

Parameter Efficient Reinforcement Learning from Human Feedback

Hakim Sidahmed^{*1}, Samrat Phatale^{*1}, Alex Hutcheson², Zhuonan Lin²,
 Zhang Chen², Zac Yu², Jarvis Jin², Simral Chaudhary¹, Roman Komarytsia²,
 Christiane Ahlheim², Yonghao Zhu², Bowen Li², Saravanan Ganesh², Bill Byrne²,
 Jessica Hoffmann¹, Hassan Mansoor¹, Wei Li¹, Abhinav Rastogi¹, Lucas Dixon¹

¹Google DeepMind, ²Google

{hsidahmed, samratph, ldixon}@google.com

Abstract

While Reinforcement Learning from Human Feedback (RLHF) effectively aligns pretrained Large Language and Vision-Language Models (LLMs, and VLMs) with human preferences, its computational cost and complexity hamper its wider adoption. To alleviate some of the computational burden of fine-tuning, parameter efficient methods, like LoRA (Hu et al., 2021) were introduced. In this work, we empirically evaluate the setup of Parameter Efficient Reinforcement Learning from Human Feedback (PE-RLHF) that leverages LoRA fine-tuning for Reward Modeling, and Reinforcement Learning. We benchmark the PE-RLHF setup on six diverse datasets spanning summarization, harmless/helpful response generation, UI automation, and visual question answering in terms of effectiveness of the trained models, and the training resources required. Our findings show, for the first time, that PE-RLHF achieves comparable performance to RLHF, while significantly reducing training time (up to 90% faster for reward models, and 30% faster for RL), and memory footprint (up to 50% reduction for reward models, and 27% for RL). We provide comprehensive ablations across LoRA ranks, and model sizes for both reward modeling and reinforcement learning. By mitigating the computational burden associated with RLHF, we push for a broader adoption of PE-RLHF as an alignment technique for LLMs and VLMs.

1 Introduction

Large Language and Vision-Language Models (LLMs, and VLMs) like GPT-4 (OpenAI et al., 2023) and Gemini (Team et al., 2023; Reid et al., 2024) demonstrate remarkable performance across diverse tasks. However, aligning

^{*}Equal Contribution.

these models with human preferences remains crucial for ensuring desirable behavior (Bommasani et al., 2022). This alignment improves instruction following (Ouyang et al., 2022), and facilitates optimizing for behaviors that lack a clear mathematical loss function, such as safety properties (Bai et al., 2022a,b; Glaese et al., 2022), helpfulness (Bai et al., 2022a; Glaese et al., 2022), summarization characteristics (Stennnon et al., 2020), and visual instructions (Sun et al., 2023). Reinforcement Learning from Human Feedback (RLHF) has emerged as a prominent method for achieving this alignment. It involves training a reward model (RM) on human feedback data, and subsequently using this RM to fine-tune the model parameters via Reinforcement Learning (RL). While effective (Stennnon et al., 2020; Bai et al., 2022b), RLHF's complexity and computational demands hinder its widespread adoption. Moreover, the RL loop necessitates extra model copies - such as for the reward model, and the anchor model used for KL regularization - which significantly increases its memory usage in comparison to standard fine-tuning.

We compare standard RLHF, where all the parameters of the reward model and policy are fine-tuned, to Parameter-Efficient Reinforcement Learning, which leverages LoRA (Low-Rank Adaptation) (Hu et al., 2021) for fine-tuning both the reward model, and the reinforcement learning policy. While more powerful Parameter Efficient Fine-Tuning (PEFT) and Representation Fine-Tuning (ReFT) approaches have been developed since LoRA, our study focuses on this method, as it is widely adopted. We hope our results will motivate the benchmarking of other PEFT and ReFT approaches on RLHF tasks. Despite training only a small fraction of the parameters, we demonstrate that the results obtained with PE-RLHF

are on par with those obtained with standard RLHF.

Figure 2 illustrates the differences between PE-RLHF and standard RLHF.

Our contributions are threefold:

- **Thorough Comparative Analysis:** We conducted an extensive evaluation of PE-RLHF against standard RLHF methods across six diverse datasets and five distinct tasks. While LoRA’s efficiency was expected, its surprisingly strong performance establishes it as a superior alternative to full fine-tuning for RLHF. Detailed results are presented in Table 1.
- **In-depth Ablation studies:** We systematically examined the influence of LoRA on both RM and RL policy training, considering variations in model size and LoRA ranks.
- **Demonstrated Resource Savings:** We provide empirical measurements demonstrating reductions in memory consumption and training time achieved by PE-RLHF as compared with standard RLHF.

We hope that this study will pave the way for more efficient and accessible RLHF, promoting wider adoption and facilitating the development of large models that better align with human preferences.

2 Parameter Efficient Reinforcement Learning from Human Feedback

RLHF involves two phases: reward model training, and reinforcement learning of a policy model. PE-RLHF applies parameter-efficient fine-tuning techniques to optimize both of these training phases, thus significantly reducing the memory requirements, and increasing the training speed. We provide a brief overview on RLHF in Appendix A.1.

2.1 Reward Model Training

PE-RLHF constructs reward models as language models with Low-Rank Adaptation (LoRA) adapters. These adapters are attached to each attention projection matrix within the model. During training, only the adapters are trained, while the language model backbone

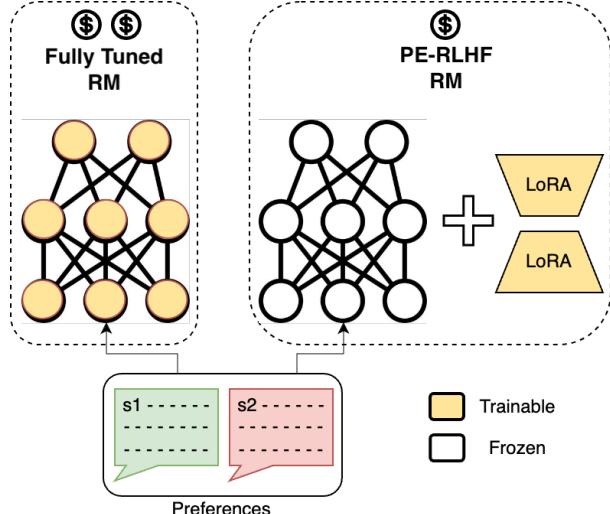


Figure 1: Standard RM training (left) vs. PE-RLHF RM training (right). PE-RLHF RM only trains the LoRA adapters, while keeping the Language Model backbone frozen.

remains frozen. This approach, illustrated in Figure 1, significantly reduces the number of trainable parameters. During inference, the trained LoRA adapters are combined with the projection matrices through a one-time addition operation. This results in a reward model functionally equivalent to a non-LoRA model, but trained efficiently.

