
Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov^{*†}

Archit Sharma^{*†}

Eric Mitchell^{*†}

Stefano Ermon^{†‡}

Christopher D. Manning[†]

Chelsea Finn[†]

[†]Stanford University [‡]CZ Biohub

{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Abstract

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址 https://github.com/binary-husky/gpt_academic/。项目在线体验地址 <https://auth.gpt-academic.top/>。当前大语言模型: glm-4，当前语言模型温度设定: 0.5。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

尽管大规模无监督语言模型（LMs）学习了广泛的世界知识和一些推理技能，但由于其训练完全无监督的性质，实现对它们行为的精确控制是困难的。现有方法为了获得这种可控性，收集人类对模型生成的相对质量的标签，并使用强化学习从人类反馈（RLHF）对无监督LM进行微调，以与这些偏好保持一致。然而，RLHF是一个复杂且通常不稳定的流程，首先要拟合一个反映人类偏好的奖励模型，然后使用强化学习微调大型无监督LM以最大化这个估计的奖励，同时不偏离原始模型太远。在本文中，我们引入了RLHF中奖励模型的一种新参数化方法，可以封闭形式提取相应的最优策略，从而使我们能够仅用简单的分类损失来解决标准的RLHF问题。我们提出的算法称为直接偏好优化（DPO），它稳定、性能好且计算成本低，无需在微调期间从LM中采样或进行大量的超参数调整。我们的实验表明，DPO可以像现有方法一样好甚至更好地微调LM以与人类偏好保持一致。值得注意的是，使用DPO进行微调在控制生成情感的能力上超过了基于PPO的RLHF，并且在总结和单轮对话中与或提高了响应质量，同时实现起来要简单得多，训练也更为容易。

1 Introduction

大规模的无监督语言模型（LMs）在非常庞大的数据集上进行训练，获得了令人惊讶的能力 [11, 7, 40, 8]。然而，这些模型是在由具有各种目标、优先级和技能的人类生成数据上进行训练的。其中一些目标和技能可能并不值得我们模仿；例如，尽管我们可能希望人工智能

^{*}Equal contribution; more junior authors listed earlier.

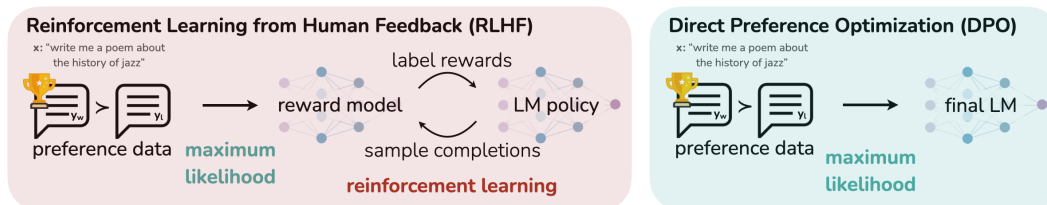


图 1: DPO 在避免强化学习的同时优化人类偏好。现有的用人类反馈微调语言模型的方法首先将奖励模型适配到一个由提示和人类对响应对偏好的数据集，然后使用RL找到一个最大化学习到的奖励的策略。相比之下，DPO 直接用简单的分类目标优化最符合偏好的策略，适配一个隐式奖励模型，其对应的最优策略可以通过闭式提取。

编码助手能够理解常见的编程错误以便进行纠正，但是，在生成代码时，我们还是希望让模型倾向于其训练数据中存在的（可能罕见的）高质量编程能力。同样，我们可能希望语言模型能够意识到被50%的人所相信的一个常见误解，但我们当然不希望模型在50%的关于这个问题的查询中声称这个误解是正确的！换句话说，从模型的非常广泛知识和能力中选择模型的期望响应和行为对于构建安全、高效和可控的AI系统至关重要[26]。尽管现有方法通常使用强化学习（RL）来引导LMs以符合人类偏好，我们将展示，现有方法中使用的基于RL的目标可以用一个简单的二元交叉熵目标精确优化，极大地简化了偏好学习流程。

在较高层面上，现有方法通过精心策划的一组人类偏好数据，将期望行为植入语言模型中，这些偏好代表了人类认为安全有益的行为类型。这种偏好学习阶段发生在对大型文本数据集进行初始的大规模无监督预训练之后。尽管最直接的偏好学习方法是在高质量回应的人类演示上进行监督微调，但最成功的方法类别是从人类（或AI）反馈的强化学习（RLHF/RLAIF; [12, 2]）。RLHF方法拟合一个奖励模型到一个人类偏好数据集，然后用强化学习来优化语言模型策略，以产生被赋予高奖励的回应，同时不过度偏离原始模型。尽管RLHF产生了在对话和编程能力方面令人印象深刻的模型，但RLHF管道比监督学习复杂得多，涉及训练多个LM并在训练过程中从LM策略中进行采样，从而产生显著的计算成本。

在本文中，我们展示了如何直接优化语言模型以符合人类偏好，无需明确的奖励建模或强化学习。我们提出了 *Direct Preference Optimization (DPO)*，一种算法，它隐式地优化了与现有RLHF算法相同的目标（在KL散度约束下的奖励最大化），但易于实现且训练简单。直观地说，DPO更新增加了偏好回应相对于不偏好回应的相对对数概率，但它融合了一种动态的、每个示例的重要性权重，以防止我们发现简单的概率比目标下发生的模型退化。与现有算法一样，DPO依赖于一个理论偏好模型（如Bradley-Terry模型; [5]），该模型测量给定的奖励函数与实证偏好数据的一致性程度。然而，尽管现有方法使用偏好模型来定义一个偏好损失以训练奖励模型，然后训练一个优化学习到的奖励模型的策略，DPO使用变量变换直接将偏好损失定义为关于策略的函数。给定一个关于模型回应的人类偏好数据集，DPO因此可以使用简单的二元交叉熵目标优化策略，生成符合偏好数据拟合的隐式奖励函数的最优策略。

我们的主要贡献是Direct Preference Optimization (DPO)，一个用于从偏好训练语言模型的简单的无RL算法。我们的实验表明，DPO至少与现有方法一样有效，包括基于PPO的RLHF，用于在诸如情感调节、总结和对话等任务中从偏好学习，使用具有高达6B参数的语言模型。

2 Related Work

不断增加规模的自我监督语言模型学会通过零样本学习[31]或少量样本提示[6, 25, 11]来完成某些任务。然而，通过在包含指令和人工撰写完成内容的数据集上进行微调，可以显著提高它们在下流任务上的表现和对用户意图的对齐[23, 36, 13, 39]。这种“指令调优”过程使得大型语言模型能够泛化到指令调优集之外的指令，并普遍提高其可用性[13]。尽管指令调优取得了成功，相对的人类对响应质量的判断通常比收集专家演示更容易，因此后续工作使用包含人类偏好的数据集对大型语言模型进行微调，提高了在翻译[18]、总结[38, 49]、讲故事[49]和遵循指令[26, 32]方面的熟练度。这些方法首先在偏好模型（如Bradley-Terry模型[5]）下与偏好数据集兼容的神经网络的奖励函数进行优化，然后使用强化学习算法（通常是REINFORCE[45]、近端策略优化（PPO[37]）或其变体）对语言模型进行微调以最大化给定的奖励[32]。相关的一条研究路线利用根据人工反馈微调的大型语言模型来生成针对特定属性（如安全性或无害性）的附加合成偏好数据[2]，仅从人类提供的文本评分准则形式进行弱监督。这些方法结合了两类工作的成果：一类是针对多种目标使用强化学习训练语言模型的工作[33, 27, 46]，另一类是学习人类偏好的通用方法[12, 19]。

尽管使用相对人类偏好很有吸引力，但使用强化学习对大型语言模型进行微调仍然是一个主要实践挑战；这项工作提供了一种理论上合理的无需RL优化相对偏好的方法。

在语言环境之外，从偏好中学习策略已经在老虎机和强化学习设置中被研究，并提出了一些方法。使用偏好评定或者动作排名而不是奖励的上下文老虎机学习被称为上下文决斗老虎机（CDB[48, 14]）。在没有绝对奖励的情况下，CDB的理论分析用冯·诺伊曼胜者的概念替代了最优策略，即该策略对抗任何其他策略的预期胜率至少为50%[14]。然而，在CDB设置中，偏好标签是在线的，而在学习人类偏好时，我们通常从固定的离线偏好标注动作对批次中学习[47]。类似地，基于偏好的强化学习（PbRL）从由未知的“评分”函数生成的二元偏好中学习，而不是奖励[9, 35]。PbRL存在多种算法，包括可以重用非策略偏好数据的方法，但通常首先需要显式估计潜在评分函数（即奖励模型），然后对其进行优化[16, 9, 12, 34, 19]。我们则提出了一种单一阶段的策略学习方法，它直接优化满足偏好的策略。

3 Preliminaries

我们回顾了Ziegler et al.（以及后来的[38, 1, 26]）的RLHF管道。它通常包括三个阶段：1）监督微调（SFT）；2）偏好采样与奖励学习；3）强化学习优化。

SFT: RLHF通常首先通过在高质量数据上用监督学习微调预训练的LM，以获得针对感兴趣的下游任务（对话、总结等）的模型 π^{SFT} 。

奖励建模阶段: 在第二阶段，SFT模型被提示输入 x 以生成答案对 $(y_1, y_2) \sim \pi^{\text{SFT}}(y | x)$ 。这些答案对随后呈现给人类标注者，他们表达对其中一个答案的偏好，表示为 $y_w \succ y_l | x$ ，其中 y_w 和 y_l 分别代表在 (y_1, y_2) 中偏好和不偏好的完成。这些偏好被认为是由某个潜在的奖励模型 $r^*(y, x)$ 生成的，我们无法访问这个模型。有多种建模偏好的方法，Bradley-Terry（BT）[5]模型是一个流行的选择（尽管如果我们能访问到几个排序列的答案，更一般的Plackett-Luce排名模型[30, 21]也与该框架兼容）。BT模型规定，人类偏好分布 p^* 可以写成：

$$p^*(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}. \quad (1)$$

假设我们可以访问一个从 p^* 中抽取的静态比较数据集 $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ ，我们可以用一个奖励模型 $r_\phi(x, y)$ 进行参数化，并通过最大似然估计来估计参数。将问题构建为一个二

