SCALABLE ENSEMBLING FOR MITIGATING REWARD OVEROPTIMISATION

Ahmed M. Ahmed, Rafael Rafailov, Stepan Sharkov, Xuechen Li, Sanmi Koyejo [*]
Department of Computer Science
Stanford University
{ahmedah, rafailov, stpshrkv, lxuechen, sanmi}@stanford.edu

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) has enabled significant advancements within language modeling for powerful, instruction-following models. However, the alignment of these models remains a pressing challenge as the policy tends to overfit the learned "proxy" reward model past an inflection point of utility as measured by a "gold" reward model that is more performant – a phenomenon known as *over-optimization*. Prior work has mitigated this issue by computing a pessimistic statistic over an ensemble of reward models, which is common in Offline Reinforcement Learning but incredibly costly for language models with high memory requirements, making such approaches infeasible for sufficiently large models. To this end, we propose using a shared encoder but separate linear heads. We find this leads to similar performance as the full ensemble while allowing tremendous savings in memory and time required for training for models of similar size.

1 INTRODUCTION

Modern language models have been ubiquitous in discussions of general purpose AI systems that can accomplish myriad tasks across many disciplines and with rapidly increasing capabilities OpenAI (2023); Manyika & Hsiao (2023); Touvron et al. (2023); Bommasani et al. (2022); Wei et al. (2022). However, alongside this increase in capabilities, there has been growing concern around the risks of such systems as they are not entirely interpretable and could be misused to cause substantial harm either maliciously or inadvertently (Hendrycks et al., 2023; Wang et al., 2023; Hendrycks et al., 2022).

A salient research question is that of *alignment*; given a set of human values, how do we imbue these principles within the behavior of such systems? Russell (2022). Suppose we can access a ground truth reward model. One observed phenomenon is that training a reward model from preference feedback eventually reaches an inflection point where increasing the policy performance on the proxy stagnates and ultimately degrades the reward credited by the ground truth model – even though both are consistent with human labels! (Gao et al., 2023).

This reward over-optimization problem has been identified as a significant technical issue in scaling up learning from human feedback Ouyang et al. (2022); Casper et al. (2023); Coste et al. (2023); Eisenstein et al. (2023). Prior work has noted that because the reward models trained from user feedback are only a proxy for their underlying preferences, opting for a high reward can lead to performance degradation. Recent work has proposed mitigations either through ensembling or transforming the RL objective but these methods are either too computationally expensive or analytically intractable to be used with modern language models given limited compute (Coste et al., 2023; Eisenstein et al., 2023).

In our work, we propose a simple modification to a common strategy of ensembling: instead of maintaining multiple separate reward models, we opt for a shared backbone with different linear heads with the hypothesis that different initialization and training procedure will generate enough

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

diversity. Our contributions show initial experiments validating the utility of such an approach in RLHF as it is just as performant as using a full ensemble for mitigating overoptimisation, while requiring less time for training during reward modeling, and less memory and time for PPO.

2 Related Work

Reinforcment Learning from Human Feedback (RLHF) RL is a powerful framework for learning diverse and performant policies that can tackle a wide array of tasks Sutton & Barto (2018); Abbeel & Ng (2004); Bellman (1957); Haarnoja et al. (2017); Drake (2005). Of particular recent interest are language models, which have shown impressive base capabilities in terms of instruction following. RLHF is a core component of the tuning processes for many of the currently highest-performing and most widely-used language models, improving their ability to follow instructions and align their responses with human preferences Stiennon et al. (2022); Christiano et al. (2023). However, tuning through RLHF can introduce a risk of overfitting to a proxy of true reward, since the reward model is learned. Prior work analyzes this phenomenon using a fixed "gold-standard" reward model in the place of humans, training proxy reward models from labels it provides (Gao et al., 2023).

