RLHF/DPO 小話（その３）

和地 瞭良 / *Akifumi Wachi*

April 25, 2024

RLHF [1]や DPO [2]に関して論文 [Wachi et al., 2024] を最近執筆した。やはり自分で研究すると理解がかなり深まるわけで、巷ではあまり議論されていないことが色々と分かった。なので、すでに公開済みの内容の中から話せる範囲で情報共有して、誰かの役に立てばいいな、と思っている次第である。やる気が持続すればその 4 くらいまでいく予定（多分力尽きる）。

[1] RLHF = Reinforcement Learning from Human Feedback [Ouyang et al., 2022]

[2] DPO = Direct Preference Optimization [Rafailov et al., 2024]

これまでのコラム（その１-２）

- `https://akifumi-wachi-4.github.io/website/jp.html`

言語モデルのアライメント $\neq$ 「調整」

その 1 とその 2 で再三述べてきたとおりだが、これまで紹介してきた手法（RLHF, DPO, KTO など）は、すべて

$$\max_{\pi_\theta} \quad \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[\, r(x,y)\,\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\, \pi_\theta(y \mid x) \,\|\, \pi_{\mathrm{ref}}(y \mid x)\,\right], \tag{1}$$

という問題を解いており、この問題の解析解は

$$\pi_r^\star(y \mid x) = \frac{1}{Z_r(x;\pi_{\mathrm{ref}})}\pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right), \tag{2}$$

という形で書ける。

改めてこの式を見て、どのような印象を持つだろうか?RLHF/DPO/KTO いずれの手法でも、β は大抵の場合 0.1 前後の値が選ばれる。Z_r は単なる正規化のための分配関数 Z_r を一旦無視して $\beta = 0.1$ を代入すると

$$\pi_r^\star(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x)\exp(10 \cdot r(x,y)) \tag{3}$$

を得る。リファレンス方策 $\pi(y \mid x)$ に対し $\exp(10 \cdot r(x,y))$ というなんとも恐ろしい関数を掛けていることに気づくだろうか？ [3]辞書を引くと、アライメント（Alignment） は日本語で「調整」と訳されることが多いが、この式を見るとそんな優しい言葉では表現できないほどの強制力を感じずにはいられない。このような強制力をもってアライメントをするからこそ、ChatGPT をはじめとする RLHF/DPO が施された大規模言語モデルは人間と見違えるほどの対話能力を持つわけである。しかし、こんな過激な操作をしてなにか弊害はないのだろうか？

[3] 我々の論文では安全性に関してアライメントしているが、さらに小さな $\beta \sim 0.01$ を用いると良い、という実験結果が得られている。なんと $\exp(100 \cdot r(x,y))$。

RLHF/DPO の弊害

最近、RLHF や DPO が施された大規模言語モデルには色々な欠点があるらしい、ということが報告されている。代表的な例の一つが

- 出力される文章の多様性が損なわれる [4]

である。たとえば、Kirk et al. [2023] では、SFT のみが施された言語モデルに比較して、RLHF が施された言語モデルの出力の多様性が大幅に減少することが実験的に示されている（Figure 1参照）。他にも、

