RAG-Reward: Optimizing RAG with Reward Modeling and RLHF

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) enhances Large Language Models (LLMs) with relevant and up-to-date knowledge, improving their ability to answer knowledge-intensive questions. It has been shown to enhance both generation quality and trustworthiness. While numerous works have focused on improving retrieval, generation, and evaluation, the role of reward models in reinforcement learning for optimizing RAG remains underexplored. In this paper, we introduce **RAG-Reward**, a framework designed to develop reward models to enable *hallucination-free, comprehensive, reliable, and efficient RAG*. We define four key metrics to assess generation quality and develop an automated benchmarking pipeline to evaluate the outputs of multiple LLMs across a variety of RAG scenarios. Using **RAG-Reward**, we train reward models and apply reinforcement learning with human feedback (RLHF) to improve LLMs' effectiveness in RAG. Experimental results demonstrate that our reward model achieves state-of-the-art performance in automatic benchmarking and aligns closely with human evaluations. Furthermore, the improved generation quality of the trained policy model highlights the feasibility and efficiency of using RLHF to enhance RAG outputs [1] .

1 Introduction

Retrieval-augmented generation (RAG) has been widely adopted in research and real-world applications on domain-specific or knowledge-intensive tasks (Gao et al., 2024; Fan et al., 2024; Yu et al., 2023). By leveraging up-to-date external knowledge, Large Language Models (LLMs) can incorporate relevant information during text generation, significantly mitigating hallucination issues (Zhang et al., 2023b; Peng et al., 2023) and ensuring higher answer quality (Lewis et al., 2020).

However, many open-source LLMs are not yet fully optimized for use in RAG scenarios. Simply applying RAG to these models often leads to suboptimal outcomes. For example, even with access to external knowledge, these LLMs may still generate misinformation (Niu et al., 2024), which can be especially problematic given the timeliness of the generated content. While many LLMs excel in conversational tasks, they often struggle in information-intensive scenarios, where both trustworthiness and efficiency are paramount (Wang et al., 2024). Therefore, a solution that can provide effective supervision and evaluation of RAG systems is needed.

Fortunately, significant progress has been made in evaluating the quality of LLM-generated content these days. Unlike many previous benchmark data sets that are created to assess the overall quality of generations (Yang et al., 2018; Kwiatkowski et al., 2019), recent studies suggest that the construction of evaluation data sets tailored to specific domains or scenarios with detailed criteria (Zhu et al., 2024; Friel et al., 2024) can lead to improved in-domain performance. (Friel et al., 2025) shows the feasibility to build such a preference dataset that reflects the RAG quality based on multiple key metrics carefully selected and defined by human experts.

At the same time, reward modeling has recently emerged as a widely adopted strategy to align text generation with human values by learning preference signals from human-annotated high-quality data (Ouyang et al., 2022; Lambert et al., 2024). Multiple well-crafted datasets have demonstrated the value of reward models and high-quality preference datasets. For example, HH-RLHF (Bai et al., 2022a) evaluates the helpfulness and harmlessness of the language, PRM800K (Lightman et al., 2023) assesses the stepwise correctness during mathematical reasoning. And (Song et al., 2024a) develop hallucination detection models using RAG-Truth (Niu et al., 2024) dataset, and achieves hallucina-

[1] The dataset will be made publicly available upon publication.

Figure 1: An overview of our data labeling method and our experiments based on the preference data in RAG Scenario. We use o3-mini as the judge to evaluate the quality of the generation from multiple models. We then train the reward models and use them for Reinforcement Learning.

tion reduction in RAG output.

While human annotation has proven to be effective in evaluating the quality of the generated responses and constructing datasets, it is both expensive and time-consuming. These challenges have led researchers to explore reliable annotation strategies that leverage the capabilities of LLMs. Previous studies have demonstrated the feasibility of using LLMs to assess response quality and build robust reward models (Tan et al., 2024). For instance, Bai et al. (2022b) introduced Constitution AI, which uses LLMs to guide the generation of fine-tuning data and provide preference pairs for reinforcement learning.

