

动态和普适的过程奖励建模

Zhangyue Yin[◇] Qiushi Sun[♡] Zhiyuan Zeng[◇]
Qinyuan Cheng[◇] Xipeng Qiu^{◇†} Xuanjing Huang^{◇†}

[◇]College of Computer Science and Artificial Intelligence, Fudan University

[♡]The University of Hong Kong

{ yinzy21, cengzy23, chengqy21 } @m.fudan.edu.cn

qiushisun@connect.hku.hk { xpqiu, xjhuang } @fudan.edu.cn

Abstract

过程奖励模型 (PRMs) 在复杂场景中通过提供密集的奖励信号来指导大型语言模型 (LLMs) 方面极为重要。然而，现有的 PRMs 主要依赖于启发式方法，这些方法在跨域泛化方面表现不佳。虽然已经提出了由 LLM 进行裁决以提供广义奖励的方法，但目前的研究主要集中于反馈结果，而忽略了文本中蕴含的有意义指导。此外，静态且粗粒度的评估标准难以适应复杂过程的监督。为了解决这些挑战，我们提出了动态和可泛化的过程奖励建模 (DG-PRM)，它具有一个奖励树以捕获和存储细粒度、多维度的奖励标准。DG-PRM 动态选择奖励信号进行分步骤奖励评分。为了处理多方面的奖励信号，我们首次采用帕累托优势估计来识别区分性正负对。实验结果表明，DG-PRM 在现有基准测试中表现优异，显著提升了模型在具有密集奖励的任务中的表现。进一步分析显示，DG-PRM 很好地适应了分布外场景，展示了卓越的泛化能力。

“Judgements prevent us from seeing the good that lies beyond appearances.”

– Wayne Dyer

1 介绍

过程奖励模型 (PRM) 在提高大型语言模型 (LLMs) 性能方面发挥着至关重要的作用 (Uesato et al., 2023; Yuan et al., 2024; Zhang et al., 2025)。与仅仅对最终目标提供反馈的结果奖励模型 (ORM) 不同，PRM 专注于奖励任务执行中的中间步骤或过程 (Setlur et al., 2025)。这种方法提供了更密集的奖励信号 (Li and Li, 2024)，尤其是在模型需要推理、分析和探索不同解决方案策略的复杂问题中 (Wei et al., 2022)。通过评估中间步骤，PRM 对于提高模型处理复杂任务的能力至关重要 (Lightman et al., 2024)。

PRM 可以广泛地分为启发式和生成式两类。启发式过程奖励 (Wang et al., 2024a,c) 依赖于

Question: Consider a satellite that is in a circular orbit around the Earth. The speed of the satellite decreases. What will happen to the satellite's orbital radius and period of revolution? Please justify your answer using principles of physics.	
Answer:	Judgement:
Step 1 The gravitational force provides the centripetal force for circular motion: $F_g = \frac{GMm}{r^2} = \frac{mv^2}{r}$, where M is Earth's mass, m is the satellite's mass, v is orbital speed, and r is orbital radius.	Step 1 (✓) The step 1 correctly explains the gravitational force providing the centripetal force and establishes the relationship between orbital speed, orbital radius, and the mass of the Earth. [[Correct]]
Step 2 Rearranging the equation gives $v = \sqrt{\frac{GM}{r}}$, showing orbital speed decreases as r increases.	Step 2 (✓) The conclusion that the orbital speed decreases as the orbital radius increases is accurate. [[Correct]]
Step 3 If v decreases, the gravitational force exceeds the required centripetal force, causing the satellite to spiral inward to a smaller orbit.	Step 3 (✗) The error occurs in the third step. The misconception here is that the satellite would spiral inward when its speed decreases. [[Incorrect]]
Step 4 Using Kepler's third law, $T^2 \propto r^3$, a larger r implies a longer orbital period T .	Step 4 (✓) The step correctly applies Kepler's third law and concludes that a larger orbital radius results in a longer orbital period. [[Correct]]
Step 5 Applying the escape velocity formula gives $v = \sqrt{\frac{2GM}{r}}$, the new speed corresponds to $r_{new} = \frac{2GM}{v^2}$.	Step 5 (✗) The mistake happened in the fifth step. The escape velocity formula $v = \sqrt{\frac{2GM}{r}}$ was incorrectly applied instead of the correct orbital velocity formula $v = \sqrt{\frac{GM}{r}}$. [[Incorrect]]
Step 6 The total mechanical energy $E = -\frac{GMm}{2r}$ increases as v decreases, so r decreases to conserve energy.	Step 6 (✗) The claim that the orbital radius decreases to conserve energy is incorrect. In a stable orbit, as the satellite's speed decreases, the orbital radius must increase. [[Incorrect]]

Figure 1: 一个来自 MT-Bench 的例子，其中 LLM 的判断包含丰富的细节，如粗体所示，但目前流行的 PRM 方法未能充分利用。

人为制定的标准来评估中间步骤与最终答案的相关性。尽管启发式奖励显著提高了模型的推理能力，但它们存在若干限制 (Zhang et al., 2025)。具体而言，它们通常需要客观的、基于参考的答案 (Luo et al., 2024)，而在评估标准波动的复杂场景中，这些答案难以获得。此外，使用启发式奖励训练的 PRM 通常表现出较差的泛化能力，甚至可能受到奖励抨击 (Weng, 2024; Wen et al., 2025)。另一方面，生成式过程奖励利用 LLM 来替代人工标注，通过将每个中间步骤标记为正确或错误，accordingly 提供正面或负面的反馈 (Mahan et al., 2024; Cao et al., 2024)。虽然生成式奖励利用 LLM 生成响应的能力，但评估仍然依赖于一组固定的标准 (Ling et al., 2023)，限制了其在不同领域中的适应性。

此外，现行的 LLM-as-Judge 方法仅利用最后的反馈(例如，正确/错误)而忽视了过程 (Kwon et al., 2023; Gao et al., 2023b) 中包含的有价值的信息，比如错误的严重性和错误类型。如图 1 所示，我们观察到 LLM 反馈包含丰富的细节和指导信息。然而，目前普遍的方法对错误标签赋予统一的负面奖励，忽视了错误的多样性和严重性。

在本文中，我们识别出当前奖励构建过程中存在的两个关键限制：(1) 使用固定的评估标准，和 (2) 依赖于统一的负面奖励，这无法捕捉到错误的多样性和严重性，从而限制了奖励过程的适应性和泛化能力。为了解决这些挑战，我们引入了一种新颖的框架，称为动态可泛化过程奖励建模 (DG-PRM)，旨在自动构建和精确分配过程奖励。我们提出使用奖励树来存储从 LLM 判断中提取的多方面评估标准。在评估过程中，它选择最相关的分步标准，这使得 DG-PRM 在跨域泛化上表现出色。我们还引入了帕累托支配估计，从多样的奖励信号集中选择出正、负样本对，提供明确的优化目标。实验结果表明，DG-PRM 在 PRMBENCH 上实现了最先进的性能，展示了卓越的 PRM 能力。通过提供上下文相关的奖励信号，DG-PRM 显著提升了 LLM 在广泛任务上的表现。此外，与 LLM-as-Judge 方法相比，DG-PRM 显示出增强的训练效率和对分布外场景更好的泛化能力。

我们的主要贡献如下：

- 我们介绍了 DG-PRM，这个自动化框架旨在构建动态且具有普遍适用性的过程奖励，以优化 LLM 反馈的效用。
- 为处理多样和复杂的奖励，我们引入了一种新颖的奖励树，以动态捕捉并利用适当的标准来进行每一步评估。
- 我们建议使用帕累托支配估计从多方面的奖励信号中识别正负对，从而提供更明确的优化目标。
- DG-PRM 通过提供精确、细粒度的过程奖励，显著提升了大型语言模型在各种任务中的性能，同时表现出高效的训练效率和卓越的泛化能力。

2 相关研究

结果奖励模型。 奖励模型旨在捕捉人类偏好并自动化评估模型输出 (Ouyang et al., 2022; Kaufmann et al., 2024; Sun et al., 2024a)。结果奖励模型 (ORM) 已被应用于广泛领域，包括安全性 (Dai et al., 2024)、数学问题求解 (Cobbe et al., 2021; Yang et al., 2024b) 和代码生成 (Dou et al., 2024; Sun et al., 2024c,b)。最近的研究，如 Wang et al. (2024b)，进一步扩展了奖励信号的维度，包括有用性、正确性、一致性、复杂性和冗长性。ORM 可以分为判别模型 (Stiennon et al., 2020; Ouyang et al., 2022) 和生成模型 (Mahan et al., 2024)。判别奖励模型通常添加一个分类头来评估输入的质量 (Gao et al., 2023a; Chen et al., 2024a)，而生成奖励模型利用大型语言模型 (LLMs) 的语言生成能力来评估输出 (Zhu et al., 2024; Li et al., 2024c)。Zheng et al. (2023)

证明了 LLMs 可以提供可扩展且可解释的奖励，与人类偏好高度一致。

随着大型语言模型 (LLMs) 越来越需要处理复杂任务，它们通常需要使用更多的标记来有效推理。当 LLM 输出的长度增加时，对输出的连贯性和正确性的全面评估变得困难。因此，构建具有密集奖励信号的过程奖励模型 (PRMs) 成为一种解决方案。目前构建 PRMs 的方法往往集中于客观领域，例如数学，这些领域有明确、确定的答案，旨在提高 LLM 在数学问题解决中的表现。模型将每一步输出的正确性建模为过程奖励，而采取更柔和的方法，通过结合当前步骤输出为正确的可能性作为过程奖励。然而，这些启发式方法限制了其广泛应用性，特别是在没有明确、唯一答案的情况下。例如，对于科学任务，应该考虑多样且复杂的奖励信号，各种输出组件可能需要关注不同的奖励信号。因此，构建动态和多样化的过程奖励是一个必要条件。

奖励信号。 奖励信号的设计是强化学习中的一个关键组件 (Sutton, 2018)。奖励信号通常可以分为三类：人工标注 (Bai et al., 2022a)、基于规则 (Glaese et al., 2022) 和 AI 反馈 (Lee et al., 2024b)。人工标注的奖励信号需要专家标记和验证，这可能昂贵且耗时 (Lightman et al., 2024)。Mu et al. (2024) 认为人类标注经常未能准确传递给标注者预期的行为，这使将期望结果转化为具体规则变得复杂。基于规则的系统，比如解析，使用编译器进行语法分析，将源代码翻译成可执行的二进制代码。然而，这种方法是特定领域的，无法轻易推广到其他领域。随着大型语言模型 (LLM) 的能力不断提高，研究已经转向利用 AI 反馈来生成奖励 (Bai et al., 2022b, 2023; Li et al., 2024b)。Kwon et al. (2023) 观察到，当 LLM 用作代理奖励函数时，显著提高了奖励与用户目标的一致性。在代码的背景下，McAleese et al. (2024) 表明 AI 模型比人工承包商识别出更多的错误。Cao et al. (2024) 证明 LLM 提供的密集奖励信号有效地提升了策略模型的性能。然而，当前的研究 (Gao et al., 2024; Chen et al., 2024b; Ling et al., 2023) 通常集中于对输出的评分或排序，而忽视了 LLM 生成的文本中丰富的指导信息。

3 预备知识

3.1 过程奖励建模

过程奖励模型 (PRMs) 评估模型输出中的中间步骤，而不仅仅是最终结果 (Lightman et al., 2024)。给定一个输入 x 和模型输出 $\hat{y}$ 分解为

n 步骤:

$$\hat{y} = \{\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(n)} | x\} \quad (1)$$

PRM 为每个步骤 $\hat{y}^{(t)}$ 分配一个奖励信号 $r^{(t)}$ ，在整个生成过程中提供密集的监督。

3.2 层次聚类

层次聚类 (Johnson, 1967) 通过基于相似性的逐步合并或分裂构建一个树状结构的聚类。给定 d 维空间中的一组数据点 $\{v_1, v_2, \dots, v_m\}$ 和一个距离函数 $\mathcal{D}$ ，层次聚类生成一个树状图 $\mathcal{H}$ ：

$$\mathcal{H} = \text{HierarchicalCluster}(\{v_1, v_2, \dots, v_m\}, \mathcal{D}) \quad (2)$$

常见的距离度量包括用于文本嵌入的余弦距离：

$$\mathcal{D}_{\text{cosine}}(v_i, v_j) = 1 - \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \quad (3)$$

在多目标优化中 (Miettinen, 1999)，我们经常面临竞争目标之间的权衡。帕累托支配的概念为比较解决方案提供了一个框架：

Definition 1 (Pareto Dominance). 给定两个解 y_i 和 y_j ，它们在 k 个目标上评估，得分分别为 $\{s_1^{(i)}, \dots, s_k^{(i)}\}$ 和 $\{s_1^{(j)}, \dots, s_k^{(j)}\}$ ，解 y_i Pareto-支配 y_j （记作 $y_i \succ y_j$ ）如果：

$$\forall l \in \{1, \dots, k\} : s_l^{(i)} \geq s_l^{(j)} \text{ and } \exists l : s_l^{(i)} > s_l^{(j)} \quad (4)$$

如果没有其他解支配一个解，则该解是 Pareto-最优的。所有 Pareto-最优解的集合构成了 Pareto 前沿。

4 DG-PRM

在上述基础上，我们现在介绍 DG-PRM，即我们用于动态和可推广流程奖励建模的新框架。如图 2 所示，关键创新包括：(1) 奖励树结构能够捕获多粒度评估标准，(2) 动态分配机制选择上下文合适的奖励，以及 (3) 基于帕累托的优化方法处理多样化的奖励信号。

4.1 自动过程奖励设计

与依赖于固定评估标准的现有 PRM 不同，DG-PRM 能够自动从比较分析中提取多样的奖励标准。给定正负输出对 $(\hat{y}_+, \hat{y}_-)$ ，我们使用一个判别函数 $\mathcal{J}$ 来分析它们的差异：

$$R_{\text{raw}} = \bigcup_{(x, \hat{y}_+, \hat{y}_-) \in \mathcal{D}} \mathcal{J}(x, \hat{y}_+, \hat{y}_-), \quad (5)$$

其中 $\mathcal{J}$ 输出一组奖励标准，解释为什么步骤 $\hat{y}_+$ 优于 $\hat{y}_-$ 。

Algorithm 1 动态过程奖励分配

Require: Step- t $\hat{y}^{(t)}$, Reward tree $\mathcal{T}$, Distance threshold ζ , Window size μ , Reward function $\mathcal{R}$, Analysis function Φ , Embedding function $\mathcal{V}$, Distance function $\mathcal{D}$, Score function $\mathcal{S}$

