OpenGenAlign: A Preference Dataset and Benchmark for Trustworthy Reward Modeling in Open-Ended, Long-Context Generation

Anonymous ACL submission

Abstract

Reward Modeling is critical in evaluating and improving the generation of Large Language Models (LLMs). While numerous recent works have shown its feasibility in improving safety, helpfulness, reasoning, and instruction-following ability, its capability and generalization to open-ended long-context generation is still rarely explored. In this paper, we introduce **OpenGenAlign**, a framework and a high-quality dataset designed to develop reward models to evaluate and improve *hallucination-free, comprehensive, reliable, and efficient open-ended long-context generation*. We define four key metrics to assess generation quality and develop an automated pipeline to evaluate the outputs of multiple LLMs across long-context QA, Data-to-Text, and Summarization scenarios using o3, ending up with 33K high-quality preference data with a human agreement rate of 81%. Experimental results first demonstrate that existing reward models perform suboptimally on the held-out benchmark. And Our trained reward model achieves superior performance in the benchmark and effectively improves the generation quality of the policy models using Reinforcement Learning (RL). Additionally, **OpenGenAlign** could be used for effective guided generation in existing datasets. Furthermore, we demonstrate that the **OpenGenAlign** could be integrated with reward data from other domains to achieve better performance.

1 Introduction

Reward models have emerged as a critical component in aligning Large Language Models (LLMs) with human preferences, serving as the backbone of Reinforcement Learning from Human Feedback (RLHF) pipelines (Bai et al., 2022; Lambert et al., 2024; Christiano et al., 2017; Dong et al., 2024; Cui et al., 2025). RLHF with explicit reward modeling has established itself as a standard paradigm for enhancing LLM capabilities and is now widely adopted across state-of-the-art models, including both proprietary systems such as GPT (OpenAI et al., 2024) and Claude (Anthropic, 2025), as well as open-source models such as Llama (Grattafiori et al., 2024), Qwen (Yang et al., 2025), and Gemma (Team et al., 2025). In RLHF pipelines, reward models guide training by evaluating outputs and providing feedback signals for reinforcement learning optimization. This approach has proven effective across diverse objectives, from improving response helpfulness and safety (Bai et al., 2022; Wang et al., 2024b; Liu et al., 2024a) to enhancing reasoning capabilities (Shao et al., 2024; Cui et al., 2025; Yuan et al., 2024).

High-quality training data is fundamental to building effective reward models. Numerous datasets have been developed for this purpose, including HH-RLHF (Bai et al., 2022), UltraFeedback (Cui et al., 2024), Skywork-Reward (Liu et al., 2024a), and HelpSteer-3 (Wang et al., 2025). These datasets construct preference pairs by assessing response quality through either human annotation (Stiennon et al., 2020; Köpf et al., 2023) or frontier LLM evaluation (Cui et al., 2024; Xu et al., 2024), both of which have demonstrated effectiveness. Correspondingly, various benchmarks have been developed to evaluate reward model performance, including RewardBench (Lambert et al., 2024; Malik et al., 2025) and RM-Bench (Liu et al., 2024b). However, existing datasets and benchmarks leave a critical gap: there is neither comprehensive training data nor dedicated evaluation benchmarks for reward models in various open-ended, long-context generation settings.

While reward models have become standard, recent work has explored alternatives. Rule-based methods (DeepSeek-AI et al., 2025; He et al., 2025; Wen et al., 2025) and Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2025) bypass reward models by leveraging answer verification in domains like mathematical reasoning and