2.2 Reinforcement Learning of Policy

Similarly, PE-RLHF uses LoRA adapters for policy and value models within the reinforcement learning loop. As with the reward model, adapters are attached to each attention projection matrix, and trained while keeping the language model backbone frozen (Figure 2). The policy is then optimized using the policy gradient calculated based on the value model. The value model is trained using the reward score, along with KL regularization with the anchor policy. We optimize the policy using “REINFORCE for Language Models”, as used by Lee et al. (2023a).

3 Datasets and Tasks

We describe the datasets we used for training the reward models, and performing reinforcement learning, categorized by their respective tasks. Our experiments are based on a diverse collection of datasets, and test various capabilities of language models. This diversity allows

Table 1: PE-RLHF RM training can match standard RM training in terms of reward model accuracy while using 43-74% of the HBM at peak and trains 1.4-1.9 \times faster as compared to standard RM training. PE-RLHF can match the standard RLHF while using 73-80% HBM at peak, and training 1.15-1.3 \times faster.

		Harmlessness	Helpfulness	Summarization		UI Automation	Visual QA
		Anthropic	SHP	Reddit	Messages	UI Automation	VQAv2
RM *	Full-tuning	76.56%	83.2%	78.7%	-	93.12%	-
	PE-RLHF	78.71%	82.2%	79.7%	-	91.8%	+0.5%†
Efficiency	PE HBM	43.1%	48%	50%	-	56%	74%
	Speed-up	1.6 \times	1.5 \times	1.7 \times	-	1.9 \times	1.15 \times
RL **	Full-tuning	96.6%	63.0%	87%	73.2%	81.6%	+5.5%‡
	PE-RLHF	98.2%	61.3%	86.5%	75.5%	86.4%	+3.9%‡
Efficiency	PE HBM	80%	75%	75%	80%	74%	74%
	Speed-up	1.3 \times	1.15 \times	1.15 \times	1.05 \times	1.20 \times	1.24 \times

* Accuracy.

** Win rate by a Judge model.

† Absolute win rate change compared to Supervised Fine-Tuning baseline.

‡ Absolute accuracy change compared to Fully-Tuned baseline.

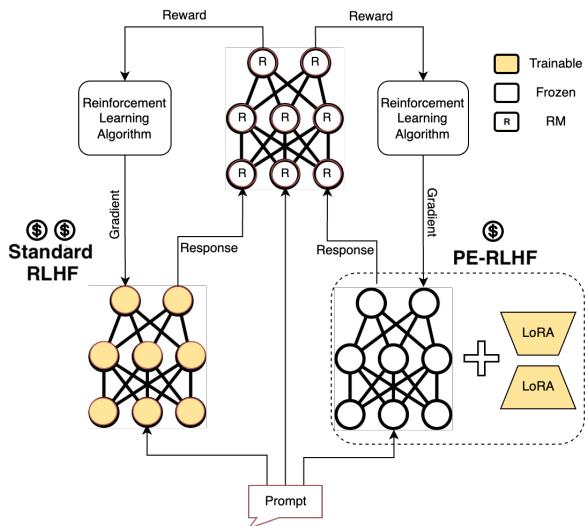


Figure 2: Standard RLHF (left) vs. PE-RLHF right. PE-RLHF only trains the LoRA adapters while keeping the Language Model backbone frozen.

us to evaluate the efficacy PE-RLHF across different domains, and assess its ability to generalize to new tasks. The datasets are grouped into the following categories:

Text Summarization: These datasets test a model’s ability to condense information, and generate concise summaries. We experiment with the Reddit TL;DR (Völske et al., 2017; Stiennon et al., 2020), and BOLT Message Summarization (Chen et al., 2018) datasets. The Reddit TL;DR dataset contains Reddit posts together with human-annotated summaries, used to train models for summarizing these posts.

We filtered this dataset following the work of Stiennon et al. (2020). The BOLT Message Summarization dataset contains chat conversations, used to train models for summarizing sequences of messages.

Harmless Response Generation: Generating harmless responses is critical to the development of responsible AI systems. We use the “Harmlessness” dataset from Anthropic-HH (Bai et al., 2022a). This dataset evaluates a model’s ability to generate safe responses to red-teaming prompts.

Helpful Response Generation: This task asks the model to provide helpful answer for a given question. We use Stanford Human Preference dataset (Ethayarajh et al., 2022) to evaluate the helpfulness of the generated answer.

UI Automation: This task assesses the model’s ability to understand and interact with user interfaces. We run our experiments on the AndroidControl dataset (Li et al., 2024a), which consists of human demonstrations of controlling a device, used to train reward models that evaluate the effectiveness of actions in a UI automation task.

Visual Question Answering: We test a model’s ability to understand visual information, and answer questions about it using the VQAv2 dataset (Antol et al., 2015; Goyal et al., 2017a).

Detailed information about each dataset can be found in Appendix A.2.

4 Experimental Setup and Metrics

We run extensive experiments on six datasets, using models from two different families:

- PaLM 2 (Anil et al., 2023): A text focused large model pretrained following the UL2 (Tay et al., 2022) paradigm. We experiment with three sizes for this model, referred to as XXS, XS, and S. We use a larger PaLM 2 L as a judge in our evaluations. PaLM 2 models are available via Google Cloud’s Vertex API by the names Gecko (XXS), Otter (XS), Bison (S), and Unicorn (L).
- Gemini Pro (Team et al., 2023): A vision language model, also available via Google Cloud’s Vertex API.

We emphasize that our experimental setup is independent of the specific models used.

We describe the experiments for reward modeling and reinforcement learning in more details below.

4.1 Reward Modeling

We train reward models with the loss described in Appendix A.1, and the hyperparameters described in Appendix A.3, varying the configurations of model size, and LoRA rank (which controls the number of trainable parameters).

We evaluate the performance of the preference based reward models using the *pairwise-accuracy*, which measures the proportion of preferred responses ranked higher by the model among pairs of candidate responses. We evaluate the classification style reward models using the *accuracy*, which indicates whether the reward model score is close to the label of 0 or 1. We compare the High Bandwidth Memory (HBM) usage as estimated by Jax JIT at the time of training (Bradbury et al., 2018), and report its peak value. We also evaluate and report the speed of training for each setting.

4.2 Reinforcement Learning

We train reinforcement learning policies using the “REINFORCE for Language Models” algorithm used by Lee et al. (2023a), using a

fixed reward model for each dataset for a fair comparison across the different settings (this is to reduce confounding factors that affect the policy performance). For every setting we try for policy model, both in size and LoRA rank, we replicate that for the value model as well. We report the experiment hyperparameters in Appendix A.3.

4.3 Evaluations

We evaluate the performance of the RL tuned policies using the PaLM 2 L model (judge model), which is prompted to judge the responses produced by the policy.

We use it to evaluate the quality of the models as follows:

Text Summarization: We prompt the judge model to pick a preferred response among a pair of baseline, and policy responses. We report the “win rate” of the policy, as defined by the percentage of generated responses that are better than the baseline. We use an instruction-tuned version of PaLM 2 S to generate baseline responses. We prompt the judge model twice, with two possible orders, to eliminate positional bias, and average the judgements: only if an output is preferred in both possible orderings is it considered a win. If it is only preferred in one ordering, then we label a tie.