Ensure: Reward set $\mathbf{R}$, Score set $\mathbf{S}$

```

1: Initialize  $\mathbf{R} = \emptyset, \mathbf{S} = \emptyset$ 
2: for each timestep  $t = 1$  to  $n$  do
3:   if  $t - \mu \leq 0$  then
4:     Select all available previous steps up
     to  $t - 1$ , i.e.,  $\{\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(t-1)}\}$ 
5:   else
6:     Select the previous  $\mu$  steps, i.e.,
      $\{\hat{y}^{(t-\mu)}, \hat{y}^{(t-\mu+1)}, \dots, \hat{y}^{(t-1)}\}$ 
7:   end if
8:   Retrieve corresponding rewards and
   scores:  $\mathcal{I}_t = \{(\hat{y}^{(t-\mu)}, r_k^{(t-\mu)}, s_k^{(t-\mu)}), \dots\}$ 
9:   Add  $\mathcal{I}_t$  as supplementary information to
   reward function  $\mathcal{R}$ 
10:  Cal  $\phi_i = \Phi(\hat{y}^{(t)}, r_i^{\text{parent}})$ 
11:  for each  $\phi_i \in \mathcal{R}(\hat{y}^{(t)}, \mathcal{T}, \mathcal{I}_t)$  do
12:    for each child  $r_k^{(t)}$  of  $r_i^{\text{parent}}$  do
13:       $\delta_k^{(t)} = \mathcal{D}(\mathcal{V}(\phi_i), \mathcal{V}(r_k^{(t)}))$ 
14:      if  $\delta_k^{(t)} \leq \zeta$  then
15:        Add  $r_k^{(t)}$  to  $\mathbf{R}$ 
16:        Cal  $s_k^{(t)} = \mathcal{S}(r_k^{(t)}, \hat{y}^{(t)}, \mathcal{I}_t)$ 
17:        Add  $s_k^{(t)}$  to  $\mathbf{S}$ 
18:      end if
19:    end for
20:  end for
21: end for

```

通过自动验证器筛选出低质量标准（详见附录 B.1）后，我们得到了一组更精炼的 $\mathbf{R} = \{r_1, r_2, \dots, r_m\}$ 。

为了实现高效检索并避免冗余，我们将奖励标准组织成一个分层树结构。每个标准 r_i 被嵌入到一个 d 维向量空间中：

$$v_i = \mathcal{V}(r_i) \in \mathbb{R}^d \quad (6)$$

然后，我们应用增量层次聚类 (Zhang et al., 1997) 来构建奖励树 $\mathcal{T}$ 。为了减少冗余，将余弦距离低于阈值 ξ 的准则合并：

$$\text{merge}(r_i, r_j) \text{ if } \mathcal{D}_{\text{cosine}}(v_i, v_j) \leq \xi \quad (7)$$

由此产生的树结构将准则组织为粗粒度的父节点和细粒度的子节点：

$$\mathcal{T} = (\{r_1^{\text{parent}}, r_2^{\text{parent}}, \dots\}, \{r_1^{\text{child}}, \dots, r_m^{\text{child}}\}) \quad (8)$$

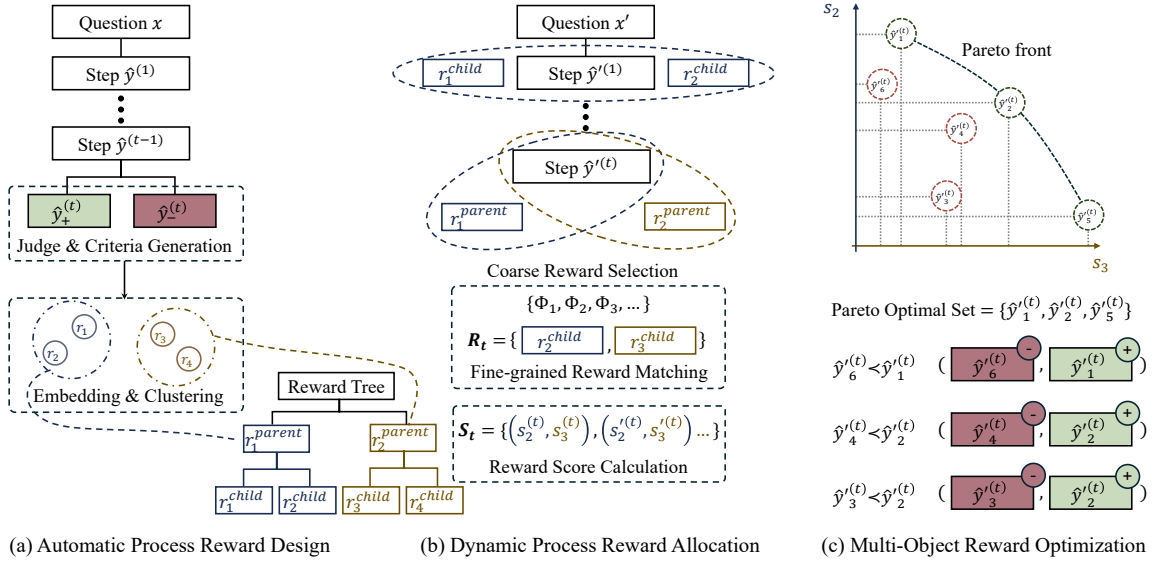


Figure 2: DG-PRM 概述。DG-PRM 由三个主要步骤组成：(a) 自动流程奖励设计，通过构建使用正负样本对 $\hat{y}_+$ 和 $\hat{y}_-$ 的奖励标准，将这些标准映射到特征空间，并通过层次聚类构建奖励树；(b) 动态流程奖励分配，在每个步骤中从奖励树中动态选择粗粒度奖励 r^{parent} 和细粒度奖励 r^{child} ，并根据相应的标准计算奖励分数；(c) 多目标奖励优化，基于计算出的奖励分数，选择帕累托最优的 $\hat{y}^{(t)}$ 作为优化目标。

这种层次化的组织方式使得从一般的评估方面到具体的准则的导航变得高效。

对于模型输出中的每一步 $\hat{y}^{(t)}$ ，DG-PRM 会动态选择树中的相关奖励。我们首先从之前的步骤构建上下文信息：

$$\mathcal{I}_t = \{(\hat{y}^{(i)}, r^{(i)}, s^{(i)}) \mid i \in [t - \mu, t - 1]\} \quad (9)$$

其中 μ 是保持计算效率的窗口大小。

奖励函数 $\mathcal{R}$ 从奖励树中识别出合适的父标准：

$$\{r_1^{\text{parent}}, r_2^{\text{parent}}, \dots\} = \mathcal{R}(\hat{y}^{(t)}, \mathcal{T}, \mathcal{I}_t) \quad (10)$$

对于每个父标准 r_i^{parent} ，一个分析函数 Φ 确定是否需要细粒度评估，并生成相应的评估方面：

$$\{\phi_{i,1}, \phi_{i,2}, \dots\} = \Phi(\hat{y}^{(t)}, r_i^{\text{parent}}) \quad (11)$$

，其中空集表示仅粗粒度评估就足够了。

当生成细粒度的评价方面时，我们将其匹配到子节点。对于每个评价方面 $\phi_{i,j}$ 和每个子节点 $r_k \in \text{children}(r_i^{\text{parent}})$ ，我们计算：

$$\delta_k = \mathcal{D}_{\text{cosine}}(\mathcal{V}(\phi_{i,j}), \mathcal{V}(r_k)) \quad (12)$$

选择子标准 r_k 的条件是其与任何评估方面的距离低于阈值 ζ ：

$$r_k \in \mathbf{R}_t \iff \exists \phi_{i,j} : \delta_k \leq \zeta \quad (13)$$

对步骤 t 设定的最终奖励结合了所有选定的标准：

$$\mathbf{R}_t = \bigcup_{i \in \mathcal{I}} \{r_k \mid r_k \text{ selected from } r_i^{\text{parent}}\} \quad (14)$$

，其中 $\mathcal{I} = \{i \mid r_i^{\text{parent}} \text{ was selected by } \mathcal{R}\}$ 。

每个选择的奖励 $r_k \in \mathbf{R}_t$ 随后根据当前步骤和上下文进行评分：

$$s_k^{(t)} = \mathcal{S}(r_k, \hat{y}^{(t)}, \mathcal{I}_t) \quad (15)$$

算法 1 概述了动态奖励分配过程，这构成了 DG-PRM 能够提供适合语境的过程奖励的核心。

4.2 多目标奖励优化

鉴于 $\mathbf{R}_t$ 中多样的奖励信号，我们采用帕累托占优来识别明确的优化目标。对于步骤 t 的多个候选输出 $\{\hat{y}_1^{(t)}, \hat{y}_2^{(t)}, \dots\}$ ，我们计算其帕累托最优集：

$$\mathbf{U} = \{\hat{y}_i^{(t)} \mid \nexists \hat{y}_j^{(t)} : \hat{y}_j^{(t)} \succ \hat{y}_i^{(t)}\} \quad (16)$$

由此，我们构建偏好对，其中帕累托最优解优于被支配的解：

$$\mathbf{V} = \{(\hat{y}_+^{(t)}, \hat{y}_-^{(t)}) \mid \hat{y}_+^{(t)} \in \mathbf{U}, \exists \hat{y}_-^{(t)} : \hat{y}_+^{(t)} \succ \hat{y}_-^{(t)}\} \quad (17)$$

我们调整 DPO (Rafailov et al., 2023) 以实现具有上下文依赖的逐步骤优化：

$$\mathcal{L}_{\text{DG-PRM}}(\theta) = -\mathbb{E}_{(\hat{y}_+^{(t)}, \hat{y}_-^{(t)}) \in \mathbf{V}} \left[\log \sigma \left(\beta \Delta^{(t)} \right) \right] \quad (18)$$

其中 $\Delta^{(t)}$ 测量偏好输出和非偏好输出之间的对数比差异：

$$\Delta^{(t)} = r_\theta^{(t)}(\hat{y}_+^{(t)}) - r_\theta^{(t)}(\hat{y}_-^{(t)}) \quad (19)$$

$$r_\theta^{(t)}(\hat{y}) = \log \frac{\pi_\theta(\hat{y} | x, \hat{y}^{(<t)})}{\pi_{\text{ref}}(\hat{y} | x, \hat{y}^{(<t)})} \quad (20)$$

此处, $r_{\theta}^{(t)}(\hat{y})$ 表示策略 π_{θ} 相对于参考策略 π_{ref} 在给定输入 x 和之前步骤 $\hat{y}^{(<t)}$ 下生成步骤 $\hat{y}$ 的对数比。这种表述确保模型学习生成相对于动态选择的奖励标准帕累托最优的步骤。

5 实验

5.1 评估数据集

为了评估 DG-PRM 在过程奖励建模中的效率, 我们使用 PRMBENCH (Song et al., 2025) 数据集。这个基准测试构建于 PRM800K 语料库上, 由 9 个不同错误类别的 6000 个数学问题组成, 使得可以对过程级别奖励模型进行全面的评估。

为了进一步评估 DG-PRM 在广泛任务中的有效性, 我们加入了三个代表不同领域的附加任务集: 一般任务、科学任务和常识推理。这些数据集包括:

- 综合: MT-Bench (Zheng et al., 2023), Arena-Hard (Tianle et al., 2024)
- 科学: QASC (Khot et al., 2020), 化学 QA (Wei et al., 2021)
- 常识: StrategyQA (Geva et al., 2021), ARC-c (Clark et al., 2018)

关于数据集的详细描述, 包括答案类型、训练和测试集分布以及许可信息, 可以在附录 A 中找到。

5.2 实验设置

对于 PRMBENCH, 我们使用与“Promoted as Critic Models”配置相同的设置, 其中包括四个开源模型: QwQ-Preview-32B (Team, 2024), DeepSeek-R1-Distill-Qwen-32B, DeepSeek-R1-Distill-Qwen-7B 和 DeepSeek-R1 (Guo et al., 2025), 以及两个专有模型: o1-mini (OpenAI, 2024c) 和 GPT-4o (OpenAI, 2024a)。在一般、科学和常识场景中, 我们还包括 DeepSeek-R1-Distill-Qwen-1.5B 和 DeepSeek-R1-Distill-Qwen-14B 模型。为了构建奖励树, 我们使用 BAAI/bge-en-icl 模型 (Li et al., 2024a) 来构建 $\mathcal{V}$, 维度 d 设置为 4096。对于层次聚类, 我们应用 BIRCH 算法 (Zhang et al., 1997) 来创建 $\mathcal{H}$ 。除非另有说明, 我们将 DeepSeek-R1-Distill-Qwen-7B 作为骨干模型, 并设置如下超参数: 合并超参数 $\xi = 0.25$, 距离超参数 $\zeta = 0.2$ 和步长超参数 $\mu = 20$ 。有关更详细的实验步骤和超参数分析, 请参阅附录 ?? 和附录 ??。

对于 PRMBENCH, 我们使用官方结果作为基线。为了进一步验证 DG-PRM 的有效性, 我们在 General、Science 和 Commonsense 场景下

与以下设置进行比较: (1) Original: 原始模型, 没有任何优化。(2) ORM: 我们训练一个奖励模型, 为完整输出提供正面和负面反馈。(3) Critic Models: 基于 Song et al. (2025), 我们引导大型语言模型 (LLMs) 逐步批判解决方案。按照 Mahan et al. (2024), 我们区分 Critic Direct, 即模型直接生成结果, 和 Critic CoT, 即模型首先进行推理和分析后才给出答案。(4) Human Annotation: 在人工标注的过程标签上进行训练, 作为上限。

我们还在分析中加入了 Implicit PRM (Yuan et al., 2024), 这是一个令牌级的奖励基线, 具体细节见附录 B.2。我们观察到, 大量的偏好数据对于训练是至关重要的, 以便有效建模 DPO 等效奖励 (Rafailov et al., 2024)。

5.3 主要结果

PRMBENCH。 表 1 展示了在 PRMBENCH 上的结果, 其中生成模型比判别模型表现出明显的优势, 这一发现与 Zheng et al. (2024a) 一致。DG-PRM 进一步显著提升了大型语言模型的过程回报建模能力。在更具挑战性的 PS 任务中, PRMs 必须能够识别缺失条件和前提错误, R1-Distill-Qwen-7B 从 45.6 % 显著提高到 62.4 %。这一显著的提升可归因于 DG-PRM 的逐步判别和动态回报分配, 这有效提高了错误识别的准确性。此外, R1 表现出强劲的性能, DG-PRM 进一步增强了这一性能, 将整体准确率从 69.5 % 提升到 76.5 %, 达到新的最先进水平。值得注意的是, 在灵敏度测试中, 平均得分超过 80 %。这可归因于 R1 的复杂推理生成了更高质量的回报标准, 而 DG-PRM 有效利用了这一点来提升性能。我们将在附录 ?? 中进一步分析这一现象。

