fully supported. (Example: "The essay cites examples of U.S. policies but does not fully explain how they do not contribute to social backwardness.")  
- 3: The essay provides some relevant evidence and reasoning, but there are significant gaps or weak points. (Example: "The essay mentions U.S. policies but does not provide specific examples or
```

explanations.")

- 2: The essay provides weak or irrelevant evidence and reasoning. (Example: "The essay mentions U.S. policies but does not explain how they are unrelated to social backwardness.")
- 1: The essay does not provide any evidence or reasoning to support the argument. (Example: "The essay makes claims without providing any supporting evidence.")
- 0: The essay provides evidence and reasoning that contradicts the argument. (Example: "The essay provides examples of U.S. policies that contributed to social backwardness.")

</item>

<item>

Counterarguments and Refutation

- 5: The essay addresses potential counterarguments and provides strong refutations. (Example: "The essay acknowledges that U.S. policies might have some indirect influence but explains why this does not make the U.S. responsible for social backwardness.")
- 4: The essay addresses some counterarguments and provides mostly strong refutations. (Example: "The essay acknowledges some counterarguments but does not fully refute them.")
- 3: The essay addresses some counterarguments but provides weak or incomplete refutations. (Example: "The essay mentions counterarguments but does not fully address them.")
- 2: The essay addresses some counterarguments but does not provide any refutations. (Example: "The essay mentions counterarguments without explaining why they do not undermine the argument.")
- 1: The essay does not address any counterarguments. (Example: "The essay does not mention any counterarguments.")
- 0: The essay addresses counterarguments but fails to refute them or provides weak refutations. (Example: "The essay mentions counterarguments but does not provide any refutations or provides weak refutations.")

</item>

<item>

Structure and Coherence

- 5: The essay is well-organized and coherent, with a clear structure and logical flow. (Example: "The essay has an introduction, body paragraphs, and a conclusion that are logically connected.")
- 4: The essay has a mostly clear structure, but some points may be slightly disconnected or lack smooth transitions. (Example: "The essay has a clear structure but some points are not well-connected or lack transitions.")
- 3: The essay has a somewhat clear structure, but the points are not well-connected, and the transitions are weak. (Example: "The essay has a structure but the points are disjointed and the transitions are weak.")
- 2: The essay lacks a clear structure, and the points are disjointed and disconnected. (Example: "The essay has no clear structure and the points are not well-connected.")
- 1: The essay is poorly structured, with no clear introduction, body, or conclusion. (Example: "The essay lacks a clear structure and the points are entirely disjointed.")
- 0: The essay is completely disorganized and lacks any structure. (Example: "The essay is a collection of unrelated points with no structure.")

</item> <item>

Depth of Analysis

- 5: The essay demonstrates a deep and nuanced analysis of the issue, addressing complexities and providing specific examples and data. (Example: "The essay discusses the historical roots of the conflict, the cultural identity of Palestinians, and the potential repercussions of the proposed solution.")
- 4: The essay provides some depth and nuance in the analysis but lacks specific examples or data. (Example: "The essay mentions that the conflict is about identity and survival but does not provide substantial evidence or reasoning.")
- 3: The essay provides minimal depth and nuance in the analysis, lacking specific examples or data. (Example: "The essay mentions the conflict is about identity but does not provide substantial historical or cultural context.")
- 2: The essay lacks depth and nuance in the analysis, providing vague statements without specific examples or data. (Example: "The essay is vague and does not provide any substantial argument or reasoning.")
- 1: The essay lacks any meaningful analysis or is superficial. (Example: "The essay does not provide any meaningful analysis and remains superficial in its discussion.")
- 0: The essay is completely lacking in analysis and is purely superficial. (Example: "The essay is vague and does not provide any substantial content or context.")

</item>

<item>

Language and Mechanics

- 5: The essay demonstrates clear and effective use of language with minimal grammatical errors. (Example: "The essay uses clear and precise language with no significant grammatical errors.")
- 4: The essay has some grammatical errors but is still generally clear. (Example: "The essay has a few grammatical errors but is still understandable.")
