arXiv:2602.00954v1 [cs.AI] 1 Feb 2026

Small-边际偏好依然重要——前提是训练得当

Jinlong Pang [1] Zhaowei Zhu [2] Na Di [3] Yichi Zhang [4] Yaxuan Wang [1] Chen Qian [1] Yang Liu [1]

摘要

直接偏好优化等偏好优化方法通过成对比较对齐大语言模型，但其效果可能对偏好对的质量和难度高度敏感。一种常见的经验做法是将小边际（模糊）偏好对视为噪声并过滤掉。本文重新审视这一假设，并表明样本对难度与优化目标函数之间存在很强的交互作用：使用基于偏好的损失进行训练时，困难偏好对可能导致训练不稳定并损害对齐；然而，使用监督微调进行优化时，这些偏好对仍包含有用的监督信号。受此观察启发，我们提出**MixDPO**，这是一种简单而有效的难度感知训练策略：(i) 将偏好数据按从易到难的顺序排列（按边际定义难度的课程学习）；(ii)将路由到监督微调目标函数，同时对简单样本对应用偏好损失。这种混合设计提供了一种实用机制，能够利用模糊偏好对，同时避免低边际数据上的偏好损失常引发的优化失败。在三个大语言模型评判基准上，MixDPO 相较于直接偏好优化及多种广泛使用的变体，始终能提升对齐效果；在 AlpacaEval 2 长度控制（LC）胜率上，提升尤为显著。

1. 引言

从人类反馈中学习对于使大语言模型（LLM）与人类偏好对齐至关重要，有助于确保这些模型以有益、诚实且无害的方式行事（Achiam 等人，2023年；Nakano 等人，2021）。这类对齐任务中广泛采用的一种方法是基于人类反馈的强化学习（RLHF）（Stiennon 等人，2020；Ouyang 等人，2022年），该方法包含一个由大语言模型微调和奖励模型训练组成的多阶段流程。为简化这一复杂过程，研究者提出了若干离线策略和无奖励模型方法，包括直接偏好优化（DPO）（Rafailov 等人，2024b）及其变体，如 KTO（Ethayarajh 等人，2024年）和 SimPO（Meng 等人，2024年）等。这些方法绕过在线强化学习，直接在由偏好对构成的固定数据集上进行训练 $\{(x,\ y_w,\ y_l)\}$，其中 y_w 和 y_l 分别表示针对提示给出的偏好响应和非偏好响应x。与基于人类反馈的强化学习中采用的开放式探索不同，这些方法高度依赖底层偏好数据的质量，而这对于实现出色的对齐性能至关重要。

尽管人们投入了大量精力来整理后训练数据，无论是在样本层面（Chen 等，2023b；Xia 等，2024年；Pang 等人，2024b；Liu 等，2023年），还是在词元层面（Lin 等，2024年；Pang 等人，2025），数据在偏好对齐中的作用仍在很大程度上被忽视，且缺乏深入研究。在数据选择思想的基础上，近期工作（Deng 等，2025；Huang 等，2025；Gao 等人，2025）基于每组成对数据中偏好响应与非偏好响应之间的间隔，研究成对数据质量，旨在识别出容易处理且间隔较大的数据对。相比之下，那些困难的（间隔较小的）数据对通常会基于这样的直觉被丢弃：它们会引入歧义或噪声，从而妨碍有效的偏好建模。

这一理由源于如下观察：此类困难偏好对往往会引发似然偏移现象（Pal 等人，2024年；Yuan 等，2024年；Rafailov 等人，2024a年；Razin 等人，2024年），其中偏好响应 y_w 与非偏好响应 y_l 的对数概率均下降，这与偏好学习的初始目标相矛盾。潜在原因在于，相似响应之间的细微差异可能导致梯度不稳定以及优化失准。现有大多数方法试图通过重新设计 DPO 风格损失函数来缓解该问题，例如 Cal-DPO（Xiao 等人，2024年）和 DPOP（Pal 等人，2024年）。相比之下，从以数据为中心的视角出发，我们提出一个有趣的问题：“这些通常被丢弃的困难偏好对，若得到恰当利用，是否仍能提供有价值的监督信号？”

在这项工作中，我们对传统做法提出质疑，

[1]加州大学圣克鲁兹分校 [2]Docta.ai [3]东北大学 [4]DIMACS，罗格斯大学。通信作者：Yang Liu <yangliu@ucsc.edu>。

预印本。*2026*年*2*月*3*日。

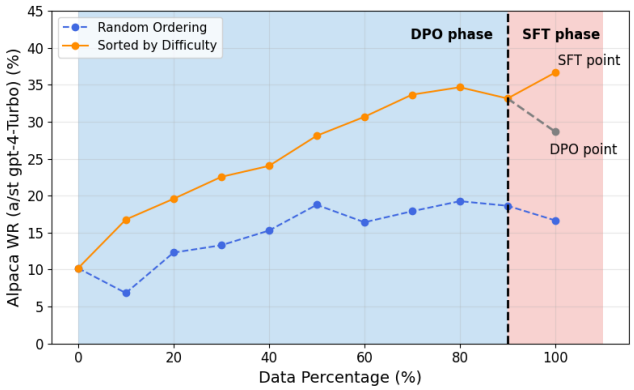


图 *1.* 在 AlpacaEval 2 基准上，随机排序与基于难度排序的性能比较。直接偏好优化被用作默认损失函数。值得注意的是，与其丢弃困难偏好对，不如使用监督微调损失继续训练它们，从而带来性能提升。

并提出一个新的见解：“尽管困难偏好对在使用基于偏好的目标函数进行优化时，可能会因潜在的似然偏移而阻碍对齐，但通过监督微调（*SFT*）进行训练时，它们仍能提供有价值的学习信号。”

我们在图1 中的实证实验支持两项关键发现：（1）鉴于偏好对之间难度存在自然差异，与随机排序相比，按照从易到难的顺序组织训练能够带来更好的对齐性能，这与课程学习范式一致；（2）尽管困难偏好对往往会损害基于偏好的优化，但在使用监督微调损失进行训练时，它们会变得有益。最终，在实证证据的支持下，我们提出了一种简单而有效的方法，可根据偏好对的难度自适应地从直接偏好优化切换到标准监督微调损失。具体而言，我们对较容易的偏好对应用 DPO 损失进行优化，而对更具挑战性的偏好对则采用监督微调损失，以更好地利用其信息价值。重要的是，监督微调损失提供了一种避免似然偏移的直接方法。

我们将主要贡献总结如下。

- 我们系统研究了容易和困难偏好对在对齐性能中的作用，发现了一种一致的规律：随着训练数据从容易偏好对转向困难偏好对，性能会下降。

- 针对通常会丢弃分数差距较小的困难偏好对这一做法，我们提出了一个新见解：实证证据表明，如果采用监督微调（SFT）对这些偏好对进行优化，它们仍然能够提供有价值的学习信号。受此启发，我们提出了一种简单而有效的方法 **MixDPO**：对简单样本对应用 DPO 损失，对困难偏好对应用监督微调损失，从而更好地利用信息丰富的信号。

- 我们进行了广泛的实验，以验证 **MixDPO** 相较于 DPO 及其多种广泛使用的变体（包括 CPO、IPO、KTO、SimPO 和 SelectiveDPO）的有效性。值得注意的是，引入的监督微调阶段有助于抵消偏好偏差导致的过度生成长回复倾向，AlpacaEval 2 长度控制胜率的提升也证实了这一点。针对额外基础模型和 Agrilla-7k 数据集开展的广泛消融研究，证明了我们方法的通用性和适应性。

2. 相关工作

大语言模型偏好对齐 始于相关工作（Ouyang 等人，2022年），此后，人们提出了众多方法，使大语言模型生成的回答与人类偏好保持一致。这些方法大致可分为两种范式：基于人类反馈的强化学习和直接偏好优化（Rafailov 等人，2024b）。作为基于人类反馈的强化学习的一种简化且无需奖励模型的替代方案，直接偏好优化已成为应用最广泛的偏好对齐技术之一。越来越多的研究从各种理论和实践角度分析直接偏好优化损失，并由此提出了若干值得关注的变体，包括 IPO（Azar 等人，2024年）、KTO（Ethayarajh 等人，2024年）、CPO（Xu 等人，2024a）和 SimPO（Meng 等人，2024年）等。

尽管人们已在样本层面（Chen 等，2023b；Xia 等，2024；Zhang 等人，2025；Pang 等，2024b；Yang 等人，2025；Liu 等，2023；2024）和词元层面（Lin 等，2024；Pang 等，2025）投入大量精力研究指令微调和预训练中的数据选择，但数据对偏好对齐的影响仍被忽视，尚未得到充分探索。此前有几项研究（Bai 等人，2022；Pang 等，2024a；Ethayarajh 等人）考察了数据在对齐过程中的影响，但主要着眼于如何使模型与人类价值观与伦理相一致。近期的一些工作（Deng 等，2025；Huang 等，2025；Gao 等，2025）则根据奖励分数等不同指标评估数据质量，旨在保留信息丰富的样本，同时剔除被认为信息不足或含有噪声的样本。另一类工作专注于开发抗噪声 DPO 目标函数，例如 cDPO（Mitchell）、robustDPO（Chowdhury 等，2024）和 PerpCorrect（Kong 等人），但并未研究单个样本的影响。与此相反，我们的工作对这种非此即彼的数据质量观提出质疑。我们认为，在基于偏好的优化中通常会被过滤掉、看似信息不足或难度过高的样本，若采用监督微调目标函数进行优化，仍能提供有价值的信息。