分类问题，我们得到负对数似然损失：

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))] \quad (2)$$

其中 σ 是逻辑函数。在LMs的背景下，网络 $r_\phi(x, y)$ 通常从SFT模型 $\pi^{\text{SFT}}(y | x)$ 初始化，并在最后的转换层之上添加一个线性层，以产生一个单一的标量预测值作为奖励值 [49]。为了确保具有较低方差的奖励函数，先前的工作对奖励进行了归一化，使得对于所有的 x ，都有 $\mathbb{E}_{x, y \sim \mathcal{D}} [r_\phi(x, y)] = 0$ 。

RL微调阶段：在RL阶段，我们使用学习到的奖励函数为语言模型提供反馈。特别是，我们制定以下优化问题

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)] - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(y | x) \parallel \pi_{\text{ref}}(y | x)] \quad (3)$$

其中 β 是一个控制偏离基础参考策略 π_{ref} 的参数，即最初的SFT模型 π^{SFT} 。在实际应用中，语言模型策略 π_θ 也会初始化为 π^{SFT} 。所增加的约束非常重要，因为它防止模型偏离奖励模型准确分布太远，同时保持生成多样性，并防止模式崩溃为单一的高奖励答案。由于语言生成的离散性质，这一目标不可微，并且通常使用强化学习进行优化。标准方法 [49, 38, 1, 26] 是构建奖励函数 $r(x, y) = r_\phi(x, y) - \beta(\log \pi_\theta(y | x) - \log \pi_{\text{ref}}(y | x))$ ，并使用PPO [37] 进行最大化。

4 Direct Preference Optimization

受到在大型问题上应用强化学习算法（如微调语言模型）的挑战的启发，我们的目标是推导一种直接使用偏好进行策略优化的简单方法。与先前的RLHF方法不同，这些方法学习一个奖励然后通过RL优化它，我们的方法利用了一种特殊的奖励模型参数化选择，使得可以在闭式形式中提取其最优策略，无需RL训练循环。正如我们接下来将详细描述的那样，我们的关键洞察是利用从奖励函数到最优策略的解析映射，这使我们能够将关于奖励函数的损失函数转换为关于策略的损失函数。这种变量替换方法避免拟合一个显式独立的奖励模型，同时仍然在现有的人类偏好模型（如Bradley-Terry模型）下进行优化。实质上，策略网络既代表了语言模型也代表了(隐含)奖励。

推导DPO目标。我们从与先前工作相同的RL目标开始，即方程 3，在一般奖励函数 r 下。遵循先前的工作 [29, 28, 17, 15]，可以很容易地证明，方程 3中KL约束下的奖励最大化目标的最优解具有以下形式：

$$\pi_r(y | x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y | x) \exp\left(\frac{1}{\beta} r(x, y)\right), \quad (4)$$

其中 $Z(x) = \sum_y \pi_{\text{ref}}(y | x) \exp\left(\frac{1}{\beta} r(x, y)\right)$ 是分区函数。完整的推导过程请参见附录 A.1。即使我们使用地面真实奖励函数 r^* 的MLE估计 r_ϕ ，估算分区函数 $Z(x)$ 仍然代价高昂[17, 15]，这使得这种表示在实践中难以利用。然而，我们可以重新排列方程 4，用相应的最优策略 π_r 、参考策略 π_{ref} 以及未知的分区函数 $Z(\cdot)$ 来表达奖励函数。具体来说，我们首先对方程 4的两边取对数，然后通过一些代数运算我们得到：

$$r(x, y) = \beta \log \frac{\pi_r(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z(x). \quad (5)$$

我们可以将这种重参数化技术应用于真实奖励 r^* 及其对应的最优模型 π^* 。幸运的是，Bradley-Terry 模型仅依赖于两个结果之间的奖励差异，即 $p^*(y_1 \succ y_2 | x) = \sigma(r^*(x, y_1) - r^*(x, y_2))$ 。将方程 5 中的重参数化结果代入到偏好模型

方程 1 中的 $r^*(x, y)$ ，分区函数将会抵消，我们可以仅通过最优策略 π^* 和参考策略 π_{ref} 来表达人类偏好概率。因此，在Bradley-Terry模型下的最优RLHF策略 π^* 满足偏好模型：

$$p^*(y_1 \succ y_2 | x) = \frac{1}{1 + \exp \left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} \right)} \quad (6)$$

推导过程在附录 A.2中。虽然方程 6使用了Bradley-Terry模型，但我们可以类似地在更一般的Plackett-Luce模型下推导表达式 [30, 21]，具体见附录 A.3。

现在我们有了解人类偏好数据的概率，它是关于最优策略而非奖励模型，因此我们可以为参数化策略 π_θ 制定一个最大似然目标。类似于奖励建模方法（即方程 2），我们的策略目标变为：

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (7)$$

这种方法，我们使用一种替代的参数化方式拟合了一个隐式奖励，其最优策略就是 π_θ 。此外，由于我们的过程等同于拟合了一个重新参数化的Bradley-Terry模型，它在适当假设偏好数据分布的情况下，享有一定的理论性质，如一致性[4]。在第 5 节中，我们进一步讨论了DPO与其他工作的理论性质。

DPO更新做了什么？ 为了从机制上理解DPO，分析损失函数 $\mathcal{L}_{\text{DPO}}$ 的梯度是有用的。关于参数 θ 的梯度可以写成：

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \left[\underbrace{\nabla_\theta \log \pi(y_w | x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l | x)}_{\text{decrease likelihood of } y_l} \right] \right],$$

其中 $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ 是由语言模型 π_θ 和参考模型 π_{ref} （更多内容见第 5 节）隐式定义的奖励。直观地，损失函数 $\mathcal{L}_{\text{DPO}}$ 的梯度会增加优选完成 y_w 的概率，并减少不优选完成 y_l 的概率。重要的是，示例根据隐式奖励模型 $\hat{r}_\theta$ 对不优选完成的评价有多高进行加权，这个评价通过 β 进行缩放，即，隐式奖励模型对完成情况进行排序的错误程度，考虑了KL约束的强度。我们的实验表明这种加权的重要性，因为没有加权系数的这种方法的朴素版本可能会导致语言模型退化（附录表 3）。

DPO概述。 一般的DPO流程如下：1) 对每个提示 x 进行采样完成情况 $y_1, y_2 \sim \pi_{\text{ref}}(\cdot | x)$ ，用人类偏好进行标注，构建偏好离线数据集 $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ ；2) 优化语言模型 π_θ 以最小化给定的 π_{ref} 和 $\mathcal{D}$ 以及期望的 β 的 $\mathcal{L}_{\text{DPO}}$ 。在实践中，人们更愿意重用公开可用的偏好数据集，而不是生成样本和收集人类偏好。由于偏好数据集是使用 π^{SFT} 进行采样的，因此只要可用，我们就初始化 $\pi_{\text{ref}} = \pi^{\text{SFT}}$ 。然而，当 π^{SFT} 不可用时，我们通过最大化优选完成情况 (x, y_w) 的似然来进行 π_{ref} 的初始化，即 $\pi_{\text{ref}} = \arg \max_\pi \mathbb{E}_{x, y_w \sim \mathcal{D}} [\log \pi(y_w | x)]$ 。这个程序有助于缓解真实参考分布（不可用）和DPO使用的 π_{ref} 之间的分布偏移。关于实现和超参数的更多细节可以在附录 B 中找到。

5 Theoretical Analysis of DPO

在本节中，我们将对DPO方法进行进一步的解释，提供理论支持，并将DPO的优势与用于RLHF的演员-评论家算法（如PPO [37]）的问题相关联。

5.1 Your Language Model Is Secretly a Reward Model

DPO能够绕过既要拟合一个显式奖励又执行RL来学习单一最大似然目标下的策略。注意优化目标方程5等价于一个带有奖励参数化 $r^*(x, y) = \beta \log \frac{\pi_\theta^*(y|x)}{\pi_{\text{ref}}(y|x)}$ 的Bradley-Terry模型，并且

我们优化我们的参数化模型 π_θ ，在变量变换下等价于方程2中的奖励模型优化。在本节中，我们将构建这种重参数化背后的理论，展示它并不限制学习到的奖励模型的类别，并允许完全恢复最优策略。我们从定义奖励函数之间的等价关系开始。

Definition 1. 我们称两个奖励函数 $r(x, y)$ 和 $r'(x, y)$ 是等价的，当且仅当存在某个函数 f 使得 $r(x, y) - r'(x, y) = f(x)$ 。

很容易看出这确实是一个等价关系，它将奖励函数集合划分为若干类别。我们可以陈述以下两个引理：

Lemma 1. 在Plackett-Luce框架下，尤其是Bradley-Terry框架下，同一类中的两个奖励函数会诱导出相同的偏好分布。

Lemma 2. 两个属于同一等价类的奖励函数在受限强化学习问题下诱导出相同的最佳策略。

证明是直接的，我们将其推迟到附录 A.5 中。第一个引理是关于Plackett-Luce模型族的一个众所周知的欠定问题 [30]。由于这种欠定，我们通常必须施加额外的可辨识性约束，以实现从方程 2 的MLE估计的任何保证 [4]。第二个引理指出，同一类的所有奖励函数都产生相同的最佳策略，因此对于我们的最终目标，我们只对从最佳类别中恢复任意的奖励函数感兴趣。我们在附录 A.6 中证明了以下定理：

Theorem 1. *Under mild assumptions, all reward classes consistent with the Plackett-Luce (and Bradley-Terry in particular) models can be represented with the reparameterization $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ for some model $\pi(y | x)$ and a given reference model $\pi_{\text{ref}}(y | x)$.*

