人类与*PO 的实践分析

Kian Ahrabian[1]* **Xihui Lin**[2] **Barun Patra**[2]

Vishrav Chaudhary[2] **Alon Benhaim**[2] **Jay Pujara**[1]

Xia Song[2]

[1]University of Southern California, Information Sciences Institute

[2]Microsoft

ahrabian@usc.edu,{xihlin,barun.patra@microsoft.com}

{vchaudhary,alonbenhaim}@microsoft.com,jpujara@isi.edu,xiaso@microsoft.com

Abstract

处于最先进的对齐人类方法前沿的是偏好优化方法 (*PO)。以往的研究通常集中在识别表现最佳的方法上，这通常涉及超参数的网格搜索，对于一般从业者来说可能不切实际。本文探讨了现有最先进方法在现实世界分布外 (OOD) 场景下的稳健性，该场景模拟了对齐人类的实际应用。我们的目标是通过各种度量标准（如 KL 散度和响应长度）的经验分析找到能增加获得更好结果可能性的方法。我们还介绍了 LN-DPO，这是 DPO 的一种简单的长度归一化版本，在超参数变化中更稳定，有效减少了平均响应长度并提升了性能。我们对最先进的无参考 (即，SimPO) 和有参考 (即，DPO 和 LN-DPO) 方法的分析表明，它们在最佳可能情况下的表现相似 (即，最佳可能情况)。然而，我们发现当偏离最佳可能情况时，性能变化模式有很大的不同。

1 介绍

近年来，大型语言模型（LLMs）的质量一直在不断提高 (Chiang et al., 2024)，在各种任务和基准测试中取得了令人印象深刻的结果 (Abdin et al., 2024; AI@Meta, 2024; Achiam et al., 2023; Team, 2023; Yang et al., 2024)。然而，即使是最严格的过滤启发式方法也无法避免训练数据 (Computer, 2023; Penedo et al., 2024) 通常被不希望的内容污染，这些内容可能导致不可接受的行为 (Bender et al., 2021; Gehman et al., 2020)。为了提高模型与人类偏好的一致性，事实上的做法是从人/AI 生成的偏好数据中学习（例如，每个提示选择一个被接受和一个被拒绝的响应）。特别是，鉴于其良好的性能和易于实现，非策略偏好优化方法 (*PO) 一直很流行 (Rafailov et al., 2024; Hong et al., 2024; Meng et al., 2024)。

	DPO	LN-DPO	相似性优化
平均得分	1.6	+0.3%	+2.7%
平均长度	119.8	-15.9%	-22.9%
KL 散度	55.0	-26.0%	-20.7%
胜者对阵。选定的	77.1%	+0.8%	+3.1%
胜者 vs. SFT	60.7%	+2.1%	+5.0%

Table 1: **最佳*采购订单绩效**指标通过相应的 DPO 性能进行了标准化。下划线的值表示最佳性能。

在报告新方法的性能时，一个常见的做法是比较它们的最佳表现变体（经过超参数网格搜索后）与一组固定超参数的默认基线。然而，从未来用户的实际角度来看，这些比较并没有很好地回答这样一个问题：给定固定的超参数搜索预算，哪种方法有望实现更高的性能。因为广泛的网格搜索通常对于许多从业者来说计算上不可行。为此，在这项工作中，我们旨在通过实验识别出在面对超参数变化时更为稳健且仍然具有竞争力的方法。

我们在现实的分布外（OOD）设置中建立

*Work done during an internship at Microsoft.

方法	目标	超参数
DPO	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$	$\beta\in\{0.01,0.05,0.1,0.3,0.5\}$
SimPO	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\pi_\theta(y_w\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$	$\beta\in\{1.0,1.5,2.0,2.5\}$ $\gamma\in\{0.5,0.8,1.0,1.2,1.4,1.6\}$
LN-DPO	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\frac{\beta}{\|y_l\|}\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$	$\beta\in\{1.0,1.5,2.0,2.5,3.0,3.5\}$