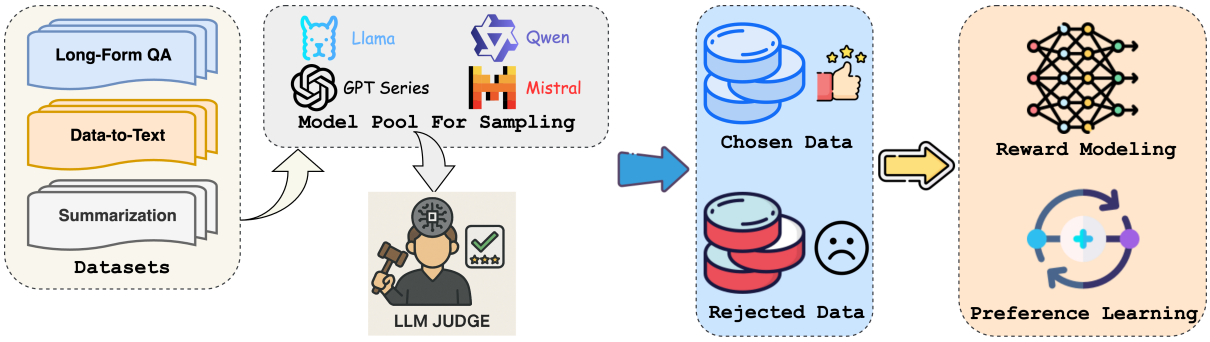

Figure 1: Overview of our data labeling method and our experiments based on it in the open-ended long-context Scenario. We use o3 as the judge to evaluate the quality of the generation from multiple models. We then train the reward models and use them for Reinforcement Learning.

code generation (Zeng et al., 2025; Zhang et al., 2025). However, such approaches cannot extend to long-form generation, where the absence of ground truth and the subjective nature of quality assessment necessitate learned reward models to encode human preferences.

To address this gap, we introduce **OpenGenAlign**, a dataset for reward modeling in open-ended, long-context generation. **OpenGenAlign** comprises three components: a training set containing 33K preference pairs for reward modeling, a development set of 9K samples for policy optimization, and a held-out benchmark consisting of 1.5K samples for evaluating both existing reward models and our proposed model (Table 9). Specifically, we curate datasets spanning three domains: Question Answering, Data-to-Text, and Summarization. Our data construction procedure operates as follows. We have a pool of 12 open-source and proprietary LLMs. For each prompt, we randomly select two models from this pool to generate responses. Subsequently, we employ OpenAI's o3 (OpenAI, 2025) model as the judge, augmented with majority voting across multiple comparisons, to assess response pairs according to four critical dimensions: **Hallucination**, **Comprehensiveness**, **Verbosity**, and **Attribution**. This evaluation framework enables us to construct preference pairs—each comprising a chosen response and a rejected response. Figure 1 provides an overview of our complete pipeline.

Empirical results validate the effectiveness of our approach: reward models trained on **OpenGenAlign** achieve about 86% accuracy on the held-out benchmark. Furthermore, we demonstrate that our reward model successfully guides policy optimization using the PPO algorithm (Schulman et al., 2017), showing substantial improvements in long-context generation quality. Our key contributions are summarized as follows:

- We introduce a high-quality reward modeling dataset and benchmark for open-ended, long-context generation and release it to facilitate future research in this domain.
- We conduct comprehensive experiments demonstrating our reward model's effectiveness, including reward model training, policy training, out-of-distribution generalization, and integration with existing datasets.

2 Related Work

2.1 Reward Modeling 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 both enhance their trustworthiness and helpfulness (Bai et al., 2022; Wang et al., 2023; Cui et al., 2024), and improve their problem-solving abilities (Dai et al., 2024; Yuan et al., 2024; Zhang et al., 2024a; Cui et al., 2025). Many high-quality datasets for reward modeling have been introduced, such as HH-RLHF (Bai et al., 2022), Ultra-Feedback (Cui et al., 2024), Code-UltraFeedback (Weyssow et al., 2024), Ultra-Interact (Yuan et al., 2024), HelpSteer (Wang et al., 2023), PKU-SafeRLHF (Ji et al., 2024), Sky-Reward-Preference-80K (Liu et al., 2024a), and HelpSteer-3 (Wang et al., 2025). The reward signal can be trained as a discriminative model to generate one or a few scalar values (Bradley and Terry, 1952; Liu et al., 2024a; Wang et al., 2024a), or directly generated as next token prediction from language models (Zhang et al., 2024b; Zheng et al., 2023; Xiong et al., 2024). Inspired by Deepseek-

Table 1: Overview of several representative **Open-Sourced** preference datasets in chronological order used in reward model training. We compare **OpenGenAlign** with several popular datasets.