- 3: The essay has some grammatical errors that affect clarity but is still mostly comprehensible. (Example: "The essay has some errors in grammar or style but is still generally clear.")
- 2: The essay has significant grammatical errors that affect clarity and coherence. (Example: "The essay has multiple grammatical or stylistic errors that make it difficult to understand.")
- 1: The essay has significant grammatical errors that severely affect clarity and coherence. (Example: "The essay has multiple grammatical or stylistic errors that make it difficult to understand.")
- 0: The essay is completely unclear and incoherent due to significant grammatical or stylistic errors. (Example: "The essay is difficult to read due to numerous grammatical or stylistic errors.")

</item>



## MPO Iteration 40 – Essay Writing

<item>

Task Alignment and Argument Clarity (10 points)

- 10: The essay clearly and consistently addresses the specific claim made in the prompt and presents a well-defined, coherent argument supported by relevant examples and evidence. The argument is fully developed and addresses the core issue of the prompt. (Example: "The essay argues that higher education is counterproductive by providing specific examples of student debt, income gaps, and over-specialization in certain fields.")
- 8: The essay mostly addresses the specific claim and presents a coherent argument, but there are minor inconsistencies or lack of depth. The argument is somewhat developed but could benefit from more specific examples or a deeper analysis. (Example: "The essay argues that higher education is counterproductive but lacks substantial examples or a clear connection to the context.")
- 6: The essay partially addresses the specific claim and presents an argument, but there are significant inconsistencies or lack of depth. The argument is weak and lacks substantial examples or a clear connection to the core issue. (Example: "The essay argues that higher education is counterproductive but fails to address the specific aspects of student debt, income gaps, and over-specialization.")
- 4: The essay addresses the specific claim but the argument is weak and poorly supported. The argument lacks substantial examples or a clear connection to the core issue. (Example: "The essay argues that higher education is counterproductive but lacks supporting evidence or explanation.")
- 2: The essay fails to address the specific claim or presents an argument that is contradictory or irrelevant. The argument is not aligned with the prompt and does not address the core issue. (Example: "The essay argues that higher education is beneficial, which contradicts the prompt.")
- 0: The essay does not address the specific claim or is completely off-topic. The argument is not related to the prompt and does not address the core issue. (Example: "The essay discusses unrelated topics without addressing the claim.")

</item>

<item>

Evidence and Reasoning (10 points)

- 10: The essay provides strong, relevant evidence and reasoning to support the argument, with specific examples and data from current events and historical contexts. The evidence is well-connected to the core issue and supports the argument effectively. (Example: "The essay cites specific studies showing the burden of student debt, the widening income gap, and the over-specialization in certain fields, providing detailed examples and data.")
- 8: The essay provides mostly relevant evidence and reasoning, but some points are weak or not fully supported. The evidence is somewhat connected to the core issue but could benefit from more substantial examples or a deeper analysis. (Example: "The essay mentions studies but does not provide substantial evidence or explanation.")
- 6: The essay provides some relevant evidence and reasoning, but there are significant gaps or weak points. The evidence is not well-connected to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay mentions studies but does not provide specific examples or explanations.")
- 4: The essay provides weak or irrelevant evidence and reasoning. The evidence is not well-connected to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay mentions studies but does not explain how this supports the argument or provide substantial evidence.")
- 2: The essay does not provide any evidence or reasoning to support the argument. The evidence is not related to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay makes claims without providing any supporting evidence.")
- 0: The essay provides evidence and reasoning that contradict the argument. The evidence is not related to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay provides examples that support the claim that higher education is beneficial.")

</item>

<item>

Counterarguments and Refutation (10 points)

- 10: The essay addresses potential counterarguments and provides strong refutations. The counterarguments are acknowledged and effectively refuted with specific examples and reasoning. (Example: "The essay acknowledges that some argue higher education is beneficial by explaining the burden of student debt, the widening income gap, and the over-specialization in certain fields.")