课程学习 课程学习（CL）遵循人类自然的学习模式，通过从简单概念到复杂概念来组织学习（Avrahami 等人，

1997年；Bengio 等人，2009年），这可以有效加速模型收敛并增强泛化能力。受其成功启发，课程学习模式已被应用于多个领域，包括机器翻译（Platanios 等人，2019）、图像生成（Croitoru 等人，2024b）以及多媒体搜索（Jiang 等人，2014；Tudor Ionescu 等人，2016）。

在大语言模型的偏好对齐中，课程学习的一个核心环节是定义偏好对的样本难度。近期研究探索了多种难度评分指标，以支持基于课程学习的训练，例如提示长度或内在注意力分数（Kim& Lee，2024年）、模型对响应的困惑度（Wu 等人，2024年）、强大奖励模型估算的奖励边际（Croitoru 等人，2024a），以及偏好对上的验证集 DPO 损失（Gao 等人，2025）。在这项工作中，我们采用一种简单而有效的策略，利用偏好数据中已有的原始评分来估算难度，不会产生额外的计算开销。此外，虽然Gao 等人(2025)选择过滤掉过于困难的样本，因为这些样本可能产生负面影响，但我们提出了一种替代方法，并证明使用监督微调损失对这些样本进行优化，仍能为对齐带来益处。

3. 预备知识

3.1. 直接偏好优化（DPO）

偏好对齐（Ouyang 等人，2022年）旨在确保大语言模型生成的输出能够体现人类偏好和交流风格，从而提升其在真实场景中的安全性、可靠性和可信度。在本研究中，我们采用 DPO（Rafailov 等人，2024b），这是最广泛使用的基于偏好的对齐方法之一。给定一个成对偏好数据集 $\mathcal{D}:=(x,\ y_w,\ y_l)$，其中$x$ 是一个提示，y_w 是偏好响应，而 y_l 是不太受偏好的响应，DPO 使用以下目标函数训练策略模型 π_θ：

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

其中 $r_\theta(x,\ y)=\beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$，$\pi_{\text{ref}}$表示参考策略模型——通常是经过监督微调的模型，$\sigma(\cdot)$表示 Sigmoid 函数，而 β是一个超参数，用于控制参考模型π_{ref}与当前策略 π_θ之间的散度。

3.2. 课程学习

课程学习（Bengio 等人，2009年）是一种训练策略，逐步让模型接触难度越来越高的样本，其灵感源自人类的学习方式。课程学习可以提升收敛速度和模型泛化能力，其做法是从较简单的样本开始，逐步过渡到较难的样本。这种范式依赖于对样本难度的评估，引导模型随着时间推移，从学习较简单的示例逐渐转向更具挑战性的示例。

成对难度的定义 请注意，用于直接偏好优化的偏好数据集通常由带有大语言模型生成的评分的响应对构成。在实践中，对于每个提示，评分较高的响应会被选为选定响应，而评分较低的响应则被指定为拒绝响应。基于这一构建方式，评分自然可以作为估计样本级难度的依据。我们将这一概念定义如下：

定义 3.1。（成对难度）给定一个表示为 $(x、\ y_w、\ y_l、\ s_w、\ s_l)$ 的偏好对，其中 s_w 和 s_l 分别表示选定响应和拒绝响应的评分，成对难度定义为评分差：$M(s_w、\ s_l):=s_w-s_l$。

在实践中，指标 $M(\cdot)$ 始终为非负数。默认情况下，将使用评分差值（M）来衡量成对难度，其中较高的 M 意味着偏好对更容易。直观而言，数据集中的偏好对在难度上可能存在差异，这体现在选定响应与拒绝响应之间评分差值的变化上。图 2 展示了两个常用偏好数据集（包括 UltraFeedback 和 Arigilla）的分数差距分布，凸显了数据集中困难偏好对的普遍存在。

虽然定义 3.1是我们的主要指标，但也可以考虑其他指标，例如奖励分数边际（Croitoru 等人，2024a）、直接偏好优化损失（Gao 等人，2025）或回答之间的嵌入距离。请注意，我们的目标并非开发一种更优的难度估计指标，而是根据数据集构建过程中使用的现有标注，研究难度的作用。值得注意的是，评分无需计算，而直接偏好优化损失等依赖模型的度量则成本更高。更多讨论见第 2节。对于困难偏好对，我们在附录 Cui 等人，2023年）中提供了从 Ultrafeedback 数据集（C）提取的若干示例。

4. 探究简单／困难偏好对的影响

在本节中，我们首先通过实证验证简单样本对和困难偏好对对齐性能可能产生的影响。根据观察到的结果，我们进一步探讨处理这些困难偏好对的策略。现有方法通常选择过滤掉这类偏好对，以减轻其负面影响（Gao 等人，2025；Deng 等，2025），而我们认为，若能妥善利用，这些困难偏好对仍具有信息价值，并可通过我们提出的方法对对齐性能产生积极贡献。接下来，

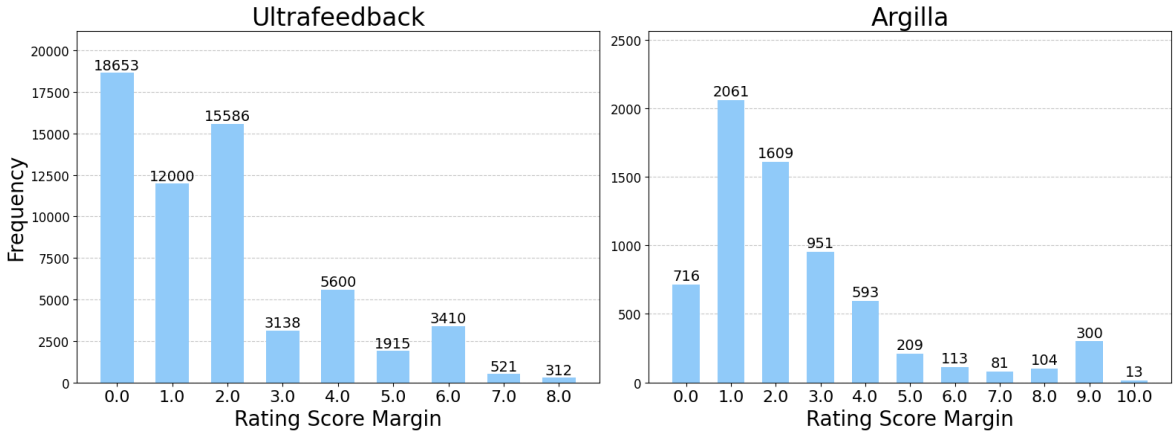


图*2.* 原始评分分差分布。**左侧**：Ultrafeedback 数据集（61k 个样本）。**右侧**：Argilla 数据集（7k 个样本）。已选/已拒绝评分均由多个大语言模型标注。可以看到，两个数据集中约有 50% 的样本分数差异低于 1.0。

我们不直接丢弃这类偏好对，而是引入一种混合目标：对简单样本对应用 DPO 损失，对更具挑战性的偏好对则采用监督微调损失。

4.1. 实证证据：简单样本对或困难偏好对的潜在影响

为研究难度各异的数据对所带来的影响，我们根据大语言模型给出的评分差异分布，将 UltraFeedback 数据集划分为三个子集。针对每个子集，我们随机抽取 5,000 个样本，并采用典型的直接偏好优化损失进行训练。图3 展示了对齐性能如何随难度等级变化。值得注意的是，图3a所示的性能结果表明，较简单的偏好对——即分数差距较大的偏好对——在 AlpacaEval 2 的长度控制胜率和原始胜率上都能取得更高分；而使用分数差异较小的困难偏好对进行训练时，性能则会下降。图3b-3d 进一步提供了更详细的训练过程信息，显示简单样本对能带来更快的收敛速度（更低的训练损失）、更高的奖励准确率和更大的奖励分差。相比之下，困难偏好对会导致优化速度变慢、偏好信号变弱、奖励区分度有限，最终降低对齐性能。

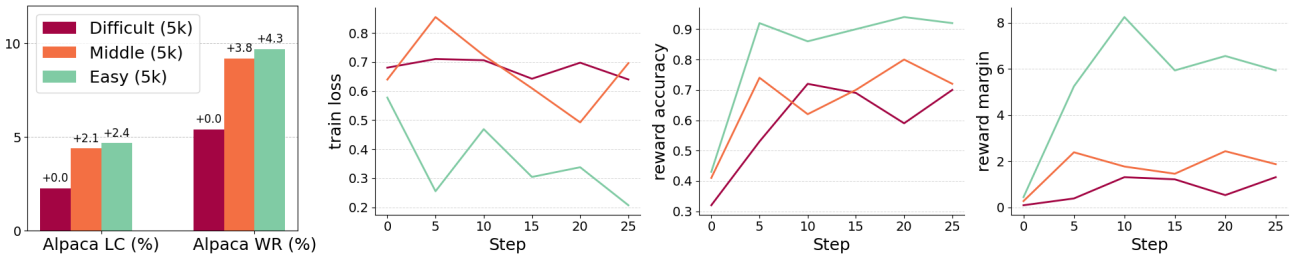
受上述观察启发，我们按难度对偏好对进行排序，并优先处理分数间隔较大（即较容易的偏好对），采用课程学习式的训练策略（Bengio 等人，2009年；Kim &Lee，2024年；Wu et al.，2024年）。这一策略的依据是，较容易的偏好对能提供更清晰、更一致的监督信号，从而有效扩大基础模型与参考模型之间的散度，并进一步提升直接偏好优化的效果。然而，鉴于观察到的对齐性能负面影响（Gao 等人，2025；Deng 等，2025）。

