The Accuracy Paradox in RLHF: When Better Reward Models Don't Yield Better Language Models

Yanjun Chen[1,3], **Dawei Zhu**[2], **Yirong Sun**[3], **Xinghao Chen**[1,3], **Wei Zhang**[3*], **Xiaoyu Shen**[3*]
[1] Department of Computing, The Hong Kong Polytechnic University
[2] Saarland University, Saarland Informatics
[3] Digital Twin Institute, Eastern Institute of Technology, Ningbo, China
yan-jun.chen@connect.polyu.hk {zhw,xyshen}@eitech.edu.cn

Abstract

Reinforcement Learning from Human Feedback significantly enhances Natural Language Processing by aligning language models with human expectations. A critical factor in this alignment is the strength of reward models used during training. This study explores whether stronger reward models invariably lead to better language models. In this paper, through experiments on relevance, factuality, and completeness tasks using the QA-FEEDBACK dataset and reward models based on Longformer, we uncover a surprising paradox: *language models trained with moderately accurate reward models outperform those guided by highly accurate ones*. This challenges the widely held belief that stronger reward models always lead to better language models, and opens up new avenues for future research into the key factors driving model performance and how to choose the most suitable reward models.

1 Introduction

Language models (LMs) have made remarkable progress, achieving close-to-human capabilities in a wide range of tasks (Shen et al., 2017; Radford et al., 2019; Brown et al., 2020; Su et al., 2022; Achiam et al., 2023; Yuan et al., 2024). While traditional fine-tuning has been effective, it often suffers from exposure bias, where models are trained on ground truth data rather than their own predictions, leading to inconsistencies during generation (Shen et al., 2019; Wang and Sennrich, 2020). Additionally, fine-tuning lacks the ability to optimize for sequence-level rewards, limiting its effectiveness in capturing complex, human-like preferences (Zhu et al., 2024). RLHF addresses these limitations by incorporating feedback from humans, allowing models to generate more contextually relevant and aligned outputs (Stiennon et al., 2020; Ouyang et al., 2022; Su et al., 2024; Madaan et al., 2024).

It is commonly assumed that higher accuracy in reward models enhances language model performance because these models provide precise feedback during training (Chaudhari et al., 2024). This perspective suggests that accurate feedback directly improves the effectiveness of LMs, especially in complex tasks like machine translation and question answering (Bai et al., 2022).

In this paper, we conducted extensive experiments using the QA-FEEDBACK dataset (Wu et al., 2024). Reward models based on Longformer (Beltagy et al., 2020) were evaluated for their binary classification accuracy in predicting task relevance, factuality, and completeness. To ensure fair evaluation, the performance of LMs trained with these reward models was assessed using independent high-accuracy models tailored to each task. Surprisingly, our findings reveal a paradox: LMs achieve their best performance *not with the most accurate reward models, but with those of moderate accuracy* (Casper et al., 2023), challenging the prevailing assumption that higher reward model accuracy directly correlates with improved outcomes. This result raises important questions about the relationship between reward model accuracy and language model performance in RLHF, warranting further investigation.

The main contributions of this study include:

- Demonstrating that moderate reward model accuracy and balanced training lead to better language model performance, contradicting the assumption that higher accuracy is invariably beneficial.
- Providing insights into reward dynamics, revealing that moderately accurate reward models offer more task-appropriate rewards, which are intuitively more beneficial for training LMs than those provided by the most accurate models.

*Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2980–2989
November 12-16, 2024

- Analyzing KL divergence trends, showing that moderately accurate reward models facilitate a balanced and stable training process, promoting better generalization and challenging the notion that higher accuracy alone ensures optimal training outcomes.

2 Motivation and Problem Settings

Motivation. Findings indicate that the strength of reward models in RLHF does not consistently correlate with improved language model performance, challenging the assumption that stronger reward models always lead to better outcomes (Casper et al., 2023). Understanding the dynamic relationship between reward model accuracy and language model performance is essential for optimizing RLHF in complex NLP tasks (Ouyang et al., 2022). This study posits that there exists an optimal range of reward model accuracy that maximizes language model performance (Wu et al., 2024). Therefore, the primary aim of this research is to identify this optimal range and examine its implications for various NLP applications.