- 8: The essay addresses some counterarguments and provides mostly strong refutations. The counterarguments are acknowledged but not fully refuted. (Example: "The essay acknowledges some counterarguments but does not fully refute them.")
- 6: The essay addresses some counterarguments but provides weak or incomplete refutations. The counterarguments are acknowledged but not fully addressed. (Example: "The essay mentions counterarguments but does not fully address them.")
- 4: The essay addresses some counterarguments but does not provide any refutations. The counterarguments are acknowledged but not addressed. (Example: "The essay mentions counterarguments without explaining why they do not undermine the argument.")
- 2: The essay does not address any counterarguments. The counterarguments are not acknowledged or addressed. (Example: "The essay does not mention any counterarguments.")
- 0: The essay addresses counterarguments but fails to refute them or provides weak refutations. The counterarguments are acknowledged but not effectively refuted. (Example: "The essay mentions counterarguments but does not provide any refutations or provides weak refutations.")

</item>

<item>

Structure and Coherence (10 points)

- 10: The essay is well-organized and coherent, with a clear structure and logical flow. The introduction sets up the argument, body paragraphs provide evidence and reasoning, and the conclusion summarizes the argument and provides a final thought. (Example: "The essay has an introduction

that sets up the argument, body paragraphs that provide evidence and reasoning, and a conclusion that summarizes the argument and provides a final thought.")

- 8: The essay has a mostly clear structure, but some points may be slightly disconnected or lack smooth transitions. The introduction sets up the argument, body paragraphs provide evidence and reasoning, and the conclusion summarizes the argument but lacks smooth transitions. (Example: "The essay has a clear structure but some points are not well-connected or lack transitions.")
- 6: The essay has a somewhat clear structure, but the points are not well-connected, and the transitions are weak. The introduction sets up the argument, body paragraphs provide evidence and reasoning, but the transitions are weak. (Example: "The essay has a structure but the points are disjointed and the transitions are weak.")
- 4: The essay lacks a clear structure, and the points are disjointed and disconnected. The introduction sets up the argument, but the body paragraphs are disjointed and the conclusion is unclear. (Example: "The essay has no clear structure and the points are not well-connected.")
- 2: The essay is poorly structured, with no clear introduction, body, or conclusion. The essay lacks a clear structure and the points are entirely disjointed. (Example: "The essay lacks a clear structure and the points are entirely disjointed.")
- 0: The essay is completely disorganized and lacks any structure. The essay is a collection of unrelated points with no structure. (Example: "The essay is a collection of unrelated points with no structure.")

</item>

<item>

Depth of Analysis (10 points)

- 10: The essay demonstrates a deep and nuanced analysis of the issue, addressing complexities and providing specific examples and data from current events and historical contexts. The analysis is well-connected to the core issue and supports the argument effectively. (Example: "The essay discusses the complexities of higher education, providing specific examples of student debt, income gaps, and over-specialization in certain fields, and provides detailed data from reliable sources.")
- 8: The essay provides some depth and nuance in the analysis but lacks specific examples or data. The analysis is somewhat connected to the core issue but could benefit from more substantial examples or a deeper analysis. (Example: "The essay mentions that higher education has issues but does not provide substantial evidence or reasoning.")
- 6: The essay provides minimal depth and nuance in the analysis, lacking specific examples or data. The analysis is not well-connected to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay mentions that higher education has issues but does not provide specific examples or explanations.")
- 4: The essay lacks depth and nuance in the analysis, providing vague statements without specific examples or data. The analysis is not well-connected to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay is vague and does not provide any substantial argument or reasoning.")
- 2: The essay lacks any meaningful analysis or is superficial. The analysis is not related to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay does not provide any meaningful analysis and remains superficial in its discussion.")
- 0: The essay is completely lacking in analysis and is purely superficial. The analysis is not related to the core issue and lacks substantial examples or a clear connection to the context. (Example: "The essay is vague and does not provide any substantial content or context.")

</item>

<item>

Language and Mechanics (10 points)

- 10: The essay demonstrates clear and effective use of language with minimal grammatical errors. The writing is clear, coherent, and free of significant errors that affect clarity or coherence. (Example: "The essay uses clear and precise language with no significant grammatical errors.")