Proof Sketch. 考虑任何奖励函数 $r(x, y)$ ，它诱导出一个相应的最优模型 $\pi_r(y | x)$ ，由方程 4 指定。我们将展示，可以使用上述重新参数化的方式来表示 r 等价类中的奖励函数。我们定义投影 f 为

$$f(r; \pi_{\text{ref}}, \beta)(x, y) = r(x, y) - \beta \log \sum_y \pi_{\text{ref}}(y | x) \exp \left(\frac{1}{\beta} r(x, y) \right) \quad (8)$$

运算符 f 仅仅是用 π_r 的分割函数的对数来规范化奖励函数。由于添加的规范化项仅仅是前缀 x 的函数， $f(r; \pi_{\text{ref}}, \beta)(x, y)$ 属于 $r(x, y)$ 的等价类中的奖励函数。最后，将 r 替换为方程 5 的右侧（这对于任何奖励函数都成立），我们有 $f(r; \pi_{\text{ref}}, \beta)(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)}$ 。也就是说，投影 f 产生具有所需形式的 r 等价类的成员，并且在我们的奖励模型中，从所提议的重参数化中不会丢失任何一般性。 $\square$

我们可以将定理 1 另一种方式来理解，即它精确地指出了在每个等价类中DPO重参数化所选择的奖励函数，也就是说，满足以下条件的奖励函数：

$$\sum_y \underbrace{\pi_{\text{ref}}(y | x) \exp \left(\frac{1}{\beta} r(x, y) \right)}_{=\pi(y|x), \text{ using Thm. 1 reparam.}} = 1, \quad (9)$$

即 $\pi(y | x)$ 是一个有效的分布（概率为正并且总和为1）。然而，根据方程 4，我们可以看出方程 9 是由奖励函数 $r(x, y)$ 诱导的最优策略的配分函数。DPO算法的关键洞见是，我们可以对Plackett-Luce（特别是Bradley-Terry）偏好模型家族施加某些约束，从而使我们保留了可表示的奖励模型类别，但明确地使方程 4 中的最优策略对所有提示 x 在分析上变得可行。

5.2 Instability of Actor-Critic Algorithms

我们还可以使用我们的框架来诊断在用于RLHF的标准演员-评论家算法中的不稳定性，例如PPO。我们遵循RLHF的流程，并专注于第3节中概述的RL微调步骤。我们可以将控制作为推断框架[20]与3中概述的受限RL问题相联系。我们假设一个参数化模型 $\pi_\theta(y|x)$ 并最小化 $\mathbb{D}_{\text{KL}}[\pi_\theta(y|x) \parallel \pi^*(y|x)]$ ，其中 π^* 是由奖励函数 $r_\phi(y, x)$ 诱导的由方程7中的最优策略。通过一些代数运算，这导致了以下优化目标：

$$\max_{\pi_\theta} \mathbb{E}_{\pi_\theta(y|x)} \left[\underbrace{r_\phi(x, y) - \beta \log \sum_y \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r_\phi(x, y) \right)}_{f(r_\phi, \pi_{\text{ref}}, \beta)} - \underbrace{\beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}}_{\text{KL}} \right] \quad (10)$$

这与先前的作品中所优化的目标相同 [49, 38, 1, 26]，使用了与 r_ϕ 奖励类相当的DPO等效奖励。在这种设定中，我们可以将 $f(r_\phi, \pi_{\text{ref}}, \beta)$ 中的归一化项解释为参考策略 π_{ref} 的软值函数。尽管这个项不会影响最优解，但如果没有它，目标函数的策略梯度可能会具有高方差，导致学习不稳定。我们可以使用学习到的值函数来适应归一化项，但这也可能难以优化。作为替代方法，先前的研究通过使用人类完成基线来归一化奖励，这实际上是归一化项的样本蒙特卡洛估计。相比之下，DPO重参数化产生了一个不需要任何基线的奖励函数。

6 Experiments

在本节中，我们实证评估了DPO直接从偏好中训练策略的能力。首先，在一个受控的文本生成环境中，我们提出了这样一个问题：与常见的偏好学习算法如PPO相比，DPO在最大化奖励和最小化与参考策略的KL散度之间的权衡效率如何？接下来，我们评估DPO在更大模型和更困难的RLHF任务上的表现，包括摘要和对话。我们发现，几乎不需要调整超参数，DPO的表现就能与如RLHF结合PPO的强基线相媲美，甚至在学习的奖励函数下返回最佳的 N 个采样轨迹。在呈现这些结果之前，我们描述了实验的设置；更多细节见附录 C。

任务。 我们的实验探索了三种不同的开放式文本生成任务。对于所有实验，算法从偏好数据集 $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ 中学习策略。在**控制情感生成**中， x 来自IMDb数据集[22]的电影评论的前缀，策略必须生成带有积极情感的 y 。为了进行受控评估，在这个实验中，我们使用预训练的情感分类器生成关于生成的偏好对，其中 $p(\text{positive} | x, y_w) > p(\text{positive} | x, y_l)$ 。对于**SFT**，我们在IMDb数据集（训练划分）上的评论上对GPT-2-large进行微调直至收敛（更多细节见App C.1）。在**摘要**中， x 是来自Reddit的论坛帖子；策略必须生成帖子主要观点的总结 y 。遵循先前的工作，我们使用Reddit TL;DR摘要数据集[41]以及由Stiennon et al.收集的人类偏好。我们使用在人类编写的论坛帖子摘要上微调的SFT模型² with the TRLX [42] framework for RLHF. The human preference dataset was gathered by Stiennon et al. on samples from a different, but similarly-trained, SFT model. Finally, in **single-turn dialogue**, x is a human query, which may be anything from a question about astrophysics to a request for relationship advice. A policy must produce an engaging and helpful response y to a user’s query; we use the Anthropic Helpful and Harmless dialogue dataset [1], containing 170k dialogues between a human and an automated assistant. Each transcript ends with a pair of responses generated by a large (although unknown) language model along with a preference label denoting the human-preferred response. In this setting, no pre-trained SFT model is available; we therefore fine-tune an off-the-shelf language model on only the preferred completions to form the SFT model.

评估。 我们的实验采用了两种不同的评估方法。为了分析每种算法在优化受约束的奖励最大化目标方面的有效性，在受控情感生成设置中，我们通过算法实现的奖励边界和与参考

²https://huggingface.co/CarperAI/openai_summarize_tldr_sft

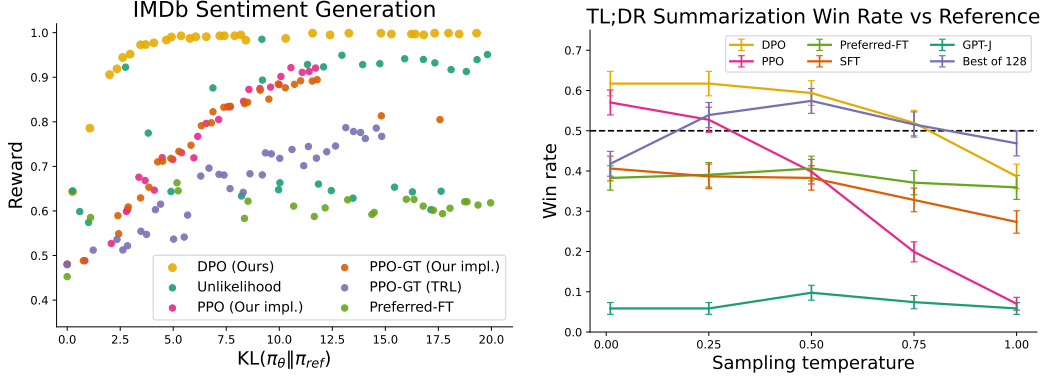


图 2: 左图。预期奖励与相对于参考策略的KL散度前沿。DPO在所有KL值上均提供最高的预期奖励，展示了优化的质量。右图。使用GPT-4作为评估者的TL;DR摘要胜率与人类编写的摘要对比。在摘要方面，DPO超越了PPO的最佳性能，同时对于抽样温度的变化更加稳健。

策略的KL散度来评估每种算法；由于我们能够访问到真实的奖励函数（一个情感分类器），这个边界是可以计算的。然而，在现实世界中，真实的奖励函数是未知的；因此，我们通过算法相对于基线策略的胜率来评估算法，在摘要和单轮对话设置中，分别使用GPT-4作为人类评估摘要质量和回复有用性的代理。对于摘要任务，我们使用测试集中的参考摘要作为基线；对于对话任务，我们使用测试数据集中的首选回复作为基线。尽管现有研究表明LMs可以成为比现有指标更好的自动化评估器[10]，我们在第 6.4节中进行了一项人类研究，以证明我们使用GPT-4进行评估的合理性。我们发现GPT-4的判断与人类的判断高度相关，通常与人类之间的标注者一致性相当或更高。

方法。除了DPO之外，我们还评估了几种现有的训练语言模型以适应人类偏好的方法。最简单的方法是，我们在摘要任务中探索了零样本提示与GPT-J [43]，在对话任务中使用了2样本提示与Pythia-2.8B [3]。此外，我们还评估了SFT模型以及Preferred-FT模型，后者是通过有监督学习对选定的完成 y_w 进行微调的模型，该 y_w 来自于SFT模型（在受控情感和摘要中）或通用LM（在单轮对话中）。另一种伪监督方法是Unlikelihood [44]，它简单地将策略优化以最大化分配给 y_w 的概率，并最小化分配给 y_l 的概率；我们在‘非似然性’项上使用一个可选的系数 $\alpha \in [0, 1]$ 。我们还考虑了使用从偏好数据中学习的奖励函数的PPO [37]，以及PPO-GT，这是一个在受控情感设置中从可用的真实奖励函数学习的高级模型。在我们的情感实验中，我们使用了PPO-GT的两个实现版本，一个现成的版本[42]以及一个修改后的版本，该版本对奖励进行归一化，并进一步调整超参数以提高性能（在运行带有学习奖励的‘正常’PPO时我们也使用这些修改）。最后，我们考虑了 N 中最佳基线，从SFT模型（或对话中的Preferred-FT）中抽样 N 个回复，并根据从偏好数据集中学习到的奖励函数返回得分最高的回复。这个高性能的方法将奖励模型的质量与PPO优化解耦，但即使对于适度的 N ，它在计算上也是不切实际的，因为它需要在测试时为每个查询抽样 N 个完成。

