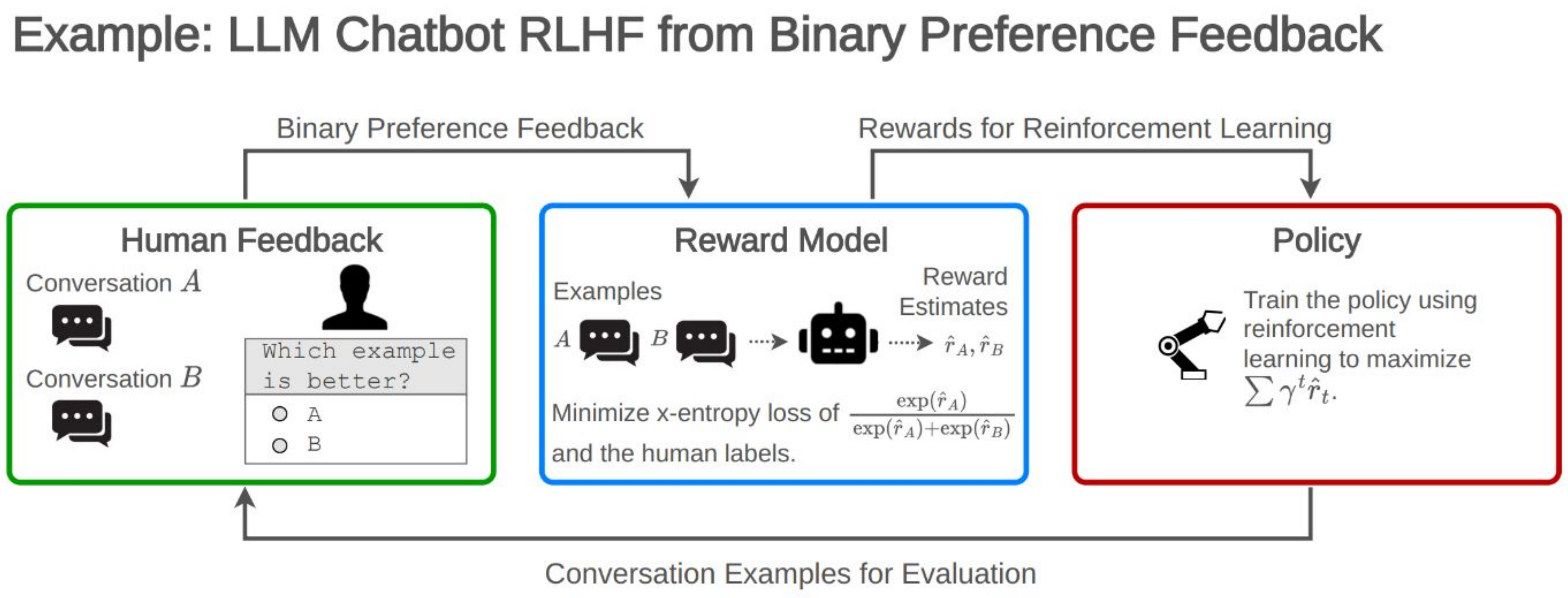
Extensions to RLHF : RLAIF and DPO

Ayush Agarwal MT22095
Janak Kapuriya MT22095
Kritarth Prasad 202084

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Background: RLHF

Reinforcement Learning from Human Feedback.