Ensembles for Overoptimisation Recent work has shown multiple ways to mitigate this overoptimization issue, but the suggestions only apply during training or require expensive copies of multiple reward models for ensembling without deeper analysis as to why they mitigate this issue Coste et al. (2023). Related work further discovered that the ensembles are more effective when pretraining from scratch with separate seeds Eisenstein et al. (2023). Separately, other work has investigated uncertainty across ensemble members for RLHF showing that using the same backbone with different linear heads can lead to significant improvement in calibration, but that this weakly correlated with performance for summarization tasks Gleave & Irving (2022). We draw on this literature to propose a multi-head reward modeling scheme, for which we use each head as a separate reward function, utilizing the shared features from supervised fine-tuning but enabling the diversity of reward ensembles in a scalable manner.

3 Background & Methods

3.1 PPO

Proximal Policy Optimization (PPO) is an algorithm in reinforcement learning (RL) that is particularly adept at ensuring gentle policy updates, crucial for the complex dynamics of language models by optimizing reward with a KL constraint to the original policy.

$$\max_{\pi} \mathop{\mathbb{E}}_{s,a\sim\pi_{\text{old}}} \left[\frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} R(s,a) - \beta \text{KL}(\pi_{\text{old}}||\pi) \right] \tag{1}$$

3.2 Reward Learning

In standard reward modelings given a prompt completion x and a human pairwise binary preference label y, we instantiate the reward model using the backbone feature extractor from a pre-trained or model $\mathcal{F}$. We then initialize a linear reward head $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}$ with feature dim d giving a reward

$$r(x) = \mathcal{H}(\mathcal{F}(x)) \tag{2}$$

3.2.1 Multi-head Reward Learning

In our approach, the multi-head reward model is structured upon a shared base neural architecture derived from the pre-trained and supervised fine-tuned language model. **Everything is fixed except instead of a singular head we design the model to incorporate multiple heads**. Formally let $\mathcal{F}$ is the feature extractor and $\mathcal{H}_i$ is the i^{th} head linear head, the reward r_i for input x using head i can be described as:

$$r_i(x) = \mathcal{H}_i(\mathcal{F}(x)) \tag{3}$$

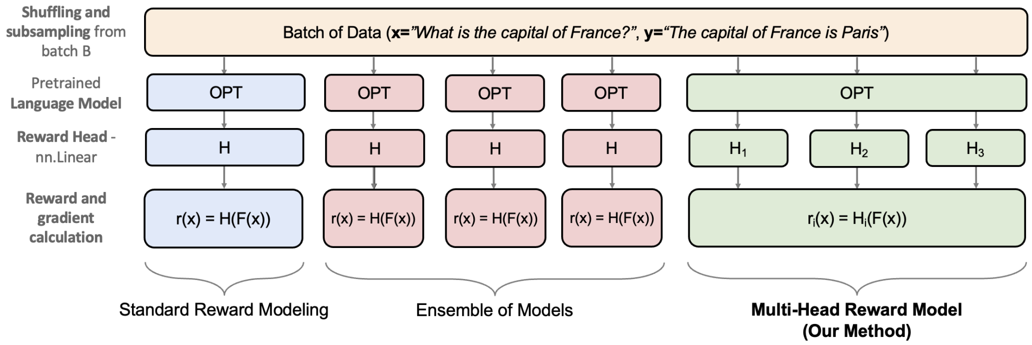
Shuffling and subsampling from batch B
Pretrained Language Model
Reward Head - nn.Linear
Reward and gradient calculation
Batch of Data (x="What is the capital of France?", y="The capital of France is Paris")
OPT
H
H_1
H_2
H_3
r(x) = H(F(x))
$r_i(x) = H_i(F(x))$
Standard Reward Modeling
Ensemble of Models
Multi-Head Reward Model (Our Method)

Figure 1: Comparison of Reward Modeling methods to ours (right)

4 Experiments

4.1 Datasets and Motivation

We select the **Alpaca Instructions** dataset which emphasizes naturalistic, closed-form questions and answers, usually having a well-defined correct response, e.g., the number of planets in the solar system, but also included more open questions such as famous actors who started on broadway, why certain states are named, how to play kickball, etc. Taori et al. (2023). Given the most relevant prior work studying overoptimisation focused on this dataset, it is our primary dataset for training and evaluation (Coste et al., 2023).