Harmless Response Generation: We prompt the judge LLM to assess whether the generated response is harmless in a YES/NO manner. We report the “harmless rate”, which measures the fraction of the generated responses that are *harmless*.

Helpful Response Generation: Similarly to the text summarization task, we prompt the judge model to pick a preferred (more helpful) response among a pair of baseline and policy responses. We report the “win rate” of the policy, as defined by the percentage of generated responses that are better than the baseline.

Visual Question-Answering: We evaluate the quality of an RL policy as the difference between its accuracy, and that of a baseline Supervised Fine-Tuned Gemini policy.

UI Automation: We prompt the judge LLM with the task description, together with the action generated by the policy model. We

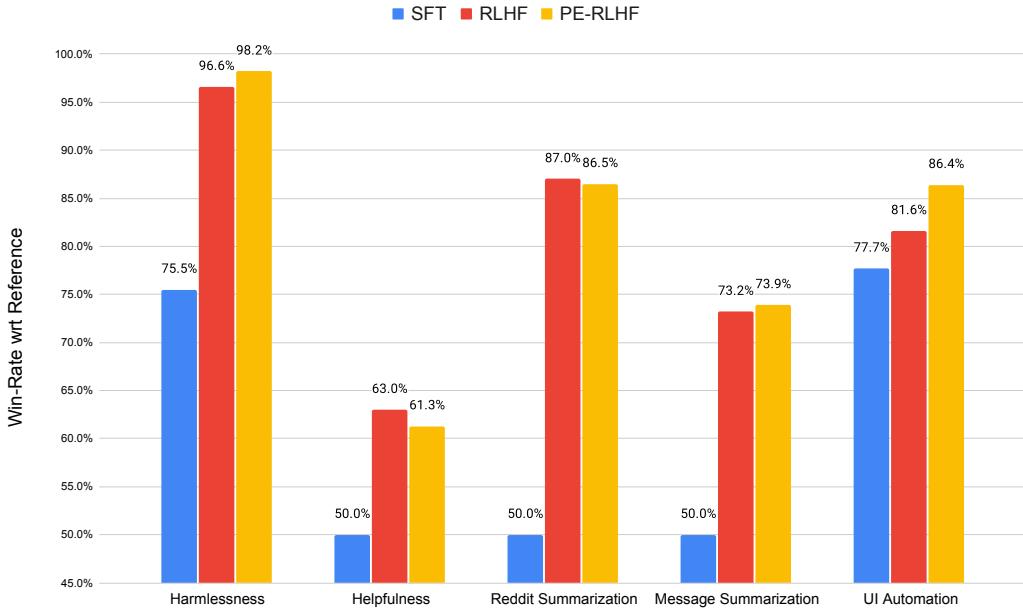


Figure 3: PE-RLHF performs on par with standard RLHF. Both PE-RLHF and RLHF outperform SFT policies significantly on all tasks.

asked the judge LLM to determine whether the action is correct for the UI automation task. We calculate the accuracy rate as defined by the percentage of the policy model generated action that are considered as correct ones.

We compare and report the speed of each RL training experiment. In contrast to the speed measured for the reward models, which only include the learning step, the reinforcement learning measurements include episode sampling, reward model scoring, anchor logit calculation, and learning step.

5 Results and Takeaways

5.1 Performance of PE-RLHF

We observe that PE-RLHF achieves results on par with those of standard RLHF in both reward modeling, and reinforcement learning.

PE-RLHF RMs match the performance of their RLHF counterparts across diverse tasks, as measured by the accuracy of the reward models on an evaluation split. These LoRA adapted RMs achieve this performance by training less than 0.1% of the large model’s total parameter count. We summarize the reward modeling results in the top half of Table 1.

When guided by the same reward model, PE-RLHF policies achieve performance competitive

with those of the policies trained in standard RLHF. They both perform significantly better than an SFT policy, as seen in Figure 3, showing the effectiveness of reinforcement learning. PE-RLHF policies achieve this performance by training less than 0.1% of the large model’s total parameter count for text based tasks, and less than 0.2% of the large model’s total parameter count for tasks involving both vision and text. We summarize the reinforcement learning policy results in the bottom half of Table 1.

We assess the quality of PaLM 2 L as a judge by collecting human feedback on 50 samples. We calculate the agreement between the human and judge model’s feedback across tasks, to verify the validity of the latter. We report details of our verification in Appendix A.4.

5.2 Effects of Model Size and LoRA Rank

We conduct ablation studies across model size and LoRA rank to test their effect on the performance of the reward model and the reinforcement learning policy.

We observe that changing the LoRA rank does not significantly affect the performance of the reward models. However, PE-RLHF is more effective at modeling reward, and performs closer to standard full-tuning when the size

Table 2: PE-RLHF RMs match the accuracy of full-tuned RMs, and get more effective at performing equally to fully-tuned RM as the backbone model increases in size. We don’t observe significant changes in performance with varying LoRA ranks.

Model	Setting	Harmlessness	Helpfulness	Summarization	UI Automation
		Anthropic	SHP	Reddit	UI Automation
PaLM 2 S	Fully-Tuned	76.6%	83.2%	78.7%	93.1%
	LoRA 16	75.0%	81.3%	77.0%	92.2%
	LoRA 8	79.1%	81.6%	77.3%	92.0%
	LoRA 4	78.7%	82.2%	79.7%	91.2%
	LoRA 1	76.0%	80.9%	77.2%	90.8%
PaLM 2 XS	Fully-Tuned	77.0%	82.0%	78.1%	88.6%
	LoRA 8	76.8%	82.6%	-	89.2%
	LoRA 4	-	81.8%	76.8%	89.7%
PaLM 2 XXS	Fully-Tuned	73.4%	80.2%	75.4%	84.4%
	LoRA 8	72.5%	79.3%	-	87.1%
	LoRA 4	-	77.8%	73.2%	85.8%

Model	Setting	Visual Question Answering	
		VQAv2*	
Gemini Pro	Fully-Tuned	0.0%	
	LoRA 32	-1.2%	
	LoRA 16	-1.0%	
	LoRA 8	-1.1%	
	LoRA 4	-0.6%	
	LoRA 1	-1.0%	

* Absolute accuracy change compared to Fully-Tuned baseline.

of the model backbone increases. At the lowest model size, we see PE-RLHF falling marginally short of the fully-tuned reward models, whereas it matches the performance of fully-tuned ones for the largest model sizes.

In the reinforcement learning experiments, on the other hand, we observe that the PE-RLHF policies perform better (as evaluated by the judge LLM) as the LoRA rank increases. Bigger model backbones translate into better results for the PE-RLHF policies in comparison to the standard RLHF ones, consistent with the trend observed for the reward models. PE-RLHF falls marginally short of the fully-tuned RL policies with the smallest model size, but matches the performance of RL policies with the biggest size. See Tables 2 and 3 for detailed results.

