Toronto Metropolitan University
Data Science & Analytics
Graduate Studies

Reinforcement Learning with Human Feedback(RLHF)-Bridging Gap between AI and Human Expertise

Roshan Issac
Amarpreet Kaur

Introduction to the Problem

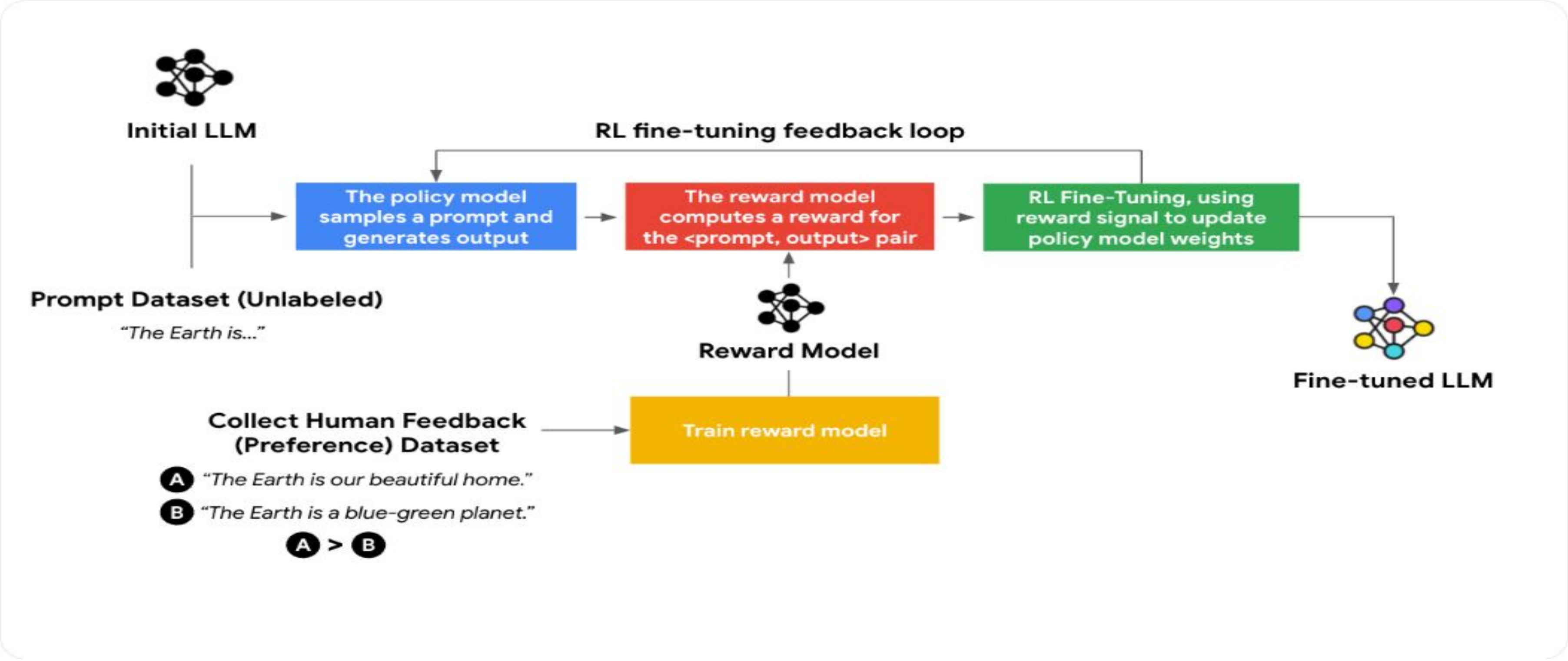
- Abstractive english text summarization is a subjective task and it is difficult to quantify summary quality without humans judgements.
- Automatic metrics like ROUGE score for evaluating summary quality is not a suitable metric as it has poor correlation with human judgements.
- In this work, we show that it is possible to improve summary quality by training a model to optimize for human preferences.
- Due to limitations in the infrastructure and budget we used a preprocessed small subset of the dataset of human comparisons between summaries provided by Google to train a model to predict the human-preferred summary, and use that model as a reward function to fine-tune a summarization policy using reinforcement learning(RL).
- We used RLHF Vertex AI pipeline template provided Google to submit the RLHF fine tuning job which has the RLHF Algorithm embedded in it.

