Improving Fairness of Large Language Models in Multi-document Summarization

Anonymous ACL submission

Abstract

Fairness in multi-document summarization (MDS) is crucial for providing comprehensive views across documents with diverse social attribute values. Previous works measure fairness in MDS at two levels: summary-level and corpus-level. While summary-level fairness focuses on individual summaries, corpus-level fairness focuses on a corpus of summaries. Recent approaches using prompting or policy gradients primarily focus on summary-level fairness. We propose FairPO, a preference tuning method that improves both summary-level and corpus-level fairness in MDS. To improve summary-level fairness, we propose to generate preference pairs by perturbing document sets based on social attributes. To improve corpus-level fairness, we propose fairness-aware preference tuning by dynamically adjusting the weights of preference pairs based on overrepsentation and underrepresentation of social attributes. Our experiments show that FairPO outperforms strong baselines while maintaining the critical qualities of summaries.

1 Introduction

Multi-document summarization (MDS) aims to summarize the salient information from multiple documents about an entity, such as reviews of a product. Each of these documents is generally associated with a *social attributes* such as sentiments in reviews. These documents with different social attribute values e.g. positive sentiment or negative sentiment tend to have diverse information or conflicting opinions. It is crucial that the summary fairly represents conflicting information since it can significantly impact decision-making.

Previous works (Shandilya et al., 2018; Olabisi et al., 2022; Huang et al., 2024) measure fairness in MDS at two levels: summary-level or corpus-level. Summary-level fairness measures how fairly a summary represents documents with different social attribute values. Corpus-level fairness measures how fairly a corpus of summaries as a whole represents different social attribute values.

Recent studies (Zhang et al., 2023; Li et al., 2024) find that modern summarization methods like LLMs face challenges in generating fair summaries according to summary-level or corpus-level fairness. To improve the summary-level fairness, Zhang et al. (2023) prompt LLMs to generate summaries based on the distribution of social attributes among documents. However, it relies on users' prior knowledge of fairness issues and social attributes, limiting its effectiveness in practice. Huang et al. (2024) improve the summary-level fairness of the finetuned T5 (Raffel et al., 2020) by policy gradient, but their method may not generalize to modern models like LLMs. Furthermore, both methods focus exclusively on summary-level fairness, overlooking the corpus-level fairness.

We propose FairPO (**F**air **P**reference **O**ptimization), a method that improves both summary-level and corpus-level fairness of LLMs in MDS through preference tuning (Ziegler et al., 2019). While previous works (Stiennon et al., 2020; Roit et al., 2023) uses preference tuning to improve the overall quality or faithfulness of summarization, FairPO is the first to use preference tuning for the fairness in MDS. FairPO is based on Direct Preference Optimization (DPO) (Rafailov et al., 2024). To optimize summary-level fairness, FairPO generates preference pairs given perturbed input document sets by removing a small subset of documents with certain social attribute values. To further improve corpus-level fairness, FairPO performs *fairness-aware preference tuning* by dynamically adjusting the weights of preference pairs.

We conduct an empirical evaluation of FairPO using three LLMs: Llama3.1 (AI@Meta, 2024), Mistral (Jiang et al., 2023), and Gemma2 (Team et al., 2024), on the Amazon (Ni et al., 2019), MITweet (Liu et al., 2023), and SemEval datasets

(Mohammad et al., 2016). Our experiments show that FairPO outperforms strong baselines while maintaining other critical qualities of summaries, such as relevance and factuality.

Our contributions are as follows:

- We propose FairPO to improve the fairness of LLMs in MDS;
- We propose to improve summary-level and corpus-level fairness by perturbation-based preference pair generation and fairness-aware preference tuning;
- We perform comprehensive experiments to show the effectiveness of FairPO.

2 Background

In this section, we describe the background knowledge related to fairness in MDS. Let G denote all input document sets in a corpus for MDS. Each document set $D \in G$ contains a list of document $\{d_1, ..., d_n\}$, where d_i denotes the i-th document labeled with a social attribute $a_i \in \{1, ..., K\}$. For each input document set D, a MDS system is supposed to generate a summary S.

To evaluate fairness in MDS, we use Equal Coverage $EC(D, S)$, a summary-level measure, and Coverage Parity $CP(G)$, a corpus-level measure, proposed by Li et al. (2024). Specifically, Equal Coverage $EC(D, S)$ examines whether each social attribute value has equal chances of being covered by the summary S for each document set D. Coverage Parity $CP(G)$ examines whether documents with certain social attribute values are systematically overrepresented or underrepresented based on all documents in the corpus G. These measures are based on the **coverage probability difference** $c(d_i, S)$, which measures differences between the probability of the document d_i being covered by the summary S and the average probability of any document being covered. For both measures, lower values indicate better fairness. Please refer to Li et al. (2024) for more details.