Additionally, LLMs can be guided by human-defined criteria and examples to generate preference scores tailored to specific requirements(Sun et al., 2023, 2024). For example, Cui et al. (2024) introduced Ultra-Feedback, a framework that leverages GPT-4 to assign scores based on attributes such as helpfulness, truthfulness, honesty, and instruction-following ability. LLMs have also shown their effectiveness in providing feedback across various reasoning tasks, including coding and solving mathematical problems (Weyssow et al., 2024; Yuan et al., 2024).

Inspired by these works, we developed **RAG-Reward**, a framework designed to create reward models that enable hallucination-free, comprehensive, reliable, and efficient RAG. The framework leverages the capabilities of LLMs to evaluate RAG responses across diverse domains, train reward models, and enhance RAG response quality through RLHF. Figure 1 provides an overview of our framework.

Several related works have explored various aspects of this approach, such as using reward models to measure the relevance of queries and passages (Nguyen et al., 2024), improving trustworthiness of RAG using DPO (Song et al., 2024b), leveraging LLM to build RAG benchmark dataset (Friel et al., 2025), and comparing efficiency of various LLMs serving as RAG reward models (Jin et al., 2024). However, to the best of our knowledge, this is the first attempt to optimize RAG with reward modeling and RLHF pipeline.

Specifically, we select datasets from domains of Question-Answering, Data-to-Text, and Summarization. First, we sample diverse responses from a pool of 12 open-source and proprietary LLMs, including GPT-4 and the Llama-3 series. For each prompt in the datasets, we randomly select 2 LLMs to generate responses. Next, we use o3-mini as the judge to compare them based on four key metrics carefully selected by the human experts: hallucination, comprehensiveness, verbosity, and attribution. This enables us to construct preference pairs, consisting of a chosen response and a rejected one, based on the weighted selection across 4 metrics. Overall, we collect 35K high-quality training samples for the reward model training. Evaluations demonstrate that our reward model achieves over 80% accuracy on the held-out test set. Additionally, we develop a policy model using the RAFT algorithm (Dong et al., 2023), leading to notable performance improvements. Our key contributions are summarized as follows:

- We introduce a reward modeling method for the RAG scenario to assess generation quality. Additionally, we release a high-quality dataset of 35K preference annotations to support future research.
- We define a comprehensive set of metrics that effectively evaluate RAG quality and guide the dataset construction process.
- We conducted extensive experiments to evalu-

ate our reward model, train a policy model, and demonstrate the effectiveness of our approach in enhancing RAG performance.

2 Related Work

2.1 Reward Modeling for Alignment and Reinforcement Learning

Training reward models have become a widely used approach to align language models with human preference (Ouyang et al., 2022). The alignment can enhance various aspects of LLM performance, such as increasing their trustworthiness and helpfulness (Bai et al., 2022a; Wang et al., 2023; Cui et al., 2024), or improving their problem-solving abilities (Dai et al., 2024; Yuan et al., 2024; Zhang et al., 2024a). The reward signal can be trained as a discriminative model to generate a scalar value (Bradley and Terry, 1952), or directly generated as critics from language models (Zhang et al., 2024b; Zheng et al., 2023). Many high-quality datasets for reward modeling have been introduced, such as HH-RLHF (Bai et al., 2022a), Ultra-Feedback (Cui et al., 2024), Code-UltraFeedback (Weyssow et al., 2024), Ultra-Interact (Yuan et al., 2024), and PKU-SafeRLHF (Ji et al., 2024), which could be either labeled by human or by powerful LLMs.

Reinforcement Learning from human feedback (RLHF) is a widely used strategy to enhance policy models after the reward model is developed (Kaufmann et al., 2024). RLHF plays a critical role in aligning LLMs with human values and achieving improved performance (Christiano et al., 2023). Proximal Policy Optimization (PPO) is a commonly used algorithm for alignment tasks to enhance the policy models (Schulman et al., 2017), although it is computationally intensive. Consequently, several alternative, more efficient algorithms have been proposed, such as Direct Preference Optimization (DPO) (Rafailov et al., 2024), Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), Group Relative Policy Optimization (Shao et al., 2024), and Rejection Sampling Fine-tuning (RAFT) (Dong et al., 2023). These approaches have also been widely adopted in state-of-the-art models like Llama-3 (Grattafiori et al., 2024), and Qwen-2 (Yang et al., 2024a).