Background

Reference	Explanation	Dataset/Input	Weakness
W. Zhou et al. [1]	The paper proposes a method to evaluate natural language generation models by learning to compare generated text pairs using fine-tuned BERT	Two datasets: the WritingPrompts dataset for story generation and the Dailydialog dataset for open-domain dialogue response generation	Inaccurately evaluate similar-quality texts and inherit BERT's biases
D. M. Ziegler et al. [2]	The method used is reinforcement learning combined with reward models trained from human feedback	Uses the CNN/Daily Mail and TL;DR datasets for summarization tasks, and the BookCorpus dataset for stylistic continuation tasks	Limited by data quality issues and models' excessive reliance on copying input text.
Nisan et al. [3]	The method involves training a policy with the PPO algorithm via reinforcement learning to optimize summaries based on human preferences, using a reward model trained on human judgments	The TL;DR (Too Long; Didn't Read) dataset from Reddit, containing about 3 million posts across a variety of topics	The high costs and computational resources required for training models using human feedback.

Methodology

Toronto Metropolitan University
Data Science & Analytics
Graduate Studies



Methodology

```
'I live right next to a huge university, and have been applying for
a variety of jobs with them through their faceless electronic jobs p
ortal (the "click here to apply for this job" type thing) for a few
months. \n\nThe very first job I applied for, I got an interview tha
t went just so-so. But then, I never heard back (I even looked up th
e number of the person who called me and called her back, left a voi
cemail, never heard anything).\n\nNow, when I\'m applying for subseq
uent jobs - is it that same HR person who is seeing all my applicati
ons?? Or are they forwarded to the specific departments?\n\nI\'ve ap
plied for five jobs there in the last four months, all the resumes a
nd cover letters tailored for each open position. Is this hurting my
chances? I never got another interview there, for any of the positio
ns. [summary]: '
```

```
preference_data[2]['input_text'][-50:]
```

```
'plan something in those circumstances. [summary]: '
```

```
print(f"candidate_0:\n{sample_1.get('candidate_0')}\n")
print(f"candidate_1:\n{sample_1.get('candidate_1')}\n")
```

```
candidate_0:
 When applying through a massive job portal, is just one HR person s
eeing ALL of them?

candidate_1:
 When applying to many jobs through a single university jobs portal,
is just one HR person reading ALL my applications?
```

```
print(f"choice: {sample_1.get('choice')}")
```

```
choice: 1
```

Preference Dataset

This dataset is used to fine tune reward model, it has 4 keys:

{Input prompt, candidate_0, candidate _1, choice (candidate 0 or 1)}

Prompt Dataset

Once the reward mode has been trained, we are using the Reinforcement Learning (RL) Loop to tune the base LLM it requires a base dataset consisting of sample prompts.

It has Input prompt key only, with no response.

Reward Model

We train this model to predict which summary (y∈ (y1, y2)) is better as judged by a human, given a post *x.* If the summary preferred by the human is yi, we can write the RM loss as,

$$\text{loss}(r_\theta) = -E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))]$$

where *r θ* (*x; y*) is the scalar output of the reward model for post *x* and summary *y* with parameters *θ*, and *D* is the dataset of human judgments.

Reinforcement Learning

We want to use the reward model trained above to train a policy that generates higher-quality outputs as judged by humans. We primarily do this using reinforcement learning, by treating the output of the reward model as a reward for the entire summary that we maximize with the PPO algorithm, where each time step is a BPE token.

Inference

We use the final tuned model to get the summarizations for the test dataset.

Proximal policy optimization (PPO)

PPO is a first-order optimization method, and it is designed to update policies in a way that doesn't stray too far from the previous iteration. There are two primary variants of PPO: PPO-Penalty and PPO-Clip. We'll focus only on PPO-Clip (the primary

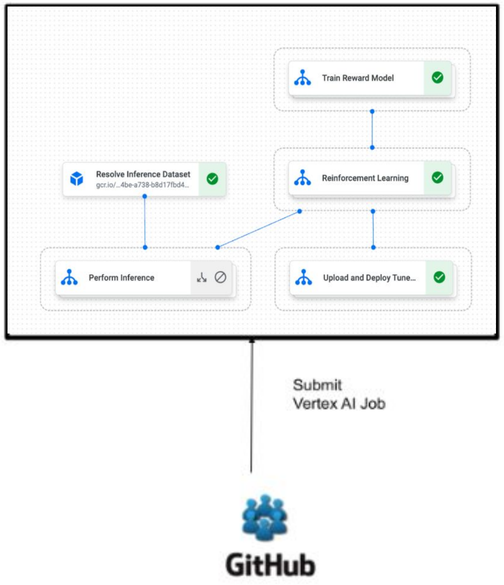
$$\theta_{k+1} = \arg\max_{\theta} \underset{s,a\sim\pi_{\theta_k}}{\mathrm{E}} [L(s,a,\theta_k,\theta)]$$

$$L(s,a,\theta_k,\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s,a),\;\; \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1-\epsilon, 1+\epsilon\right) A^{\pi_{\theta_k}}(s,a)\right)$$