3 FairPO

In this section, we describe our proposed preference tuning method, FairPO.

3.1 Perturbation-based Preference Pair Generation

In this section, we describe how to generate preference pairs based on perturbation. A preference pair for FairPO contains a chosen summary S_c and a rejected summary S_r for the document set D. Ideally, the chosen and rejected summaries should differ significantly in representing documents with different social attribute values. To this end, FairPO generates summaries for perturbed input document sets, where small subsets of documents with specific social attribute values are removed.

Specifically, FairPO first generates a summary S for the full input document set D. For the summary S, FairPO identifies its most overrepresented, k^+, and underrepresented, k^-, social attribute value. These are determined based on the highest or lowest average coverage probability differences, $\mathbb{E}(\{c(d_i, S)|a_i = k\})$. Then, FairPO generates summary S^+ and S^- given the perturbed input document that removes α percent of randomly sampled documents whose social attribute value a_i is k^+ and k^-. Among the summaries S, S^+, S^-, FairPO selects the summary with the lowest Equal Coverage value, indicating the best summary-level fairness, as the chosen summary S_c. The summary with the highest Equal Coverage value is selected as the rejected summary S_r.

3.2 Fairness-aware Preference Tuning

In this section, we describe fairness-aware preference tuning that optimizes summary-level and corpus-level fairness. To achieve this, FairPO dynamically assigns separate weights for the chosen summary S_c and the rejected summary S_r based on corpus-level fairness estimated during training.

FairPO modifies the DPO objective (more explanations in App. A.3) and introduces separate weights, w_c and w_r, for the chosen summary S_c and rejected summary S_r respectively:

$$\sigma(-m)\beta(w_r log\frac{\pi_\theta(S_r|D)}{\pi_{ref}(S_r|D)} - w_c log\frac{\pi_\theta(S_c|D)}{\pi_{ref}(S_c|D)}) \tag{1}$$

where σ is the sigmoid function, π_θ is the policy model, π_{ref} is the reference model, and m is the reward margin as in DPO:

$$m = \beta log\frac{\pi_\theta(S_c|D)}{\pi_{ref}(S_c|D)} - \beta log\frac{\pi_\theta(S_r|D)}{\pi_{ref}(S_r|D)} \tag{2}$$

The term $\sigma(-m)$ in Eq. 1 serves as a scaling factor and FairPO does not consider its gradient.

We now describe how FairPO assigns weights w_c and w_r. It assigns high weights w_c to chosen summaries that improve corpus-level fairness by balancing the overrepresentation and underrepresentation of social attribute values. Conversely, it assigns high weights w_r to rejected summaries that hurt corpus-level fairness. To estimate corpus-level fairness, FairPO computes the sum of coverage probability differences for documents with different social attribute values, $C_k(D, S_*) = \sum_{d \in \{d_i|a_i=k\}} c(d, S_*)$ for each chosen or rejected summary, S_*. A summary S_* is considered overrepresenting or underrepresenting the social attribute value k if the sum of coverage probability differences, $C_k(D, S_*)$, is greater or less than zero respectively. In each training step, FairPO estimates the overrepresentation $O(k)$ of social attribute value k:

$$O(k) = \frac{\sum_{(D,S) \in T_k^+} |C_k(D, S)| \cdot \pi_\theta(S|D)/|S|}{\sum_{(D,S) \in T_k^+} \pi_\theta(S|D)/|S|} \quad (3)$$

where T_k^+ is the set of document sets D and corresponding chosen or rejected summaries that overrepresent social attribute value k ($C_k(D, S_*) > 0$) in recent training steps. Similarly, FairPO estimates the underrepresentation $U(k)$ using the set T_k^- of document sets and summaries that underrepresents social attribute value k ($C_k(D, S_*) < 0$) as Eq. 3.

Using the overrepresentation $O(k)$ and underrepresentation $U(k)$, FairPO assigns weight w_c and w_r. Chosen summaries that help balance overrepresentation $U(k)$ and underrepresentation $O(k)$ receive higher weights and vice versa for rejected summaries. For example, the weight w_c should be higher if the chosen summary S_c overrepresents the social attribute value k systematically underrepresented ($U(k) > O(k)$). For each social attribute value k, FairPO calculates an intermediate weight $w_{c,k}$ for the chosen summary S_c:

$$\begin{cases} \dfrac{2}{1 + (O(k)/(U(k))^{|C_k(D,S_c)|/\tau}}, C_k(D, S_c) > 0 \\ \dfrac{2}{1 + (U(k)/O(k))^{|C_k(D,S_c)|/\tau}}, C_k(D, S_c) < 0 \end{cases} \quad (4)$$

where τ is the temperature. The weight w_c for chosen summaries is the average intermediate weight $w_{c,k}$ across all social attribute values. Similarly, FairPO calculates the weight w_r for the rejected summary S_r with the inverse condition of Eq. 4. The design ensures that summaries enhancing corpus-level fairness are prioritized during training.

