Assignment 3

1. Baseline Evaluation (15%)

(Q1.1)

Final Average Reward: 0.6298

(Q1.2)

Example text #1: Young Elijah Wood and Joseph Mazzello are outstanding performers in the film, and they are also the best performers in the film.
Example text #2: This is one of the best sequels around and a great sequel to the original.
Example text #3: That's what me and my friends kept asking each other about. I was like, 'What's the best way to do this?' And I was
Example text #4: "Mame" is a disgrace to many things.
Example text #5: The Tender Hook, or, Who Killed The President?

Sample 1-2 are positive, samples 3 and 5 do not have a clear sentiment, and example 4 has negative sentiments. The overall quality of the generation is pretty natural, but it seems to have a hard time generate clear positive sentiments.

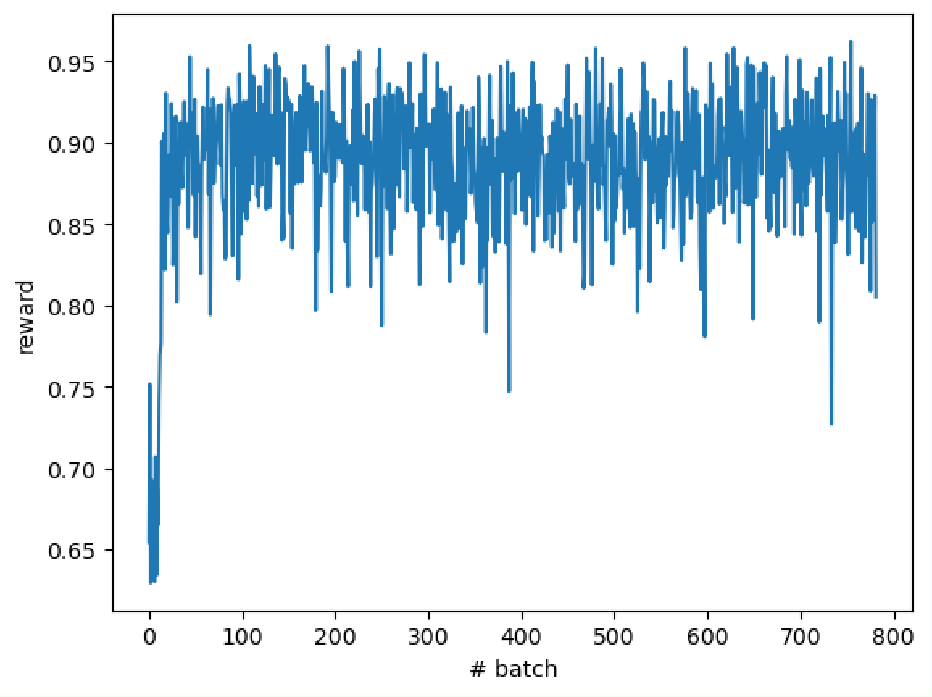
(Q1.3)

The loop variable is the progress bar which uses tqdm. It keeps track of the evaluation progress as it iterates through the test batches and displays the current average reward.

2. REINFORCE Implementation (20%)

(Q2.1)

Final Average Reward: 0.8898



(Q2.2)

Example text #1: Young Elijah Wood and Joseph Mazzello are outstanding, and I'm pleased to see the world, and I'm pleased to see the world, and
Example text #2: This is one of the best sequels around and a world, and I'm pleased to see the world, and I'm pleased to see the world,
Example text #3: That's what me and my friends kept asking each, and I'm pleased to see the world, and I'm pleased to see the world, and
Example text #4: "Mame" is a disgrace to many things, and Iḿ pleased to see the world, and Iḿ pleased to see the world, and
Example text #5: The Tender Hook, or, Who Killed The world, and I'm pleased to see the world, and I'm pleased to see the world,

The generations are more positive, but they achieved the positive sentiment by repeating the same positive phrases over and over again, which is not natural.