- 8: The essay has some grammatical errors but is still generally clear. The writing is mostly clear and coherent, with a few minor errors that do not significantly affect clarity. (Example: "The essay has a few grammatical errors but is still understandable.")
- 6: The essay has some grammatical errors that affect clarity but is still mostly comprehensible. The writing is generally clear but has some errors that affect clarity. (Example: "The essay has some errors in grammar or style but is still generally clear.")
- 4: The essay has significant grammatical errors that affect clarity and coherence. The writing is unclear and difficult to follow due to significant errors. (Example: "The essay has multiple grammatical or stylistic errors that make it difficult to understand.")
- 2: The essay has significant grammatical errors that severely affect clarity and coherence. The writing is unclear and difficult to follow due to multiple significant errors. (Example: "The essay has multiple grammatical or stylistic errors that make it difficult to understand.")
- 0: The essay is completely unclear and incoherent due to significant grammatical or stylistic errors. The writing is difficult to read due to numerous grammatical or stylistic errors. (Example: "The essay is difficult to read due to numerous grammatical or stylistic errors.")

</item>

## C.2 伦理推理

### MPO Iteration 1 – Ethical Reasoning

<item>

Logical Coherence (0-5)

- 5: The reasoning is logically consistent, free from contradictions, and well-supported by evidence or examples. The argument is clear, structured, and avoids repetition or disjointedness.
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable

response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. Trusting someone with your body is a significant decision, and I believe it's important to prioritize safety and quality. <verdict>WRONG</verdict>"

- 4: The reasoning is mostly logical, with a few minor contradictions or gaps in reasoning. The argument is coherent but could be more clear or structured.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate and disrespectful. <verdict>WRONG</verdict>"

- 3: The reasoning is somewhat logical but contains some contradictions or gaps in reasoning. The argument is coherent but could be more clear or structured.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate. <verdict>WRONG</verdict>"

- 2: The reasoning is partially logical, with significant contradictions or gaps in reasoning. The argument is somewhat coherent but lacks clarity and structure.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience. <verdict>WRONG</verdict>"

- 1: The reasoning is illogical, with major contradictions or gaps in reasoning. The argument is incoherent and lacks clarity.

- Example: "The girlfriend's reaction is wrong because she overreacted. The act of forgetting to defrost the shrimp is a minor inconvenience. <verdict>WRONG</verdict>"

- 0: The reasoning is completely illogical, with no coherent argument or supporting evidence. The argument is incoherent and lacks clarity.

- Example: "The girlfriend's reaction is wrong. <verdict>WRONG</verdict>"

</item>

<item>

Depth of Moral Insight (0-5)

- 5: The analysis demonstrates a deep understanding of the ethical issues involved, providing nuanced insights and thoughtful consideration of multiple perspectives. The reasoning explores multiple ethical principles and their implications.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. This response reflects a lack of consideration for others and a disregard for the impact of one's actions. It also violates deontological principles, which emphasize the importance of treating others with respect and dignity. Additionally, her response fails to consider the utilitarian principle of maximizing overall happiness and well-being. <verdict>WRONG</verdict>"

- 4: The analysis demonstrates a good understanding of the ethical issues involved, with some thoughtful consideration of multiple perspectives. The reasoning explores some ethical principles and their implications.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. This response reflects a lack of consideration for others and a disregard for the impact of one's actions. It also violates deontological principles, which emphasize the importance of treating others with respect and dignity. <verdict>WRONG</verdict>"

- 3: The analysis demonstrates a basic understanding of the ethical issues involved, with some consideration of multiple perspectives. The reasoning touches on some ethical principles and their implications.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate and disrespectful. <verdict>WRONG</verdict>"

- 2: The analysis demonstrates a limited understanding of the ethical issues involved, with little consideration of multiple perspectives. The reasoning touches on some ethical principles but lacks depth.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience. <verdict>WRONG</verdict>"

- 1: The analysis demonstrates a minimal understanding of the ethical issues involved, with no consideration of multiple perspectives. The reasoning touches on some ethical principles but lacks depth.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. <verdict>WRONG</verdict>"

- 0: The analysis demonstrates no understanding of the ethical issues involved. The reasoning lacks any exploration of ethical principles or ethical theories.