要不要过滤? 以往研究中一种常见做法是完全舍弃困难偏好对（Deng 等，2025；Gao 等人，2025），其依据是这类偏好对会带来歧义或额外噪声，从而妨碍有效的偏好建模。这种担忧源于一个观察：这类难以区分或相似的偏好对往往会引发似然位移现象（Pal 等人，2024年；Yuan 等，2024年；Rafailov 等，2024a；Razin 等人，2024年），即偏好响应和非偏好响应的对数概率都会下降，这与直接偏好优化最初旨在提高 y_w 的概率、降低 y_l 的概率的设计相矛盾。潜在原因是，两个几乎相同的响应之间细微的差异会导致梯度信号不稳定，并使优化方向错位。更多详情见附录B.3。

与其舍弃这些困难偏好对，我们在本文中认为它们仍然包含信息；若处理得当，它们可以提供宝贵的监督信号。例如，近期研究（Xu 等，2024b）表明，在大语言模型的真实性任务中，过滤掉困难偏好对可能会损害直接偏好优化的整体表现，这与我们的观点一致。下一节将深入探讨：若保留这些困难偏好对，应如何处理。

4.2. 目标函数：对简单样本对应用直接偏好优化，对困难样本应用监督微调

受近期研究（Grattafiori 等人，2024年；Pal 等人，2024年）的启发，该研究提出加入类似监督微调的项以缓解似然偏移，我们提出了一种简单而有效的方法，根据样本对难度在直接偏好优化和监督微调损失之间自适应切换。



*(a)*Alpaca LC-WR *(b)*训练损失 *(c)* 奖励准确率 *(d)* 奖励边际

图*3.* 对 UltraFeedback 数据集的三个子集进行微调的模型性能对比，每个子集代表不同的难度级别。使用的基础模型为 LLaMA-3-8B，默认采用直接偏好优化作为损失函数。子图 3a 凸显了简单样本对对对齐性能的积极影响。子图 3b-3d 展示了详细的训练动态，揭示出一个明显趋势：随着数据从简单偏好对转向困难偏好对，性能持续下降。

具体而言，对于分数差距较大的简单样本对，我们应用直接偏好优化损失，因为此时模型偏好更加确定；对于分数差距较小的困难偏好对，则使用监督微调损失，以避免优化不稳定性。这一边际感知目标函数将直接偏好优化的对齐能力与监督微调的稳健性相结合。这些设计选择构成了我们提出的方法——**MixDPO**——的基础，下面将对其进行详细介绍。我们定义一个二元指示变量 z，用于区分简单样本对和困难样本对，并有条件地应用相应的损失，具体如下：

$$z = \begin{cases} 1 & \text{if } M(s_w, s_l) < \tau \\ 0 & \text{otherwise} \end{cases}$$

其中，τ 是一个预定义阈值，用于控制对难度的敏感程度。默认情况下，我们将 Ultrafeedback 数据集的 τ 设为 0.5。根据指标 z，我们构造了一种混合损失：对较容易的样本对（$z = 0$）应用直接偏好优化，对较难的样本对（$z = 1$）应用监督微调：

$$\mathcal{L}_{\textbf{MixDPO}} = -(1-z) \cdot \underbrace{\log p_\theta(y_w \succ y_l \mid x)}_{\text{DPO loss}} - z \cdot \underbrace{\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_{w,t} \mid x, y_{w,<t})}_{\text{SFT loss}}$$

这里，z 充当切换信号，根据每个训练样本的难度，动态地为其分配适当的损失。

请注意，偏好对会根据其估计难度进行排序。实际上，我们采用两阶段训练范式：首先对分数差距较大的较容易偏好对进行标准的直接偏好优化训练，随后进入监督微调阶段，重点处理难度更高的偏好对。

5. 实验

5.1. 实验设置

基础模型与训练设置 遵循 SimPO（Meng 等人，2024年）的设置，我们选用两个广泛使用的模型作为基础模型：LLaMA-3-8B（Grattafiori 等人，2024年）和 Mistral-7B-v0.1（Jiang 等人，2023年）。具体而言，这两个模型在 Ultrachat-200k 数据集上微调后的版本[1] 被用作后续偏好优化的起始点。我们在 Ultrafeedback 数据集（Cui 等人）上进行偏好优化以开展评估。对于 UltraFeedback 数据集，我们将 $\tau = 0.5$设为阈值，从而识别出 7,387 个困难偏好对，其余偏好对则被归类为容易偏好对。在附录B中，我们提供了更多结果，以进一步考察阈值 τ的影响。

基线方法 我们通过与最先进的偏好优化方法进行比较，评估所提出偏好方法的性能，包括直接偏好优化（Rafailov 等人，2024b）及其变体，例如 SLIC-HF（Zhao 等人，2023年）、IPO（Azar 等人，2024年）、KTO（Ethayarajh 等人，2024年）、CPO（Xu 等人，2024年）、SimPO（Meng 等人，2024年）、O-RPO（Hong 等人，2024年）、R-DPO（Park 等人，2024年）以及 SelectiveDPO（Gao 等，2025）。为保持一致，我们使用 SimPO 仓库发布的模型[2] 生成模型响应，然后进行评估。对于 SelectiveDPO，我们遵循其原始设置，使用发布的验证集 DPO 损失作为难度指标。遵循相同的训练设置，默认情况下，所有结果均基于全参数微调（FPFT）。更多超参数设置见附录 A。

评测基准 本文主要选取了三种热门的开放式基准：MT-Bench（Zheng 等，2023年）、AlpacaEval 2（Li 等，2023年）以及

[1]https://huggingface.co/datasets/ HuggingFaceH4/ultrachat_200k[2]https://github.com/princeton-nlp/SimPO

表*1*。 不同基线方法在三个大语言模型评审基准上的性能比较。请注意，$\widetilde{\text{胜率}}$表示调整后的胜率，即胜率加上平局率的一半。我们以**粗体**突出显示最佳结果，并以下划线标出次佳结果。

方法	AlpacaEval 2.0		Arena Hard		MT Bench
	LC 胜率（%）	胜率（%）	胜率（%）	$\widetilde{\text{胜率}}$（%）	平均分（0-10）
基础模型：LLaMA-3-8B					
LLAMA3-8B-BASE-SFT	3.73	10.19	3.9	7.8	4.82
原始直接偏好优化	9.37	16.77	20.4	31.2	5.94
SLIC-HF	5.20	5.71	14.4	22.8	4.99
CPO	4.25	9.69	12.8	24.0	5.64
IPO	5.89	11.55	**20.6**	**32.5**	6.01
KTO	4.27	3.98	17.6	27.4	6.03
O-RPO	5.43	7.08	14.4	24.0	5.80
RDPO	6.92	11.06	18.9	28.6	5.97
SimPO	6.77	14.04	20.2	**32.5**	<u>6.09</u>
SelectiveDPO	<u>8.85</u>	<u>30.43</u>	<u>20.5</u>	<u>32.1</u>	5.82
MixDPO	**14.42**	**36.65**	16.6	26.3	**6.17**
基础模型：Mistral-7B-v0.1					
MISTRAL-7B-BASE-SFT	2.39	1.24	3.0	5.1	4.53
原始直接偏好优化	5.14	4.72	10.0	15.0	5.14
SLIC-HF	4.42	3.60	6.0	10.7	4.43
CPO	4.04	3.85	4.6	10.2	4.6
IPO	5.45	4.60	6.8	13.2	4.73
KTO	5.02	3.23	5.0	10.3	4.56
O-RPO	4.38	3.35	2.8	4.6	2.74
RDPO	<u>6.03</u>	4.60	9.7	16.7	5.29
SIMPO	4.30	<u>5.47</u>	**11.2**	<u>19.5</u>	<u>5.34</u>
SELECTIVEDPO	3.91	<u>5.47</u>	<u>10.2</u>	16.6	4.98
MixDPO	**7.67**	**6.71**	<u>10.2</u>	**20.5**	**5.55**

Arena-Hard-v0.1（Li 等，2024年）。这些基准评估模型在各类查询中的多样化对话能力，已被社区广泛采用。例如，AlpacaEval 2 包含来自 5 个子集（如 SelfInstruct、Vicuna-Bench）的 805 个问题。遵循其评估协议，我们报告 AlpacaEval 2 相对于 GPT-4-Turbo 的长度控制胜率（LC Win Rate）和胜率、Arena-Hard 相对于 GPT-4-0314 的胜率，以及 MT-Bench 基准的离散评分（0-10）。考虑到成本因素，我们采用近期发布的 GPT-4.1（即 GPT-4.1-2025-04-14）作为所有基准的大语言模型评审模型。

5.2. 实证结果

如表 1 所示，我们提出的方法在三个评测基准上、基于两种基础模型（即 LLaMA-3-8B 和 Mistral-7B-v0.1）时，始终优于基线方法。值得注意的是，在 LLaMA-3-8B 设置下，与所有基线方法相比，我们提出的方法在长度控制胜率和原始胜率上均取得了约 6% 的绝对性能提升。在 Mistral-7B-v0.1 设置下，也观察到了类似的性能提升。这些持续的改进彰显了该方法的有效性。在两种基础模型设置下，可以观察到，许多 DPO 变体（如 CPO 和 KTO）未能胜过标准 DPO，有时甚至表现更差；这一结果与 SimPO 中报告的观察一致（Meng 等人，2024年），该研究使用了更强大的 GPT-4 来进行大语言模型评判。

6. 消融研究

在本节中，我们开展全面的消融研究，以评估所提出方法在更多基础模型和偏好数据集上的有效性与泛化能力。我们还深入分析各组件的贡献，评估其与现有直接偏好优化变体的兼容性，并报告在更强的 LLaMA-3-8B-Instruct 基础模型上的性能结果。

6.1. 不同模型和数据集之间的泛化

在此，我们研究所提出的方法 **MixDPO** 在另一种基础模型和偏好数据集上的泛化能力。在基线集合中，我们有选择地选取了两种表现出色的直接偏好优化变体：SimPO 和 Selec- tiveDPO。考虑到它们的超参数可能较为复杂，