Problem Settings. This study investigates the effect of reward model strength on language model performance in RLHF, focusing on tasks that evaluate the factuality, relevance, and completeness of generated text (Wu et al., 2024). Specifically, reward model strength is defined by binary classification accuracy on test sets (Wu et al., 2024), and language model performance is measured using high-accuracy, independent reward models.

Formally, for a language model trained with RLHF, this study analyzes how the reward model's classification accuracy ($\mathcal{S}_{\text{RM}}$) and the number of training steps (τ) affect language model performance ($\mathcal{P}_{\text{LM}}$) (Qin et al., 2024). This relationship is mathematically represented by:

$$\mathcal{P}_{\text{LM}} = f(\mathcal{S}_{\text{RM}}, \tau) \quad (1)$$

The objective is to determine the optimal conditions that maximize language model performance across various tasks, providing insights for the development of more effective RLHF strategies in NLP (Li et al., 2023).

3 Experiment and Results

3.1 Basic Experimental Setup

Models. We examine three models from the T5 language model family (Raffel et al., 2020; Kaplan et al., 2020): T5-small[1], T5-base[2], and T5-large[3]. Each model underwent supervised fine-tuning (SFT). Reward models were based on Longformer-base-4096, suitable for processing long sequences, necessary for tasks requiring extensive context (Beltagy et al., 2020). These models were trained for tasks involving factuality, relevance, and completeness, with training steps and accuracy ranges summarized in Table 1.

Task Type	Steps Range	Accuracies Range
Factuality	2–1256	0.64–0.77
Relevance	2–2852	0.49–0.69
Completeness	30–5730	0.44–0.70

Table 1: Training steps and accuracy ranges for reward models by task type.

Datasets. The QA-FEEDBACK dataset (Wu et al., 2024), derived from the ASQA dataset (Stelmakh et al., 2022), is used for this study. This dataset focuses on generating long-form answers to ambiguous factual questions in an open-domain setting. The data is split into 3,853/500/948 for training, validation, and testing, requiring the generation of detailed answers from multiple knowledge passages (Min et al., 2020).

Hyperparameter Settings. We follow the hyperparameter settings recommended by Wu et al. (Wu et al., 2024), whose configuration has been specifically designed and empirically validated for RLHF tasks involving QA-feedback. These settings are selected to ensure an optimal trade-off between model performance and training stability, based on prior experimental findings. For a detailed description of all hyperparameters used in the experiments, please refer to Appendix D.

Training and Evaluation Paradigm. Following common practice (Schulman et al., 2017), we begin by fine-tuning LMs, followed by applying RLHF using Proximal Policy Optimization (PPO). In addition, a separate instance of the T5-base model was specifically initialized as the value model for the PPO algorithm. Finally, we evaluate the trained LMs using three independent, highly accurate reward models, which assess various aspects of the LMs' outputs, including relevance, factuality, and

[1] https://huggingface.co/t5-small
[2] https://huggingface.co/t5-base
[3] https://huggingface.co/t5-large

completeness. A summary of the reward models' performance is provided in Table 2.

Reward Model	Accuracy (%)	F1 Score (%)
Rϕ_1 (Relevance)	69.6	68.5
Rϕ_2 (Factuality)	77.8	67.5
Rϕ_3 (Completeness)	70.9	N/A

Table 2: Summary of independent high-accuracy reward models used for evaluation.

A common pitfall in performing RLHF is reward gaming, where LMs maximize rewards in unintended ways, such as finding shortcuts in generation that attain high reward scores from the reward models, yet misalign with human preferences (Pang et al., 2022). To mitigate this, we following (Wu et al., 2024) and set a KL threshold. When the divergence between the current policy and the reference policy exceeded this threshold, the training process was interrupted. This approach ensured that the model did not deviate excessively from the reference policy, effectively reducing the likelihood of reward manipulation.

