Direct Preference Optimization Report

Edward Ibarra & Aniket Chaudhry

November 4, 2025

1 Introduction

This report details an experiment in aligning a large language model using **Direct Preference Optimization (DPO)**. The primary objective was to fine-tune a GPT-2 Large model to generate text with a consistently positive sentiment, using the IMDb dataset as a benchmark.

2 Code Organization

Our project is organized into three main parts. The first file, `sft.py`, handles supervised fine-tuning. It trains the GPT-2 Large model on the IMDb training data for one epoch using the TRL SFTTrainer and saves the fine-tuned model as `./fine-tuned-gpt2-large`, which we later use as the reference model for DPO. The second file, `preference_pairs.py`, generates synthetic preference pairs by taking 1,000 random prompts from the IMDb dataset, generating four completions for each, and ranking them using the pretrained `siebert/sentiment-roberta-large-english` sentiment classifier. The ranked results are saved as JSON files in the `./generations` folder. Finally, the `dpo.py` performs the Direct Preference Optimization training and evaluation. It loads all preference pairs, runs DPO training for beta values of 0.05, 0.1, and 1.0, and logs the results to Weights & Biases. The `eval.py` file evaluates each trained model using test prompts and generates the Reward vs. KL plot to visualize the trade-off between performance and divergence.

3 SFT Run Metrics

Link to Wandb Report for Metrics

4 DPO Run Metrics

The training for this model is split into two parts because the initial session (Part 1) was automatically stopped at approximately step 690 after exceeding the job's time limit. The job was then resumed from a checkpoint at step 600 (Part 2) and ran to completion at step 750.

Link to Wandb Report for Metrics

5 Reward vs. KL Divergence

Our Reward vs. KL plot shows a clear upward trend similar to what was described in the paper. As the KL divergence increases, the sentiment reward also goes up, meaning that models with higher divergence from the reference model produce text with stronger positive sentiment.

The SFT model starts with a positive probability of around 0.74, while all three DPO models ($\beta = 0.05, 0.1, 1.0$) improve to nearly 1.0. The $\beta = 1.0$ model increases the fastest but has more variation between checkpoints, while $\beta = 0.05$ stays more stable overall.

In general, our results follow the expected pattern from the paper, where the reward rises quickly with moderate KL divergence and then levels off close to 1.0. This shows that the DPO models learned to generate more positive text while still staying relatively close to the reference model.

6 Training Methods

For training, we used GPT-2 Large with the TRL library. In SFT, the model was trained on the IMDb training split for one epoch using the default learning rate of 2×10^{-5}. Checkpoints were saved every 500 steps, and all metrics were logged to Weights & Biases.

In DPO, we used the fine-tuned model as both the reference and starting model. It was trained for one epoch with the same 2×10^{-5} learning rate, using three β values (0.05, 0.1, 1.0). Checkpoints were saved every 100 steps, and each run was tracked in Weights & Biases.

7 Results

Our DPO results showed a big improvement over the SFT baseline. The SFT model had an average sentiment reward of about 0.74, while all DPO models were close to 1.0. The final KL divergence values were around 43.3 for $\beta = 0.05$, 29.2 for $\beta = 0.1$, and 18.2 for $\beta = 1.0$.

As β got larger, the models stayed closer to the reference model but had slightly lower rewards. This shows the expected trade-off where smaller β values lead to stronger positive sentiment but higher divergence.

