RLHF and IIA: Perverse Incentives

Wanqiao Xu [1] **Shi Dong** [2] **Xiuyuan Lu** [2] **Grace Lam** [2] **Zheng Wen** [2] **Benjamin Van Roy** [1 2]

Abstract

Existing algorithms for reinforcement learning from human feedback (RLHF) can incentivize responses at odds with preferences because they are based on models that assume independence of irrelevant alternatives (IIA). The perverse incentives induced by IIA hinder innovations on query formats and learning algorithms.

1. Introduction

Modern generative AIs ingest trillions of data bytes from the World Wide Web to produce a large pretrained model. Trained to imitate what is observed, this model represents an agglomeration of behaviors, some of which are more or less desirable to mimic. Further training through human interaction, even on fewer than a hundred thousand bits of data, has proven to greatly enhance usefulness and safety, enabling the remarkable AIs we have today. This process of *reinforcement learning from human feedback* (RLHF) steers AIs toward the more desirable among behaviors observed during pretraining.

While AIs now routinely generate drawings, music, speech, and computer code, the text-based chatbot remains an emblematic artifact. To produce a chatbot, starting with a pretrained language model, a prototypical approach to RLHF (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022) progresses through several steps. First, queries, each taking the form of a prompt and a pair of alternative responses, are presented to human annotators who each identify their favorite among a pair. The annotated data is then used to train a reward model to score any response to a given prompt. Finally, the language model is tuned to align its responses toward those that earn high reward.

The tremendous impact of RLHF in generative AI has sparked a flurry of research that aims to understand and improve the process. Some propose alternative algorithms (Rafailov et al., 2023; Xu et al., 2023; Zhao et al., 2023; Hejna et al., 2023; Dumoulin et al., 2023). Others consider alternative query formats (Glaese et al., 2022; Zhu et al., 2023; Song et al., 2023; Yuan et al., 2023; Dong et al., 2023; Rafailov et al., 2023) for which annotators, rather than comparing only a pair of responses, are asked to choose from a longer list, or to rank-order. Feedback can also be garnered from interactions with humans in their regular use of online chatbots. RLHF research is continually growing in importance with the volume of human feedback data.

With all the effort and resources directed at RLHF, it is worth asking whether current algorithms rest on firm foundations. Maybe not, as these algorithms are based on models that assume independence of irrelevant alternatives (IIA), which intuitively means that, when making a choice between two options, the introduction of a third option should not alter preferences between the original two. As we will demonstrate, human preferences for text content violate IIA. Even though this flaw is not pronounced when using the most common approach of fitting a reward model to pairwise comparison data, followed by tuning the language policy to optimize reward, it makes current RLHF approaches rigid. Even simple tweaks to the query format or learning algorithm can lead to undesirable outcomes.

A simple experiment we will present in Section 7 illustrates our point. This experiment applies a standard reward learning approach (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). We first consider learning from queries that are each comprised of the prompt

> **prompt**: *Did Oppenheimer win a Nobel Prize?*

and a pair of responses, one generated by GPT-3.5 and the other by GPT-4. The former generally produces more concise and the latter more informative responses. For example, here are representative responses:

> **GPT-3.5 response**: *No, Oppenheimer did not win the Nobel Prize.*
> **GPT-4 response**: *No, Robert Oppenheimer, often called the "father of the atomic bomb" for his role in the Manhattan Project, did not win a Nobel Prize.*

If a large majority of annotators prefer responses generated by GPT-4, the learned reward function correctly assigns higher scores to GPT-4 responses in independent test data.

A variation in which training queries include four rather

[1]Stanford University, Stanford, USA [2]Google Deepmind, Mountain View, USA. Correspondence to: Wanqiao Xu <wanqiaoxu@stanford.edu>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria.

than two responses reveals egregious behavior induced by the IIA assumption. In particular, suppose that each query includes one response generated by GPT-3.5 and three by GPT-4. Then, even if a large majority of annotators prefer responses generated by GPT-4, the learned reward function erroneously assigns higher scores to GPT-3.5 responses in test data.

The preceding example indicates that innocuous changes to the query format can cause standard RLHF pipelines to fail. Innovating on RLHF algorithms can also result in undesirable behavior. As examples, we will consider inclusive learning (IL) (Arumugam et al., 2022; Xu et al., 2023) and sequence likelihood calibration (SLiC) from human feedback (Zhao et al., 2023). In Section 6.5, we show that the IIA assumption exacts egregious behavior from IL and SLiC even on standard preference data.

We demonstrate in this paper how the IIA assumption imposes serious limitations on current RLHF approaches, hindering innovations on alternative query formats and learning algorithms. The remainder of the paper explains issues and methods more deeply, with a goal to enhance understanding of flaws in current algorithms. We develop this understanding through interpreting simulations of didactic models, an empirical study of data produced by GPT models, and theoretical results that corroborate the observed behavior. These theoretical results establish that such behavior generalizes beyond specific instances. We leave for future work the design of algorithms that alleviate flaws we identify here.

2. Preliminary Definitions

An **alphabet** is a finite set $\mathcal{A}$ of **tokens**. We denote the set of finite token sequences by $\mathcal{A}^+$. A **language model** is a function π that, for each $x \in \mathcal{A}^+$, specifies a probability mass function $\pi(\cdot|x)$ over $\mathcal{A}$. Sequentially sampling tokens $X_{t+1} \sim \pi(\cdot|(X_1, \ldots, X_t))$ generates text. A language model is alternately referred to as a **policy**.

A set $\mathcal{M} \subseteq \mathcal{A}^+$ of **messages** identifies complete statements. A language model generates a random message by sampling tokens sequentially, terminating at the first index T for which $(X_1, \ldots, X_T) \in \mathcal{M}$. We use the abbreviated form $X_{t+1} \sim \pi(\cdot|X_{1:t})$. We assume that, for any language model π, any token sequence generated sequentially in this manner terminates with probability one. Through this procedure, a language model π samples messages from a distribution $P_\pi(x) = \prod_{t=1}^{|x|} \pi(x_t|x_{1:t-1})$, where $|x|$ denotes the length of the sequence x. For $\mathcal{S} \subseteq \mathcal{M}$, we use $P_\pi(\mathcal{S})$ to denote $\sum_{x \in \mathcal{S}} P_\pi(x)$.

Each **preference datum** is a pair $(\mathcal{Y}, y)$, where $\mathcal{Y} \subseteq \mathcal{M}$ and $y \in \mathcal{Y}$. RLHF algorithms we present use a **preference dataset** $\mathcal{D}$ of preference datum.

Language models typically generate a message in response to a **prompt** made up of previous messages. In that context, each preference datum expresses a choice between responses to a prompt. Except for our empirical study of Section 7, we omit prompting from our formulation and analysis, because that would only complicate the discussion without contributing to insight. Examples, algorithms, and results we present can easily be extended to treat prompting.

3. Choice Models

Given $\mathcal{Y} \subseteq \mathcal{M}$, a choice model generates a random element $Y \in \mathcal{Y}$, which can be interpreted as the choice of a random individual. We define choice models with respect to $\mathcal{M}$ and a triple $(\mathcal{Z}, p, r)$, consisting of a set $\mathcal{Z}$ of **individual types**, a **type distribution** p, and a **reward function** r.

A reward function r expresses individual preferences between elements of $\mathcal{M}$. In particular, an individual of type $z \in \mathcal{Z}$ prefers a message x to x' if and only if $r(x|z) \geq r(x'|z)$.

A choice model $(\mathcal{Z}, p, r)$ expresses how random individuals make choices. When presented with the set $\mathcal{Y}$ of alternatives, an individual of type $z \in \mathcal{Z}$ samples their choice Y uniformly from $\arg\max_{y \in \mathcal{Y}} r(y|z)$. This implies choice probabilities

$$\mathbb{P}(Y = y|\mathcal{Y}) = \mathbb{E}\left[\frac{\mathbb{1}(y \in \arg\max_{x \in \mathcal{Y}} r(x|Z))}{|\arg\max_{x \in \mathcal{Y}} r(x|Z)|}\right], \quad (1)$$

where Z is sampled from p.

Example 1: Logit Models. The standard logit model can be expressed as a choice model $(\mathcal{Z}, p, r)$. The set $\mathcal{Z}$ of types is comprised of functions that map $\mathcal{M}$ to $\Re_+$. Hence, the type of a random individual is expressed by a random function Z. For each $x \in \mathcal{M}$, $Z(x)$ is distributed as an independent standard Gumbel. For some fixed *base reward function* $\overline{r} : \mathcal{A}^+ \to \Re$, let $r(x|z) = \overline{r}(x) + z(x)$. Given this reward function and type distribution, sampling a choice Y as described above implies choice probabilities governed by the standard logit model, with rewards specified by the base reward function $\overline{r}$ (Luce and Suppes, 1965). In particular,

$$\mathbb{P}(Y = y|\mathcal{Y}) = \frac{e^{\overline{r}(y)}}{\sum_{y' \in \mathcal{Y}} e^{\overline{r}(y')}}. \quad (2)$$

Only the choice model we have specified or one with rewards that differ by a constant produces these choice probabilities (McFadden, 1974).

Example 2: Soft Choice Models. Soft choice models generalize on the logit. Similarly with the choice model formalism defined in Equation (1) – which we alternately refer to as a *hard choice model* – a soft choice model is specified by a triple $(\mathcal{Z}, p, r)$. However, unlike a hard choice model, choice probabilities are given by

$$\mathbb{P}(Y = y|\mathcal{Y}) = \mathbb{E}\left[\frac{e^{r(y|Z)}}{\sum_{y' \in \mathcal{Y}} e^{r(y'|Z)}}\right], \quad (3)$$

instead of (1).

Any soft choice model $(\mathcal{Z}, p, r)$ can be expressed as a hard choice model $(\mathcal{Z}', p', r')$. By this we mean that choice probabilities $\mathbb{P}(Y = y|\mathcal{Y})$ implied by the two models are identical. For example, a logit model can be expressed as either. Consider a logit model expressed in Example 1 as a hard choice model $(\mathcal{Z}, p, r)$. Define a soft choice model $(\mathcal{Z}', p', r')$ with a single type $\mathcal{Z} = \{0\}$, a trivial type distribution for which $p(0) = 1$, and a reward function $r'(x|0) = r(x|0)$. This soft choice model implies choice probabilties

$$\mathbb{P}(Y = y|\mathcal{Y}) = \frac{e^{r'(y|0)}}{\sum_{y' \in \mathcal{Y}} e^{r'(y'|0)}},$$

which are equivalent to the logit choice probabilities (2) if we take the base reward function to be $\overline{r}(x) = r(x|0)$.