3.2 Are High-Accuracy and Deeply Trained Reward Models Always the Best?

Setup. Building on the Basic Experimental Setup, reward models for relevance, factuality, and completeness from the QA-FEEDBACK dataset were used in PPO training. Performance was assessed at regular intervals, and top-performing instances were identified and visualized in three-dimensional plots.

Results. Figures 1 to 3 show that optimal language model performance is achieved using reward models with moderate accuracy and an appropriate number of trained steps. For the relevance task, the T5-small model performed best with moderately accurate reward models, effectively mitigating the risk of overfitting. Similarly, the results for factuality emphasized the importance of maintaining balanced reward model accuracy to prevent overfitting and ensure reliable outcomes. These findings suggest that overly accurate reward models can result in overfitting, which impairs the generalization ability of LMs. These trends were consistent across the T5-base and T5-large models, further supporting the conclusion that moderate accuracy in reward models strikes the best balance between training stability and performance. Detailed results for T5-base and T5-large are available in the Appendix A.

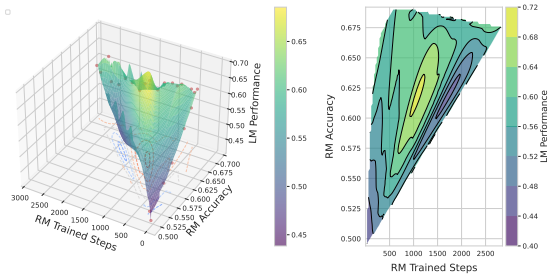

Figure 1: 3D surface plot evaluating relevance ratios for T5-small. Optimal performance was achieved with reward models having moderate accuracy.

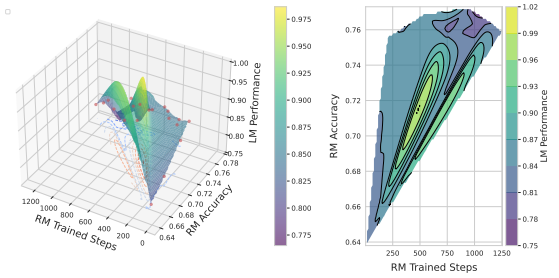

Figure 2: 3D surface plot evaluating factuality ratios for T5-small. The best performance was seen with reward models of moderate accuracy.

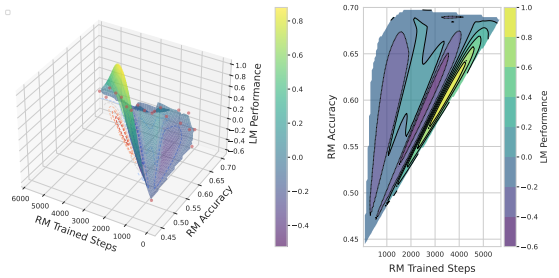

Figure 3: 3D surface plot evaluating completeness rewards for T5-small. Intermediate reward model strength yielded the best language model performance.

3.3 How Do Best and Most Accurate Reward Models Differ?

Setup. This evaluation utilized three models: T5-small, T5-base, and T5-large, to compare the best-performing and most accurate reward models across relevance, factuality, and completeness tasks. The analysis focused on understanding the differences in reward behavior during training for each model. While the primary analysis in this section is based on the T5-small model, similar trends were observed with the T5-base and T5-large models, whose results are provided in the appendix for reference.

Results. Figures 4, 5, and 6 illustrate the distinct strategies of the best-performing reward models compared to the most accurate models using the T5-small model. For the relevance task, the best-performing reward model provided higher and more variable rewards (Figure 4), indicating an aggressive approach that likely stimulated the generation of more relevant outputs. In the factuality task, this model maintained higher mean rewards with less variability (Figure 5), promoting factual accuracy. Conversely, for the completeness task, it employed a conservative strategy with lower average rewards but greater variability (Figure 6).