- Example: "The girlfriend's reaction is wrong because she overreacted. <verdict>WRONG</verdict>"

</item>

<item>

Alignment with Established Ethical Principles (0-5)

- 5: The response references and applies relevant ethical principles or theories comprehensively and accurately. The reasoning explicitly identifies and applies specific ethical theories or principles, including how they apply to the specific situation and why they are relevant.

- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is

a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. This response reflects a lack of consideration for others and a disregard for the impact of one's actions. It also violates deontological principles, which emphasize the importance of treating others with respect and dignity. Additionally, her response fails to consider the utilitarian principle of maximizing overall happiness and well-being. <verdict>WRONG</verdict>"

- 4: The response references and applies relevant ethical principles or theories with some accuracy. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. This response reflects a lack of consideration for others and a disregard for the impact of one's actions. It also violates deontological principles, which emphasize the importance of treating others with respect and dignity. <verdict>WRONG</verdict>"
- 3: The response references some ethical principles or theories but with limited accuracy. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate and disrespectful. <verdict>WRONG</verdict>"
- 2: The response references some ethical principles or theories but with significant inaccuracies. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience. <verdict>WRONG</verdict>"
- 1: The response references relevant ethical principles or theories but with major inaccuracies. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. <verdict>WRONG</verdict>"
- 0: The response does not reference any relevant ethical principles or theories. The reasoning does not reference any ethical principles or theories.  
- Example: "The girlfriend's reaction is wrong. <verdict>WRONG</verdict>"

</item>  
<item>  
Clarity of Verdict (0-5)

- 5: The final verdict is clear, properly formatted, and well-supported by the ethical reasoning. The verdict is clearly stated within the <verdict></verdict> tags, and the reasoning supports the verdict.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. This response reflects a lack of consideration for others and a disregard for the impact of one's actions. It also violates deontological principles, which emphasize the importance of treating others with respect and dignity. Additionally, her response fails to consider the utilitarian principle of maximizing overall happiness and well-being. <verdict>WRONG</verdict>"
- 4: The final verdict is clear and properly formatted but lacks some supporting evidence. The verdict is clearly stated within the <verdict></verdict> tags, but the reasoning does not fully support the verdict.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate, disrespectful, and potentially harmful to the relationship. This response reflects a lack of consideration for others and a disregard for the impact of one's actions. It also violates deontological principles, which emphasize the importance of treating others with respect and dignity. <verdict>WRONG</verdict>"
- 3: The final verdict is somewhat clear and properly formatted but lacks supporting evidence. The verdict is clearly stated within the <verdict></verdict> tags, but the reasoning does not fully support the verdict.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience, and the girlfriend's response is disproportionate and disrespectful. <verdict>WRONG</verdict>"
- 2: The final verdict is partially clear and properly formatted but lacks significant supporting evidence. The verdict is clearly stated within the <verdict></verdict> tags, but the reasoning does not fully support the verdict.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. The act of forgetting to defrost the shrimp is a minor inconvenience. <verdict>WRONG</verdict>"
- 1: The final verdict is unclear or improperly formatted. The verdict is not clearly stated within the <verdict></verdict> tags, and the reasoning does not support the verdict.  
- Example: "The girlfriend's reaction is wrong because she overreacted. This is not a justifiable response as it lacks empathy and understanding. <verdict>WRONG</verdict>"
- 0: The final verdict is completely unclear or improperly formatted. The verdict is not clearly stated within the <verdict></verdict> tags, and the reasoning does not support the verdict.  