Concurrently, the work introduced in Jin et al. (2024) utilizes existing reward models to evaluate Question-Answering tasks in RAG scenarios with fine-grained metrics, highlighting the limitation of the general reward models. It also shows the feasibility of constructing RAG scenario data using Large Language Models. Our work is built upon these existing works to train the RAG-specific reward model and use it for alignment training.

2.2 Large Language Models and Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) has proven to be an effective method for enhancing language models with real-world knowledge to address a wide range of tasks, thereby improving the accuracy and credibility of the generated output (Lewis et al., 2021). In the era of LLMs, which possess a strong ability to understand and utilize in-context information, RAG can significantly enhance their capabilities (Fan et al., 2024; Gao et al., 2024). RAG addresses common challenges of LLMs, such as hallucinations and outdated knowledge, by grounding their outputs in external knowledge bases (Peng et al., 2023; Li et al., 2024). RAG-based LLMs can be trained to effectively adapt and integrate retrieved information (Schick et al., 2023; Shao et al., 2023), or use training-free methods that directly insert the retrieved context into the prompt (Ram et al., 2023). These LLMs have been widely adopted in real-world applications. For instance, retrievers are integrated into LLM-based chatbots to increase the helpfulness and trustworthiness of the conversations (Komeili et al., 2021). RAG-based models have also been deployed as domain-specific experts, such as finance (Zhang et al., 2023a) and medicine (Xiong et al., 2024).

In this project, we are the first to systematically construct RAG-scenario preference datasets and develop reward models, paving the way for evaluating and enhancing the generation quality of LLMs within the RAG framework.

3 Dataset Construction

We construct our dataset based on existing RAG datasets to ensure its relevance and applicability. To reflect the diverse use cases of RAG scenarios, we include three common types: Question Answering, Data-to-Text and Summarization. Specifically, we use WebGLM (Liu et al., 2023), Yelp (Yelp, 2021), and XSum (Narayan et al., 2018) as experimental datasets, each dataset corresponding to one of the three RAG scenarios.

For the WebGLM dataset, LLMs are tasked with reasoning over web-retrieved reference data to answer real-world questions, generating concise re-

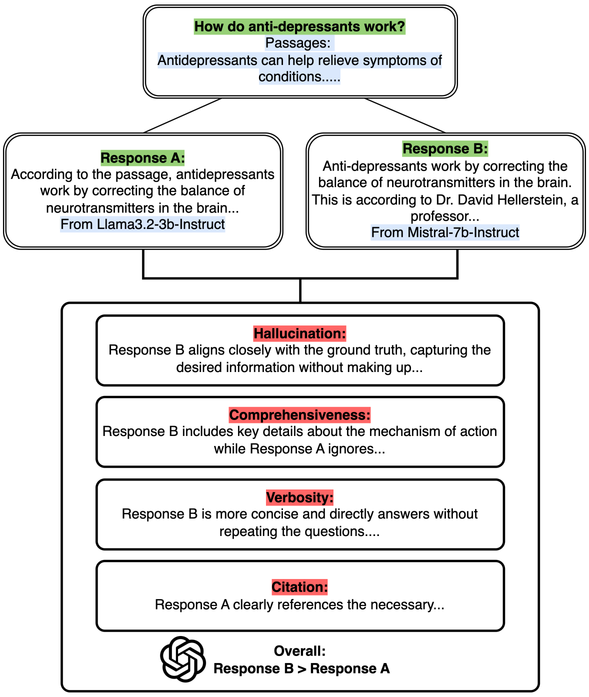

Figure 2: An illustration of our data annotation method. Given a sample and two responses, we prompt o3-mini to provide a judgment based on each metric separately. We then aggregate the results and construct the pairs.