Table 2: ***优化目标**。偏好数据被定义为 $D=(x,y_w,y_l)$，其中 x 是提示，y_w 和 y_l 分别是选择的和拒绝的响应。

了实验，专注于安全性和有用性领域，在该设置下训练和测试数据集具有共同的核心目标，但它们的样本来自不同的分布（例如，AI 和人类专家）。这种设置类似于现实世界的情况，因为它模拟了大型生成模型对公众发布的场景。此外，为了更好地理解最先进的模型的行为，我们选取了表现最佳的无参考和有参考模型（如 Meng et al. (2024) 所报告），并通过标准指标如 KL 散度、响应长度和胜率来分析它们。我们还引入了一个非常简单的长度归一化扩展的普通直接偏好优化（DPO）(Rafailov et al., 2024)，即 LN-DPO，它有效地解决了生成内容过长的问题而不降低任何明显的性能 [1]。总结来说，我们的贡献如下：

- 我们考察了在真实环境设置中广泛范围的超参数下的最先进的无参考和有参考依赖的偏好优化方法。
- 我们分析了这些方法在关键指标上的性能，如平均响应长度、在黄金奖励模型上的平均得分、与选定方法和 SFT 的胜率比较，以及与 SFT 的 KL 散度比较。
- 我们介绍了并研究了 LN-DPO，这是 DPO 的一个简单的长度归一化版本，在超参数上更稳定，有效减少了平均响应长度，并提高了性能。

2 相关工作

自引入 DPO (Rafailov et al., 2024) 以来，出现了一系列具有新优化目标的作品，这些作品改进了性能和效率 (Azar et al., 2024; Tang et al., 2024; Hong et al., 2024; Rosset et al., 2024; Meng et al., 2024; Xu et al., 2024a; Ethayarajh et al., 2024)。这些方法可以分为两类：无参考的 (Meng et al., 2024; Hong et al., 2024) 和有参考的 (Rafailov et al., 2024; Park et al., 2024)。无参考的方法通常得益于快速的训练运行，而有参考的方法则在它们的目标中包含控制与参考模型偏差的项。在这项工作中，我们将最近的状态-of-the-art 无参考方法 SimPO (Meng et al., 2024) 与作为有参考的方法 DPO 和 LN-DPO 进行了比较（有关扩展的相关工作，请参阅 Appendix A）。

3 实验设置

3.1 数据集

对于我们的数据集，我们遵循 Xu et al. (2024b) 中引入的设置。具体来说，我们在 SafeRLHF (Dai et al., 2024) 的安全/不安全过滤训练子集上进行训练，并在 HH-RLHF (Ganguli et al., 2022) 的测试子集上进行评估。这种设置与现实世界的情景非常相似，在这些情景中，尽管模型是在各种领域（我们在实验中的例如，安全性及有用性）上训练的，但它们必

[1] 同时，Meng et al. (2024) 在其实验中添加了类似的方法（更新于 2024 年 7 月 7 日）。这里，我们呈现了一个更全面的分析和比较。

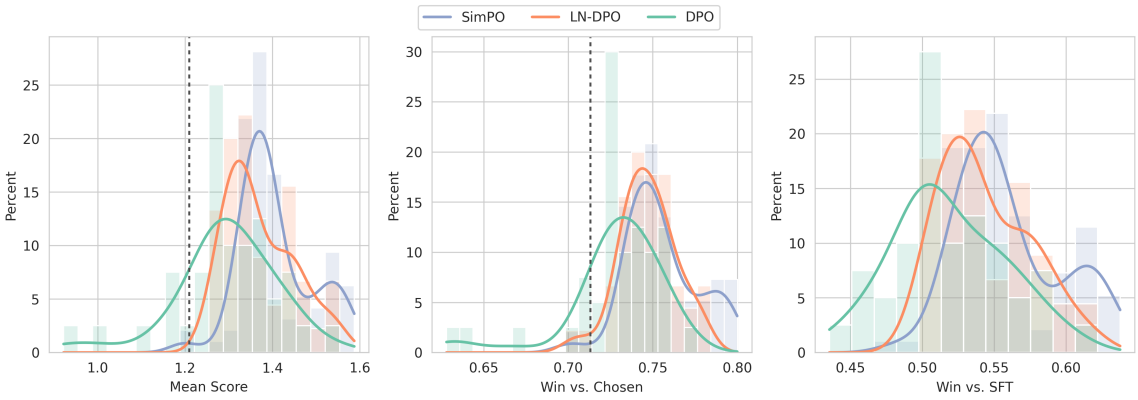

Figure 1: ***PO 绩效分布**分布中的每个样本代表一组超参数在指定指标上的性能。虚线表示初始 SFT 模型的性能。

须能够推广到类似的新颖查询并在与用户交互时做到这一点。

3.2 模型

对于我们的所有实验，我们选择了 Phi-3 Medium 模型 (Abdin et al., 2024)，因为其在基准测试中的高性能和较小的尺寸，确保了计算上的可行性。为了评估训练好的模型，我们使用 OpenAssistant 奖励模型 (Köpf et al., 2024) 来评分它们生成响应的质量。我们选择此模型是因为它的尺寸较小并且在先前的工作中已被使用 (Xu et al., 2024b)，确保了快速且正确的评估。