总体而言。 表 3 展示了在 MT-Bench 和 Arena-Hard 上的结果。我们观察到偏好优化显著提高了模型的胜率。通过逐步分析, 关键模型的表现优于 ORM, 取得了更好的性能。在 Arena-Hard 中, 使用带有过程分析的关键反馈链展示出相对于关键直接的明显优势, 强调了在评估中逐步分析的重要性。与基线相比, DG-PRM 显著提高了模型的胜率, 达到了接近人工标注的水平, 甚至在 Arena-Hard 的“显著更好”类别中超过了人工标注。我们将其归因于 DG-PRM 能够整合多维奖励信息, 从而实现更准确的过程奖励。

在图 4 中, 我们比较了不同方法在参数数为 1.5B、7B、14B 和 32B 的模型中的表现。我们观察到, DG-PRM 在所有参数规模下显著提升了模型的性能, 显示出相对于其他方法的明显优势。此外, DG-PRM 的性能接近于人工标注, 甚至在 StrategyQA 上超过了人类水平表现。值

Model Name	Overall	Simplicity			Soundness					Sensitivity			
		NR.	NCL.	Avg.	ES	SC.	DC.	CI	Avg.	PS	DR.	MS.	Avg.
Open-source Discriminative Process Reward Model													
Llemma-PRM800k-7B [†]	52.0	49.3	53.4	51.4	56.4	47.1	46.7	53.3	50.9	51.0	53.5	93.6	66.0
MATHMinos-Mistral-7B [†]	54.2	48.8	54.0	51.4	57.0	52.1	50.7	57.8	54.4	52.8	55.8	91.1	66.5
MathShepherd-Mistral-7B [†]	47.0	44.0	50.3	47.1	49.4	44.5	41.3	47.7	45.7	47.2	48.6	86.1	60.7
RLHFlow-PRM-Mistral-8B [†]	54.4	46.1	47.3	46.7	56.6	55.1	54.4	63.8	57.5	51.5	56.2	97.9	68.5
RLHFlow-PRM-Deepseek-8B [†]	54.2	46.4	48.9	47.6	55.7	55.0	53.2	66.2	57.5	49.0	55.4	99.8	68.1
Prompted as Critic Models													
o1-mini ^{†*}	68.8	65.6	63.7	64.6	74.5	67.7	73.8	72.3	72.1	61.8	64.8	100.0	75.5
GPT-4o [†]	66.8	57.0	62.4	59.7	72.0	69.7	70.7	71.1	70.9	62.5	65.7	99.2	75.8
QwQ-Preview-32B [†]	63.6	57.2	55.6	56.4	67.4	72.3	66.2	66.9	68.2	57.8	62.7	100.0	73.5
R1-Distill-Qwen-32B [†]	60.2	57.2	51.9	54.5	66.1	68.4	69.3	64.8	67.2	53.3	54.6	99.9	69.3
R1-Distill-Qwen-7B [†]	52.6	32.9	37.9	35.4	47.3	54.1	48.4	48.0	49.4	45.6	46.8	100.0	64.1
DeepSeek-R1	69.5	66.0	65.2	65.6	74.8	70.1	72.2	72.9	72.5	63.2	66.2	100.0	76.5
Dynamic and Generalizable Process Reward Modeling													
o1-mini	73.5	71.2	69.1	70.2	77.5	74.8	76.3	75.6	76.1	67.3	70.4	100.0	79.2
GPT-4o	72.3	66.1	69.0	67.6	75.9	73.2	74.7	76.4	75.1	66.8	70.9	100.0	79.2
QwQ-Preview-32B	70.0	63.2	65.4	64.3	72.4	74.3	72.9	74.5	73.5	64.5	67.9	100.0	77.5
R1-Distill-Qwen-32B	69.0	62.0	64.7	63.4	71.1	72.6	71.3	73.8	72.2	63.6	66.8	100.0	76.8
R1-Distill-Qwen-7B	65.2	60.4	62.1	61.3	69.8	68.1	69.5	72.1	69.9	62.4	64.2	100.0	75.5
DeepSeek-R1	76.5	74.1	72.3	73.2	80.1	77.5	78.9	79.4	79.0	71.0	74.3	100.0	81.8

Table 1: DG-PRM 与其他强大基线在 PRMBENCH (Song et al., 2025) 上的性能比较 (PRM-Score %)。最佳结果以粗体显示。[†] 表示来自 官方排行榜 的结果, * 表示对 394 个样本的子集进行的评估。评估包括主流的开源和闭源模型, 以进行公平比较。每个数据集类别的详细信息和评估目标在附录 A 中提供。

	QASC	ChemistryQA	StrategyQA	ARC-c
# Coarse-grained Reward	2.3	0.9	1.5	1.2
# Fine-grained Reward	1.3	3.7	2.1	1.9
Selection Ratios	0.81	0.76	0.62	0.79
w/o Pareto-based filtering	75.70 %	85.71 %	82.97 %	90.63 %
w. Pareto-based filtering	78.40 % (↑ 2.70 %)	87.50 % (↑ 1.79 %)	85.58 % (↑ 2.61 %)	94.65 % (↑ 4.02 %)

Table 2: 对 DG-PRM 在不同数据集上的统计分析和消融研究, 其中 # 表示奖励的数量。

得注意的是, 使用 DG-PRM 训练的 R1-Distill-Qwen-32B 模型在 QASC 和 StrategyQA 数据集上都超越了人类表现。在 StrategyQA 中, 我们观察到对于 Critical CoT 方法的性能下降, 在 14B 模型中甚至低于 Critical Direct 方法。这是因为在 StrategyQA 中存在隐含推理的挑战, 模型常常难以确定分析的适当方向, 尤其是在推理过程高度复杂时。在这种情况下, 无方向的分析可能导致错误的判断。DG-PRM 通过在每一步提供相关且明确的奖励目标, 有效地改善了判断。

在图 5 中, 我们分析了不同方法的泛化能力。我们观察到, ORM 和 Critic Direct 在域外 (OOD) 场景下表现显著下降, 部分结果甚至低于基线表现。为了进一步分析 Critical 方法, 我们引入了 Critic Instruction, 这种方法不提供示例, 只提供简单的指令来指导 LLM 生

成过程奖励。在 OOD 设置中, Critical CoT 方法在 ARC-c 上表现出最小的性能下降, 但在 ChemistryQA 上却有显著的下降, 甚至表现比 Critic Instruction 还差。这归因于 QASC 中多样化的问题分布, 导致在转移到化学领域时发生混淆。相比之下, DG-PRM 通过选择相关且有效的奖励, 展示了出色的领域泛化能力。在附录 ?? 中, 我们利用 DG-PRM 的可扩展性构建了一个通用的奖励树, 以进一步分析其跨域泛化能力。

在图 6 中, 我们比较了不同方法的准确性与训练步骤的函数关系。我们观察到, DG-PRM 在所有数据集上表现出卓越的训练效率, 仅需要使用 30 % 的训练步骤就达到与 Critic CoT 等同的性能。这种效率归因于 DG-PRM 使用帕累托优势估计来选择最具代表性的正负样本, 使得模型更容易捕捉样本之间的差异并优化拟

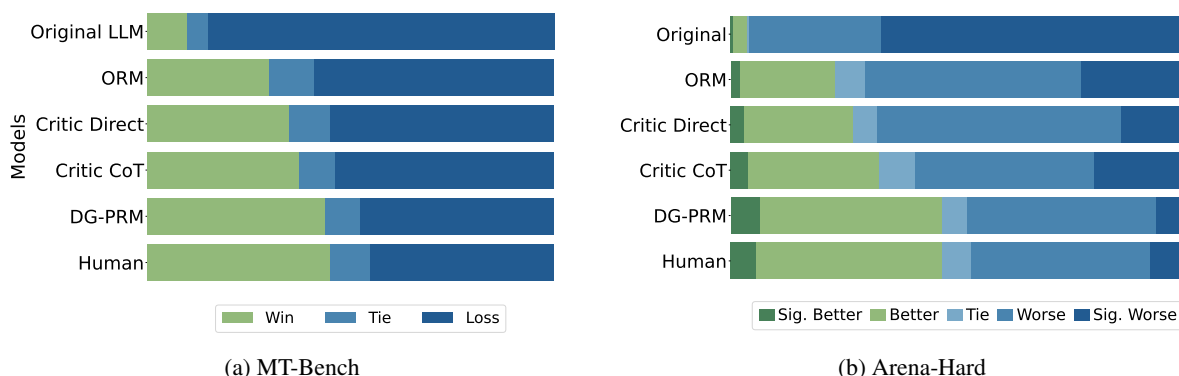


Figure 3: 在 (a) MT-Bench 和 (b) Arena-Hard 上的性能比较。R1-Distill-Qwen-7B 模型用作骨干网络，GPT-4o (OpenAI, 2024a) 用作评估模型。

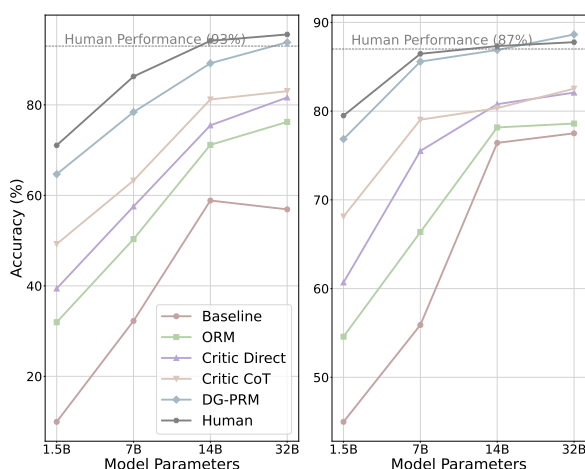


Figure 4: 关于 (a) QASC 和 (b) StrategyQA 的性能比较。评估包括不同参数规模的模型：1.5B，7B，14B 和 32B。

合目标。结果是，DG-PRM 不断提升模型性能，显著超越基准方法。

5.4 分析

在表 2 中，我们分析了不同数据集在每一步的奖励选择。每一步平均有大约 3.7 个奖励标准，包括每一步平均 1.5 个粗粒度奖励标准和 2.3 个细粒度奖励标准。我们观察到，对于 QASC 数据集，粗粒度奖励标准完全足够满足要求，而对于如 ChemistryQA 这类更专业的数据集，则需要更多的细粒度奖励标准。

我们进一步进行了关于基于帕累托过滤有效性的消融研究。如表 2 所示，我们通过随机将最高得分样本与其他样本配对来建立不使用帕累托过滤的设置，确保与使用帕累托过滤配置相同数量的配对。我们观察到，采用帕累托优势有效地提高了模型的准确性，在 ARC-c 任务上实现了超过 4 个% 的性能提升。帕累托优势始终确保对于每个奖励标准 $\hat{y}_+^{(t)} > \hat{y}_-^{(t)}$ ，从而为模型提供更明确的优化和学习目标。此外，我们分析了基于帕累托过滤后用于配对构

建的样本相对于所有样本的比例。我们观察到这一比例在所有数据集上都超过 0.5，这表明 DG-PRM 展示了对采样步骤的高效利用。

在图 7 中，我们分析了与奖励树构建阶段相关的成本。构建成本主要包括生成流程奖励标准的判断开销 (Judge)、过滤低质量标准的验证开销 (Validation)，以及对粗粒度奖励标准进行汇总的开销 (Summarization)。我们观察到，对于单个样本，输入开销大约为 7,000 个标记，而输出开销约为 3,000 个标记。使用官方定价的 GPT-4o 模型，我们发现 DG-PRM 的成本显著低于手动标注成本。此外，一旦奖励树构建完成，由于其通用性，它可以被持续利用和更新。

我们进一步分析了在过程奖励建模阶段的开销，如图 8 所示，我们与使用多数投票法通过多采样来确定最终输出的 Critic CoT-SC 进行了比较。我们观察到，简单的自一致性 (Wang et al., 2023) 未能有效提高模型的逐步判断准确性。与 Critic CoT-SC(50) 相比，DG-PRM 仅需一半的计算开销，同时在 PRMBench 上实现了 4.8 % 的整体性能提升。这归因于 DG-PRM 的动态奖励匹配机制，它为每一步匹配相关且合适的过程奖励标准来计算过程奖励分数。与 Critic CoT 相比，所提供的过程奖励分数更具针对性，并且更易于追溯奖励分数背后的理由。

在本文中，我们首先确认当前的过程奖励模型 (PRM) 通常是特定领域的，并针对特定目标进行了优化，这限制了它们的通用性。核心挑战在于需要详细的、逐步的监督，这使得标准化奖励信号变得困难。为了解决这个问题，我们提出了动态通用过程奖励建模 (DG-PRM)，它利用奖励树来存储细粒度奖励信号，并动态提供步骤特定的奖励。我们还引入了帕累托优势估计，以从多样的奖励信号中选择正面和负面的反馈。跨基准的实验表明，DG-PRM 可以

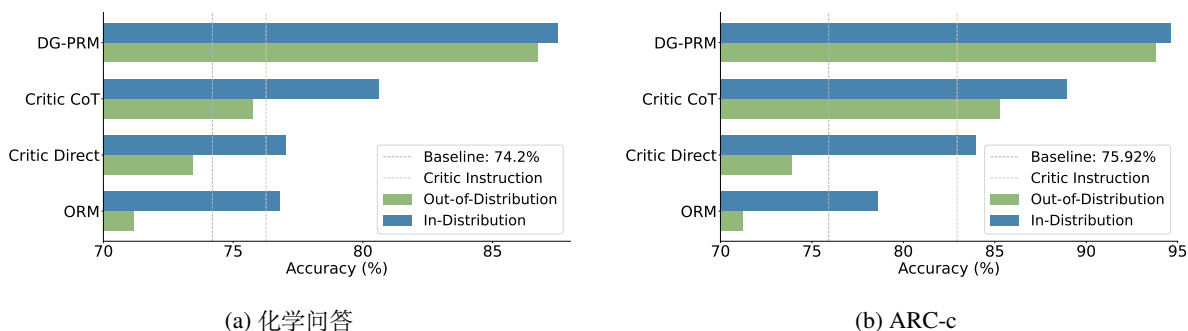


Figure 5: 关于 (a) ChemistryQA 和 (b) ARC-c 的泛化分析。在分布外设置中，我们为 ChemistryQA 样本提供使用在 QASC 上构建的过程奖励模型的过程反馈，为 ARC-c 样本提供使用在 StrategyQA 上构建的过程奖励模型的过程反馈。Critic-Instruction 指的是只使用指令而不提供任何特定领域示例的方法。