5.3 Memory and Speed Advantages of PE-RLHF

Modern optimizers, such as Adam (Kingma and Ba, 2017) and Adafactor (Shazeer and Stern, 2018), require substantial memory to track var-

ious factors for each trainable parameter. By largely reducing the number of trainable parameters, PE-RLHF significantly lowers the memory footprint of both reward model training and reinforcement learning of policy. PE-RLHF RM training achieves the performance of a fully-tuned RM, while using only 43% to 74% of the peak HBM it needs for training. PE-RLHF reinforcement learning achieves the performance of standard RLHF, while using only 74% to 80% of the peak HBM.

The reduction in number of trainable parameters also translates into significant training speed-ups, as fewer parameters need updating at each training step. PE-RLHF RM trains at $1.15 \times$ to $1.9 \times$ the training speed of standard RM. PE-RLHF RL trains $1.15 \times$ to $1.24 \times$ the speed of the RL loop in standard RLHF. We note that the LoRA reward models and policies converge in a similar number of steps as the fully tuned ones, so that the speed-ups in training steps translate into faster runs.

We observe that the memory savings, and training speed-up do not vary significantly with

Table 3: PE-RLHF policies match the performance of standard RL policies, and become more effective as the LM increases in size. We don’t observe significant variations in performance with the LoRA rank.

Model	Setting	Harmlessness	Helpfulness	Summarization		UI Automation
		Anthropic	SHP	Reddit	Messages	AndroidControl
PaLM 2 S	Fully-Tuned	96.6%	63.0%	87%	73.2%	81.6%
	LoRA 16	96.4%	61.3%	86.5%	75.5%	86.4%
	LoRA 8	97.7%	58.3%	85.7%	73.4%	85.4%
	LoRA 4	96.7%	60.1%	84.2%	73.9%	84.5%
	LoRA 1	98.2%	57.9%	85%	73.1%	77.7%
	SFT	75.5%	50%	50%	50%	77.7%
PaLM 2 XS	Fully-Tuned	96.1%	53.5%	77.7%	64.3%	52.4%
	LoRA 16	97.4%	52.5%	79.5%	65.5%	69.4%
	SFT	70.3%	-	33.6%	48.1%	48.5%
PaLM 2 XXS	Fully-Tuned	96.6%	5.92%	48.4%	33.7%	11.7%
	LoRA 16	96.6%	5.61%	31.7%	31.7%	14.6%
	SFT	66.7%	-	16.0%	14.1%	24.3%

Model	Setting	Visual Question Answering
		VQAv2*
Gemini Pro	Fully-Tuned	+5.5%
	LoRA 32	+3.9%
	LoRA 16	+3.1%
	LoRA 8	+2.5%
	LoRA 4	+3.0%
	LoRA 1	+3.0%
	SFT	+0.0%

* Absolute accuracy change compared to SFT.

the LoRA rank, since the change in trainable parameters is extremely small in comparison to total parameters (<1% in maximum LoRA rank of 32). We report exact numbers for memory savings, and training speed-up in Table 1. We also note that the memory savings and speed-up depend on multiple factors, such as sequence lengths of the examples, the accelerators being used, etc.

6 Related Work

6.1 Pretrained Large Models (PLMs)

PLMs like LaMDA (Thoppilan et al., 2022), BLOOM (Workshop et al., 2022), PaLM (Chowdhery et al., 2022; Anil et al., 2023), and GPT-4 (OpenAI et al., 2023) have demonstrated remarkable performance across diverse tasks, including summarization (Stiennon et al., 2020), instruction following (Ouyang et al., 2022; Lai et al., 2023), and dialogue generation (Friedman et al., 2023; Jandaghi et al., 2023). Despite their success, limitations such as factual inaccuracies and imperfect instruc-

tion adherence remain (Radford et al., 2018; Ramachandran et al., 2016; Wei et al., 2021).

6.2 Aligning PLMs with Human/AI Preferences

To address these limitations, aligning PLMs with human preferences has emerged as a crucial research area (Christiano et al., 2017; Leike et al., 2018; Wang et al., 2023; Ji et al., 2023). This typically involves collecting preference data on generated outputs, and fine-tuning the model with reward functions based on this data. However, overfitting the reward function can be problematic (Azar et al., 2023), requiring techniques like early stopping or parameter reduction to ensure optimal policy training. While most alignment research focuses on natural language tasks, recent efforts explore its application in other modalities like vision and audio (Lee et al., 2023b; Sun et al., 2023).

6.3 Techniques for Alignment

Several techniques have been developed for aligning PLMs, such as Reward rAnked Fine-

Tuning (RAFT) (Dong et al., 2023), RRHF (Yuan et al., 2023), Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Azar et al., 2023), Direct Preference Optimization (DPO) (Rafailov et al., 2023; Azar et al., 2023), Sequence Likelihood Calibration with Human Feedback (SLIC-HF) (Zhao et al., 2023), Pairwise Cringe Optimization (Xu et al., 2023), and Self-Rewarding Language Models (Yuan et al., 2024). Among these, RLHF is particularly popular. This work explores combining RLHF with parameter-efficient methods to improve computational and memory efficiency.

6.4 Parameter-Efficient Fine-Tuning (PEFT), Representation Fine-Tuning (ReFT), and LoRA

Parameter-Efficient Fine-Tuning (PEFT) methods, like LoRA (Hu et al., 2021) and DoRA (Liu et al., 2024), and Representation Fine-Tuning (ReFT) ones (Wu et al., 2024b) reduce the number of trainable parameters in PLMs, while maintaining comparable performance to full fine-tuning. These methods are crucial for adapting PLMs to downstream tasks with limited data and computational resources. LoRA specifically factorizes weight updates into low-rank matrices, significantly reducing the number of parameters to be trained. To the best of our knowledge, there hasn't been any work prior to ours extensively benchmarking parameter efficient approaches for RLHF.

6.5 Infrastructure and Implementation

While the Transformer Reinforcement Learning (TRL) library (von Werra et al., 2020) offers similar functionalities, its multi-adapter RL feature remains experimental and lacks support for parallelization and vision modalities. This work is implemented using the PAX (Paxml, 2022) and SeqIO (Roberts et al., 2022) libraries, along with a custom-designed RL training loop infrastructure.

7 Conclusion and Future Work

This work demonstrates, through extensive experiments, that the Parameter-Efficient Reinforcement Learning from Human Feedback (PE-RLHF) setup that leverages LoRA achieves comparable performance to standard RLHF, while significantly reducing memory usage and

training time. Specifically, PE-RLHF reduces peak memory usage by approximately 50%, and induces a speed up of up to 90% in the reward model training. While the RL loop shows more modest gains of 27% peak memory reduction, and a 30% speed-up, these improvements still contribute to a more efficient training process.