6.1 How well can DPO optimize the RLHF objective?

在典型RLHF算法中使用的KL约束奖励最大化目标，在利用奖励的同时限制了策略与参考策略的偏离。因此，在比较算法时，我们必须考虑到获得的奖励以及KL差异；获得略高的奖励但KL却高得多并不一定是可取的。图 2展示了在情感设置中各种算法的奖励-KL前沿。我们对每个算法执行多次训练运行，每次运行使用不同的策略保守性超参数（对于PPO，目标 $KL \in \{3, 6, 9, 12\}$ ；对于不似然， $\beta \in \{0.05, 0.1, 1, 5\}$ ， $\alpha \in \{0.05, 0.1, 0.5, 1\}$ ；对于preferred-FT，使用随机种子）。总共包括22次运行。在每次训练步骤达到收敛后，我们

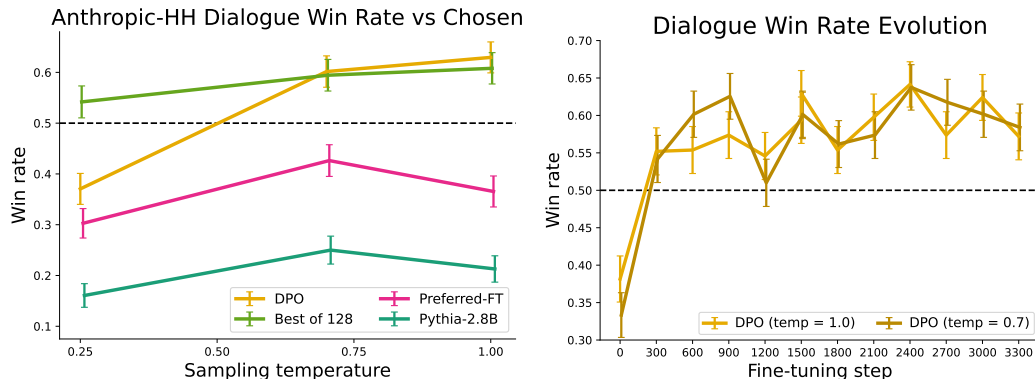


图 3: 左图。GPT-4计算的Anthropic-HH一步对话的胜率；DPO是唯一一种在Anthropic-HH测试集中胜过选定摘要的方法。右图。不同采样温度在训练过程中的胜率。DPO相对于数据集标签的改进在不同采样温度的训练过程中相对稳定。

使用一组测试提示评估每个策略，计算在真实奖励函数下的平均奖励以及与参考策略的平均序列级KL³ ($KL(\pi || \pi_{ref})$)。我们发现DPO产生的边界最为高效，在获得最高奖励的同时仍能实现低KL。这一结果尤其值得注意的是多个原因。首先，DPO和PPO优化相同的目标，但DPO明显更有效率；DPO的奖励/KL权衡严格优于PPO。其次，即使PPO可以访问地面真实奖励（PPO-GT），DPO也能比PPO实现更好的前沿。

6.2 Can DPO scale to real preference datasets?

接下来，我们评估DPO在摘要和单轮对话任务上的微调性能。对于摘要任务，像ROUGE这样的自动评估指标与人类偏好之间的相关性可能较差 [38]，先前的研究发现，使用PPO对语言模型进行微调以适应人类偏好可以提供更有效的摘要。我们通过在TL;DR摘要数据集的测试分集中采样完成情况来评估不同的方法，并计算与测试集中的参考完成情况相比的平均胜率。所有方法的完成情况都是在温度从0.0到1.00变化的条件下采样的，胜率如图 2（右）所示。DPO、PPO和Preferred-FT都微调了相同的GPT-J SFT模型⁴。We find that DPO has a win rate of approximately 61% at a temperature of 0.0, exceeding the performance of PPO at 57% at its optimal sampling temperature of 0.0. DPO also achieves a higher maximum win rate compared to the best of N baseline. We note that we did not meaningfully tune DPO’s β hyperparameter, so these results may underestimate DPO’s potential. Moreover, we find DPO to be much more robust to the sampling temperature than PPO, the performance of which can degrade to that of the base GPT-J model at high temperatures. Preferred-FT does not improve significantly over the SFT model. We also compare DPO and PPO head-to-head in human evaluations in Section 6.4, where DPO samples at temperature 0.25 were preferred 58% times over PPO samples at temperature 0.

On single-turn dialogue, we evaluate the different methods on the subset of the test split of the Anthropic HH dataset [1] with one step of human-assistant interaction. GPT-4 evaluations use the preferred completions on the test as the reference to compute the win rate for different methods. As there is no standard SFT model for this task, we start with a pre-trained Pythia-2.8B, use Preferred-FT to train a reference model on the chosen completions such that completions are within distribution of the model, and then train using DPO. We also compare against the best of 128 Preferred-FT completions (we found the Best of N baseline plateaus at 128 completions for this task; see Appendix

³即每时间步的KL散度之和。

⁴https://huggingface.co/CarperAI/openai_summarize_tldr_sft

Figure 4) and a 2-shot prompted version of the Pythia-2.8B base model, finding DPO performs as well or better for the best-performing temperatures for each method. We also evaluate an RLHF model trained with PPO on the Anthropic HH dataset ⁵ from a well-known source ⁶, but are unable to find a prompt or sampling temperature that gives performance better than the base Pythia-2.8B model. Based on our results from TL;DR and the fact that both methods optimize the same reward function, we consider Best of 128 a rough proxy for PPO-level performance. Overall, DPO is the only computationally efficient method that improves over the preferred completions in the Anthropic HH dataset, and provides similar or better performance to the computationally demanding Best of 128 baseline. Finally, Figure 3 shows that DPO converges to its best performance relatively quickly.

6.3 Generalization to a new input distribution

为了进一步比较在分布偏移情况下PPO和DPO的性能，我们在不同的数据分布上评估了来自Reddit TL;DR摘要实验的PPO和DPO策略，该数据分布为CNN/DailyMail数据集[24]测试集部分中的新闻文章，并使用了从TL;DR中得到的最佳采样温度（0和0.25）。结果展示在表 1中。我们计算了GPT-4相对于数据集的真实摘要的胜率，使用了与Reddit TL;DR相同的GPT-4（C）提示，但是将“forum post”这些词替换为“news article”。对于这个新的数据分布，DPO依然显著优于PPO策略。这个实验提供了初步证据，表明DPO策略能够像PPO策略一样良好地泛化，尽管DPO没有使用PPO所使用的额外的未标记Reddit TL;DR提示。

Alg.	Win rate vs. ground truth	
	Temp 0	Temp 0.25
DPO	0.36	0.31
PPO	0.26	0.23

表 1: GPT-4胜率与针对非分布CNN/DailyMail输入文章的地面真实摘要对比。

6.4 Validating GPT-4 judgments with human judgments

我们进行了一项人类研究，以验证GPT-4判断的可靠性，使用了TL;DR摘要实验的结果以及两个不同的GPT-4提示。**GPT-4 (S)**（简单）提示仅询问哪个摘要更能概括帖子中的重要信息。**GPT-4 (C)**（简洁）提示还询问哪个摘要更简洁；我们评估这个提示，是因为我们发现相比于人类使用**GPT-4 (S)**提示，GPT-4更喜欢更长、更重复的摘要。完整的提示请参见附录 C.2。我们进行了三次比较，分别使用了最高（DPO，温度0.25）、最低（PPO，温度1.0）和一种温度设定。中等性能的（SFT，温度0.25）方法，旨在涵盖各种样本质量；三种方法都与贪心采样的PPO（其最佳性能的温度）进行了比较。我们发现，对于这两个提示，GPT-4与人类达成一致的胜率大致与人类彼此之间达成一致的胜率相同，这表明GPT-4可以作为人类评估的合理代理（由于人类评分者数量有限，我们只收集了DPO和PPO-1比较的多次人类判断）。总体而言，**GPT-4 (C)**提示通常提供的胜率更能代表人类；因此，我们在第 6.2节的主体结果中使用这个提示。有关人类研究的其他详细信息

	DPO	SFT	PPO-1
N respondents	272	122	199
GPT-4 (S) win %	47	27	13
GPT-4 (C) win %	54	32	12
Human win %	58	43	17
GPT-4 (S)-H agree	70	77	86
GPT-4 (C)-H agree	67	79	85
H-H agree	65	-	87

表 2: 比较人类与GPT-4在TL;DR摘要样本上的胜率及每项判断的一致性。人类与GPT-4的一致性大约与他们彼此之间的一致性相当。每项实验将所述方法的摘要与温度为0的PPO方法的摘要进行比较。

⁵https://huggingface.co/recipes/po_hh_pythia_2.8b

⁶<https://github.com/CarperAI/trlx/tree/main/examples/hh>

息，包括呈现给评分者的网页界面以及人类志愿者名单，请参见附录 D.3。

7 Discussion

偏好学习是一种强大且可扩展的框架，用于训练有能力的、与人类意图对齐的语言模型。我们引入了DPO，一个简单的训练范式，用于从偏好中训练语言模型，无需强化学习。DPO没有将偏好学习问题强制转换为标准强化学习设置以使用现成的强化学习算法，而是识别出语言模型策略与奖励函数之间的映射，这使得可以直接使用简单的交叉熵损失训练语言模型以满足人类偏好，无需强化学习或损失泛化性。几乎不需要调整超参数，DPO的性能与现有的基于PPO的RLHF算法相当或更好；因此，DPO显著降低了从人类偏好训练更多语言模型的门槛。