3.3 优化目标

考虑到由 Meng et al. (2024) 报告的性能，我们选择 DPO 作为我们的依赖参考方法，并将 SimPO 作为我们的不依赖参考的方法。虽然 DPO 通过参考模型具有隐式长度归一化，但奖励（即，$\log \frac{\pi_\theta}{\pi_{\text{ref}}}$）的方差随响应长度增加。因此，受 SimPO 和 R-DPO (Park et al., 2024) 中的显式长度正则化的启发，我们进一步类似于 SimPO 用响应长度对其进行归一化，我们将这种方法称为 LN-DPO（请参见 Section 3.4以获取更多详细信息）。

3.4 LN-DPO 与 SimPO 之间的联系

LN-DPO 类似于 SimPO 的自适应边距版本，其中每个样本的边距定义为

$$\gamma_{w,l} = \log \frac{\pi_{\text{ref}}(y_w|x)}{|y_w|} - \log \frac{\pi_{\text{ref}}(y_l|x)}{|y_l|}. \quad (1)$$

本质上，这种自适应边距鼓励对参考策略中具有大边距的配对使用更大的边距。根据参考模型的质量和标签，与 SimPO 的固定边距相比，这一变化可能是有益的。自适应边距更关注“较容易”的配对（即，配对具有某些先验证据表明它们不同），而较少关注“较难”的配对（即配对距离较近），这意味着 LN-DPO 潜在地不太容易过拟合，且对错误标签的敏感度较低。

4 训练方案

遵循常见做法，在偏好优化步骤之前，我们进行一个监督微调（SFT）步骤。具体来说，我们首先在以下超参数上运行网格搜索：epochs$\in \{1, 3\}$ 和学习率 $\in \{1e-6, 3e-6, 1e-5, 2e-5\}$。然后我们将最终的检查点与测试集进行评估，并选择性能最高的那个。这个过程确保了偏好优化方法从一个好的检查点开始初始化。对于偏好优化方法，我们使用以下范围运行网格搜索：1）epochs 和学习率与 SFT 相同的范围和 2）如先前工作中所用的方法特定超参数的常见值 (Meng et al., 2024; Rafailov et al., 2024; Hong et al., 2024)。Table 2展示了我们在实验中使用