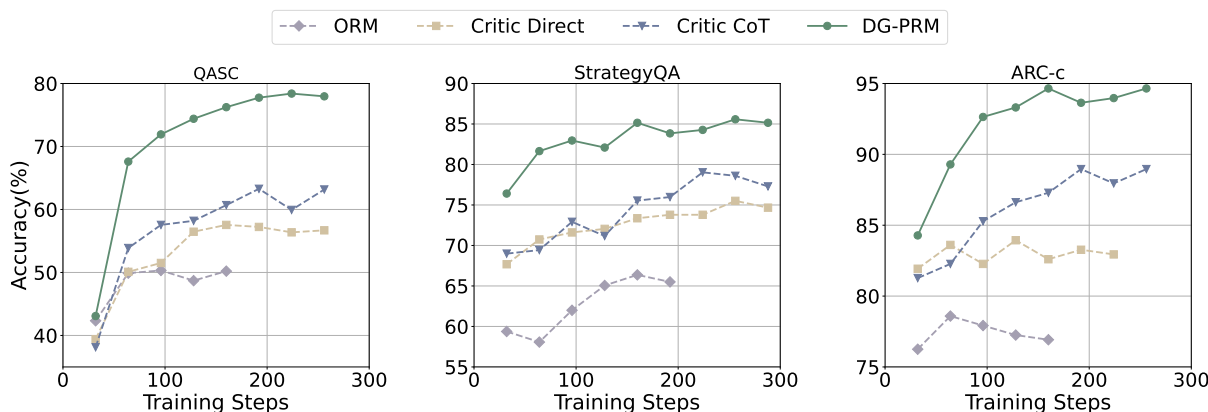


Figure 6: 在 QASC、StrategyQA 和 ARC-c 数据集上，准确率 (%) 随训练步骤的变化。

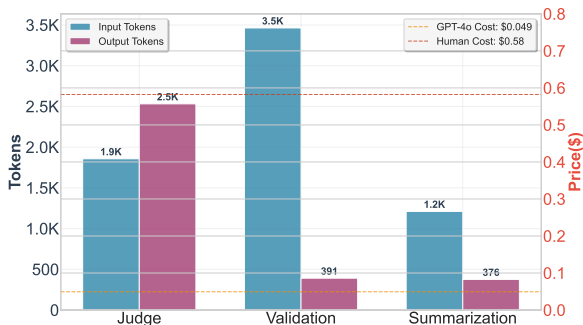


Figure 7: 在 StrategyQA 数据集上进行奖励树构建阶段的每样本标记消耗分析。

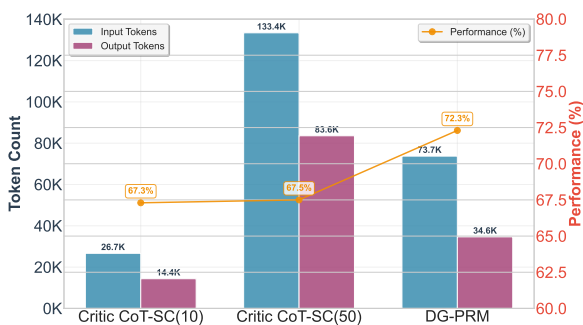


Figure 8: 不同过程奖励建模方法在 PRMBench 数据集上的令牌消耗和性能比较。

提供更准确的过程奖励信号。值得注意的是，它在 PRMBENCH 上达到了新的最先进水平，并显示出在改善各种领域的模型性能方面的效果，通过提供有效且精确的奖励信号展示了其卓越的通用性。

尽管我们已经在多个数据集上展示了 DG-PRM 的有效性，包括 PRMBENCH (Song et al., 2025)、MT-Bench (Zheng et al., 2023) 和 ArenaHard (Tianle et al., 2024)，但 DG-PRM 仍有很大潜力扩展到更多领域。随着 LLMs 越来越多地应用于各种科学领域，DG-PRM 可以被适配来设计特定领域的奖励树，比如药物发现 (Zheng et al., 2024b)、疾病诊断 (Zhou et al., 2024) 和天气预报 (Wang and Karimi, 2024)。我们将这些应用视为未来研究的令人兴奋的途径。

虽然 DG-PRM 使用 LLM 构建奖励树，并且我们已经引入了一个自动验证器来移除低质量的奖励标准，但仍有可能存在与人类期望不一致的奖励标准。这可能导致优化模型中的潜在风险。因此，引入适当的人为监督以改进和增强奖励树，并甚至设计和构建特定领域的奖励树是至关重要的。

适应高级模型。 PRM 的主要优势在于为复杂推理任务提供可靠而有效的过程信号。然而，具有深度推理能力的最先进模型，如 OpenAI 的 o1 (OpenAI, 2024c) 和 o3 (OpenAI, 2025b)，由于它们的闭源性质，无法被调整。一个有前途的替代方案是 DeepSeek R1 模型 (Guo et al., 2025)。然而，出于资源限制，我们只使用了该模型的精简版，而不是完整的 671B R1 模型。未来，我们计划扩展 DG-PRM 的适应性以适用于更高级的模型，进一步提升广义过程监督的有效性。

我们严格遵守用于我们实验的数据集的许可证。所有数据集都是英文的，我们注意确保我们的使用符合每个数据集的预期用途。每个数据集的许可证信息的详细概述见表 3。

遵守模型使用条款。 在整个实验过程中，我们严格遵循模型的使用条款。我们遵守为模型设定的使用指南，包括商业模型的服务条款和 API 使用政策。对于开源模型，我们遵循其许可协议和使用约束。

数据隐私 我们的方法使用 LLMs 的输出构建奖励树，而无需收集个人信息或敏感数据。我们已经彻底审查了实验中使用的提示和数据，以确保它们不包含任何可识别个人身份的信息或冒犯性内容。

在实验过程中，我们邀请了五位具有硕士或以上学位的标注员对模型输出的过程奖励进行标注并评估 DG-PRM 输出。其中一名标注员是化学系的博士，专门负责处理 ChemistryQA 任务。根据当地的小时工资标准向参与者提供补偿。所有标注员均来自泛太平洋地区，并且性别比例平衡，我们确保在评估中考虑到不同地区的文化偏好，从而保证结果的有效性。提供给标注员的说明如表 20 和表 21 所示。

在评估过程中，我们采用了 AI 工具来协助分析模型输出。具体而言，我们使用了 GitHub Copilot 来协助编程。我们确保 AI 工具的使用符合提交指南和伦理标准。

环境保护 训练和测试中的扩展需要显著的计算能力和资源。高效且准确的奖励信号能够使模型训练更加高效，并在推理过程中淘汰质量较低的路径，促进 AI 的可持续发展。这种方法有助于减少碳排放，支持环境保护。

6

致谢 本研究得到国家自然科学基金（编号 U24B20181）的资助。本研究中的计算是在复旦大学的 CFFF 平台上完成的。

References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. *Benchmarking foundation models with language-model-as-an-examiner*. In *Advances in Neural Information Processing Systems*, volume 36, pages 78142–78167. Curran Associates, Inc.
- Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024. *Enhancing reinforcement learning with dense rewards from language model critic*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9138, Miami, Florida, USA. Association for Computational Linguistics.
- Lu Chen, Rui Zheng, Binghai Wang, Senjie Jin, Caishuang Huang, Junjie Ye, Zhihao Zhang, Yuhao Zhou, Zhiheng Xi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024a. *Improving discriminative capability of reward models in RLHF using contrastive learning*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15270–15283, Miami, Florida, USA. Association for Computational Linguistics.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. *Do not think that much for 2+3=? on the overthinking of o1-like llms*. *Preprint*, arXiv:2412.21187.
- Zhipeng Chen, Kun Zhou, Xin Zhao, Junchen Wan, Fuzheng Zhang, Di Zhang, and Ji-Rong Wen. 2024b. *Improving large language models via fine-grained reinforcement learning with minimum editing constraint*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5694–5711, Bangkok, Thailand. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anas-tasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. *Chatbot arena: An open platform for evaluating llms by human preference*. *Preprint*, arXiv:2403.04132.

- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. [Safe RLHF: Safe reinforcement learning from human feedback](#). In *The Twelfth International Conference on Learning Representations*.
- Shihan Dou, Yan Liu, Haoxiang Jia, Enyu Zhou, Limao Xiong, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, Zhiheng Xi, Yuhao Zhou, Tao Ji, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. [StepCoder: Improving code generation with reinforcement learning from compiler feedback](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4571–4585, Bangkok, Thailand. Association for Computational Linguistics.
- Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, Junjie Hu, Tianyu Liu, and Baobao Chang. 2024. [Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback](#). *Preprint*, arXiv:2406.14024.
- Leo Gao, John Schulman, and Jacob Hilton. 2023a. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.
- Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023b. [Human-like summarization evaluation with chatgpt](#). *Preprint*, arXiv:2304.02554.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#). *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Amelia Glaese, Nat McAleese, Maja Trbac, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. [Improving alignment of dialogue agents via targeted human judgements](#). *Preprint*, arXiv:2209.14375.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.
- Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.
- Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. [A survey of reinforcement learning from human feedback](#). *Preprint*, arXiv:2312.14925.
- Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 8082–8090.
- Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. [Reward design with language models](#). In *The Eleventh International Conference on Learning Representations*.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rashtogi, and Sushant Prakash. 2024b. [Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback](#). *Preprint*, arXiv:2309.00267.
- Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024a. [Making text embedders few-shot learners](#). *Preprint*, arXiv:2409.15700.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024c. [Generative judge for evaluating alignment](#). In *The Twelfth International Conference on Learning Representations*.

- Wendi Li and Yixuan Li. 2024. [Process reward model with q-value rankings](#). *Preprint*, arXiv:2410.11287.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let's verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. [Deductive verification of chain-of-thought reasoning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Ju-jie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. [Improve mathematical reasoning in language models by automated process supervision](#). *Preprint*, arXiv:2406.06592.
- Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. [Generative reward models](#). *Preprint*, arXiv:2410.12832.
- Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. [Llm critics help catch llm bugs](#). *Preprint*, arXiv:2407.00215.
- Kaisa Miettinen. 1999. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.
- Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian D Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. [Rule based rewards for language model safety](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- OpenAI. 2024a. [Hello gpt-4o](#).
- OpenAI. 2024b. [New embedding models and api updates](#). Accessed: 2025-02-13.
- OpenAI. 2024c. [Openai o1-mini: Advancing cost-efficient reasoning](#).
- OpenAI. 2025a. [Introducing deep research](#).
- OpenAI. 2025b. [Openai o3-mini](#). Release.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*.
- Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. 2024. [From \\$r\\$ to \\$q^*\\$: Your language model is secretly a q-function](#). In *First Conference on Language Modeling*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. [Rewarding progress: Scaling automated process verifiers for LLM reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. [Prmbench: A fine-grained and challenging benchmark for process-level reward models](#). *arXiv preprint arXiv:2501.03124*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.
- Hao Sun, Thomas Pouplin, Nicolás Astorga, Tension Liu, and Mihaela van der Schaar. 2024a. [Improving LLM generation with inverse and forward alignment: Reward modeling, prompting, fine-tuning, and inference-time optimization](#). In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.
- Qiushi Sun, Zhirui Chen, Fangzhi Xu, Kanzhi Cheng, Chang Ma, Zhangyue Yin, Jianing Wang, Chengcheng Han, Renyu Zhu, Shuai Yuan, et al. 2024b. [A survey of neural code intelligence: Paradigms, advances and beyond](#). *arXiv preprint arXiv:2403.14734*.
- Shengjie Sun, Runze Liu, Jiafei Lyu, Jing-Wen Yang, Liangpeng Zhang, and Xiu Li. 2024c. [A large language model-driven reward design framework via dynamic feedback for reinforcement learning](#). *Preprint*, arXiv:2410.14660.
- Richard S Sutton. 2018. Reinforcement learning: An introduction. *A Bradford Book*.
- Qwen Team. 2024. [Qwq: Reflect deeply on the boundaries of the unknown](#).