4 Experiments

In this section, we describe experiments of finetuning models with FairPO.

	Domain	Soci. Attr.	Soci. Attr. Val.	Doc. Set Size	Doc. Len
Amazon	Review	Sentiment	negative, neutral, positive	8	40
MiTweet	Tweet	Ideology	left, center, right	20	34
SemEval	Tweet	Stance	support, against	30	17

Table 1: Dataset statistics. Doc. Set Size means size of document set and Doc. Len. means average length of documents.

4.1 Datasets

We experiment on three datasets: Amazon (Ni et al., 2019), MITweet (Liu et al., 2023), SemEval (Mohammad et al., 2016) datasets. For each dataset, we use 1000 samples for training, 300 samples for validation, and 300 samples for testing. Tab. 1 shows the statistics of these datasets along with their social attribute values. The summary length is 50 words for all datasets. More details of preprocessing are in App. A.1.

4.2 Implementation Details

We perform experiments with three LLMs: Llama3.1-8b-Instruct (AI@Meta, 2024) , Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Gemma-2-9b-it (Team et al., 2024). LLMs are trained for 2 epochs using LoRA (Hu et al., 2021) with a learning rate of $5e-5$ and batch size of 16. The proportion of documents α removed for generating preference pairs is 10%. The temperature τ is 1 on the MITweet dataset, 2 for Mistral and 1 for other LLMs on the Amazon dataset, 3 for Mistral and 2 for other LLMs on the SemEval dataset. All hyperparameters are tuned on the validation set. More details are in App. A.4.

4.3 Automatic Evaluation of Fairness

We evaluate summary-level and corpus-level fairness of summaries using Equal Coverage (EC) and Coverage Parity (CP) (Li et al., 2024). Using these measures, we compare FairPO with other methods. Specifically, we consider DPO (Rafailov et al., 2024), where the chosen and rejected summaries are selected among three randomly sampled summaries based on EC values like FairPO. We consider OPTune (Chen et al., 2024), which weights preference pairs based on fairness differences between chosen and rejected summaries. We also consider policy gradients (Lei et al., 2024) and the prompting method (Zhang et al., 2023). Details of their implementations are in App A.5. Tab.2 reports the average measures from three runs for

	Amazon		MITweet		SemEval		Overall	
	$EC\downarrow$	$CP\downarrow$	EC	$CP\downarrow$	$EC\downarrow$	$CP\downarrow$	$\overline{EC}\downarrow$	$\overline{CP}\downarrow$
Llama3.1	7.90	1.92	4.43	0.26	2.94	1.33	5.09	1.17
+DPO	6.87	1.04	**4.03**	0.31	2.55	0.91	4.49	0.75
+OPTune	6.58	0.75	4.22	**0.23**	2.50	0.81	4.43	0.60
+Prompt	7.71	1.84	4.33	0.38	2.53	**0.26**	4.86	0.83
+Policy G.	7.71	2.10	4.46	0.31	2.95	1.32	5.04	1.24
+FairPO	**6.57**	**0.37**	4.20	0.26	**2.39**	0.56	**4.39**	**0.39**
Mistral	8.18	2.98	3.98	0.42	2.67	1.07	4.94	1.49
+DPO	7.17	1.55	3.60	0.28	2.21	0.64	4.33	0.82
+OPTune	7.48	1.56	3.60	0.25	2.00	0.67	4.36	0.83
+Prompt	7.67	1.93	4.02	0.23	2.38	0.38	4.69	0.85
+FairPO	**6.98**	**0.89**	**3.56**	**0.21**	**1.97**	**0.36**	**4.17**	**0.49**
Gemma2	8.44	2.75	4.17	0.34	2.74	0.91	5.12	1.33
+DPO	6.87	1.04	4.04	0.29	2.42	0.70	4.44	0.68
+OPTune	6.90	1.15	3.86	0.45	**2.40**	0.65	4.39	0.75
+Prompt	7.21	1.13	4.28	**0.24**	2.62	**0.30**	4.70	0.56
+FairPO	**6.09**	**0.33**	**3.84**	0.47	2.53	0.59	**4.15**	**0.46**

Table 2: Summary-level fairness (EC) and corpus-level fairness (CP) of summaries generated by different methods. The best performing method is in **bold**. The second-best performing method is underlined. FairPO has the best overall performance.