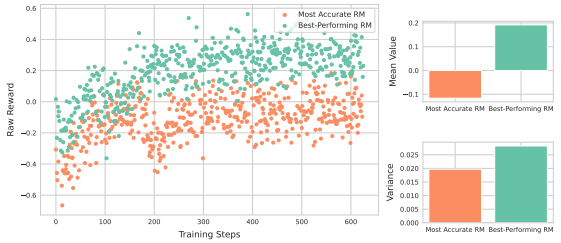

Figure 4: Reward analysis for relevance task (T5-small model): training steps vs. rewards (left), mean and variance of rewards (right).

Figure 5: Reward analysis for factuality task (T5-small model): training steps vs. rewards (left), mean and variance of rewards (right).

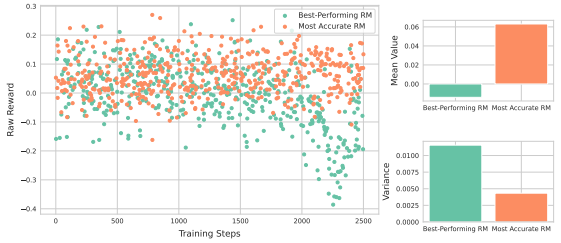

Figure 6: Reward analysis for completeness task (T5-small model): training steps vs. rewards (left), mean and variance of rewards (right).

Analysis. Moderately accurate best-performing reward models typically align rewards with task requirements. In both relevance and factuality tasks, these models provide higher and more varied rewards, thus encouraging the generation of more relevant and accurate outputs. This variability allows LMs to explore a broader range of responses, improving the quality of the generated text. Conversely, in completeness tasks, a conservative strategy with lower average rewards but greater variability helps ensure thorough and comprehensive text evaluation. The trends observed in T5-small models are consistent with those seen in T5-base and T5-large models, further supporting the conclusion that moderate accuracy in reward models effectively balances overfitting and underfitting. Detailed results for T5-base and T5-large can be found in the Appendix B.

3.4 How Do Best and Most Accurate Rewards Impact Models?

Setup. This section evaluates the impact of reward models on the training dynamics of T5-small, T5-base, and T5-large models in relevance, factuality, and completeness tasks, with a focus on KL divergence trends to assess stability and adaptability. While the results presented here focus on the T5-small model, similar trends were observed for the T5-base and T5-large models, whose results are provided in the appendix.

KL Divergence and Its Role in RLHF KL divergence (Kullback-Leibler divergence) is a measure of how one probability distribution P diverges from a second, expected probability distribution Q (Kullback and Leibler, 1951). It is commonly used in reinforcement learning to constrain the difference between the current policy and a reference policy during training. Mathematically, KL divergence is defined as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \quad (2)$$

In the context of RLHF, KL divergence serves as a regularization term to prevent the trained policy from deviating excessively from the reference policy. This constraint helps to stabilize the training process by reducing the chance of reward hacking or reward gaming, where the model could exploit the reward system without truly improving performance (Pang et al., 2022).

Results. Comparing KL divergence trends revealed significant differences in how LMs aligned with the training data. For the relevance task, the

best reward model resulted in consistently lower KL divergence and variance, indicating stable alignment (Figure 7). In the factuality task, the best reward model exhibited higher mean KL divergence but lower variance, suggesting a consistent yet varied alignment process (Figure 8). For the completeness task, the best reward model showed higher mean and variance in KL divergence, indicating a flexible approach suitable for evaluating complex texts (Figure 9).

Figure 7: Relevance task KL divergence (T5-small model): training steps vs. KL divergence (left), mean and variance of rewards (right).

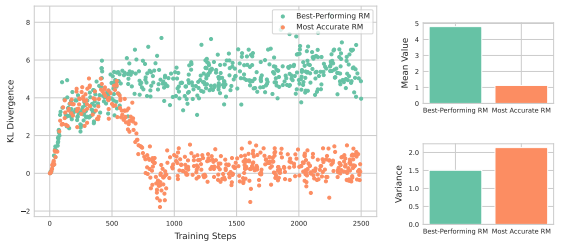

Figure 8: Factuality task KL divergence (T5-small model): training steps vs. KL divergence (left), mean and variance of rewards (right).

