FINE-TUNING NEWS SUMMARIZATION MODEL WITH HUMAN PREFERENCES

Rimika Dhara, Nikhitha Gollamudi, Akhil Mavallapalli & Xifan Li
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
{dhara015, golla063, maval003, li003646} @umn.edu

1 INTRODUCTION AND PROBLEM STATEMENT

Traditional news summarization models lack optimization with respect to human preference. Currently, transformer-based summarization models are facing issues including inaccurate information and verbose summaries that are not aligned with how humans would normally prefer their news to be summarized. Given these issues, the problem we aim to tackle is improving how accurately, precisely, and concisely we can tune our model to align more closely with human preferences. We also want to understand how much difference this tuning would make to the established pre-trained models.

Our approach to this problem is to implement reinforcement learning (RL) by training a reward model on human-labeled preferences. More specifically, we employ two RL algorithms, Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO), in hopes of increase the performance of the summarization model. This will help directly tackle current issues by optimizing the model to create summaries that humans prefer. Building on the work of Ziegler et al. (2019) and Stiennon et al. (2020), we apply RL from human feedback (RLHF) specifically to the news domain, with particular attention to factual accuracy and conciseness.

2 SIGNIFICANCE OF THE PROBLEM

Automated summarization plays a vital role in modern information consumption, enabling quick understanding of large texts by highlighting key ideas. This is especially impactful in news summarization, where journalism can benefit from faster, more digestible content. However, current systems face major issues—most notably, factual inaccuracy and verbosity—which limit their effectiveness. Inaccurate or overly long summaries can lead to misinformation, reduce readability, and erode trust in AI-generated content, ultimately defeating the purpose of aiding information processing. Beyond journalism, accurate and concise summarization has critical applications in domains like healthcare, legal analysis, and finance, where mistakes can be costly and time is a premium resource. Although transformer-based models have advanced the field, existing limitations show the need for better alignment with human preferences. Incorporating human feedback into training—through methods like RL—can improve both reliability and utility. This work has the potential to reshape how AI contributes to information consumption, setting new standards for trustworthy, human-aligned summarization systems.

3 PREVIOUS AND RELATED WORK

The field of language model fine-tuning from human preferences has seen significant developments in recent years. Notably, Ziegler et al. (2019) demonstrated that RL from human feedback (RLHF) can successfully fine-tune language models for tasks including summarization. Their work on the TL;DR and CNN/Daily Mail datasets showed that models trained with 60,000 human comparisons outperformed supervised baselines according to human evaluators, although they observed that the resulting models often acted as ”smart copiers” that selected relevant sentences from the input. Building on this foundation, Stiennon et al. (2020) further refined the approach for summarization tasks, showing that human feedback can be more effective than traditional supervised learning methods. Their work emphasized the importance of high-quality human feedback and introduced iterative

data collection procedures to improve model performance. In parallel, recent work by B"ohm et al. (2019) also explored using human evaluations to learn reward functions for summarization, providing detailed investigations of learned policies on the CNN/Daily Mail dataset. Meanwhile, advances in large language models like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) have established new state-of-the-art results on summarization benchmarks using traditional supervised approaches. More recently, researchers have explored various techniques to improve RLHF, including Constitutional AI (Bai et al., 2022), which combines rule-based and preference-based learning, and Direct Preference Optimization (Rafailov et al., 2023), which reformulates the preference learning problem to avoid the instabilities associated with traditional RL approaches. These advancements provide significant evidence that this is a promising direction.

4 Our Goals

Our fine-tuning framework is designed with two key objectives: (1) benchmark summarization models trained with traditional supervised methods, and (2) explore how preference-based feedback can improve summary alignment with human expectations. As illustrated in Figure 1, our pipeline branches into two tracks: one for supervised fine-tuning using the CNN/DailyMail dataset, and the other for preference-driven learning using OpenAI's TL;DR dataset. The TL;DR dataset consists of Reddit posts and two candidate summaries, with binary preference annotations indicating which summary was preferred by a human annotator. We use this data to both train a reward model and directly fine-tune models using Direct Preference Optimization (DPO). We also simulate Proximal Policy Optimization (PPO) by sampling multiple summaries and reranking them using reward scores. The full codebase for this project is available at https://github.com/rimikadhara67/HumanPref-NewsSummary.