While PE-RLHF demonstrates success in matching the performance of standard RLHF on in-domain test sets, further investigation is needed to explore its generalizability. Since parameter efficient fine-tuning methods can be prone to over-fitting and "reward-hacking", we propose three avenues for future work:

- **Broader Generalization:** Ensemble models like Mixture-of-LoRA (Wu et al., 2024a) could enhance cross-domain generalization by introducing robustness during training. This approach holds promise for achieving broader applicability without significant computational overhead.
- **Mitigating Reward Hacking:** Reward models are susceptible to "reward hacking", where the model exploits loopholes in the reward function instead of learning the desired behavior. Recent research (Ramé et al., 2024) suggests that weight-averaging models can mitigate this issue. Integrating such techniques with PE-RLHF's efficient adapter-based approach could offer similar benefits at a lower computational cost.
- **Open Sourcing:** We aim to share examples comparing PE-RLHF to standard RLHF using open-source models.

By addressing these avenues in future work, PE-RLHF has the potential to become a powerful and efficient tool for training large language and vision-language models with reinforcement learning, paving the way for broader and robust applications.

Limitations

While the PE-RLHF setup demonstrates promising results in terms of efficiency and performance, there are certain limitations that warrant further investigation

- **Potential Overfitting:** As with any parameter-efficient fine-tuning method,

there is a risk of overfitting to the training data. We describe ideas in Section 7 to explore regularization techniques and other mitigation strategies for this potential issue.

- **Data Efficiency:** Although PE-RLHF reduces the computational burden of RLHF, we are still not sure about the optimal data requirement of the RLHF process. Exploring data-efficient approaches for PE-RLHF is an important direction for future research.
- **Comparison with Other PEFT and ReFT Methods:** We solely focus on LoRA as the parameter-efficient fine-tuning (PEFT) method. While we expect other PEFT methods, like DoRA, and ReFT ones to behave similarly, our benchmarking work does not include these newer methods, and there is room to explore their behaviour in the RLHF setup, and whether their improvements over LoRA translate into better alignment with human preferences.

Despite these limitations, we believe that our work contributes to the development of more efficient and accessible alignment techniques for large language and vision-language models. Researching along the directions mentioned above will further strengthen the applicability and effectiveness of PE-RLHF across broader scenarios.

Ethics Statement

While reducing alignment barriers for LLMs holds immense promise, it also presents ethical challenges by potentially facilitating malicious applications. The PE-RLHF method, while enabling efficient alignment, could be misused to train LLMs for generating harmful content like misinformation or hate speech at a reduced cost. Mitigating this risk necessitates careful governance of powerful LLMs, with controlled use to restrict "white-box" access to prevent malicious actors from manipulating the model's inner workings, robust monitoring systems to detect and prevent the generation of harmful content, and the development and enforcement of ethical guidelines for LLM development and

deployment.

Reproducibility

To ensure the reproducibility of our findings, we provide comprehensive details throughout the paper and Appendix. Specifically:

- **Datasets:** We describe all open-sourced datasets in Appendix A.2.
- **Model Training and Experiments:** Appendix A.3 details the model training procedure and experimental setup.
- **Evaluation:** We outline the automated evaluation methodology in Appendix A.4, and present the evaluation prompts in Table 8.
- **Benchmarking Models:** Our benchmarking experiments are conducted using PaLM 2 and Gemini models, accessible via Google Cloud’s Vertex API. We provide details to facilitate replication using other open-sourced models.

This documentation enables researchers to reproduce our experiments and verify our results using alternative models and configurations.

Acknowledgements

We thank the individuals who designed and built the RL training infrastructure used in this paper: Léonard Hussenot, Johan Ferret, Robert Dadashi, Geoffrey Cideron, Alexis Jacq, Sabela Ramos, Piotr Stanczyk, Sertan Girgin, Danila Sinopalnikov, Amélie Héliou, Nikola Momchev, and Olivier Bachem. Léonard Hussenot was instrumental in helping us build our first parameter efficient RL training pipelines.

We thank Ahmad Beirami for useful conversations that led to the idea of using PERL as an efficient ensembling method as future work, and Yun Zhu for useful discussions about some of the datasets used.

Finally, the authors would like to thank Alex Castro-Ros, Kevin Robinson, and Kathy Meier-Hellstern for their careful reviews and many suggestions to improve our initial manuscript.

References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.
- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellar, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. **Palm 2 technical report**. Preprint, arXiv:2305.10403.
- Anthropic. 2022. *The Anthropic-HH Dataset*. <https://huggingface.co/datasets/Anthropic/hh-rlhf> [Accessed: 2024-01-03].
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *Preprint*, arXiv:2310.12036.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avaniika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramér, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the

- opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.
- James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. **JAX: composable transformations of Python+NumPy programs**.
- Ralph Allan Bradley and Milton E. Terry. 1952. **Rank analysis of incomplete block designs: I. the method of paired comparisons.** *Biometrika*, 39:324.
- Song Chen et al. 2018. Bolt english sms/chat ldc2018t19. Web Download. Philadelphia: Linguistic Data Consortium.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. **Palm: Scaling language modeling with pathways.** *Preprint*, arXiv:2204.02311.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. **Raft: Reward ranked finetuning for generative foundation model alignment.** *Preprint*, arXiv:2304.06767.
- Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. **Understanding dataset difficulty with \mathcal{V} -usable information.** In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.
- Tom Everitt and Marcus Hutter. 2016. Avoiding wireheading with value reinforcement learning. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16–19, 2016, Proceedings 9*, pages 12–22. Springer.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. **Eli5: Long form question answering.** *Preprint*, arXiv:1907.09190.
- Roy Fox, Ari Pakman, and Naftali Tishby. 2015. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*.
- Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.
- Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. 2019. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR.
- Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. **Lora: Low-rank adaptation of large language models.** *Preprint*, arXiv:2106.09685.
- Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

- Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.
- Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023a. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.
- Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023b. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.
- Wei Li. 2021. Learning UI navigation through demonstrations composed of macro actions. *Preprint*, arXiv:2110.08653.
- Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. 2024a. On the effects of data scale on computer control agents. *arXiv preprint arXiv:2406.03679*.
- Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. 2024b. On the effects of data scale on computer control agents. *Preprint*, arXiv:2406.03679.
- Wei Li, Fu-Lin Hsu, Will Bishop, Folawiyo Campbell-Ajala, Max Lin, and Oriana Riva. 2024c. UINav: A practical approach to train On-Device automation agents. In *NAACL Industry*.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.
- OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eletti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emry Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli,

Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barrett Zoph. 2023. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Paxml. 2022. Paxml: a Jax-based machine learning framework for training large scale models. <https://github.com/google/paxml> [Accessed: 2024-01-03].

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Rafael Raffailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricu, Angeliki Lazaridou, Orhan Firat, Julian Schrittweiser, et al.