Figure 9: Completeness task KL divergence (T5-small model): training steps vs. KL divergence (left), mean and variance of rewards (right).

Analysis. Best-performing reward models, which are typically of moderate accuracy, create a balanced training environment that facilitates both stability and adaptability. In relevance and factuality tasks, these models encourage stable learning, enhancing the relevance and accuracy of outputs. For the completeness task, the flexibility in handling complex texts is demonstrated by higher variance in KL divergence. The observed trends in T5-small models were consistent with those seen in T5-base and T5-large models, further validating the conclusion that moderate accuracy in reward models effectively balances overfitting and underfitting. Detailed results for T5-base and T5-large models can be found in the Appendix C.

4 Conclusion and Future Work

This study demonstrates that LMs trained with moderately accurate reward models in RLHF achieve optimal performance, challenging the conventional belief that higher accuracy is always more beneficial. The results show that moderately accurate reward models offer more task-aligned feedback and foster a balanced, stable training process, promoting better generalization. This research highlights the limitations of relying exclusively on highly accurate reward models, as excessive focus on accuracy may lead to suboptimal outcomes. In future work, it will be crucial to further explore the potential overfitting of reward models, particularly in their ability to generalize to out-of-distribution (OOD) tasks. Techniques such as regularization, data augmentation, and explicit OOD evaluation will be key areas of investigation to enhance the robustness of reward models across diverse scenarios and ensure their effectiveness in guiding LMs in broader, more complex NLP tasks.

Limitations

Dataset Constraints. The conclusions are drawn from the QA-FEEDBACK dataset (Wu et al., 2024), which is specialized in generating long-form responses to factual inquiries. This focus may limit the generalizability of the results, necessitating validation across various datasets, including those pertaining to conversational and question-answering contexts.

Model Scope. The evaluation utilized T5 models of different scales for initial validation (Raffel et al., 2020). Future investigations should incorporate more complex models, such as Llama2 (Touvron et al., 2023), to gain deeper insights and verify the robustness of the proposed methodologies across a broader range of model architectures.

Reward Model Variations. This study did not explore the impact of different reward model sizes

and architectures on RLHF performance. The reward models used were based on a single architecture, which may limit the applicability of the findings. Future research should systematically investigate how variations in reward model size, capacity, and design affect the learning process, generalization, and overall RLHF performance, particularly in diverse NLP tasks. Understanding the influence of these factors will be crucial for developing more robust and scalable reward models that can generalize across a wider range of applications.

Acknowledgements

We thank Xingluan (AI Cloud computing service), EIT and IDT High Performance Computing Center for providing computational resources for this project. This work is supported by 2035 Key Research and Development Program of Ningbo City under Grant No.2024Z127.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Forty-first International Conference on Machine Learning*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *Cornell University - arXiv,Cornell University - arXiv*.

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.

Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur P Parikh, and He He. 2022. Reward gaming in conditional text generation. *arXiv preprint arXiv:2211.08714*.

Bowen Qin, Duanyu Feng, and Xi Yang. 2024. Towards understanding the influence of reward margin on preference model performance. *arXiv preprint arXiv:2404.04932*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. 2017. Estimation of gap between current language models and human performance. In *INTERSPEECH*, pages 553–557.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

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.

Hui Su, Zhi Tian, Xiaoyu Shen, and Xunliang Cai. 2024. Unraveling the mystery of scaling laws: Part i. *Preprint*, arXiv:2403.06563.

Hui Su, Xiao Zhou, Houjin Yu, Xiaoyu Shen, Yuwen Chen, Zilin Zhu, Yang Yu, and Jie Zhou. 2022. Welm: A well-read pre-trained language model for chinese. *arXiv preprint arXiv:2209.10372*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

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

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024. A preference-driven paradigm for enhanced translation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3385–3403.