Dataset	Size	Domain
OpenAI Summarization (Stiennon et al., 2020)	93K	Summarization
WebGPT-Comparisons (Nakano et al., 2021)	20K	General Web-based QA
HH-RLHF (Bai et al., 2022)	161K	Chat (Helpfulness, Harmlessness)
UltraFeedback (Cui et al., 2024)	64K	Chat (Helpfulness, Honesty, Truthfulness, Instruction-Following)
WildGuard (Han et al., 2024)	87K	Chat (Safety, Jailbreaks, Refusals)
UltraInteract (Yuan et al., 2024)	220K	Reasoning (Math, Code)
HelpSteer2 (Wang et al., 2024b)	10K	Chat (Helpfulness, Correctness, Coherence, Verbosity, Complexity)
Skywork-Reward-80K-v0.2 (Liu et al., 2024a)	80K	Chat, Format Bias, Safety
HelpSteer3 (Wang et al., 2025)	40K	General Chat, Multi-lingual, Coding
OpenGenAlign	33K	Open-ended, Long-form Generation

Table 2: The number of preference pairs constructed from the three datasets used in our experiments.

Dataset	Training	Dev	Testing
WebGLM (Liu et al., 2023a)	11,000	3,000	500
Yelp (Yelp, 2021)	11,000	3,000	500
XSum (Narayan et al., 2018)	11,000	3,000	500

R1 (DeepSeek-AI et al., 2025), there has been R1-like reasoning reward models which integrates long CoT and RL to further enhance the performance on various tasks (Chen et al., 2025; Liu et al., 2025; Xiong et al., 2025).

Reinforcement Learning from human feedback (RLHF) is a widely used strategy to enhance policy models after the reward model is developed (Bai et al., 2022; 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). Alternative variants include Rejection Sampling Fine-tuning (RAFT) (Dong et al., 2023), Direct Preference Optimization (DPO) (Rafailov et al., 2024), Group Relative Policy Optimization (GRPO) (Shao et al., 2024), Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025), and TreePO (Li et al., 2025). Rule-based reward without reward models has also been proven an effective way in reinforcement learning on verifiable domains (Zeng et al., 2025; Zhang et al., 2025; Wen et al., 2025).

2.2 Open-Ended Generation and Long Context Understanding for LLMs

There has been much research towards the evaluation of the long context understanding, such as Needle-in-a-Haystack (Kamradt, 2023), Ruler (Hsieh et al., 2024), NeedleBench (Li et al., 2024), where important pieces of information are placed inside long context and the LLMs are required to locate them. However, these benchmarks fail to evaluate the overall generation quality. Other works, such as ZeroSCROLLS (Shaham et al., 2023), LongBench (Bai et al., 2024), and LongBench-v2 (Bai et al., 2025), focus more on real-world scenarios like long Question-Answering and Summarization tasks. But they either limit the format in multiple-choice questions, or adopt metrics like F1 and ROUGE (Lin, 2004), which might not be completely accurate for long-form generation and might deviate from human judgment (Tan et al., 2024). Early works (Liu et al., 2023b; Chiang and yi Lee, 2023) also attempt to incorporate LLMs as evaluators via prompting. However, there are no available domain-specific datasets for effective training of such evaluators.