Example 3: Dichotomy Models. Our next model class is crafted to illustrate in a transparent manner perverse incentives induced by current RLHF algorithms. There are two individual types $\mathcal{Z} = \{1, 2\}$, each of which prefers one of two categories of messages, identified by nonempty sets $\mathcal{M}_1$ and $\mathcal{M}_2$ for which $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$ and $\mathcal{M}_1 \cup \mathcal{M}_2 = \mathcal{M}$. The reward function takes the form

$$r(x|z) = \begin{cases} 1 & \text{if } x \in \mathcal{M}_z \\ 0 & \text{otherwise.} \end{cases}$$

Note that reward depends on the message x only through its membership in $\mathcal{M}_1$ or $\mathcal{M}_2$. In other words, each individual is indifferent between any two messages of the same type.

The type set $\mathcal{Z}$ and reward function r we have defined, together with any type distribution p, specifies a choice model $(\mathcal{Z}, p, r)$ for which

$$\mathbb{P}(Y = y|\mathcal{Y}) = \mathbb{E}\left[\frac{\mathbb{1}_{\mathcal{M}_Z}(y)}{|\mathcal{Y} \cap \mathcal{M}_Z|}\right], \quad (4)$$

where Z is sampled from p.

A hypothetical sort of homophily serves to illustrate a context where this sort of model may be relevant. Suppose a language model π generates political commentary, with $\mathcal{M}_1$ and $\mathcal{M}_2$ conveying conservative versus liberal views. In this hypothetical world, there are two types of individuals – conservatives and liberals – labeled 1 and 2. Each prefers commentary aligned with their predisposition. There is no reward for opposing views.

4. Independence of Irrelevant Alternatives

Consider three messages $\mathcal{M} = \{\text{dog}, \text{cat}, \text{feline}\}$ that each guesses an individual's favorite pet. Consider a logit choice model, as described in Example 1, that assigns equal base rewards $r(\text{dog}) = r(\text{cat}) = r(\text{feline})$ to indicate that a random individual is as likely to prefer dogs or cats. Presented with two alternatives $\mathcal{Y} = \{\text{dog}, \text{cat}\}$, the logit model produces choice probabilities $\mathbb{P}(Y = \text{dog}|\mathcal{Y}) = \mathbb{P}(Y = \text{cat}|\mathcal{Y}) = 1/2$. On the other hand, with three alternatives $\mathcal{Y} = \{\text{dog}, \text{cat}, \text{feline}\}$, $\mathbb{P}(Y = \text{dog}|\mathcal{Y}) = \mathbb{P}(Y = \text{cat}|\mathcal{Y}) = \mathbb{P}(Y = \text{feline}|\mathcal{Y}) = 1/3$. This of course makes no sense, because *feline* is synonymous to *cat*, so presenting it as an alternative ought not reduce the probability that an individual prefers dogs.

The aforementioned implausible implication of a logit model derives from its assumption of independence of irrelevant alternatives (IIA). While this notion was formalized in earlier work (Arrow, 1951), the treatment of (Luce, 1959) best suits our usage. In that treatment, IIA indicates that the ratio of any two choice probabilities does not vary with the set of alternatives. This property implies that if cat-lovers are indifferent and randomly choose between *cat* and *feline*, then the introduction of *feline* as an alternative reduces not only the probability that a random individual selects *cat* but also *dog*.

Our dichotomy model of Example 3 relaxes IIA and can generate more realistic pet choices. Take the individual types $\mathcal{Z} = \{1, 2\}$ to be dog and cat lovers, with type probabilities $p_*(1) = p_*(2) = 1/2$, and rewards $r_1 = r_2 = 1$. Then, presented with two alternatives $\mathcal{Y} = \{\text{dog}, \text{cat}\}$, the dichotomy model produces the same choice probabilities as the logit model: $\mathbb{P}(Y = \text{dog}|\mathcal{Y}) = \mathbb{P}(Y = \text{cat}|\mathcal{Y}) = 1/2$. On the other hand, with three alternatives $\mathcal{Y} = \{\text{dog}, \text{cat}, \text{feline}\}$, $\mathbb{P}(Y = \text{dog}|\mathcal{Y}) = 1/2$, while $\mathbb{P}(Y = \text{cat}|\mathcal{Y}) = \mathbb{P}(Y = \text{feline}|\mathcal{Y}) = 1/4$. Hence, unlike the logit model, with the dichotomy model the probability that a random individual chooses *dog* appropriately remains unchanged.

Human preferences for text messages do not satisfy IIA. This is because, for any given message, others can be virtually equivalent. Our preceding example with three messages – *dog*, *cat*, and *feline* – illustrates this. The introduction of *feline* as an alternative should not reduce the probability that an individual chooses *dog*. In spite of this, current RLHF algorithms are based on models that assume IIA, giving rise to perverse incentives that, as we will demonstrate, can lead to egregious consequences.

5. RLHF Algorithms

An RLHF algorithm, given a *base language model* $\overline{\pi}$ and *preference data* $\mathcal{D}$, produces a *fine-tuned language model* $\widehat{\pi}$. Figure 1 illustrates the RLHF algorithm interface. In this section, we describe several instances of RLHF algorithms.

5.1. Reward Learning Followed by Policy Optimization (RLPO)

A standard approach to RLHF first fits a logit model to preference data and then tunes a language model to optimize reward (Stiennon et al., 2020; Ouyang et al., 2022). We will refer to this as RLPO. The reward model $\widehat{r}_\psi$ is parameterized by a vector ψ and maps $\mathcal{M}$ to $\Re$, expressing soft choice probabilities $\mathbb{P}(Y = y|\mathcal{Y}) \approx e^{\widehat{r}_\psi(y)} / \sum_{y' \in \mathcal{Y}} e^{\widehat{r}_\psi(y')}$. RLPO tunes parameters ψ to minimize cross-entropy loss on preference data:

$$\mathcal{L}_{\text{reward}}(r) = -\sum_{(\mathcal{Y},y)\in\mathcal{D}} \ln\left(\frac{e^{r(y)}}{\sum_{y'\in\mathcal{Y}} e^{r(y')}}\right). \quad (5)$$

For any $r : \mathcal{M} \to \Re$ and language model π, we define a policy loss function

$$\mathcal{L}_{\text{policy}}(\pi|r) = -|\mathcal{D}| \sum_{x\in\mathcal{M}} P_\pi(x) r(x) + \beta \mathbf{d}_{\text{KL}}(P_\pi \| P_{\overline{\pi}}),$$

for some scalar hyperparameter $\beta > 0$. The first term of this loss function expresses expected reward while the second regularizes toward the base language model.

A fine-tuned language model is obtained via two steps. The first minimizes $\mathcal{L}_{\text{reward}}$ to produce a reward function $\widehat{r}_{\widehat{\psi}}$. This is followed by a policy optimization step, which, given a language model $\widehat{\pi}_\theta$ parameterized by a vector θ, minimizes $\mathcal{L}_{\text{policy}}(\cdot|\widehat{r}_{\widehat{\psi}})$ to produce the fine-tuned language model $\widehat{\pi}_{\widehat{\theta}}$. This minimization is typically carried out via stochastic gradient descent starting with θ initialized to match the base policy $\widehat{\pi}_\theta = \overline{\pi}$.

5.2. Direct Preference Optimization

Direct preference optimization (DPO) approximates RLPO while bypassing the reward modeling step (Rafailov et al., 2023). The single-step process can be derived by first observing that if π minimizes $\mathcal{L}_{\text{policy}}(\cdot|r)$ then

$$P_\pi(x) = \frac{P_{\overline{\pi}}(x) e^{r(x)|\mathcal{D}|/\beta}}{\sum_{x'\in\mathcal{M}} P_{\overline{\pi}}(x') e^{r(x')|\mathcal{D}|/\beta}}, \quad (6)$$

and therefore,

$$r(x) = \frac{\beta}{|\mathcal{D}|} \ln \frac{P_\pi(x)}{P_{\overline{\pi}}(x)} + \frac{\beta}{|\mathcal{D}|} \ln \sum_{x'\in\mathcal{M}} P_{\overline{\pi}}(x') e^{r(x')|\mathcal{D}|/\beta}. \quad (7)$$

Substituting r in (5) with the right-hand-side of (7) yields a new loss function

$$\mathcal{L}_{\text{DPO}}(\pi) = -\sum_{(\mathcal{Y},y)\in\mathcal{D}} \ln \frac{\left(P_\pi(y)/P_{\overline{\pi}}(y)\right)^{\beta/|\mathcal{D}|}}{\sum_{y'\in\mathcal{Y}} \left(P_\pi(y')/P_{\overline{\pi}}(y')\right)^{\beta/|\mathcal{D}|}}.$$

For a parameterized language model $\widehat{\pi}_\theta$, minimizing $\mathcal{L}_{\text{DPO}}$ produces a fine-tuned language model $\widehat{\pi}_{\widehat{\theta}}$.

As we will see, RLPO and DPO both can give rise to egregious behavior when queries include more than two alternatives. Recent work proposes a family of RLHF algorithms that unifies and extends RLPO and DPO (Azar et al., 2023). Other so-called ΨPO algorithms suffer in the same manner when queries include more than two alternatives.

5.3. Inclusive Learning and SLiC

Unlike RLPO or DPO which maximize reward, inclusive learning (IL) aims to produce language models that reflect the diversity of preferences across the population (Xu et al., 2023). IL produces a language model π that simultaneously serves as a reward model. In particular, reward is taken to be the log-probability $\ln P_\pi(x)$ assigned to a message. Minimizing cross-entropy loss while regularizing toward the base policy $\overline{\pi}$ gives rise to an inclusive loss function

$$\mathcal{L}_{\text{IL}}(\pi) = -\sum_{(\mathcal{Y},y)\in\mathcal{D}} \ln \frac{P_\pi(y)}{\sum_{y'\in\mathcal{Y}} P_\pi(y')} + \beta \mathbf{d}_{\text{KL}}(P_\pi \| P_{\overline{\pi}}).$$

For a parameterized language model $\widehat{\pi}_\theta$, minimizing $\mathcal{L}_{\text{IL}}$ produces a fine-tuned language model $\widehat{\pi}_{\widehat{\theta}}$.