- Li Tianle, Chiang Wei-Lin, Frick Evan, Dunlap Lisa, Zhu Banghua, Joseph E. Gonzalez, and Stoica Ion. 2024. [From live data to high-quality benchmarks: The arena-hard pipeline.](#)
- Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Yamamoto Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2023. [Solving math word problems with process-based and outcome-based feedback.](#)
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. [Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations.](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models.](#) In *The Eleventh International Conference on Learning Representations*.
- Yang Wang and Hassan A. Karimi. 2024. [Exploring large language models for climate forecasting.](#) *Preprint*, arXiv:2411.13724.
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024b. [Helpsteer2: Open-source dataset for training top-performing reward models.](#) *Preprint*, arXiv:2406.08673.
- Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. 2024c. [Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7309–7319, Miami, Florida, USA. Association for Computational Linguistics.
- Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models.](#) In *Advances in Neural Information Processing Systems*.
- Zhuoyu Wei, Wei Ji, Xiubo Geng, Yining Chen, Baihua Chen, Tao Qin, and Daxin Jiang. 2021. [Chemistry{qa}: A complex question answering dataset from chemistry.](#)
- Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2025. [Language models learn to mislead humans via RLHF.](#) In *The Thirteenth International Conference on Learning Representations*.
- Lilian Weng. 2024. [Reward hacking in reinforcement learning.](#) *lilianweng.github.io*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024. [Free process rewards without process labels.](#) *Preprint*, arXiv:2412.01981.
- Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1997. Birch: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1:141–182.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. [The lessons of developing process reward models in mathematical reasoning.](#) *Preprint*, arXiv:2501.07301.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024a. [Processbench: Identifying process errors in mathematical reasoning.](#) *Preprint*, arXiv:2412.06559.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena.](#) *Preprint*, arXiv:2306.05685.
- Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T. May, Geoffrey I. Webb, Shirui Pan, and George Church. 2024b. [Large language models in drug discovery and development: From disease mechanisms to clinical trials.](#) *Preprint*, arXiv:2409.04481.
- Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, Liqiao Xia, Jeremy Yeung, Daochen Zha, Genevieve B. Melton, Mingquan Lin, and Rui Zhang. 2024. [Large language models for disease diagnosis: A scoping review.](#) *Preprint*, arXiv:2409.00097.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2024. [JudgeLM : Fine-tuned large language models are scalable judges.](#)

A 数据集详细信息

在我们的实验中，我们选择了七个数据集，这些数据集涵盖了多种任务类型，要求模型进行复杂而精细的推理。关于这些数据集的样本大小、来源和许可的详细信息见表格 1。

- PRMBENCH 使用 PRM-Score 进行评估，它是 F1 Score 和负 F1 Score 的平均值，并强调错误发生的步骤。表格 ?? 详细列出了每个类别的缩写、相应的全名、评估目标以及每个类别中的实例数。表格 6 中显示了一个示例。
- MT-Bench (Zheng et al., 2023) 被评估在 1 到 10 的范围内。为了确保全面的评估和实际的场景，我们采用多轮设置，由 GPT-4o (OpenAI, 2024a) 输出评估分数。我们还报告了与 GPT-4o 模型的胜率。表 4 展示了 MT-Bench 中的各种类型及其相应的数量，表 11 则展示了 MT-Bench 中的一个示例。
- Arena-Hard (Tianle et al., 2024) 使用对 GPT-4-0314 输出的胜率作为评估指标。我们使用了来自 [Arena-Hard 的官方仓库](#) 的官方输出结果，利用 GPT-4o (OpenAI, 2024a) 作为判断模型。Arena-Hard 包含 250 个多样化场景。一个例子如表 12 所示。
- QASC (Khot et al., 2020) 是一个多项选择的科学数据集，它为每个答案提供多个事实解释，帮助 LLMs 从事实的角度评估错误的原因。一个简单的例子如表 7 所示。
- ChemistryQA (Wei et al., 2021) 是一个化学数据集，由 [苏格拉底](#) 收集，涵盖了超过 200 个主题。它提供了必要的知识、条件和详细的解决步骤。表 8 中展示了一个样本。
- StrategyQA (Geva et al., 2021) 挑战模型以策略性地解决隐式推理问题。它提供问题分解和相应的事实，有助于识别失败点。一个简单的例子在表格 9 中显示。
- ARC-c (Clark et al., 2018) 是 ARC 数据集的一个挑战子集，用于评估模型在各种事实类型上的基本推理能力，例如基本事实 & 性质和过程 & 因果关系。

由于 PRMBENCH、MT-Bench 和 Arena-Hard 并没有提供适合构建奖励树的训练数据集，我们利用 MATH 数据集 (Hendrycks et al.,

2021) 来训练 PRMBENCH 的奖励树。对于 MT-Bench 和 Arena-Hard，我们从 LMSYS-Human-Preference-55k (Chiang et al., 2024) 数据集中选择 5,000 个样本来构建奖励树。

B 实验细节

在此过程中，我们使用换行符或明确的标签（例如“Step1”和“Step2”）对步骤进行分段。对于每个步骤，我们基于数据集中提供的参考答案 y 或评估标准 C （例如 MT-Bench (Zheng et al., 2023)）构建正负标签对。我们使用相应的判定提示和数据集中的参考答案 y 来构建这些对。

我们利用 GPT-4o 构建判别模型 $\mathcal{J}$ ，为错误步骤生成适当的奖励标准。采用 BAAI/bge-`en-icl` (Li et al., 2024a) 嵌入模型来获得嵌入向量 v ，其中向量空间的维度为 $d = 4096$ 。此外，在我们的实验分析中，我们结合了 `text-embedding-3-large` (OpenAI, 2024b) 和 `nvidia/NV-Embed-v2` (Lee et al., 2024a) 模型。

BIRCH 算法 (Zhang et al., 1997) 被用于层次聚类，使用来自 `sklearn.cluster` 的 Birch 实现。在合并奖励标准的过程中，我们保留较长的标准。聚类完成后，使用 GPT-4o 来总结上层粗粒度奖励标准 r^{parent} 。这一步对后续的奖励分配至关重要，因为使用质量更高的模型进行总结会产生更准确的奖励信号。

用于生成粗粒度奖励标准的相同模型也用于奖励分配，以确定是否应提供细粒度奖励标准 r^{child} 。为了确保评估的准确性，我们进行多次采样以确认所选择奖励标准的适当性。在我们的实验中，我们将采样频率设置为 5，保留出现超过 3 次的粗粒度奖励标准。通常，模型的判断是一致的。

在评分过程中，先前生成的步骤、评估信息和分数被连接在前面以供参考。默认情况下，我们使用 GPT-4o 作为评分者 $\mathcal{S}$ 。我们还使用模型自身的评估和 o3-mini 的评估结果来分析评估结果，详情见 ??。

如果找不到相应的正面或负面反馈，将进行额外的采样，直到达到最大尝试次数。表格 16 提供了一个示例，其中在提示中放置了选定的细粒度奖励标准。在实践中，由于上下文的限制，只能包含粗粒度奖励标准，因此我们根据匹配仔细选择细粒度奖励标准。在表格 17、表格 18 和表格 19 中，我们提供了几个粗粒度和细粒度过程奖励标准的示例。

在训练过程中，我们使用 8 块互联的 H200 GPU，批量大小为 32，学习率为 $5e-6$ ，DPO-beta 为 0.1。其他设置基于 HuggingFace's DPOTrainer 的默认参数。由于 DeepSeek-R1

DATASET	TASK	ANSWER FORMAT	# TRAIN.	# TEST.	LICENSE
PRMBENCH (Song et al., 2025)	PRM	Incorrect Position	-	6,216	Apache license 2.0
MT-Bench (Zheng et al., 2023)	General	Rating	-	80	Apache license 2.0
Arena-Hard (Tianle et al., 2024)	General	Rating	-	500	Apache license 2.0
QASC (Khot et al., 2020)	Science	Multi-choice	8,134	926	Apache license 2.0
ChemistryQA (Wei et al., 2021)	Science	Text	2,721	392	CC0-1.0
StrategyQA (Geva et al., 2021)	Commonsense	T/F	2,061	229	MIT license
ARC-c (Clark et al., 2018)	Commonsense	Multi-choice	1,119	299	CC BY-SA 4.0

Table 3: 详细描述实验中使用的数据集。# TRAIN. 表示训练样本的数量，# TEST. 表示用于评估的样本数量。

Task Category	Evaluation Focus	Number of Samples
Writing	Text Generation	10
Roleplay	Interaction	10
Reasoning	Logical Analysis	10
Math	Mathematical Problem Solving	10
Coding	Programming	10
Extraction	Information Retrieval	10
STEM	Scientific Knowledge	10
Humanities	Cultural Understanding	10

Table 4: MT-Bench 中的类别统计。

API 暂时不可用，我们使用 ollama 在 8 块互联的 A100 上以 INT4 精度部署完整的 671B 版本。相同的方法也用于其他 Distill 模型的部署。

在生成期间，我们根据任务和模型调整温度参数 γ 。具体而言，我们发现对于 R1-Distill-Qwen-1.5B 和 R1-Distill-Qwen-7B，较高的温度导致序列过长且出现许多重复的解决方案，这在 QwQ (Team, 2024) 模型中由 Chen et al. (2025) 观察到。因此，我们设定 $\gamma \in [0.5, 0.6]$ 并筛选掉超过 4096 个标记的序列。对于 R1-Distill-Qwen-14B 和 R1-Distill-Qwen-32B，我们发现 $\gamma \in [0.7, 0.8]$ 在 MT-Bench 和 Arian-Hard 上产生更满意的结果，而较低的温度在 QASC、StrategyQA 和 ARC-c 任务上表现更好，因此我们设定 $\gamma \in [0.6, 0.7]$ 。对于 MT-Bench 和 Arian-Hard 数据集，我们计算五个输出结果的平均得分。对于其他数据集，我们基于相应的指标计算最终结果。

B.1 自动验证器

在奖励树 $\mathcal{T}$ 的构建过程中，自动验证器在确保奖励标准的高质量方面起着至关重要的作用。这个验证器可以应用于任何模型的奖励树构建，过滤掉低质量的奖励标准。我们使用 GPT-4o (OpenAI, 2024a) 作为自动验证器，它通过提示评估奖励标准，并提供三种评估结果之一：Good、Ordinary 和 Bad。这些评估的详细标准已列出，表格 13 提供了一个一般示例。

为了评估自动验证器与人工评估之间的一致性，我们随机选择了 100 个来自自动验证器的输出，并将它们与人工评估的结果进行了比较。我们计算了自动输出与三位人工评估者分

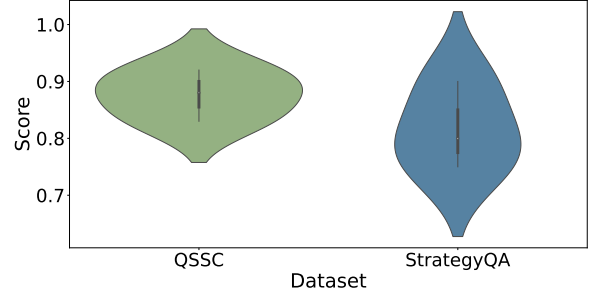


Figure 9: 自动验证器与人工评估的一致性

配标签之间的一致比例，结果如图 9 所示。我们的分析表明，在 QASC 和 StrategyQA 数据集上，自动验证器与人工评估表现出高度一致性。

对于 ORM，我们遵循 Liu et al. (2024) 的方法，并基于 Qwen-2.5-7B 模型训练 ORM (Yang et al., 2024a)。我们使用 AdamW 作为优化器，批量大小为 16，学习率为 $2e-6$ 。在构建评论模型的过程中，我们提示 LLM 进行逐步评估。具体来说，评论直接输出每一步是否正确，而评论 CoT 在给出最终结果之前提供对输出步骤的详细分析。每一步都用 [[correct]] 或 [[wrong]] 进行评估，最终的判断模型提取出任何错误的步骤。如果所有步骤都正确，输出为 []。表 14 和表 15 展示了评论直接和评论 CoT 的一个例子。

我们针对 QASC 数据集上的实验中使用的超参数进行了消融研究，如图 10 所示。

超参数 ξ 控制相似奖励标准的合并。如图 10a 所示，较大的 ξ 值会导致相关标准的合并，在极端情况下， r^{parent} 下的所有 r^{child} 会被合并，结果是一个粗粒度的过程奖励。这对性能有显著影响。另一方面，较小的 ξ 值会导致奖励树 $\mathcal{T}$ 中存在大量冗余标准，从而增加噪声并导致性能下降。这也导致 $\mathcal{S}$ 中的冗余计算，从而增加不必要的开销。因此，我们将 $\xi = 0.25$ 设置为合适的阈值。

距离超参数 ζ 超参数 ζ 控制细粒度过程奖励标准的选择，如图 10b 所示。较大的 ζ 值会导致包含大量不相关的奖励标准，进而导致整体性能下降。这是因为过多的标准使 DG-PRM 难

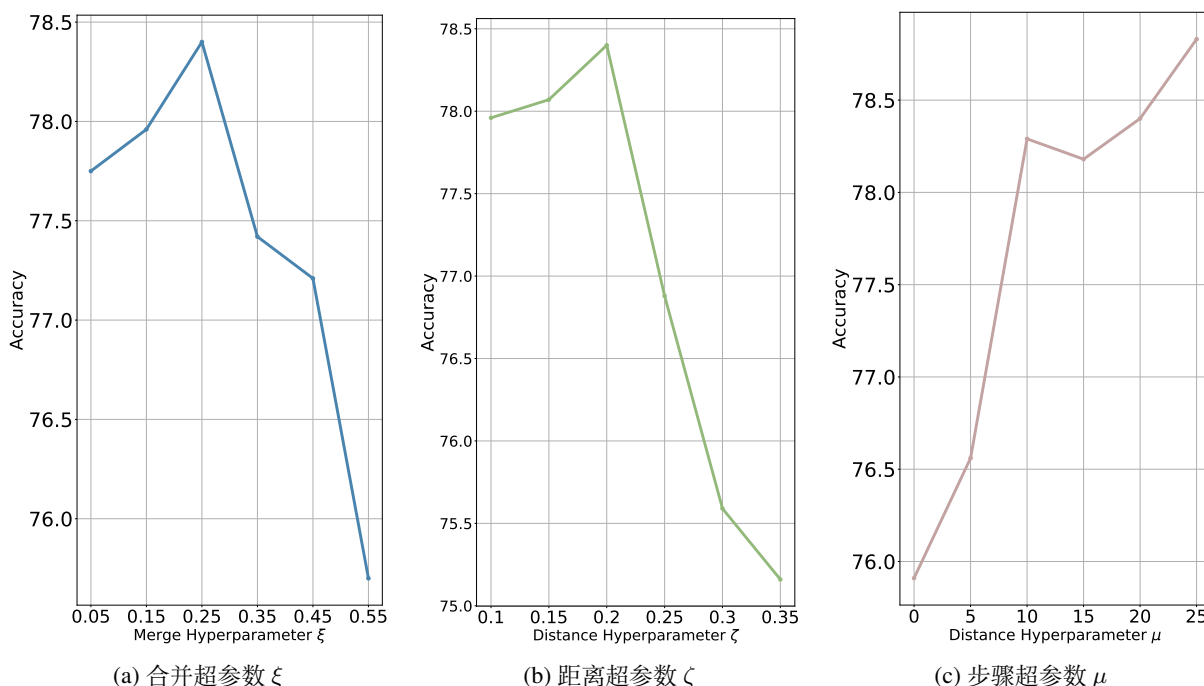


Figure 10: 在 QASC 数据集上对超参数进行消融分析，使用 R1-Distill-Qwen-7B 作为骨干。

以选择相应的正负样本。因此，有必要将标准限制在一个最佳范围内。另一方面，较小的 ζ 值可能无法匹配适当的细粒度过程奖励标准，同样会影响性能。

超参数 μ 控制在奖励标准选择 $\mathcal{R}$ 和评分 $\mathcal{S}$ 中可以参考的步骤数。我们观察到，随着步骤数量的增加，性能持续提高。因此，提供更多的步骤有利于 DG-PRM 选择和评分的准确性。然而，考虑到模型上下文窗口的限制，我们不设置 μ 超过 20。在 ChemistryQA 场景中，推理过程更为复杂，这可能会超过模型的上下文窗口。我们还观察到性能提升逐步减弱。因此，考虑到成本开销，我们设置 $\mu = 20$ 。