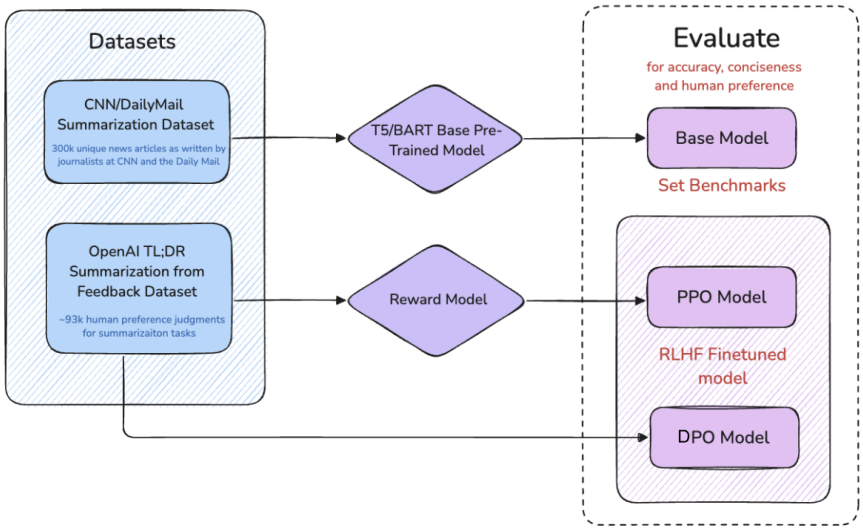

Figure 1: Project Workflow Overview

5 Our Work on Human Preference-Aligned Fine-Tuning

This section presents our empirical work to align news summarization models with human preferences. We begin with supervised fine-tuning of baseline models on the CNN/DailyMail dataset Hermann et al. (2015), followed by evaluations using both a custom-trained reward model and `OpenAssistant/reward-model-deberta-v3-large-v2` that is trained on the TL;DR dataset. We then describe our simulated PPO approach, where multiple candidate summaries are reranked using preference scores. Finally, we implement Direct Preference Optimization (DPO) Rafailov et al. (2023), a method that directly incorporates human preference comparisons during

fine-tuning using the TL;DR dataset Ziegler et al. (2019). All experiments are conducted using the `trl` library Face (2023).

5.1 Baseline Metrics and Supervised Fine-Tuning

We fine-tuned two baseline models—`t5-base` on 2K samples and `t5-large` on 1K samples from the CNN/DailyMail dataset Hermann et al. (2015)—using standard supervised learning. The purpose of this baseline was to establish a performance reference point for later reinforcement learning methods. Evaluation was conducted using ROUGE, METEOR, and reward scores from two separate models: our custom-trained reward model and OpenAssistant's DeBERTa-v3 model available via Hugging Face OpenAssistant (2023).

With Our Custom-Trained Reward Model: We trained a binary classifier reward model based on `distilbert-base-uncased` architecture using 93K human preference judgments from OpenAI's TL;DR dataset Ziegler et al. (2019). Each training example consisted of a Reddit post with two summaries, and a binary label indicating which was preferred. We converted this into pairwise format and optimized using margin ranking loss.

In our initial 5-epoch training, training loss dropped from 1.02 to 0.6128 while validation loss improved until epoch 3 (0.5053) before plateauing—suggesting overfitting. The epoch 3 checkpoint was selected for downstream scoring. Figure 2 shows the learning curve from this run.

Figure 2: Training and validation loss per epoch for our custom-trained reward model (default schedule). Overfitting begins around epoch 4.