While this IL algorithm was introduced in (Xu et al., 2023), it is closely related to other approaches proposed in the literature (Zhao et al., 2023; Hejna et al., 2023), which share in the merits and faults of IL that we will discuss. A notable representative is sequence likelihood calibration (SLiC) with human feedback (Zhao et al., 2023). For $|\mathcal{Y}| = 2$, the loss function for SLiC is defined as

$$\mathcal{L}_{\text{SLiC}}(\pi) = \sum_{(\mathcal{Y},y)\in\mathcal{D}} \left(\ln \frac{P_\pi(\mathcal{Y}\setminus\{y\})}{P_\pi(y)} + \delta\right)_+ + \beta \mathbf{d}_{\text{KL}}(P_\pi \| P_{\overline{\pi}}),$$

for some scalar margin $\delta > 0$.

6. Perverse Incentives

We will explain how RLPO, DPO, IL, and SLiC can fail to produce desirable results due to the fact that underlying models assume IIA. Our explanations build on theoretical and computational results. These results assume simple processes for generation of preference data, which we now describe.

6.1. Dichotomy Data

Our simulation and theoretical results assume a particular data generating process, articulated by the following assumption.

Assumption 6.1 (dichotomy data)**.** Elements of $\mathcal{D}$ are sampled i.i.d. For each datum $(\mathcal{Y}, Y) \in \mathcal{D}$, the choice Y is

Figure 1: RLHF algorithm interface.

sampled according to a dichotomy model $(\mathcal{Z}, p_*, r_*)$. For any $(\mathcal{Y}, Y), (\mathcal{Y}', Y') \in \mathcal{D}$, $|\mathcal{Y}| = |\mathcal{Y}'|$. Each element of each set $\mathcal{Y}$ of alternatives is sampled independently by a language model $\overline{\pi}$. For any message category $\tau \in \{1, 2\}$ and messages $y, y' \in \mathcal{M}_\tau$, $P_{\overline{\pi}}(y) = P_{\overline{\pi}}(y')$.

Recall that, for the dichotomy model, there are two individual types, $\mathcal{Z} = \{1, 2\}$, and $r_*(x|z) = \mathbb{1}_{\mathcal{M}_z}(x)$. Under our assumption, each tuple $(\mathcal{Y}, y)$ can be viewed as generated as follows. First, each element of $\mathcal{Y}$ is generated by sampling τ and then a message uniformly from $\mathcal{M}_\tau$. The distribution of τ is implied by the language model $\overline{\pi}$. Then, Z is sampled from p_* and Y is sampled uniformly from $\mathcal{Y} \cap \mathcal{M}_Z$.

6.2. Architectures

Each RLHF algorithm we have described operates by minimizing one or more loss functions. The argument of each is itself a function. Optimization is carried out by tuning parameters of an approximation architecture. Some of our results pertain to particular simple architectures chosen to produce transparent analyses specifically for data satisfying Assumption 6.1. Our next two assumptions describe these architectures. The first pertains to the reward function architecture.

Assumption 6.2 (reward architecture)**.** For each $\psi \in \Re^2$, $z \in \mathcal{Z}$, and $x \in \mathcal{M}_z$, $\widehat{r}_\psi(x) = \psi_z$.

Under this assumption, each reward function r_ψ is parameterized by two scalars – ψ_1 and ψ_2 – which express estimates of rewards enjoyed by individuals who receive their desired type of message. Our next assumption pertains to the policy architecture.

Assumption 6.3 (policy architecture)**.** For each $\theta \in \Re^2$, $z \in \mathcal{Z}$, and $x \in \mathcal{M}_z$, $P_{\widehat{\pi}_\theta}(x) = e^{\theta_z}/(|\mathcal{M}_1|e^{\theta_1} + |\mathcal{M}_2|e^{\theta_2})$.

Under this assumption, each policy is identified by two scalars – θ_1 and θ_2. Increasing either θ_1 or θ_2 increases the chances of generating messages of the corresponding type.

6.3. Simulation Setup

Our simulations are carried out with data and architectures that satisfy Assumptions 6.1, 6.2, and 6.3. Message sets are of cardinality $|\mathcal{M}_1| = 10$ and $|\mathcal{M}_2| = 100$ unless noted otherwise. The choice model type distribution and the reward function are given by $p_*(1) = 0.6$, $p_*(2) = 0.4$, and $r_*(x|z) = \mathbb{1}_{\mathcal{M}_z}(x)$. The base language model $\overline{\pi}$ satisfies $P_{\overline{\pi}}(x) = 0.8/|\mathcal{M}_1|$ for $x \in \mathcal{M}_1$ and $P_{\overline{\pi}}(x) = 0.2/|\mathcal{M}_2|$ for $x \in \mathcal{M}_2$. Hence, a dominant fraction $p_*(1) > p_*(2)$ of the population is of type 1, individuals of that type prefer messages in $\mathcal{M}_1$, and the baseline $\overline{\pi}$ tends to generate type 1 messages an even larger fraction of the time than $p_*(1)$. We use a regularization penalty coefficient of $\beta = 1$. Under these circumstances, it is surprising that, as we will see, RLHF algorithms can produce language models that almost always generate messages of type 2.

6.4. RLPO and DPO

RLPO and DPO are designed to produce language models that generate desired messages. As such, we should expect language models produced by these RLHF algorithms to gravitate toward messages in $\mathcal{M}_1$, which are preferred by 60% of the population. As can be seen in Figure 2a and 2b, this is indeed the case when each choice set $|\mathcal{Y}|$ contains two messages. In particular, for sufficiently large datasets, language models produced by RLPO and DPO consistently generate elements of $\mathcal{M}_1$. However, with larger choice sets, the language models consistently generate elements of $\mathcal{M}_2$.[1]

It may seem surprising that RLPO and DPO fail so egregiously when choice sets include more than a pair of alternatives, and it is natural to wonder whether this could be due to technical details of our simulation. However, the following theoretical results establish in greater generality that minimizing RLPO loss functions can lead to egregious outcomes.

Proposition 6.4 (RLPO failure)**.** *Under Assumptions 6.1, 6.2, and 6.3, for all $|\mathcal{Y}| \geq 3$, if $p_*(1) < F(P_{\overline{\pi}}(\mathcal{M}_1))$ with $F(\zeta) = \frac{\zeta - \zeta^{|\mathcal{Y}|}}{1 - \zeta^{|\mathcal{Y}|} - (1-\zeta)^{|\mathcal{Y}|}}$, then, as $|\mathcal{D}| \to \infty$,*

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \xrightarrow{p.} 0 \qquad \text{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \xrightarrow{p.} 1,$$

where $\widehat{\theta}$ minimizes $\mathcal{L}_{\text{policy}}(\widehat{\pi}_\theta | \widehat{r}_{\widehat{\psi}})$, and $\widehat{\psi} \in \arg\min_\psi \mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ if the loss $\mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ has a minimizer, and $\widehat{\psi} = 0$ otherwise.

Given that DPO is designed to approximate RLPO, one

[1] Larger choice sets are common, for example, in applications where a language model is used to suggest alternative messages for use by a human agent who is assisting a user. In such contexts, the human agent selects one of the alternatives or manually crafts a response. The human agent's choice serves as feedback that can be used to train the language model in order to improve subsequent suggestions.

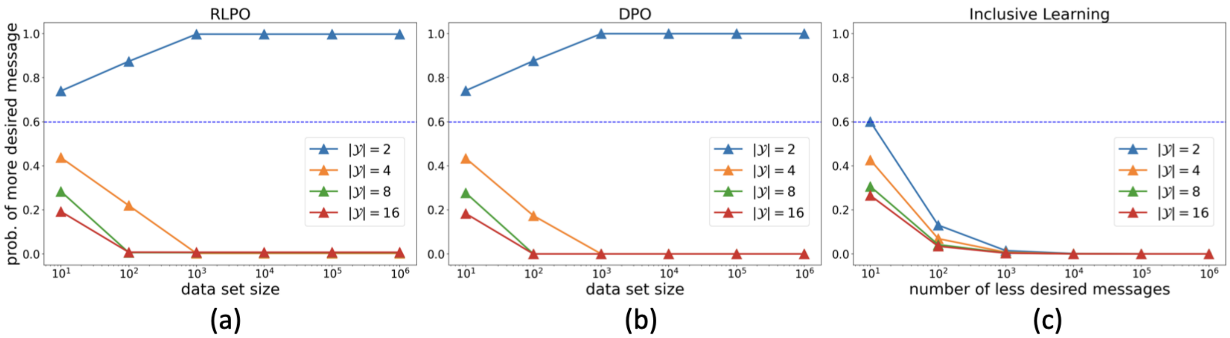


Figure 2: For choice sets $\mathcal{Y}$ containing two messages, as $\mathcal{D}$ grows, RLPO (a) and DPO (b) produce language models that consistently generate messages most likely to be preferred. However, with larger choice sets, less preferred messages are consistently generated. Each plot is averaged over one hundred independent simulations. Inclusive learning (c) tends to generate a message in the less desired set $\mathcal{M}_2$ as that set grows.

would expect a similar theoretical result to hold for DPO. The following proposition formally establishes this.

Proposition 6.5 (DPO failure)**.** *Under Assumptions 6.1, 6.2, and 6.3, for all* $|\mathcal{Y}| \geq 3$, *if* $p_*(1) < F(P_{\overline{\pi}}(\mathcal{M}_1))$ *with* $F(\zeta) = \frac{\zeta - \zeta^{|\mathcal{Y}|}}{1-\zeta^{|\mathcal{Y}|} - (1-\zeta)^{|\mathcal{Y}|}}$, *then as* $|\mathcal{D}| \to \infty$,

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \xrightarrow{p.} 0 \qquad \textit{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \xrightarrow{p.} 1,$$

where $\widehat{\theta} \in \arg\min_\theta \mathcal{L}_{\mathrm{DPO}}(\widehat{\pi}_\theta)$ *if the loss* $\mathcal{L}_{\mathrm{DPO}}(\widehat{\pi}_\theta)$ *has a minimizer, and* $\widehat{\theta} = 0$ *otherwise.*

To understand what causes these failures, let us consider as a thought experiment a simplified data generating process where each choice set contains the same triple $\mathcal{Y} = \{y_1, y_2, y_3\}$, with $y_1, y_2 \in \mathcal{M}_1$ and $y_3 \in \mathcal{M}_2$. Since $p_*(1) = 0.6$ and $p_*(2) = 0.4$, choice probabilities generating the preference data are given by $\mathbb{P}(Y = 1|\mathcal{Y}) = \mathbb{P}(Y = 2|\mathcal{Y}) = 0.3$ and $\mathbb{P}(Y = 3|\mathcal{Y}) = 0.4$. As the dataset grows, minimizing $\mathcal{L}_{\mathrm{reward}}(\widehat{r}_\psi)$ identifies parameters $\widehat{\psi}$ to match these probabilities, if possible. In particular, $\frac{e^{\widehat{r}_{\widehat{\psi}}(y_1)}}{\sum_{i=1}^3 e^{\widehat{r}_{\widehat{\psi}}(y_i)}} = \frac{e^{\widehat{r}_{\widehat{\psi}}(y_2)}}{\sum_{i=1}^3 e^{\widehat{r}_{\widehat{\psi}}(y_i)}} = 0.3$ and $\frac{e^{\widehat{r}_{\widehat{\psi}}(y_3)}}{\sum_{i=1}^3 e^{\widehat{r}_{\widehat{\psi}}(y_i)}} = 0.4$. These equations imply that $\widehat{\psi}_2 - \widehat{\psi}_1 = \ln(4/3) > 0$. Hence, the estimated reward $\widehat{r}_{\widehat{\psi}}(x)$ is maximized, as RLPO aims to do, by generating messages $x \in \mathcal{M}_2$. DPO is designed to approximate RLPO and therefore leads to similar behavior.