4.2 Methodology and Tools

We follow the RLHF pipeline of supervised fine-tuning (SFT), reward learning, and proximal policy optimization (PPO). The experiments are implemented using the OPT model family Zhang et al. (2022). We opted for the Alpaca Farm codebase to serve as our framework (Dubois et al., 2023).

- **Supervised Fine-Tuning (SFT):** Given a prompt (”Tell me a bedtime story”) and some ideal completions, the base model is fine-tuned to minimize perplexity on a split of 52k instructions.
- **Reward Learning:** Given the SFT model and the same set of prompts but now with pairs of completions (preferred and dispreferred), we use the backbone and fine-tune a linear head on the Bradley-terry loss with the preferred completion as the target (Bradley & Terry, 1952).
- **PPO** Finally, we further tune the SFT language model as a policy in the RL framework against the reward model through PPO.

We modify this pipeline by training the multi-head reward models with the base model initialized from SFT. When selecting the reward for a given sample, since we produce multiple predictions, we are presented with multiple choices for computing our final prediction Coste et al. (2023), but we **simply take the minimum over the ensemble**. A pessimistic estimator might reduce performance in principle, although for offline RL and RLHF preferences, it has been either conjectured or empirically validated that a minimum is optimal, and prior work demonstrated uncertainty penalties do not improve the efficiency of ensembles in preventing over-optimization over a simple min Zhu et al. (2023); Coste et al. (2023); Kumar et al. (2020); Chen et al. (2021). Our approach is shown in Figure 1, and all hyperparameters and further training details are in the Appendix.

4.3 Results

To test the efficacy of our approach, we run 1) Standard PPO 2) PPO with an ensemble of three reward models and 3) PPO with a multi-head reward model with three linear heads. We fix ensemble size at three as prior work found no gains when increasing to four or five members due to fixed

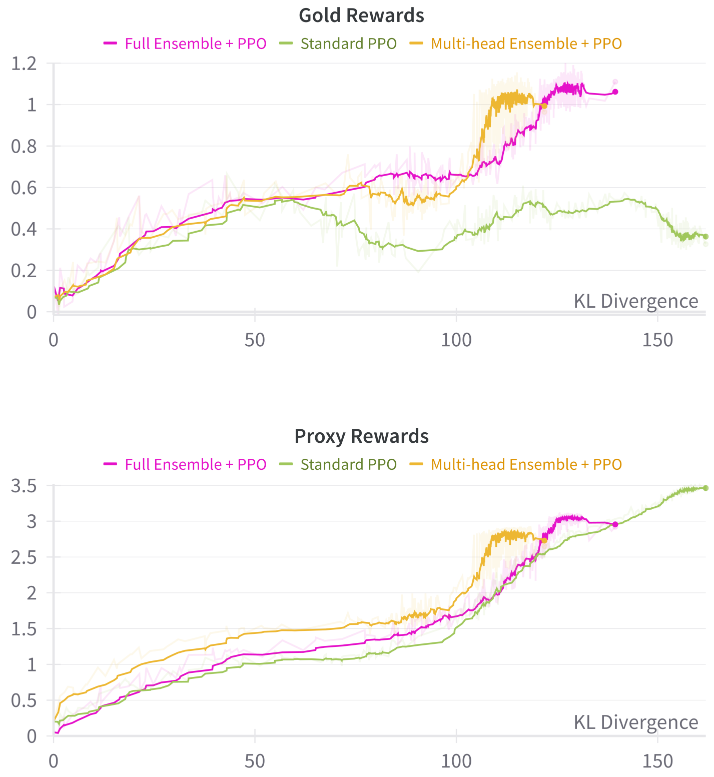

Figure 2: Gold analysis on top, Proxy metrics below.

compute, although we emphasize our method is amenable to a much larger number of ensemble members with minimal overhead Coste et al. (2023). We select the 1.3 B parameter OPT model as the proxy and the 6.7B variant as the gold model and run PPO for 15 epochs to extensively test overoptimisation.