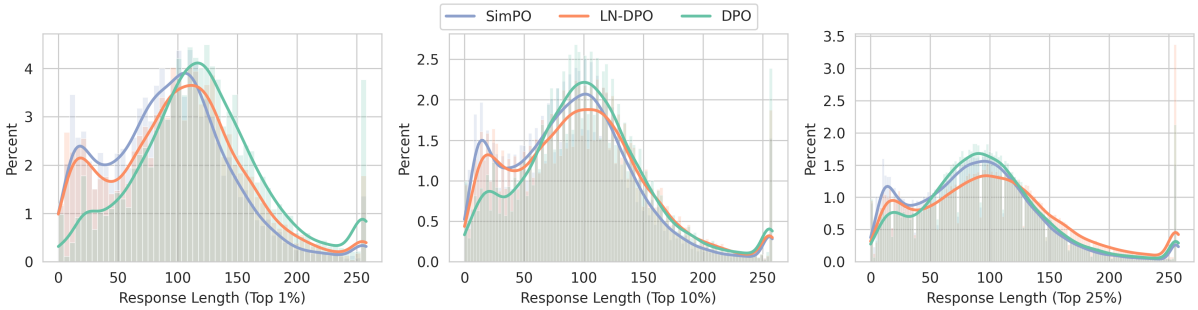

Figure 2: **响应长度**前 k%（$k \in \{1, 10, 25\}$）表示从每种方法的运行中选取的最佳超参数的百分比。

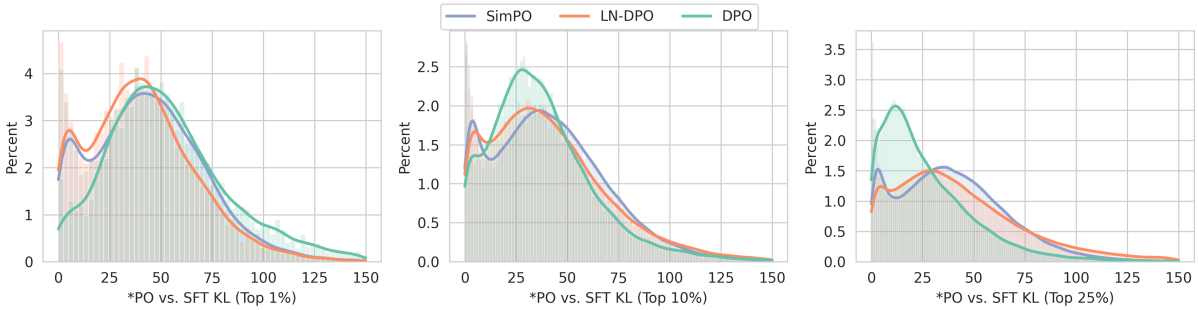

Figure 3: **KL 散度**前 k%（$k \in \{1, 10, 25\}$）表示从每种方法的运行中选取的最佳超参数的百分比。

的方法特定范围。在我们所有的实验中，批量大小设置为 256。

5 指标

我们的分析集中在以下五个指标上：

- **平均得分：**由黄金奖励模型判定的生成响应的平均分数。
- **胜者 vs. 选定：**奖励模型分配给生成响应的分数高于数据集中选定响应的比例。
- **胜者 vs. SFT:** 样本中黄金奖励模型对生成的响应评分高于初始 SFT 模型响应的比例。
- **KL 散度：**SFT 和训练模型在样本上的对数概率之差的总和。
- **响应长度：**在基础模型的标记化空间中生成响应的标记数量。

6 实现细节

我们通过采样生成所有响应，使用了温度 $= 0.7$，和顶部_p $= 0.95$。此外，在所有实验中，最大生成长度设置为 256，遵循 Xu et al. (2024b) 的设置。我们的所有实验都在一个集群上进行，该集群配备了 256 个 ×A100 80GB GPU。最后，我们使用了 Transformers (Wolf et al., 2020)、TRL (von Werra et al., 2020) 和 PyTorch (Paszke et al., 2019) 库实现了我们的代码。

7 实验结果

7.1 超参数鲁棒性

最佳性能。 遵循常见做法，我们将每种方法达到的最佳性能进行比较，在 Table 1中。显而易见，在峰值时，SimPO、LN-DPO 和 DPO 的得分相似（平均相差不到 0.05 分）。然而，SimPO 和 LN-DPO 在其他指标上表现出优势。具体来说，我们可以观察到长度归一化项的有效性。我们还注意到 KL 散度有显著下降。但是 SimPO 的 KL 下降幅度小于 LN-DPO，显示出与 SFT 有更大的偏离。有关调整这些模型的更多细节，请参见 Appendix B。

一对一性能。 尽管通常情况下，仅比较在所需指标上实现的纯性能足以对比不同的方法，但在某些潜在情况下，可以利用平均值（例如，

%	动态规划优化	LN-双酚氧杂环丁烷	相似性优化
DPO	-	49.04	47.51
LN-DPO	49.47	-	46.43
相似性优化	51.12	51.09	-

(a) 最佳

%	DPO	LN-DPO	模拟 PO
动态规划优化	-	45.72	44.33
LN-DPO	51.77	-	47.28
相似性优化	54.34	50.13	-

(b) 第 75 百分位数

Table 3: **头对头*PO 比较**。每个单元格表示行方法相对于列方法的胜率。下划线值表示行方法击败了列方法。

奖励高的异常值)。因此，进行一对一的样本比较也至关重要，这提供了更细致的见解。Table 3 比较了每种方法的最佳性能和第 75 百分位数的性能。值得注意的是，我们观察到 DPO 从最佳模型到前 25%模型的性能急剧下降，与其他两种情况相反。这一现象突显了仅比较最佳性能的实际缺陷。