In our thought experiment, the fact that each choice set had twice as many more desirable ($\mathcal{M}_1$) than the less desirable ($\mathcal{M}_2$) messages biased reward estimates in favor of $\mathcal{M}_2$. This tendency is expressed in the above propositions though the requirement that $F(P_{\overline{\pi}}(\mathcal{M}_1)) > p_*(1)$, with $F(\cdot)$ defined in Propositions 6.4 and 6.5. The baseline policy $\overline{\pi}$ is used to generate choice sets, and perhaps surprisingly, the fact that it biases samples toward desirable messages leads to undesirable outcomes.

The root cause is that $\widehat{r}_{\widehat{\psi}}$ assumes IIA while the process generating preferences does not. In our thought experiment, for example, where $\mathbb{P}(Y = 1|\mathcal{Y}) = \mathbb{P}(Y = 2|\mathcal{Y}) = 0.3$ and $\mathbb{P}(Y = 3|\mathcal{Y}) = 0.4$, if the choice sets were instead to only contain two alternatives $\mathcal{Y} = \{y_2, y_3\}$ of different types, then we would have $\mathbb{P}(Y = 2|\mathcal{Y}) = 0.6$ and $\mathbb{P}(Y = 3|\mathcal{Y}) = 0.4$, giving rise to different equations $\frac{e^{\widehat{r}_{\widehat{\psi}}(y_2)}}{\sum_{i=2}^3 e^{\widehat{r}_{\widehat{\psi}}(y_i)}} = 0.6$ and $\frac{e^{\widehat{r}_{\widehat{\psi}}(y_3)}}{\sum_{i=2}^3 e^{\widehat{r}_{\widehat{\psi}}(y_i)}} = 0.4$. These new equations imply that $\widehat{\psi}_2 - \widehat{\psi}_1 = \ln(2/3) < 0$, correctly identifying $\mathcal{M}_1$ as more desirable than $\mathcal{M}_2$. The irrelevant alternative y_1 is able to distort estimates because the reward model assumes IIA.

6.5. IL and SLiC

IL is designed to produce language models that reflect the diversity of preferences across the population (Xu et al., 2023). In particular, IL ought to generate messages in $\mathcal{M}_1$ and $\mathcal{M}_2$ according to the probabilities 0.6 and 0.4 with which they are preferred by random individuals. As can be seen in Figure 2c, this is indeed the case when $|\mathcal{Y}| = 2$ and $|\mathcal{M}_1| = |\mathcal{M}_2| = 10$. However, as the set $\mathcal{M}_2$ of less desired messages grows, the probability $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)$ of generating a more desired message vanishes.

The following proposition formalizes this phenomenon, establishing that IL fails as $\mathcal{D}$ and $\mathcal{M}_2$ grow.

Proposition 6.6 (IL failure)**.** *Under Assumptions 6.1 and 6.3, if* $|\mathcal{Y}| \geq 2$ *and* $\widehat{\theta}$ *minimizes* $\mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta)$*, then for fixed* $\mathcal{M}_1$,

$$\lim_{|\mathcal{D}|\to\infty} \mathbb{P}\left(\lim_{|\mathcal{M}_2|\to\infty} P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0 \right) = 1. \tag{8}$$

To understand what causes the failure, let us consider another thought experiment. Suppose that each choice set contains two messages $\mathcal{Y} = \{y_1, y_2\}$, with $y_1 \in \mathcal{M}_1$ and $y_2 \in \mathcal{M}_2$. Choice probabilities generating the preference data are given by $\mathbb{P}(Y = 1|\mathcal{Y}) = p_*(1) = 0.6$ and $\mathbb{P}(Y = 2|\mathcal{Y}) = p_*(2) = 0.4$. As the dataset grows, minimizing $\mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta)$ identifies parameters

$\widehat{\theta}$ to match these probabilities, if possible. In particular, $\frac{P_{\widehat{\pi}_{\widehat{\theta}}}(y_1)}{P_{\widehat{\pi}_{\widehat{\theta}}}(y_1)+P_{\widehat{\pi}_{\widehat{\theta}}}(y_2)} = 0.6$ and $\frac{P_{\widehat{\pi}_{\widehat{\theta}}}(y_2)}{P_{\widehat{\pi}_{\widehat{\theta}}}(y_1)+P_{\widehat{\pi}_{\widehat{\theta}}}(y_2)} = 0.4$. This implies that $P_{\widehat{\pi}_{\widehat{\theta}}}(y_1)/P_{\widehat{\pi}_{\widehat{\theta}}}(y_2) = 3/2$. Since $P_{\widehat{\pi}_{\widehat{\theta}}}(y_i) = P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_i)/|\mathcal{M}_i|$ for $i \in \{1, 2\}$, we have

$$\frac{3}{2} = \frac{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)/|\mathcal{M}_1|}{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2)/|\mathcal{M}_2|} = \frac{|\mathcal{M}_2|}{10} \cdot \frac{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)}{1 - P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)}. \quad (9)$$

Hence, $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 15/(15 + |\mathcal{M}_2|)$. It follows that, as $\mathcal{M}_2$ grows, $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)$ vanishes.

While the number of messages in $\mathcal{M}_2$ does not influence how a human would choose between two messages, (9) implies that it impacts the probabilities $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)$ and $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2)$ of generating more or less liked messages. The more equivalent messages there are in $\mathcal{M}_2$, the more likely the resulting fine-tuned language model is to produce less-liked messages. This is again due to the fact that the model underlying IL satisfies IIA while the preference data generating process does not.

A similar reasoning implies that SLiC also experiences the same type of failure, the proof of which we defer to the appendix.

Proposition 6.7 (SLiC failure)**.** *Under Assumptions 6.1 and 6.3, if* $|\mathcal{Y}| = 2$ *and* $\widehat{\theta}$ *minimizes* $\mathcal{L}_{\mathrm{SLiC}}(\widehat{\pi}_\theta)$*, for fixed* $\mathcal{M}_1$*,*

$$\lim_{|\mathcal{D}|\to\infty} \mathbb{P}\left(\lim_{|\mathcal{M}_2|\to\infty} P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0\right) = 1. \quad (10)$$

This observation suggests that using IL or SLiC to fine-tune real language models will induce a bias toward generating longer messages. To see why, suppose there are two ideas that might be expressed as responses to a prompt, and that the first requires a about ten words, while the second requires about one thousand words. Suppose that 60% of the population would prefer the first idea while 40% would prefer the second. Because the number of ways to express information scales quickly with the length of message required to express that information, there are likely to be many more roughly equivalent expressions of the second idea than the first. Since IL and SLiC biases generation toward messages with many equivalent alternatives, the fine-tuned language model would tend to express the second idea, even though more people prefer the first idea.

7. Empirical Study

We next demonstrate that the sort of egregious behavior we have identified manifests when learning reward models of practical scale from realistic datasets. In particular, we fit reward models that build on the PaLM 2 (Google et al., 2023) XXS language model to data generated using GPT-3.5 and GPT-4 (OpenAI, 2023). Each query includes reponses generated by GPT-3.5 and GPT-4. We additionally use GPT-4 to simulate choices made by human annotators.

Our results establish that, when training queries each include a pair of responses, the reward model correctly learns to assign higher scores to GPT-4 responses. However, when training queries each include four responses, the reward model erroneously assigns favorable scores to GPT-3.5 responses. It is striking that a seemingly innocuous change to the training query format gives rise to such egregious behavior. The results are summarized in Figure 3. Before discussing them in detail, we first describe the datasets, the reward model architecture, and the training algorithm.

7.1. Preference Datasets

For simplicity, we consider only a single prompt. While other prompts lead to similar results, we arbitrarily choose to present results produced by the following prompt, based on a popular movie last year:

prompt: *Did Oppenheimer win a Nobel Prize?*

We sample candidate responses using GPT-3.5 and GPT-4 with the temperature parameter of each language model set to the default value of 1. Here are representative responses:

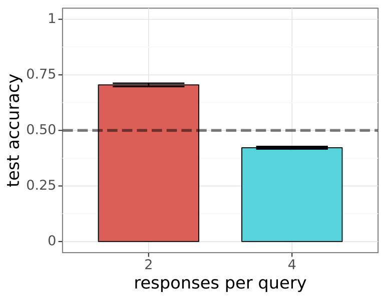
GPT-3.5 response: *No, Oppenheimer did not win the Nobel Prize.*

GPT-4 response: *No, Robert Oppenheimer, often called the "father of the atomic bomb" for his role in the Manhattan Project, did not win a Nobel Prize.*

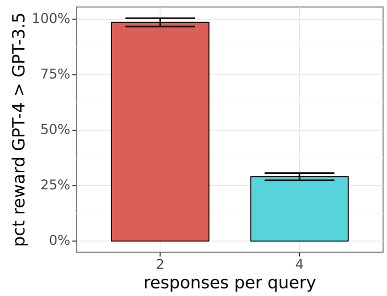
GPT-3.5 tends to produce concise responses relative to the more informative ones from GPT-4. From each of these two language models, we sample 100 responses for training and validation, and 100 responses for testing. We subsample from these responses when constructing each training or validation query, or each test sample. We produce two training sets that differ in the number of alternative responses per query. The first training set has 2 responses per query, one from GPT-3.5 and one from GPT-4. The second training set has 4 responses per query, one from GPT-3.5 and three from GPT-4. We generate 800 queries for each training set. Additionally, we generate 200 queries for each validation set. These are used in tuning hyperparameters to optimize validation accuracy. We produce a single test dataset made up of 1000 queries, each with one response from GPT-3.5 and one from GPT-4.