Phase 1. Pre-training for completion

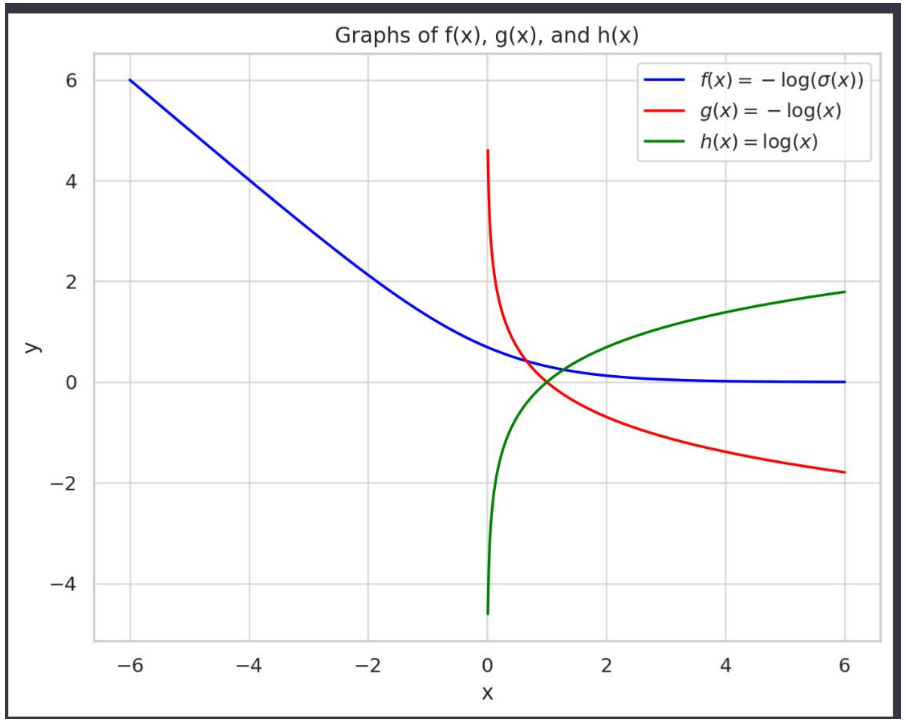
- LLM_ϕ: the language model being trained, parameterized by ϕ. The goal is to find ϕ for which the cross entropy loss is minimized.
- $[T_1, T_2, \ldots, T_V]$: vocabulary – the set of all unique tokens in the training data.
- V: the vocabulary size.
- $f(x)$: function mapping a token to its position in the vocab. If x is T_k in the vocab, $f(x) = k$.
- Given the sequence $(x_1, x_2, \ldots, x_n)$, we'll have n training samples:
  - Input: $x = (x_1, x_2, \ldots, x_{i-1})$
  - Ground truth: x_i
- For each training sample (x, x_i):
  - Let $k = f(x_i)$
  - Model's output: $LLM(x) = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_V]$. Note: $\sum_j \bar{y}_j = 1$
  - The loss value: $CE(x, x_i; \phi) = -\log \bar{y}_k$
- Goal: find ϕ to minimize the expected loss on all training samples. $CE(\phi) = -E_x \log \bar{y}_k$

3.1 Training a Reward Model

- r_θ: the reward model being trained, parameterized by θ. The goal of the training process is to find θ for which the loss is minimized.
- Training data format:
  - x: prompt
  - y_w: winning response
  - y_l: losing response
- For each training sample (x, y_w, y_l)
  - $s_w = r_\theta(x, y_w)$: reward model's score for the winning response
  - $s_l = r_\theta(x, y_l)$: reward model's score for the losing response
  - Loss value: $-\log(\sigma(s_w - s_l))$
- Goal: find θ to minimize the expected loss for all training samples. $-E_x \log(\sigma(s_w - s_l))$

To get more intuition how this loss function works, let's visualize it.

Let $d = s_w - s_l$. Here's the graph for $f(d) = -\log(\sigma(d))$. The loss value is large for negative d, which incentivizes the reward model to not give the winning response a lower score than the losing response.



3.2 Policy Optimization

- RM: the reward model obtained from phase 3.1.
- LLM^{SFT}: the supervised finetuned model obtained from phase 2.
  - Given a prompt x, it outputs a distribution of responses.
  - In the InstructGPT paper, LLM^{SFT} is represented as π^{SFT}.
- LLM^{RL}_{ϕ}: the model being trained with reinforcement learning, parameterized by ϕ.
  - The goal is to find ϕ to maximize the score according to the RM.
  - Given a prompt x, it outputs a distribution of responses.
  - In the InstructGPT paper, LLM^{RL}_{ϕ} is represented as π^{RL}_{ϕ}.
- x: prompt
- D_{RL}: the distribution of prompts used explicitly for the RL model.
- $D_{pretrain}$: the distribution of the training data for the pretrain model.