To address overfitting, we explored multiple regularization strategies. First, we extended training to 10 epochs and observed marginal improvements, though the validation loss continued to fluctuate. Second, we added dropout layers to reduce reliance on specific activations. Most notably, we incorporated a cosine annealing learning rate scheduler, which yielded smoother convergence and modest improvements in both training and validation loss across 5 epochs (validation loss decreased from 0.6721 to 0.5400). While this training run outperformed our earlier default-schedule attempts, the validation curve flattened toward later epochs, indicating diminishing returns (shown in Figure 3).

Despite this improvement, the model continued to show limited generalization beyond the training distribution—likely due to the compact architecture of DistilBERT and the domain mismatch between Reddit-based preferences and news article summaries. We also experimented with longer training and dropout-based regularization, but performance gains remained marginal. Ultimately, we chose to evaluate all summarization models using the more robust HuggingFace reward model, `OpenAssistant/reward-model-deberta-v3-large-v2` OpenAssistant (2023), which demonstrated stronger alignment with human preference and more reliable discriminative scoring (discussed later in this section). While KL divergence was not used during reward model training, it remains a promising direction for future work to enhance stability and downstream utility.

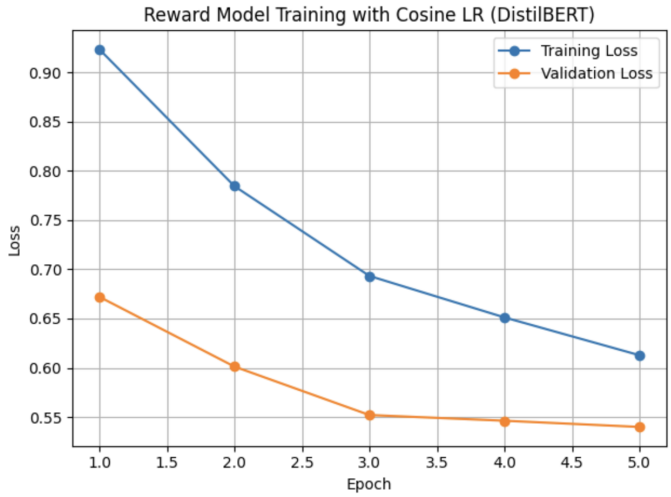

Figure 3: Training and validation loss for reward model with cosine annealing LR. Curve flattens toward later epochs.

Model	Dataset Size	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR
`t5-base`	2K	0.3302	0.1345	0.2523	-	0.1495
`t5-large`	1K	0.2468	0.0890	0.1998	-	0.1495

Table 1: Baseline summarization performance evaluated using our custom-trained reward model. ROUGE-Lsum not available.

With HuggingFace Reward Model: Given our model's limited generalization despite multiple interventions, we turned to `OpenAssistant/reward-model-deberta-v3-large-v2` OpenAssistant (2023), a robust reward model trained on a much larger corpus and deeper architecture. This model was used to score both our supervised baselines and later preference-tuned models. The DeBERTa reward model provided consistent, discriminative scores that penalized low-preference outputs and better captured human alignment. It also allowed us to use the same evaluation model across experiments for consistent comparisons, and integrates cleanly with the `trl` library Face (2023) for future reinforcement learning compatibility.

Model	Dataset Size	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR
`t5-base`	2K	0.3744	0.1690	0.2706	0.3207	0.3236
`t5-large`	1K	0.3755	0.1672	0.2621	0.3201	0.3204

Table 2: Baseline summarization performance evaluated using HuggingFace's DeBERTa reward model.

In summary, while our custom reward model showed expected learning trends and marginal improvement with cosine annealing, it lacked the reliability and precision of the HuggingFace model. These results underscore the importance of architecture scale, data quality, and training strategies in reward model performance.