- Example: "The girlfriend's reaction is wrong. <verdict>WRONG</verdict>"

</item>



```

<item>
Logical Coherence (0-5)
- 5: The reasoning is logically consistent, free from contradictions, and well-supported by evidence
or examples. The argument is clear, structured, and avoids repetition or disjointedness. It
considers multiple perspectives and the broader ethical implications of the action. The verdict
is presented within the required tags. <verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning is clear, structured, and avoids
repetition. <verdict>RIGHT</verdict>"
- 4: The reasoning is mostly logical, with a few minor contradictions or gaps in reasoning.
The argument is coherent but could be more clear or structured. The reasoning considers some
perspectives but lacks a comprehensive analysis. The verdict is presented within the required tags.
<verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning is mostly coherent but could be more
clear or structured. <verdict>RIGHT</verdict>"
- 3: The reasoning is somewhat logical but contains some contradictions or gaps in reasoning.
The argument is coherent but could be more clear or structured. The reasoning touches on some
perspectives but lacks a comprehensive analysis. The verdict is presented within the required tags.
<verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning is somewhat coherent but lacks clarity.
<verdict>RIGHT</verdict>"
- 2: The reasoning is partially logical, with significant contradictions or gaps in reasoning. The
argument is somewhat coherent but lacks clarity and structure. The reasoning is disjointed and lacks
coherence. The verdict is presented within the required tags. <verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning is disjointed and lacks coherence.
<verdict>RIGHT</verdict>"
- 1: The reasoning is illogical, with major contradictions or gaps in reasoning. The argument is
incoherent and lacks clarity. The reasoning is incoherent and lacks logical structure. The verdict
is presented within the required tags. <verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning is incoherent and lacks logical
structure. <verdict>RIGHT</verdict>"
- 0: The reasoning is completely illogical, with no coherent argument or supporting evidence.
The argument is incoherent and lacks clarity. The reasoning is completely illogical, with no
coherent argument or supporting evidence. The verdict is presented within the required tags.
<verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. <verdict>RIGHT</verdict>"
</item>
<item>
Depth of Moral Insight (0-5)
- 5: The analysis demonstrates a deep understanding of the ethical issues involved, providing
nuanced insights and thoughtful consideration of multiple perspectives. The reasoning explores
multiple ethical principles and their implications, including the balance between personal needs and
others' well-being. The reasoning explores the ethical principles of personal safety, respect for
autonomy, and the impact on the relationship. The verdict is presented within the required tags.
<verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning explores the ethical principles of
personal safety, respect for autonomy, and the impact on the relationship. <verdict>RIGHT</verdict>"
- 4: The analysis demonstrates a good understanding of the ethical issues involved, with some