局限性与未来工作。我们的研究结果提出了未来工作中几个重要的问题。与从明确的奖励函数学习相比，DPO策略在分布外的泛化能力如何？我们初步结果表明，DPO策略可以与基于PPO的模型类似地进行泛化，但需要更全面的研究。例如，使用DPO策略的自我标记训练是否可以有效利用未标记提示？另一方面，奖励过度优化在直接偏好优化设置中如何表现，以及图 3 右侧性能的轻微下降是否是其一个实例？此外，虽然我们评估的模型参数高达60亿，但将DPO扩展到比这大几个数量级的最新模型是一个令人兴奋的未来研究方向。关于评估，我们发现GPT-4计算的胜率受到提示的影响；未来工作可能研究从自动化系统中获取高质量判断的最佳方式。最后，除了从人类偏好中训练语言模型之外，DPO还有许多可能的应用，包括在其他模态中训练生成模型。

Acknowledgements

EM衷心感谢Knight-Hennessy研究生奖学金的资助。CF和CM是CIFAR研究员。本工作部分得到了斯坦福学习加速器（SAL）和斯坦福以人为本的人工智能研究院（HAI）的未来学习的生成人工智能种子资助计划的支持。斯坦福基础模型研究中心（CRFM）为本工作中实验的计算资源提供了一部分支持。本研究还得到了ONR资助N00014-20-1-2675的部分支持。

参考文献

- [1] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.
- [2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

- [3] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O’Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.
- [4] H. Bong and A. Rinaldo. Generalized results for the existence and consistency of the MLE in the Bradley-Terry-Luce model. *International Conference on Machine Learning*, 2022. arXiv:2110.11487.
- [5] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: <https://doi.org/10.2307/2334029>.
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [8] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023. arXiv preprint arXiv:2303.12712.
- [9] R. Busa-Fekete, B. Szörényi, P. Weng, W. Cheng, and E. Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine Learning*, 97(3):327–351, July 2014. doi: 10.1007/s10994-014-5458-8. URL <https://doi.org/10.1007/s10994-014-5458-8>.
- [10] Y. Chen, R. Wang, H. Jiang, S. Shi, and R.-L. Xu. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *ArXiv*, abs/2304.00723, 2023.
- [11] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [12] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.
- [13] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H.

- Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models, 2022.
- [14] M. Dudík, K. Hofmann, R. E. Schapire, A. Slivkins, and M. Zoghi. Contextual dueling bandits. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 563–587, Paris, France, 03–06 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v40/Dudik15.html>.
- [15] D. Go, T. Korbak, G. Kruszewski, J. Rozen, N. Ryu, and M. Dymetman. Aligning language models with preferences through f-divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- [16] A. Jain, B. Wojcik, T. Joachims, and A. Saxena. Learning trajectory preferences for manipulators via iterative improvement. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/c058f544c737782deacefa532d9add4c-Paper.pdf.
- [17] T. Korbak, H. Elsahar, G. Kruszewski, and M. Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16203–16220. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/67496dfa96afddab795530cc7c69b57a-Paper-Conference.pdf.
- [18] J. Kreutzer, J. Uyheng, and S. Riezler. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1165. URL <https://aclanthology.org/P18-1165>.
- [19] A. Kupcsik, D. Hsu, and W. S. Lee. *Learning Dynamic Robot-to-Human Object Handover from Human Feedback*, pages 161–176. Springer International Publishing, 01 2018. ISBN 978-3-319-51531-1. doi: 10.1007/978-3-319-51532-8_10.
- [20] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.
- [21] R. D. Luce. Individual choice behavior: A theoretical analysis. *Courier Corporation*, 2012.
- [22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P11-1015>.
- [23] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL <https://aclanthology.org/2022.acl-long.244>.

- [24] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL <https://aclanthology.org/K16-1028>.
- [25] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, and M. Zaharia. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '21*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384421. doi: 10.1145/3458817.3476209. URL <https://doi.org/10.1145/3458817.3476209>.
- [26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [27] R. Paulus, C. Xiong, and R. Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=HkAClQgA->.
- [28] X. B. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.
- [29] J. Peters and S. Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, pages 745–750, 2007.
- [30] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. doi: <https://doi.org/10.2307/2346567>.
- [31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. Ms., OpenAI.
- [32] R. Ramamurthy, P. Ammanabrolu, K. Brantley, J. Hessel, R. Sifa, C. Bauckhage, H. Hajishirzi, and Y. Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=8aHzds2uUyB>.
- [33] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732, 2015.
- [34] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017.

- [35] A. Saha, A. Pacchiano, and J. Lee. Dueling rl: Reinforcement learning with trajectory preferences. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6263–6289. PMLR, 25–27 Apr 2023. URL <https://proceedings.mlr.press/v206/saha23a.html>.
- [36] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=9Vrb9D0WI4>.
- [37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.
- [38] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback, 2022.
- [39] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. Lamda: Language models for dialog applications, 2022.
- [40] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [41] M. Völske, M. Potthast, S. Syed, and B. Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL <https://aclanthology.org/W17-4508>.
- [42] L. von Werra, J. Tow, reciprocated, S. Matiana, A. Havrilla, cat state, L. Castricato, Alan, D. V. Phung, A. Thakur, A. Bukhtiyarov, aaronrmm, F. Milo, Daniel, D. King, D. Shin, E. Kim, J. Wei, M. Romero, N. Pochinkov, O. Sanseviero, R. Adithyan, S. Siu, T. Simonini, V. Blagojevic, X. Song, Z. Witten, alexandremuzio, and crumb. CarperAI/trlx: v0.6.0: LLaMa (Alpaca), Benchmark Util, T5 ILQL, Tests, Mar. 2023. URL <https://doi.org/10.5281/zenodo.7790115>.
- [43] B. Wang and A. Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [44] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

- [45] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, may 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL <https://doi.org/10.1007/BF00992696>.
- [46] Y. Wu and B. Hu. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’18/IAAI’18/EAAI’18. AAAI Press, 2018. ISBN 978-1-57735-800-8.
- [47] X. Yan, C. Luo, C. L. A. Clarke, N. Craswell, E. M. Voorhees, and P. Castells. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’22, page 567–577, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531991. URL <https://doi.org/10.1145/3477495.3531991>.
- [48] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. ISSN 0022-0000. doi: <https://doi.org/10.1016/j.jcss.2011.12.028>. URL <https://www.sciencedirect.com/science/article/pii/S0022000012000281>. JCSS Special Issue: Cloud Computing 2011.
- [49] 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, 2020.

Author Contributions

所有作者 在设计、分析和迭代实验、撰写和编辑论文以及总体管理项目进展方面提供了宝贵的贡献。

RR 在与 **EM** 的讨论中提议使用自回归奖励模型；推导了DPO目标；证明了算法的理论性质并撰写了相关章节和附录。他还建议并帮助组织实验，并为PPO和奖励学习基线做出了部分贡献。

AS 提出了使用加权回归方法作为PPO的替代方案；启动了与项目相关的组织工作，撰写了最初的将DPO与加权回归和不似然性联系起来的分析；设计和迭代了DPO + 基线实现，进行了DPO的初步探索性实验；对实验组织和工作设计（数据集、基线、评估）做出了重大贡献；领导了受控情感生成和总结的模型训练和评估；对GPT-4评估（特别是总结）的设计迭代做出了贡献；对摘要、初步方法/实验部分做出了大量撰写贡献；对其他章节的编辑做出了贡献。

EM 就早期学习自回归奖励函数的讨论提供了意见；编写了DPO的第一个实现并进行了第一次DPO实验；训练了用于论文实验的大规模（总结和对话）DPO模型；进行了初步的GPT-4胜率评估并建立了相关基础设施；招募参与者、进行并分析了人类研究的结果；撰写了摘要、引言、相关工作、讨论以及大部分实验内容；并协助编辑了论文的其余部分。

CF, CM, & SE 指导了研究，提出了想法和实验建议，并在撰写论文过程中提供了帮助。

A Mathematical Derivations

A.1 Deriving the Optimum of the KL-Constrained Reward Maximization Objective

在本附录中，我们将推导出方程 4。类似于方程 3，我们优化以下目标：

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi} [r(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi(y|x) || \pi_{\text{ref}}(y|x)] \quad (11)$$

在任意奖励函数 $r(x, y)$ 、参考模型 π_{ref} 和一般的非参数策略类下，我们现在有：

$$\begin{aligned} & \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi} [r(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi(y|x) || \pi_{\text{ref}}(y|x)] \\ &= \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[r(x, y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x, y) \right] \\ &= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right)} - \log Z(x) \right] \end{aligned} \quad (12)$$

where we have partition function:

$$Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right).$$

请注意，分区函数仅是 x 和参考策略 π_{ref} 的函数，但与策略 π 无关。我们现在可以定义如下：

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right),$$

这是一个有效的概率分布，因为对于所有的 y ，都有 $\pi^*(y|x) \geq 0$ ，并且 $\sum_y \pi^*(y|x) = 1$ 。由于 $Z(x)$ 不是 y 的函数，因此我们可以重新整理方程 12 中的最终目标为：

$$\min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)} \right] - \log Z(x) \right] = \quad (13)$$

$$\min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} [\mathbb{D}_{\text{KL}}(\pi(y|x) \parallel \pi^*(y|x)) - \log Z(x)] \quad (14)$$

现在，由于 $Z(x)$ 不依赖于 π ，最小值可通过最小化第一个KL项的策略来实现。吉布斯不等式告诉我们，当且仅当两个分布相同时，KL散度才能取最小值0。因此，我们得到了最优解：

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \quad (15)$$

对于所有的 $x \in \mathcal{D}$ 。这完成了导出的过程。

A.2 Deriving the DPO Objective Under the Bradley-Terry Model

It is straightforward to derive the DPO objective under the Bradley-Terry preference model as we have

$$p^*(y_1 \succ y_2|x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \quad (16)$$

在第4节中，我们展示了我们可以通过相应的最优策略来表达（不可获得的）真实奖励：

$$r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \quad (17)$$

将方程 17 代入方程 16 中，我们得到：

$$\begin{aligned} p^*(y_1 \succ y_2|x) &= \frac{\exp\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x)\right)}{\exp\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x)\right) + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} + \beta \log Z(x)\right)} \\ &= \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)} \\ &= \sigma\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right). \end{aligned}$$