	Llama3.1			Mistral			Gemma2		
	flu.↑	rel.↑	fac.↑	flu.↑	rel.↑	fac.↑	flu.↑	rel.↑	fac.↑
DPO	7.56	8.33	2.78	5.11	11.56	11.56	5.11	1.11	8.67
OPTune	1.00	0.44	-6.89	-0.78	6.78	8.89	7.00	11.67	11.67
Prompt	-15.33	-19.22	-24.44	-0.44	-6.00	-5.56	-42.67	-50.78	-51.44
FairPO	5.78	3.11	2.89	2.11	5.33	9.11	11.44	16.11	9.44

Table 3: Pairwise comparison of quality between summaries generate by LLMs before and after tuning. Statistical significant differences ($p < 0.05$) according to paired bootstrap resampling (Koehn, 2004) are underlined. FairPO does not affect summary quality.

each dataset, and *Overall* scores, which is the average across all datasets. We additionally report the results for ablated versions of FairPO in App.A.6. A lower value indicates better fairness.

From the table, we observe that FairPO outperforms other methods for most LLMs on all datasets and yields the best overall performance for all LLMs. The results show that FairPO improves both summary-level and corpus-level fairness.

4.4 Human Evaluation of Fairness

We perform a human evaluation to compare the fairness of summaries generated by LLMs tuned with DPO and FairPO. For each LLM, we randomly select 10 pairs of summaries generated by the LLM tuned with DPO or FairPO, yielding a total of 30 pairs. Each pair is annotated by three annotators recruited from Amazon Mechanical Turk. Annotators are asked to read all corresponding documents and select the fairer summary. We perform experiments on the Amazon dataset since each document set only contains eight reviews (Tab. 1) and judging the sentiment of an opinion is relatively easy for common users. To simplify the evaluation, we consider document sets with only negative and positive reviews. The Randolph's Kappa (Randolph, 2005) between annotations of three annotators is 0.40, which shows a moderate correlation. The correlation is expected considering the subjectivity of the task. More details are in App. A.2.

Out of 30 pairs, summaries generated by FairPO-tuned LLMs are fairer in 18 pairs and summaries generated by DPO-tuned LLMs are fairer in 9 pairs.The difference is statistically significant ($p < 0.05$) using bootstrap (Koehn, 2004). The results show that FairPO performs better than DPO in improving fairness. We additionally show example summaries generated by FairPO in App. A.7.

4.5 Evaluation of Summary Quality

To evaluate the impact of FairPO on the quality of summaries, we perform a pairwise comparison between summaries generated by LLMs before and after tuning to improve fairness. Specifically, for a pair of summaries, we instruct Prometheus 2 (7B) (Kim et al., 2024) to select the better summary in three dimensions: fluency, relevance, and factuality. To prevent the impact of position bias (Huang et al., 2023), we perform the pairwise comparison twice with different orders of summaries and only consider consistent results among different orders. We report the differences between the winning and losing rates of different methods to improve fairness in Tab. 3. A positive value indicates summary quality is better compared to original LLMs.

From the table, we observe that the quaility of summaries generated by LLMs tuned with FairPO is comparable with summaries generated by original LLMs. Contrarily, prompting significantly hurt the quality of summaries. The results show that FairPO improves the fairness of summaries while maintaining their quality.

5 Conclusion

We propose FairPO, a preference tuning method that optimizes summary-level fairness and corpus-level fairness in MDS. Specifically, FairPO generates preference pairs using perturbed document sets to improve summary-level fairness and performs fairness-aware preference tuning to improve corpus-level fairness. Our experiments show that FairPO outperforms strong baselines while maintaining critical qualities of summaries.

6 Limitation

Our experiments demonstrate FairPO's effectiveness in improving both summary-level and corpus-level fairness of summaries within individual domains. While this work focuses on optimizing fairness within a single domain, extending FairPO to improve fairness simultaneously across multiple domains with diverse social attributes presents a promising future direction. Besides, FairPO currently selects the two summaries with the largest fairness differences among the three generated summaries for preference tuning, following commonly used practices of DPO. Exploring approaches to utilize all three summaries generated by FairPO can be another interesting future direction.

7 Ethical Consideration

The datasets we use are all publicly available. We do not annotate any data on our own. All the models used in this paper are publicly accessible. The inference and finetuning of models are performed on one Nvidia A6000 or Nvidia A100 GPU.