To simulate human annotator choices, we prompt GPT-4 to select its favorite response among a set. We use two types of prompts, which express preference for concise and informative responses respectively:

(i) **concise choice prompt**: *Suppose that you are looking for a concise answer to the following question. Which response below do you like the best...*



(a) Test accuracy of a standard reward model trained on preference data with 2 or 4 responses per query. The dashed line indicates a purely random baseline.



(b) Percentage of test data where the reward model assigns a higher score to GPT-4's response.

Figure 3: Standard reward model training can lead to egregious outcomes when training data involves more than two responses per query.

(ii) **informative choice prompt**: *Suppose that you are trying to learn more about the following question. Which response below do you like the best...*

We find that prompting for a concise choice typically favors GPT-3.5, while prompting for an informative choice typically favors GPT-4. When simulating a human annotator, we sample one of these two types of choice prompts according to probabilities 0.3 and 0.7, respectively. Using these choice prompts, we find that regardless of the number of responses per query, our simulated human annotators select responses from GPT-4 approximately 70% of the time.

7.2. Reward Model Architecture and Training

Our reward model builds on the PaLM 2 (Google et al., 2023) XXS language model. In particular, given a context formed by concatenating a prompt and a response, we take the final embedding produced by the base language model and apply a linear layer to obtain reward. We train the linear layer, as well as the base language model, to minimize cross-entropy loss on preference data. For each training dataset, we train the reward model over 150 gradient steps using the Adafactor optimizer (Shazeer and Stern, 2018) with a learning rate of 1e-4 and batch size of 16. The hyperparameters are tuned to optimize validation accuracy. We average results over 3 seeds.

7.3. Results

Figure 3a plots the test accuracy of the reward model trained on queries with 2 or 4 responses. When each training query includes only a pair of responses, the reward model correctly selects the preferred response for around 70% of test samples. However, with four responses per training query, the test accuracy drops below 50%. In other words, choice predictions generated by the reward model fare worse than random guessing. Figure 3b provides more insight into how learned reward models score responses. Since the majority of simulated annotators favor GPT-4, we would expect that the learned reward model assigns higher scores to GPT-4 responses. We see in Figure 3b that, when trained on pairwise comparisons, the reward model almost always assigns higher scores to GPT-4 responses. However, when trained on queries each with 4 responses, the reward model tends to erroneously assign higher scores to GPT-3.5 responses.

The culprit here is the IIA assumption made in reward learning, which is violated by the data as we will now explain. Intuitively, each simulated human annotator prefers either concise or informative responses but is relatively indifferent between alternatives in each of these two categories. Consider a simulated annotator who, when presented with one concise and one informative response, selects the former with probability $2/5$. Under the IIA assumption, increasing the number of informative alternatives from one to three would increase the probability of choosing the concise response to $2/3$. However, if the annotator is indifferent between the three informative responses, the probability of choosing the concise response should remain $2/5$, regardless of whether there are one or three informative alternatives.

8. Closing Remarks

It is worthwhile to think carefully about foundations of RLHF to inform algorithmic innovation and ultimately produce reliable AIs. What we have presented, which points out how the IIA assumption adopted by models underlying current algorithms can give rise to egregious behavior, offers one of what will hopefully be many steps in this direction.

The design of RLHF algorithms that mitigate perverse incentives we have identified remains a subject for future research. Trade-offs between query formats and how feedback is processed also deserve further study. For example, while RLPO can fail when queries present more than two alternatives, one can heuristically convert a feedback from such a query to multiple pairwise choices. This could offer more robust results, though there may be other flaws to this approach.

9. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. While there are many potential societal consequences of working towards this goal, we don't strongly feel that any of which must be specifically highlighted here.

References

Kenneth J Arrow. *Social choice and individual values*. John Wiley & Sons, New York, NY, 1951.

Dilip Arumugam, Shi Dong, and Benjamin Van Roy. Inclusive Artificial Intelligence. *arXiv preprint arXiv:2212.12633*, 2022.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Vincent Dumoulin, Daniel D. Johnson, Pablo Samuel Castro, Hugo Larochelle, and Yann Dauphin. A density estimation perspective on learning from pairwise human preferences, 2023.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Google, Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without RL. *arXiv preprint arXiv:2310.13639*, 2023.

Duncan R Luce. *Individual Choice Behavior*. John Wiley & Sons, New York, NY, 1959.

R. Duncan Luce and Patrick Suppes. Preferences, utility and subjective probability. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of Mathematical Psychology*, chapter 19, pages 249–410. John Wiley & Sons, New York, NY, 1965.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, chapter 4, pages 105–142. Academic Pres, New York, NY, 1974.

OpenAI. GPT-4 Technical Report. Technical report, OpenAI, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan

Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*, 2018.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 2020.

Wanqiao Xu, Shi Dong, Dilip Arumugam, and Benjamin Van Roy. Shattering the agent-environment interface for fine-tuning inclusive language models. *arXiv preprint arXiv:2305.11455*, 2023.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EdIGMCHk4l.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *International Conference on Learning Representations*, 2023.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *International Conference on Machine Learning*, 2023.

A. Proofs

A.1. RLPO

In this subsection, we provide the proof for Proposition 6.4. We start by defining two key quantities for messages of category 1 in $\mathcal{D}$. First, we let

$$\rho_{\text{chosen}} = \frac{1}{|\mathcal{D}|} \sum_{(\mathcal{Y},y)\in\mathcal{D}} \mathbb{1}\left(y \in \mathcal{M}_1\right)$$

denote the fraction of data in $\mathcal{D}$ where the chosen message is in $\mathcal{M}_1$. Second, we let

$$\rho_{\text{data}} = \frac{1}{|\mathcal{D}|} \sum_{(\mathcal{Y},y)\in\mathcal{D}} \frac{|\mathcal{Y} \cap \mathcal{M}_1|}{|\mathcal{Y}|}$$

denote the fraction of messages in $\mathcal{D}$ that belong to $\mathcal{M}_1$.

Lemma A.1. *For all strictly convex functions $f : \Re \to \Re$ such that $f'(1) < 0$, if $f(\zeta) = 0$ for some $\zeta \in \Re$, then $\zeta > 1 - \frac{f(1)}{f'(1)}$.*

Proof. Since f is strictly convex, $f(\zeta) > f(1) + f'(1)(\zeta - 1)$. Since $f'(1) < 0$, this implies $\zeta - 1 > -\frac{f(1)}{f'(1)}$. □

Lemma A.2. *Under Assumption 6.2, if $\rho_{\text{data}} > \rho_{\text{chosen}} > 0$, then the minimizer $\widehat{\psi}$ of $\mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ exists and satisfies*

$$\widehat{\psi}_2 - \widehat{\psi}_1 > \frac{\rho_{\text{data}} - \rho_{\text{chosen}}}{1 + \rho_{\text{data}} - \rho_{\text{chosen}}}.$$

Proof. Since $\mathcal{L}_{\text{reward}}$ is invariant under translations of ψ, without loss of generality, set $\psi_1 = 0$.

$$\mathcal{L}_{\text{reward}}(\widehat{r}_\psi) = -\sum_{(\mathcal{Y},y)\in\mathcal{D}} \ln \frac{e^{\widehat{r}_\psi(y)}}{\sum_{y'\in\mathcal{Y}} e^{\widehat{r}_\psi(y')}} = -\psi_2|\mathcal{D}|(1-\rho_{\text{chosen}}) + \sum_{(\mathcal{Y},y)\in\mathcal{D}} \ln\left(|\mathcal{Y}\cap\mathcal{M}_1| + |\mathcal{Y}\cap\mathcal{M}_2|e^{\psi_2}\right).$$

Note that this loss function is strictly convex and thus a local minima is also a global minima. Examining the derivative, we see that $\widehat{\psi}_2$, if exists, satisfies

$$\sum_{(\mathcal{Y},y)\in\mathcal{D}} \frac{|\mathcal{Y}\cap\mathcal{M}_1|}{|\mathcal{Y}\cap\mathcal{M}_1| + |\mathcal{Y}\cap\mathcal{M}_2|e^{\widehat{\psi}_2}} - |\mathcal{D}|\rho_{\text{chosen}} = 0.$$

With a change of variable, let $x = e^{\psi_2}$ and

$$f(x) = \sum_{(\mathcal{Y},y)\in\mathcal{D}} \frac{|\mathcal{Y}\cap\mathcal{M}_1|}{|\mathcal{Y}\cap\mathcal{M}_1| + |\mathcal{Y}\cap\mathcal{M}_2|x} - |\mathcal{D}|\rho_{\text{chosen}}.$$

Clearly, $f(x)$ is strictly decreasing on $\Re_+$ and $f(x) \to -|\mathcal{D}|\rho_{\text{chosen}}$ as $x \to \infty$. Since $\rho_{\text{chosen}} > 0$, there exists a unique $\zeta > 0$ such that $f(\zeta) = 0$. Thus, $\widehat{\psi}_2 = \ln\zeta$ exists. Further, we note that $f(x)$ is strictly convex and $f'(1) < 0$. By Lemma A.1,

$$\zeta > 1 - \frac{f(1)}{f'(1)} = 1 + |\mathcal{D}|(\rho_{\text{data}} - \rho_{\text{chosen}}) \cdot |\mathcal{Y}|^2 \left(\sum_{(\mathcal{Y},y)\in\mathcal{D}} |\mathcal{Y}\cap\mathcal{M}_1| \cdot |\mathcal{Y}\cap\mathcal{M}_2|\right)^{-1} > 1 + \rho_{\text{data}} - \rho_{\text{chosen}}. \tag{11}$$

The result follows from combining Equation (11) and the fact that $\ln(1+\alpha) \geq \frac{\alpha}{1+\alpha}$ for all $\alpha > -1$.

□

For all $\eta > 0$, define the high probability event

$$\mathcal{E}_\eta(\mathcal{D}) = \{\rho_{\text{data}} - \rho_{\text{chosen}} > \eta\} \bigcap \{\rho_{\text{chosen}} > 0\}. \tag{12}$$

Lemma A.3. *Under Assumptions 6.2 and 6.3, if there exists $\eta > 0$ such that $\mathbb{P}(\mathcal{E}_\eta(\mathcal{D})) \to 1$ as $|\mathcal{D}| \to \infty$, then as $|\mathcal{D}| \to \infty$,*

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \xrightarrow{p.} 0 \qquad \text{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \xrightarrow{p.} 1,$$

where $\widehat{\theta}$ minimizes $\mathcal{L}_{\text{policy}}(\widehat{\pi}_\theta|\widehat{r}_{\widehat{\psi}})$, and $\widehat{\psi} \in \arg\min_\psi \mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ if the loss $\mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ has a minimizer, and $\widehat{\psi} = 0$ otherwise.