A Supplementary 3D Surface Plots

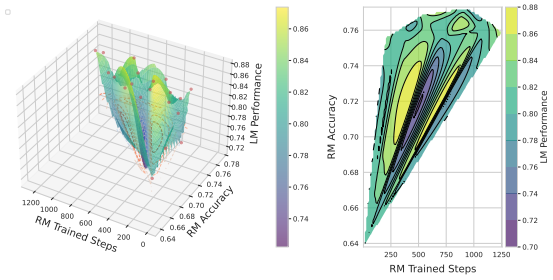

Figure 10: 3D surface plot evaluating factuality ratios for T5-base. Optimal performance was achieved with reward models having moderate accuracy.

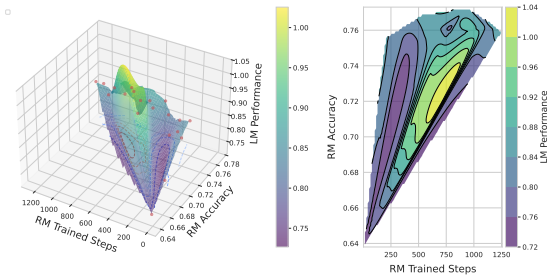

Figure 11: 3D surface plot evaluating factuality ratios for T5-large. The best performance was seen with reward models of moderate accuracy.

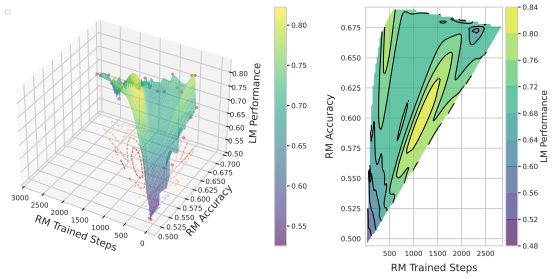

Figure 12: 3D surface plot evaluating relevance ratios for T5-base. Optimal performance was achieved with reward models having moderate accuracy.

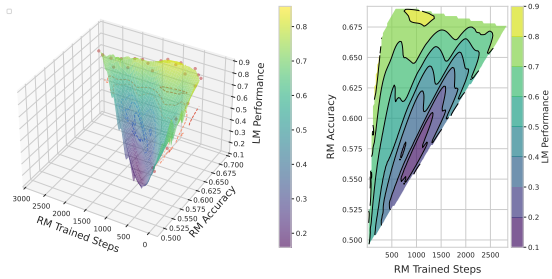

Figure 13: 3D surface plot evaluating relevance ratios for T5-large. Optimal performance was achieved with reward models having moderate accuracy.

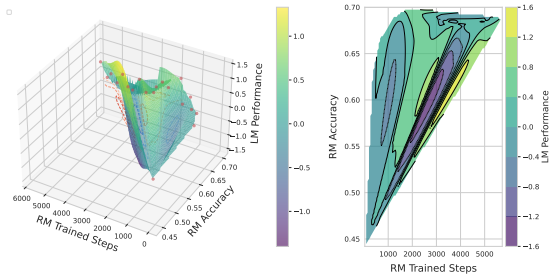

Figure 14: 3D surface plot evaluating completeness rewards for T5-base. Intermediate reward model strength yielded the best language model performance.

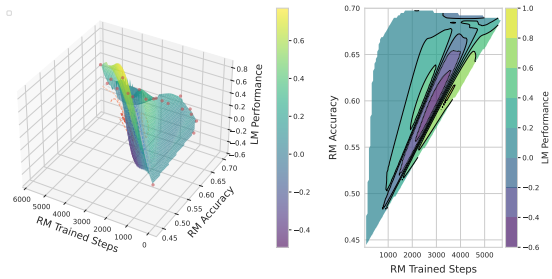

Figure 15: 3D surface plot evaluating completeness rewards for T5-large. Intermediate reward model strength yielded the best language model performance.