We perform human evaluation experiments on Amazon Mechanical Turk. The annotators were compensated at a rate of $20 per hour. During the evaluation, human annotators were not exposed to any sensitive or explicit content.

References

AI@Meta. 2024. Llama 3 model card.

Lichang Chen, Jiuhai Chen, Chenxi Liu, John Kirchenbauer, Davit Soselia, Chen Zhu, Tom Goldstein, Tianyi Zhou, and Heng Huang. 2024. Optune: Efficient online preference tuning. *arXiv preprint arXiv:2406.07657*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kung-Hsiang Huang, Philippe Laban, Alexander R Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. *arXiv preprint arXiv:2309.09369*.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pre-training to adaptation: A case study in political bias. *arXiv preprint arXiv:2402.00322*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity calibration for opinion summarization. *arXiv preprint arXiv:2404.01706*.

Haoyuan Li, Yusen Zhang, Rui Zhang, and Snigdha Chaturvedi. 2024. Coverage-based fairness in multi-document summarization. *Preprint*, arXiv:2412.08795.

Songtao Liu, Ziling Luo, Minghua Xu, LiXiao Wei, Ziyao Wei, Han Yu, Wei Xiang, and Bang Wang. 2023. Ideology takes multiple looks: A high-quality dataset for multifaceted ideology detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, et al. 2023. Fair abstractive summarization of diverse perspectives. *arXiv preprint arXiv:2311.07884*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Datasets

In this section, we describe how we preprocess the datasets.

Amazon (Ni et al., 2019) consists of reviews with labels of their ratings of different products. We filter out reviews that are non-English or without ratings. We obtain the social attribute of each review based on its rating provided in the dataset. The social attribute of a review will be positive if its rating is 4 or 5, neutral if its rating is 3, and negative if its rating is 1 or 2. We sample 1000 products and their corresponding reviews for training, 300 products for validation, and 300 products for testing.

MITweet (Liu et al., 2023) consists of tweets with labels of political ideologies on different facets about different topics. The social attribute of a tweet will be left if it is left on most facets, right if it is right on most facets, otherwise neutral. First, we evenly divide all tweets into two parts. For each part, we cluster tweets about the same topic based on their TFIDF similarity into clusters. We then divide these clusters into input document sets of 20 tweets about the same topic. We generate 1000 input document sets for training from the first part of the tweets. Similarly, we generate 300 input document sets for validation and 300 input document sets for testing from the second part of the tweets.

Tweet Stance (Mohammad et al., 2016) consists of tweets with labels of stance toward a short phrase such as Climate Change or Hillary Clinton. First, we evenly divide all tweets into two parts. We cluster tweets about the same short phrase based on their TFIDF similarity into clusters. We then divide these clusters into input document sets of 30 tweets about the same short phrase. We generate 1000 input document sets for training from the first part of the tweets. Similarly, we generate 300 input document sets for validation and 300 input document sets for testing from the second part of the tweets.

A.2 Human Evaluation

We perform a human evaluation to compare the fairness of summaries generated by LLMs tuned with DPO and FairPO. For each LLM, we randomly select 10 pairs of summaries generated by the LLM tuned with DPO or FairPO, yielding a total of 30 pairs. Each pair is annotated by three annotators recruited from Amazon Mechanical Turk. The annotators should be from English-speaking countries and have HIT Approval Rates greater than 98%. For each pair, annotators are first asked to read corresponding reviews and unique opinions automatically extracted by GPT-4o-mini (Ouyang et al., 2022). They then evaluate whether each summary reflects these opinions and classify the summary as leaning negative, fair, or leaning positive. Eventually, they are asked to select the fairer summary in each pair. The interface of human evaluation is shown in Fig. 1.

A.3 Relation between FairPO and DPO

The FairPO objective (Eq. 1) is motivated by the derivate of the DPO objective with respect to the model parameters θ:

$$\begin{aligned}\sigma(-m)\beta(\pi_\theta(S_r|D)^{-1}\frac{\partial\pi_\theta(S_r|D)}{\partial\theta}\\ -\pi_\theta(S_c|D)^{-1}\frac{\partial\pi_\theta(S_c|D)}{\partial\theta})\end{aligned} \tag{5}$$

where σ is the sigmoid function, π_θ is the policy model, π_{ref} is the reference model, and m is the reward margin in DPO:

$$\beta log\frac{\pi_\theta(S_c|D)}{\pi_{ref}(S_c|D)} - \beta log\frac{\pi_\theta(S_r|D)}{\pi_{ref}(S_r|D)} \tag{6}$$

The reward margin m can be viewed as a measure of the model's ability to distinguish between the chosen summary S_c and the rejected summary S_r. A larger value of m indicates that the model is already proficient at differentiating S_c from S_r. Consequently, DPO assigns lower weights, $\sigma(-m)$, to chosen and rejected summaries where the model is confident in their differences and higher weights to chosen and rejected summaries where the differences is more challenging. The term $\sigma(-m)$ can help the model focuses more on difficult cases.