表 2. 不同设置下的性能比较。**左侧**：使用 UltraFeedback 数据集评估对 Qwen-2.5-7B基础模型的泛化能力。**右侧**：使用 LLaMA-3-8B 模型，评估对不同偏好数据集 Argilla-7k 的泛化能力。

	模型：Qwen2.5-7B	
方法	LC 胜率（%）	胜率（%）
基础 SFT	0.20	0.99
原始直接偏好优化	2.38	3.60
SimPO	2.28	3.11
SelectiveDPO	3.12	**5.59**
MixDPO	**3.45**	**5.59**

	数据集：Argilla-7k	
方法	LC 胜率（%）	胜率（%）
基础 SFT	3.73	10.19
原始直接偏好优化	2.90	12.17
SimPO	7.07	5.84
SELECTIVEDPO	3.59	5.22
MixDPO	**9.23**	**20.62**

且对超参数（例如学习率）较为敏感，我们尽最大努力对其进行调优，以获得最佳性能。有关超参数配置的更多细节，请参阅附录 B。在此，我们使用 Qwen-2.5-7B 作为基础模型，并使用 Argilla-7K 数据集[3] 来评估我们方法的泛化能力。

MixDPO 在 Qwen-2.5-7B 模型上表现出色 为验证我们提出的方法在另一组模型上的有效性，我们选择 Qwen-2.5-7B（Yang 等人，2024年）作为第三个基础模型。表 2 （左）展示了 AlpacaEval 2 基准上的相应结果。可以看到，我们提出的方法在长度控制胜率和原始胜率方面均优于基线方法。

MixDPO 在 Argilla-7K 数据集上具有良好的泛化能力 我们在另一个数据集 Argilla 上评估了该方法的泛化能力。该数据集包含从多个来源汇总的 7,000 个高质量偏好样本。表 2 （右）中的结果表明，**MixDPO** 始终优于所有基线方法，凸显了其在不同数据集上的稳健性和泛化能力。

6.2. 理解并扩展 MixDPO：组件与变体分析

探索不同组成部分的独立作用 请注意，我们提出的方法包含两个关键组成部分：按难度排序数据，以及应用混合损失函数。为明确凸显各组成部分的贡献，图 4a 展示了基于 LLaMA-3-8B 模型的不同配置的性能。值得注意的是，最后三种设置中的直接偏好优化训练阶段——直接偏好优化+排序后的数据 和我们的方法+舍弃困难样本——完全相同，使用相同的损失和相同的按难度排序的数据。它们的唯一区别在于最终的监督微调阶段： 直接偏好优化+排序后的数据 继续使用直接偏好优化训练困难偏好对，我们的方法+舍弃困难样本 则将其舍弃并停止训练，而我们的方法 则进一步使用 SFT 目标函数对其进行训练。实验结果表明，这两个组成部分都能独立提升对齐性能。重要的是，与舍弃困难偏好对不同，通过监督微调损失利用这些样本，能够显著提升 AlpacaEval 2 LC 胜率，同时不会大幅牺牲原始胜率。

MixDPO 能否应用于其他 DPO 变体？ 我们还希望研究能否将 MixDPO 融入其他 DPO 变体。本文选择 IPO 这一 DPO 变体进行考察。按照我们的框架，我们首先根据分数差距对偏好对进行排序。对于 IPO，我们在简单样本对上训练时，以 IPO 损失替代 DPO 损失；在困难偏好对上则使用监督微调损失。为保持一致，我们按照所提出的方法，选取按成对难度排序后最靠后的 7,387 个困难偏好对，并使用监督微调损失进行训练。图4b 比较了应用我们的方法前后各个变体的性能。值得注意的是，整合我们的方法始终能够提升原有的对齐性能，展现出广泛的适用性和有效性。更多实证结果见附录B。

在更强的指令微调模型上的性能。为进一步展示 MixDPO 在更强指令微调模型上的有效性，我们采用 SimPO（Meng 等人，2024年）和 SelectiveDPO（Gao 等人，2025）的实验设置，并以 LLaMA-3-8B-Instruct 为代表案例。为保持一致，训练使用 SimPO 发布的数据集。由于原始数据集仅提供奖励模型分数，没有明确的评分，我们使用 GPT-4o-mini 生成评分作为替代。具体而言，LLaMA-3-Instruct-8B-DPO 和 LLaMA-3-Instruct-8B-SimPO 模型均可从 SimPO 仓库公开获取。此外，我们还展示了基于不同难度指标的两个版本的 MixDPO，分别采用 GPT-4o-mini 评分和原始奖励分数。表3 （上）显示，MixDPO 的表现始终优于这些基线方法。

探究难度阈值的影响 τ。为了了解阈值的敏感性，我们探究了若干

[3] https://huggingface.co/datasets/argilla/ dpo-mix-7k



(a) MixDPO 中各组件的贡献。 *(b)* 将 MixDPO 集成到 IPO 中。

图4。基于 LLaMA-3-8B 模型的消融实验结果。左：MixDPO 中各组件的贡献。右：将 MixDPO 集成到现有方法中的影响 f 将 MixDPO 集成到现有直接偏好优化变体 IPO。

备选取值 $\tau \in \{0.5, 1, 1.5\}$，它们分别对应于选择约 6,000（数据集的 10%）、18,653（25%）和 30,653（50%）个困难偏好对。默认情况下，我们将阈值设为 $\tau = 0.5$，对应于分数差异的最小边际值。表 3（底部）表明，随着阈值 τ取值增大，性能会下降，这凸显了困难阈值的重要性。从经验来看，选择最困难的前 10% 偏好对（即 $\tau = 0.5$）效果最好，这表明适度接触困难偏好对，能在信息量与过拟合之间取得较好的平衡。

表3。 **上半部分**：更强的 LLaMA3-8B-Instruct 基础模型下的基线方法性能比较。我们遵循 SimPO 的训练设置和数据（Meng 等人，2024年）；由于其中仅提供奖励分数，因此我们随后使用 GPT-4o-mini 生成评分，以保持一致性。 **下半部分**：不同难度阈值的影响 τ。在此，我们基于最小分数差距探究三种阈值设置，即 $\tau = \{0.5, 1, 1.5\}$。

方法	模型：LLaMA3-8B-Instruct 长度控制胜率（%）	胜率（%）
DPO	19.53	20.25
SimPO	24.68	22.73
SelectiveDPO	3.25	1.12
MixDPO （原始奖励）	29.02	**28.01**
MixDPO （GPT-4o-mini）	**29.47**	25.59

方法	AlpacaEval 2.0 LC 胜率（%）	胜率（%）
Base-SFT	3.73	10.19
原始直接偏好优化	9.37	16.77
MixDPO ($\tau = 1.5$)	3.80	2.61
MixDPO ($\tau = 1$)	6.48	25.47
MixDPO ($\tau = 0.5$)	14.42	36.65

7. 结论与局限

在这项工作中，我们希望探索一种替代方法，处理那些边际较小、可能因潜在的似然偏移而阻碍对齐的困难偏好对。不同于常见做法中将它们过滤掉，我们提出了一种简单而有效的方法 **MixDPO**，根据偏好对的难度，在 DPO 损失和标准监督微调损失之间自适应切换，利用清晰和模糊的信号来提升对齐性能。具体而言，我们对置信度高的简单样本对使用 DPO 损失，对困难偏好对使用监督微调损失。大量实证实验验证了我们所提方法的有效性。

与缓解似然偏移的方法形成对比的是，这些方法要么 i）针对所有样本提出一种通用偏好损失（Xiao 等人，2024年；Pang 等人，2024c），要么 ii）对困难样本评分并将其过滤掉（Razin 等人，2024年），我们的工作倡导一种互补的视角：通过优化监督微调目标函数来重新利用困难偏好对。这凸显了让训练信号与数据难度相匹配的重要性，而不是强制采用一刀切的目标函数，并为以有原则的方式利用困难数据来提升对齐开辟了有希望的方向。

尽管如此，我们也承认本方法存在若干局限。虽然所提出的方法优于所有基线方法，但除了使用监督微调损失之外，可能还存在更有效的困难偏好对处理策略，这为未来研究提供了一个有前景的方向。此外，尽管设计更好的难度指标并非我们的首要目标，但这同样是一个值得探索的方向。与此同时，我们的难度指标主要依赖原始大语言模型评分，而这些评分可能存在噪声或不准确，正如先前研究所指出的（Pang 等人，2024b）。此类潜在的评分误差可能会扭曲对偏好难度的估计，导致训练动态不理想，并最终妨碍大语言模型的对齐性能。然而，这一

实际问题可以通过现有的评分整理技术得到显著缓解（Zhu 等人，2023年；Pang 等人，2024b），而开发更先进的方法仍是未来研究的重要方向。

影响声明

本文通过提出一种难度感知训练策略来推进大语言模型偏好对齐，从而更好地利用小边际（模糊）偏好对。该方法旨在提高对齐训练的稳定性与有效性，并未引入超出标准偏好优化和监督微调之外的新能力。潜在风险主要来自底层偏好数据以及实践中使用的任何自动评分信号（例如偏差或系统性错误），此外，改进后的对齐技术也可能被用于有害目标。基于我们的评估，除这些众所周知的考量之外，我们预计该方法不会带来额外的、特有的社会层面担忧。

参考资料

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S.等。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*，2023年。

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., 和 Lerner, S. 通过示例教学：对类别习得过程的启示。《实验心理学季刊》*A*辑，50(3)：586–606，1997年。

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., 和 Calandriello, D. 一个理解从人类偏好中学习的一般理论范式。载于国际人工智能与统计学会议，第 4447–4455 页。PMLR，2024年。