thoughtful consideration of multiple perspectives. The reasoning explores some ethical principles
and their implications. The reasoning considers the ethical principles of personal safety and respect
for autonomy. The verdict is presented within the required tags. <verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning considers the ethical principles of
personal safety and respect for autonomy. <verdict>RIGHT</verdict>"
- 3: The analysis demonstrates a basic understanding of the ethical issues involved, with some
consideration of multiple perspectives. The reasoning touches on some ethical principles and their
implications. The reasoning touches on the ethical principles of personal safety and respect for
autonomy. The verdict is presented within the required tags. <verdict>RIGHT</verdict>
- Example: "The action of not using the GPS app is justified. It respects personal autonomy and
privacy, which are fundamental human rights. The decision to not use the app is a reasonable response
to concerns about personal privacy and autonomy. The reasoning touches on the ethical principles of
personal safety and respect for autonomy. <verdict>RIGHT</verdict>"
- 2: The analysis demonstrates a limited understanding of the ethical issues involved, with little
consideration of multiple perspectives. The reasoning touches on some ethical principles but lacks
depth. The reasoning touches on the ethical principles of personal safety and respect for autonomy
but lacks depth. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

```

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning touches on the ethical principles of personal safety and respect for autonomy but lacks depth. <verdict>RIGHT</verdict>"

- 1: The analysis demonstrates a minimal understanding of the ethical issues involved, with no consideration of multiple perspectives. The reasoning touches on some ethical principles but lacks depth. The reasoning touches on the ethical principles of personal safety and respect for autonomy but lacks depth. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning touches on the ethical principles of personal safety and respect for autonomy but lacks depth. <verdict>RIGHT</verdict>"

- 0: The analysis demonstrates no understanding of the ethical issues involved. The reasoning lacks any exploration of ethical principles or ethical theories. The reasoning lacks any exploration of ethical principles or ethical theories. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. <verdict>RIGHT</verdict>"

</item>

<item>

Alignment with Established Ethical Principles (0-5)

- 5: The response references and applies relevant ethical principles or theories comprehensively and accurately. The reasoning explicitly identifies and applies specific ethical principles or theories, including how they apply to the specific situation and why they are relevant. The reasoning explicitly identifies and applies specific ethical principles such as personal safety, respect for autonomy, and the impact on the relationship. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning explicitly identifies and applies specific ethical principles such as personal safety, respect for autonomy, and the impact on the relationship. <verdict>RIGHT</verdict>"