We present our results in Figure 6. Given the difference in reward scales, and multiple works validating the benefit of using a full ensemble for overoptimisation we opt for two figures that demonstrate 1) the efficacy of the multi-head ensemble against standard PPO and 2) the efficacy of using a multi-head or full ensemble. Following prior work we also plot reward against KL divergence Gao et al. (2023); Coste et al. (2023); Eisenstein et al. (2023). Figure 6 clearly shows the replication of overoptimisation as we reproduce the concave-down nature of gold rewards under standard PPO and then summarily show how using a multi-head reward model bridges this gap. Figure 6 shows that the multi-head and full ensemble nature to similar gold performance, however we emphasize that our approach allows an ensemble with much larger reward models and following the prior work suggesting the full ensemble **we train each member for three epochs** to maximize performance whereas we find the **multi-head approach needs only one epoch**. We include further details and ablations in the appendix.

4.4 Calibration

4.5 Calibration Analysis and Implications

We investigate the calibration of our multi-head reward models, as accurate representation of uncertainty is crucial in RLHF to prevent overoptimistic predictions Casper et al. (2023). Prior work found that ensembling improves calibration but is weakly related to model error when using a shared backbone with different linear heads, suggesting that separate reward models might be necessary for efficient ensembling Gleave & Irving (2022). However, we hypothesize that the difference in improvement is due to the diversity of tasks, as the AlpacaFarm dataset focuses on more open-ended

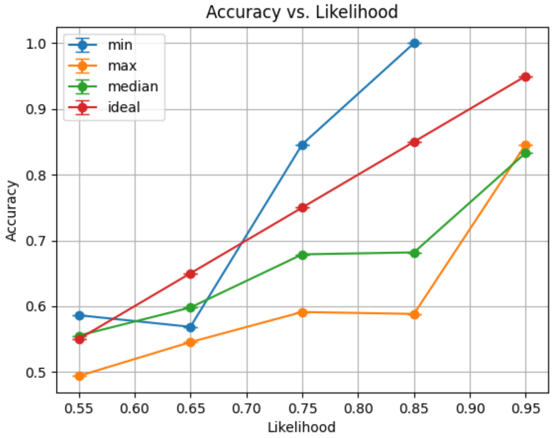

Figure 3: Multi-head model calibration over different objectives with respect to the probabilities (taking min, max over ensemble etc.)

questions compared to the summarization tasks in the prior work, which might increase model error due to epistemic uncertainty. We assess the calibration quality of the models using different ensemble objectives, such as mean and minimum reward estimation. The multi-head models, particularly with the minimum objective, demonstrate an enhanced ability to capture uncertainty in the reward signals without incurring significant computational overhead. While the minimum estimate is technically mis-calibrated, it leads to better than predicted performance at high levels of certainty. This suggests a regularization effect from the pessimistic estimator, as overoptimized models typically reach a local minima, explaining the ”U-shaped” curves that result as KL divergence from the base policy increases. Our findings align with previous work on pessimistic estimates in ensemble models, underscoring the benefits of a conservative approach in certain RL contexts. Furthermore, our experiments suggest that a smaller number of heads (3) might be optimal for mitigating overoptimisation, providing a valuable guideline for future research in ensemble-based models for RLHF. An increase in the number of heads beyond a certain threshold might introduce more noise than beneficial diversity, potentially leading to overfitting on dataset nuances, which is consistent with existing literature on the effectiveness of smaller ensembles in specific scenarios Galdran et al. (2023). See Figure 4 for details.

4.6 Conclusion

Our research contributes a novel method to improve the robustness of aligning language models by utilizing a robust pessimistic statistics while avoiding the cost of materializing a full ensemble. By addressing the challenges of overoptimization in a computationally efficient manner, we move a step closer to developing AI systems that can reliably align with human values and preferences.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML ’04, pp. 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL `https://doi.org/10.1145/1015330.1015430`.

Richard E Bellman. *Dynamic programming.* Princeton University Press, Princeton, NJ, USA, 1957.

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, Avanika 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 Tramèr, 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. On the opportunities and risks of foundation models, 2022.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model, 2021.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization, 2023.

Alvin Drake. Observation of a markov process through a noisy channel. 08 2005.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking, 2023.

Adrian Galdran, Johan Verjans, Gustavo Carneiro, and Miguel A. González Ballester. Multi-head multi-loss model calibration, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models, 2022.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies, 2017.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety, 2022.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning, 2020.