预期性能。 鉴于大多数用户资源有限，进行广泛的超参数搜索以找到最佳组合极其困难。因此，分析超参数的鲁棒性变得至关重要，这提供了从有限搜索中找到良好超参数集的期望洞察。Figure 1展示了在 Table 2和 Section 4所表示的超参数上进行网格搜索后 *PO 方法的表现分布。显然，SimPO 和 LN-DPO 有效提高了平均性能（即向右移动分布），展示了它们的优势。请注意，我们扩展了超参数范围直到观察到平台期或极端方差。

7.2 响应长度

由于长度利用是一个关键问题 (Park et al., 2024)，我们将由每种方法的最佳超参数生成的样本响应长度进行了比较（$k \in \{1, 10, 25\}$ 的前 k%）。如 Figure 2所示，在最佳超参数集（即，前 1%）中，非 DPO 方法显示出长度分布向左偏移的效果（与 DPO 相比），这是期望的结果。然而，当我们包含表现较差的超参数时，这种现象开始减弱。例如，在顶部 25%分布的尾端，LN-DPO 的比例高于 DPO。总体而言，我们观察到两种长度归一化的模型都优于 DPO，其中 SimPO 在整个分布中产生的响应最短。

7.3 KL 散度（与 SFT 比较）

由于无参考方法没有以参考策略（例如，SFT 模型）进行归一化，可能会发生奖励操纵（即，性能下降但仍降低损失）。因此，我们比较了由每种方法的最佳超参数生成的前 k%（$k \in \{1, 10, 25\}$）样本中的 KL 散度（Figure 3）。显然，SimPO 和 LN-DPO 在其峰值时都实现了较低的 KL 值。然而，当我们转向性能较差的模型时，DPO 在 10%处实现了较低的 KL 值。这一现象是由于许多 DPO 运行未能超越 SFT 模型的学习。

8 何时使用 LN-DPO 而非 SimPO？

虽然 SimPO 在大多数指标上相较于 LN-DPO 表现出更优的性能，但缺乏参考策略正则化可能导致与初始检查点有极大的偏离，这一点也在我们的实验中得到体现。这一问题进而可能会导致在其他基准测试上的性能下降，这是一个关键的缺陷（同样在 Korbak et al. (2022) 中观察到）。因此，我们认为存在多种场景下应优先选择 LN-DPO 而非 SimPO。我们将在未来的工作中进一步探索这一方向的实验。

9 结论

在这项工作中，我们介绍了 LN-DPO，这是一种长度归一化的 DPO 变体，在保持依赖参考的同时减少了平均响应长度。此外，我们在模拟的真实世界场景中对安全性和有用性领域内的 LN-DPO 以及两种最先进的依赖参考和不依赖参考的偏好优化方法进行了全面分析。具体来说，我们在诸如平均响应长度、KL 散度（相对于 SFT）和胜率（相对于选定项和 SFT）等指标下，探讨了这些方法在广泛超参数范围

内的行为。我们的实验展示了最先进方法的优势和劣势，并为其他从业者提供了见解。

限制条件

由于此类实验的运行成本极高（即，大约 86000 个 GPU 小时用于当前实验），在这项工作中，我们仅对一小部分模型、方法和数据集进行了实验。虽然这可能会限制可泛化性，但我们认为存在这样的分析对于帮助从业者节省成本至关重要。此外，自我们的实验得出结论以来，已经发布了性能更高的新奖励模型（例如，ArmoRM (Wang et al., 2024)）；然而，我们仍然依赖于较旧、较小的模型，以在如此多的运行次数上保持评估的可追踪性。

致谢

本工作部分由国防高级研究计划局通过拨款 HR00112220046 资助。

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

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*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

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*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. 2024a. The n+ implementation details of rlhf with ppo: A case study on tl; dr summarization. *arXiv preprint arXiv:2403.17031*.