Implementation

Inputs

- Base Model
- Prompt Dataset
- Preference Dataset
- Test Dataset
- Parameters



Outputs

- Tuned Model
- Predictions based on test data
- logs

Base Model: Llama-2 7B
Prompt Dataset:gs://nlp-project-edplato-ds-dev-ds8008/datasets/prompt_dataset.jsonl
Preference Dataset:gs://nlp-project-edplato-ds-dev-ds8008/datasets/preference_dataset.jsonl
Test Dataset:gs://nlp-project-edplato-ds-dev-ds8008/datasets/validate_dataset.jsonl

Compute Engine	Time to Execute	Cost
TPU	1 day 4 hour	~400-600 CAD

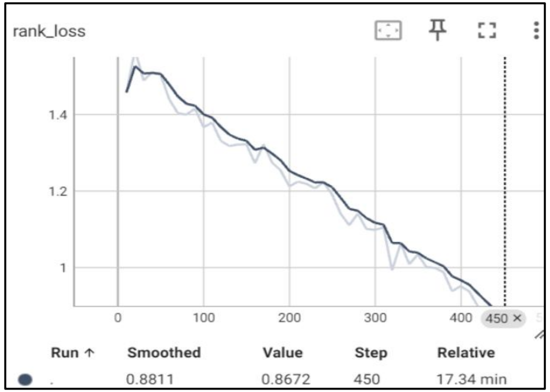
Implementation(Contd)

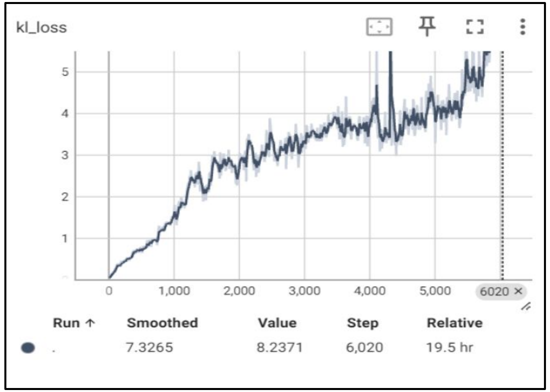
Toronto Metropolitan University
Data Science & Analytics Graduate Studies

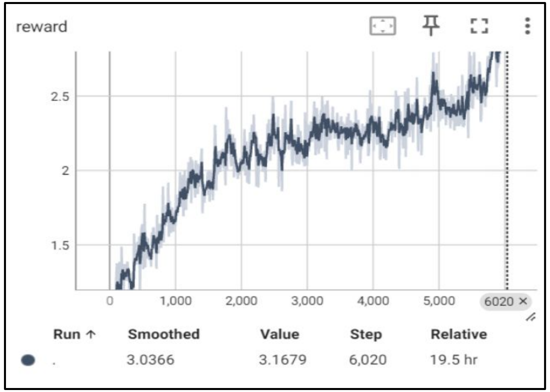
Training Curves(from TensorBoard Metrics)

Rank Loss(Reward Model) **KL Loss** **Rewards**

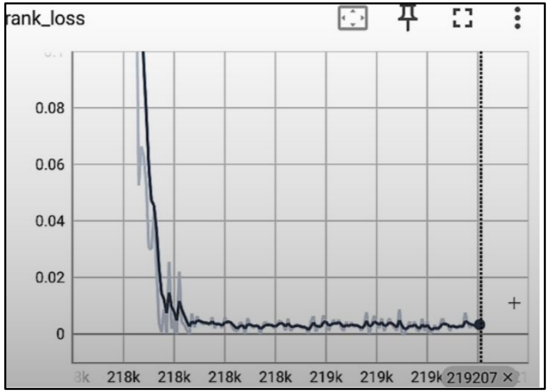
Our Model Results

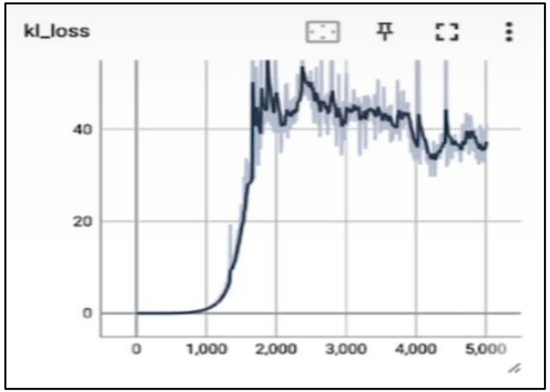


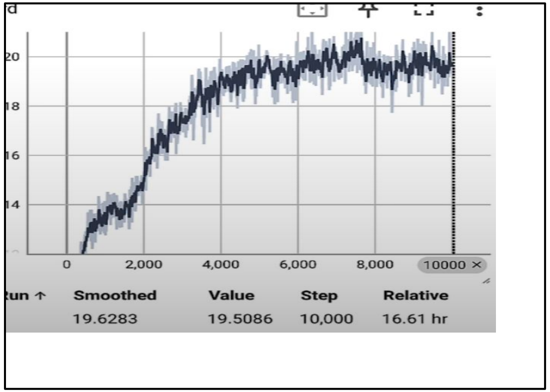




Full data results (Source: Google)