- 4: The response references and applies relevant ethical principles or theories with some accuracy. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. <verdict>RIGHT</verdict>"

- 3: The response references some ethical principles or theories but with limited accuracy. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. <verdict>RIGHT</verdict>"

- 2: The response references some ethical principles or theories but with significant inaccuracies. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. <verdict>RIGHT</verdict>"

- 1: The response references relevant ethical principles or theories but with major inaccuracies. The reasoning references some ethical principles or theories but does not provide a clear explanation of how they apply to the situation. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. It respects personal autonomy and privacy, which are fundamental human rights. The decision to not use the app is a reasonable response to concerns about personal privacy and autonomy. The reasoning references some ethical principles such as personal safety and respect for autonomy but does not provide a clear explanation of how they apply to the situation. <verdict>RIGHT</verdict>"

- 0: The response does not reference any ethical principles or theories. The reasoning lacks any exploration of ethical principles or ethical theories. The reasoning lacks any exploration of ethical principles or ethical theories. The verdict is presented within the required tags. <verdict>RIGHT</verdict>

- Example: "The action of not using the GPS app is justified. <verdict>RIGHT</verdict>"

</item>

## C.3 训练细节

### C.3.1 总结

**设置。** 为了进行总结，我们使用了 BillSum 基准 [Kornilova and Eidelman, 2019]，这是一个专为总结美国国会和加利福尼亚州立法而设计的语料库。该数据集包含超过 22,000 篇中等长度的国会法案及其人类撰写的摘要，并附加了一个加利福尼亚的测试集，以支持跨领域的泛化。其技术性和层级结构的性质呈现出独特的挑战，使其适合于特定领域和通用的总结研究。训练集包含大约 18.9K 个样本，而合并的测试集包括 4.5K 个样本。我们提供完整的法案文本作为输入，并要求撰写大约 400 字的摘要。训练在数据集上进行一个周期，每 20 个批次步骤执行一次 MPO 步骤，整个训练过程中总共进行 29 次准则优化。

总共有 4,500 个法案摘要由三个模型在测试集上生成。这些生成的摘要通过计算 ROUGE [Lin, 2004] 分数与其对应的人类撰写的参考进行评估。此外，基于所生成的摘要的 5,000 次成对比较，我们为三个模型计算了 Elo 评分，其中 GPT-4o 作为裁判。结果如表 2 所示。

### C.3.2 伦理推理

**设置。** 对于伦理推理，我们利用 [Lourie et al., 2020] 数据集中的 Scruples 的例子，该数据集包含来自 Reddit 社区的超过 32,000 个真实生活故事，其中用户描述了他们经历或考虑的伦理相关情境。每个故事都包含一个标题、详细的故事，以及社区判断的分布，指出谁被认为有错，例如作者、另一方、所有人或无一人。这些叙述通常具有道德上的模糊性，并通过众包的伦理评估进行标注，使其非常适合建模社区规范并捕捉伦理推理的多样性。

该数据集包括 27.8K 训练轶事，我们从中随机抽样了 13K 进行实验，此外还有 4.7K 轶事用于测试。训练在一个单一的迭代中进行，每 10 个批处理步骤执行一次 MPO 步骤，总计进行 20 次评分细化。尽管每个轶事都包括来自人工标注者的二元判断结果，但我们在 RM 或 MRM 中都没有使用这些真实标签。相反，奖励分数是仅根据响应中表现出的伦理推理质量来分配的。这个决定是基于两个原因：(1) 标签分布不平衡，和 (2) 我们的目的是通过生成更强的推理痕迹来鼓励策略模型改善，而不是仅仅优化标签预测。

### C.3.3 数学推理

**设置。** 对于数学推理，我们使用了 MATH 数据集 [Hendrycks et al., 2021]，该数据集由 12,500 道高中竞赛风格的数学题组成，这些题目来源于类似 AMC 10、AMC 12 和 AIME 的比赛。每个题目都附有用 LaTeX 编写的详细的逐步解决方案，既支持最终答案的评估，也支持学习解决问题的过程。该数据集涵盖七个学科——包括代数、几何和数论，并按 1 到 5 的难度级别进行标注，提供对广泛的数学推理任务的细粒度评估。

该数据集由 7.5K 的训练样本和 5K 的测试样本组成。因为数学推理需要相当深的评估深度，我们在更细的粒度上应用 MPO。具体来说，我们将每个数学科目内的问题基于语义嵌入分为三组，形成 21 ( $7 \times 3$ ) 个不同类别。在 MPO 训练过程中，我们为每个类别维护一个单独的评估提示，并根据样本的科目和簇索引来触发优化步骤。奖励模型遵循“规划然后执行”的策略：首先根据问题、参考解答和元级别的指南制定一个评估计划，然后应用该计划来评估学生的回答。该方法基于 Saha et al. [2025] 提出的原用于成对判断的方法，我们将其调整为使用（元级别）标准引导的评估进行绝对评分。训练在一个单周期内进行，每 30 批次步骤执行一次 MPO 步骤——使用较长的间隔以确保在每次优化之前收集到具有不同质量水平的足够多的回答。

## D 评价提示的话语主题分布

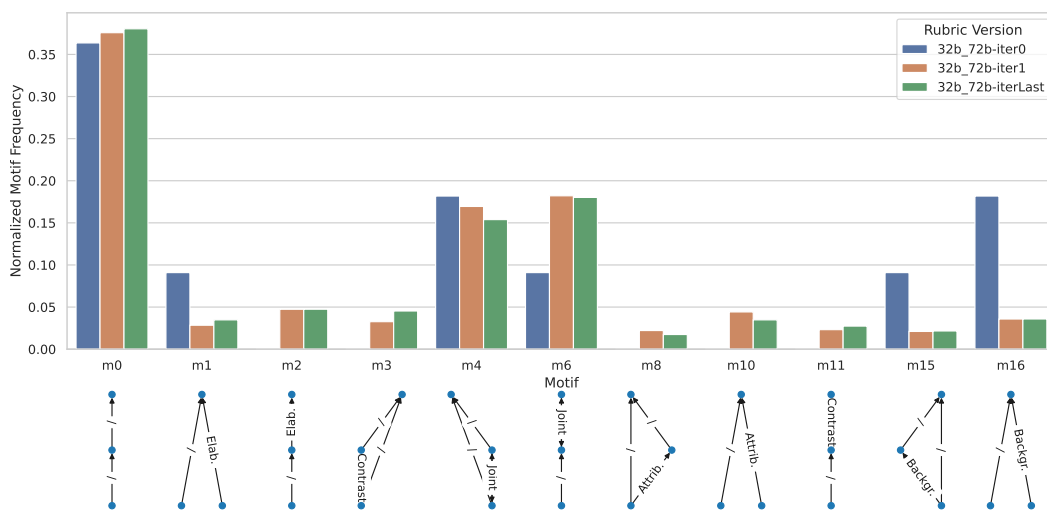


Figure 9: 在初始阶段、第一阶段之后和最后一个 MPO 步骤的训练中，用于作文任务的评分标准提示中发现的语篇模式的比较。