The objective of FairPO is designed so that chosen and rejected summaries have separate weight while preserving the effect of the term $\sigma(-m)$ in Eq.5. The derivative of FariPO objective with respect to the model parameters θ is as follows:

$$\begin{aligned}\sigma(-m)\beta(w_r\pi_\theta(S_r|D)^{-1}\frac{\partial\pi_\theta(S_r|D)}{\partial\theta}\\ -w_c\pi_\theta(S_c|D)^{-1}\frac{\partial\pi_\theta(S_c|D)}{\partial\theta})\end{aligned} \tag{7}$$

Comparing with the derivative of DPO objective (Eq. 5), the term $\sigma(-m)$ remains consistent in the derivative of FairPO objective.

Suppose we directly add seperate weights w_c and w_r for chosen and rejected summaries to DPO objective. The corresponding objective is as follows:

$$\begin{aligned}-log\sigma(\beta w_c log\frac{\pi_\theta(S_c|D)}{\pi_{ref}(S_c|D)} -\\ \beta w_r log\frac{\pi_\theta(S_r|D)}{\pi_{ref}(S_r|D)})\end{aligned} \tag{8}$$

The corresponding derivative is as follows:

$$\begin{aligned}\sigma(-m')\beta(w_r\pi_\theta(S_r|D)^{-1}\frac{\partial\pi_\theta(S_r|D)}{\partial\theta}\\ -w_c\pi_\theta(S_c|D)^{-1}\frac{\partial\pi_\theta(S_c|D)}{\partial\theta})\end{aligned} \tag{9}$$

where m' is a weighted reward margin:

$$\beta w_c log\frac{\pi_\theta(S_c|D)}{\pi_{ref}(S_c|D)} - \beta w_r log\frac{\pi_\theta(S_r|D)}{\pi_{ref}(S_r|D)} \tag{10}$$

Comparing with m, m' is less effective as a measure of the model's ability to distinguish between the chosen summary S_c and the rejected summary S_r since the term $log\frac{\pi_\theta(S_c|D)}{\pi_{ref}(S_c|D)}$ and $log\frac{\pi_\theta(S_r|D)}{\pi_{ref}(S_r|D)}$ have different weights. We additionally provide empirical evidences in App.A.6.

A.4 Implementation Details

To reduce training cost, we perform LoRA (Hu et al., 2021) tuning. Specifically, the rank for LoRA tuning is 16 and the scaling factor is also 16. All models are quantized in 8-bit to additionally reducing training cost.

We prompt these LLMs to generate summaries for the input document sets of different datasets. The prompt are tuned so that the average length of generated summaries are 50 words. We show the summarization prompts for the Amazon dataset in Fig. 2. The temperature for generation is 0.6 for all LLMs.

The set T_k^+ in Eq.3 is updated so that recent training steps have higher impacts. Specifically, at the end of each training step, the impacts of all the samples already in the set T_k^+ are reduced with a discount factor γ. Then, all the samples that over-represents social attribute value k ($C_k(D, S_*)$>0) in current training steps are added to the set T_k^+. The discount factor γ is 0.75 for Llama3.1 and 0.5 for other LLMs.

The goal of the exponent, $|C_k(D, S_C|$, of $O(k)/(U(k)$ or $U(k)/(O(k)$ in Eq. 4 is to adjust the weight w_c such that it more deviate from