3 Dataset Construction

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

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

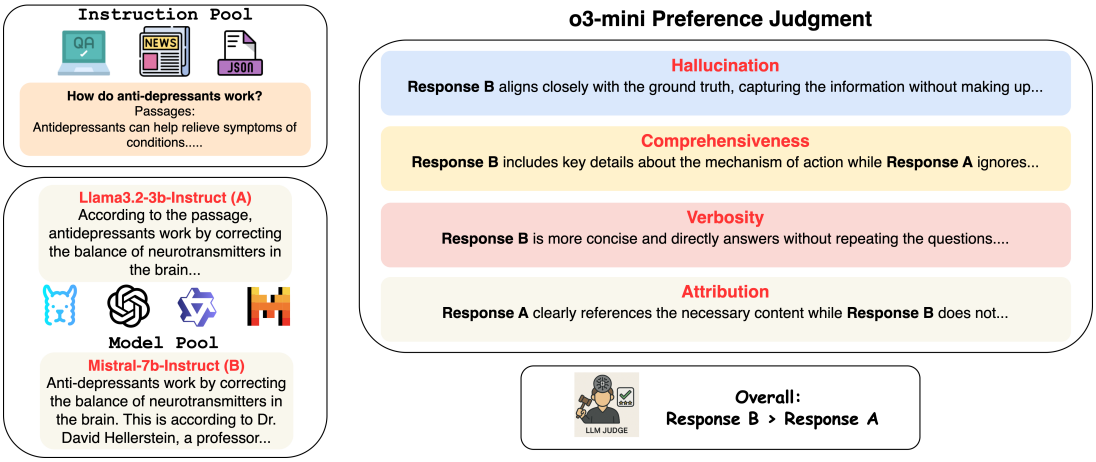

Figure 2: Illustration of our data annotation method. Given a sample and two responses, we prompt o3 to provide judgments based on each metric and aggregate the results to construct pairs.

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 open-ended long-context understanding. Table 2 presents the number of data samples used in our experiments. And examples of these data sets are presented in Table 9 Appendix.

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 context. If the context contradicts the model's parametric knowledge, the model should adhere to the reference, ensuring that the response is accurate and contextually relevant (Niu et al., 2024).

Comprehensiveness: The response should fully utilize the context provided by the retrieved content and address all aspects of the instruction (Zhu et al., 2024). This requires the model to extract and integrate all relevant information from the retrieval context to ensure the response is thorough and complete.

Verbosity: Zhu et al. (2024) also adopts **Irrelevance** as a metric and we modify a bit. 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. This is especially important for the summarization task.

Attribution: This metric is specifically applied to the WebGLM-QA dataset and this is important to ensure the generations are trustworthy and verifiable (Huang and Chang, 2024). The response should explicitly cite or point to the relevant part of the context when generating factual content.

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 have first compared the labeling quality of the reasoning model (o3) and non-reasoning model (GPT-4o), and identified the superior performance of the reasoning models. So we end up using o3 (OpenAI, 2025) to label the data. An illustration of our labeling methods is shown in Figure 2. Given a question x and a pair of responses (y_1, y_2) from

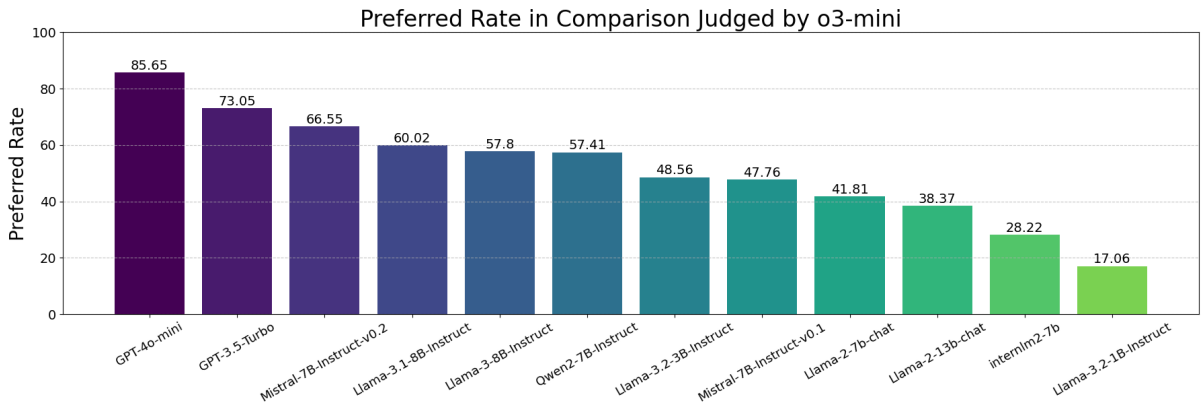

Figure 3: Statistics of the preferred rate for each model during preference pairs construction phase.

Table 3: Human agreement rate of the o3-labeled dataset (measured as proportion).