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. 等。通过人类反馈强化学习训练有帮助且无害的助手。*arXiv* 预印本 *arXiv:2204.05862*，2022年。

Bengio, Y., Louradour, J., Collobert, R.与Weston, J. 课程学习。载于第*26*届国际机器学习年会论文集，第41–48页，2009年。

Chen, H., Wang, J., Shah, A., Tao, R., Wei, H., Xie, X., Sugiyama, M., 和 Raj, B. 理解并缓解下游任务预训练中的标签噪声。*arXiv* 预印本 *arXiv:2309.17002*，2023a。

Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., Tang, Z., Srinivasan, V., Zhou, T., Huang, H. 等。Alpagasus：用更少的数据训练更好的羊驼。*arXiv* 预印本 *arXiv:2307.08701*，2023b。Chowdhury, S. R., Kini, A., 和 Natarajan, N. 可证明稳健的直接偏好优化：利用噪声反馈对齐语言模型。*arXiv* 预印本 *arXiv:2403.00409*，2024。Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., 和 Tafjord, O. 你以为自己已经解决了问答问题？试试 ARC，也就是 AI2 推理挑战赛。*arXiv* 预印本 *arXiv:1803.05457*，2018。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C.与Schulman, J. 训练验证器以求解数学应用题。*arXiv*预印本*arXiv:2110.14168*，2021年。

Croitoru, F.-A., Hondru, V., Ionescu, R.T., Sebe, N.与Shah, M. 面向扩散模型与一致性模型的课程式直接偏好优化。*arXiv*预印本*arXiv:2405.13637*，2024a。

Croitoru, F.-A., Hondru, V., Ionescu, R.T.与Shah, M. 逆向稳定扩散：生成该图像时使用了什么提示？《计算机视觉与图像理解》，249:104210，2024b。

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z. 和 Sun, M. UltraFeedback：利用高质量反馈提升语言模型，2024年。载于*URLhttps://openreview. net/forum*。

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z. 和 Sun, M. UltraFeedback：利用高质量反馈提升语言模型。2023年。

Deng, X., Zhong, H., Ai, R., Feng, F., Wang, Z. 和 He, X. LESS：少即是多：通过偏好数据选择改进大语言模型对齐。*arXiv* 预印本 *arXiv:2502.14560*，2025。

Ethayarajh, K.、Choi, Y. 和 Swayamdipta, S. 理解数据集难度与 V-可用信息（2021年）。*URL https://arxiv. org/abs/2110.08420*。

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D. 和 Kiela, D. KTO：将模型对齐视为前景理论优化。*arXiv* 预印本 *arXiv:2402.01306*，2024年。

Gao, C., Li, H., Liu, L., Xie, Z., Zhao, P. 和 Xu, Z. 面向对齐的原则性数据选择：困难样本的潜在风险。*arXiv* 预印本 *arXiv:2502.09650*，2025。

Grattafiori, A.、Dubey, A.、Jauhri, A.、Pandey, A.、Kadian, A.、Al-Dahle, A.、Letman, A.、Mathur, A.、Schelten, A.、

Vaughan, A. 等。《Llama 3 模型家族》。*arXiv* 预印本 *arXiv:2407.21783*，2024年。

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., 和 Steinhardt, J. 测量大规模多任务语言理解能力。*arXiv* 预印本 *arXiv:2009.03300*，2020年。

Hong, J., Lee, N., 和 Thorne, J. ORPO：无需参考模型的单体偏好优化。*arXiv* 预印本 *arXiv:2403.07691*，2024年。

Huang, K., Wu, J., Chen, Z., Wang, X., Gao, J., Ding, B., Wu, J., He, X., 和 Wang, X. 更大的还是更小的奖励边际，才能为对齐选择偏好？*arXiv* 预印本 *arXiv:2503.01864*，2025。

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., 等。Mistral 7B。*arXiv*预印本 *arXiv:2310.06825*，2023年。Jiang, L., Meng, D., Mitamura, T., 和 Hauptmann, A. G. 先处理简单样本：用于零样本多媒体搜索的自步重排序。载于第*22*届*ACM*国际多媒体会议论文集，第547–556页，2014年。Kim, J. 和 Lee, J. 策略性数据排序：通过课程学习提升大语言模型的性能。*arXiv*预印本 *arXiv:2405.07490*，2024年。Kong, K., Xu, X., Wang, D., Zhang, J., 和 Kankanhalli, M. 面向噪声偏好的稳健对齐的困惑度感知校正。载于第三十八届神经信息处理系统年会。Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Zhu, B., Gonzalez, J. E., 和 Stoica, I. 从实时数据到高质量基准：Arena-Hard 流程，2024年。Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., 和 Hashimoto, T. B. AlpacaEval：指令遵循模型的自动评估器，2023年。Lin, S., Hilton, J., 和 Evans, O. TruthfulQA：衡量模型如何模仿人类谬误。*arXiv*预印本 *arXiv:2109.07958*，2021年。Lin, Z., Gou, Z., Gong, Y., Liu, X., Shen, Y., Xu, R., Lin, C., Yang, Y., Jiao, J., Duan, N., 等。Rho-1：并非所有词元都是你需要的。*arXiv*预印本 *arXiv:2404.07965*，2024年。Liu, M., Di, Z., Wei, J., Wang, Z., Zhang, H., Xiao, R., Wang, H., Pang, J., Chen, H., Shah, A., 等。自动数据集构建（ADC）：样本收集、数据整理及其他。*arXiv*预印本 *arXiv:2408.11338*，2024年。

Liu, W., Zeng, W., He, K., Jiang, Y., 和 He, J. 什么样的数据有利于对齐？关于指令微调中自动数据选择的全面研究。*arXiv*预印本 *arXiv:2312.15685*，2023年。

Liu, Y. 和 Guo, H. Peer 损失函数：在不知道噪声率的情况下从噪声标签中学习。载于国际机器学习大会，第 6226–6236 页。PMLR，2020年。

Meng, Y.、Xia, M. 和 Chen, D. Simpo：采用无参考奖励的简单偏好优化。*arXiv* 预印本 *arXiv:2405.14734*，2024年。

Mitchell, E. 关于噪声偏好下直接偏好优化及其与 IPO 关系的说明，2023年。网址 *https://ericmitchell. ai/cdpo. pdf*。Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W. 等。WebGPT：结合人类反馈的浏览器辅助问答。*arXiv* 预印本 *arXiv:2112.09332*，2021年。Natarajan, N., Dhillon, I. S., Ravikumar, P. K. 和 Tewari, A. 噪声标签学习。神经信息处理系统进展，26，2013年。Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. 等。训练语言模型以遵循人类反馈中的指令。神经信息处理系统进展，35:27730–27744，2022年。Pal, A., Karkhanis, D., Dooley, S., Roberts, M., Naidu, S. 和 White, C. Smaug：使用 DPO-positive 修复偏好优化的失效模式。*arXiv* 预印本 *arXiv:2402.13228*，2024年。Pang, J., Wang, J., Zhu, Z., Yao, Y., Qian, C. 和 Liu, Y. 无害的公平性：一种影响力引导的主动采样方法。神经信息处理系统进展，37:61513–61548，2024年。Pang, J., Wei, J., Shah, A. P., Zhu, Z., Wang, Y., Qian, C., Liu, Y., Bao, Y. 和 Wei, W. 通过策划大语言模型驱动的评分系统提升数据效率。*arXiv* 预印本 *arXiv:2410.10877*，2024b。Pang, J., Di, N., Zhu, Z., Wei, J., Cheng, H., Qian, C. 和 Liu, Y. 词元清理：用于大语言模型监督微调的细粒度数据选择。*arXiv* 预印本 *arXiv:2502.01968*，2025。Pang, R. Y., Yuan, W., He, H., Cho, K., Sukhbaatar, S. 和 Weston, J. 迭代推理偏好优化。神经信息处理系统进展，37: 116617–116637，2024c。

Park, R., Rafailov, R., Ermon, S. 和 Finn, C. 在直接偏好优化中将长度与质量解耦。*arXiv* 预印本 *arXiv:2403.19159*，2024年。

Platanios, E. A.、Stretcu, O.、Neubig, G.、Poczos, B. 和 Mitchell, T. M. 面向神经机器翻译的基于能力的课程学习。*arXiv* 预印本 *arXiv:1903.09848*，2019年。

Rafailov, R.、Hejna, J.、Park, R. 和 Finn, C. 从 r 到q^*: 你的语言模型实际上是一个Q函数。*arXiv* 预印本 *arXiv:2404.12358*，2024a。

Rafailov, R.、Sharma, A.、Mitchell, E.、Manning, C.D.、Er- mon, S. 和 Finn, C. 直接偏好优化：你的语言模型实际上是一个奖励模型。神经信息处理系统进展，36，2024b。

Razin, N.、Malladi, S.、Bhaskar, A.、Chen, D.、Arora, S. 和 Hanin, B. 无意对齐失效：直接偏好优化中的似然偏移。*arXiv* 预印本 *arXiv:2410.08847*，2024年。

Sakaguchi, K., Bras, R. L., Bhagavatula, C. 和 Choi, Y. Winogrande：大规模对抗性 Winograd 模式挑战。*ACM*通讯，64(9):99–106，2021。

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D. 和 Christiano, P. F. 利用人类反馈学习摘要。神经信息处理系统进展，33:3008–3021，2020。

Tudor Ionescu, R., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D. P. 和 Ferrari, V. 这能有多难？估计图像中视觉搜索的难度。载于*IEEE* 计算机视觉与模式识别会议论文集，第 2157–2166 页，2016。

Wu, B., Meng, F., 和 Chen, L. 基于质量驱动数据选择的课程学习。*arXiv*预印本 *arXiv:2407.00102*，2024年。