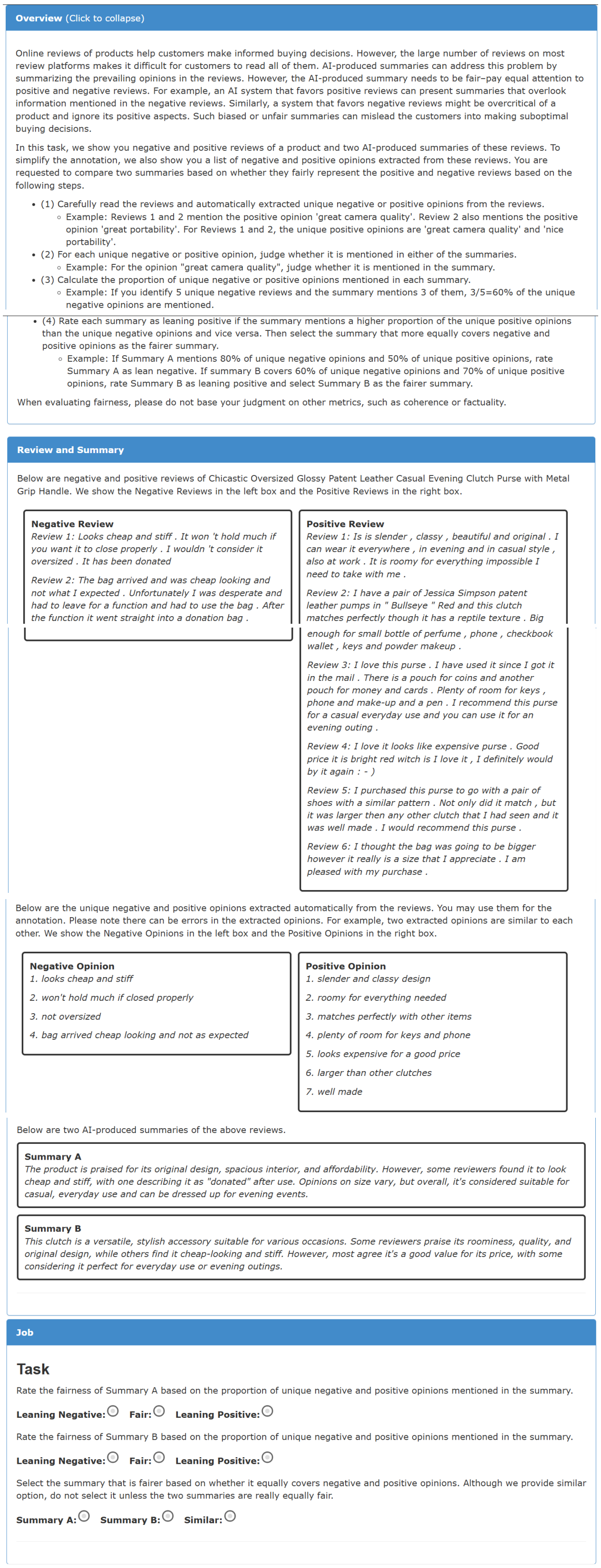

Figure 1: Interface for Human Evaluation

Below is a list of product reviews:

1.This is a card reader that does everything I needed it to . My adapters for the micro SD cards were defective so I have no complaints only praise . It reads any Compact Flash , Memory Stick , SD , and XD cards . Well that is all I wanted to say except this is a great product overall , and thank you .

2.The pins in the CF slot are very flimsy and get bent out of alignment easily , making it impossible to insert the card (until you perform delicate surgery on the pins with small tweezers) . Do not buy this product if you will ever use the CompactFlash slot . It will just lead to frustration .

3.So far I only use this for SM and SD cards , but it installed (USB) quickly , easily and reads the cards I need read .

4.Initially it worked great but after the 5th time it stopped working . It also helped fry my SD-card will all my pictures and video clips . Not happy at all with this product .

5.Reads 64 cards is quite deceiving . It only reads four types of cards made by 64 different manufacturers . Also , the connector port is difficult to plug in .

6.good product , reads quite fast. only issue is that the card reader does not have a satisfying ' click ' when the card is inserted. you kinda have to stick the card in the slot and hope it is lodged properly .

7.I can get it to read SD cards , but I bought it to read my CF 's and it won 't read a single one . My experience is in line with others . Go check out similar reviews on newegg.com.

8.The card reader comes in retail packaging and totally lacks instructions on how best to put 68 types of cards into 4 slots . It did read an SD card successfully . The micro usb plug on the usb cord broke after 1 use .

Please write a single summary around 50 words for all the above reviews.

Figure 2: Summarization prompt for the Amazon Dataset.

one as $|C_k(D, S_C|$ increases. Therefore, FairPO does not directly use the raw value of the absolute value of the sum of coverage probability differences $|C_k(D, S_C|$ as the exponent. Instead, FairPO uses the normalized $|C_k(D, S_C|$ among all training samples with the average of 1 as the exponent.

A.5 Implementation of Baseline

We implement the policy gradient method proposed by Lei et al. (2024) as a baseline. In the original implementation, there is a loss that maximize the probability for reference summary in addition to the policy gradients. Since datasets used in this paper do not contain reference summary, we only consider the policy gradients. Besides, for a fair comparison with other methods, we implement the policy gradient method in an offline setting. The learning rate for the policy gradeint is $1e-6$ following the original paper. We only implement the policy gradient method for Llama3.1 since the training is very unstable even if we lower the learning rate to $1e-9$ for Mistral and Gemma2. For OPTune and DPO, they use the same hyperparameters as FairPO.