- 学習データのバイアスが RLHF/DPO によって助長される

などの報告が多数報告されている。バイアスと一口に言ってもさまざまで、政治的志向 [Liu et al., 2021] やジェンダー [Sun et al., 2019] などのいわゆる「バイアス」はもちろんこと、言語モデルで特に顕著なバイアスなど多岐にわたる。せっかくなので、面白いバイアスを 2 つ紹介しよう。1 つ目が「冗長性バイアス（verbosity bias）」 [Singhal et al., 2023, Saito et al., 2023] である。これは、言語モデルが**「長い回答をより好む」**ということであり、その傾向は人間と比較してより強い、ということが実験的に分かっている [Saito et al., 2023]。たしかに ChatGPT は「必要以上にペラペラ喋るなぁ」という印象を持っている人は多いのではないだろうか？この現象、Preference data 自体が、長い回答を好むように人間がラベル付けされて作成されているからでは、というような仮説が立てられていたりする [Zheng et al., 2023]。また、自己強化バイアス（self-enhancement bias） [Zheng et al., 2023] もなかなか面白い。言語モデルは、**自分自身の回答を高く評価してしまう**、というものである。GPT-4 は 10%高い勝率で、Claude-v1 は 25%高い勝率で自分自身を好む [Zheng et al., 2023]、という実験的な結果が得られている。言語モデルの特性上、自分が高く評価する文章を生成するわけで、至極当然の現象と言えるが、**「ある言語モデル A が生成した文章を、言語モデル A 自身ないしはそれの派生モデルで評価するのはやめましょうね」**というキーメッセージはちゃんと理解しておいたほうがいいだろう。

上記で紹介した実験的結果は、あくまで状況証拠を集めただけであり、その真偽ははまだ定まっていない。しかし、(3) を眺めるとなんかそういう現象が起こりそうだなぁ、というのを直感的に掴んでいただけると嬉しい。$\exp\left(10 \cdot r(x,y)\right)$ で方策の確率分布を強烈に変形させたら、多様性は減りそうな気がしてくるし、自分の回答を高く評価しやすくなるのも、なんとなくそんな気がしてこないだろうか？実はバイアスに関しては、非常に簡単な式変形 1 回で、バイアスが助長されるのを説明することができる。それについてはその 4 のいちばん大事なアイデアにつながるので、ここではあえて秘密にしておく。

[4] 人間が (3) で服従させにきたのに、「今度は多様性がない！」って LLM くんも辛いですな …

exp はどこから来たのか？

式 (2) と (3) において $\exp\left(\frac{1}{\beta} \cdot r(x,y)\right)$ が弊害を生んでいるとして、なぜ exp は登場しているのだろうか？ [5]

式 (1) から (2) をどのように導出したかを復習しよう。以降は、その 1 に記載した証明から抜粋するので、証明全体を追いたい方はその 1 を参照されたい。いま、(1) に関し、Reverse KL divergence の定義から、以

[5] β を大きくすればよいのでは？という読者もいるかと思うが、実用上は β を大きくするとうまくいかないことが大半である。

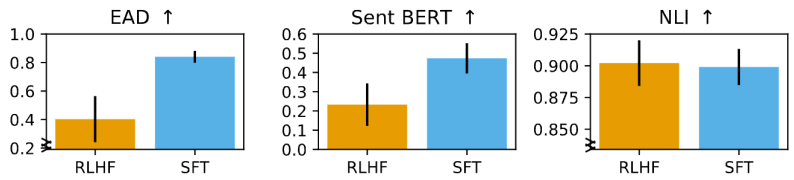

Figure 5: **Per-input diversity metrics for RLHF and SFT models**. For these scores the outputs used to calculate the diversity are a sample of outputs from the model for single input. These per-input scores are then averaged, as in Eq. (2). Error bars are standard deviation of the per-input diversity score across different inputs. Note that some plots have broken y-axis for better visualisation.

Figure 1: 図はKirk et al. [2023] から拝借。EAD, Sent BERT, NLO は多様性の指標であり、値が大きいほど多様であることを示す。それぞれ
EAD = Expectation-adjusted distinct N-grams
Sent BERT = Sentence-BERT embedding cosine similarity
NLI = Natural language inference diversity

下のような関係が成り立つ。

$$\begin{aligned}
&\max_{\pi} \mathbb{E}_{\rho,\pi}\left[r(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\,\pi(y \mid x) \,\|\, \pi_{\mathrm{ref}}(y \mid x)\,\right] \\
&\quad = \max_{\pi} \mathbb{E}_{\rho,\pi}\left[r(x,y) - \beta \log \frac{\pi(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}\right] \\
&\quad = \min_{\pi} \mathbb{E}_{\rho,\pi}\left[\log \frac{\pi(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} - \frac{1}{\beta} r(x,y)\right] \\
&\quad = \min_{\pi} \mathbb{E}_{\rho,\pi}\left[\log \frac{\pi(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta} r(x,y)\right)}\right]. \qquad (4)
\end{aligned}$$

