

## A Broader Impacts

RRHF can align with not only human preferences but also any preferences. One may use RRHF to align with harmful preferences like sexual and criminal preferences which are discouraged by us.

## B Safeguards of Wombat

As a large language model, Wombat has the possibility to generate unsafe responses. Wombat is only used for research and is not intended for use in production systems. We will use RRHF to further improve the safety of Wombat to align to a helpful and harmless AI.

## C IMDB Sentiment

We also conduct experiments on the IMDB dataset for assessing positive movie reviews generation. The task expects the model to give positive and fluent movie review completions based on given partial review input texts. The dataset contains 25k training samples and each 5k sample set for validation and testing. Following Ramamurthy et al. [25], we use a partial movie review as the input for each sample, and the lengths of partial text are set up to 64 tokens. During both training and evaluation, we set the maximum generated completion length to 48 tokens.

**Table 10:** In the Setting Column, for RRHF, BP represents the same training workflow as the top-most workflow in Figure 2 in the main texts. B represents the same workflow while it excludes the text completion labels in the dataset. RRHF-OP-128 follows the bottommost workflow in Figure 2 in the main texts.

	Setting	Reward	Perplexity
SFT	-	0.539	35.472
PPO	w/o KL penalty	0.796	42.916
NLPO	w/o KL penalty	0.777	41.035
RRHF	BP	0.861	32.083
RRHF	B	0.799	32.077
RRHF-OP-128	w/o KL penalty	0.990	32.081
PPO	0.1 KL penalty	0.626	35.049
NLPO	0.1 KL penalty	0.620	34.816
RRHF-OP-128	0.1 KL penalty	0.635	32.088

For detailed experiment settings, in order to conduct a fair comparison with PPO and NLPO from Ramamurthy et al. [25]. For the reward model, we use the same sentiment classifier as Ramamurthy et al. [25] which is provided by Sanh et al. [26], and the same SFT GPT-2 model as the starting language model provided by Ramamurthy et al. [25]. For generation settings, we also use top-k sampling with K=50 across our experiments for RRHF and RRHF-OP. We set the training batch size to be 64 and set the total training epochs to be 5 which is far less than Ramamurthy et al. [25] and is enough for RRHF to achieve good performance. We also experiment using reward designs with and without KL penalty against SFT model distribution for RRHF-OP.

Results of IMDB sentiment generation are listed in Table 10. We use the reward score of the reward model and perplexity by GPT-2 [23] to demonstrate the performance of alignment. We can conclude from the results that: (1) PPO, NLPO and RRHF(-OP) can align the SFT model to the preference of the reward model (increasing the reward score); (2) RRHF performs better in terms of reward score and perplexity than both PPO and NLPO with and without KL penalty; (3) RRHF-OP-128 outperforms PPO and NLPO with and without KL penalty; (4) With KL penalty in training reward design, RRHF-OP-128 shows less progressive increase in reward score compared with RRHF-OP-128 trained without KL penalty in reward designs.

Although we keep the input and output lengths and generation settings consistent with Ramamurthy et al. [25], we do not observe fluctuations in perplexity as measured by GPT-2 for RRHF. Therefore we conduct a case study on the samples generated by models trained with RRHF-OP-128 without KL penalty. Cases in Table 11 show that without KL penalty, the model trained with RRHF-OP-128 learns to generate positive reviews such as " It's a great film and I highly recommend it to anyone."

for different review inputs. This pattern may explain the extremely high reward score while still maintaining a perplexity score by GPT-2.

**Table 11:** Case Studies. Texts in red are the models generated completions

... knowing how AWFUL Drew’s character was (ostrich feathers?) at the start of the school year would <b>have made it a lot more satisfying. It’s a great film and I highly recommend it to anyone. It’s a great film and I highly recommend it to anyone.</b>
... Maybe it was from a gynecological experiment gone wrong.   <b>The film is great. It’s a great film and I highly recommend it to anyone. It’s a great film and I highly recommend it to anyone.</b>
... feeling and atmosphere perfectly, helped in part with some incredible archival footage. Tony Alva is <b>a great film, it is a great film, I highly recommend it to anyone.</b>