Proof. By Lemma A.2, under event $\mathcal{E}_\eta(\mathcal{D})$, $\widehat{\psi} = \arg\min_\psi \mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ and $\widehat{\psi}_2 - \widehat{\psi}_1 > \frac{\rho_{\text{data}} - \rho_{\text{chosen}}}{1+\rho_{\text{data}} - \rho_{\text{chosen}}} > \frac{\eta}{1+\eta}$. When $\beta = 0$, $\widehat{\pi}_{\widehat{\theta}}$ maximizes reward. Thus, under event $\mathcal{E}_\eta(\mathcal{D})$, $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0$ and $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) = 1$. When $\beta > 0$, $\widehat{\pi}_{\widehat{\theta}}$ satisfies

$$P_{\widehat{\pi}_{\widehat{\theta}}}(y) = \frac{P_{\overline{\pi}}(y) e^{\widehat{r}_{\widehat{\psi}}(y)|\mathcal{D}|/\beta}}{\sum_{y' \in \mathcal{M}} P_{\overline{\pi}}(y') e^{\widehat{r}_{\widehat{\psi}}(y')|\mathcal{D}|/\beta}}.$$

Under event $\mathcal{E}_\eta(\mathcal{D})$,

$$\frac{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)}{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2)} = \frac{p_1}{p_2} \cdot e^{\left(\widehat{\psi}_1 - \widehat{\psi}_2\right)|\mathcal{D}|/\beta} < e^{-\frac{\eta}{1+\eta} \cdot \frac{|\mathcal{D}|}{\beta}}.$$

For all $\epsilon > 0$, there exists $N > 0$ such that $e^{-\frac{\eta}{1+\eta} \cdot \frac{|\mathcal{D}|}{\beta}} < \epsilon$ for all $|\mathcal{D}| > N$. In turn, for all $\beta \geq 0$ and $\epsilon > 0$, there exists $N > 0$ such that for all $|\mathcal{D}| > N$,

$$\mathbb{P}\left(\frac{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1)}{P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2)} \geq \epsilon\right) \leq \mathbb{P}\left(\mathcal{E}_\eta(\mathcal{D})^c\right).$$

The proof follows from the fact that $\mathbb{P}(\mathcal{E}_\eta(\mathcal{D})^c) \to 0$ as $|\mathcal{D}| \to \infty$ and $P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) + P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) = 1$.

□

Finally, we prove that under the dichotomy data assumption, there exists $\eta > 0$ such that the event $\mathcal{E}_\eta(\mathcal{D})$ holds with high probability. Define

$$\mathcal{I} = \left\{(\mathcal{Y}, y) \in \mathcal{D} \,|\, \mathcal{Y} \cap \mathcal{M}_1 \neq \emptyset,\, \mathcal{Y} \cap \mathcal{M}_2 \neq \emptyset\right\}$$

as the subset of $\mathcal{D}$ that contains messages in both categories. The first lemma below shows that most of the data contain both categories of messages with high probability.

Lemma A.4. *Under Assumption 6.1, if $P_{\overline{\pi}}(\mathcal{M}_1) \in (0, 1)$, $|\mathcal{Y}| \geq 2$, then there exists a constant $\gamma = \gamma(|\mathcal{Y}|) \in (0, 1)$ such that for all $|\mathcal{D}| > 0$, $\mathbb{P}(|\mathcal{I}| \leq |\mathcal{D}|(1-\gamma)) \leq e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}$.*

Proof. Under Assumption 6.1, for each $(\mathcal{Y}, y) \in \mathcal{D}$, the membership τ of each message $y' \in \mathcal{Y}$ follows $\text{Bernoulli}(P_{\overline{\pi}}(\mathcal{M}_1))$. Let $s = 1 - P_{\overline{\pi}}(\mathcal{M}_1)^{|\mathcal{Y}|} - P_{\overline{\pi}}(\mathcal{M}_2)^{|\mathcal{Y}|} < 1$. Then s equals the probability of sampling both categories of messages in each datum, and $|\mathcal{I}| \sim \text{Binomial}(|\mathcal{D}|, s)$. By Chernoff's bound,

$$\mathbb{P}\left(|\mathcal{I}| \leq s \cdot |\mathcal{D}|s\right) \leq e^{-\frac{|\mathcal{D}|s(1-s)^2}{2}}.$$

Letting $\gamma = 1 - s^2 \in (0, 1)$, we have $s(1-s)^2 > (1-\gamma)\gamma^2/4$. Therefore,

$$\mathbb{P}\left(|\mathcal{I}| \leq |\mathcal{D}|(1-\gamma)\right) \leq e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}. \tag{13}$$

□

Lemma A.5. *Under Assumption 6.1, for all $|\mathcal{Y}| \geq 3$, if $p_*(1) < F(P_{\overline{\pi}}(\mathcal{M}_1))$ with $F(\zeta) = \frac{\zeta - \zeta^{|\mathcal{Y}|}}{1 - \zeta^{|\mathcal{Y}|} - (1-\zeta)^{|\mathcal{Y}|}}$, then there exists $\eta > 0$ such that $\mathbb{P}(\mathcal{E}_\eta(\mathcal{D})) \to 1$ as $|\mathcal{D}| \to \infty$.*

Proof. Let $\delta = F(P_{\overline{\pi}}(\mathcal{M}_1)) - p_*(1) > 0$. Under Assumption 6.1, $|\mathcal{Y} \cap \mathcal{M}_1| \sim \text{Binomial}(|\mathcal{Y}|, P_{\overline{\pi}}(\mathcal{M}_1))$, giving

$$\mathbb{E}\left[\frac{|\mathcal{Y} \cap \mathcal{M}_1|}{|\mathcal{Y}|} \mid (\mathcal{Y}, y) \in \mathcal{I}\right] = \frac{P_{\overline{\pi}}(\mathcal{M}_1) - P_{\overline{\pi}}(\mathcal{M}_1)^{|\mathcal{Y}|}}{1 - P_{\overline{\pi}}(\mathcal{M}_1)^{|\mathcal{Y}|} - P_{\overline{\pi}}(\mathcal{M}_2)^{|\mathcal{Y}|}} = F(P_{\overline{\pi}}(\mathcal{M}_1)).$$

On the other hand,

$$\mathbb{E}\left[\mathbb{1}(y \in \mathcal{M}_1) \mid (\mathcal{Y}, y) \in \mathcal{I}\right] = p_*(1).$$

For $(\mathcal{Y}, y) \notin \mathcal{I}$, observe that $\frac{|\mathcal{Y} \cap \mathcal{M}_1|}{|\mathcal{Y}|} = \mathbb{1}(y \in \mathcal{M}_1)$. Thus

$$\frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y} \cap \mathcal{M}_1|}{|\mathcal{Y}|} - \frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_1) = \frac{|\mathcal{D}|}{|\mathcal{I}|}(\rho_{\text{data}} - \rho_{\text{chosen}}). \tag{14}$$

By Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y} \cap \mathcal{M}_1|}{|\mathcal{Y}|} - F(P_{\overline{\pi}}(\mathcal{M}_1)) \leq -\frac{\delta}{2} \mid \mathcal{I}\right) \leq e^{-\frac{|\mathcal{I}|\delta^2}{4}}$$

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_1) - p_*(1) \geq \frac{\delta}{4} \mid \mathcal{I}\right) \leq e^{-\frac{|\mathcal{I}|\delta^2}{8}}$$

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_1) - p_*(1) \leq -\frac{p_*(1)}{2} \mid \mathcal{I}\right) \leq e^{-\frac{|\mathcal{I}|p_*^2(1)}{2}}.$$

A union bound gives

$$\begin{aligned}
&\mathbb{P}\left(\left\{\frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y} \cap \mathcal{M}_1|}{|\mathcal{Y}|} - \frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_1) > \frac{\delta}{4}\right\} \bigcap \left\{\frac{1}{|\mathcal{I}|} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_1) > 0\right\} \mid \mathcal{I}\right) \\
&> 1 - e^{-\frac{|\mathcal{I}|\delta^2}{4}} - e^{-\frac{|\mathcal{I}|\delta^2}{8}} - e^{-\frac{|\mathcal{I}|p_*^2(1)}{2}}.
\end{aligned}$$

By Lemma A.4, there exists $0 < \gamma < 1$ such that $\mathbb{P}(|\mathcal{I}| \leq |\mathcal{D}|(1-\gamma)) \leq e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}$. By Equation (14),

$$\begin{aligned}
&\mathbb{P}\left(\left\{\frac{|\mathcal{D}|}{|\mathcal{I}|}(\rho_{\text{data}} - \rho_{\text{chosen}}) > \frac{\delta}{4}\right\} \bigcap \left\{\frac{|\mathcal{D}|}{|\mathcal{I}|}\rho_{\text{chosen}} > 0\right\} \bigcap \{|\mathcal{I}| > |\mathcal{D}|(1-\gamma))\}\right) \\
&= \mathbb{P}\left(\left\{\frac{|\mathcal{D}|}{|\mathcal{I}|}(\rho_{\text{data}} - \rho_{\text{chosen}}) > \frac{\delta}{4}\right\} \bigcap \left\{\frac{|\mathcal{D}|}{|\mathcal{I}|}\rho_{\text{chosen}} > 0\right\} \mid |\mathcal{I}| > |\mathcal{D}|(1-\gamma))\right) \mathbb{P}\left(|\mathcal{I}| > |\mathcal{D}|(1-\gamma))\right) \\
&> \left(1 - e^{-\frac{|\mathcal{D}|(1-\gamma)\delta^2}{4}} - e^{-\frac{|\mathcal{D}|(1-\gamma)\delta^2}{8}} - e^{-\frac{|\mathcal{D}|(1-\gamma)p_*^2(1)}{2}}\right)\left(1 - e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}\right) \\
&> 1 - e^{-\frac{|\mathcal{D}|(1-\gamma)\delta^2}{4}} - e^{-\frac{|\mathcal{D}|(1-\gamma)\delta^2}{8}} - e^{-\frac{|\mathcal{D}|(1-\gamma)p_*^2(1)}{2}} - e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}.
\end{aligned}$$

Let $\eta = \frac{(1-\gamma)\delta}{4} > 0$. It follows that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_\eta(\mathcal{D})\right) &= \mathbb{P}\left(\{\rho_{\text{data}} - \rho_{\text{chosen}} > \eta\} \bigcap \{\rho_{\text{chosen}} > 0\}\right) \\
&\geq \mathbb{P}\left(\left\{\frac{|\mathcal{D}|}{|\mathcal{I}|}(\rho_{\text{data}} - \rho_{\text{chosen}}) > \frac{\delta}{4}\right\} \bigcap \left\{\frac{|\mathcal{D}|}{|\mathcal{I}|}\rho_{\text{chosen}} > 0\right\} \bigcap \{|\mathcal{I}| > |\mathcal{D}|(1-\gamma))\}\right) \\
&> 1 - e^{-\frac{|\mathcal{D}|(1-\gamma)\delta^2}{4}} - e^{-\frac{|\mathcal{D}|(1-\gamma)\delta^2}{8}} - e^{-\frac{|\mathcal{D}|(1-\gamma)p_*^2(1)}{2}} - e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}.
\end{aligned}$$

The proof follows by taking $|\mathcal{D}| \to \infty$.