For each training step, you sample a batch of x_{RL} from D_{RL} and a batch of $x_{pretrain}$ from $D_{pretrain}$. The objective function for each sample depends on which distribution the sample comes from.

1. For each x_{RL}, we use LLM^{RL}_{ϕ} to sample a response: $y \sim LLM^{RL}_{\phi}(x_{RL})$. The objective is computed as follows. Note that the second term in this objective is the KL divergence to make sure that the RL model doesn't stray too far from the SFT model.

$$\text{objective}_1(x_{RL}, y; \phi) = RM(x_{RL}, y) - \beta \log \frac{LLM^{RL}_{\phi}(y|x)}{LLM^{SFT}(y|x)}$$

2. For each $x_{pretrain}$, the objective is computed as follows. Intuitively, this objective is to make sure that the RL model doesn't perform worse on text completion – the task the pretrained model was optimized for.

$$\text{objective}_2(x_{pretrain}; \phi) = \gamma \log LLM^{RL}_{\phi}(x_{pretrain})$$

The final objective is the sum of the expectation of two objectives above. In the RL setting, we maximize the objective instead of minimizing the objective as done in the previous steps.

$$\text{objective}(\phi) = E_{x \sim D_{RL}} E_{y \sim LLM^{RL}_{\phi}(x)} [RM(x, y) - \beta \log \frac{LLM^{RL}_{\phi}(y|x)}{LLM^{SFT}(y|x)}] + \gamma E_{x \sim D_{pretrain}} \log LLM^{RL}_{\phi}(x)$$

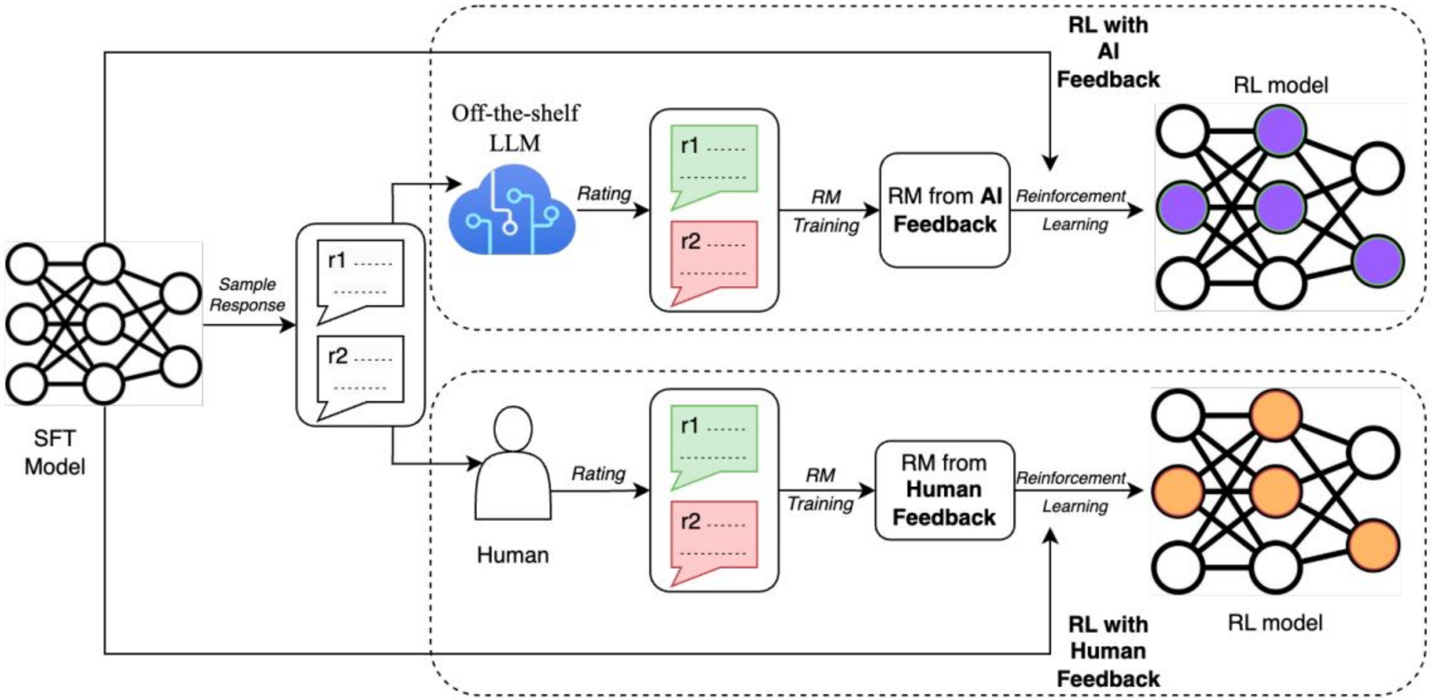
Note:

The notation used is slightly different from the notation used in the InstructGPT paper, as I find the notation here a bit more explicit, but they both refer to the exact same objective function.

$$\begin{aligned} \text{objective}(\phi) = & E_{(x,y) \sim D_{\pi^{\mathrm{RL}}_{\phi}}} \left[r_\theta(x, y) - \beta \log \left(\pi^{\mathrm{RL}}_{\phi}(y \mid x) / \pi^{\mathrm{SFT}}(y \mid x) \right) \right] + \\ & \gamma E_{x \sim D_{\text{pretrain}}} \left[\log(\pi^{\mathrm{RL}}_{\phi}(x)) \right] \end{aligned} \quad (2)$$

RLHF vs RLAIF

- Using **“Off-the-shelf”** LLM (not trained) to generate preference labels, instead of human feedback.
- Across different tasks, RLAIF achieves comparable or superior performance to RLHF.



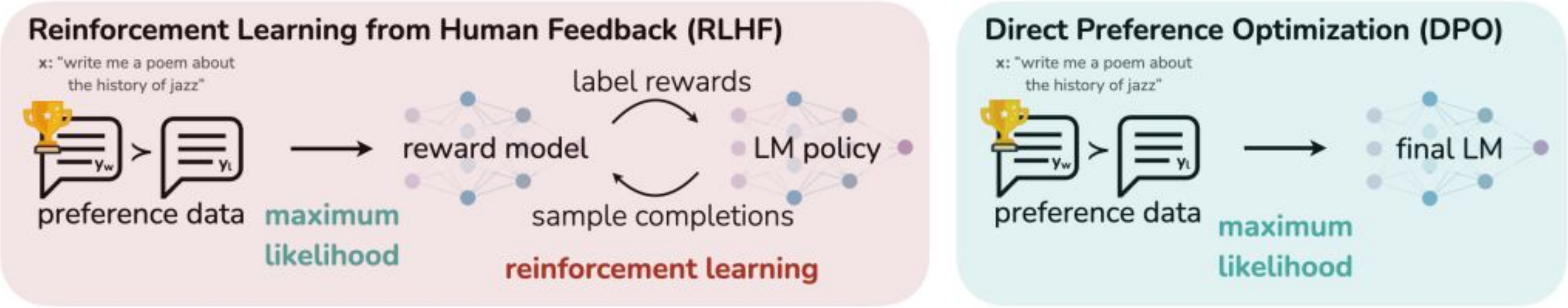
Prompt for Preference Labeling with LLMs (AI Feedback)

Example Prompt to generate AI Feedback (Reward Model) - Which response in better?