## D Details of Human Evaluation on HH Dataset

A total of 330 comparison pairs were sampled for RRHF evaluation, involving comparisons between RRHF and good responses (110 pairs), RRHF and PPO (110 pairs), and RRHF and RRHF<sub>IP-2</sub> (110 pairs). Out of these, 30 pairs were used to calculate agreement, while the remaining 300 pairs were used for reporting scores. Each crowd-sourced worker was tasked with labeling 130 pairs, consisting of 100 random pairs and 30 common pairs. The average consistency between each pair of reviewers was calculated, revealing that they provided the same annotations for 57.7% of pairs and their annotations did not contradict each other for 84.4% of pairs.

## E Wombat: Learn from ChatGPT comparison

**Sampling Policy** We use training data from Alpaca as sampling queries. We sample five different responses for training:  $\rho_1, \rho_2$  are generated by ChatGPT,  $\rho_3$  is generated by text-davinci-003,  $\rho_4$  is generated by LLaMA and  $\rho_5$  is generated by Alpaca.

**Reward Prompt** We require ChatGPT to read five responses and score by four different dimensions: Relevance, Correctness, Coherence, and Safety from 1 to 5. We parse the outputs of ChatGPT and calculate the sum as the response reward score. The used prompt is:

*Score different responses separately without explanation and without copying any input from these respects, please scores start from response 1: Relevance (does it relevant to user’s query), Correctness (does it contain correct knowledge), Coherence (does it generate fluently and without grammar problems), Safety (does it refuse to answer sex or criminal queries) and give a score for each respect 1-5. Query: ... Response 1: ... Response 2: ... Response 3: ... Response 4: ... Response 5: ...*

**Results** For 52k training data from Alpaca, we successfully parsed 46k data with reward scores. We use Alpaca as our initial checkpoint and train it by RRHF. We name this model Wombat. The total training time of Wombat is only 4 hours on 8 A100 GPUs. Here we give generated examples of Wombat in Appendix C to show its ability. We can find it has general world knowledge, the ability to answer complex machine learning problems, and the ability to follow instructions.

## F Example responses generated by Wombat

**Table 12:** Example responses generated by Wombat.

<b>Query:</b> Give a introduction of wombat
<b>Response:</b> Wombat is a marsupial found in Australia, Tasmania, and New Zealand. They have a distinctive appearance with a short, stocky body, a broad head, and a short tail. Wombats are herbivores and primarily graze on grasses, leaves, and bark. They are known for their burrowing habits and can dig burrows up to 18 meters deep and 10 meters long. Wombats are a protected species and are considered vulnerable due to habitat loss and predation by introduced species.
<b>Query:</b> How to design an algorithm to rank responses of language model outputs with aligned human preferences?
<b>Response:</b> One possible algorithm to rank responses of language model outputs with aligned human preferences is: <ol style="list-style-type: none"> <li>1. Collect data: Collect data on human preferences for a given task, such as rating the importance of different features or characteristics of a product or service.</li> <li>2. Preprocess data: Preprocess the data by removing outliers, missing values, and irrelevant information.</li> <li>3. Train a language model: Train a language model using the collected data to generate responses based on the preferences of humans.</li> <li>4. Evaluate model performance: Evaluate the performance of the language model by comparing its responses to the collected data on human preferences.</li> <li>5. Design a ranking algorithm: Design a ranking algorithm that takes into account the preferences of humans as well as the performance of the language model.</li> <li>6. Apply the ranking algorithm: Apply the ranking algorithm to the language model's responses to rank them based on their alignment with human preferences.</li> <li>7. Evaluate the results: Evaluate the results of the ranking algorithm to determine its effectiveness in aligning language model outputs with human preferences.</li> <li>8. Improve the algorithm: Continuously improve the algorithm by incorporating feedback and refining the preferences data and the ranking criteria.</li> </ol>
<b>Query:</b> Assume you are a teacher, you are teaching a 6-year-old child. Do not give the children direct answers, but try to teach him. Children: How to calculate 6+7?
<b>Response:</b> Great! Let's count up to 13 together. 6+7 equals 13. Starting with 6, we can count up by 1 each time. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. We've counted up to 13! So, the answer is 13.