(Q2.3)

The model tends to reinforce over the same sequence of tokens once they receive a positive reward, and fails to give diverse generations. The model also fails to generate coherent sentences, all the example texts are nonsensical.

This is likely due to lack of regularization. The model exploit the reward signal to achieve high reward by repeating the same sequence of tokens and ignoring the fluency and coherence of the generated text (reward hacking).

(Q2.4)

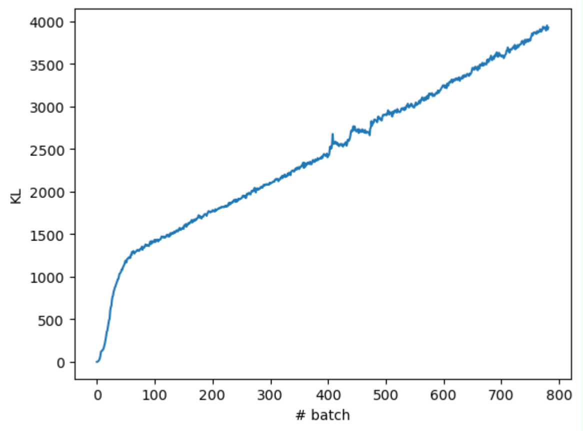
reset_model_optimizer() is used to reinitialize the model and optimizer, which is to make sure that each training run starts with the same initial conditions without any learning from previous runs.

(Q2.5)

The shape of log_probs is torch.Size([32, 30, 50257]), where 32 is the batch size, 30 is the sequence length (max_length=seed_token_length+max_new_tokens=10 + 20 = 30), and 50257 is the vocabulary size.

3. Regularization (15%)

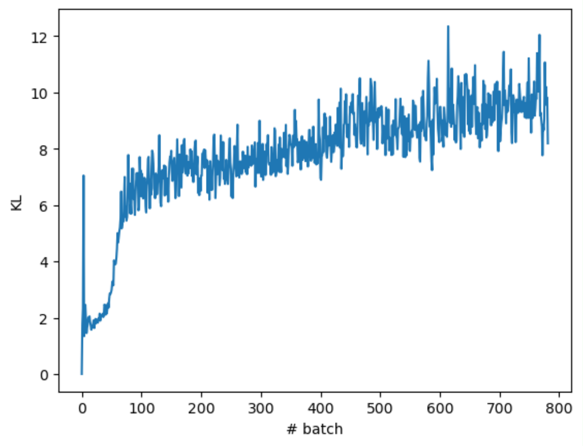
(Q3.1)

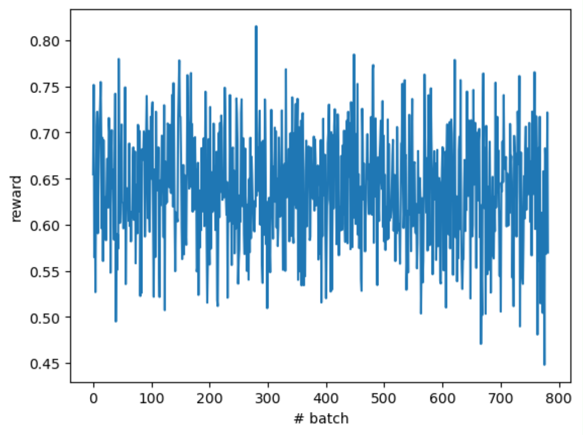


The KL-divergence increases over time as training progresses.
This is undesirable because, as the KL-divergence increases, the model is drifting away from the pre-trained distribution (might be an indicator of catastrophic forgetting), and the generated text might lose fluency, coherence, or other desirable properties that the model learned from pre-training.