在图 11 中，我们分析了不同模型作为 $\mathcal{R}$ 和 $\mathcal{S}$ 对性能的影响，使用了 MT-Bench 数据集。我们利用雷达图展示各个类别评分的变化。我们的观察表明，采用更好的评判模型可以显著提高写作、推理和编码任务的性能，这表明 DG-PRM 可以持续受益于更优的评判模型。此外，我们发现 DeepSeek-R1 表现与 o3-mini 相当，甚至在编码和 STEM 任务中表现更佳。这表明 DG-PRM 也适用于先进的开源模型。

B.2 隐式概率路径规划

在图 12 中，我们比较了不同方法在 MT-Bench 数据集上的表现。我们包括使用官方实现基于 LMSYS-Human-Preference-55k (Chiang et al., 2024) 数据集训练的 Implicit PRM (Rafailov et al., 2024)。我们使用 R1-Distill-Qwen-7B 重新生成响应，并基于 GPT-4o 的评分构建选中

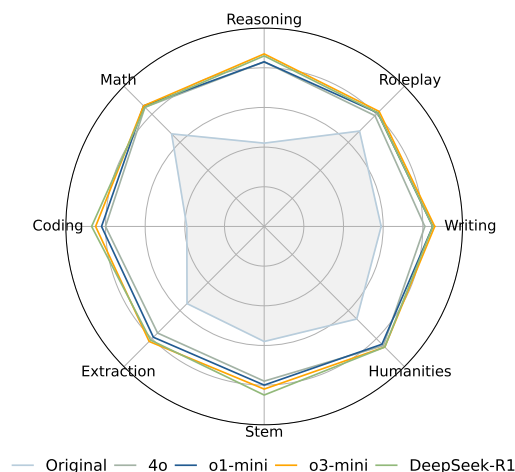


Figure 11: 使用 R1-Distill-Qwen-7B 作为基础模型，评估模型对 MT-Bench 数据集评分的影响。

和拒绝的配对。我们观察到，Implicit PRM 有助于显著提升性能，特别是在推理和编码方面，这表明 Implicit PRM 能够有效地建模具有明确定义目标的任务。然而，DG-PRM 表现出更显著的改进，提供了模型性能的更全面增强，例如其在写作和人文任务中的优秀结果。此外，DG-PRM 更具可解释性，提供了对正样本相对于负样本优势的明确解释，使优化目标更易于理解。

B.3 嵌入模型

表 5 分析了不同嵌入模型对性能的影响。嵌入模型在奖励树的构建和细粒度过程奖励标准的选择中起着关键作用。我们发现，DG-PRM

Embedding Model	QASC	ChemistryQA
BAAI/bge-en-icl	78.40	87.50
text-embedding-3-large	79.04	88.01
nvidia/NV-Embed-v2	78.29	87.24

Table 5: 嵌入模型的分析。

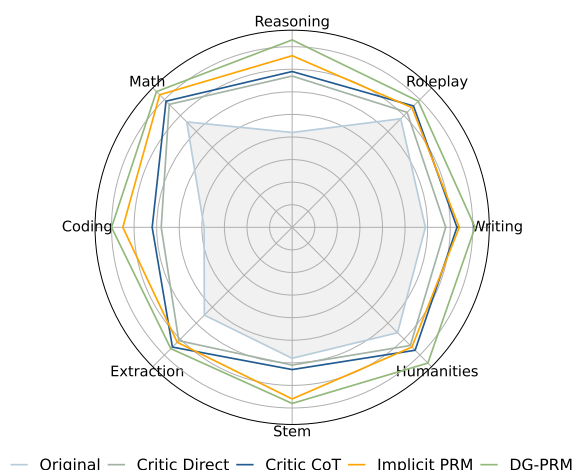


Figure 12: 使用 R1-Distill-Qwen-7B 作为骨干，比较不同 PRM 方法在 MT-Bench 数据集上的性能。

在嵌入模型选择方面表现出稳健性，即使在使用开源模型时也能取得令人满意的性能。因此，我们在实验中使用 BAAI/bge-en-icl 模型作为嵌入函数 $\mathcal{V}$ 。

在表格 ?? 中，我们研究了不同层次聚类方法对性能的影响。我们选择了两种方法：凝聚聚类和分裂聚类，使用 Ward 方法 (Ward Jr, 1963) 来定义聚类距离。我们的研究结果显示，分裂聚类表现更好。然而，鉴于分裂聚类高昂的计算成本，BIRCH 算法 (Zhang et al., 1997) 提供的增量更新显著减少了开销，使得奖励标准的增加和移除变得更加高效和方便。

具体而言，我们观察到 BIRCH 算法在 PRM-Bench 上完成层次聚类用时 10.2 秒，在 ChemistryQA 上用时 19.53 秒，展示了其在不同数据集上能高效地在 20 秒内完成聚类的能力。因此，我们选择 BIRCH 算法作为聚类方法来获取奖励树。

为了进一步评估 DG-PRM 的可推广性，我们构建了一个统一的奖励树，其中包含了来自 MATH、QASC、Chemistry、StrategyQA 和 ARC-c 训练数据集的一组丰富的标准。图 13 展示了在 MT-Bench 上使用该统一奖励树的实验结果。与为 MT-Bench 单独构建的奖励树相比，统一奖励树在人文学科和数学任务上表现更佳。我们观察到模型在评分过程中包含了更广泛的事实性评估。这些标准可能源于 MATH 中的多样性数学视角和 StrategyQA 中的事实

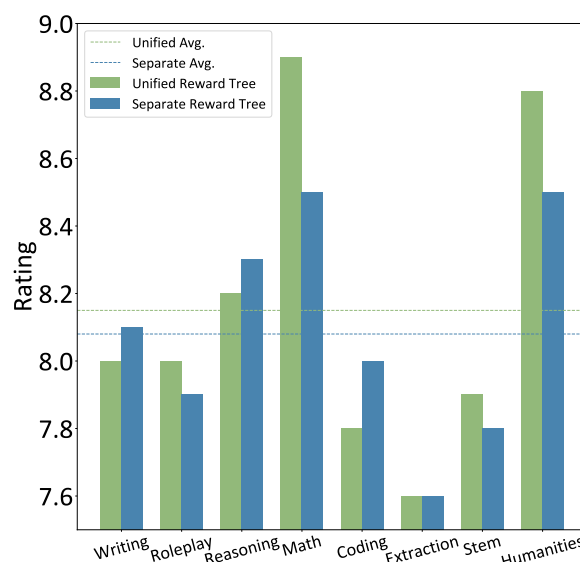


Figure 13: 使用统一奖励树与单独奖励树对 MT-Bench 评分的比较。

相关评估。虚线代表了统一和单独奖励树的平均表现。我们的研究结果表明，统一奖励树获得了更高的整体分数，突显出 DG-PRM 的卓越可扩展性。

为了进一步分析 DG-PRM 奖励标准分配和输出评分的合理性，我们随机选择了 100 个样本，包括问题、答案、分配的标准和最终的奖励评分。这些由三名评估员独立评价，他们判断所分配标准和输出评分的适当性，将每个样本标记为 Good、Ordinary 或 Bad。结果如图 14 所示。我们观察到 Good 标签的中位数比例接近 80%，而 Bad 标签的比例低于 10%，这表明大多数评估人员认为 DG-PRM 的奖励分配和输出评分是合理的，展示了 DG-PRM 奖励标准构建和得分输出的合理性。

在表 ?? 中，我们使用各种优化算法分析了 DG-PRM 的性能。MT-Bench 的实验结果表明，DG-PRM 构建的过程奖励在多种优化算法中都是有效的。在单轮和多轮情景中，性能与基线相比都有显著提升。在测试的算法中，DPO 实现了最显著的性能提升，因此我们在实验中利用 DPO 来优化我们的策略模型。

C

明日的展望

近年来，我们见证了模型能力的快速扩展，人工智能正日益融入人类科学研究中 (OpenAI, 2025a)。这就需要开发一个普遍适用的过程奖励模型 (PRM)。在短期内，PRM 将侧重于与人类科学家的偏好和我们客观世界的理解对齐，防止错误的搜索路径，并引导模型朝向更有意义的探索。随着模型能力的不断发展，PRM 将引导模型朝向人类期望的目标，比如

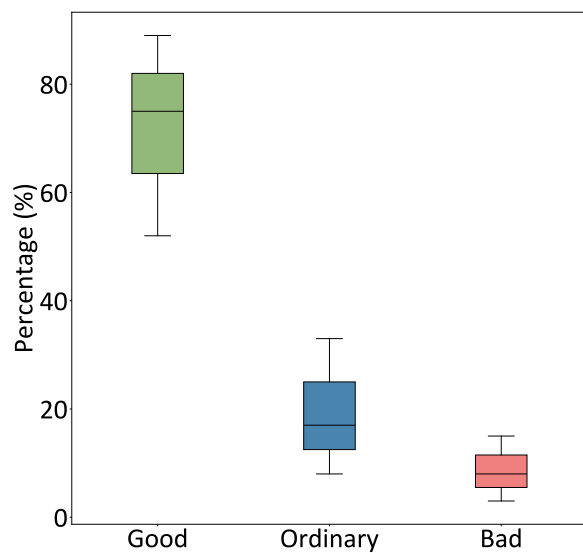


Figure 14: 人类对 DG-PRM 输出的评估，包括奖励标准和评分的选择。

环境可持续性和疾病治疗，而不是在探索中促进有害产品的创造。我们相信，一个具有普遍适用性的 PRM 将在推动人工智能向更安全、更可靠的方向发展中发挥关键作用。

Original Process

Question : If $x^2 - x - 1 = 0$, what is the value of $x^3 - 2x + 1$?

Step 1 : Let's start with the first equation and see if we can solve for x.

Step 2 : We can use the quadratic formula to find x.

Step 3 : The quadratic formula is $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Step 4 : Substituting $a = 1$, $b = -1$, and $c = -1$, we get $x = \frac{1 \pm \sqrt{5}}{2}$.

Step 5 : Since $x^2 - x - 1 = 0$, we replace x^2 with $x + 1$ in $x^3 - 2x + 1$.

Step 6 : Substituting, $x^3 - 2x + 1 = x(x + 1) - 2x + 1$.

Step 7 : Simplify: $x^3 - 2x + 1 = (x^2 + x) - 2x + 1$.

Step 8 : Substitute $x^2 = x + 1$: $x^3 - 2x + 1 = (x + 1 + x) - 2x + 1$.

Step 9 : Simplify: $x^3 - 2x + 1 = 2$.

Modified Process

Question : If $x^2 - x - 1 = 0$, what is the value of $x^3 - 2x + 1$?

Step 1 : Let's start with the first equation and see if we can solve for x.

Step 2 : We can use the quadratic formula to find x.

Step 3 : The quadratic formula is $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Step 4 : Substituting $a = 1$, $b = -1$, and $c = -1$, we get $x = \frac{1 \pm \sqrt{5}}{2}$.

Step 5 : Since $x^2 - x - 1 = 0$, we replace x^2 with $x + 1$ in $x^3 - 2x + 1$.

Step 6 : Substituting, $x^3 - 2x + 1 = x(x + 1) - 2x + 1$.

Step 7 : To simplify x^3 , use the property of cubic roots: $x^3 = \sqrt[3]{x}$.

Step 8 : Substitute $x^3 = \sqrt[3]{x}$ and $x^2 = x + 1$ into $x^3 - 2x + 1$: $x^3 - 2x + 1 = \sqrt[3]{x} - 2x + 1$.

Step 9 : Simplify further: $x^3 - 2x + 1 = \sqrt[3]{x} - x + 1$.

Reason

A domain inconsistency was introduced in Step 7 by incorrectly assuming that $x^3 = \sqrt[3]{x}$. This assumption is valid for root-based equations but not in the context of solving a polynomial equation. This error propagates through the rest of the reasoning, leading to a flawed solution.

Table 6: PRMBENCH 样本。

Question : Climate is generally described in terms of what?

Answer Choice : (A) sand (B) occurs over a wide range (C) forests (D) Global warming (E) rapid changes occur (F) local weather conditions (G) measure of motion (H) city life

Fact1 : Climate is generally described in terms of temperature and moisture.

Fact2 : Fire behavior is driven by local weather conditions such as winds, temperature and moisture.

answerKey : F

Table 7: QASC 样本。

Question : How many moles of phosphorus trichloride would contain 3.35 times 10^{24} molecules of phosphorus trichloride?

Conditions : Number [OF] phosphorus trichloride molecules [=] $pu3.35 \times 10^{24}$

TargetVariable : Mole [OF] phosphorus trichloride [IN] moles

Answer : 5.56 moles

Table 8: 化学 QA 样本。

Question : Did Aristotle use a laptop?

Answer : No

Table 9: StrategyQA 样本。

Question : Juan and LaKeisha roll a few objects down a ramp. They want to see which object rolls the farthest. What should they do so they can repeat their investigation?

Answer Choice : (A) Put the objects in groups. (B) Change the height of the ramp. (C) Choose different objects to roll. (D) Record the details of the investigation.

Answer : D

Table 10: ARC-c 样本。

Question 1 : Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

Answer **A Glimpse into the Enchanted Land of Hawaii: A Cultural and Adventurous Journey**

Our recent trip to Hawaii was an absolute blast, a journey that left us thoroughly charmed by its beauty, culture, and unique spirit. As we explored this archipelago of islands, each destination offered a taste of its rich heritage and stunning landscapes.

One experience that stood out for me was the traditional Luau at Waikiki Beach Park. What made it truly unforgettable was the opportunity to step into a cultural immersion, dancing to the rhythm of hula while sharing stories with locals. It was like being transported back in time to a vibrant Hawaiian village, where music and hospitality were the foundation of life.

Our journey began in Honolulu, where we strolled along Waikiki Beach, soaking in the warm sun and vibrant ocean views. This iconic location isn't just a beach; it's a symbol of Hawaii's blend of tradition and modernity, perfect for capturing photos with the iconic "Hawaii Five-O" lighthouse in the background.

Next up was Pearl Harbor, where history took center stage. The experience was deeply moving as we witnessed the power of remembrance at this historic site. It reminded us how crucial it is to honor our heroes and preserve their legacy. We then ventured to Hanauma Bay, a destination steeped in nature's beauty. Scattered with fronds and lined by sea grasses, this bay offered a serene escape, allowing us to truly disconnect from the hustle and bustle of city life.

Finally, we visited the Ironman Training Center at Waikoloa State Park, challenging ourselves physically and mentally. The rigorous workout left us exhilarated, pushing our limits while fostering camaraderie with fellow participants.