(4) の最後によく見た $\exp\left(\frac{1}{\beta}\cdot r(x,y)\right)$ の形が登場している。逆を追っていくと、exp が登場しているのは、log が存在していたからで、それは Reverse KL divergence $\mathbb{D}_{\mathrm{KL}}$ の定義が log を用いているから、であることが分かる。言い換えるならば、**Reverse KL divergence を他のものに変えれば、exp を登場させずに済む**、ということになる。

他の *divergence* を使うことはできるのか？

では、Reverse KL divergence 以外の divergence を使うことは出来るのか？結論から言えば可能である。

Wang et al. [2024] は、Reverse KL divergence を、f-divergence に拡張した手法を提案している。具体的には、$\mathbb{D}_{\mathrm{KL}}$ を 任意の f-divergence $\mathbb{D}_f$ に置き換えて、

$$\max_{\pi_\theta} \quad \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[\,r(x,y)\,\right] - \beta\mathbb{D}_f\left[\,\pi_\theta(y \mid x) \,\|\, \pi_{\mathrm{ref}}(y \mid x)\,\right] \qquad (5)$$

のような問題を解くことを目指す。なお、$f(1) = 0$ を満たし $\mathrm{x} = 1$ [6]の周辺で狭義凸なる任意の凸関数 $f : \mathbb{R}_+ \to \mathbb{R}$ に対し、f-divergence は以下のように定義される。

$$\mathbb{D}_f\left[\,p \,\|\, q\,\right] := \mathbb{E}_{q(x)}\left[f\left(\frac{p(x)}{q(x)}\right)\right]. \qquad (6)$$

ただし、p と q は任意の確率分布である。f-divergence の例とそれに対応する関数 f としては Figure 2 にまとめたので参照していただきたい。

[6] この x はプロンプトの x とは異なることに注意。

f-divergence	$f(u)$	$f'(u)$	$0 \notin$ Domain of $f'(u)$
α-divergence ($\alpha \in (0,1)$)	$(u^{1-\alpha} - (1-\alpha)u - \alpha)/(\alpha(\alpha-1))$	$(1-u^{-\alpha})/\alpha$	✓
Reverse KL ($\alpha = 0$)	$u \log u$	$\log u + 1$	✓
Forward KL ($\alpha = 1$)	$-\log u$	$-1/u$	✓
JS-divergence	$u \log u - (u+1)\log((u+1)/2)$	$\log(2u/(1+u))$	✓
Total Variation	$\frac{1}{2}\lvert u-1\rvert$	$u > 1 \,?\, \frac{1}{2} : -\frac{1}{2}$	✗
Chi-squared	$(u-1)^2$	$2(u-1)$	✗

Figure 2: 図はWang et al. [2024] から拝借。

ここで (6) を (7) に代入すると、以下の式を得る。

$$\max_{\pi_\theta} \quad \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right] - \beta\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}\left[f\left(\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\right)\right] \tag{7}$$

導出が結構複雑なので気になる方は元論文を追っていただきたいのだが、DPO のときと同様に、(7) の最適解は解析的に得られて、

$$\pi_r^\star(y \mid x) = \frac{1}{\widehat{Z}_r(x; \pi_{\text{ref}})}\pi_{\text{ref}}(y \mid x)(f')^{-1}\left(\frac{1}{\beta}r(x,y)\right), \tag{8}$$

のようになる。$\widehat{Z}_r$ は Z_r とは微妙に形は異なるものの依然として正規化のための分配関数である。また、$(f')^{-1}$ は、関数 f の一階微分の逆関数であることに留意されたい。