5.2 Proximal Policy Optimization (PPO)

To explore reinforcement learning from human feedback (RLHF), we implemented a simulated PPO setup using the `facebook/bart-large-cnn` model and a reward reranking pipeline. While `trl`'s PPOTrainer natively supports causal models (e.g., GPT-2), it is not compatible with encoder-decoder models like T5 or BART. Thus, we used a proxy PPO simulation by generating multiple candidate summaries for each article and scoring them using a reward model trained on human preferences. This allowed us to simulate PPO's effect on summary quality by selecting outputs that maximize human-aligned reward without direct gradient updates.

Training Setup. We first fine-tuned the `facebook/bart-large-cnn` model using supervised learning on 2K samples from CNN/DailyMail. We then selected 100 validation articles and generated 4 candidate summaries per article using different decoding configurations: beam search (length penalties of 1.0 and 2.0), top-k sampling (k=50), and nucleus sampling (top-p=0.9). This gave us a total of 400 summaries. Each set of 4 candidates per article was then passed to our preference-trained reward model (`OpenAssistant/reward-model-deberta-v3-large-v2`), and the summary with the highest reward was selected as the "PPO-improved" output.

Evaluation. We computed standard metrics (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR) for the top-selected summaries, as well as the mean reward score using the DeBERTa model. These metrics are available in Table 3.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR	Reward
PPO	0.3830	0.1722	0.2750	0.3315	0.3250	+0.0412

Table 3: Evaluation metrics for PPO-generated summaries using HuggingFace DeBERTa reward model.

PPO outputs showed moderate improvements over the supervised baseline in all metrics. ROUGE-1 increased from 0.3744 to 0.3830, and reward score improved from -2.6000 to $+0.0412$. However, gains were lower than DPO, which achieved a higher reward score (+0.1290) and stronger improvements across metrics. This suggests that while PPO-based reranking can nudge models toward human preferences, it lacks the direct gradient-based updates that DPO benefits from. Qualitatively, summaries selected via PPO were slightly more concise and less repetitive than beam search outputs, often removing trailing phrases or off-topic information. However, since PPO was not directly fine-tuned, its ability to generalize beyond sampled generations is limited.

This PPO setup is a simulation rather than full policy optimization. No backpropagation occurred to update the model weights. A more complete PPO implementation would require switching to causal language models and integrating the reward model directly into training via `trl.PPOTrainer`. Simulated PPO showed clear alignment improvements over baseline, both in terms of reward and lexical metrics. However, DPO proved to be more effective and robust in our setting (explained in the next section). PPO remains promising but requires architecture-compatible models and larger-scale experimentation to realize its full potential.

5.3 Direct Preference Optimization (DPO)

Given the limitations of PPO for encoder-decoder models, we fine-tuned using Direct Preference Optimization (DPO). Unlike PPO, DPO directly uses (prompt, chosen, rejected) triplets and updates model weights to increase the log-likelihood of preferred summaries relative to rejected ones. It avoids sampling and is more stable and efficient for preference learning.

Training Procedure: We used `trl.DPOTrainer` to fine-tune `t5-base` on 1K examples from the TL;DR comparisons dataset. Each Reddit post was prepended with the `summarize:` prompt. The chosen and rejected summaries were fed into the trainer, and training was run for 3 epochs with `beta=0.1` for KL regularization. Importantly, **no reward model was used during training**—DPO relies solely on the relative log-likelihood of the chosen versus rejected summaries, without computing scalar reward scores.

Training Curve: We tracked the training loss across 75 checkpoints and visualized the full progression in Figure 4. The plot includes actual loss points, a smoothed curve using Savitzky-Golay filtering, min and max markers, and a linear trend line. While the curve exhibits noise, the general downward slope with smoother trends shows effective learning over time.