(Q3.2)

A high α ($\alpha = 10$): Final Average Reward: 0.6297



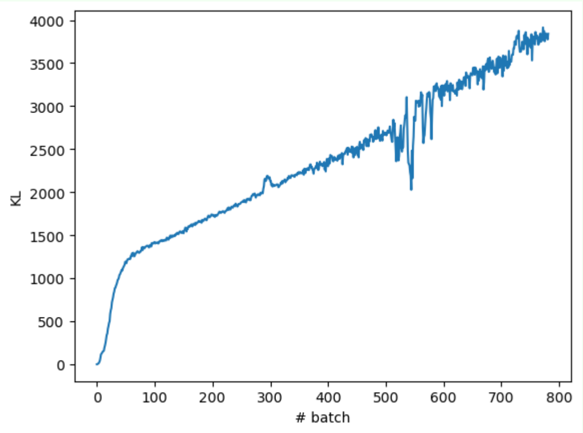


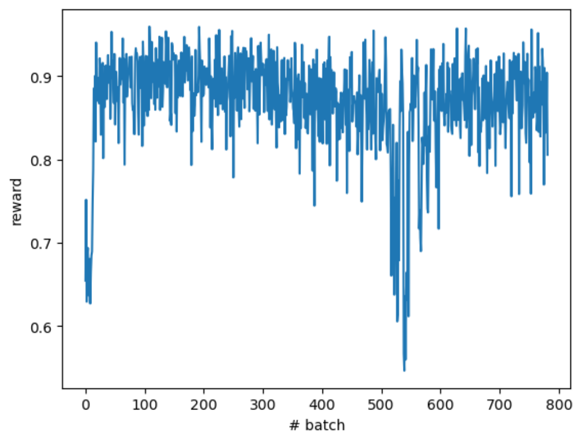
(Q3.2 contd.)

Example texts:
1. Young Elijah Wood and Joseph Mazzello are outstanding performers. They are the best performers in the world. They are the best performers in the world.
2. This is one of the best sequels around and a must-have for any indie game. It's a must-have for any indie game
3. That's what me and my friends kept asking each other about. I'm a big fan of the show, and I'm a big fan of the
4. ”Mame” is a disgrace to many things.
5. The Tender Hook, or, Who Killed The Man, is a series of short stories about the life of a man who was killed by a man

Comments: The quality of the samples are similar to the pretrained model. Sample 1,2, and 5 are positive, sample 4 is negative, and sample 3 is neutral.

A low α ($\alpha = 0.001$): Final Average Reward: 0.8756





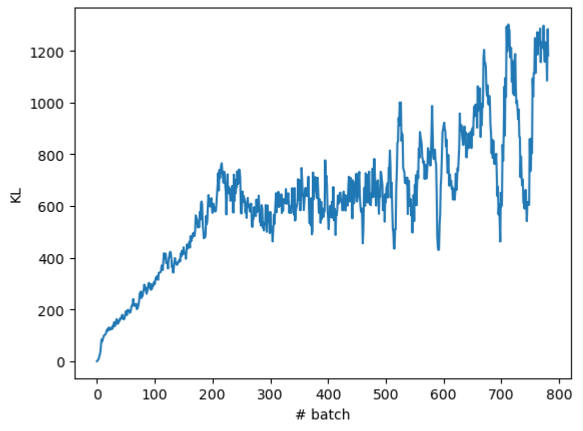
Example texts:
1. Young Elijah Wood and Joseph Mazzello are outstanding world, and I'm pleased to see the world, and I'm pleased to see the world,
2. This is one of the best sequels around and a, and I'm pleased to see the world, and I'm pleased to see the world, and
3. That's what me and my friends kept asking each world, and I'm pleased to see the world, and I'm pleased to see the world,
4. ”Mame” is a disgrace to many things world, and Iḿ pleased to see the world, and Iḿ pleased to see the world,
5. The Tender Hook, or, Who Killed The world, and I'm pleased to see the world, and I'm pleased to see the world,