Dataset	Number of Samples		
	Training	Dev	Testing
WebGLM	11000	1000	500
Yelp	12000	1000	500
XSum	12000	1000	500

Table 1: The number of preference pairs that we construct from the 3 datasets in our experiments.

sponses in a few sentences. For the Yelp dataset, our experiments focus on data from the restaurant category, represented in JSON format. Each sample includes information such as a restaurant's location and ambiance. Based on the structured JSON input, LLMs generate descriptive text about the restaurant. The XSum dataset contains diverse articles from the British Broadcasting Corporation (BBC), with models tasked with summarizing these articles. These three datasets cover a broad range of circumstances, ensuring that the reward model trained on them can significantly improve the development and evaluation of RAG systems. Table 1 presents the number of data samples used in our experiments. And examples of these data sets are presented in Table 2.

When evaluating the quality of the responses, we consider the following metrics:

Hallucination: The models should generate responses strictly based on the context provided, without introducing information not grounded in the retrieval context. If the retrieval context contradicts the model's parametric knowledge, the model should adhere to the retrieval reference, ensuring that the response is accurate and contextually relevant.

Comprehensiveness: The response should fully utilize the context provided by the retrieved content and address all aspects of the instruction. This requires the model to extract and integrate all relevant information from the retrieval context to ensure the response is thorough and complete.

Verbosity: While the response should be detailed and comprehensive, it should also be concise, relevant, and straight to the point. Striking the right balance between detail and brevity is essential to providing informative answers without overwhelming the user.

Attribution: This metric is specifically applied to the WebGLM-QA dataset to ensure the generations are trustworthy and verifiable. The response should explicitly refer to the context retrieved to improve credibility and allow users to trace information sources.

3.1 Dataset Sampling

We utilize a combination of open source instruction models, the GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI et al., 2024) series to generate data, ensuring diversity and inclusion of both high-quality and relatively low-quality responses. The open-source models consist of various sizes of the instruction-tuned versions of Llama-3, Llama-3.1, Llama-3.2 (Grattafiori et al., 2024), Llama-2 (Touvron et al., 2023), Qwen-2 (Yang et al., 2024a), InternLM-2 (Cai et al., 2024), and Mistral (Jiang et al., 2023).

In total, we include 12 candidate models for generation. For each question and its corresponding reference in the dataset, we randomly select two models' generations to form preference pair.

3.2 Dataset Labeling

We use o3-mini (OpenAI, 2025) to label the data. An illustration of our labeling methods is shown in Figure 2. Given a question and a pair of responses from different models, we prompt o3-mini to compare and select the preferred response.

Dataset	Data Example
WebGLM (Liu et al., 2023)	***Question:*** Why are different tiers (regular < mid < premium) of gas' prices almost always 10 cents different? ***References:*** [The gap between premium and regular gas has..., According to national averages, the price...] ***Answer:*** The 10 cent difference between the different tiers of gas prices is likely due to a convention...
	Our Prompt: Answer the following question: *{question}* Your response should be based on the following passages: *{passages}* When you respond, you should refer to the source of information...
Yelp (Yelp, 2021)	[***Name:*** The Green Pheasant ***Address:*** 215 1st Ave S ***City:*** Nashville ***State:*** TN ***Attributes:***{ HappyHour: True, DogsAllowed: False, ... }]
	Our Prompt: Write an overview about the following business based only on the provided structured data in the JSON format...
XSum (Narayan et al., 2018)	***Document:*** The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed. Repair work is ongoing in Hawick and many roads... ***Summary:*** Clean-up operations are continuing across the Scottish Borders and Dumfries and Galloway after...
	Our Prompt: Summarize the following document: *{document}*...

Table 2: Illustration and statistics of the original datasets and the prompts used to construct the preference data. For WebGLM, LLMs will generate responses based on the reference. For Yelp, LLMs will convert the JSON data into a descriptive overview. For XSum, LLMs will summarize the given document.