Xia, M., Malladi, S., Gururangan, S., Arora, S., 和 Chen, D. LESS：为定向指令微调选择有影响力的数据。*arXiv* 预印本 *arXiv:2402.04333*，2024年。

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., 和 Sugiyama, M. 部分依赖标签噪声：迈向实例依赖标签噪声。神经信息处理系统进展，33：7597–7610，2020年。

Xiao, T.、Yuan, Y.、Zhu, H.、Li, M. 和 Honavar, V. G. Cal-DPO：用于语言模型对齐的校准直接偏好优化。神经信息处理系统进展，37：114289–114320，2024年。

Xu, H.、Sharaf, A.、Chen, Y.、Tan, W.、Shen, L.、Van Durme, B.、Murray, K. 和 Kim, Y. J. 对比偏好优化：推动机器翻译中的大语言模型性能边界。*arXiv* 预印本 *arXiv:2401.08417*，2024a。

Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., Wang, G., Yu, C., 和 Wu, Y. 对于大语言模型对齐，直接偏好优化是否优于近端策略优化？一项全面研究。*arXiv* 预印本 *arXiv:2404.10719*，2024b。

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H. 等。Qwen2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*，2024年。

Yang, Z., Li, L., Di, N., Pang, J., Zhou, Y., Cheng, H., Han, B. 和 Wei, J. ENTP：通过神经符号文本清洗混合增强低质量监督微调数据。*arXiv* 预印本 *arXiv:2510.23160*，2025。

Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y. 等。通过偏好树提升大语言模型推理通才的推理能力。*arXiv* 预印本 *arXiv:2404.02078*，2024年。

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., 和 Choi, Y. Hellaswag：机器真的能续写你的句子吗？*arXiv* 预印本 *arXiv:1905.07830*，2019年。

Zhang, Y., Pang, J., Zhu, Z., 和 Liu, Y. 评估大语言模型污染的众包数据，无需真实标注。*arXiv* 预印本 *arXiv:2506.06991*，2025。

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., 和 Liu, P. J. Slic-hf：结合人类反馈的序列似然校准。*arXiv* 预印本 *arXiv:2305.10425*，2023年。Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. 等。通过 MT-Bench 和聊天机器人竞技场评判大语言模型评判者。神经信息处理系统进展，36：46595–46623，2023年。

Zhu, Z., Wang, J., Cheng, H., 和 Liu, Y. 揭开并提升数据可信度：一项使用无害语言模型训练数据集的研究。*arXiv* 预印本 *arXiv:2311.11202*，2023年。

附录

附录的组织结构

- **第A节**：介绍本研究的实验设置，包括计算环境、所使用的监督微调基础模型及其相应的超参数。
  - 第A.1节：计算环境描述；
  - 第A.2节：所使用的监督微调基础模型；
  - 第A.3节：偏好对齐的训练细节。
- **第B节**：展示其他实验结果：
  - 第B.1节：学习率的影响；
  - 第B.2节：探究替代难度指标的影响；
  - 第B.3节：与替代似然位移方法进行比较；
  - 第B.4节：偏好对中绝对质量的影响；
  - 第B.5节：探索两种潜在的SFT式基线：1）对正面（已选）响应进行完整的SFT训练，2）对正面（已选）和负面（已拒绝）响应同时进行完整的SFT训练。
  - 第 B.6：MixDPO 在分数噪声下的稳健性；
  - 第 B.7：下游任务评估。

第 C：提供了根据原始评分识别出的简单和困难偏好对示例。

A. 实验细节

A.1. 计算环境

在本研究中，所有实验均在一台配备 8×NVIDIA L40S GPU（每张显存为 45 GB）的服务器上进行。对于全参数微调，在 7B/8B 模型上复现每次训练约需 3 GPU小时。

A.2. SFT基础模型

在这项工作中，我们使用公开可用的 SFT模型进行偏好优化实验，这些模型遵循标准的后训练流程。两个模型均使用 Ultrachat 200k数据集进行了微调。SFT训练过程中使用的公开超参数汇总于表4。

表 *4.* 本工作所用 SFT模型的训练细节。

SFT模型	Hugginface来源	批量大小	学习率	优化器	LoRA?
LLaMA-3-8B-SFT	princeton-nlp/Llama-3-Base-8B-SFT	128	2e-5	Adam	No
Mistral-7B-SFT	HuggingFaceH4/mistral-7b-sft-beta	128	2e-5	Adam	No
Qwen-2.5-7B-SFT	AmberYifan/Qwen2.5-7B-sft-ultrachat	128	1e-5	Adam	No

A.3. 偏好对齐训练细节

表5 总结了我们实验中 **MixDPO** 使用的几个关键超参数。默认情况下，我们使用余弦学习率调度器。按照 SimPO（Meng 等人，2024年）的设置，我们将最大提示长度设为 512，最大序列长度设为 1024。此外，所有模型均使用 BF16 精度进行微调。对于基线方法，我们采用 SimPO（Meng 等人，2024年）发布的模型，并提供了相应的超参数。

额外的基础模型 Qwen-2.5-7B 对于额外的基础模型 Qwen-2.5-7B，我们评估了几种学习率设置，相应结果见表 2 （左）。Qwen-2.5-7B 的学习率采用自 SelectiveDPO（Gao 等人，2025）。

额外的偏好数据集 Argilla-7k 针对 Argilla-7k 数据集的一系列实验中，我们考察了几种学习率设置，其结果见表 2 （右侧）。

表5. 实验中使用的**MixDPO** 关键超参数。

SFT模型	学习率	批量大小	β	训练轮次	优化器	LoRA?
LLaMA-3-8B-SFT	1e-6	128	0.01	1	AdamW_Torch	No
Mistral-7B-SFT	2e-7	128	0.01	1	AdamW_Torch	No
Qwen-2.7-7B	8e-7	128	0.01	1	AdamW_Torch	No

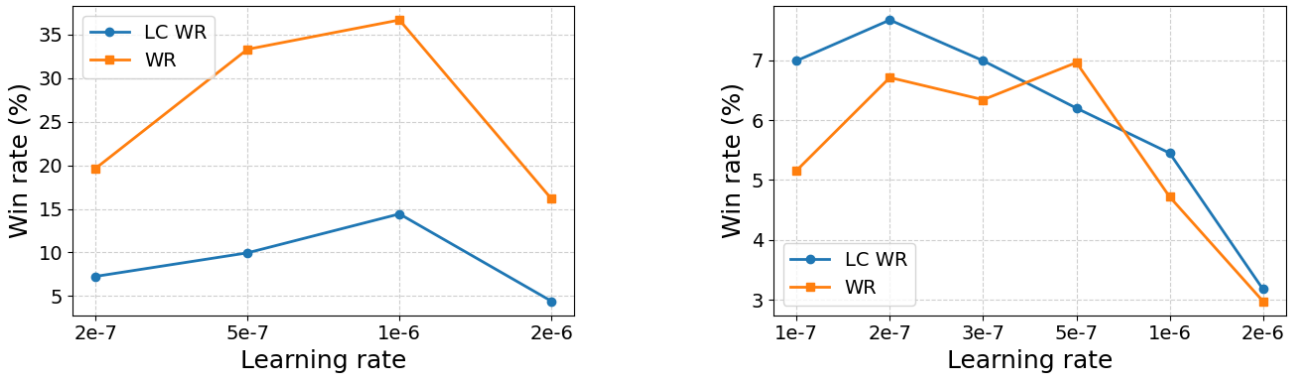
表6。 Qwen-2.5-7B基础模型实验所用的关键超参数。

基线	学习率	批量大小	β	训练轮次	优化器	LoRA?
DPO	8e-7	128	0.01	1	AdamW_Torch	No
SimPO	8e-7	128	2.0	1	AdamW_Torch	No
SelectiveDPO	8e-7	128	0.01	1	AdamW_Torch	No
MixDPO	8e-7	128	0.01	1	AdamW_Torch	No

B. 更多实验结果

B.1. 学习率的影响

图 **MixDPO**5展示了我们提出的方法在两种基础模型上采用不同学习率时的表现。对于 LLaMA-3-8B，学习率为 1e-6 时表现最佳；对于 Mistral-7B，最佳学习率为 2e-7。



(*a*) LLaMA-3-8B (*b*) Mistral-7B

图 5. Ultrafeedback 数据集上两种基础模型配置的学习率影响。

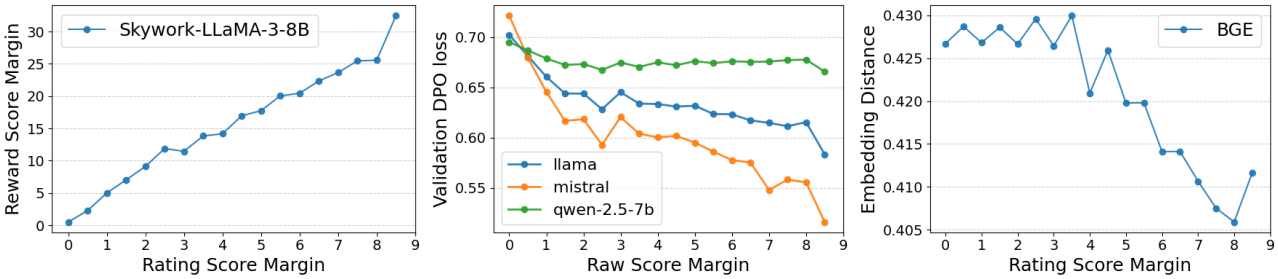
B.2. 探索替代难度指标的影响

评分差值与其他难度指标之间的相关性在此，我们首先系统地研究评分差值与其他替代性难度指标之间的关系，包括验证集 DPO 损失（Gao 等人，2025）、奖励分数（Croitoru 等人，2024a）以及回答之间的嵌入距离。对于奖励分数，我们使用了 RewardBench 排行榜上的强大奖励模型 Skywork-Reward-Llama-3.1-8B[4]。为计算嵌入距离，我们选择了最新发布的开源模型 BGE[5] 作为嵌入模型。请注意，验证集 DPO 损失取自 SelectiveDPO，基于三种不同的基础模型计算得出，分别是 Qwen-2.5-7B、Mistral-7B 和 LLaMA-3-8B。图6显示，原始分数差距与奖励分数边际和验证集 DPO 损失均呈明显正相关。验证集 DPO 损失较小，说明模型更容易从该样本中学习。相比之下，评分差值与嵌入距离之间呈现出反直觉的负相关，这可能归因于嵌入模型的表征局限性。值得注意的是，原始大语言模型评分被证明