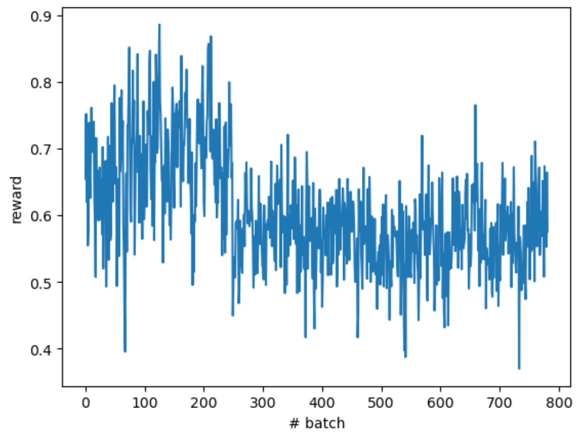
Comments: The quality of the generations are similar to the un-regularized model. The generation are all positive, but they are not natural and have many repetitions.

(Q3.2 contd.)

I tried 20 different α values, but wasn't able to find a good value that get a ≥ 0.8 reward while getting natural text. However, I do found a general trend of high α values tend to perform more similar to the original pre-trained model, but the reward is lower, and low α values tend to perform more similar to the un-regularized model, but the reward is higher.

The following is the best model I found which generates reasonably natural texts with pretty good amount of positive sentiments (judged by me):
Final Average Reward: 0.6869
$\alpha = 0.5$





Example texts:
1. Young Elijah Wood and Joseph Mazzello are outstanding musicians, and they're still getting closer to the music of the 1970s. They're still getting
2. This is one of the best sequels around and a lot of them, but I'm not sure why they're so good. I think they're going
3. That's what me and my friends kept asking each other about their experiences with the game. I've always been fascinated by the game, but I've
4. "Mame" is a disgrace to many things: The Simpsons, which is a great example of how badly flawed the Simpsons were. But thereś
5. The Tender Hook, or, Who Killed The Beatles, is a fascinating collection of modern music. It is a fascinating collection of music, and it

Comments: The quality of the generations lie in between the high and low α models. Sample 1, 2, 3 and 5 are positive, sample 4 is negative (but it did try to use the word "great" which is generally a positive word). The generations are also much more natural and coherent than the high α models, but still not as good as the original pre-trained model.

(Q3.3)

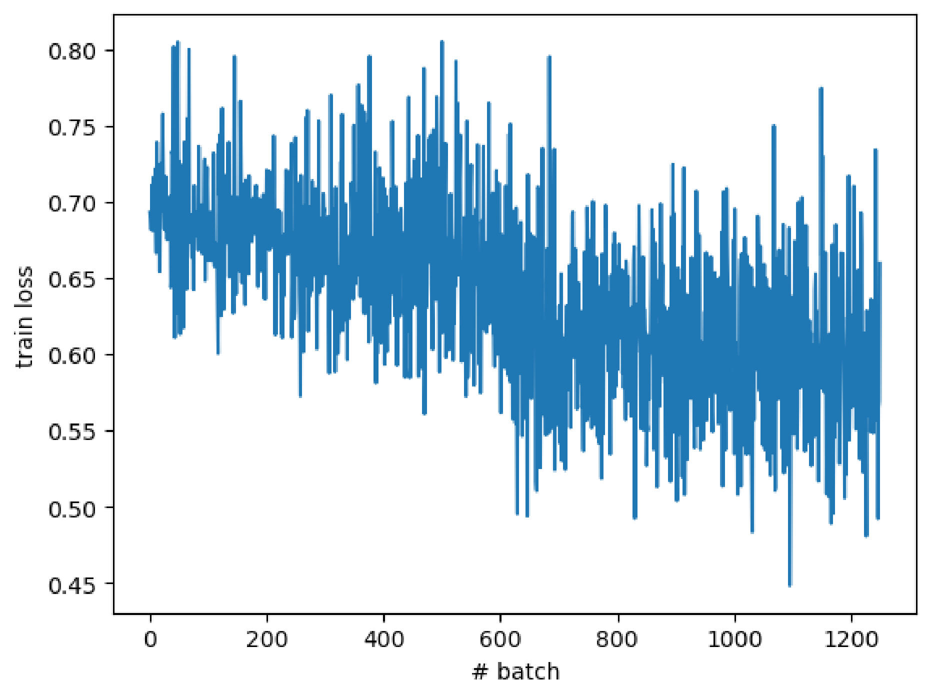
One limitation is that both model have high variance in loss and reward, which makes it hard to tune the hyperparameters and the the training is very unstable (might be because the reward signal is sparse). One possible solution is to use a baseline function to normalize the rewards or scales the rewards towards a center value, which can reduce the variance and make the training more stable.