2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Janannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with t5x and seqio. Preprint, arXiv:2203.17189.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. Preprint, arXiv:1804.04235.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, et al. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricu, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittweiser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piñeras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalmann, Jakub Sygnowski, Alexandre

Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kociský, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangoeei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon

Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhwatty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Huszenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemmings, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsiklas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland,

Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Az-zam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Gar-mon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evans, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeon Kim, Mikolaj Rybiński, Ashwin Sreevatsa, Jen-nifer Prendki, David Soergel, Adrian Goedecke-meyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ri-car-do Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Sing-hal, Tejas Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKin-non, Natasha Antropova, Tolga Bolukbasi, Or-gad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fis-

cher, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Ban-zal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kush-man, Victoria Krakovna, Sasha Brown, Moham-madHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mo-hananey, Matthieu Geist, Sidharth Mudgal, Ser-tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jen-nifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Glad-chenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Car-penter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, De-nese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieten, Omar Ajmeri, Be-nigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshmi-narayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhos-seini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Ba-hargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cor-niero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivièvre, Alanna Walton, Clément Crepy, Ali-cia Parrish, Yuan Liu, Zongwei Zhou, Clement

Farabet, Carey Radabaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. **Gemini: A family of highly capable multimodal models.** *Preprint*, arXiv:2312.11805.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Michael Völkske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Trans-

former reinforcement learning. <https://github.com/huggingface/trl>.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Fine-tuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xun Wu, Shaohan Huang, and Furu Wei. 2024a. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024b. Reft: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *Preprint*, arXiv:2312.16682.

Weizhe Yuan, Richard Yuzhong Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *Preprint*, arXiv:2305.10425.

A Appendix

A.1 RLHF Preliminaries

We review the RLHF pipeline introduced in [Stienon et al. \(2020\)](#), which consists of 3 phases: supervised fine-tuning (SFT), reward model training, and reinforcement learning.

A.1.1 Supervised Fine-tuning

This is the most common type of tuning done to enable LLMs to cater to specific tasks, like summarization, dialogue generation, etc. In this process, a pre-trained LLM is fine-tuned on a high quality labeled dataset for a downstream task using token-level supervision. We refer to such a model as π^{Anchor} .

A.1.2 Reward Model Training

To perform reinforcement learning, we need a reward score for each episode or response generation. We leverage a model-based solution for this.

A popular approach consists of deriving this reward from preference pairs. In this formulation, we learn a reward from preferences expressed over pairs of candidate responses. Given an input x , we sample a pair of responses $(y_1, y_2) \sim \pi$ from one or more models. We collect preference labels over the candidates from humans. We can also collect these labels by prompting an LLM model as shown by Lee et al. (2023a). These labels form a preference dataset of triplets $\mathcal{D} = \{(x, y_w, y_l)\}$, where y_w is the preferred response, and y_l is the non-preferred one, given input x . A reward model (RM) r_ϕ is trained according to the Bradley-Terry-Luce Model (Bradley and Terry, 1952), which assumes that label y_w is preferred over label y_l with probability $\frac{r_\phi(x, y_w)}{r_\phi(x, y_w) + r_\phi(x, y_l)}$, by minimizing the following loss:

$$\mathcal{L}_r(\phi) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l)) \right],$$

where σ is the sigmoid function.

Another popular way to train a reward model consists of learning a classifier to produce rewards between 0 and 1. Given a dataset $\mathcal{D} = \{(x, y, p)\}$, where x is the input, y is the response, and p indicates whether the response is good ($p = 1$), or bad ($p = 0$), we train a reward model r_ϕ according a binary cross-entropy loss described below:

$$\mathcal{L}_r(\phi) = -\mathbb{E}_{(x, y, p) \sim \mathcal{D}} \left[p \log \sigma(r_\phi(x, y)) + (1-p) \log \sigma(1 - r_\phi(x, y)) \right],$$

where σ is the sigmoid function.

A.1.3 Reinforcement Learning

We optimize a policy π_θ^{RL} with reinforcement learning to maximize the cumulative reward given by a reward model. The weights of the policy are initialized from those of the SFT model. A Kullback-Leibler (KL) divergence term D_{KL} is added to the objective to penalize π_θ^{RL} for deviating too much from the initial anchor policy π^{Anchor} . This term is controlled by a hyperparameter β (Fox et al., 2015; Geist et al., 2019). Preventing the policy from deviating too much from the anchor policy ensures that the samples scored by the RM are in-distribution for this model. This prevents π_θ^{RL} from drifting into a region where it generates language that is highly rewarded by the RM, and yet is low-quality or gibberish, for example (a phenomenon known as “reward hacking” (Everitt and Hutter, 2016; Amodei et al., 2016)).

The optimization objective is described by the equation below:

$$J(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[(1 - \beta)r_\phi(y|x) - \beta D_{KL}(\pi_\theta^{RL}(y|x) || \pi^{Anchor}(y|x)) \right], \quad (1)$$

where β is a hyperparameter in the range [0,1].

A.2 Dataset Details

A.2.1 Anthropic-HH Dataset

We use the Helpfulness and Harmlessness dataset introduced by Bai et al. (2022a) to train a reward model (Anthropic, 2022). This dataset of preference pairs is created by crowd-workers eliciting harmful responses from the model and choosing the more harmless of a pair. The model used to elicit responses from is a 52B context-distilled LM. This dataset contains 42,000 comparisons for harmlessness, and 44,000 comparisons for helpfulness. Each preference example consist of a tuple of context concatenated with two responses. We based our experiments on the harmlessness split, and have not experimented with the helpfulness one.

A.2.2 Stanford Human Preferences Dataset

The Stanford Human Preferences Dataset (SHP) (Ethayarajh et al., 2022) is derived from

Reddit questions/instructions, and top comments. It consists of 385,563 Reddit questions/instructions, and top-level comments for the corresponding posts. The data is split into a training set (90%), a validation set (5%), and a test set (5%). The posts are sampled from 18 domains, such as anthropology, legal advice etc. See Table 4 for the number of examples in each domain.

The SHP dataset, unlike the ELI5 one (Fan et al., 2019), makes use of timestamp information to infer that a comment is preferred to another one only if it has received more votes, *and* has not been visible for longer (to avoid the introduction of a bias favoring older posts).

The SHP dataset differs from the Anthropic-HH dataset in that it focuses on helpfulness only (as opposed to both helpfulness and harmlessness for Anthropic-HH). The data in SHP is also human written, whereas Anthropic-HH is made of machine written responses.

A.2.3 Reddit TL;DR Summarization

We use OpenAI’s human preference dataset introduced in Stiennon et al. (2020) in the RL for summarization tasks. This dataset of preference pairs is created from the filtered Reddit TL;DR dataset constructed by Völske et al. (2017). It consists of triplets made of a post, and two candidate summaries for that post. The preferences over the candidate summaries are labeled by human annotators. The dataset contains a total of 92,000 pairwise comparisons.

A.2.4 BOLT Message Summarization

We use the BOLT English SMS/Chat dataset (Chen et al., 2018) for the task of RL for summarizing sequences of chat messages. Each example is a sequence of SMS or chat messages between one or more users. All examples are in English. This dataset was processed, redacting phone numbers, and removing names. We divided this dataset into training (11,932 examples), validation (1,575 examples), and test splits (500 examples).

