CS 542: Reward Modelling

Q: How do the mean winning and losing rewards change during training? How does the change in these values compare to the change in the respective rewards for Reponses A and B of the specific randomly-chosen validation sample?

A: Over 10 epochs, the model successfully learned to distinguish between good and bad summaries.

- **Mean Winning Reward:** Increased consistently, starting at **0.139** (Epoch 1) and plateauing around **3.623** (Epoch 10). This indicates the model learned to assign high rewardd scores to correct summaries.
- **Mean Losing Reward:** Decreased consistently, dropping from **-0.063** (Epoch 1) to **-2.859** (Epoch 10). This shows the model learned to penalize incorrect summaries.

Q: Compare the actual and predicted winning and losing responses from the 10 randomly-chosen test samples that print out after model training. What do you observe about the assigned rewards for the correctly predicted samples vs. the incorrectly predicted samples? Consider both the raw pointwise reward values and the differences between winning and losing rewards.

A: The most striking difference between the correct and incorrect predictions is the magnitude of the difference between the winner's score and the loser's score. For correctly Predicted Samples, the model is highly confident. In clear cases, the margins are large.

For Incorrectly Predicted Samples, the model seems confused. The margins are narrow, often less than 1.0 point. When the model sees both options as equally good (or equally bad), making the final choice leads to incorrect prediction.

Q: Look at the text of the winning and losing responses from the 10 randomly-chosen test samples, particularly the 5 incorrectly labeled samples. What do you observe about the texts of the responses? What, if anything, may have made the incorrectly labeled samples hard to label?

A: When the semantic gap between the two options is unambiguous, the model effectively pushes the rewards to their maximum and minimum extremes. For example, in the "Hong Kong Cookies" prompt, Response A (regarding African millionaires) is entirely unrelated to the prompt context. The model correctly identified this, assigning the relevant Response B a high reward (**5.032**) and the irrelevant Response A a negative reward (**-3.710**). This demonstrates the model's ability to confidently distinguish relevant summaries from unrelated text.

In case of incorrectly labelled samples, I found semantic twin problem. That means two responses are close in similarity. For example, the Americal pie prompt’s responses. Because the negative sampling used cosine similarity, it selected a summary from a different article covering the exact same event. The model correctly identified both summaries as excellent matches (scores > 5.0). It just assigned a bit more reward to the wrong response. It is not a failure of sematic understanding, rather a failure to choose more appropriate wrong response.

In cases where neither summary aligned with the prompt, the model correctly recognized the irrelevance of both options. For instance, in the "Boston Marathon" example, the prompt described the race's atmosphere, while both responses focused on the legal trial of the bomber. The model assigned highly negative rewards (~-4.2) to both, indicating that neither was a suitable summary. The prediction failed simply because the model was forced to make an arbitrary choice between two "bad" options.