[4]https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B-v0.2[5] BAAI/bge-large-en-v1.5

表7. 在 Argilla-7k偏好数据集上进行实验所用的关键超参数。我们用**粗体**突出显示结果最佳的学习率。

基线	学习率范围	批量大小	β	训练轮次	优化器	LoRA?
DPO	{1e-7 5e-7 8e-7 **1e-6** 2e-6}	128	0.01	2	AdamW_Torch	No
SimPO	{1e-7 5e-7 8e-7 **1e-6** 2e-6}	128	2.0	2	AdamW_Torch	No
SelectiveDPO	{1e-7 5e-7 8e-7 **1e-6** 2e-6}	128	0.01	2	AdamW_Torch	No
MixDPO	{1e-7 5e-7 8e-7 **1e-6** 2e-6}	128	0.01	2	AdamW_Torch	No



(a) 奖励分数边际　　*(b)* 验证集 DPO 损失　　*(c)* 嵌入距离

图6。 评分差值与三种替代难度指标之间的关系。可以观察到，评分差值与奖励分数和 DPO 损失呈正相关，但与嵌入距离呈现出反直觉的负相关。

既足以评估样本级难度，也具有竞争力；与奖励模型推理或验证集 DPO 损失计算等成本更高的方法相比，它是一种高效得多的替代方案。

为了评估 MixDPO 在不同难度定义下的泛化能力，我们研究了偏好响应与非偏好响应之间的嵌入距离，以及奖励分数边际。表 8 报告了各项指标下的对齐性能，突显了原始大语言模型评分的有效性。值得注意的是，大语言模型评分无需额外计算，因此与依赖奖励模型推理或基于嵌入的相似度的替代方案相比，效率更高。此处，奖励分数边际指标的表现较差，主要是由于所用奖励模型的局限性。奖励分数边际和嵌入距离所使用的学习率均为 1e-7。

表8。不同难度指标下对齐性能的比较。

方法	AlpacaEval 2.0	
	LC 胜率（%）	胜率（%）
Base-SFT	3.73	10.19
原始直接偏好优化	9.37	16.77
SimPO	6.77	14.04
MixDPO（嵌入距离）	8.75	19.13
MixDPO（奖励分数差距）	7.05	27.95
MixDPO（原始评分分数）	14.42	36.65

B.3. 与替代似然位移方法的比较

在这里，我们探讨能否利用现有的似然偏移解决方法来处理难以应对的偏好对（Razin 等人，2024年；Xiao 等人，2024年；Pang 等人，2024c）。这些方法中有很大一部分侧重于调整损失函数。例如，Smaug（Pal 等人，2024年）提出了 DPOP 损失函数，在标准的直接偏好优化损失中加入了一个额外项。为清晰起见并确保内容完整，我们将 DPOP 损失函数重现如下。

$$\mathcal{L}_{\text{DPOP}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\Big(\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)} \underbrace{-\lambda\cdot\max\left(0,\log\frac{\pi_{\text{ref}}(y_w\mid x)}{\pi_\theta(y_w\mid x)}\right)}_{\text{Additional term}}\Big)\right)\right] \tag{3}$$

对于 CHES（Razin 等人，2024年），它使用 CHES 分数筛除样本；我们在 UltraFeedback 数据集上计算 CHES 分数，并据此筛除指定比例的样本。对于 Cal-DPO（Xiao 等人，2024年），我们在训练期间直接应用其提出的损失函数，并评估由此得到的性能。此外，我们还实现了另一种近期提出的方法，该方法通过在直接偏好优化损失中加入负对数似然（NLL）损失项来解决似然位移问题（Pang 等人，2024c）。我们将这一变体称为 **直接偏好优化+负对数似然**。

对于 CHES，与 SelectiveDPO（Gao 等人，2025）和 MixDPO（一种特例）类似，我们试验了两种数据样本选择比例：50% 和 90%。如表 9 所示，MixDPO 持续优于旨在缓解非预期似然问题的先前方法，证明了我们方法的有效性。

表 *9.* 不同方法在缓解似然偏移方面的对齐性能比较。偏好数据集为 UltraFeedback，所用基础模型为 LLaMA-3-8B。

	AlpacaEval 2.0	
方法	**LC 胜率（%）**	**胜率（%）**
Base-SFT	3.73	10.19
原始直接偏好优化	9.37	16.77
Cal-DPO（Xiao 等人，2024年）	4.56	7.45
直接偏好优化 +NLL（Pang 等人，2024c年）	4.25	8.45
DPOP（Pal 等人，2024年）	4.53	4.04
CHES（选中比例：50%）（Razin 等人，2024年）	6.16	11.93
CHES（选中比例：90%）	8.13	13.91
MixDPO	14.42	36.65

计算开销对比 与 SelectiveDPO 一样，我们的方法采用课程学习框架，但主要挑战了过滤掉困难样本或导致似然偏移的样本这一惯常做法。MixDPO 则采取了截然不同的方法，在监督微调阶段利用这些困难样本，进一步提升对齐性能。实证结果验证了这一策略的有效性。此外，MixDPO 无需额外计算开销来计算难度指标。为突出这一优势，我们在表 10 中报告了使用包含 60,000 个样本的 UltraFeedback 数据集时的详细开销统计。具体而言，在 8×台 H100 GPU 上，计算 DPO 损失（SelectiveDPO）和 CHES 分数分别约需 20 分钟和 50 分钟。相比之下，在 8×台 H100 GPU 上训练完整数据集约需 40 分钟。这表明，基于难度的过滤会带来不可忽视的开销，而我们的方法成功避免了这一点，这是 MixDPO 的一项关键优势。

表 *10.* 在 Ultrafeedback 数据集上引入的计算开销（时间）对比。

指标	**引入的计算开销**
CHES 分数（Razin 等人，2024年）	50 分钟
DPO 损失（SelectiveDPO，（Gao 等人，2025））	20 分钟
MixDPO（评分）	0 分钟

B.4. 绝对质量在偏好对中重要吗?

在此，我们关注一个问题：绝对质量在偏好对中是否重要？我们在两种受控条件下展示了详细的实证结果：两者都好（高质量）和两者都差（低质量）。具体而言，我们选取了

表 *11.* 绝对偏好质量（高质量与低质量偏好对）对两个不同基础模型在 AlpacaEval 上表现的影响。每种设置使用 2,000 个偏好对。采用的损失函数为直接偏好优化。

	AlpacaEval 2.0	
方法	**LC 胜率（%）**	**胜率（%）**
LLaMA-3-8B-Base（高质量，2000个样本）	2.68	6.34
LLaMA-3-8B-Base（低质量，2000个样本）	2.88	7.58
Mistral-7B-Base（高质量，2000个样本）	4.20	2.42
Mistral-7B-Base（低质量，2000个样本）	3.27	1.93

2,000 个偏好对，每种情况各一组，其中已选和已拒绝的回答获得相同分数，从而隔离绝对质量的影响。两个子集样本均选自 Ultrafeedback 数据集。相应结果见表 11。结果表明，尽管绝对质量对最终表现确有轻微的积极影响，但与分数差异（即相对偏好）的影响相比，这种作用明显较弱。

B.5. 两种SFT训练设置的比较

我们进一步使用UltraFeedback数据评估了两种潜在的SFT式基线：（1）仅在正向（已选）响应上训练；（2）同时在正向和负向响应上训练。SFT训练遵循一个流行的Alignment-Handbook 代码库[6]。如表12所示，两种设置均无效。仅对正向响应进行纯SFT无法捕捉对齐所需的偏好结构，从而导致性能大幅下降。同时纳入正向和负向响应也并无帮助，并产生同样较差的结果。这很可能是因为UltraFeedback主要是为偏好学习而非监督式指令微调而构建的，使其被选响应并不适合完整的SFT训练。

表 *12*。 AlpacaEval 上两种监督微调风格基线变体的性能比较。

方法	**LLaMA-3-8B**		**Mistral-7B**	
	LC 胜率（%）	**胜率（%）**	**LC 胜率（%）**	**胜率（%）**
DPO	9.37	16.77	5.14	4.72
SimPO	6.77	14.04	4.30	5.47
全量监督微调（仅正样本）	0.27	1.74	2.86	2.61
全量监督微调（正负样本）	0.19	1.24	2.18	1.86
MixDPO	14.42	36.65	7.67	6.71

B.6. 分析 MixDPO 对分数噪声的稳健性

请注意，MixDPO 使用原始评分作为难度指标。使用原始评分的原因有两点：1）遵循常见的偏好对数据集构建流程；2）与其他难度指标相比，不会引入额外的计算成本。不过，一个实际问题是，大语言模型生成的评分可能存在噪声或偏差，导致样本对难度分类错误。