Specifically, we ask o3-mini to compare the responses based on the four metrics outlined earlier, assessing them individually. In the prompt, we explicitly ask o3-mini to put heavier weights on hallucination and comprehensiveness metrics, as they are crucial to the answer quality, while the other two mainly improves the readability. After the o3-mini has made the individual judgments on the 4 metrics, it will generate an overall preference for the pair data based on the judgments above. We may end up getting a preferred answer with no hallucination but a bit verbose. And this approach acknowledges real-world scenarios where responses are rarely perfect, and trade-offs are often necessary.

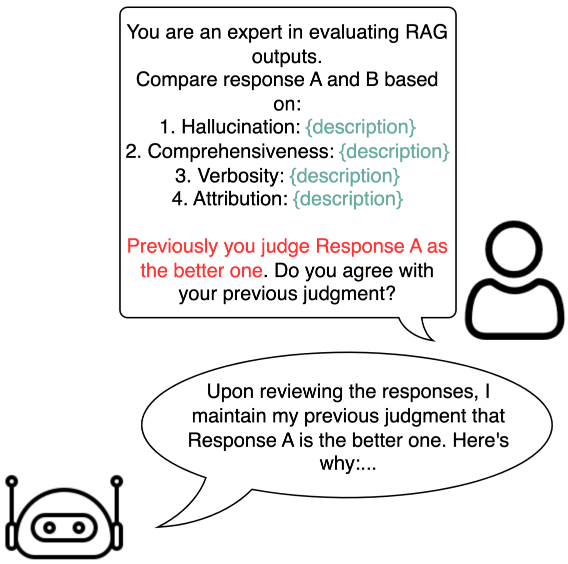

Figure 3: An illustration of our self-evaluation method. For a sample of the constructed data pairs, we provide o3-mini with the previous judgment, and ask it to re-evaluate the comparison result.

4 Dataset Evaluation

4.1 Self Evaluation

In this subsection, we evaluate the consistency of the evaluations provided by o3-mini, a key metric

	WebGLM	Yelp	XSum	Avg
Consist.	97.9	98.8	95.2	97.3

Table 3: The consistency rate for the self-evaluation of the constructed dataset. We measure whether the o3-mini agrees with the comparison results previously made.

commonly used to demonstrate the reliability of the annotations. To objectively assess the quality of our constructed dataset, we design a self-evaluation method that measures the consistency of o3-mini's responses. An illustration of this method is shown in Figure 3. Specifically, for each prompt, a chosen response, and a rejected response in the dataset, we prompt o3-mini to revisit its previous comparison result and verify whether it maintains its original judgment. We define the consistency rate as the proportion of samples where the evaluation results remain unchanged across both assessments. In this experiment, we randomly select 1,000 samples for re-evaluation.

The results of this experiment are presented in Table 3. We observe a very high consistency rate across the three tasks and the overall consistency rate exceeds 97%. The results demonstrate that o3-mini produces stable and consistent labels for most of the data according to well-defined criteria, and reflects the quality of the dataset.

	WebGLM	Yelp	XSum	Avg
Consist.	0.79	0.80	0.83	0.81

Table 4: The consistency rate between human evaluation and the o3-mini labeled dataset.

4.2 Human Evaluation

We also performed human evaluations to assess the alignment of AI annotations with human preferences. Specifically, we randomly select 100 samples with paired responses from each dataset and ask the annotators to evaluate using the same pipeline described in Section 3.2 and illustrated in Figure 2. Annotators compare the responses based on each metric and determine the preference pair. We calculate the agreement ratio between the human annotators and o3-mini in the preference pairs, and the results are shown in Table 4. We observe an overall agreement rate of 81%, with consistent agreement across the three tasks. This metric is comparable to the figures reported in (Jin et al., 2024), where RAG is evaluated based on helpfulness and harmlessness criteria. These results highlight proprietary LLMs' ability to effectively capture human preferences in assessing RAG response quality.

5 Limitation of Existing Reward Models

In this section, we evaluate several existing reward models in our test set. We select models from RewardBench (Lambert et al., 2024) known for their strong performance in assessing aspects such as helpfulness, safety, and reasoning. We examine their performance on diverse RAG scenarios using our curated test set (Table 1), and demonstrate their limitation of the evaluation on RAG domains.