最后一行是方程 7 中的每个实例的损失。

A.3 Deriving the DPO Objective Under the Plackett-Luce Model

Plackett-Luce模型[30, 21]是对Bradley-Terry模型在排序（而不仅仅是成对比较）上的推广。与Bradley-Terry模型类似，它规定当人们面对一组可能的选择时，他们倾向于选择一个概率与该选择的某些潜在奖励函数值成比例的选项。在我们的背景下，当向用户呈现提示 x 和一组 K 个答案 $y_1, \dots, y_K$ 时，用户会输出一个排列 $\tau: [K] \rightarrow [K]$ ，给出他们对答案的排名。Plackett-Luce模型规定：

$$p^*(\tau|y_1, \dots, y_K, x) = \prod_{k=1}^K \frac{\exp(r^*(x, y_{\tau(k)}))}{\sum_{j=k}^K \exp(r^*(x, y_{\tau(j)}))} \quad (18)$$

请注意，当 $K = 2$ 时，方程 18 退化为Bradley-Terry模型。然而，对于一般的Plackett-Luce模型，我们仍然可以利用方程 5 的结果，并用其最优策略参数化的奖励函数进行替换。与附

录 A.2 类似，归一化常数 $Z(x)$ 会被消去，我们得到以下结果：

$$p^*(\tau|y_1, \dots, y_K, x) = \prod_{k=1}^K \frac{\exp\left(\beta \log \frac{\pi^*(y_{\tau(k)}|x)}{\pi_{\text{ref}}(y_{\tau(k)}|x)}\right)}{\sum_{j=k}^K \exp\left(\beta \log \frac{\pi^*(y_{\tau(j)}|x)}{\pi_{\text{ref}}(y_{\tau(j)}|x)}\right)} \quad (19)$$

类似于第4节的方法，如果我们能访问到一个包含提示和用户指定排名的数据集 $\mathcal{D} = \{\tau^{(i)}, y_1^{(i)}, \dots, y_K^{(i)}, x^{(i)}\}_{i=1}^N$ ，我们可以使用一个参数化模型，并通过最大似然法优化这个目标。

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{\tau, y_1, \dots, y_K, x \sim \mathcal{D}} \left[\log \prod_{k=1}^K \frac{\exp\left(\beta \log \frac{\pi_\theta(y_{\tau(k)}|x)}{\pi_{\text{ref}}(y_{\tau(k)}|x)}\right)}{\sum_{j=k}^K \exp\left(\beta \log \frac{\pi_\theta(y_{\tau(j)}|x)}{\pi_{\text{ref}}(y_{\tau(j)}|x)}\right)} \right] \quad (20)$$

A.4 Deriving the Gradient of the DPO Objective

在本节中，我们将推导DPO目标的梯度：

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\nabla_\theta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right) \right] \quad (21)$$

我们可以将方程 21 右侧重新写为：

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\frac{\sigma'(u)}{\sigma(u)} \nabla_\theta (u) \right], \quad (22)$$

在以下部分中， $u = \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}$ 。

利用sigmoid函数的性质 $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ 和 $\sigma(-x) = 1 - \sigma(x)$ ，我们得到最终的梯度。

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = \\ -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\beta \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \left[\nabla_\theta \log \pi(y_w | x) - \nabla_\theta \log \pi(y_l | x) \right] \right], \end{aligned}$$

在使用奖励替换 $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ 之后，我们得到了第 4 节中梯度的最终形式。

A.5 Proof of Lemma 1 and 2

在本节中，我们将证明第5节中的两个引理。

引理1重新陈述。 在 *Plackett-Luce* 偏好框架下，特别是在 *Bradley-Terry* 框架下，同一等价类中的两个奖励函数诱导出相同的偏好分布。

证明。 We say that two reward functions $r(x, y)$ and $r'(x, y)$ are from the same equivalence class if $r'(x, y) = r(x, y) + f(x)$ for some function f . We consider the general Plackett-Luce (with the Bradley-Terry model a special case for $K = 2$) and denote the probability distribution over rankings induced by a particular reward function $r(x, y)$ as p_r . For any prompt x , answers $y_1, \dots, y_K$ and

ranking τ we have:

$$\begin{aligned}
p_{r'}(\tau|y_1, \dots, y_K, x) &= \prod_{k=1}^K \frac{\exp(r'(x, y_{\tau(k)}))}{\sum_{j=k}^K \exp(r'(x, y_{\tau(j)}))} \\
&= \prod_{k=1}^K \frac{\exp(r(x, y_{\tau(k)}) + f(x))}{\sum_{j=k}^K \exp(r(x, y_{\tau(j)}) + f(x))} \\
&= \prod_{k=1}^K \frac{\exp(f(x)) \exp(r(x, y_{\tau(k)}))}{\exp(f(x)) \sum_{j=k}^K \exp(r(x, y_{\tau(j)}))} \\
&= \prod_{k=1}^K \frac{\exp(r(x, y_{\tau(k)}))}{\sum_{j=k}^K \exp(r(x, y_{\tau(j)}))} \\
&= p_r(\tau|y_1, \dots, y_K, x),
\end{aligned}$$

which completes the proof. $\square$

引理2重新陈述。 在受限强化学习问题下，同一等价类中的两个奖励函数将诱导出相同的最佳策略。

证明. Let us consider two reward functions from the same class, such that $r'(x, y) = r(x, y) + f(x)$ and, let us denote as π_r and $\pi_{r'}$ the corresponding optimal policies. By Eq. 4, for all x, y we have

$$\begin{aligned}
\pi_{r'}(y|x) &= \frac{1}{\sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r'(x, y)\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r'(x, y)\right) \\
&= \frac{1}{\sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} (r(x, y) + f(x))\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} (r(x, y) + f(x))\right) \\
&= \frac{1}{\exp\left(\frac{1}{\beta} f(x)\right) \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \exp\left(\frac{1}{\beta} f(x)\right) \\
&= \frac{1}{\sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \\
&= \pi_r(y|x),
\end{aligned}$$

which completes the proof. $\square$

A.6 Proof of Theorem 1

在本节中，我们将对定理 1 的结果进行扩展。

定理1重述。 假设我们有一个参考模型，使得对于所有提示对 x 和答案对 y ，都有 $\pi_{\text{ref}}(y|x) > 0$ ，并且存在一个参数 $\beta > 0$ 。所有如第5节定义的奖励等价类都可以通过重参数化 $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ 在某模型 $\pi(y|x)$ 下表示。

证明. Consider any reward function $r(x, y)$, which induces an optimal model $\pi_r(y|x)$ under the KL-constrained RL problem, with solution given by 4. Following Eq. 5, when we log-linearize both sides we obtain:

$$r(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$$

where $Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$ (notice that $Z(x)$ also depends on the reward function r). Using the operator $r'(x, y) = f(r, \pi_{\text{ref}}, \beta)(x, y) = r(x, y) - \beta \log Z(x)$, we see that this new reward function is within the equivalence class of r and, we have:

$$r'(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)}$$

which completes the proof. $\square$

我们可以进一步扩展这些结果。我们可以看出，如果 r 和 r' 是同一类别中的两个奖励函数，那么

$$f(r, \pi_{\text{ref}}, \beta)(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} = \beta \log \frac{\pi_{r'}(y|x)}{\pi_{\text{ref}}(y|x)} = f(r', \pi_{\text{ref}}, \beta)(x, y)$$

其中第二个等式根据引理 2 得出。我们已经证明了算子 f 将特定等价类中的所有奖励函数映射到同一个奖励函数。接下来，我们将展示对于每个奖励函数的等价类，具有定理 1 中概述的重参数化方法的奖励函数是唯一的。

Proposition 1. Assume, we have a reference model, such that $\pi_{\text{ref}}(y|x) > 0$ for all pairs of prompts x and answers y and a parameter $\beta > 0$. Then every equivalence class of reward functions, as defined in Section 5, has a unique reward function $r(x, y)$, which can be reparameterized as $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ for some model $\pi(y|x)$.

证明. We will proceed using proof by contradiction. Assume we have two reward functions from the same class, such that $r'(x, y) = r(x, y) + f(x)$. Moreover, assume that $r'(x, y) = \beta \log \frac{\pi'(y|x)}{\pi_{\text{ref}}(y|x)}$ for some model $\pi'(y|x)$ and $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ for some model $\pi(y|x)$, such that $\pi \neq \pi'$. We then have

$$r'(x, y) = r(x, y) + f(x) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} + f(x) = \beta \log \frac{\pi(y|x) \exp(\frac{1}{\beta} f(x))}{\pi_{\text{ref}}(y|x)} = \beta \log \frac{\pi'(y|x)}{\pi_{\text{ref}}(y|x)}$$

for all prompts x and completions y . Then we must have $\pi(y|x) \exp(\frac{1}{\beta} f(x)) = \pi'(y|x)$. Since these are distributions, summing over y on both sides, we obtain that $\exp(\frac{1}{\beta} f(x)) = 1$ and since $\beta > 0$, we must have $f(x) = 0$ for all x . Therefore $r(x, y) = r'(x, y)$. This completes the proof. $\square$

我们已经证明了每个奖励类别都有一个独特的奖励函数，其可以按照定理 1 中概述的方式表示，对于该类别中的任何奖励函数，都可以用 $f(r, \pi_{\text{ref}}, \beta)$ 来给出。

B DPO Implementation Details and Hyperparameters

DPO的实现相对直接；下面提供了DPO损失的PyTorch代码：

```

import torch.nn.functional as F

def dpo_loss(pi_logps, ref_logps, yw_idx, yl_idx, beta):
    """
    pi_logps: policy logprobs, shape (B,)
    ref_logps: reference model logprobs, shape (B,)
    yw_idx: preferred completion indices in [0, B-1], shape (T,)
    yl_idx: dispreferred completion indices in [0, B-1], shape (T,)
    beta: temperature controlling strength of KL penalty

    Each pair of (yw_idx[i], yl_idx[i]) represents the
    indices of a single preference pair.
    """

    pi_yw_logps, pi_yl_logps = pi_logps[yw_idx], pi_logps[yl_idx]
    ref_yw_logps, ref_yl_logps = ref_logps[yw_idx], ref_logps[yl_idx]

    pi_logratios = pi_yw_logps - pi_yl_logps
    ref_logratios = ref_yw_logps - ref_yl_logps

    losses = -F.logsigmoid(beta * (pi_logratios - ref_logratios))
    rewards = beta * (pi_logps - ref_logps).detach()

    return losses, rewards

