用信息论引导消除奖励模型中的归纳偏置

Zhuo Li[1,2,3], Pengyu Cheng[1], Zhechao Yu[1], Feifei Tong[1], Anningzhe Gao[3], Tsung-Hui Chang[2,3], Xiang Wan[3], Erchao Zhao[1], Xiaoxi Jiang[1], Guanjun Jiang[1]

[1] Qwen Large Model Application Team, Alibaba, [2] The Chinese University of Hong Kong, [3] Shenzhen Research Institute of Big Data

* 本工作完成于阿里巴巴实习期间。 † 通讯作者。

奖励模型（RM）是在人类反馈强化学习（RLHF）中使大语言模型（LLM）与人类价值对齐的关键组件。然而，RM 训练数据普遍质量不高，包含容易导致过拟合与奖励黑客的归纳偏置。例如，更详细、更全面的回答通常更受人类偏好，但也往往更长，从而使回答长度成为一种常见偏置。现有少量去偏方法要么仅针对单一偏置类型，要么只建模简单线性相关（如 Pearson 系数）。为缓解更复杂、更多样的归纳偏置，我们提出一种新的信息论去偏方法 DIR（Debiasing via Information optimization for RM）。受信息瓶颈（IB）启发，DIR 在最大化 RM 分数与人类偏好对之间互信息的同时，最小化 RM 输出与偏置属性之间的互信息。基于信息论的理论支撑，DIR 能处理具备非线性相关的复杂偏置，扩展 RM 去偏在真实场景中的适用性。实验在长度、迎合（sycophancy）与格式三类偏置上验证了 DIR 的有效性。结果显示，DIR 不仅能有效缓解目标偏置，还能在多项基准上提升 RLHF 表现并增强泛化能力。代码与训练配方：https://github.com/Qwen-Applications/DIR。

1. 引言

将大语言模型（LLM）与人类价值对齐，是保证模型输出有用性与无害性的核心技术，且已在开放域对话场景中广泛应用（OpenAI, 2024; Touvron et al., 2023; Yang et al., 2024; Ouyang et al., 2022b; Kimi et al., 2025; Gemini, 2025）。为了获得更符合人类偏好的 LLM 行为，人类反馈强化学习（RLHF）已成为主流路线：先在人类偏好响应对上训练奖励模型（RM），再用 RM 作为奖励信号进行强化学习（RL）（Ouyang et al., 2022b; Rafailov et al., 2024b; DeepSeek-AI, 2025）。尽管 RLHF 在后训练中被广泛采用（DeepSeek-AI, 2025; Touvron et al., 2023），其训练稳定性长期受到质疑，容易导致策略崩溃与过拟合（Rafailov et al., 2024b; Zhu et al., 2024; Yu et al., 2025）。

在导致 RLHF 不稳定的诸多因素中，奖励模型黑客问题尤其关键：由于偏好数据质量有限，常包含冲突标注与归纳偏置（Zeng et al., 2024; Liu et al., 2025; Wang et al., 2025b; Liu et al., 2024a），RM 可能学习到与真实内容质量无关的表面属性（Skalse et al., 2025; Gao et al., 2023; Amodei et al., 2016）。例如，标注者通常被要求偏好信息更充分的回答，而更详细的回答往往更长，这会使 RM 误学到“越长越好”的启发式（Singhal et al., 2023）。除长度偏置外，风格与

格式模式（Zhang et al., 2025）以及迎合性措辞（Sharma et al., 2023; Denison et al., 2024）也被视为奖励建模中的典型归纳偏置。这些属性本身与回答质量并无直接关系，却与偏好标注高度相关（Sharma et al., 2023; Liu et al., 2024c）。在含偏数据上训练会扰动 RM 的学习目标，并显著损害 RLHF 的可靠性与泛化能力（Gao et al., 2023; Coste et al., 2023）。

为缓解奖励建模中的归纳偏置，已有研究进行了初步探索。Bu et al. (2025)、Chen et al. (2024)、Zhang et al. (2025) 将 Pearson 系数（Benesty et al., 2009）作为偏置度量，与 RM 损失联合最小化；但 Pearson 仅能捕捉线性相关，难以覆盖更一般场景。Shen et al. (2023) 通过新增 RM 头预测长度分数，仅适用于标量偏置且缺乏理论保证。Wang et al. (2025a) 施加较强外部约束，例如在 chosen/rejected 分布间最小化 MMD（Gretton et al., 2012），虽可抑制偏置却可能扭曲奖励地形，使功能差异明显的响应组得分塌缩。相较之下，仅做一般压缩的策略（如 InfoRM 中的信息瓶颈，Tishby et al., 2000; Miao et al., 2024）由于没有显式约束偏置属性，也无法保证有效去偏。