James Manyika and Sissie Hsiao. An overview of bard: An early experiment with generative ai. https://ai.google/static/documents/google-about-bard.pdf, 2023. Accessed: 2023-11-26.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback, 2022.

Stuart Russell. *Artificial Intelligence and the Problem of Control*. Springer, 2022.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model, 03 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html. Accessed: insert date you accessed here.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or -wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023. URL https://arxiv.org/abs/2301.11270.

A Appendix

A.1 D Additional experimental details

A.1.1 D.1 Hyperparameters

We give the hyperparameters here for different components of our RLHF pipeline:

Table 1: SFT hyperparameters for 6.7B model

Parameter	Value
Learning rate	2e-5
Epochs	3
Batch size	128

Table 2: SFT hyperparameters for 1.3B

Parameter	Value
Learning rate	8e-6
Epochs	1
Batch size	128

A.1.2 D.2 AlpacaFarm Dataset Details

The AlpacaFarm dataset Dubois et al. (2023) employed in our experiments uses the Alpaca data Taori et al. (2023) made up of 52,000 samples. This data is chosen due to its large size and success in training instruction-following models. AlpacaFarm contains five splits: a labeled 10k ”sft” split for supervised fine-tuning, a 10k ”pref” split containing pairwise preference labels, a 20k ”unlabeled” split for training algorithms such as PPO, a 2k validation split, and an unused 10k split. We use the noisy variant of alpaca instructions with 25% label noise to more closely model real-world data distributions.

A.1.3 Reward model training ablation

Interestingly, we find that if we only train each reward model in the full ensemble for one epoch the ensemble approach is not sufficient to prevent over-optimisation, and that at least three epochs are necessary while prior work had five or more. We hypothesize that the multi-head approach is more amenable to underspecified features for reward modeling.

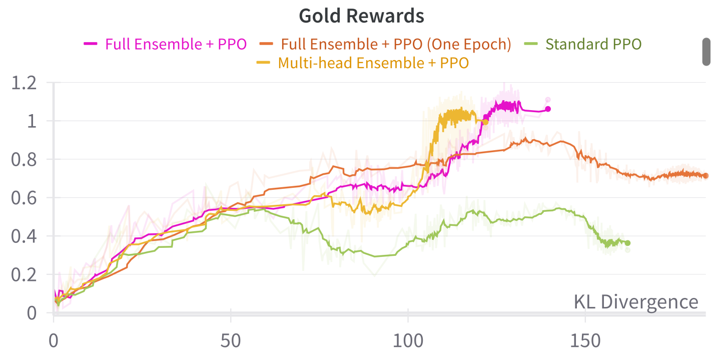

Figure 4: Gold analysis.

Table 3: RM hyperparameters

Parameter	Value
Learning rate	1e-5
Epochs	3
Batch size	64

Table 4: PPO hyperparameters

Parameter	Value
Max instruction length	520
Max new tokens (answer length)	256
PPO epochs	4
Top-p	0.9 (1.0 for PPO training)
Top-k	0
Temperature	1.0
Rollout Batch size	512
Gradient Step Batch size	256
Learning Rate	6e-6

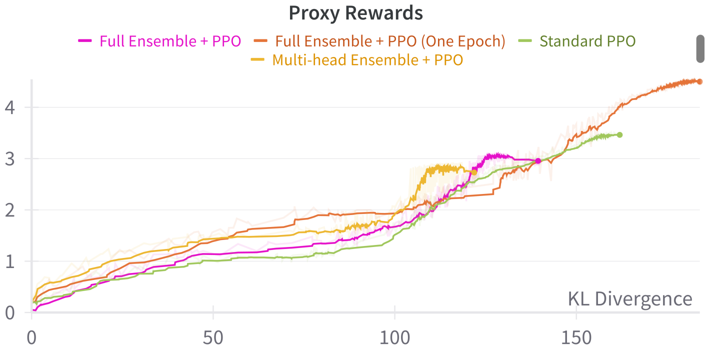

Figure 5: Proxy metrics.

Figure 6: Gold analysis on top, Proxy metrics below.