具体的な例として Reverse KL と Forward KL を見ていくと、π と π_{ref} について以下のような関係式が得られる。なお、簡単のため定数項は省略していることに留意されたい。

- Reverse KL divergence

$$\pi_r^\star(y \mid x) \propto \pi_{\text{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

- Forward KL divergence

$$\pi_r^\star(y \mid x) \propto -\pi_{\text{ref}}(y \mid x)\left(\frac{1}{\beta}r(x,y)\right)^{-1}$$

そうすると、Forward KL については Reverse KL と比較すると、穏やかな操作をしていることが分かるだろう。事実、Wang et al. [2024] において、Forward KL を用いたほうが、Reverse KL を用いたときよりも アライメント後の言語モデルの出力が多様である、ということが実験的に示されている（Figure 3参照）。一方で、アライメントの度合いが小さいため、生成された文章の有用性（すなわち報酬）は低い、ということが分かっている。

データセットの分布が性能に与える理論的考察

この RLHF/DPO 小話シリーズ、どちらかというとアルゴリズムにばかり焦点が当てられてきたが、大事なものを一個忘れていないだろうか？

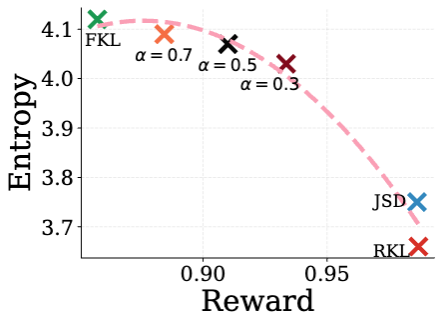

Figure 3: 図はWang et al. [2024] から拝借。α の値が小さいときほど、方策のエントロピーが大きくなり、報酬が低下することが示されている。Forward KL divergence と Reverse KL divergence はそれぞれ、$\alpha = 0$ と $\alpha = 1$ のときの α-divergence であることに注意。

「データ」である。データは機械学習の肝であり、この研究分野でも結果的に得られる言語モデルの性能を大きく左右する。では、どの程度言語モデルの性能が影響されるか理論的に見てみよう。以降の理論解析は、Xiong et al. [2023] および Wachi et al. [2024] をベースにして、キーとなるアイデアが伝わりやすいよう簡略化したものである。もし詳細が気になる方は元論文を御覧いただきたい。

いま、言語モデルの性能を表す関数として、任意の方策 π に対し

$$J(\pi) := \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x)}[r(x,y)] \tag{9}$$

を定義する。この関数は、「ある方策 π が真の報酬 r をどのくらい平均的に獲得できますか？」というのを意味する。すると、今我々が気になるのは、最適な方策 $\pi^\star$ と自分の方策（言語モデル）$\hat{\pi}$ の（真の報酬 r に関する）性能差であり、それは

$$\begin{aligned} J(\pi^\star) - J(\hat{\pi}) &:= \mathbb{E}_{x\sim\rho, y\sim\pi^\star(\cdot|x)}[r(x,y)] - \mathbb{E}_{x\sim\rho, y\sim\hat{\pi}(\cdot|x)}[r(x,y)] \\ &= \mathbb{E}_{x\sim\rho}\Big[\mathbb{E}_{y\sim\pi^\star(\cdot|x)}[r(x,y)] - \mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r(x,y)]\Big] \end{aligned}$$

と表せる。これは、簡単な式変形（同じものと足して引いてるだけ）から

$$\begin{aligned} &\mathbb{E}_{y\sim\pi^\star(\cdot|x)}[r(x,y)] - \mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r(x,y)] \\ &= \mathbb{E}_{y\sim\pi^\star(\cdot|x)}[r(x,y) - \hat{r}(x,y)] - \mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r(x,y) - \hat{r}(x,y)] \\ &\quad + \mathbb{E}_{y\sim\pi^\star(\cdot|x)}[\hat{r}(x,y)] - \mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] \end{aligned}$$