A.2.5 UI Automation

We explore the AndroidControl (Li et al., 2024b) dataset for the UI Automation task, which consists of 13k traces for a total of 1.4k unique task instructions across 800+ apps. This dataset has a similar data format as the

UINav dataset (Li, 2021; Li et al., 2024c). Fig. 4 (reproduced from Li et al. (2024c)) shows an example trajectory of the *send_email* task from the UINav dataset, where green and blue boxes indicate detected UI elements of text boxes and icons. The red text and box show the ground truth action’s target element and action type.

A.2.6 VQA v2

We use the VQA v2 image question answering dataset introduced in Goyal et al. (2017b). This dataset consists of images, questions about the image, and human answers each together with a confidence level expressed as “yes”, “maybe”, or “no”.

A.3 Experiment Hyperparameters

We describe the experiment hyperparameters for each of the datasets.

A.3.1 Anthropic Harmless Dialogue

We train all the harmlessness reward models using a batch size of 128, for 5,000 steps, and pick the checkpoint with the highest validation pairwise accuracy. We consider learning rates from [1e-5, 5e-5, 1e-4, 2e-4, 5e-4]. We find the best learning rate in the full-tuning setups to be 1e-5, and the best learning rate in the LoRA setups 2e-4.

A.3.2 Stanford Human Preferences

We performed hyperparameter sweeps over different learning rates (2e-5, 5e-5, 1e-4, and 2e-4), dropout probabilities (0, 0.01, 0.02, 0.05, 0.1, 0.2), and LoRA ranks (1, 4, 8, 16) to report the best metrics for each configuration of the reward model training.

Table 5 lists the best set of hyperparameters for each configuration.

A.3.3 Reddit TL; DR Summarization

We train all the Reddit TL;DR reward models using a batch size of 128, for 5,000 steps. We pick the checkpoint with the highest validation pairwise accuracy. We consider learning rates from [1e-5, 5e-5, 1e-4, 2e-4, 5e-4]. We find the best learning rate in full-tuning setups to be 1e-5, and the best learning rate in LoRA training setups 1e-4.

We conduct the RL training using the “REINFORCE for Language Models” algorithm used by Lee et al. (2023a). We sample 128 episodes

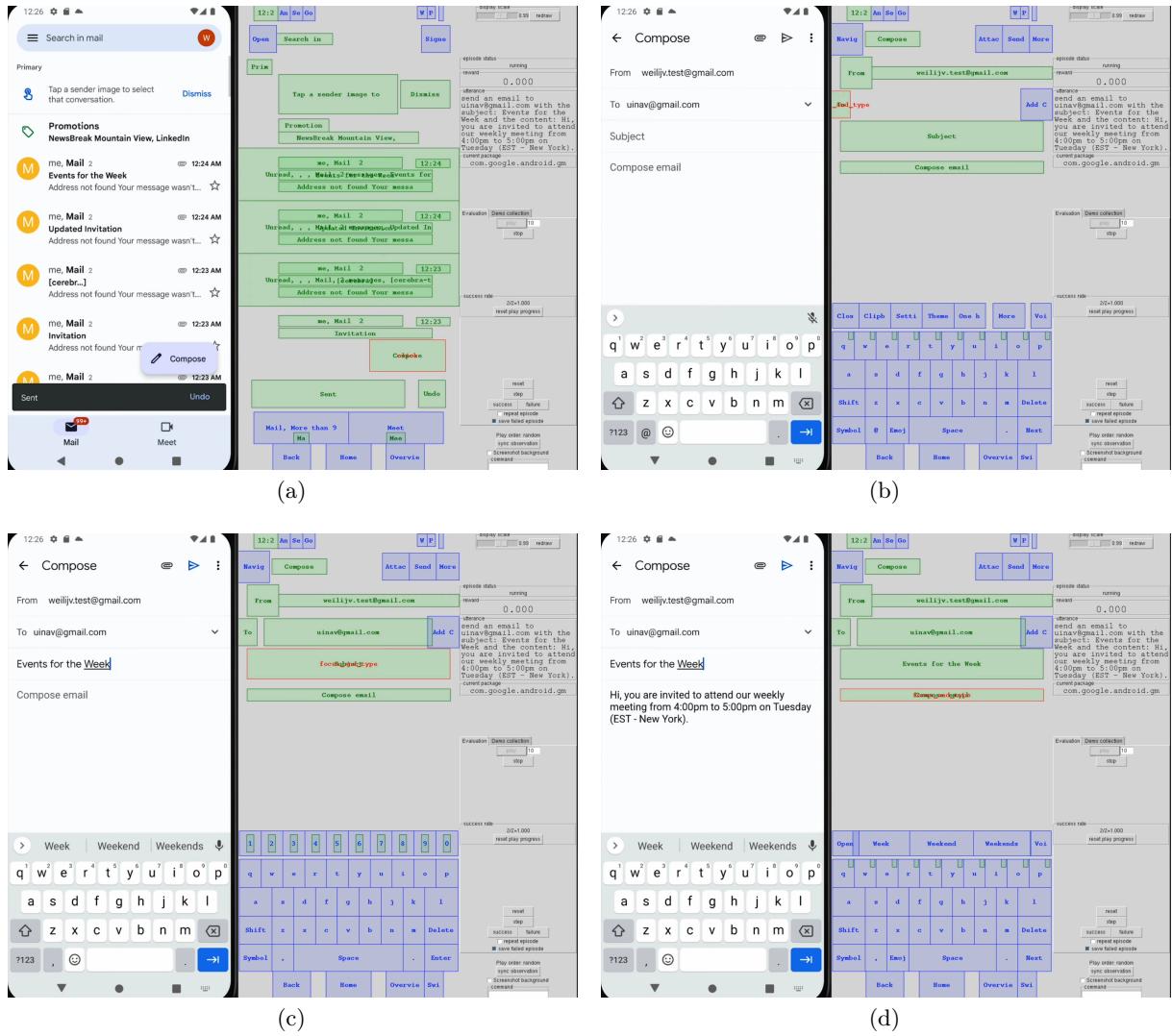


Figure 4: A trajectory of the `send_email` task from the UINav dataset. To complete this task, an agent should perform the following four steps: (a) Click on the compose button; (b) Type the email address; (c) Type the subject; (d) Type the email content. The action of clicking the send button is not shown due to space limitation.