	WebGLM	**Yelp**	**XSum**	**Avg**
Agreement rate (%)	0.77	0.83	0.82	0.81

different models, we prompt o3 to compare and select the preferred response. Specifically, we ask o3 to compare the responses based on the four metrics outlined earlier, assessing them individually. In the prompt, we explicitly ask o3 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 has made the individual judgments on the 4 metrics, it will generate an overall preference for the pair data based on the judgments above. To increase reliability, we collect three independent judgments per question pair and use **majority voting** to determine the final preference. And we end up getting the preference dataset of triplets (x, y_w, y_l). In our experiments, we observe no significant performance difference between evaluating attributes individually versus holistically in a single prompt; therefore, we adopt the latter approach to reduce computational costs. Our prompts for each dataset are shown in Figure 4.

3.3 Human Evaluation

We also performed human evaluations to assess the alignment of AI annotations with human preferences. To ensure the expertise and reliability, we hired **U.S. citizens with a major in Journalism** in college to compare and judge the quality of the dataset. Specifically, we randomly select 100 samples with paired responses from each dataset and ask the annotators to evaluate using the same 4 metrics illustrated in Figure 2. To ensure the consistency and quality of human annotation results, we ask 3 annotators to compare the pair and use majority voting results as the final preference. The agreement ratio between the human annotators and o3 is shown in Table 3. We observe an overall agreement rate of 81%, with consistent agreement across the three tasks. These results highlight proprietary LLMs' ability to effectively capture human preferences of subjective questions in assessing the response quality.

4 Benchmark of Existing Reward Models

In this section, we benchmark several existing reward models in our test set. We select top models from RewardBench (Lambert et al., 2024) known for their strong performance in assessing aspects such as helpfulness and instruction-following. We examine their performance on diverse **OpenGenAlign** scenarios using our curated benchmark (test set) (Table 2), and demonstrate their limitation in the evaluation on these domains.

The experiment results are shown in Table 4. In addition to performance on the **OpenGenAlign** benchmark, we evaluate these models on RewardBench, which assesses their capabilities across chat, safety, and reasoning scenarios. While all listed reward models achieve accuracy near or above 90% on RewardBench domains, their performance drops below 80% on **OpenGenAlign**, This underscores a significant gap between mainstream reward models and the unique requirements of tasks in **OpenGenAlign**, highlighting the distinct challenges posed by long-context generation tasks. We further observe significant performance variation across tasks: reward models demonstrate stronger results on data-to-text and summarization tasks (e.g., Yelp and XSum datasets) compared to

Table 4: The evaluation results of the existing reward models on the 3 tasks. They achieve SOTA performance on helpfulness and instruction-following, but do not excel in Long-Context Understanding.

Models	WebGLM	Yelp	XSum	Average	RewardBench
Llama-3.1-8B-Instruct-RM-RB2 (Malik et al., 2025)	**70.0**	**83.8**	**80.4**	**78.1**	88.8
QRM-Gemma-2-27B (Dorka, 2024)	67.4	83.2	77.4	76.0	**94.4**
Skywork-Reward-Gemma-2-27B-v0.2 (Liu et al., 2024a)	67.8	80.8	78.8	75.8	94.3
QRM-Llama3.1-8B-v2 (Dorka, 2024)	66.6	77.4	78.2	74.1	93.1
URM-LLaMa-3.1-8B (Lou et al., 2024)	64.6	83.6	73.0	73.7	92.9
Llama-3-OffsetBias-RM-8B (Park et al., 2024)	65.6	78.2	77.4	73.7	89.4
GRM-Llama3-8B-rewardmodel-ft (Yang et al., 2024b)	66.6	74.8	79.4	73.6	90.9
Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024a)	65.0	77.4	76.4	72.9	93.1

question-answering tasks (e.g., WebGLM dataset), indicating that current reward models lack uniform capability across different long-context scenarios.

Notably, we observe that strong performance on RewardBench does not guarantee comparable results on **OpenGenAlign**. For instance, Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024a) and URM-LLaMa-3.1-8B (Lou et al., 2024) achieve relatively high overall scores on RewardBench but underperform on **OpenGenAlign**. Conversely, Llama-3.1-8B-Instruct-RM-RB2 (Malik et al., 2025), which scores lower on RewardBench, attains the highest accuracy on our benchmark. This disparity suggests that reward models optimized for general objectives (e.g., chat and safety) may not effectively generalize to long-context generation tasks, which require assessing different quality dimensions such as comprehensiveness and hallucination.