を得る。実は、$J(\pi^\star) - J(\hat{\pi})$ を考えるうえで支配的なのは第二項の

$$\mathbb{E}_{x\sim\rho}\Big[\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r(x,y) - \hat{r}(x,y)]\Big]$$

なので、これに焦点を絞って分析してみよう。まず問題なのは、この項が $\hat{\pi}$ に関して期待値が取られていることである。したがって、

$$\begin{aligned} &\mathbb{E}_{x\sim\rho}\Big[\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r(x,y) - \hat{r}(x,y)]\Big] \\ &= \mathbb{E}_{x\sim\rho}\left[\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\left[\frac{\hat{\pi}(y \mid x)}{\pi^\star(y \mid x)}(r(x,y) - \hat{r}(x,y))\right]\right] \end{aligned}$$

というふうに式変形[7]してみる。

[7] これはオフ方策評価（Off-policy Evaluation, OPE）とかでよく見る式変形ですね。

$\pi^\star$ と $\widehat{\pi}$ がそれぞれ、r と $\widehat{r}$ に関する最適解であることから、任意の f-divergence に対し、

$$\pi^\star(y \mid x) \propto \pi_{\text{ref}}(y \mid x)(f')^{-1}\left(\frac{1}{\beta}r(x,y)\right),$$
$$\widehat{\pi}(y \mid x) \propto \pi_{\text{ref}}(y \mid x)(f')^{-1}\left(\frac{1}{\beta}\widehat{r}(x,y)\right)$$

を満たす。したがって、Reverse KL と Forward KL それぞれに対して、以下のような関係式が得られる。

- Reverse KL divergence のとき

$$\frac{\widehat{\pi}(y \mid x)}{\pi^\star(y \mid x)} \propto \exp\left(\frac{1}{\beta}(r(x,y) - \widehat{r}(x,y))\right)$$

- Forward KL divergence のとき

$$\frac{\widehat{\pi}(y \mid x)}{\pi^\star(y \mid x)} \propto \frac{r(x,y)}{\widehat{r}(x,y)}$$

やはり、ここでも Reverse KL divergence の過激さが分かるだろうか？真の報酬関数とデータから推定した報酬関数の間の誤差の絶対値

$$\varepsilon = |\, r(x,y) - \widehat{r}(x,y)\,|$$

を定義すると、Reverse KL divergece を用いたときの真の方策 $\pi^\star$ と得られた方策 $\widehat{\pi}$ の間の性能差は、

$$J(\pi^\star) - J(\widehat{\pi}) \lessapprox \mathbb{E}_{x\sim\rho}\left[\mathbb{E}_{y\ \sim\pi^\star(\cdot|x)}\left[\exp\left(\frac{1}{\beta}\varepsilon(x,y)\right)\varepsilon(x,y)\right]\right]$$

と大雑把に見積もることができる。なかなかやばい上界になっているのが分かるだろうか？このとき、注目すべき点が二点ある。

一点目は、最適方策からの性能の低下が、$\exp\left(\frac{1}{\beta}\varepsilon(x,y)\right)\varepsilon(x,y)$ で特徴づけられていること。つまり、データが疎であったり、ラベル付けが不正確な (x,y) の領域があったりすると、$\varepsilon(x,y)$ が大きくなるわけだが、その推定誤差が $\exp\left(\frac{1}{\beta}\varepsilon(x,y)\right)\varepsilon(x,y)$ のレートで性能悪化に効いてしまう、ということを意味する。

二点目は、期待値が $\pi^\star$ に関して計算されている、ということである。つまり、最適方策からの性能の乖離を防ぐためには、最適方策が生成しうるような高品質なデータをまんべんなく集めて、正確にラベル付けする必要があるということを意味する。理論的には、低品質なデータをただただ大量に集めるだけでは不十分である、ということを示唆しているのである。[8]