Prompt: In the context of a Human Assistant dataset, the relevance of responses is crucial for ensuring effective assistance. To evaluate response relevance, we focus on several key factors: coherence, accuracy, coverage, and overall quality.\n\n**Coherence:** This factor assesses whether the response is easily understandable and free of English errors. A coherent response is clear and straightforward, making it easy for the user to grasp the intended message. It's more important for the response to be understandable than completely free of grammatical errors.\n\n**Accuracy**: This factor evaluates the factual correctness of the response. An accurate response aligns with the facts and details of the user's query, avoiding misinformation or confusion.\n\n**Coverage**: This factor examines how well the response addresses the key aspects of the user's query. A response with good coverage includes all the crucial information needed to fully understand and address the user's needs. It should not leave out significant details that are vital for the user's understanding.\n\n**Overall Quality:** This factor considers the response's effectiveness in representing the user's query and providing a helpful answer. It encompasses coherence, accuracy, and coverage, as well as other relevant aspects. A high-quality response not only meets these criteria but also enhances the user's experience by being informative, relevant, and user-friendly.\n\nGiven a user's query and two potential responses, your task is to determine which response better adheres to these criteria of coherence, accuracy, coverage, and overall quality.\n\n**Query** - Human: How do scientists test which medicines are effective?\n\nAssistant:\n**Response 1** - One way that scientists test whether a medicine is effective is to perform a clinical trial. In a clinical trial, scientists will randomly assign a group of patients to receive the medicine being tested, and a group of patients to receive a placebo (a dummy pill with no therapeutic effect). The scientists will then compare the outcomes for the two groups.\n**Response 2** - Scientists test which medicines are effective by conducting clinical trials, where the medicine is tested on a group of people with the condition or illness it is meant to treat. The trials are ethically designed to ensure the safety of participants and that the results are scientifically accurate. The results can then be analyzed to determine the effectiveness of the medicine.\n\n**Evaluate each response based on coherence, accuracy, coverage, and overall quality, and explain which one is more relevant and effective in the context of the Human Assistant dataset.\n\nRationale:**

"rank": [2]

DPO – Direct Preference Optimization



Policy objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

DPO – Salient Features

1.**What -** This method obviates the need for an explicit reward model, as the LLM itself acts as a reward model. The key advantage is that it's a straightforward loss function optimized using backpropagation.

2. **Impact -** The stability, performance, and computational efficiency of DPO are significant improvements over traditional methods. It eliminates the need for sampling from the LM during fine-tuning, fitting a separate reward model, or extensive hyperparameter tuning.

3. **How -** Our key insight is to leverage an analytical mapping from reward functions to optimal policies, which enables us to transform a loss function over reward functions into a loss function over policies. This change-of-variables approach allows us to skip the explicit reward modeling step, while still optimizing under existing models of human preferences"

Comparison of PPO vs DPO Responses: Summarization Task

"**Reddit Post**": My side of this story is as the customer. Sort of. The other day, I was riding shotgun with one of my buddies on the way home from one of our other friend's houses. It was getting close to dinner time, so we decide to get some fast food.\n\nMy friend pulls into a restaurant and drives up to the little intercom thing where the employee's can take your order. Without even waiting for the voice to ask, \"How can I help you today?\" or whatever, my friend says, without any hesitation,\n\n\"Let me get two McChickens, a big-ass sweet tea, and McFlurry, BOUNCE.\"\n\nAnd then immediately drives up to the first window. I'm cracking up at this point, and we are both watching the guy through the window quickly looking around and trying to figure out what happened, as he was somewhat unaware an order had even been placed.\n\nOnce the employee saw the kid with the wife beater and backwards snapback hat on staring expectantly at him through the window, he started laughing and then took his order again.\n\nNot exactly the most exciting or rewarding story, but I was laughing uncontrollably throughout the majority of the whole ordeal.\n\nSo, what are some of your drive-thru stories?