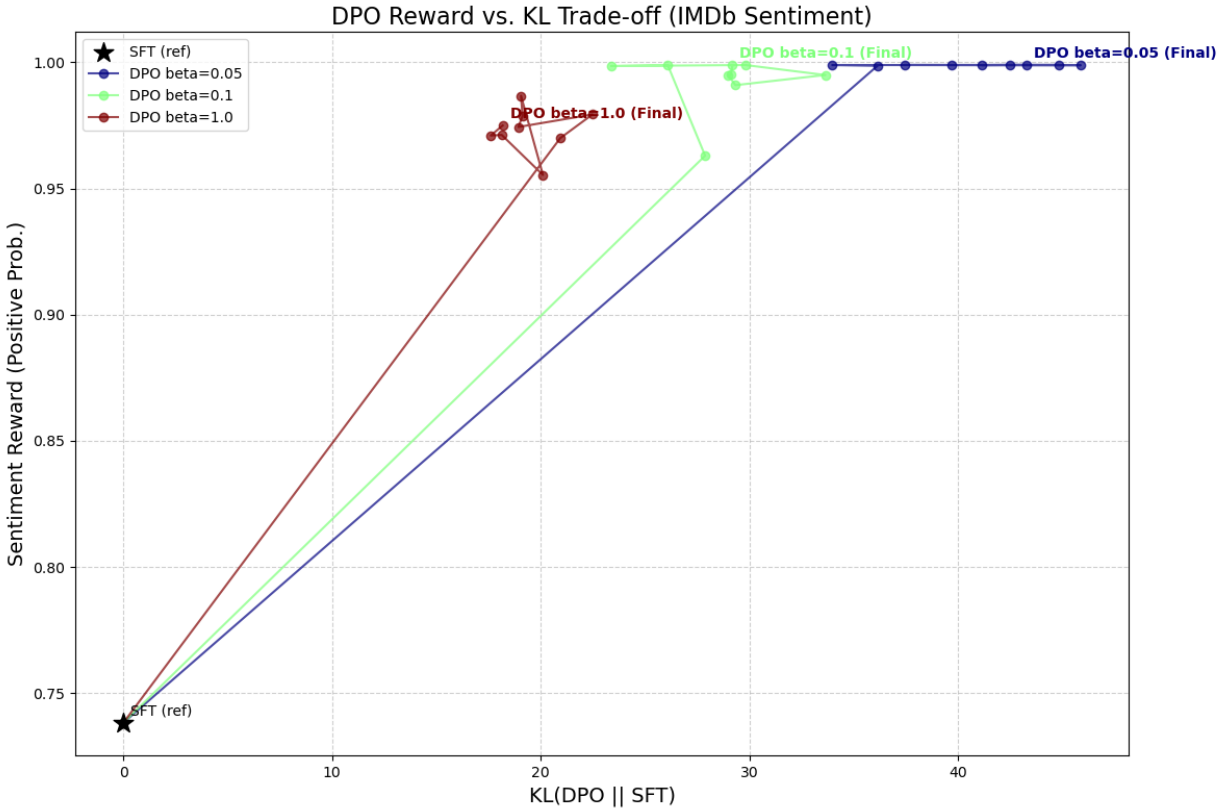

Figure 1: DPO Reward vs. KL Trade-off Plot

8 Model Sample Generations

8.1 SFT

Prompt: "I love"
Generation: "I love this show. I was a little skeptical at first, but I now know why. I have watched the entire first season, and I'm just amazed at the way this show is filmed and developed. All the performances are excellent, and the writing is spot on. The 'wonderland' is just an amazing place to watch this show. The only thing that I didn't love was the writing. I thought the characters were too simple. It was very similar to the way sitcoms are written. However, I think that the writers are so talented that they can come up with a new way to portray a certain character. I could probably have done with more of my favorite characters, but I still think this show is great."

Prompt: "Worth the entertainment value of a"
Generation: "Worth the entertainment value of a Saturday afternoon (especially with a few friends over) or a half-day (especially with a few friends over) and a few beers. 'The Man in the Moon' is the kind of film that could be made by anybody, with a camera. And it's surprisingly well-made. The acting is good, the dialogue is good, and the story is good. And the cinematography is good. The film is set in the mid-1980s, and is set in a small, rural area of Pennsylvania. It's one of those rural films that you could watch every day of the week. I saw it last night, and it was really well done. It's very similar to the movie 'The Hills Have Eyes' – a very good film, and one that probably deserves more attention than it has gotten."

Prompt: "its a"
Generation: "its a classic. all of the actors who played the lead roles are outstanding. i loved this movie. the film was very different from anything i've ever seen. the direction was fantastic. the music was great. i also liked the way that the actors were portraying the characters. the movie is very interesting and fun. it's also a very good horror movie that is very realistic."

8.2 DPO Beta = 0.05

Prompt: "I love"
Generation: "I love this movie, and I have seen it many times. It's a great movie for all ages. It's a great movie for all ages, and I think that it's a great movie for all ages, and it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it's a great movie for all ages, and I think that it..."

Prompt: "Worth the entertainment value of a"
Generation: "Worth the entertainment value of a great cast. A very good movie, and I highly recommend it. I think this movie is very well done, and I would recommend it to anyone who likes the genre. I think it is a very good movie for all ages. I think this movie is very well done, and I would recommend it to anyone who likes the genre. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages. I think it is a very good movie for all ages."

Prompt: "its a"
Generation: "its a great film for all ages. The film is an excellent portrayal of the lives of a family and the lives of the family members. It is a great story and a great movie for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all

ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The film is a great film for all ages. The...”

8.3 DPO Beta = 0.1

Prompt: “I love”
Generation: “I love the movie. The movie is well acted, the story is well told, and the actors are great. The movie was very well done and the actors are all very good. The movie is a great movie.”

Prompt: “Worth the entertainment value of a”
Generation: “Worth the entertainment value of a good movie, it’s a great movie to watch with your friends. The acting is excellent, and the movie is a good way to watch. The movie is very well done, and I recommend it to anyone who likes a good movie.”

Prompt: “its a”
Generation: “its a great film for those of all ages, and I’m glad I saw it. I’m very impressed with the acting, the direction, the cinematography, the overall quality of the film. I’m very well done, and I’m glad I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased I saw it. I’m very pleased...”

8.4 DPO Beta = 1.0

Prompt: “I love”
Generation: “I love this show. The acting is great, the story is great, the plot is great, the characters are great, the humor is great, the characters are great, the acting is great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot is great, the characters are great, the plot...”

Prompt: “Worth the entertainment value of a”
Generation: “Worth the entertainment value of a show that is so funny and charming. The show is also a reminder of how to make the best of the best of the best. I’m also a fan of the show. I also love the show. I also love the show. I’m a fan of the show. I also love the show. I’m a fan of the show. I also love the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan of the show. I’m a fan...”

Prompt: “its a”
Generation: “its a good movie that is very well acted. The plot is very well done and the story is very well told. The characters are well developed and the plot is very well written. The editing is excellent. The acting is excellent. The production is excellent.”

9 Appendix

9.1 Links

Project Link: Weights & Biases Project
Report Link: Weights & Biases Report