为在理论上有保证地统一消除归纳偏置，我们提出信息论去偏框架 DIR（Debiasing via Information optimization for RMs）。受信息瓶颈方法（Tishby et al., 2000）启发，我们用互信息（MI）（Kullback, 1997）建模“无关属性 - 人类偏好”之间的复杂偏置关系。DIR 同时最大化“响应内容质量 - 真实偏好标签”的互信息，并最小化“RM 偏好预测 - 偏置属性”的互信息。针对 MI 难以直接计算的问题（Poole et al., 2019），我们采用 BA 下界（Barber and Agakov, 2004）实现 MI 最大化估计，采用 CLUB 上界（Cheng et al., 2020）实现 MI 最小化估计。进一步地，我们引入“响应对相对偏置属性”的比较式正则，而非单样本偏置属性约束，从而在不扭曲奖励地形的前提下处理更复杂偏置，拓展了方法适用范围。我们在 LLM 能力基准（GSM8K, MMLU, ArenaHard, MT-Bench）及 RM 基准（RM-Bench, RewardBench）上进行了系统实验，覆盖长度、格式与迎合偏置设置，结果显示 DIR 稳定优于现有去偏方法。

2. 预备知识

人类反馈强化学习（RLHF）已成为对齐 LLM 的核心训练流程之一（Ouyang et al., 2022a）。给定奖励模型 $r_\phi(x, y)$ 对响应 $y \in Y$（在提示 $x \in X$ 下）的偏好程度打分，RLHF 优化策略 $\pi_\theta(y|x)$ 的目标为

$$\mathbb{E}_{x\sim X,\, y\sim\pi_\theta(\cdot|x)}\left[r(x,y) - \beta \cdot \mathrm{KL}\big(\pi_\theta(y|x) \,\|\, \pi_{\text{ref}}(y|x)\big)\right], \tag{1}$$

其中 $\pi_{\text{ref}}(y|x)$ 为参考策略，$\beta > 0$ 控制与参考策略的 KL 正则强度（Csiszár, 1975）。PPO（Schulman et al., 2017）是该目标最常用优化器（OpenAI, 2024; Bai et al., 2023; Rafailov et al., 2024b）；GRPO（Shao et al., 2024）进一步移除 critic，用组相对优势近似，在工程上更简洁（DeepSeek-AI, 2025; Yang et al., 2025）。

奖励建模（Reward Modeling）旨在学习人类偏好分布。设参数化 RM 为 $r_\phi : X \times Y \to \mathbb{R}$，其中 $r_\phi(x, y)$ 表示提示 x 与回答 y 的奖励分数。给定一对响应 $(y, \bar{y})$，若 $r(x, y) > r(x, \bar{y})$，则预测 $y \succ \bar{y}$；反之为 $y \prec \bar{y}$。定义二值指示变量 $1_{y \succ \bar{y}}$ 表示“人类偏好”事件，则

$$q_\phi(1_{y \succ \bar{y}} = 1 \mid x, y, \bar{y}) = \frac{\exp(r_\phi(x, y))}{\exp(r_\phi(x, y)) + \exp(r_\phi(x, \bar{y}))} = \sigma(r_\phi(x, y) - r_\phi(x, \bar{y})), \tag{2}$$

其中 $\sigma(\cdot)$ 为 Sigmoid。由于真实偏好分布 $p^*(1_{y \succ \bar{y}} \mid x, y, \bar{y})$ 不可得，训练时在偏好数据 $D_{\text{Pref}} = \{(x^i, y_w^i, y_l^i)\}_{i=1}^N$ 上最大化 q_ϕ 的对数似然：

$$\begin{aligned} L_{\text{RM}}(\phi) &= -\mathbb{E}_{1_{y \succ \bar{y}} \sim p^*} \log q_\phi(1_{y \succ \bar{y}} \mid x, y, \bar{y}) \\ &\approx -\frac{1}{N} \sum_{i=1}^{N} \log \sigma\big(r_\phi(x^i, y_w^i) - r_\phi(x^i, y_l^i)\big). \end{aligned} \tag{3}$$

该式即 Bradley-Terry 排序损失（Bradley and Terry, 1952）。

信息论方法（Information-theoretic Methods）从信息论视角优化深度模型（Chen et al., 2016; Hjelm et al., 2019; Yuan et al., 2021; Cheng et al., 2021）。其核心是将神经网络前向过程视作信息通道，用互信息衡量变量相关性：

$$I(x; y) = \mathbb{E}_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)} = \text{KL}\big(p(x, y) \,\|\, p(x)p(y)\big). \tag{4}$$