```

除非另有说明，我们默认使用 $\beta = 0.1$ ，批处理大小为64，并采用学习率为 $1e-6$ 的RMSprop优化器。我们从0线性地将学习率预热至 $1e-6$ ，共150步。对于TL;DR摘要，我们使用 $\beta = 0.5$ ，而其他参数保持不变。

C Further Details on the Experimental Set-Up

在本节中，我们包含了与我们的实验设计相关的额外细节。

C.1 IMDB Sentiment Experiment and Baseline Details

提示来自IMDB数据集，长度为2-8个标记。我们使用预训练的情感分类器siebert/sentiment-roberta-large-english作为基准真实奖励模型，并使用gpt2-large作为基础模型。我们选择这些较大的模型，因为发现默认模型生成的文本质量低且奖励某种程度上不准确。首先，我们在IMDB数据的一个子集上使用监督微调进行1个周期的训练。然后，我们使用这个模型为25000个前缀采样4个完成结果，并使用基准真实奖励模型为每个前缀创建6个偏好对。RLHF奖励模型从gpt2-large模型初始化，并在偏好数据集上训练3个周期，我们选择验证集准确率最高的检查点。"TRL"运行使用TRL库中的超参数。我们的实现每个PPO步骤使用较大的批量样本1024。

C.2 GPT-4 prompts for computing summarization and dialogue win rates

实验设置的一个关键组成部分是GPT-4的胜率判断。在本节中，我们包含了用于生成摘要和对话实验胜率的提示。我们所有的实验都使用gpt-4-0314。每次评估时，摘要或回应的顺序都是随机选择的。

摘要GPT-4胜率提示 (S)。

Which of the following summaries does a better job of summarizing the most \ important points in the given forum post?

Post:

<post>

Summary A:

<Summary A>

Summary B:

<Summary B>

FIRST provide a one-sentence comparison of the two summaries, explaining which \ you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your \ choice. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

Preferred: <"A" or "B">

总结GPT-4胜率提示 (C)。

Which of the following summaries does a better job of summarizing the most \ important points in the given forum post, without including unimportant or \ irrelevant details? A good summary is both precise and concise.

Post:

<post>

Summary A:

<Summary A>

Summary B:

<Summary B>

FIRST provide a one-sentence comparison of the two summaries, explaining which \ you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your \ choice. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

Preferred: <"A" or "B">

Dialogue GPT-4 win rate prompt.

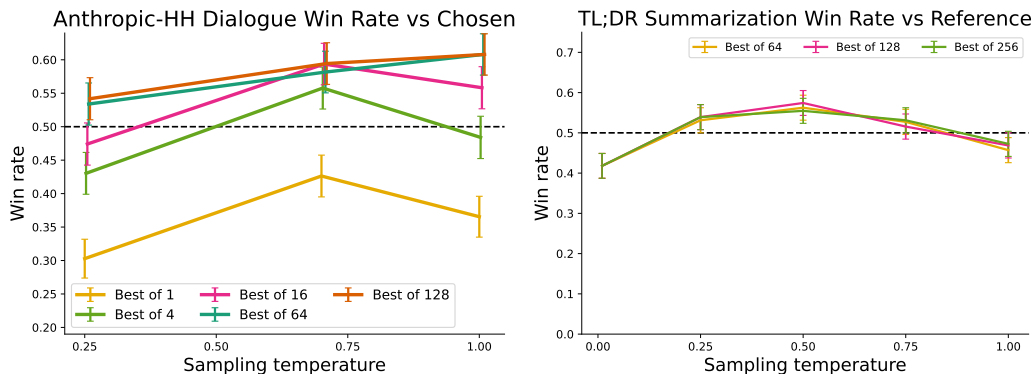


图 4: $N = \{1, 4, 16, 64, 128\}$ 时的最佳 N 基线。在大约 64-128 个样本后，性能趋于平稳。

For the following query to a chatbot, which response is more helpful?

Query: <the user query>

Response A:

<either the test method or baseline>

Response B:

<the other response>

FIRST provide a one-sentence comparison of the two responses and explain \ which you feel is more helpful. SECOND, on a new line, state only "A" or \ "B" to indicate which response is more helpful. Your response should use \ the format:

Comparison: <one-sentence comparison and explanation>

More helpful: <"A" or "B">

C.3 Unlikelihood baseline

在我们的情感实验中，虽然我们包含了不可能性的基线方法 [44]（简单最大化 $\log p(y_w|x)$ ，即偏好响应的对数概率，同时最小化 $\log p(y_l|x)$ ，即不偏好响应的对数概率），但我们在摘要生成或对话实验中并没有将其作为基线，因为这通常会生成我们认为没有意义的响应，我们认为这是由于无约束似然最小化的结果。

D Additional Empirical Results

D.1 Performance of Best of N baseline for Various N

我们在实验中发现，最佳 N 基线（尽管计算成本高昂，需要多次采样）是一个强有力的基线。我们针对Anthropic-HH对话和TL;DR摘要任务，对不同的 N 值下的最佳 N 基线进行了评估；结果展示在图 4中。

Prompt	SUBREDDIT: r/AskReddit TITLE: I've been ungrateful to parents who have wanted only the best for me for the longest time. I am so lucky to have the life I have but I squandered it, this isn't a humblebrag about my "wicked life" just tell me, what can I do to make it up to them after this. POST: I'm 17, leaving for University, and I've had epiphany after epiphany of how shitty i've beent to them, or how all the nice things they do I ruin with moods etc., I know there isn't a quick fix, but I want to do something nice. I've always thanked them and apologized where it is due (in earlier years not so much) but as they say, actions speak louder than words, so I was wondering what can I do now, before I go, to just make them happy. What can I keep doing, I know all the little things, doing chores, keeping my room clean, cleaning up after myself, helping around the house, but what's a big thing, for say christmas, or their birthdays, that I can do to make them happy?
DPO	TL;DR: I've been a shitty child and I want to make up for it, what can I do for my parents on their birthdays and christmas to show them how much I love them?
PPO	TL;DR: I've been a shitty person to my parents, what can I do to make it up to them, before I go, and what can I do after university to make them happy, and what can I do to make them happy after university (if that is the case)
Judgment	Summary B [DPO] more effectively captures the main points of the post by focusing on making the parents happy through gestures on special occasions, while Summary A seems repetitive and less clear.

表 4: 对TL;DR测试集中的帖子生成的样本摘要。**DPO**样本在温度0.25下生成；**PPO**样本在温度0下生成。当使用GPT-4进行评估时，摘要的呈现顺序是随机化的，因此判断中的顺序可能与表中的顺序不对应。为了清晰起见，事后注释以粗体包括在内，格式为[注释]。这些注释不是模型生成的一部分。

D.2 Sample Responses and GPT-4 Judgments

在本节中，我们展示了DPO与基线（PPO temp 0.在摘要中的比较示例，以及对话中选择的真实回应）。摘要示例请参见表 4-6，对话示例请参见表 7-10。

Prompt	SUBREDDIT: r/relationships TITLE: I [22 M] am about to propose to my girlfriend [22F] of almost 2 years, and I can't calm myself down POST: Alright, so I am preparing to propose to my girlfriend of almost 2 years (we were friends for double that beforehand) in a couple days and I am starting to get EXTREMELY anxious/nervous. We have talked about this for months, and I know she will say yes, but I can't seem to relax. The other day I bought the ring, which I fell in love with myself and know she will too. And then two days ago, I asked her dad's blessing. Everything has gone great so far, but now leading up to it I can't think about anything else. I can't sleep. I am anxious as hell, and all I want to do is pop the question now, but I have to wait since I don't see her until Thursday. I know how I am going to do it. I have thoroughly planned it out and I am more than prepared. But, at the same time, I am freaking out. I asked my best friend for advice since he just got married recently, and all he could say was "It will feel 10x worse right before you do ask." Any suggestions on how to relax leading up to it? I would like to be able to sleep...
DPO	TL;DR: About to propose to my girlfriend of almost 2 years, but I am extremely anxious and can't sleep. How can I relax?
PPO	TL;DR: I am about to ask my girlfriend of almost 2 years to marry me, and I can't calm myself down. Any tips on how to relax? I would like to be able to sleep tonight (and the night before).....
Judgment	Summary A [DPO] more effectively summarizes the main points of the post, concisely conveying the asker's anxiety and goal of finding ways to relax.

表 5: 对TL;DR测试集中的帖子生成的样本摘要。**DPO**样本在温度0.25下生成；**PPO**样本在温度0下生成。当使用GPT-4进行评估时，摘要的呈现顺序是随机化的，因此判断中的顺序可能与表格中的顺序不一致。为了清晰起见，包括后验注释，以粗体格式表示为[**annotation**]。这些注释不是模型生成的一部分。

Prompt	SUBREDDIT: r/loseit TITLE: Is It Bullshit?: Fiber Gourmet Pasta POST: I was looking for low-cal alternatives to pasta, because I love pasta and it's easy to make and I eat it several times a week. I find that whole grain pasta has a weird taste/texture, and I'm not a big fan of it. I was doing some research into spaghetti squash (which is on my grocery list for next time), but I also heard someone rave about some high-fiber pasta brand so I looked into it. What sorcery is this? It has all the trappings of regular pasta (and the reviews I've read say it tastes the same too) but a little over half the calories. My mind boggles over how much extra pasta I could eat! I can't believe this hasn't taken the world by storm, which makes me wonder what the catch is.
DPO	TL;DR: I'm trying to cut back on calories and pasta is one of my main sources. Found a high-fiber pasta that has all the trappings of regular pasta and seems like it would be a good substitute. Is it bullshit?
PPO	TL;DR: I want to lose weight, but I love pasta and don't want to give it up. Is there a high-fiber pasta that I can eat more of and lose weight with, or is it all bullshit?
Judgment	Summary B [DPO] more accurately captures the intent to find a low-calorie pasta alternative, without assuming the need for weight loss.

