Week 5 Assignment

Generative AI

Saarland University – Winter Semester 2024/25

Martínez
7057573
cama00005@stud.uni-saarland.de

1 Exercise Assignment: E1

Table 1 was constructed using the Bradley-Terry model, which given a prompt x_p and two responses y_1 and y_2, models the probability that $y_1 1$ is preferred over y_2:

$$\Pr(y_1 \succ y_2 | x_p) = \sigma(r(x_p, y_1) - r(x_p, y_2)) = \frac{1}{1 + e^{-(r(x_p,y_1)-r(x_p,y_2))}} \tag{1}$$

Table 1: Pairwise comparisons using the Bradley-Terry model in Eq. (1).

Subpopulation	**Pairwise comparisons**		
	$\Pr(y_1 \succ y_2)$	$\Pr(y_1 \succ y_3)$	$\Pr(y_2 \succ y_3)$
1	0.5	0.27	0.27
2	0.73	0.5	0.27
3	0.88	0.95	0.73
4	0.12	0.047	0.27

Based on Table 1, we can determine the most preferred within each subpopulation:

$\Rightarrow$ **Subpopulation 1:** y_3 is the most preferred response.

$\Rightarrow$ **Subpopulation 2:** y_1 and y_3 are equally preferred.

$\Rightarrow$ **Subpopulation 3:** y_1 is the most preferred response.

$\Rightarrow$ **Subpopulation 4:** y_3 is the most preferred response.

2 Exercise Assignment: E3

For RLHF, the loss function is given by:

$$\mathcal{L}_R(r_\phi) = -\mathbb{E}_{(x_p,y_w,y_l)\sim\mathcal{D}_p}\left[\log \sigma(r_\phi(x_p, y_w) - r_\phi(x_p, y_l))\right] \tag{2}$$

Plugging in the values, we get:

$$\begin{aligned}\mathcal{L}_R(r_\phi) &= -\mathbb{E}_{(x_p,y_w,y_l)\sim\mathcal{D}_p}\left[\log \sigma(0.5 - 0.8)\right] \\ &= -\log \sigma(-0.3) = -\log \frac{1}{1 + e^{-(-0.3)}} \approx 0.85\end{aligned}$$

On the other hand, the loss function for DPO is given by:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{SFT}) = -\mathbb{E}_{(x_p,y_w,y_l)\sim\mathcal{D}_p}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x_p)}{\pi_{SFT}(y_w|x_p)} - \beta \log \frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)}\right)\right] \tag{3}$$

The DPO loss for the given datapoint is thus:

$$\begin{aligned}\mathcal{L}_{DPO}(\pi_\theta;\pi_{SFT}) &= -\mathbb{E}_{(x_p,y_w,y_l)\sim\mathcal{D}_p}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x_p)}{\pi_{SFT}(y_w|x_p)}-\beta\log\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)}\right)\right]\\ &= -\mathbb{E}_{(x_p,y_w,y_l)\sim\mathcal{D}_p}\left[\log\sigma\left(2\log\frac{0.2}{0.2}-2\log\frac{0.3}{0.4}\right)\right]\\ &= -\log\sigma\left(-2\log\frac{0.3}{0.4}\right)\approx 0.44\end{aligned}$$

3 Exercise Assignment: E4

In the case of Reinforcement Learning from Human Feedback (RLHF):

- Looking at the reward values, we see that $r_\phi(x_p,y_l)=0.8>r_\phi(x_p,y_w)=0.5$, which is problematic since y_w is supposed to be the preferred response.
- When updating π_θ through RLHF, we aim to make the policy encouraged to increase the probability of generating responses that lead to higher rewards. This means that, in this case, we would push π_θ to generate more y_l-like responses since they have higher reward according to r_ϕ.
- This is obviously counter-productive since y_l is not the preferred response and highlights the failure of having poorly trained the reward model r_ϕ.

On the other hand, with Direct Preference Optimization (DPO):

- DPO directly uses the preference data without relying on the learned reward model, which is an advantage over RLHF in this case, where the reward model is not reliable.
- Through DPO, we aim to modify π_θ to better match the preference data while staying close to π_{SFT} through the maximum-likelihood objective under the KL constraint, $\max_\theta \mathcal{L}_{DPO}(\pi_\theta;\pi_{SFT})$, where $\mathcal{L}_{DPO}$ is defined in Eq. 3 [1].
- Since y_w is the preferred response, DPO would encourage increasing $\pi_\theta(y_w|x_p)$ from its current value of 0.2 and decreasing $\pi_\theta(y_l|x_p)$ from its current value of 0.3, while not deviating too far from π_{SFT}, moderated by the temperature parameter $\beta=2$.