B Supplementary Reward Analysis

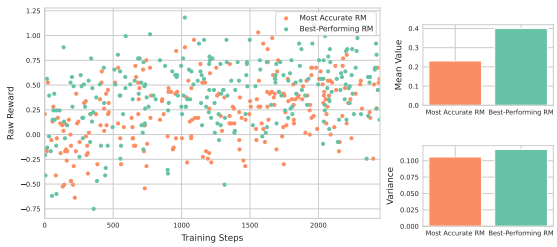

Figure 16: Reward analysis for relevance task (T5-base model): training steps vs. rewards (left), mean and variance of rewards (right).

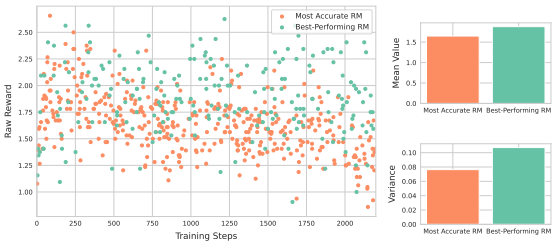

Figure 17: Reward analysis for factuality task (T5-base model): training steps vs. rewards (left), mean and variance of rewards (right).

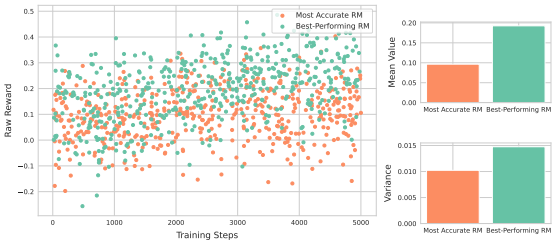

Figure 18: Reward analysis for completeness task (T5-base model): training steps vs. rewards (left), mean and variance of rewards (right).

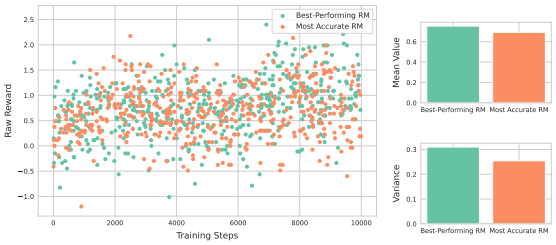

Figure 19: Reward analysis for relevance task (T5-large model): training steps vs. rewards (left), mean and variance of rewards (right).

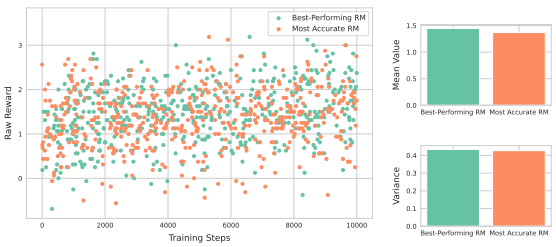

Figure 20: Reward analysis for factuality task (T5-large model): training steps vs. rewards (left), mean and variance of rewards (right).

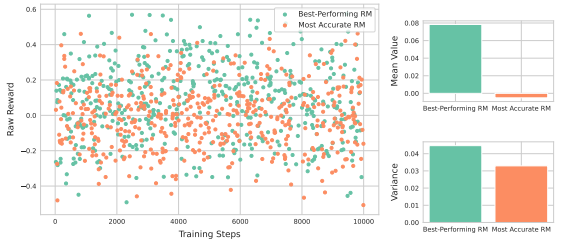

Figure 21: Reward analysis for completeness task (T5-large model): training steps vs. rewards (left), mean and variance of rewards (right).

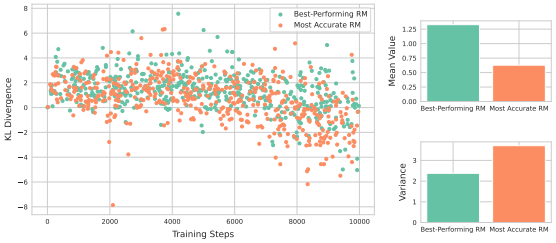

Figure 25: Relevance task KL divergence (T5-large model): training steps vs. KL divergence (left), mean and variance (right).