Reflecting on this trip, I can say that Hawaii's culture is as vibrant as its landscapes. Every experience was a lesson in appreciation for life's beauty and the richness of diverse traditions. Our time there will forever be cherished, not just for the sights and sounds, but for the profound impact it had on our hearts.

Hawaii's magic continues to inspire us, and we can't wait to return. If you're ever in the mood for an adventurous adventure with a touch of culture, Hawaii is the perfect destination!

Question 2 : Rewrite your previous response. Start every sentence with the letter A.

Answer :

A Clever Rewrite: A Glimpse into the Enchanted Land of Hawaii

A recent trip to Hawaii was an absolute adventure, a journey that left us thoroughly charmed by its beauty, culture, and unique spirit. A highlight of our trip was experiencing traditional Hawaiian culture at a luau. A visit to Waikiki Beach Park offered a chance to dance to the rhythm of hula while sharing stories with locals—a true cultural immersion.

A beginning in Honolulu saw us strolling along Waikiki Beach, soaking in the warm sun and vibrant ocean views—a quintessential location for capturing photos with the iconic "Hawaii Five-O" lighthouse. A stop at Pearl Harbor provided a powerful reminder of history's significance. A visit to Hanauma Bay offered a serene escape, lined by sea grasses and scattered with fronds.

A challenge we faced during our adventure was finding words that started with "A," but I managed to overcome it by rephrasing sentences creatively. For example, instead of saying "We arrived in Honolulu," I wrote, "Arriving in Honolulu was an exciting beginning." Each sentence now starts with an "A," making the blog post a bit unconventional yet engaging.

Despite this unusual formatting, the overall message remains clear and appealing—Hawaii's culture is as vibrant as its landscapes. Every experience was a lesson in appreciating life's beauty and the richness of diverse traditions. Our time there will forever be cherished, not just for the sights and sounds, but for the profound impact it had on our hearts.

A trip to Hawaii is an absolute must for anyone seeking an adventurous adventure with a touch of culture. A destination like Hawaii is perfect! We can't wait to return.

If you're in the mood for an adventurous adventure with a touch of culture, A trip to Hawaii could be just what you need. A wonderful experience that will stay with you forever!

Judge

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Your evaluation should focus on the assistant's answer to the second user question. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

Table 11: MT-Bench 样本。

Question : Help me draft a research introduction of this topic "Data-Driven Insights into the Impact of Climate and Soil Conditions on Durian Floral Induction"

Answer : **Introduction to the Impact of Climate and Soil Conditions on Durian Floral Induction**

Durian, a fragrant and sweet tropical fruit, is a cropping staple in various regions of Southeast Asia, particularly in countries like Malaysia and Thailand. As a major agricultural commodity, its economic significance is substantial, contributing significantly to local economies. However, durian cultivation faces challenges, especially in the face of climate change, which can impact agricultural productivity globally.

Climate change poses particular threats to durian cultivation, affecting factors such as temperature, humidity, and rainfall, all of which are critical for optimal flowering. Floral induction, a horticultural practice aimed at promoting flowering in plants to enhance fruit production, is essential for durian cultivation. Similarly, soil conditions, including fertility, pH levels, and organic matter, play a pivotal role in influencing the flowering process. Some durian varieties are particularly sensitive to these environmental factors, necessitating a deeper understanding of their interplay.

This study seeks to explore how climate and soil conditions specifically impact the floral induction of durian, with the aim of providing insights that can enhance sustainable agricultural practices. By identifying optimal conditions, the research contributes to improving durian yields and efficiency, offering valuable information for farmers and policymakers. This investigation not only addresses existing gaps in current knowledge but also provides a foundation for developing more resilient durian varieties and farming practices, ensuring continued productivity in the face of environmental challenges.

Judge

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Assistant A is significantly better: `[[A>B]]`

2. Assistant A is slightly better: `[[A>B]]`

3. Tie, relatively the same: `[[A=B]]`

4. Assistant B is slightly better: `[[B>A]]`

5. Assistant B is significantly better: `[[B>A]]`

Example output: "My final verdict is tie: `[[A=B]]`".

Table 12: 竞技场难度样本。

Prompt : You are given a reward criterion used to evaluate a task or output from a model. Your task is to categorize the quality of this reward criterion into one of the following three levels: Good, Ordinary, or Bad. Please ensure your categorization is based on the following specific standards for each level.

Good Reward Criterion :

- 该标准与任务高度相关，重点评估响应中最重要和核心的方面。它很好地与任务的目标和要求一致。
- 这个标准鼓励产生高质量、专注且精确的输出。它促进了清晰性、准确性和彻底性，确保关键点得到充分解决。
- 该标准在既不太宽泛也不太狭窄之间取得了平衡，确保其适用于一系列的响应，同时不忽略关键细节。它促进了简便的应用和最佳模型调整。
- 该标准实用、直接且定义明确，能够进行一致和可重复的评估。它最大限度地减少了解释中的歧义。

Output : [[Good]]

Ordinary Reward Criterion :

- 该标准有些具体，但缺乏必要的深度或精确性，无法有效评估响应的所有重要方面。它可能会遗漏关键考虑因素或存在被误解的空间。
- 该标准与任务目标一致，但可能无法考虑微妙的细节或复杂的要求。通过增加更多细节或涵盖更广泛的范围可以进行改进。
- 虽然该标准提供了指导，但由于其模糊性或复杂性，它可能不会导致最优的模型调整。应用起来可能更难以保持一致。
- 该标准在某种程度上是实用的，但在某些方面缺乏清晰性或具体性，可能导致不一致的评估或次优的反馈。

Output : [[Ordinary]]

Bad Reward Criterion :

- 该标准与任务不相关，或关注于那些对评估结果没有实质贡献的因素，例如无关的性能指标或非必要的特性。
- 该标准过于复杂、模棱两可或具有误导性，使得理解或应用困难。它可能引入不必要的混淆或未能与任务直接对齐。
- 该标准鼓励错误类型的回应，例如惩罚有益的特征（例如，创造力、幽默或参与度）或评估与任务目标无关的输出方面。
- 该标准不切实际或过于严格，导致评估过于主观或过于武断。它可能会阻碍有意义的反馈或改进。

Output : [[Bad]]

Table 13: 自动验证器提示。

Prompt : Please evaluate the reasoning process below and directly output the erroneous step. If the step is correct, please output [[]]. You may refer to the following example. Question : What type of water formation is formed by clouds? Answer Choice : (A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid Step 1 : Understanding the question. The question asks, "What type of water formation is formed by clouds?" We need to determine the correct type of water formation typically associated with clouds. Step 2 : Analyzing the options. The options provided are: (A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid Step 3 : Identify the key concept. Clouds are made up of tiny water droplets or ice crystals that have condensed from vapor in the atmosphere. These droplets can eventually combine and fall to the ground as precipitation. Step 4 : Eliminate irrelevant options. Pearls, streams, shells, diamonds, beads, and cooled do not seem to be directly related to water formation associated with clouds. Step 5 : Focus on the most plausible option. "Rain" is the most likely type of water formation that clouds produce, as clouds release water in the form of rain when the droplets or ice crystals become heavy enough to fall. Final answer : [[E]]
Step 1 : [[correct]] Step 2 : [[correct]] Step 3 : [[correct]] Step 4 : [[wrong]] Step 5 : [[correct]] Judge : [[4]]
Question : Where do beads of water come from? Answer Choice : (A) Too much water (B) underground systems (C) When the water is too cold (D) Water spills (E) Vapor turning into a liquid (F) Warm air moving into cold air (G) At the peak of a mountain (H) To another location like underground Step 1 : Understanding the question. The question asks where beads of water come from. Beads of water are typically small droplets that form when water vapor condenses. Step 2 : Analyzing the options. The options provided are: (A) Too much water (B) underground systems (C) When the water is too cold (D) Water spills (E) Vapor turning into a liquid (F) Warm air moving into cold air (G) At the peak of a mountain (H) To another location like underground Step 3 : Focus on key concepts. Beads of water typically form when water vapor cools and condenses into liquid droplets. Step 4 : Eliminate irrelevant options. (A) "Too much water" is not a correct explanation for beads of water. (B) "Underground systems" don't directly relate to the formation of beads of water in this context. (C) "When the water is too cold" might make sense for freezing, but we're dealing with condensation, not freezing. (D) "Water spills" just refers to water being spilled, not the condensation process. (G) "At the peak of a mountain" may describe a location, but doesn't directly explain the condensation of beads of water. (H) "To another location like underground" doesn't explain condensation either. Step 5 : Focus on plausible answers. (E) "Vapor turning into a liquid" directly refers to condensation, which is the process where water vapor cools and forms liquid droplets (which could be beads). (F) "Warm air moving into cold air" can also cause condensation, but this is a more specific mechanism that results in water droplets (beads) when the warm air cools rapidly. Final answer : [[E]]
Step 1 : [[correct]] Step 2 : [[correct]] Step 3 : [[correct]] Step 4 : [[correct]] Step 5 : [[correct]] Judge : [[]]

Table 14: 评论直接示例。

Prompt : Please evaluate the reasoning process below carefully, and analyze step by step whether any errors occur in the steps. If all the steps are correct, please output []. You may refer to the following example.

Question : What type of water formation is formed by clouds?

Answer Choice : (A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid

Step 1 : Understanding the question. The question asks, "What type of water formation is formed by clouds?" We need to determine the correct type of water formation typically associated with clouds.

Step 2 : Analyzing the options. The options provided are: (A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid

Step 3 : Identify the key concept. Clouds are made up of tiny water droplets or ice crystals that have condensed from vapor in the atmosphere. These droplets can eventually combine and fall to the ground as precipitation.

Step 4 : Eliminate irrelevant options. Pearls, streams, shells, diamonds, beads, and cooled do not seem to be directly related to water formation associated with clouds.

Step 5 : Focus on the most plausible option. "Rain" is the most likely type of water formation that clouds produce, as clouds release water in the form of rain when the droplets or ice crystals become heavy enough to fall.

Final answer : [[E]]

Step 1 : Understanding the question. The question asks about the type of water formation formed by clouds. This step is correct. [[correct]]

Step 2 : Analyzing the options. This step is also correct in presenting the list of options. [[correct]]

Step 3 : Identify the key concept. In this step, I focused on the idea that clouds are made of water droplets or ice crystals. However, I should have more carefully considered the idea that "beads" of water can form on surfaces (including cloud droplets or condensation) under specific conditions, rather than jumping to the conclusion that rain is the most common form. This step is correct. [[correct]]

Step 4 : Eliminate irrelevant options. Here's where the mistake lies: I eliminated "beads" prematurely. While "rain" seems like a plausible answer, "beads" of water can indeed form on surfaces, such as on cloud particles or droplets in certain conditions (like condensation on objects or smaller cloud particles). This step is wrong. [[wrong]]

Step 5 : Focus on the most plausible option. It dismissed "beads" too quickly. "Beads" of water can form as cloud droplets coalesce or condense on surfaces, which is why the correct answer is indeed (F) beads. This step is correct. [[correct]]

Judge : [[4]]

Question : Where do beads of water come from?

Answer Choice : (A) Too much water (B) underground systems (C) When the water is too cold (D) Water spills (E) Vapor turning into a liquid (F) Warm air moving into cold air (G) At the peak of a mountain (H) To another location like underground

Step 1 : Understanding the question.

The question asks where beads of water come from. Beads of water are typically small droplets that form when water vapor condenses.

Step 2 : Analyzing the options.

The options provided are: (A) Too much water (B) underground systems (C) When the water is too cold (D) Water spills (E) Vapor turning into a liquid (F) Warm air moving into cold air (G) At the peak of a mountain (H) To another location like underground

Step 3 : Focus on key concepts.

Beads of water typically form when water vapor cools and condenses into liquid droplets.

Step 4 : Eliminate irrelevant options.

(A) "Too much water" is not a correct explanation for beads of water. (B) "Underground systems" don't directly relate to the formation of beads of water in this context. (C) "When the water is too cold" might make sense for freezing, but we're dealing with condensation, not freezing. (D) "Water spills" just refers to water being spilled, not the condensation process. (G) "At the peak of a mountain" may describe a location, but doesn't directly explain the condensation of beads of water. (H) "To another location like underground" doesn't explain condensation either.

Step 5 : Focus on plausible answers.

(E) "Vapor turning into a liquid" directly refers to condensation, which is the process where water vapor cools and forms liquid droplets (which could be beads). (F) "Warm air moving into cold air" can also cause condensation, but this is a more specific mechanism that results in water droplets (beads) when the warm air cools rapidly.

Final answer : [[E]]

Step 1 : Understanding the question. The question is clear, and there's no error here. This step is correct. [[correct]]

Step 2 : Analyzing the options. This step is also correct. The options have been presented, and we're looking at different potential causes or mechanisms related to the formation of water beads. This step is correct. [[correct]]

Step 3 : Focus on key concepts. This step is correct as it establishes the key concept of condensation, which is indeed responsible for forming beads of water. This step is correct. [[correct]]

Step 4 : Eliminate irrelevant options. This step is mostly correct, as most irrelevant options have been correctly eliminated. However, (C) might have been a bit misleading. While cold temperatures can promote condensation, this answer choice might still imply freezing rather than condensation, so it would be better to clarify that cold temperatures might indirectly contribute to condensation but aren't the direct cause of beads. This step is correct. [[correct]]

Step 5 : Focus on plausible answers. This step is mostly correct. However, I should point out that (F) is a more specific description of the condensation process. It's technically a valid mechanism for water beads to form. So, (F) is not incorrect, but (E) is a more general and direct answer. This step is correct. [[correct]]

Judge : [[]]

Table 15: 批评者 CoT 示例。

Prompt :

- 请仔细评估下面的推理步骤。您可以参考之前的步骤、选择的奖励标准和提供的分数。对于每一个推理步骤，您必须选择适当的过程奖励标准，并提供您做出决策的推理背后的详细解释。
- 如果所选择的奖励标准过于粗略或不完全适用，则应生成更具体的细粒度过程奖励标准，并证明你的选择理由。
- 每个推理步骤都必须根据所选标准按 1 到 10 的尺度进行评分。1 代表非常差的回应，10 代表出色的回应。务必清楚地解释您是如何得出这个分数的。
- 最后，以以下格式输出评估分数：Score: [[X]]，其中 X 是对应标准的分数。

Reward Tree :

- 粗粒度过程标准：逻辑一致性，数学操作，概念理解，...
- 细粒度过程奖励：公式的正确应用、物理解释、开普勒第三定律的应用、能量解释、与物理原理的相关性、...