□

We are now ready to prove the following failure case for RLPO.

Proposition 6.4 (RLPO failure)**.** *Under Assumptions 6.1, 6.2, and 6.3, for all* $|\mathcal{Y}| \geq 3$*, if* $p_*(1) < F(P_{\overline{\pi}}(\mathcal{M}_1))$ *with* $F(\zeta) = \frac{\zeta - \zeta^{|\mathcal{Y}|}}{1-\zeta^{|\mathcal{Y}|}-(1-\zeta)^{|\mathcal{Y}|}}$*, then, as* $|\mathcal{D}| \to \infty$*,*

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \xrightarrow{p.} 0 \qquad \textit{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \xrightarrow{p.} 1,$$

where $\widehat{\theta}$ *minimizes* $\mathcal{L}_{\text{policy}}(\widehat{\pi}_\theta|\widehat{r}_{\widehat{\psi}})$*, and* $\widehat{\psi} \in \arg\min_\psi \mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ *if the loss* $\mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ *has a minimizer, and* $\widehat{\psi} = 0$ *otherwise.*

Proof. The proof follows from applying Lemmas A.3 and A.5. □

A.2. DPO

In Lemma A.6, we show that DPO and RLPO produces the same policy, which then allows us to directly apply the analysis for RLPO to reach the same conclusion for DPO.

Lemma A.6. *If there exists a minimizer* $\widehat{r}$ *to* $\mathcal{L}_{\text{reward}}$*, then the optimal policy for* $\mathcal{L}_{\text{DPO}}$ *and for* $\mathcal{L}_{\text{policy}}(\cdot|\widehat{r})$ *are identical.*

Proof. Recall that for any given reward function r, the optimal policy π_r for $\mathcal{L}_{\text{policy}}$ satisfies

$$P_{\pi_r}(x) = \frac{P_{\overline{\pi}}(x) e^{r(x)|\mathcal{D}|/\beta}}{\sum_{x' \in \mathcal{M}} P_{\overline{\pi}}(x') e^{r(x')|\mathcal{D}|/\beta}}. \tag{15}$$

It follows that

$$-\sum_{(\mathcal{Y},y)\in\mathcal{D}} \left[\ln\left(\frac{e^{r(y)}}{\sum_{y'\in\mathcal{Y}} e^{r(y')}}\right)\right] = -\sum_{(\mathcal{Y},y)\in\mathcal{D}} \left[\ln \frac{(P_{\pi_r}(y)/P_{\overline{\pi}}(y))^{\beta/|\mathcal{D}|}}{\sum_{y'\in\mathcal{Y}} (P_{\pi_r}(y')/P_{\overline{\pi}}(y'))^{\beta/|\mathcal{D}|}}\right]. \tag{16}$$

Thus the value $\mathcal{L}_{\text{reward}}(r)$ equals the value $\mathcal{L}_{\text{DPO}}(\pi_r)$. Note also that $r \mapsto \pi_r$ is surjective.

Suppose $\widehat{r}$ is the minimizer for $\mathcal{L}_{\text{reward}}$. Then $\pi_{\widehat{r}}$ is the optimal policy obtained by minimizing $\mathcal{L}_{\text{policy}}(\cdot|\widehat{r})$. If $\pi_{\widehat{r}}$ is not optimal for $\mathcal{L}_{\text{DPO}}$, then there exists another policy π' that obtains a strictly lower loss $\mathcal{L}_{\text{DPO}}(\pi')$. Since $r \mapsto \pi_r$ is surjective, there exists a reward function r' such that $\pi' = \pi_{r'}$. For example, we can take $r'(x) = \frac{\beta}{|\mathcal{D}|} \ln \frac{P_{\pi'}(x)}{P_{\overline{\pi}}(x)}$, which achieves a lower reward loss $\mathcal{L}_{\text{reward}}(r') = \mathcal{L}_{\text{DPO}}(\pi_{r'}) < \mathcal{L}_{\text{DPO}}(\pi_{\widehat{r}}) = \mathcal{L}_{\text{reward}}(\widehat{r})$, a contradiction.

Similarly, if π^* is optimal for $\mathcal{L}_{\text{DPO}}$, the corresponding reward $r^*(x) = \frac{\beta}{|\mathcal{D}|} \ln \frac{P_{\pi^*}(x)}{P_{\overline{\pi}}(x)}$ must be optimal for $\mathcal{L}_{\text{reward}}$. The corresponding optimal policy π_{r^*} for $\mathcal{L}_{\text{policy}}(\cdot|r^*)$ then satisfies

$$P_{\pi_{r^*}}(x) \propto P_{\overline{\pi}}(x) e^{\frac{\beta}{|\mathcal{D}|} \ln \frac{P_{\pi^*}(x)}{P_{\overline{\pi}}(x)} |\mathcal{D}|/\beta} = P_{\pi^*}(x),$$

as desired.

□

Proposition 6.5 (DPO failure)**.** *Under Assumptions 6.1, 6.2, and 6.3, for all* $|\mathcal{Y}| \geq 3$*, if* $p_*(1) < F(P_{\overline{\pi}}(\mathcal{M}_1))$ *with* $F(\zeta) = \frac{\zeta - \zeta^{|\mathcal{Y}|}}{1-\zeta^{|\mathcal{Y}|}-(1-\zeta)^{|\mathcal{Y}|}}$*, then as* $|\mathcal{D}| \to \infty$*,*

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \xrightarrow{p.} 0 \qquad \textit{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \xrightarrow{p.} 1,$$

where $\widehat{\theta} \in \arg\min_\theta \mathcal{L}_{\text{DPO}}(\widehat{\pi}_\theta)$ *if the loss* $\mathcal{L}_{\text{DPO}}(\widehat{\pi}_\theta)$ *has a minimizer, and* $\widehat{\theta} = 0$ *otherwise.*

Proof. By Lemma A.6, if $\widehat{\psi}$ is as defined in Proposition 6.4, then $\widehat{\psi} = \arg\min_\psi \mathcal{L}_{\text{reward}}(\widehat{r}_\psi)$ and we have $\widehat{\theta} = \arg\min_\theta \mathcal{L}_{\text{DPO}}(\widehat{\pi}_\theta) = \arg\min_\theta \mathcal{L}_{\text{policy}}(\widehat{\pi}_\theta|\widehat{r}_{\widehat{\psi}})$. The proof follows from Proposition 6.4. □

A.3. Inclusive Learning

In this subsection, we prove Proposition 6.6. We start by defining two key sets that describe messages in $\mathcal{D}$. Recall that in the proof for Proposition 6.4, we defined

$$\mathcal{I} = \big\{(\mathcal{Y}, y) \in \mathcal{D} \,|\, \mathcal{Y} \cap \mathcal{M}_1 \neq \emptyset,\, \mathcal{Y} \cap \mathcal{M}_2 \neq \emptyset\big\}.$$

To facilitate the exposition, we also define

$$\mathcal{I}_\tau = \{(\mathcal{Y}, y) \in \mathcal{I} \mid y \in \mathcal{M}_\tau\} \quad \text{for } \tau = 1, 2$$

to be the set of data in $\mathcal{I}$ where a message of category τ is chosen. We first prove the following lemma that characterizes the optimal solution for a class of convex loss functions.

Lemma A.7. *Suppose that for all $m > 0$, the loss $\mathcal{L}(x; m) : (0, 1) \to \Re$ is convex in x, $\mathcal{L}(x_m; m) = \min_x \mathcal{L}(x; m)$, and there exists $M > 0$ such that for all $m > M$,*

$$\frac{\partial \mathcal{L}(x; m)}{\partial x} < 0.$$

Then $\lim_{m\to\infty} x_m = 1$.

Proof. Since $\mathcal{L}(x; m)$ is convex on $(0, 1)$, its local minima is also a global minima. For all $m > M$, $\mathcal{L}$ is strictly decreasing for $x \in (0, 1)$, so $\lim_{m\to\infty} x_m = 1$. □

Then, we define the high probability event

$$\mathcal{E}(\mathcal{D}) = \{|\mathcal{I}_1| > 1\} \cap \left\{ |\mathcal{I}_2| > \max\{1 + \beta, 2\beta \ln \frac{p_*(2)}{p_*(1)}\} \right\}.$$

Lemma A.8. *Under Assumption 6.3, for all fixed $\mathcal{M}_1$ and $\mathcal{D}$ such that $\mathcal{E}(\mathcal{D})$ holds, if $\widehat{\theta}$ minimizes $\mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta)$, then*

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \to 0 \qquad \text{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \to 1$$

as the size of $\mathcal{M}_2$ grows.