The experiment results are shown in Table 5. While many of the listed reward models achieve accuracy higher than 90% in evaluating chat, safety, and reasoning tasks, their overall accuracy in RAG scenarios is below 80%. This underscores a significant gap between mainstream reward models and the unique requirements of RAG tasks. We also observe inconsistent performance across tasks for the reward models, with better results on Data-to-Text and Summarization tasks (e.g. the Yelp dataset and the XSum dataset) compared to the QA task (WebGLM dataset), suggesting that current reward models are not uniformly capable across different RAG scenarios. Interestingly, several reward models that achieve state-of-the-art (SOTA) performance in reasoning and safety evaluations, as shown on the leaderboard[2], do not perform well on RAG tasks. For example, models like URM-LLaMa-3.1-8B (Lou et al., 2024) and Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024) underperform in the RAG domain.

In contrast, UltraRM-13b (Cui et al., 2024), which performs suboptimally in reasoning and safety evaluations but excels in assessing helpfulness and instruction-following, achieves the top accuracy on RAG tasks. This suggests that reward models trained primarily on reasoning tasks may not generalize effectively to evaluating RAG-specific generations. Most of the existing reward models could not excel in expressing the preference in RAG scenarios. Domain-specific training data are therefore essential to address this gap and improve RAG performance evaluation.

[2] https://huggingface.co/spaces/allenai/reward-bench

Models	WebGLM	Yelp	XSum	Average
UltraRM-13b (Cui et al., 2024)	71.0	77.4	79.2	75.9
llama-3-tulu-2-8b-uf-mean-rm (Ivison et al., 2024)	71.0	76.2	78.8	75.3
internlm2-7b-reward (Cai et al., 2024)	72.0	73.0	80.4	75.1
Eurus-RM-7b (Yuan et al., 2024)	71.8	74.4	77.8	74.7
FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2024)	71.6	72.4	77.8	73.9
Llama-3-OffsetBias-RM-8B (Park et al., 2024)	70.0	70.4	76.6	72.3
URM-LLaMa-3.1-8B (Lou et al., 2024)	66.8	70.2	76.0	71.0
QRM-Llama3.1-8B-v2 (Dorka, 2024)	68.4	68.0	74.6	70.3
GRM-Llama3.2-3B-rewardmodel-ft (Yang et al., 2024b)	64.6	71.6	73.4	69.9
Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024)	64.8	68.6	72.8	68.7

Table 5: The evaluation results of the existing reward models on the 3 tasks. They achieve SOTA performance on chatting, safety, and reasoning evaluation, but do not excel in RAG tasks.

6 Experiments

We conduct both reward model training and reinforcement learning using our **RAG-Reward** dataset. In total, 35K preference pairs are used for reward modeling (see Table 1). Additionally, we create a 3K-sample development set for sampling and learning during RLHF training. To evaluate the performance of the policy and reward models, a held-out test set of 1.5K samples is used.

	WebGLM	Yelp	XSum	Average
Acc.	81.4	87.6	84.0	84.3

Table 6: The evaluation results of the reward model on the 3 tasks. The accuracy is calculated as the proportion of test samples where the reward model assigns a higher score to the chosen response than to the rejected response.

6.1 Reward Modeling

We adopt the common approach to train the Bradley-Terry reward model (Bradley and Terry, 1952; Ouyang et al., 2022) to learn the reward signal from the preference data. Specifically, we use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the base model for training. We train the reward model with a learning rate of $2e^{-6}$, a global batch size of 64, a max length of 4096, and an epoch of 1 on 4 H100-80G GPUs.

During the test stage, each test sample contains a chosen response and a rejected response. The accuracy is calculated as the proportion of test samples in which the reward model assigns a higher score to the chosen response than to the rejected one. The detailed results are shown in Table 6. We observe a high accuracy of 84.3% for the reward model, demonstrating its effectiveness in aligning with the intended criteria. Compared to the strong reward models in Table 5, our model achieves the highest accuracy, highlighting the advantage of leveraging RAG-specific reward data.