(Q3.4)

The REINFORCE loss is computed over the 20 tokens (tokens 11...30) generated by the model in response to the seed. The goal of the model is to maximize the expected reward by generating the 20 tokens with the given seed. Since first 10 tokens is the fixed seed (i.e. not part of the model's output), so the loss is not computed over the first 10 tokens.

4. DPO Implementation (35%)

(Q4.1)



Average Loss on Test Set: 0.6552

(Q4.2)

Final Average Reward: 0.6466
The reward for the DPO model is 0.25 lower than the REINFORCE model. One possible reason is that the DPO model is more stable and less likely to exploit the reward signal to achieve high reward by repeating the same sequence of tokens and ignoring the fluency of the generated text (reward hacking). Another possible reason is that because of the small amount of data and limitation of the pretrained model, the DPO model is not able to learn a good policy to generate high reward text using the given hyperparameters.

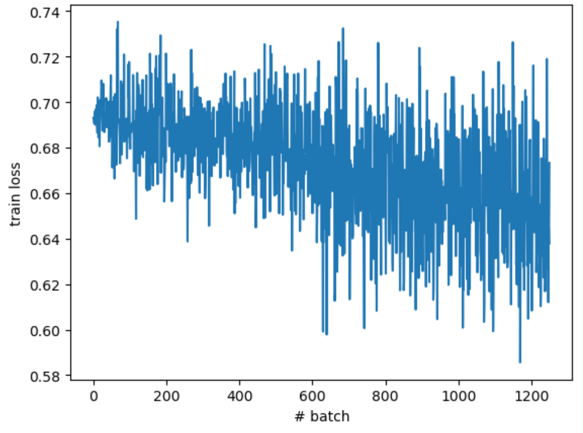
(Q4.3)

Example texts:
1. Young Elijah Wood and Joseph Mazzello are outstanding performers in the film, and they are also the best performers in the film.
2. This is one of the best sequels around and a great sequel to the original.
3. That's what me and my friends kept asking each other about. I was like, 'What's the best way to do this?' And I was
4. "Mame" is a disgrace to many things.
5. The Tender Hook, or, Who Killed The President?

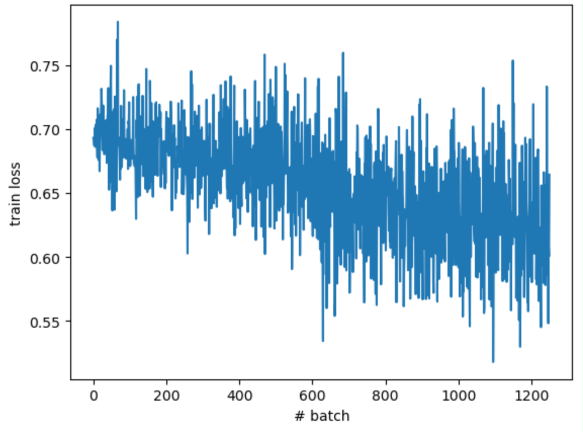
Comments: Sample 1 and 2 are positive, sample 4 is negative, and sample 3 and 5 do not have a clear sentiment. The generation are not more positive than the original pre-trained model. Actually, the generations are the same as the original pre-trained model.