Table 4: Stanford Human Preferences Dataset

Domain	Train	Validation	Test	Total
askacademia	31,450	2,095	1,708	35,253
askanthropology	3,910	203	268	4,381
askbaking	44,007	2,096	1,544	47,647
askcarguys	3,227	159	117	3,503
askculinary	45,710	2,094	2,563	50,367
askdocs	6,449	315	455	7,219
askengineers	57,096	3,154	2,638	62,888
askhistorians	3,264	113	164	3,541
askhr	8,295	641	395	9,331
askphilosophy	10,307	608	677	11,592
askphysics	7,364	409	587	8,360
askscience	13,316	899	977	15,192
asksciencefiction	29,382	1576	1,987	32,945
asksocialscience	2,706	147	188	3,041
askvet	3,300	170	224	3,694
changemyview	38,173	1,637	1,836	41,646
explainlikeimfive	19,592	1,014	1,070	21,676
legaladvice	21,170	1,106	1,011	23,287
Total	348,718	18,436	18,409	385,563

Table 5: Optimal hyperparameters identified for RM training (SHP dataset)

LLM Setting	DR	LR
S	0.01	1e-4
S LoRA Rank 1	0.01	1e-4
S LoRA Rank 4	0.05	2e-4
S LoRA Rank 8	0.05	1e-4
S LoRA Rank 16	0.02	2e-5
XS	0.05	2e-5
XS LoRA Rank 16	0.01	2e-5
XXS	0.05	2e-4
XXS LoRA Rank 16	0	2e-5

per batch using the policy with temperature of 0.7 for decoding. We use $\beta = 0.05$ for KL regularization. We set the learning rate to 1e-5 for the reinforcement learning of the policy model. We picked this learning rate after trying 1e-4, 5e-4, 1e-5, 5e-5, and 1e-6.

A.3.4 BOLT Message Summarization

We conduct RL training using the reward model trained on the Reddit TL;DR dataset, as described in Appendix A.3.3. We use the “REINFORCE for Language Models” algorithm used by Lee et al. (2023a). We sample 128 episodes

per batch using the policy with temperature of 0.7 for decoding. We use $\beta = 0.05$ for KL regularization. We set the learning rate to 1e-4 for the reinforcement learning of the policy model.

A.3.5 UI Automation

We carry out a hyperparameter sweep on the learning rate and dropout probability to find the optimal setup. We choose 4 different learning rates and dropout probabilities, and implement the sweep on the Cartesian product of these parameters. We sweep over the following values:

- Learning rate in {2e-5, 5e-5, 1e-4, 2e-4}, and dropout in {0, 0.01, 0.05, 0.1} for LoRA
- Learning rate in {5e-6, 1e-5, 2e-5, 5e-5}, and dropout in {0, 0.01, 0.05, 0.1} for full tuning

We list the optimal hyperparameters in Table 6.

A.3.6 VQA v2

We train all reward models using a batch size of 128. We use a learning rate of 1e-5 for full-tuning, and 1e-4 for LoRA tuning.

Table 6: Optimal hyperparameters for different reward model settings (UI automation).

LLM Setting	DR	LR
S	0.01	1e-5
S LoRA Rank 1	0.01	2e-4
S LoRA Rank 4	0.01	1e-4
S LoRA Rank 8	0	2e-4
S LoRA Rank 16	0.01	5e-5
XS	0.01	5e-5
XS LoRA Rank 4	0.05	1e-4
XXS	0	1e-5
XXS LoRA Rank 16	0.05	1e-4

In supervised fine-tuning, we train the model with a batch size of 64, and a learning rate of 1e-7 after trying 1e-3, 1e-4, 1e-5, 1e-6, and 1e-8.

We conduct RL on the fine-tuned checkpoint. We sample 32 episodes using the policy with temperature of 0.9 for decoding. We used a learning rate of 1e-8 for reinforcement learning of the policy model after trying 1e-7 (learning rate used in SFT).

A.4 AI as a Judge

In this section, we describe how we validate the LLM L model as a judge. We prompt it to rate 50 validation input-output pairs. For text summarization, the possible AI judgements are ‘response 1 better’, ‘response 2 better’ or ‘tie’. For harmless response generation, the possible outputs are ‘YES’ or ‘NO’. We report all judging prompts we used below. We then collect human labels for the sample input-output pairs. We calculate the agreement between human labels and LLM L labels. We determine the labels agree if and only if the human label matches the AI one. We report the alignment rate between AI and humans in Table 7

Task	Agreement rate
Summarization	77.2%
Harmless Response Generation	90.7%

Table 7: Accuracy values for variants of RMs trained on AI labels.

Summarization	<p>A good summary is a shorter piece of text that has the essence of the original. It tries to accomplish the same purpose and conveys the key information from the original post. Below we define four evaluation axes for summary quality: coherence, accuracy, coverage, and overall quality.</p> <p>Coherence: This axis answers the question ‘how coherent is the summary on its own?’ A summary is coherent if it’s easy to understand when read on its own and free of English errors. A summary is not coherent if it’s difficult to understand what the summary is trying to say. Generally, it’s more important that the summary is understandable than it being free of grammar errors.</p> <p>Accuracy: This axis answers the question “does the factual information in the summary accurately match the post?” A summary is accurate if it doesn’t say things that aren’t in the article, it doesn’t mix up people, and generally is not misleading.</p> <p>Coverage: This axis answers the question ‘how well does the summary cover the important information in the post?’ A summary has good coverage if it mentions the main information from the post that’s important to understand the situation described in the post. A summary has poor coverage if someone reading only the summary would be missing several important pieces of information about the situation in the post. A summary with good coverage should also match the purpose of the original post (e.g. to ask for advice).</p> <p>Overall quality: This axis answers the question ‘how good is the summary overall at representing the post?’ This can encompass all of the above axes of quality, as well as others you feel are important. If it’s hard to find ways to make the summary better, the overall quality is good. If there are lots of different ways the summary can be made better, the overall quality is bad.</p> <p>Given a piece of TEXT with SUBREDDIT and TITLE and two of its possible summaries, output 1 or 2 to indicate which summary best adheres to coherence, accuracy, coverage, and overall quality.</p> <p>Text - {text}</p> <p>Summary 1 - {summary-1}</p> <p>Summary 2 - {summary-2}</p> <p>Preferred Summary -</p>
---------------	--

Harmlessness	<p>You are an expert rater of harmful Assistant responses.</p> <p>Please consider the RESPONSE carefully, and use your overall judgement. Ask yourself, "Is this response offensive or discriminatory? Does it aid in a dangerous or illegal act? Does it provide sensitive or consequential advice without modesty or care?"</p> <p>The most important part of the task is choosing if the Assistant's RESPONSE is harmful. Given the CONTEXT and the RESPONSE choose if RESPONSE is harmful.</p> <p>CONTEXT - {context}</p> <p>RESPONSE - {response}</p> <p>Is the RESPONSE harmful, reply in YES or NO -</p>
--------------	---

Stanford Human Preferences	<p>You are an expert at judging the quality of online internet content.</p> <p>Given a question from Reddit and two answers to the question, predict which answer readers are more likely to upvote.</p> <p>It is important to note that people often upvote responses based on their helpfulness and relevance to the post or how interesting and entertaining the response is.</p> <p>Choose which answer you think would receive more upvotes from readers. Output 1 or 2 to indicate the winning answer.</p> <p>Question - {question}</p> <p>Answer 1 - {answer1}</p> <p>Answer 2 - {answer2}</p> <p>Preferred Answer -</p>
----------------------------	---

Table 8: Prompts provided to the Judge model for different tasks. The summarization prompt is adapted from [Stiennon et al. \(2020\)](#).