[8] 勘の鋭い読者は「あれ、ラベルは π_{SFT} が集めるのが普通と言っていたじゃないか」と思うかもしれない。まさしくそれはいまホットなトピックで、Online DPO とか Iterative DPO という名前等で研究されており、雑に言えば「アライメントとデータ収集を交互に行う」というアプローチである。データを収集する方策がより高性能になることから、最終的に得られる言語モデルの性能が向上するすることが何件か報告されている。

Reverse KL divergence は悪なのか？

Reverse KL divergence にこだわる必要はないことが分かったが、では他の divergence を使うべきなのだろうか？ [9]Reverse KL divergence は悪なのか？個人的な答えとしては NO だと思う。 [10]そのように考える理由をいくつか挙げると、

- **回答の質。**Reverse KL divergence を用いたときの回答の質がやはり良い。Figure 3で見た通り、Forward KL divergence 等を用いてしまうと、アライメントの「強制力」が低くなってしまい、質の低い回答が増えてしまうのである。
- **多様性。** 単一のモデルで多様な回答をできるようにすべきかは結構怪しいと思う。昨今、モデルマージや Mixture of Expert (MoE) の有用性が叫ばれているわけで、とても良質な回答を出力できるモデルを複数用意したうえで、なにか工夫を施す方が結果として好ましい可能性は高いと（個人的には）思う。
- **バイアス。** こればかりは Reverse KL divergence を使う限りずっと悩まされる弊害だろう。一方で（関係者が見たらこいつ馬鹿だなぁと思うかもしれないが） OpenAI とか Google が考えているのはおそらく
  - **「めっちゃいいデータめっちゃ集めればいいんじゃないの？」**

    じゃないかと思う。たしかに、良質なデータを大量に集めれば [11]、Reverse KL divergence を使って強制的なアライメントをしても全く問題ない。全く問題ないどころか、あらゆる分野のエキスパートにスパルタ的にエリート教育された極めて良質な言語モデルが完成する、というわけである。 [12]

[9] この辺の話は現在進行系で議論が進められている（ちなみに Wang et al. [2024] も ICLR24 論文で新しい）。ここから先は私の個人的な感想だと思っていただきたい。

[10] 説明が回りくどくて申し訳ないという気持ちがありながらも、上の話を理解したうえで Reverse KL を使うのと、何となく使うのとでは、（研究するなら）大きな差になるかなぁと思ってあえてこの書き方をした。

[11] 前章の言い方を使うならば理想的な方策 $\pi^\star$ が生成しうる理想的な出力をまんべんなくラベル付けすれば

[12] 一方で、そこまで高品質なデータを集めることができないときは、Forward KL divergence で緩いアライメントをすべき、という状況は生じるかもしれない。そのとき、RLHF/DPO をそもそもすべきなのか？という議論も当然すべきだが。

Reverse KL のメリットはそれだけではない

Reverse KL divergence の方がなんとなく良いような気がしてたと思う。まだ釈然としない方もいるのではないかと推察するが、実は Reverse KL divergence にはぐうの音も出ない実用的なメリットがあるのである。それが何かについては次回解説しよう。

References

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2023.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models

through reinforced calibration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

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

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

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*, 2023.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *Association for Computational Linguistics (ACL)*, 2019.

Akifumi Wachi, Thien Q Tran, Rei Sato, Takumi Tanabe, and Yohei Akimoto. Stepwise alignment for constrained language model policy optimization. *arXiv preprint arXiv:2404.11049*, 2024.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *International Conference on Learning Representations (ICLR)*, 2024.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

L Zheng, WL Chiang, Y Sheng, S Zhuang, Z Wu, Y Zhuang, Z Lin, Z Li, D Li, and E Xing. Judging llm-as-a-judge with mt-bench and chatbot arena. arxiv preprint arxiv: 230605685. 2023.