(Q4.4)

Experiment 1 ($\beta = 0.1$): Test Loss: 0.6764, Final Average Reward: 0.6526

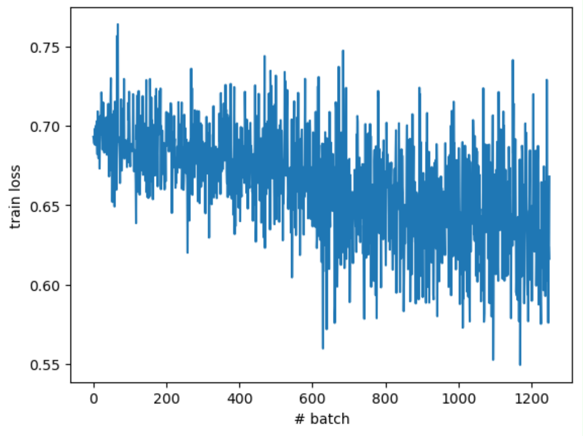


Experiment 2 ($\beta = 0.25$): Test Loss: 0.6613, Final Average Reward: 0.6500



(Q4.4 contd.)

Experiment 3 ($\beta = 0.175$): Test Loss: 0.6670, Final Average Reward: 0.6513



Comments: The best setup in terms of test loss among the three experiments is $\beta = 0.25$, but the reward is the lowest among the three. As we can see from the data of the three experiments, the higher the β, the lower the test loss, but the lower the reward. This is probably because the higher the β, the more the model is regularized and the less likely the model is to exploit the reward signal.

(Q4.5)

σ is used to squash the log probability ratios into the range between 0 and 1 which can help the loss not diverge when the differences between log probs. It can help regularize the model and make the training more stable.

(Q4.6)

The main differences/advantages of DPO are that DPO optimizes for human preferences while avoiding reinforcement learning, and DPO doesn't require external reward model (the DPO model is the reward model). Also, DPO is more stable and less likely to exploit the reward, because the examples are weighted by how much higher the implicit reward model rates the dispreferred completions, scaled by β.

5. GPT-4 Capability Forecast (15 %)

(Q5.1)

Accuracy: 64.29%
Log loss: 1.056

(Q5.2)

1. I was expecting GPT-4 could do things that seemed straightforward for me, like win a tic-tac-toe game, draw "hello" in ascii, and the birthday question but it couldn't do any of that. I wan't expecting it to solve some complex questions that took me a while to understand (the car cluster problem, card deck problem, etc.), but it did solve them.

2. One pattern is that the model struggles with the tasks that don't present clear semantic information in the question or answer. For example, interpreting a tic-tac-toe board represented by characters is challenging for GPT-4 because it can't visualize the board's layout, and without the visual understanding of the layout, it is hard to generate the correct answer. Similarly, without being able to have the visual information about the ascii output, it's hard for the model to judge the correctness of the output. One potential reason for this could be the lack of semantic continuity in the question or answer. Since the model generate the next token base on the previous tokens, it's hard for the model to generate the correct answer without clear semantic information in the tokens it generates.

3. I think I'm less confident about GPT-4 to solve my task after taking the quiz. Because the correctness of the answer from GPT-4 is quite unpredictable, and GPT-4 is also overly confident about its wrong answers.

4. I envision future models to be multi-sensory, capable of learning and decision-making based on diverse inputs like light, sound, text, smell, taste, and touch. It would be cool if they can receive information from the environment like a human. It is important because currently the model is trained on a large amount of human generated data, which can be biased or not representative of the real world. If the model can observe the world objectively and make decisions based on these observations, it might be able to produce better and more accurate outputs. To make that happen, future GPT-x models would need many many more GPUs or a new type of hardware to handle the increased complexity and amount of input information, and potentially new paradigms of multi-modal learning.

Acknowledgement

This assignment was completed by Sebastian Liu. ChatGPT 3.5 were used as a human collaborator to ask conceptual questions about RL, DPO, expected correlations between different hyperparameters, the internals of different PyTorch functions, and how to train the models locally using Mac's gpu.