Furthermore, we observe a consistent accuracy across the 3 tasks, indicating that the reward model could jointly learn the preference signal from diverse tasks and domains. Notably, the reward model achieves the highest accuracy on the Data-to-Text task, while its performance is relatively lower on the Question-Answering task. This difference suggests that comparing structured data with text data is easier for the reward model, while evaluating the quality of a long-form QA poses a greater challenge. This observation aligns with our intuition and expectations.

	Mistral-7B-v0.1	Llama-3.2-3B
WebGLM	64.2	62.6
Yelp	77.8	75.0
Xsum	65.6	64.0
Average	69.2	67.2

Table 7: The RLHF results on the 3 tasks. The win rate is calculated as the proportion of test samples where the **reward model** assigns a higher score to the response generated by the post-trained policy model.

6.2 Preference Alignment

We adopt the RAFT algorithm (Dong et al., 2023) to perform the preference alignment. RAFT utilizes the reward model to select the response with the highest reward score from N candidate responses, and then fine-tunes the policy model on this selected set of responses.

	Mistral-7B-v0.1	Llama-3.2-3B
WebGLM	61.2	60.0
Yelp	80.2	73.6
XSum	63.8	63.2
Average	68.4	65.6

Table 8: The RLHF results on the 3 tasks. The win rate is calculated as the proportion of test samples where the **o3-mini** prefers the response generated by the post-trained policy model.

We set $N = 16$ in our experiments. We use Llama-3.2-3B-Instruct (Grattafiori et al., 2024) and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) as initial policy models for sampling, followed by RAFT training. Both Mistral and Llama models are fine-tuned with a learning rate of $5e^{-6}$, an epoch of 1, a packing length of 4096, and a global batch size of 16, using the axolotl package[3].

To measure the improvement brought by alignment training, we first sample responses from our held-out test set using both the initial models and the post-trained policy models. Thus, for each prompt, we have paired responses from the two models. These pairs are then evaluated using the reward model, and we calculate the proportion of cases where the responses from the post-trained models are preferred. Furthermore, we ask o3-mini to compare the pairs based on the criteria introduced in Section 3. For both metrics, a baseline of 50% indicates no improvement in the policy models. The calculated scores are referred to as the win rate against the initial model.

The experiment results are shown in Table 7 and Table 8. We observe a clear improvement in the policy models after a single iteration of RAFT. Both the reward model and o3-mini agree that generations align more closely with the RAG metrics, as the average win rate is significantly above 50%. These results highlight the effectiveness of our dataset and the reward model. The ratings across the 3 tasks from the reward model is very similar to the ratings from o3-mini, showing that our reward model learns the rationale of rating from o3-mini.

However, we also observe some imbalance in learning for the policy model across tasks. As shown in the tables, there are differences in the win rate across 3 tasks differences. Specifically, the win rate for Yelp could reach 80% while the other 2 are only above 60%, even though they are trained on the same number of samples for each task. The comparison reveals that the difficulties are different across RAG scenarios.

[3] https://github.com/axolotl-ai-cloud/axolotl

	WebGLM	Yelp	XSum	Average
Human	62.0	70.0	66.0	66.0

Table 9: The human evaluation results of the policy model (Mistral) after RAFT on 3 tasks. The agreement is calculated as the proportion of test samples where the generation after RAFT is preferred by humans.

6.3 Human Evaluation

To further validate the improvement of the policy model, we leverage human evaluation of the generations from it. Specifically, we select the model which trained from Mistral-7B-Instruct with one iteration of RAFT, and the Mistral itself as the reference. We adopt the same evaluation strategy introduced in Section 6.2 and replace the reward model and o3-mini with human annotators. Due to the expense, we select 50 samples of each dataset for human labeling. From Table 9, we observe an agreement far above 50%, indicating the effiveness of the RAFT training to improve the policy model on RAG domains. Compared with the results from the reward model and o3-mini, we discover the same trend across 3 tasks, which further shows the alignment of our dataset with humans.