Evaluating the DPO Model. Although DPO does not use a reward model during training, we evaluate the fine-tuned model using a separate preference-trained model: `OpenAssistant/reward-model-deberta-v3-large-v2`. This reward model, also used for our baseline benchmarks, acts as a proxy for human judgment, helping us assess whether the

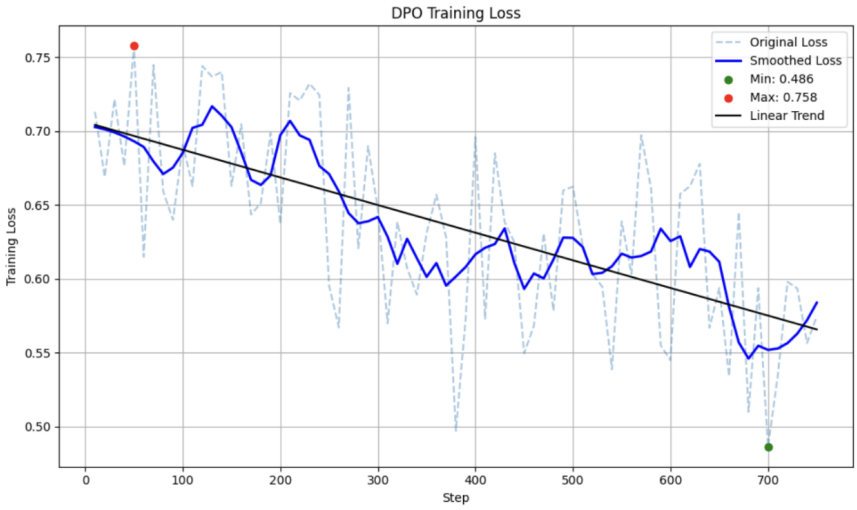

Figure 4: DPO training loss over steps, with smoothed curve, min/max points, and trendline.

outputs from our DPO model align better with human preferences. Importantly, because DPO was trained using only (prompt, chosen, rejected) pairs and never accessed this reward model, there is no risk of evaluation circularity or data leakage. In fact, using the same reward model to evaluate both baseline and DPO ensures that we maintain a consistent and controlled metric for alignment. If DPO achieves a higher reward score on the same evaluator, it suggests genuine generalization to the preferences the model represents.

We opted to evaluate using the HuggingFace reward model rather than our custom-trained reward model because the latter showed signs of overfitting and poor generalization. As reported in earlier sections, our model plateaued in validation loss after epoch 3 and yielded significantly lower ROUGE and METEOR scores. In contrast, the HuggingFace model consistently produced more nuanced and discriminative scores for both baseline and DPO outputs. Its alignment with our evaluation goals made it a more robust and trustworthy proxy for human preferences.

We also compute traditional metrics such as ROUGE and METEOR for lexical quality, shown in Table 4. Specifically, we generated summaries for 50 CNN/DailyMail articles using the DPO model. Each summary was evaluated using ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR. In parallel, we computed the average reward score from the HuggingFace DeBERTa model by passing in each generated summary and applying softmax-based scoring. we evaluate the fine-tuned model using a separate preference-trained model: `OpenAssistant/reward-model-deberta-v3-large-v2`. This reward model acts as a proxy for human judgment, helping us assess whether the outputs from our DPO model align better with human preferences. We also compute traditional metrics such as ROUGE and METEOR for lexical quality.

Specifically, we generated summaries for 50 CNN/DailyMail articles using the DPO model. Each summary was evaluated using ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR. In parallel, we computed the average reward score from the HuggingFace DeBERTa model by passing in each generated summary and applying softmax-based scoring.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR	Reward
DPO-finetuned	0.3864	0.1763	0.2798	0.3360	0.3262	0.1290

Table 4: Evaluation metrics of DPO-finetuned model.

Compared to our supervised T5 baselines, DPO significantly improved across all ROUGE metrics and METEOR. ROUGE-1 increased from 0.3744 (T5-base) to 0.3864, and ROUGE-Lsum improved

from 0.3207 to 0.3360. METEOR showed a slight improvement from 0.3236 to 0.3262, further indicating closer lexical and semantic alignment with reference summaries.