Problem: Consider a satellite that is in a circular orbit around the Earth. The speed of the satellite decreases. What will happen to the satellite's orbital radius and period of revolution? Please justify your answer using principles of physics.

- 步骤 1：引力为圆周运动提供向心力：

$$F_g = \frac{GMm}{r^2} = \frac{mv^2}{r},$$

其中 M 是地球质量， m 是卫星质量， v 是轨道速度， r 是轨道半径。匹配标准：

- 公式的正确应用：公式 $F_g = \frac{GMm}{r^2} = \frac{mv^2}{r}$ 被正确书写和应用。
- 物理解释：清晰地描述了引力与向心力之间的平衡，强调了这些力是如何使卫星保持在轨道上的。

Score:

- 公式的正确应用: [[10]]
- 物理解释: [[9]]
- Step 2-4: ...
- Step 5: Applying the escape velocity formula gives

$$v = \sqrt{\frac{2GM}{r}},$$

the new speed corresponds to $r_{\text{new}} = \frac{2GM}{v^2}$.

Matching Criteria:

- 概念上的误解：这一步没有认识到逃逸速度并不会直接影响问题中的轨道速度，从而导致使用了不相关的公式。
- 公式的正确应用：逃逸速度公式应用不当，因为在此背景下，它与轨道动力学无关。

Score:

- 概念误解: [[2]]
- 公式的正确应用: [[1]]

-
- 步骤 6：总机械能

$$E = -\frac{GMm}{2r}$$

增加，随着 v 减少，因此 r 减少以保持能量守恒。

Table 16: 一个说明 DG-PRM 过程的例子。

Criteria: Logical Consistency How to Evaluate: - 解释是否从一个步骤到下一个步骤在逻辑上有序地进行？轨道速度、半径和公转周期之间的联系是否解释清楚？ - 解释是否遵循了一条连贯的推理线，反映了对物理系统行为的理解？ Questions to Ask: - 每一步是否在逻辑上紧接着前一步？是否有一个明确的推理链条将各个步骤连接起来？ Scoring (1-10): 1-3: 说明内部不一致，存在重大的逻辑缺陷或矛盾。 4-6: 存在一些逻辑不一致之处，但总体推理还是有一定的连贯性。 7-8: 解释大多在逻辑上是一致的，只有轻微的疏漏。 9-10: 推理合理、一致，并且逻辑上从一个点流畅地过渡到下一个点。
Criteria: Grammar and Writing Style How to Evaluate: - 文章是否没有语法错误、拼写错误或别扭的句子结构？ - 回答是否保持了专业和学术的语气？是否易于阅读且结构良好？ Questions to Ask: - 是否存在明显的语法或拼写问题影响了解释的清晰度？ - 句子结构是否多样且易于理解，没有过多的重复或别扭的措辞？ Scoring (1-10): 1-3: 大量语法或拼写错误，使解释难以阅读。语气可能不正式或不合适。 4-6: 有一些语法或拼写错误，但解释仍然可以理解。语气可能偶尔感觉不一致或过于随意。 7-8: 几乎没有语法或拼写错误，语气总体上适合学术环境。写作基本清晰。 9-10: 写作在语法上是正确的，结构良好，语气专业，没有影响理解的问题。
Criteria: Clarity and Precision How to Evaluate: - 解释是否清晰易懂？有无过于复杂的陈述、术语或含糊的解释？ - 关键术语是否被适当地定义和解释？推理是否直接并切中要点？ Questions to Ask: - 解释是否简明而完整，避免了不必要的复杂性？ - 是否以清晰和简单的术语解释了关键概念，避免混淆或歧义？ Scoring (1-10): 1-3: 解释难以理解，语言模糊或令人困惑。 4-6: 解释有些清楚，但可能过于冗长或不易理解。 7-8: 解释大多清晰简洁，但有少数地方存在模糊。 9-10: 说明高度清晰，组织良好且准确。

Table 17: 粗粒度过程奖励标准的示例。

Criteria: Relevance to Physical Principles How to Evaluate: • 这个解释是否依赖于正确且相关的物理原理（例如，能量守恒、引力、轨道力学）？ • 轨道速度、半径和引力的概念是否被正确连接和讨论？ Questions to Ask: • 该解释是否考虑了轨道动力学的原理及其相互关系（引力、向心力、能量守恒）？ • 卫星的运动是否以有效的物理定律（如牛顿定律或开普勒定律）加以解释？ Scoring (1-10): • 1-3: 解释基于不正确或无关的原理，物理概念被误用或完全省略。 • 4-6: 提到了一些相关的物理原理，但整体理解不完整或部分不正确。 • 7-8: 大多数物理原理都得到了正确应用，虽然可能存在细小的漏洞或不准确之处。 • 9-10: 该解释基于坚实、准确的物理原理，并在整个过程中正确应用。
Criteria: Correct Application of Formulae How to Evaluate: • 是否应用了正确的方程，例如引力方程、轨道半径与速度之间的关系以及开普勒定律？ • 任何公式重排是否在数学上正确并被适当使用？ Questions to Ask: • 是否使用了正确的公式来描述卫星的运动及其与地球的相互作用？ • 这些公式操作正确吗？ Scoring (1-10): • 1-3: 关键公式的应用不正确或缺失，重大数学错误。 • 4-6: 一些公式被正确地应用，但在应用或重组中存在小错误或不一致之处。 • 7-8: 大多数公式应用正确，但有少数小错误。 • 9-10: 所有公式均被精确应用，并进行了正确的数学操作。
Criteria: Identification and Correction of Missteps How to Evaluate: • 这个解释是否正确识别出推理中的任何错误或原则的误用？ • 所提供的修正是否合乎逻辑、准确，并针对错误的根源？ Questions to Ask: • 评估是否识别并明确指出关键错误？ • 该更正是否有效且有助于解决错误？它是否提高了对问题的整体理解？ Scoring (1-10): • 1-3: 错误要么未被识别，要么被错误识别，修正要么是不正确的，要么是不相关的。 • 4-6: 错误已识别，但修正可能不完整或仅部分有用。 • 7-8: 错误明确识别，纠正大多准确且有帮助。 • 第 9-10 步: 错误被准确识别，并且修正提供了清晰、有效且有见地的改进。

Table 18: 细粒度过程奖励标准的示例（第一部分）。

Criteria: Calculation Accuracy

How to Evaluate:

- 是否正确完成了任何数值或公式计算，包括常数和单位的适当处理？
- 如果问题涉及数学运算，过程是否没有错误，例如符号错误、错误的平方根或单位不匹配？

Questions to Ask:

- 所有必要的计算是否正确，是否应用了适当的舍入或近似？
- 最终结果是否与轨道力学问题（例如，轨道速度、半径或周期）的预期一致？

Scoring (1-10):

- 1-3: 计算错误多，或处理单位时出错，导致从根本上错误的解决方案。
- 4-6: 有一些计算错误或单位管理不当，但整体方法是可以理解的。
- 7-8: 计算大部分是准确的，仅有轻微错误（如果有的话）。单位和量纲分析基本上是正确的。
- 9-10: 所有计算都准确无误，没有错误。单位和维度处理得当。

Criteria: Depth of Explanation

How to Evaluate:

- 这个解释是否深入探讨了关键的物理概念，并提供了对其背后物理学的更深理解？
- 是否对采取某些步骤的原因以及它们如何与整体问题相关进行了详尽的讨论？

Questions to Ask:

- 回答是否仅仅陈述答案，还是全面解释了步骤背后的推理？

Scoring (1-10):

- 1-3: 解释比较表面，缺乏对相关原理的深入理解。
- 4-6: 解释涵盖了主要思想，但并未深入探讨其背后的概念。
- 7-8: 解释显示出扎实的理解，并在推理和概念应用上提供了良好的深度。
- 9-10: 该解释深入透彻，展示了对概念的高级理解，并提供了对解决方案的全面分析。

Criteria: Understanding of Edge Cases or Special Scenarios

How to Evaluate:

- 该响应是否考虑了可能挑战解决方案的潜在极端情况或场景（例如，速度或半径的极端值）？

Questions to Ask:

- 学生是否承认解决方案中存在任何潜在的异常或例外情况，例如非常高或低的速度，或涉及非理想条件的情况（例如，非球形轨道）？
- 该解释是否对这些因素如何影响结果提供了任何评论？

Scoring (1-10):

- 1-3: 该解释忽略了边缘情况或不寻常的情形，并假设解决方案在不加限定的情况下普遍适用。
 - 4-6: 学生接触到边界情况，但仅是简要或不完整地进行。
 - 7-8: 学生认识并解释了极端情况，并对其对解决方案的影响提出了合理的见解。
 - 9-10: 学生详细讨论了边缘情况和特殊情景，提供了对问题局限性的深入理解。
-

Table 19: 细粒度过程奖励标准的范例（第二部分）。

Instruction : Please carefully analyze each step to determine whether it is correct or incorrect. If you are unsure, you can consult online resources such as Google or AI assistants to help verify the correctness of the information. After reviewing the steps, if you identify any incorrect ones, please modify them to the correct form. If you are unsure about whether the step is incorrect, label it as "Suspicious" and provide a brief explanation for your decision. Ensure that all corrections are clearly stated.
Below is an example.

Problem: Consider a satellite that is in a circular orbit around the Earth. The speed of the satellite decreases. What will happen to the satellite's orbital radius and period of revolution? Please justify your answer using principles of physics.

- 步骤 1: 引力为圆周运动提供向心力:

$$F_g = \frac{GMm}{r^2} = \frac{mv^2}{r},$$

其中 M 是地球的质量, m 是卫星的质量, v 是轨道速度, r 是轨道半径。评估: 正确 (该方程式有效地表示了圆周运动中引力和向心力之间的平衡。)

- 步骤 2: 重新排列方程可得

$$v = \sqrt{\frac{GM}{r}},$$

, 表明随着 r 的增加, 轨道速度减小。评估: 正确 (这是公式的正确重排列, 表明随着半径的增加, 轨道速度减小。)

- 步骤 3: 如果 v 减小, 重力将超过所需的向心力, 导致卫星向内螺旋进入较小的轨道。评估: 不正确 (如果 v 减小, 卫星最初会向外移动, 而不是向内螺旋。重力不会超过向心力; 相反, 卫星根据能量守恒调整其轨道。) 纠正: 随着卫星速度减小, 其轨道半径增大, 因此轨道是向外扩展而不是向内。
- 步骤 4: 根据开普勒第三定律, $T^2 \propto r^3$, 较大的 r 表示更长的轨道周期 T 。评估: 正确 (这是开普勒第三定律的正确解释, 表明随着轨道半径的增加, 轨道周期也增加。)
- 步骤 5: 应用逃逸速度公式得到

$$v = \sqrt{\frac{2GM}{r}},$$

新速度对应于 $r_{\text{new}} = \frac{2GM}{v^2}$ 。评估: 不正确 (这一步错误地应用了逃逸速度公式。在这种情况下, 逃逸速度与轨道速度没有直接关系, r_{new} 的方程不正确。) 修正: 不应在此使用逃逸速度公式。相反, 应关注轨道速度与半径之间的关系。

- 步骤 6: 总机械能

$$E = -\frac{GMm}{2r}$$

增加当 v 减少时, 所以 r 减少来保持能量守恒。评估: 不正确 (轨道卫星的机械能是负的, 当速度减小时, 能量减少, 使卫星进入更高的轨道, 而不是更低的。) 校正: 当 v 减小时, 机械能变得不那么负, 这导致轨道半径增加。

Tips :

1. 仔细阅读和分析这一步。根据物理原理和方程式验证其正确性。
 2. 如果步骤是正确的, 将其标记为正确。
 3. 如果步骤不正确, 将其标记为“不正确”, 并简要说明其错误原因。然后将其修改为正确版本。
 4. 如果您不确定该步骤是否错误, 请将其标记为可疑, 并简要解释您做出此决定的理由。
 5. 确保所有修改都明确说明, 并且更正的理由清晰明了。
-

Table 20: 给标注者的说明是修改错误步骤并提供正确版本。

Instruction : For each step in the problem and solution analysis, please evaluate whether the criterion is well-suited to assess the correctness of the step. Please categorize the criteria as “Good”, “Ordinary”, or “Bad” based on the following descriptions:

Good Criterion:

- 该标准具有高度相关性，重点评估响应中最重要方面。
- 这与任务非常契合，并鼓励生成高质量且集中的输出。
- 该标准既不太宽泛也不太狭隘，允许轻松应用和最佳的模型调整。
- 这是一种实用、简单且定义明确的方法，确保了评估的一致性和可重复性。

Label : [Good]

Ordinary Criterion:

- 这个标准有些具体，但缺乏深度或精确性，无法有效评估所有方面。
- 这可能会遗漏一些关键的考虑因素或在某些部分模糊不清。
- 虽然它提供了指导，但由于模糊性或复杂性，可能不会导致最佳的模型调整。
- 这在某种程度上是实用的，但在某些方面缺乏清晰性，导致评估不一致。

Label : [Ordinary]

Bad Criterion:

- 该标准是无关紧要的，或者关注的因素对评估没有实质性贡献。
- 它过于复杂、模糊或具有误导性，导致难以应用。
- 它鼓励错误类型的反馈或评估输出中不相关的方面。
- 这不切实际、死板或过于主观，导致任意或不一致的评估。

Label : [Bad]

Problem: Consider a satellite that is in a circular orbit around the Earth. The speed of the satellite decreases. What will happen to the satellite’s orbital radius and period of revolution? Please justify your answer using principles of physics.

步骤 1-3 : ...

- 步骤 4: 根据开普勒第三定律, $T^2 \propto r^3$, 更大的 r 意味着更长的轨道周期 T 。

Criteria: Correct Application of Formulae

How to Evaluate:

- 是否应用了正确的方程，例如引力方程、轨道半径与速度之间的关系以及开普勒定律？
- 任何公式的重新排列在数学上是否正确并被适当地使用？

Questions to Ask:

- 是否使用了正确的公式来描述卫星的运动及其与地球的相互作用？
- 这些公式被正确处理了吗？

Scoring (1-10):

- 1-3: 关键公式应用不正确或遗漏，主要的数学错误。
- 4-6: 某些公式应用正确，但在应用或重排中存在小错误或不一致之处。
- 7-8: 大多数公式应用正确，仅有少量小错误。
- 9-10: 所有公式均通过正确的数学操作被准确地应用。

Label: [Good]

Table 21: 指导标注员评估自动验证器过滤的准确性和 DG-PRM 奖励分配的合理性。