Implementation(Contd)

Toronto Metropolitan University Data Science & Analytics Graduate Studies

Side By Side Evaluation

	prompt	base_model	tuned_model
0	Summarize in less than 50 words\nWe bought a house! But the paint colors aren't great - lots of bright colors in what used to be kids' rooms and walls in need of retouching in the common areas. I'd like to get starting on the paint project right away and my family will be visiting in two weeks to help - so I'd like to develop a paint scheme to go off of. \n I get a little overwhelmed when contemplating the seemingly endless paint colors at Home Depot. I want neutral walls, but should I stick with one shade throughout or mix it up? Is beige blah, is cream too boring, and could anyone tell me what #%&! greige is? \n My couch is brown leather, the entertainment center, kitchen table, and bedroom furniture are a warm cherry, and the cabinetry is maple. Since we'll be buying furniture over time as funds allow, I think neutral walls that play well off other colors is my best bet. The house gets gorgeous light from western facing windows, but even still I don't want to go too dramatic or dark.	Congratulations on your new home! You're feeling overwhelmed by the many paint colors available at Home Depot, and you're unsure of how to choose the best scheme for your neutral-colored furniture and light-filled space. You're considering sticking with one shade throughout or mixing it up, but you want to avoid going too dramatic or dark. You're open to suggestions!	Excited about their new house, the owner is eager to repaint. They seek a neutral palette but feel overwhelmed by choices. With family visiting soon, they aim for a versatile scheme that complements existing furniture and enhances natural light without being too dramatic or dark.
1	Summarize in less than 50 words\nHello. I recently changed around a little bit on my router settings. To be more specific, only the channels. I had a problem with an extender, which had a problem where the signal would get weaker every 25th second (If anyone can recommend a solution to this as well, I will be more than happy to listen). So, I checked the internet before to se nothing was wrong, (10 ping, 25mbps) and looked through the settings later and changed the channel. When I realized all the other channels were no better, I changed back to the "auto" channel. After all this, my ping went up to 50, and mbps down to 15. I have no idea how that happened, and I appreciate your help. Thank you. \n The router I'm using is an old Linksys WRT54G with the Tomato OS, if that helps. And the extender is an Airlink2200AC	The person changed the channels on their router and noticed a significant decrease in speed and ping. They changed back to the "auto" channel and noticed a further decrease in speed and ping, despite checking the internet connection before making any changes. They are seeking help to understand why this happened and how to resolve the issue.	The user adjusted router channels due to signal issues with an extender but encountered increased ping (50) and decreased Mbps (15) after reverting to "auto" channel. Seeking assistance, they use a Linksys WRT54G router with Tomato OS and an Airlink2200AC extender. Troubleshooting guidance is requested.

Conclusion and Future Directions

Learnings

- We learned fine tune a LLM and how to execute RLHF pipeline in Google's Vertex AI Platform.
- Learned How human feedback can affect the LLM response.
- Learned fine tuned models performed better than base model in giving response aligning with human preference.
- Learned to understand various training curves in the RLHF domain

Limitations

- Infrastructure and Cost is a hindrance to run the RLHF fine tuning.
- We need very large formatted dataset to train ,especially we need human experts for their judgement in preference dataset.

Future Direction

- Train the model with large dataset and with better fine tuned parameters.
- Use methods like Reinforcement Learning from AI Feedback (RLAIF) which automates human preference annotations with the use of a generic LLM.

References

[1] Wangchunshu Zhou, Ke Xu, "Learning to compare for better training and evaluation of open domain natural language generation models", 2020, https://arxiv.org/pdf/2002.05058.pdf

[2] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu Tom B. Brown Alec Radford Dario Amodei Paul Christiano Geoffrey Irving, "Fine-tuning language models from human preferences", 2020, https://arxiv.org/pdf/1909.08593.pdf

[3]: Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul Christiano ,"Learning to summarize from human feedback ", 2022, https://arxiv.org/pdf/2009.01325.pdf

[4] Tutorial Reinforcement Learning from Human Feedback(Code Implementation) , https://learn.deeplearning.ai/login?callbackUrl=https%3A%2F%2Flearn.deeplearning.ai%2Fcourses%2Freinforcement-learning-from-human-feedback

[5] Google Cloud RLHF, https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune-text-models-rlhf

[6] Proximal Policy Optimization(PPO), https://spinningup.openai.com/en/latest/algorithms/ppo.html#id5