C Supplementary KL Divergence Analysis

Figure 22: Relevance task KL divergence (T5-base model): training steps vs. KL divergence (left), mean and variance (right).

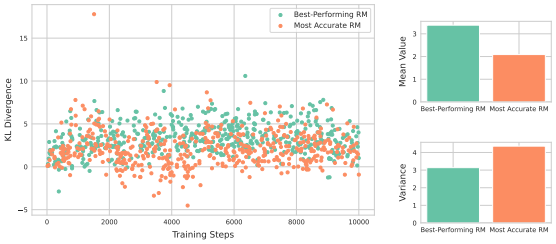

Figure 26: Factuality task KL divergence (T5-large model): training steps vs. KL divergence (left), mean and variance (right).

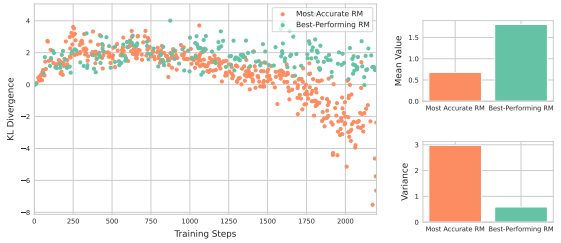

Figure 23: Factuality task KL divergence (T5-base model): training steps vs. KL divergence (left), mean and variance (right).

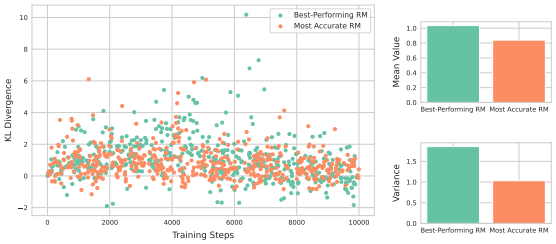

Figure 27: Completeness task KL divergence (T5-large model): training steps vs. KL divergence (left), mean and variance (right).

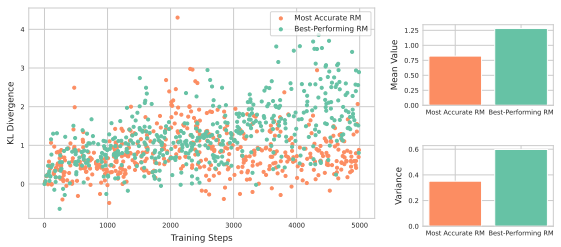

Figure 24: Completeness task KL divergence (T5-base model): training steps vs. KL divergence (left), mean and variance (right).

D Hyperparameter Settings

Model Component	Setting
Input Padding Side	Right
Top-K Sampling	20
Temperature	0.7
Value Model	T5-base
Freeze Value Model	False
Policy-Value Sharing	False

Table 3: Model Configuration Hyperparameters

Reward Model	Setting
Relevance Model (Positive Reward)	0.3
Relevance Model (Negative Reward)	-0.3
Factuality Model (Positive Reward)	0.5
Factuality Model (Negative Reward)	-0.5
Completeness Model (Mean)	-0.4468
Completeness Model (Std)	8.3012
Completeness Model (Bias)	0.0
Completeness Model (Scale)	0.3

Table 4: Reward Model Hyperparameters

Environment Parameter	Setting
Maximum Input Length	1024
Maximum Generated Length	200
Train Samples per Input	4

Table 5: Environment Configuration Hyperparameters

PPO Parameter	Setting
KL Coefficient	0.3
Lambda	0.95
Gamma	1.0
Policy Gradient Coef.	1.0
Value Function Coef.	1.0
Clip Range (Policy)	0.2
Clip Range (Value)	0.2
Whiten Rewards	True

Table 6: PPO Training Hyperparameters

Training Parameter	Setting
Total Episodes	80,000
Learning Rate	0.00001
Warmup Steps	100
PPO Epochs per Rollout	4
KL Threshold	20.0
Clip Gradients	False
Max Gradient Norm	0.5
Random Seed	42

Table 7: Training Procedure Hyperparameters