Shengyi Costa Huang, Tianlin Liu, and Leandro von Werra. 2024b. The n implementation details of rlhf with ppo. In *ICLR Blogposts 2024*. Https://d2jud02ci9yv69.cloudfront.net/2024-05-07-the-n-implementation-details-of-rlhf-with-ppo-130/blog/the-n-implementation-details-of-rlhf-with-ppo/.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. 2022. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1083–1091, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Guilherme Penedo, Hynek Kydlíek, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

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. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024b. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A 扩展相关工作

在线算法。 从人类/AI 反馈中进行强化学习（RLHF/RLAIF）是使大语言模型与人类偏好对齐的常见方法 (Christiano et al., 2017; Bai et al., 2022a; Stiennon et al., 2020; Bai et al., 2022b)，并且已被用于训练如 GPT-4 (Achiam et al., 2023) 和 Llama-3 (AI@Meta, 2024) 等模型。在大多数情况下，这些方法包括三个阶段：1) 监督微调 (Taori et al., 2023; Zhou et al., 2024; Xia et al., 2024)，2) 奖励建模 (Gao et al., 2023; Chen et al., 2024; Lightman et al., 2023)，和 3) 策略优化 (Schulman et al., 2017)。政策优化的突出方法是近端策略优化（PPO），这是一种在线在策略方法 (Schulman et al., 2017)。尽管 PPO 表现出有希望的表现 (Stiennon et al., 2020; Ouyang et al., 2022; Achiam et al., 2023)，但它存在诸如重现性细节过多 (Huang et al., 2024b)、2) 训练时间过长 (Huang et al., 2024a) 以及 3) 奖励过度优化 (Skalse et al., 2022) 等问题。

离线算法。 为了克服 RLHF/RLAIF 的缺点，最近的研究提出了更简单且高效的离线算法，特别是基于 Bradley-Terry 模型 (Bradley and Terry, 1952) 的直接偏好优化（DPO）(Rafailov et al., 2024)。这些离线算法直接在偏好数据上优化目标，使用隐式奖励模型而无需单独的阶段。一些最近的研究集中在 PPO 和 DPO 之间的广泛比较上。具体来说，它们展示了带有黄金奖励模型（$\sim +10\%$）的 PPO 的潜力，同时指出当在同一数据上训练时与 DPO（在基准测试中平均的 $\sim +1\%$）的相似性 (Ivison et al., 2024; Xu et al., 2024b)。

B 超参数调整考虑因素

DPO. 如 Figure 4所示，较低的 β 导致更高的性能；然而，当 β 减小时，性能方差增加，这展示了该方法的不稳定性。总体而言，$\beta = 0.05$ 提供了稳定性和性能的最佳平衡。

LN-DPO. 虽然我们最初从 SimPO (Meng et al., 2024) 借用了 β 的范围，但更多的实验表明进一步减少其值是有益的。Figure 5展示了不同运行中的性能分布。从这些实验中可以看出，$\beta \in [1.0, 2.0]$ 包含了大多数表现最佳的模型。