MI 能捕捉任意非线性相关，在多种任务中有效（Chen et al., 2016; Belghazi et al., 2018; Hjelm et al., 2019）。但直接估计往往不可 tractable，尤其仅有样本时。为此常用变分界近似（Oord et al., 2018; Cheng et al., 2020; Belghazi et al., 2021）。

Barber-Agakov（BA）下界（Barber and Agakov, 2004）通过变分分布 $q_\theta(y|x)$ 给出

$$I(x; y) \geq \mathbb{E}_{p(x,y)}[\log q_\theta(y|x)] + H[p] \;=:\; I_{\text{BA}}(x; y). \tag{5}$$

CLUB 上界（Cheng et al., 2020）则给出

$$I(x; y) \leq \mathbb{E}_{p(x,y)}[\log q_\theta(y|x)] - \mathbb{E}_{p(x)p(y)}[\log q_\theta(y|x)] \;=:\; I_{\text{CLUB}}(x; y). \tag{6}$$

最小化式 (6) 可减少 x, y 的信息耦合。典型应用是信息瓶颈（IB）（Tishby et al., 2000）：

$$\min_h \; I(x; h) - \lambda \cdot I(h; y), \tag{7}$$

其中 $\lambda > 0$ 控制“压缩输入”与“保留预测信息”的权衡。IB 已广泛用于表示学习（Saxe et al., 2019; Wan et al., 2021; Federici et al., 2020）。

3. 方法

我们从信息论视角重新审视奖励建模。给定输入查询 $x \in X$ 及响应对 $y, \bar{y} \in Y$，记 b 为与 $(x, y, \bar{y})$ 相关的目标偏置属性。去偏目标是学习奖励模型 $r_\phi(x, y)$：其偏好预测 $1_{y \succ \bar{y}}$ 应与响应内容质量高度相关，同时尽量不携带预定义偏置属性 b 的信息。受 IB（式 (7)）启发，我们将目标写为

$$\max_{\phi} \; I(1_{y \succ \bar{y}}; x, y, \bar{y}) - \lambda \cdot I(1_{y \succ \bar{y}}; b)\,, \tag{8}$$

其中 $\lambda > 0$ 平衡“偏好学习项”与“去偏项”。理想情况下，最小化该目标可促使 RM 从 $(x, y, \bar{y})$ 学习真实质量信号，同时降低对偏置属性的依赖。

直接优化 MI 在高维空间通常不可 tractable（Poole et al., 2019）。因此我们遵循已有工作（Oord et al., 2018; Cheng et al., 2021），分别用变分界（式 (5),(6)）估计式 (8) 中的两项。

偏好项估计。对 $I(1_{y \succ \bar{y}}; x, y, \bar{y})$ 采用 BA 下界：

$$I(1_{y \succ \bar{y}}; x, y, \bar{y}) \geq \mathbb{E}_{p^*(x,y,\bar{y},1_{y \succ \bar{y}})}[\log q_\phi(1_{y \succ \bar{y}} | x, y, \bar{y})] + H[p^*]. \tag{9}$$

其中 p^* 是偏好数据联合分布，$H[p^*]$ 对参数为常数。由式 (3) 可知，上式右侧期望项等价于 Bradley-Terry 排序损失（Bradley and Terry, 1952; Azar et al., 2024; Cheng et al., 2024）。故给定批数据 $D_{\text{Pref}} = \{(x^i, y_w^i, y_l^i)\}_{i=1}^B$，可用

$$L_{\text{Preference}}(\phi) := -\frac{1}{B} \sum_{i=1}^{B} \log \sigma\big(r_\phi(x^i, y_w^i) - r_\phi(x^i, y_l^i)\big) \tag{10}$$

近似式 (8) 中的偏好项。

去偏项估计。由于响应对 $(x, y, \bar{y})$ 足以确定偏置属性 b，RM 前向过程 $(b \to (x, y, \bar{y}) \to H \to 1_{y \succ \bar{y}})$ 可视为马尔可夫链（Shannon, 1948），其中 $H = [h_\phi(x, y), h_\phi(x, \bar{y})]$ 为 RM 主干最后隐藏状态。由数据处理不等式与 CLUB 上界（Cheng et al., 2020）可得

$$I(1_{y \succ \bar{y}}; b) \leq I(H; b) \leq I_{\text{CLUB}}(H; b). \tag{11}$$

在批数据 $D_{\text{Pref}} = \{(x^i, y_w^i, y_l^i, b^i)\}_{i=1}^B$ 上，用变分网络 $q_\psi(b|H)$ 计算

$$I_{\text{CLUB}}(H; b) \approx \frac{1}{B} \sum_{i=1}^{B} \left[\log q_\psi(b^i | H^i) - \frac{1}{B} \sum_{j=1}^{B} \log q_\psi(b^j | H^i) \right] =: L_{\text{Debiasing}}(\phi, \psi). \tag{12}$$