In this specific scenario, DPO would likely lead to better alignment with human preferences compared to RLHF, since it's not led astray by the problematic reward model.

4 Exercise Assignment: E5

First of all, the original case with $\pi_{SFT}(y_l|x_p)=0.4$ and everything else being the same, has the following form for the DPO loss:

$$\begin{aligned}\mathcal{L}_{DPO}(\pi_\theta;\pi_{SFT}) &= -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x_p)}{\pi_{SFT}(y_w|x_p)}-\beta\log\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)}\right)\\ &= -\log\sigma\left(\beta\log\frac{0.2}{0.2}-\beta\log\frac{0.3}{0.4}\right)\\ &= -\log\sigma\left(\beta\log\frac{4}{3}\right)\end{aligned}$$

On the other hand, the modified case with $\pi_{SFT}(y_l|x_p)=0.2$ and everything else being the same, has the following form for the DPO loss:

$$\begin{aligned}\mathcal{L}_{DPO}(\pi_\theta;\pi_{SFT}) &= -\log\sigma\left(\beta\log\frac{0.2}{0.2}-\beta\log\frac{0.3}{0.2}\right)\\ &= -\log\sigma\left(-\beta\log\frac{3}{2}\right)\end{aligned}$$

From the previous expressions, we can make the following analysis:

- **Original case, $\boldsymbol{\pi_{SFT}(y_l|x_p) = 0.4}$**
  - Inside the sigmoid function $\sigma(\cdot)$, we have $\beta \log \frac{4}{3}$ which is positive, since $\frac{4}{3} > 1$ and $\log(x) > 0, \forall x > 1$.
  - As β increases, the argument becomes more positive, i.e., shifted to the right on the sigmoid curve.
  - This makes $\sigma(\cdot)$ approach 1 from below, **decreasing the loss**, since the $\log$ function monotically increases and is negative for values less than 1.
  - This means that the update on π_θ that the loss is encouraging aims to maintain the current ratio $\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)} < 1$.
- **Modified case, $\boldsymbol{\pi_{SFT}(y_l|x_p) = 0.2}$**
  - Inside the sigmoid function $\sigma(\cdot)$, we have $-\beta \log \frac{3}{2}$ which is negative, since the argument of the $\log$ function is greater than 1 and we have a negative sign in front[1].
  - As β increases, the argument becomes more negative, i.e., shifted to the left on the sigmoid curve.
  - This makes $\sigma(\cdot)$ approach 0, **increasing the loss**, contrary to the original case.
  - Thus, the loss more strongly penalizes the current ratio $\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)} > 1$.

From this analysis, we can make the following conclusions:

- When $\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)} < 1$ (meaning a favorable ratio for non-preferred responses) as in the original case, higher β reduces loss.
- When $\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)} > 1$ (meaning an unfavorable ratio for non-preferred responses), higher β increases loss, amplifying the learning signal from preference comparisons and leading to stronger policy updates. This is desirable when preferences disagree with the current behavior, as in the modified case. More specifically, the modified case with a ratio $\frac{\pi_\theta(y_l|x_p)}{\pi_{SFT}(y_l|x_p)} = \frac{0.3}{0.2} = 1.5$ indicates that the model assigns too much probability to the less preferred option and higher β values will more aggressively push for correcting this misalignment.

[1] Similar to the original case, we could say that $\mathcal{L}_{DPO}(\pi_\theta; \pi_{SFT}) = -\log \sigma\left(\beta \log \frac{2}{3}\right)$, arriving to the same conclusion that the argument is negative, since $\frac{2}{3} < 1$ and $\log(x) < 0, \forall x < 1$.

5 Individual Assignment: I3

Looking at the preference samples in `pref_data_completion_20samples.txt`, we can note that the `CHOSEN` completions are more concise, focused, and direct, in contrast to the `REJECTED` ones. The latter tend to be longer, include multiple speakers/characters, and often drift into lengthy dialogues or scene descriptions. Whilst the former typically complete the immediate thought or action without unnecessary elaboration and maintain the style of the original prompt.