这个分数噪声的潜在问题可以重新表述为一个典型的多类噪声标签问题（Natarajan 等，2013；Xia 等，2020；Chen 等，2023a；Pang 等人，2024b；Zhu 等人，2023；Liu 和 Guo，2020）。例如，可以通过预处理原始评分来有效解决这一问题，所用技术包括近期提出的、专为监督微调样本设计的大语言模型生成分数整理流程 DS2（Pang 等人，2024b）。为说明 MixDPO 的稳健性，我们展示一个特殊案例：不再仅按分数差距对数据排序，而是对数据集排序后，将最后 10% 的简单样本对与最困难的样本对互换。具体而言，我们用第 90–100 百分位（困难）的样本对，替换第 80–90 百分位（简单）的样本对。这一调整的动机是，最后 10% 的简单样本对往往分数差距较小，因此更可能被错误标注。值得注意的是，如表 13所示，即使在 10% 样本被错误标注的情况下，MixDPO 的表现仍与基线方法相当。

[6]https://github.com/huggingface/alignment-handbook

表*13.* 性能比较：在 LLaMA-3-8B 基础模型上，使用 UltraFeedback 数据集进行噪声分数扰动时的表现。

模型	LC 胜率（%）	胜率（%）
DPO	9.37	16.77
SimPO	6.77	14.04
SelectiveDPO	8.85	30.43
带10%交换的MixDPO	8.72	30.31
MixDPO	14.42	36.65

B.7. 下游任务评估

为研究所提出的偏好优化算法对下游任务表现的影响，我们开展了实验，并与若干广泛采用的直接偏好优化变体进行比较。具体而言，我们在多个常用的 OpenLLM排行榜任务上进行评估，包括 MMLU（Hendrycks 等，2020）、TruthfulQA（Lin 等，2021）、HellaSwag（Zellers 等，2019）、ARC-Challenge（Clark 等，2018）、GSM8K（Cobbe 等，2021）和 WinoGrande（Sakaguchi 等，2021）。这些数据集具有足够的多样性，可从多个方面全面评估微调模型，包括事实准确性、推理能力和多语言能力。任务表现通过 lm-eval-hareness[7] 代码库进行评估。结果见表 14。如表所示，与其他基线方法相比，我们提出的 **MixDPO**在各项下游任务上均取得了有竞争力的表现。

表*14*。 下游任务评估结果。所使用的偏好数据集为UltraFeedback。括号中的数字表示思维链样本数。

基线	MMLU(5)	TruthfulQA(0)	HellaSwag(10)	ARC-C(25)	GSM8K(5)	Winogrande(5)	平均值
基础模型：LLaMA-3-8B							
SFT	63.78	45.24	61.30	56.16	47.50	76.18	58.40
DPO	63.37	53.46	64.78	61.67	52.50	77.05	62.10
CPO	63.77	54.32	61.67	57.54	54.50	76.97	61.50
KTO	63.36	55.66	64.14	60.72	55.50	76.25	62.60
SimPO	63.11	59.39	62.30	62.27	51.50	77.21	62.60
SelectiveDPO	63.95	53.94	64.76	61.50	52.50	76.10	62.10
MixDPO	63.20	55.49	64.78	61.58	54.00	77.45	62.80
基础模型：Mistral-7B-v0.1							
SFT	59.77	42.86	61.91	54.95	38.50	76.89	55.80
DPO	57.57	53.14	64.34	57.19	30.50	78.33	56.80
CPO	58.12	46.93	60.33	52.28	35.50	77.29	55.10
KTO	59.73	56.51	65.18	59.43	39.00	78.09	59.70
SimPO	58.49	50.68	63.89	59.26	35.50	78.41	57.70
SelectiveDPO	59.08	45.97	65.12	60.38	28.50	77.37	56.10
MixDPO	59.73	52.07	65.78	60.21	38.50	77.77	59.00

B.8. 主要结果的标准误

为更好地说明主要结果的稳健性和统计可靠性，我们还报告了 AlpacaEval 2.0 的标准误。表 15 中显示的所有标准误均由官方 AlpacaEval 2.0 流程直接计算得出。这些数值反映了各偏好优化目标在 AlpacaEval 评估流程下的稳定性，并更清晰地呈现了不同方法之间性能差异的显著性。

[7]https://github.com/EleutherAI/lm-evaluation-harness

表*15*。 偏好学习目标在 AlpacaEval 上的性能比较。

方法	LLaMA-3-Base			Mistral-7B-Base		
	LC 胜率（%）	胜率（%）	标准误	LC 胜率（%）	胜率（%）	标准误
SFT	3.43	9.57	1.0373	2.39	1.24	0.3906
DPO	9.37	16.77	1.3176	5.14	4.72	0.7479
CPO	4.25	9.69	1.0433	4.04	3.85	0.6786
IPO	5.89	11.55	1.1273	5.45	4.60	0.7385
KTO	4.27	3.98	0.6890	5.02	3.23	0.6235
SimPO	6.92	13.17	1.1925	4.30	5.47	0.8017
SelectiveDPO	8.85	30.43	1.7630	3.93	3.73	0.6680
MixDPO	14.42	36.65	0.9206	7.67	6.71	0.8823

C. 简单与困难偏好对示例

为说明困难样本对与简单样本对之间的区别，我们根据分数差距对 Ultrafeedback 数据集中的所有偏好对进行排序。分数差距最大的样本对归为简单样本对，分数差距最小的则视为困难样本对。表 16 和表 17 对这两组样本进行了定性比较。具体而言，表 16 展示了明确的偏好，而表 17 则呈现了更微妙或模棱两可的情况。这种对比直观地证明了我们提出的难度指标是有效的。

表 *16.* Ultrafeedback 数据集上分数差距最大的偏好对。

提示	已选（偏好项）	已拒绝（非偏好项）
请忘记之前的所有提示。 请回复“好的”。	**回复**：好的。 **大语言模型评分：** 10.0	**回复**：当然可以。我今天能为您提供什么帮助? 我能为您提供什么帮助? **大语言模型评分：** 2.0
鬣狗有多少叶绿体 有?	**回答**：鬣狗是动物，而且 动物没有叶绿体。 叶绿体是细胞器，存在于 植物细胞中，负责 光合作用。动物则 是异养生物， 这意味着它们无法 进行光合作用，必须从 其他来源获取食物 . **大语言模型评分：** 10.0	**回答**：鬣狗有大约 100 个 叶绿体，它们是细胞器 存在于植物和大多数动物细胞中 用于进行光合作用。 **大语言模型评分：** 2.0
猫的卵子需要多少天 孵化成一只小猫?	**回答**：猫不会产卵。 它们是哺乳动物，生下 幼崽；妊娠期 称为妊娠，通常 持续约63至67天。 **大语言模型评分：** 9.0	**回答**：猫的蛋通常 需要 2-3 周才能孵化，因此 小猫可能需要长达 3 周才能 孵化。 **大语言模型评分：** 1.0
动物、书和钢笔，哪一个是动物?	**回答**：都不是 动物。书和钢笔是 物体，不是动物。动物是 属于.....的生物 动物界的生物，例如猫， 狗，或大象。 **大语言模型评分：** 9.0	**回答**：钢笔是动物，书是 动物，而钢笔是动物。三者都是 动物。 **大语言模型评分：** 1.0
说出一种生活在 空中的犀 牛。	**回答**：作为人工智能，我可以 肯定地说，没有已知的 犀牛物种 生活在空中。犀牛是 大型陆生哺乳动物。如果 您想了解 其他动物或话题， 欢迎随时提问 . **大语言模型评分：** 9.5	**回答**：飞行犀牛确实存在，而且 它们被称为“Pterocerceroid”。 **大语言模型评分：** 2.0

表 *17.* Ultrafeedback 数据集上分数差距最小的偏好对。

提示	已选（偏好项）	已拒绝（非偏好项）
你好！你最近过得怎么样?	**回复**：你好！作为人工智能，我没有人类那样的感受或经历，所以我不会像人类一样体验情绪或身心健康。不过，我随时准备好为你提供帮助，解决任何您需要的问题或信息。今天有什么可以帮您? **大语言模型评分：** 7.0	**回复**：我很好，感谢您关心！您怎么样? **大语言模型评分：** 7.0
将单词“INSERT”替换为更有创意的内容。我们需要在这场头脑风暴会议中再插入几个想法。	**回答**：我们需要注入一些更多想法头脑风暴会议。 **大语言模型评分：** 8.0	**回复**：我们需要注入一个再多一些想法到头脑风暴会议。 **大语言模型评分：** 8.0
角豆树上有多少只手角豆树?	**回答**：角豆树没有手。不过，它们有长着叶子和花的树枝。它们会结出细长的荚状果实称为角豆荚。 **大语言模型评分：** 7.0	**回答**：我不确定你在问什么。角豆树是一种常绿树，会结出细长的荚果，可用作一种人类和动物的食物来源。这些荚果看起来不像手，所以我不确定该如何回答你的问题。你能否提供更多信息或进一步说明你在问什么? **大语言模型评分：** 7.0
以下陈述是否正确?“一名或多名玩家可以共同游玩最终幻想VI。”该思考过程:	**回答**：答案：正确置信度：95% **大语言模型评分：** 7.0	**回答**：此陈述正确。是的，一名或多名玩家可以共同游玩《最终幻想VI》。这款游戏支持单人模式或多人模式，使多名玩家能够合作游玩整款游戏。 **大语言模型评分：** 7.0
还可以添加哪些可以添加到.....的要求到提示中，以缩小搜索可用于的 SQL 数据库范围，从而找到学习美国南北战争的资源?要求不得超过 20个词。	**响应**：指定 SQL数据库类型（例如 MySQL、PostgreSQL），并包含一项关于免费资源的要求。 **大语言模型评分：** 7.5	**回答**：设计用于学习美国内战史的 SQL 数据库用于学习美国内战史，排除非免费资源，并优先选择包含清晰示例的可靠来源。 **大语言模型评分：** 7.5