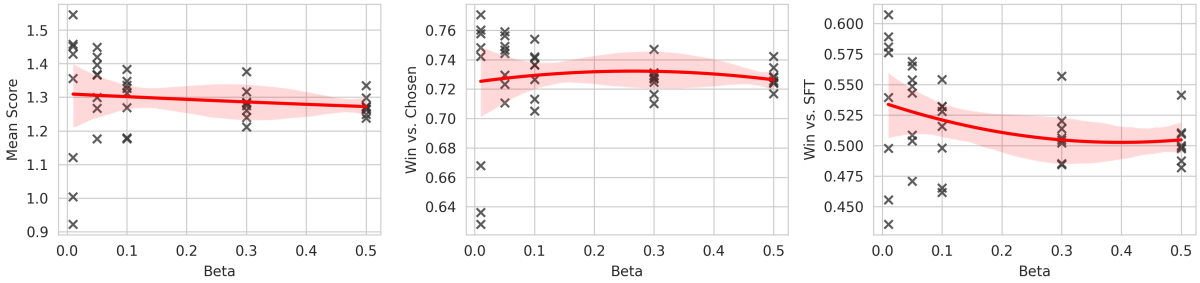

Figure 4: **DPO** β 每个点表示一个具有相应 β 值的运行。

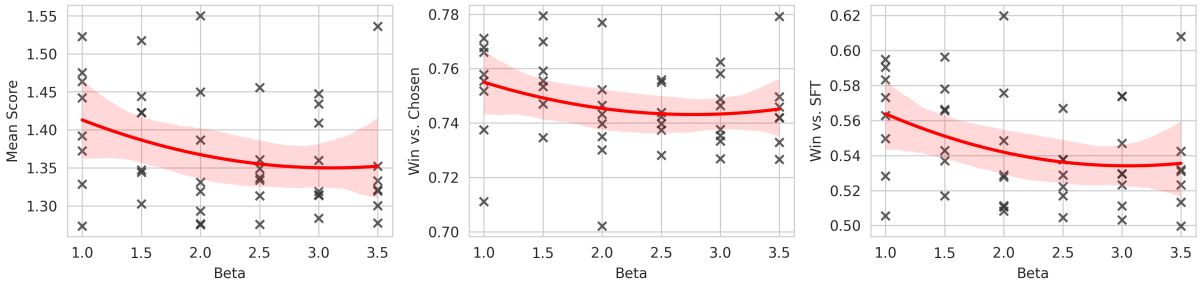

Figure 5: **LN-DPO** β 每个点表示一个具有相应 β 值的运行。

此外，我们观察到与 DPO 相比，性能方差相对较低，展示了 LN-DPO 的另一项优势。

模拟粒子优化 与另外两种方法相比，SimPO 有两个特定的方法超参数：β 和 γ。如 Figure 6 所示，平均而言，较低的 β 值导致更好的性能。我们相信在较低范围内的性能提升是由于本工作与原工作的训练集平均长度不同所致。此外，如 Figure 7所示，表现最佳的模型具有一个 $\gamma \in [1.0, 1.4]$，这与 Meng et al. (2024) 的建议一致。值得注意的是，β 和 γ 在实验中具有相对较低的方差，这是 SimPO 的另一个优点。

C 终极问题的答案

基于我们的集体实证结果，我们认为 SimPO 是在三种方法中最好的起点，主要是因为它在超参数变化方面的鲁棒性和有效的长度缩减。至于 SimPO 的超参数，我们推荐使用 $\beta \in \{1.0, 1.5\}$ 和 $\gamma \approx 1.2$。此外，虽然 LN-DPO 在我们大多数实验中始终是第二好的选择，但我们讨论了在 Section 8中选择它而非 SimPO 的情景。

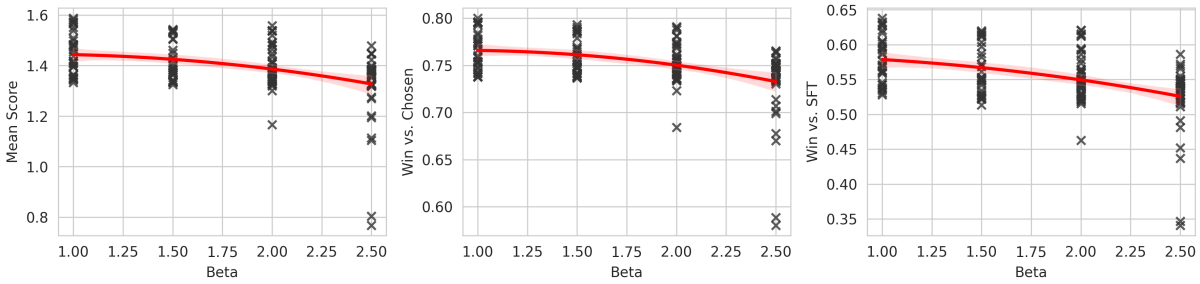

Figure 6: **相似性优化** β 每个点表示一个具有相应 β 值的运行。

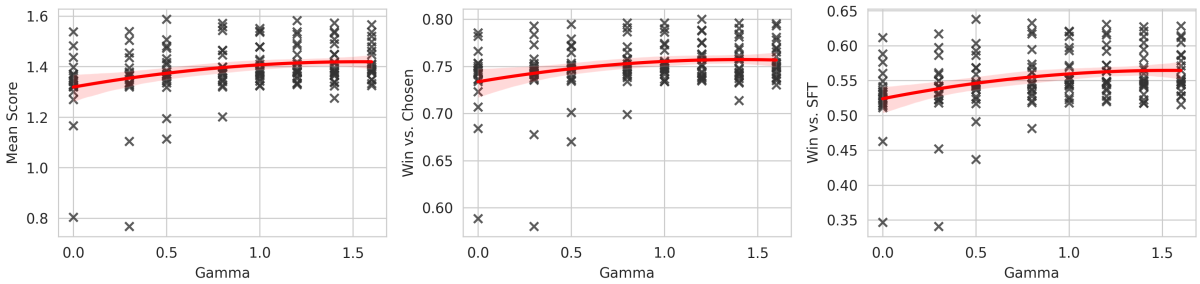

Figure 7: **相似性优化** γ 每个点表示一个具有相应 γ 值的运行。