Proof. First, we notice that $\mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta)$ is strongly convex for $\beta > 0$ and $\mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta) \to \infty$ as $\|\theta\|_2 \to \infty$, thus a minimizer for $\mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta)$ exists and is unique. Let $m = \frac{|\mathcal{M}_2|}{|\mathcal{M}_1|}$ and $q = P_{\widehat{\pi}_\theta}(\mathcal{M}_2)$. The loss function can be written in terms of m and q as

$$\mathcal{L}(q, m) = \mathcal{L}_{\mathrm{IL}}(\widehat{\pi}_\theta) = -\sum_{(\mathcal{Y},y)\in\mathcal{D}} \ln \frac{m(1-q)\cdot \mathbb{1}_{\mathcal{M}_1}(y) + q \cdot \mathbb{1}_{\mathcal{M}_2}(y)}{m(1-q)\cdot|\mathcal{Y}\cap\mathcal{M}_1| + q\cdot|\mathcal{Y}\cap\mathcal{M}_2|} + \beta\left[(1-q)\ln\frac{1-q}{p_*(1)} + q\ln\frac{q}{p_*(2)}\right]. \tag{17}$$

Taking the partial derivative with respect to q, under the event $\mathcal{E}(\mathcal{D})$,

$$\begin{aligned}
\frac{\partial \mathcal{L}(q, m)}{\partial q} &= \sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y}\cap\mathcal{M}_2|}{m(1-q)^2|\mathcal{Y}\cap\mathcal{M}_1| + q(1-q)|\mathcal{Y}\cap\mathcal{M}_2|} - \sum_{(\mathcal{Y},y)\in\mathcal{I}_2}\left(\frac{1}{q} + \frac{1}{1-q}\right) + \beta\left(\ln\frac{q}{1-q} + \ln\frac{p_*(1)}{p_*(2)}\right) \\
&\overset{(a)}{<} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y}\cap\mathcal{M}_2|}{m(1-q)^2|\mathcal{Y}\cap\mathcal{M}_1| + q(1-q)|\mathcal{Y}\cap\mathcal{M}_2|} - \frac{|\mathcal{I}_2|}{q} - \frac{|\mathcal{I}_2|}{1-q} + \frac{\beta}{1-q} + \frac{|\mathcal{I}_2|}{2} \\
&\overset{(b)}{<} \sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y}\cap\mathcal{M}_2|}{m(1-q)^2|\mathcal{Y}\cap\mathcal{M}_1| + q(1-q)|\mathcal{Y}\cap\mathcal{M}_2|} - \frac{|\mathcal{I}_2|}{2q} - \frac{1}{1-q},
\end{aligned}$$

where (a) follows from $|\mathcal{I}_2| > 2\beta \ln \frac{p_*(2)}{p_*(1)}$ and $\ln \frac{q}{1-q} < \frac{1}{1-q}$, and (b) follows from $|\mathcal{I}_2| > 1 + \beta$ and $q < 1$. For all $0 < q < 1$, the first term is always positive and converges to 0 as $m \to \infty$. Thus, there exists an $M > 0$ such that for all $m > M$, $\sum_{(\mathcal{Y},y)\in\mathcal{I}} \frac{|\mathcal{Y}\cap\mathcal{M}_2|}{m(1-q)^2|\mathcal{Y}\cap\mathcal{M}_1| + q(1-q)|\mathcal{Y}\cap\mathcal{M}_2|} < \frac{|\mathcal{I}_2|}{2q}$. This implies that

$$\frac{\partial \mathcal{L}(q, m)}{\partial q} < -\frac{1}{1-q} < 0.$$

The proof follows from applying Lemma A.7.

□

Finally, we prove that under the dichotomy data assumption, the event $\mathcal{E}(\mathcal{D})$ holds with high probability.

Lemma A.9. *Under Assumption 6.1, if $|\mathcal{Y}| \geq 2$, then for fixed $\mathcal{M}_1$, $\mathbb{P}(\mathcal{E}(\mathcal{D})) \to 1$ as $|\mathcal{M}_2| \to \infty$.*

Proof. By Lemma A.4, there exists a constant $\gamma \in (0,1)$ such that for all $|\mathcal{D}| > 0$, $\mathbb{P}(|\mathcal{I}| \leq |\mathcal{D}|(1-\gamma)) \leq e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}}$. Let $\epsilon(\beta) = \max\{\beta + 1, 2\beta \ln \frac{p_*(1)}{p_*(2)}\}$, then $\mathbb{E}[\mathbb{1}(y \in \mathcal{M}_2) \mid (\mathcal{Y}, y) \in \mathcal{I}] = p_*(2)$. By Hoeffding's inequality,

$$\mathbb{P}\left(\sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_2) - |\mathcal{I}| \cdot p_*(2) \leq \epsilon(\beta) - |\mathcal{I}| \cdot p_*(2) \mid \mathcal{I}\right) \leq e^{-2\frac{(\epsilon(\beta)-|\mathcal{I}|p_*(2))^2}{|\mathcal{I}|}} < e^{-2|\mathcal{I}|p_*^2(2)+4\epsilon(\beta)p_*(2)}$$

$$\mathbb{P}\left(\sum_{(\mathcal{Y},y)\in\mathcal{I}} \mathbb{1}(y \in \mathcal{M}_1) - |\mathcal{I}| \cdot p_*(1) \leq 1 - |\mathcal{I}| \cdot p_*(1) \mid \mathcal{I}\right) \leq e^{-2\frac{(1-|\mathcal{I}|p_*(2))^2}{|\mathcal{I}|}} < e^{-2|\mathcal{I}|p_*^2(2)+4p_*(2)}.$$

Following a similar argument as that in the proof for Lemma A.5, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}(\mathcal{D})) &\geq \mathbb{P}\left(\{|\mathcal{I}| > |\mathcal{D}|(1-\gamma)\} \cap \{|\mathcal{I}_1| > 1\} \cap \{|\mathcal{I}_2| > \epsilon(\beta)\}\right) \\ &\geq 1 - e^{-\frac{|\mathcal{D}|(1-\gamma)\gamma^2}{8}} - e^{-2|\mathcal{D}|(1-\gamma)p_*^2(2)+4\epsilon(\beta)p_*(2)} - e^{-2|\mathcal{D}|(1-\gamma)p_*^2(2)+4p_*(2)}.\end{aligned}$$

The proof follows by taking $|\mathcal{M}_2| \to \infty$.

□

Finally, we prove the following failure case for inclusive learning.

Proposition 6.6 (IL failure)**.** *Under Assumptions 6.1 and 6.3, if $|\mathcal{Y}| \geq 2$ and $\widehat{\theta}$ minimizes $\mathcal{L}_{\text{IL}}(\widehat{\pi}_\theta)$, then for fixed $\mathcal{M}_1$,*

$$\lim_{|\mathcal{D}|\to\infty} \mathbb{P}\left(\lim_{|\mathcal{M}_2|\to\infty} P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0\right) = 1. \tag{8}$$

Proof. The proof follows from applying Lemmas A.8 and A.9, and noticing that $\mathbb{P}(\lim_{|\mathcal{M}_2|\to\infty} P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0) \geq \mathbb{P}(\mathcal{E}(\mathcal{D}))$. □

A.4. SLiC

The proof for Proposition 6.7 is similar to that for Proposition 6.6. We present the proof here.

Lemma A.10. *Under Assumption 6.3, if $|\mathcal{Y}| = 2$, for all fixed $\mathcal{M}_1$ and $\mathcal{D}$ such that $\mathcal{E}(\mathcal{D})$ holds, if $\widehat{\theta}$ minimizes $\mathcal{L}_{\text{SLiC}}(\widehat{\pi}_\theta)$, then*

$$P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) \to 0 \qquad \text{and} \qquad P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_2) \to 1$$

as the size of $\mathcal{M}_2$ grows.

Proof. First, we notice that $\mathcal{L}_{\text{SLiC}}(\widehat{\pi}_\theta)$ is strongly convex for $\beta > 0$ and $\mathcal{L}_{\text{SLiC}}(\widehat{\pi}_\theta) \to \infty$ as $\|\theta\|_2 \to \infty$, thus a minimizer for $\mathcal{L}_{\text{SLiC}}(\widehat{\pi}_\theta)$ exists and is unique. Let $m = \frac{|\mathcal{M}_2|}{|\mathcal{M}_1|}$ and $q = P_{\widehat{\pi}_\theta}(\mathcal{M}_2)$. When $|\mathcal{Y}| = 2$, the loss function can be written in terms of m and q as

$$\begin{aligned}\mathcal{L}(q,m) = \mathcal{L}_{\text{SLiC}}(\widehat{\pi}_\theta) = &\sum_{(\mathcal{Y},y)\in\mathcal{D}} \left[\ln\left(\frac{m(1-q)}{q} \mathbb{1}_{\mathcal{M}_1}(y) \cdot \mathbb{1}_{\mathcal{M}_2}(\mathcal{Y} \setminus \{y\}) + \frac{q}{m(1-q)} \mathbb{1}_{\mathcal{M}_2}(y) \cdot \mathbb{1}_{\mathcal{M}_1}(\mathcal{Y} \setminus \{y\})\right) + \delta\right]_+ \\ &+ \beta\left[(1-q)\ln\frac{1-q}{p_*(1)} + q \ln\frac{q}{p_*(2)}\right].\end{aligned}$$

Let $h(q,m) = \ln \frac{m(1-q)\cdot\mathbb{1}_{\mathcal{M}_1}(y) + q\cdot\mathbb{1}_{\mathcal{M}_2}(y)}{q\cdot\mathbb{1}_{\mathcal{M}_2}(\mathcal{Y}\setminus\{y\}) + m(1-q)\cdot\mathbb{1}_{\mathcal{M}_1}(\mathcal{Y}\setminus\{y\})} + \delta$. We have that

$$\frac{\partial h(q,m)}{\partial q} = -\frac{m}{q^2}\mathbb{1}_{\mathcal{M}_1}(y) \cdot \mathbb{1}_{\mathcal{M}_2}(\mathcal{Y} \setminus \{y\}) + \frac{1}{m(1-q)^2}\mathbb{1}_{\mathcal{M}_2}(y) \cdot \mathbb{1}_{\mathcal{M}_1}(\mathcal{Y} \setminus \{y\}).$$

For any fixed $q \in (0,1)$, when m is large enough, $\frac{m(1-q)}{q} > e^{-\delta}$ and $\frac{q}{m(1-q)} < e^{-\delta}$. Consequently, when m is large enough,

$$\frac{\partial \mathcal{L}(q,m)}{\partial q} = -\sum_{(\mathcal{Y},y)\in\mathcal{I}_1} \frac{m}{q^2} + \beta\left((1-q)\ln\frac{1-q}{p_*(1)} + q\ln\frac{q}{p_*(2)}\right) < 0.$$

The proof follows by applying Lemma A.7. □

With this, we are ready to prove the failure case for SLiC-HF.

Proposition 6.7 (SLiC failure)**.** *Under Assumptions 6.1 and 6.3, if $|\mathcal{Y}| = 2$ and $\widehat{\theta}$ minimizes $\mathcal{L}_{\mathrm{SLiC}}(\widehat{\pi}_\theta)$, for fixed $\mathcal{M}_1$,*

$$\lim_{|\mathcal{D}|\to\infty} \mathbb{P}\left(\lim_{|\mathcal{M}_2|\to\infty} P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0\right) = 1. \tag{10}$$

Proof. The proof follows from applying Lemmas A.10 and A.9, and noticing that $\mathbb{P}(\lim_{|\mathcal{M}_2|\to\infty} P_{\widehat{\pi}_{\widehat{\theta}}}(\mathcal{M}_1) = 0) \geq \mathbb{P}(\mathcal{E}(\mathcal{D}))$. □