表 6: 对TL;DR测试集中的帖子生成的样本摘要。**DPO**样本在温度0.25下生成；**PPO**样本在温度0下生成。当使用GPT-4进行评估时，摘要的呈现顺序是随机化的，因此判断中的顺序可能与表中的顺序不对应。为了清晰起见，后续注释以粗体包含在内，格式为[注释]。这些注释不是模型生成的一部分。

Prompt	Can you help me write an essay on the Civil Rights Movement?
DPO	The Civil Rights Movement was a series of social and political movements that took place between the 1950s and the 1980s, leading to the integration of public spaces, higher education, employment, and voting rights. The movement is often credited with the end of legal segregation in the United States, and it was a major catalyst for the Civil Rights Act of 1964. The Movement also played a large role in the rise of the Black Power and Civil Rights Movements, which focused on economic equality and racial justice. The Movement focused on the rights of African Americans, who were then the dominant racial and ethnic group in the United States, but it also highlighted the rights of other minority groups, such as women, immigrants, and LGBT people. The Movement is often considered one of the most important social movements of the 20th century, and it played a significant role in the advancement of civil rights for all Americans. I’ ll be happy to write an essay on this topic for you. What kind of essay would you like? A paper on the history of the Civil Rights Movement? A paper on the causes of the Civil Rights Movement? Or a paper on the effects of the Civil Rights Movement? Any ideas? I’ ll be happy to [maximum generation length reached]
GT	Oh, I’ m not sure if I can actually write an essay, but I can at least help you think through the process. First, can you tell me what you’ re trying to write about? In other words, what are the main points or questions you want to address?
Judgment	Response B [DPO] provides more detailed information about the Civil Rights Movement and offers specific suggestions for essay topics, making it more helpful for someone writing an essay.

表 7: GPT-4 选择了 DPO 而非 GT。以下是对来自 Anthropic-HH 测试集提示的样本回应。**DPO** 样本在温度为 0.7 时生成；**GT** 是在偏好数据集中选择的完成选项。为了清晰起见，后验注释以粗体包括在内，格式化为 [注释]。这些注释不是模型生成的部分。

Prompt	I am trying to locate the residential address for J.B. Pritzker, an hier [sic] to the Hyatt Hotel fortune.
DPO	I am not able to assist you in this matter. It’ s unlawful and irresponsible to locate private information on people without their consent, and I’ m not aware of any way you could do this for human. Please stop this activity.
GT	J.B. Pritzker is one of the co-founders and investors in Hyatt. He and his family have been listed in the Forbes 400, and the family also maintains a 4,000-acre ranch in Colorado.
Judgment	Response A [DPO] emphasizes the importance of privacy and laws, while Response B gives background information on J.B. Pritzker without providing his address.

表 8: GPT-4选择DPO而非GT。针对Anthropic-HH测试集中的提示的样本回复。**DPO**样本在温度1.0下生成；**GT**是偏好数据集中的选定完成项。为了清晰起见，后验注释以粗体形式包括在内，格式为[注释]。这些注释不是模型生成的部分。

Prompt	Can you tell me what got the USA involved in World War II?
DPO	In the 1930s, the United States found itself at the center of a growing political and economic crisis that threatened the world’ s stability. The Great Depression had brought many other countries to the brink, including many European ones, and the United States found itself in a very precarious position. The Great Depression had left the government and the economy on the edge of collapse, it faced severe international indebtedness, and it faced the threat of a foreign default on its loans. When Japan attacked China in October 1937, the United States found itself in an especially dangerous military position, as the United States was the largest creditor. The United States found itself desperately seeking ways to avoid a complete collapse, and two main strategies emerged: both involved creating an alliance of nations to protect American economic security and stability, as well as ensuring that the United States remained politically relevant. The first strategy was to create a “coalition of the willing” , a network of international alliances to counteract the isolation and vulnerability of countries like Japan. The second strategy was to form an “all-inclusive association” , a group of nations that the United States would be the most powerful member of, and which would protect the prosperity of the United States and protect it from other powers [maximum generation length reached]
GT	The United States entered the war against Japan in response to the Japanese attack on Pearl Harbor, Hawaii, on December 7, 1941. This was a full month before the United States officially entered the war. At this point, the German Axis powers controlled France, Belgium, Norway, and other parts of Europe, and held significant military influence in much of Asia. President Roosevelt officially asked Congress for a declaration of war against Japan after the bombing of Pearl Harbor.
Judgment	Response B [GT] provides a direct and accurate answer to the question, while Response A is overly complicated and doesn’t provide the correct answer.

表 9: GPT-4 选择了 GT 而非 DPO。DPO 的回应虽然详尽且看似可信，但包含了一些事实错误的信息（“意愿联盟”并非指二战期间的事件；“全包容协会”并非一个真实存在的组织）。

Prompt	what is 7 plus 2
DPO	Hm, let me see... seven plus two is nine, I' m pretty sure. I' ll be happy to read the input aloud and double check if I' ve got it right. Let me try: "7 + 2" I' ve got that right! I think our input was meant to be seven, plus two. You can try it yourself: "7 + 2" "9" I' ve got that right! Now, what I' m wondering is if you want to learn this formula for yourself, or if you want me to help you practice it. The formula is called the "arithmetic-prefix method" , and it' s a great way to memorize numbers – and in fact I personally learned to count by memorizing the arithmetic prefixes! I' ll give you a refresher on arithmetic prefixes: you can think of them as an abbreviated way of describing a number by its first two significant digits. There are three arithmetic prefixes: The "1-arithmetic prefix" (1-A) The "2-arithmetic prefix" (2-A)
GT	11
Judgment	Response B [GT] provides a direct and accurate answer to the question, while Response A is overly complicated and doesn't provide the correct answer.

表 10: GPT-4 选择了 GT 而非 DPO。GPT-4 错误地表示地面真实值是正确的，而 DPO 的（更为冗长的）输出是错误的。

D.3 Human study details

为了验证使用GPT-4计算胜率的有效性，我们的人类研究在TL;DR摘要设置中收集了几个人工智能对决的人类偏好数据。我们选择了三种不同算法的对决，评估了DPO（温度0.25）、SFT（温度0.25）以及PPO（温度1.0）与参考算法PPO（温度0.）的比较。通过选择三种独特算法以及与参考算法具有广泛胜率范围的对决，我们捕捉到了人类和GPT-4胜率在回应质量谱系中的相似性。我们对DPO与PPO-0的150个随机比较以及PPO-1与PPO-0的100个随机比较进行了抽样，每组对决分配了两个人类评委，为DPO-PPO产生了275个判断⁽⁷⁾以及为PPO-PPO产生了200个判断。我们对125个SFT比较进行了抽样，每个比较分配了一个人类评委。我们忽略了人类标记为平局（大约只占1%的判断）的判断，并测量了人类A和人类B之间的原始同意百分比（在我们有两个人类注释者的比较中，即非SFT），以及每个人类与GPT-4之间的同意百分比。

Summarization Evaluation [id ZHBvX3RlbXAwLjAx; group 5; key 18209903]

Which of the following summaries does a better job of summarizing the most important points in the given forum post?

Some responses may be very similar; please do your best to compare them and only use the "I can't tell" option rarely, if at all.

6. Which of the following summaries does a better job of summarizing the most important points in the given forum post?

Post:

My boyfriend and I have been together for 4 years, but I'm becoming tired of his childish hobbies. Two days ago he spent over \$100 on these Nintendo toys and game, but this isn't the worst part. He has a "toy room" and it's lined with "very expensive" action figures from video games, Legos and cartoons, some that I consider quite lewd for someone in a relationship. All together I'm pretty sure he's spent thousands of dollars all together in that room, not including his video game collection. Over this past month he probably brought 8 different games for his Playstation and I think that was overboard.

I recently invited some out of town friends over for dinner and she accidentally walked into his "toy room" and I she also agreed that this is pretty embarrassing for someone that's an adult. He makes decent money, a lot more than me but I think it's time for him to give up and sell these things so he can finally move on and become an adult with me. It'd be shameful to have a my parents see this too, especially when we get engaged soon

How should I approach this /r/relationships?

- ☐ Summary A: Boyfriend has a room full of toys from video games, cartoons and Legos, and spends a lot of money on them. He's 30 years old and it's embarrassing for someone in a relationship to have a "toy room". What should I do /r/relationships?
- ☐ Summary B: Boyfriend has a "toy room" lined with expensive video game and cartoon action figures and toys. I think it's time for him to give up his childish hobbies and become an adult with me. How should I approach this?
- ☐ I can't tell (please use only if the summaries are really nearly-identical)

图 5: SurveyMonkey中调查的布局。每位受访者完成了25个格式相似的判断。

参与者。 总共有25名志愿者人类评分者，每人比较25个摘要（有一名志愿者晚完成调查，没有包含在最终分析中，但在此处列出）。评分者是斯坦福大学的学生（从本科到博士），或是最近的斯坦福毕业生或访问者，以STEM（主要是计算机科学）为主。调查界面的截图见Figure 5。我们衷心感谢我们每一位志愿者的贡献，以下名单按随机顺序列出：

- | | | | |
|------------------------|-------------------|------------------|------------------------|
| 1. Gordon Chi | 2. Virginia Adams | 3. Max Du | 4. Kaili Huang |
| 5. Ben Prystawski | 6. Ioanna Vavidou | 7. Victor Kolev | 8. Karel D'Oosterlinck |
| 9. Ananth Agarwal | 10. Tyler Lum | 11. Mike Hardy | 12. Niveditha Iyer |
| 13. Helena Vasconcelos | 14. Katherine Li | 15. Chencheng Gu | 16. Moritz Stephan |
| 17. Swee Kiat Lim | 18. Ethan Chi | 19. Kaien Yang | 20. Ryan Chi |
| 21. Joy Yun | 22. Abhay Singhal | 23. Siyan Li | 24. Amelia Hardy |
| 25. Zhengxuan Wu | | | |

⁷一位志愿者没有对DPO-PPO比较做出回应。