"**PPO**": My friend ordered a McDonald's while driving and someone else came up to the window asking if they could get something to eat. The person didn't know what was going on, and it was pretty funny. What are some of your best drive-thru stories at a fast food joint? Either as the employee or the customer?My side of this story is as the customer. Sort of. The other day, I was riding shotgun with one of my buddies on the way home from one of our other friend's houses. It was getting close to dinner time, so we decide to get some fast food.\n\nMy friend pulled into a restaurant and drove up to the little intercom thing where the employee's can take your order. Without even waiting for the voice to ask, \"How can I help you today?\" or whatever, my friend said, without any hesitation,\n\n\"Let me get two McChickens, a big-ass sweet tea, and McFlurry, BOUNCE.\"\n\nAnd then immediately drove up to the first window. I'm cracking up at this point, and we are both watching the guy through the window quickly looking around and trying to figure out what happened, as he was somewhat unaware an order had even been placed.\n\nOnce the employee saw the kid with the wife beater and backwards snapback hat on staring expectantly at him through the window, he started laughing and then took his order again.\n\nNot exactly the most exciting or rewarding story, but I was laughing uncontrollably throughout the majority of the whole ordeal.\n\nSo, what are some of your drive-thru stories?

"**DPO**": Was with a friend who decided to yell an order at a fast food place through the intercom, not realizing that it had already been taken by another person, causing a comical situation for all parties involved.

Conclusion: DPO generates short and coherent responses compared to PPO. (matches with our result)

Our Objectives from the Project

1. **Limitations of Existing Literature**
    - Explored only summarization Task.
    - Limitations of RLHF and PPO algorithm.

2. **Our Objective:**
    - Develop Human-aligned response generation LLMs for both Conversation and Summarization task.
    - Specifically used same LLM (LLaMA-2) for both RM and Policy Models.
    - Compare two different forms of reward models - Human Feedback and AI Feedback.
    - Experiment two different RL algorithms for model alignment - DPO and PPO.

Results & Conclusions

Model_Name	METEOR	BLUE-1	BLUE-2	BLUE-3	BLUE-4	ROUGE1	ROUGE2	ROUGEL	ROUGELSUM	BERTScore
LLAMA_2_DPO_RLAIF_SUMMARY	**0.2231**	0.0799	0.0499	0.0377	0.029	0.1875	0.0507	0.143	0.1472	**0.75944224**
LLAMA_2_PPO_RLAIF_SUMMARY	**0.3144**	0.0712	0.0566	0.0499	0.0428	0.1638	0.1048	0.1396	0.1468	**0.77308379**
LLAMA_2_DPO_RLHF_SUMMARY	**0.2709**	0.2118	0.1247	0.0872	0.0632	0.2924	0.1026	0.2283	0.2284	**0.78425058**
LLAMA_2_PPO_RLHF_SUMMARY	**0.3392**	0.0784	0.062	0.0541	0.0463	0.1822	0.1191	0.1564	0.1629	**0.7788625**
LLAMA_2_DPO_RLAIF_CONVERSATION	**0.2677**	0.2225	0.114	0.0685	0.0416	0.29	0.0861	0.1809	0.2068	**0.7882451**
LLAMA_2_DPO_RLHF_CONVERSATION	**0.2762**	0.2199	0.1024	0.0599	0.0377	0.2985	0.0746	0.1789	0.2034	**0.78755994**
LLAMA_2_PPO_RLAIF_CONVERSATION	**0.2845**	0.2377	0.1194	0.0715	0.0437	0.3143	0.0925	0.1952	0.2259	**0.79329664**
LLAMA_2_PPO_RLHF_CONVERSATION	**0.267**	0.2116	0.1031	0.062	0.0376	0.2881	0.0776	0.1774	0.2076	**0.78424052**

1. RLAIF (500 training samples manually annotated using GPT-3.5) is comparable to RLHF (1000 training samples). This could help obliterate human-effort and scale building of alignment models.

2. DPO algorithm is comparable with PPO, which suggests the usage of DPO as more stable and computationally efficient algorithm over PPO.

References

1. Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.
2. Lee, Harrison, et al. "Rlaif: Scaling reinforcement learning from human feedback with ai feedback." arXiv preprint arXiv:2309.00267 (2023).
3. Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." arXiv preprint arXiv:2305.18290 (2023).
4. Chip Huyen's Blog RLHF
5. https://github.com/huggingface/trl/tree/main/examples/scripts
6. https://colab.research.google.com/drive/1gAixKzPXCqjadh6KLsR5ZRUnb8VRvZl1
7. Aman AI Feedback