7 Conclusion

In this paper, we introduce **RAG-Reward**, a high-quality preference dataset designed for Retrieval-Augmented Generation (RAG). Our dataset is generated through a novel automated AI annotation pipeline, leveraging both open-source and proprietary models to enhance generalization and versatility. To ensure fair and reliable evaluations, we use o3-mini to assess generation quality based on four key metrics carefully selected by human experts. The dataset spans multiple domains, including *Question Answering, Data-to-Text, and Summarization*, resulting in a large-scale and diverse benchmark. The experimental results show strong alignment with human evaluations, demonstrating the effectiveness of RAG-Reward in reward modeling and reinforcement learning. These findings highlight the potential of our dataset to advance both the evaluation and generation of RAG systems. To foster further research, we will publicly release the dataset to the community.

8 Limitations

In this paper, we constructed a large-scale, high-quality dataset tailored for RAG scenarios and demonstrated the effectiveness of our proposed pipeline through RLHF experiments. However, due to computational constraints, we did not conduct large-scale RLHF training or implement more complex algorithms such as PPO. Future work could explore training larger reward models and incorporating iterative-DPO or PPO to further enhance performance in RAG domains.

References

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 Olsson, 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, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

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.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Ning Dai, Zheng Wu, Renjie Zheng, Ziyun Wei, Wenlei Shi, Xing Jin, Guanlin Liu, Chen Dun, Liang Huang, and Lin Yan. 2024. Process supervision-guided policy optimization for code generation. *Preprint*, arXiv:2410.17621.

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.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *Preprint*, arXiv:2405.07863.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat

Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Preprint*, arXiv:2406.09279.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *Preprint*, arXiv:2406.15513.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rag-rewardbench: Benchmarking reward models in retrieval augmented generation for preference alignment. *Preprint*, arXiv:2412.13746.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *Preprint*, arXiv:2107.07566.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. *Preprint*, arXiv:2403.13787.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021.

Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. *Preprint*, arXiv:2305.14623.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. *Preprint*, arXiv:2306.07906.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *Preprint*, arXiv:1808.08745.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2024. Reward-rag: Enhancing rag with reward driven supervision. *Preprint*, arXiv:2410.03780.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

OpenAI. 2025. Openai o3-mini: Pushing the frontier of cost-effective reasoning. OpenAI Release.

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, Mohammad 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 Eleti, 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, Jan 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, Emy 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 H. 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 B. 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 Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *Preprint*, arXiv:2407.06551.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *Preprint*, arXiv:2302.12813.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Preprint*, arXiv:2302.00083.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

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

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *Preprint*, arXiv:2305.15294.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024a. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2024b. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. *Preprint*, arXiv:2409.11242.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Salmon: Self-alignment with instructable reward models. *Preprint*, arXiv:2310.05910.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Preprint*, arXiv:2305.03047.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *Preprint*, arXiv:2402.13446.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. *Preprint*, arXiv:2402.02420.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2023. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *Preprint*, arXiv:2311.09528.

Martin Weyssow, Aton Kamanda, and Houari Sahraoui. 2024. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences. *Preprint*, arXiv:2403.09032.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *Preprint*, arXiv:2408.00727.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024b. Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Yelp. 2021. Yelp open dataset.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. *Preprint*, arXiv:2209.10063.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024. Advancing llm reasoning generalists with preference trees. *Preprint*, arXiv:2404.02078.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. *Preprint*, arXiv:2310.04027.

Hanning Zhang, Pengcheng Wang, Shizhe Diao, Yong Lin, Rui Pan, Hanze Dong, Dylan Zhang, Pavlo Molchanov, and Tong Zhang. 2024a. Entropy-regularized process reward model. *Preprint*, arXiv:2412.11006.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024b. Generative verifiers: Reward modeling as next-token prediction. *Preprint*, arXiv:2408.15240.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Rageval: Scenario specific rag evaluation dataset generation framework. *Preprint*, arXiv:2408.01262.