最小化 $L_{\text{Debiasing}}$ 即可削弱偏置属性 b 与 RM 表示 $h_\phi(x, y)$ 的相关性，从而使输出分数 $r_\phi(x, y)$ 减少偏置信号干扰。

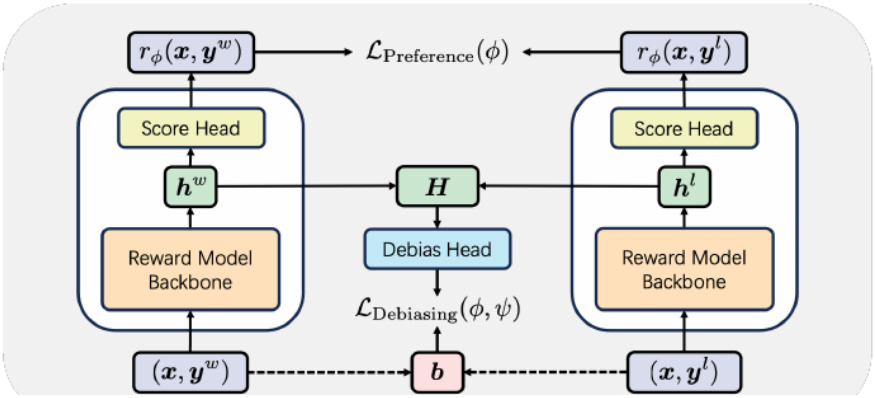

图 1: *

图 1 | DIR 框架示意。RM 架构由主干 transformer 与 score head 组成。$L_{\text{Preference}}(\phi)$ 基于偏好对的 score 差计算；主干隐藏状态 (h_w, h_l) 组成表示 H；$L_{\text{Debiasing}}(\phi, \psi)$ 在偏置标签 b 与去偏头输出之间计算。

算法 1：交替训练 $r_\phi(x, y)$ 与 $q_\psi(b|H)$。
输入：带偏置属性的偏好数据 $D_{\text{Pref}} = \{(x^i, y_w^i, y_l^i, b^i)\}_{i=1}^N$，学习率 $\alpha_1, \alpha_2 > 0$。
1) 采样一个 batch：$\{(x^i, y_w^i, y_l^i, b^i)\}_{i=1}^B$。
2) 编码得到 $H^i = [h_\phi(x^i, y_w^i), h_\phi(x^i, y_l^i)]$。
3) 进行若干步 bias estimator 更新：
$L_{\text{Estimator}}(\psi) = \frac{1}{B}\sum_{i=1}^B \log q_\psi(b^i|H^i)$，　$\psi \leftarrow \psi - \alpha_2 \nabla_\psi L_{\text{Estimator}}(\psi)$。
4) 计算 RM 偏好损失 $L_{\text{Preference}}(\phi)$（式 (10)）。
5) 计算 RM 去偏损失 $L_{\text{Debiasing}}(\phi, \psi)$（式 (12)）。
6) 总损失：$L_{\text{Total}}(\phi, \psi) = L_{\text{Preference}}(\phi) + \lambda L_{\text{Debiasing}}(\phi, \psi)$。
7) 更新 RM 参数：$\phi \leftarrow \phi - \alpha_1 \nabla_\phi L_{\text{Total}}(\phi, \psi)$。

如 Cheng et al. (2020) 所示，$q_\psi(b^i|H^i)$ 对真实条件分布 $p^*(b^i|H^i)$ 的逼近越准确，I_{CLUB} 作为 MI 上界估计器就越准确。因此在优化式 (12) 时，我们同步最大化 batch 上的对数似然，以维持估计器精度：

$$L_{\text{Estimator}}(\psi) := \frac{1}{B}\sum_{i=1}^{B} \log q_\psi(b^i|H^i). \tag{13}$$

总体目标。由以上推导，式 (8) 转化为可训练目标：

$$\min_{\phi} \; L_{\text{Preference}}(\phi) + \lambda \cdot L_{\text{Debiasing}}(\phi, \psi). \tag{14}$$

损失计算流程如图 1 所示。为保证 $L_{\text{Debiasing}}(\phi, \psi)$ 始终准确上界 $I(1_{y \succ \bar{y}}; b)$，我们在每个训练 batch 内交替更新 $r_\phi(x, y)$ 与 $q_\psi(b|H)$（算法 1），并将该方法称为 DIR（Debiasing via Information optimization for RMs）。

4. 相关工作

LLM 的奖励黑客。奖励黑客指策略模型利用奖励函数中的伪相关或目标错设，在未实现真实目标时获得较高分数。