A.6 Ablation Study

We compare FairPO with its ablated versions. We consider FairPO without perturbation-based preference generation (w/o pert.). For this version, the chosen and rejected summaries are selected among three randomly sampled summaries based on Equal coverage values. We consider FairPO without fairness-aware preference tuning (w/o fair.). For this version, FairPO uses DPO objective for preference tuning. We also consider FairPO without reward margin (w/o rew.). For this version, FairPO uses the objective function described in Eq. 8. Tab.4 reports the average measures from three runs for each dataset, and *Overall* scores, which is the average across all datasets. A lower value indicates better fairness.

	Amazon		MITweet		SemEval		Overall	
	$EC\downarrow$	$CP\downarrow$	EC	$CP\downarrow$	$EC\downarrow$	$CP\downarrow$	$\overline{EC}\downarrow$	$\overline{CP}\downarrow$
Llama3.1								
FariPO	6.57	**0.37**	4.20	**0.26**	2.39	**0.56**	**4.39**	**0.39**
w/o pert.	7.01	0.48	**4.07**	0.34	2.54	0.81	4.54	0.54
w/o fair.	6.70	0.95	4.26	0.31	**2.29**	0.65	4.42	0.64
w/o rew	**6.48**	0.79	4.19	0.27	2.60	0.86	4.42	0.64
Mistral								
FariPO	**6.98**	**0.89**	**3.56**	0.21	**1.97**	**0.36**	**4.17**	**0.49**
w/o pert.	7.29	1.64	3.81	0.21	2.30	0.26	4.47	0.71
w/o fair.	7.31	1.36	3.57	0.25	2.21	0.66	4.37	0.76
w/o rew	7.05	1.26	3.65	**0.14**	2.06	0.55	4.25	0.65
Gemma2								
FariPO	**6.09**	**0.33**	**3.84**	0.47	2.53	0.59	**4.15**	0.46
w/o pert.	6.18	0.19	4.17	**0.21**	2.43	0.53	4.26	**0.31**
w/o fair.	6.77	1.11	3.84	0.51	**2.39**	0.59	4.34	0.74
w/o rew	6.89	0.90	3.94	0.40	2.49	**0.44**	4.44	0.58

Table 4: Summary-level fairness (EC) and corpus-level fairness (CP) of summaries generated by ablated versions of FairPO. The best performing method is in **bold**. FairPO has the best overall performance.

From the table, we observe that FairPO yields the best overall performance compared to its ablated versions. The results show the effectiveness of perturbation-based preference pair generation and fairness-aware preference tuning. It also provides empirical evidences for the choice of objective of FairPO.

Llama3.1	Mistral	Gemma2
DPO: The product reviews are mixed, with some customers expressing satisfaction with the quality and price, while others find it flimsy, useless, and painful to use. Issues include bending, crookedness, and difficulty with hair removal, leading to disappointment and a recommendation to avoid the product.	**DPO:** This tablet receives generally positive reviews, praised for its portability, picture quality, speed, and features like USB and HDMI ports. Some users found the battery life lacking, while one experienced a defective product. Overall, it's considered a good value for the price and user-friendly, with a few noting its similarity to Android phones.	**DPO:** This plush bed topper is generally well-received for its comfort and ability to stay in place thanks to its straps. Many users praise its cloud-like softness and how it elevates their sleep quality. However, some users found it loses its fluffiness over time and requires manual plumping.
FairPO: The product reviews are mixed, with some users praising its availability and quality, while others criticize its flimsiness, uselessness, and high price. Some users find the tips too long, weak, or hard to use, while others appreciate its effectiveness, but note that it may not be suitable for all purposes.	**FairPO:** This Toshiba tablet receives mixed reviews. Users praise its lightweight design, fast speed, great picture quality, and affordability. Some highlight its functionality similar to laptops and its ease of use. However, concerns arise regarding battery life and a negative experience with a damaged, non-functional product.	**FairPO:** This plush bed topper receives rave reviews for its comfort, fluffiness, and stay-in-place straps. Some users find it a dream come true, while others experience flattening over time. It's praised for its luxurious feel, comparable to a feather bed without the prickliness. While some find it ideal for air mattresses and adding height, others note back pain issues.

Figure 3: Sample summaries generated by DPO and FairPO.

A.7 Qualiative Example

We show sample summaries generated by LLMs tuned with DPO and FairPO on the Amazon dataset in Fig. 3. From the figure, we observe that summaries generated by LLMs tuned FairPO tend to more balancely present negative and positive information.