Table 2: Qualitative pairwise comparison of the three completions generated by the Supervised fine-tuned Phi-1.5 model and the Preference-tuned Phi-1.5 model. Note that **Cpl.** is the completion id and the last column corresponds to whether the Preference-tuned model wins (W), ties (T), or loses (L) against the Supervised fine-tuned model.

PREFIX	Cpl.	Supervised fine-tuned Model	Preference-tuned Model	
"O, what a noble mind is here o'erthrown!"	A	Too long response about brotherhood	Concise "O, let's see"	W
	B	Too long response about seeing the King	Concise but unrelated "HOST"	T
	C	Too long response but more *dialogue-esque*	Concise, not particularly *Shakespeare-esque*	W
"look in thy glass and tell the face thou viewest"	A	Too long response	Simple "."	W
	B	Verbose dialogue	Simple "."	W
	C	Hallucinates a paragraph	Concise, coherent continuation	W
"Then let not winter's ragged hand deface"	A	Too long response	Focused completion	W
	B	Too long response	Concise "our joy"	W
	C	Hallucinates an unrelated story	Concise "our faces"	W
"Thou canst not see one wrinkle in my brow,"	A	Hallucinates a blog post	*Poetic* contradiction	W
	B	Too long response	Focused completion	W
	C	Too long response	Brief, relevant response	W
"Great lords, wise men ne'er sit and wail their loss,"	A	Too long response	Focused completion	W
	B	Too long response	Focused completion	W
	C	Too long response	Concise "but to stand on it"	W

Table 2 shows a qualitative pairwise comparison of completions generated by the Supervised fine-tuned Phi-1.5 model and the Preference-tuned Phi-1.5 model. There, we can see that the latter consistently produces completions that better align with the characteristics seen in the preference samples. In contrast to the former, it opts for *conciseness*, *relevance* and *style consistency*, even if it means only producing a simple period (".") as a completion. In summary, across all 15 comparisons, the Preference-tuned Model wins 14 times and ties 1 time against the Supervised fine-tuned Model.

6 Individual Assignment: I4

Table 3 shows a qualitative pairwise comparison of the three summaries generated by the Supervised fine-tuned GPT-J Model and the Preference-tuned GPT-J Model. Overall, the Preference-tuned Model tends to include more context and details from the original posts, but sometimes at the cost of accuracy or by adding unverified information. It wins in situations where additional context is helpful for understanding the complete situation (like financial planning questions). The Supervised fine-tuned Model generally produces more concise, focused summaries that stick closer to explicitly stated information.

In summary, across all 15 comparisons, the Preference-tuned Model wins 2 times, loses 10 times, and ties 3 times against the Supervised fine-tuned Model.

Table 3: Qualitative pairwise comparison of the three summaries generated by the Supervised fine-tuned GPT-J Model and the Preference-tuned GPT-J Model. Note that the last column corresponds to whether the Preference-tuned model wins (W), ties (T), or loses (L) against the Supervised fine-tuned model.

Post	**S.**	**Supervised fine-tuned Model**	**Preference-tuned Model**	
`TITLE: Decisions regarding a 401k Cash out`	A	Simple, focused on key question	Includes context about 401k details and alternative income sources	W
	B		Adds detail about 401k being "large" and is redundant in the end adding `"(401k)"`	L
	C		Adds specific 401k amount but it's still redundant	L
`TITLE: Race Report: First Marathon!`	A	Simple, focused on completion (no emotions)	Hallucinates irrelevant details about pictures	L
	B	Simple, but emotional response	Similar emotional response with slightly less detail	T
	C		Hallucinates about pictures and providing future updates	L
`TITLE: How can I mange this Job`	A	Simple summary of situation	Hallucinates claims about work/school history	L
	B		Extremely redundant and unprecise, and hallucinates working hours a day	L
	C	Simple summary of the situation but slightly redundant in the end	Hallucinates ADHD condition	L
`TITLE: 27 yr old planning on getting an apartment in July with my 20 yr d brother. How do I plan so We don't have to struggle?`	A	Simple summary of situation	More specific about financial needs, but slightly redundant	T
	B		Includes incorrect age information	L
	C	Comprehensive summary of the situation	Slightly more specific about financial situation	W
`TITLE: Is there a good solution to all the mass amount of usernames and passwords I need to remember for every website?`	A	Direct statement of the problem	Slightly more verbose and redundant	L
	B	Direct statement of the problem, and poses a question		L
	C	Direct statement of the problem	Direct statement of the problem	T

Acknowledgements

This week's slides and listed references.

References

[1] 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, 2024.