Most notably, the average reward score—using OpenAssistant's `deberta-v3` model—jumped from **2.6000** (baseline T5) to **+0.1290**. This sharp shift shows that DPO fine-tuning substantially aligned the model's outputs with human preferences, as encoded in the reward model. In contrast, supervised models were heavily penalized. Overall, DPO training demonstrated not only improved alignment but also preservation—and even slight enhancement—of lexical similarity. These results confirm that DPO is more effective than simulated PPO for preference-based summarization fine-tuning.

6 Conclusion and Discussion

Our study explored multiple strategies to align summarization models with human preferences, including supervised fine-tuning, PPO-based reranking, and DPO fine-tuning. As shown in Table 5, the DPO-finetuned model outperformed all other approaches across every evaluation metric. It achieved the highest scores for ROUGE-1 (0.3864), ROUGE-2 (0.1763), ROUGE-L (0.2798), ROUGE-Lsum (0.3360), and METEOR (0.3262). Most notably, the DPO model shifted the average reward score—computed using the HuggingFace DeBERTa reward model—from a significantly negative baseline of -2.6000 to a positive score of +0.1290, demonstrating a strong alignment with human preferences.

Model	Dataset Size	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR	Reward
Supervised T5-base	2K	0.3744	0.1690	0.2706	0.3207	0.3236	-2.6000
PPO	100 (val)	0.3830	0.1722	0.2750	0.3315	0.3250	+0.0412
DPO-finetuned	1K	**0.3864**	**0.1763**	**0.2798**	**0.3360**	**0.3262**	**+0.1290**

Table 5: Comparison of supervised, PPO, and DPO outputs using HuggingFace DeBERTa reward model and standard evaluation metrics. Best values are bolded.

In contrast, the PPO approach, although limited to reranking without gradient-based learning, still offered a noticeable improvement over the supervised baseline. PPO achieved a reward score of +0.0412 and moderate gains across lexical metrics. However, the lack of backpropagation in this implementation constrained its ability to fine-tune preferences deeply. The supervised T5-base model served as a baseline but underperformed in all alignment and quality metrics. While it demonstrated reasonable lexical similarity, its negative reward score reflects poor alignment with human-preferred summaries. This highlights the limitations of traditional supervised fine-tuning in optimizing for human values.

Despite promising results, our project faced several technical and methodological challenges. First, we were unable to apply PPO directly to the T5 model due to incompatibilities with the `trl.PPOTrainer`, which is currently limited to causal language models. Second, our custom-trained reward model, while effective during early training, quickly overfit and failed to generalize—despite interventions like dropout and cosine learning rate scheduling. Data preparation also required extensive effort: TL;DR comparisons had to be flattened, and CNN/DailyMail summaries required prompt formatting. Finally, the DPO implementation necessitated careful version control and tokenization setup within the `trl` and `transformers` libraries.

Looking ahead, future work could involve implementing PPO using decoder-only architectures like GPT-2 or OPT, which are compatible with gradient-based preference optimization. Reward models could also be fine-tuned on domain-specific summaries to improve alignment in the news domain. Moreover, integrating human-in-the-loop mechanisms such as active preference labeling could offer better generalization and real-time adaptability. Expanding the scale of DPO experiments—both in terms of training data and model capacity—may yield further improvements in alignment and quality. In conclusion, Direct Preference Optimization presents a powerful and stable approach for aligning summarization models with human preferences, outperforming both supervised baselines and PPO in our experiments. With thoughtful implementation and scaling, preference-based fine-tuning has the potential to set a new standard for reliable and human-aligned summarization systems.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Florian B"ohm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*, 2019.

Hugging Face. Trl: Transformer reinforcement learning, 2023. `https://github.com/huggingface/trl`.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pp. 1693–1701, 2015.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

OpenAssistant. Openassistant reward model, 2023. `https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2`.

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. *arXiv preprint arXiv:2305.18290*, 2023.

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

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

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