Overall, most of the existing reward models could not excel in expressing the preference in these scenarios. Domain-specific training data are therefore essential to address this gap and improve long-context performance evaluation.

5 Main Experiments

We conduct both reward model training and reinforcement learning using our **OpenGenAlign** dataset. In total, 33K preference pairs are used for reward modeling (see Table 2). Additionally, we use a 9K-sample development set for sampling and learning during RLHF training. To evaluate the performance of the policy and reward models, a held-out benchmark set of 1.5K samples is used.

5.1 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. A formal formulation of the Bradley-Terry reward model is in Appendix A.

Table 5: Evaluation results of the reward model on the three tasks. Accuracy is measured as the percentage of test samples where the reward model assigns a higher score to the chosen response than to the rejected one.

	WebGLM	Yelp	XSum	Avg
Accuracy (%)	80.2	92.0	85.6	85.9

We evaluate reward model performance using pairwise accuracy: for each test sample containing a chosen and rejected response, we measure whether the model assigns a higher score to the chosen response. As shown in Table 5, our reward model achieves 85.9% accuracy, demonstrating strong alignment with preference signals. Notably, our model outperforms all baselines from Table 4, achieving the highest accuracy and highlighting the benefits of training on domain-specific preference data for long-context generation.

Furthermore, we observe a consistent improvement 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.

Table 6: **Win rates (%)** of post-trained policy models on three tasks. Evaluation is done by both our reward model and o3, under two PPO settings: using our reward model vs. using the baseline reward model (Llama-3.1-8B-Instruct-RM-RB2).

Dataset	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct			
	Reward Model		o3		Reward Model		o3	
	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline
WebGLM	**89.6**	75.8	**85.2**	66.6	**83.6**	70.8	**72.4**	62.6
Yelp	**91.6**	75.6	**89.4**	74.1	**95.6**	86.4	**91.0**	85.0
XSum	**90.4**	83.4	**88.6**	75.0	**86.0**	74.6	**80.0**	67.2
Average	**90.5**	78.3	**87.7**	71.9	**88.4**	77.3	**81.1**	71.6

5.2 Reinforcement Learning

We adopt the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) to perform the Reinforcement Learning. We use Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the initial policy models and use Verl (Sheng et al., 2025) as the framework for experiments. For the baseline, we adopt Llama-3.1-8B-Instruct-RM-RB2 (Malik et al., 2025), which achieves the highest score in Table 4, to serve as the baseline reward model. The training details and mathematical formulations are described in Appendix A.2.

To quantify the effectiveness of alignment training, we evaluate our models using a **Win Rate** metric against the initial (pre-training) models. Specifically, we sample responses from a held-out test set using both the initial and post-trained policy models, generating paired responses for each prompt. These response pairs are evaluated through two methods: (1) scoring by the reward model, and (2) pairwise comparison by o3 based on the criteria defined in Section 3. For each evaluation method, we calculate the **win rate as the proportion of cases where the post-trained model's response is preferred over the initial model's response.** A win rate of 50% represents no improvement.

The experiment results are shown in Table 6. We observe a clear improvement in the policy models after PPO. Both the reward model and o3 agree that generations align more closely with the defined 4 metrics, as the average win rate is significantly above 50%. These results highlight the effectiveness of our dataset and the reward model. Additionally, the ratings across the 3 tasks from the reward model is very similar to the ratings from o3. For example, our reward model and o3 both assign the highest preference on the Yelp dataset. This shows that our reward model learns the rationale of rating from o3. It is also observed that our reward model significantly outperforms the baseline when utilized for PPO, where the average win rate for Ours is about 10% higher than the baseline reward model for PPO. This comparison demonstrates the need for the **OpenGenAlign**, which could effectively improve the performance of both the reward model and policy model in these specific domains. However, we also observe some imbalance in learning for the policy model across tasks. Specifically, for the Llama-3.2-3B-Instruct model, the win rate for Yelp could reach 91% while only above 70% for WebGLM. The comparison reveals that the difficulties are various across different domains in **OpenGenAlign**.

5.3 Guided Generation

We also evaluate our reward model on Out-Of-Distribution (OOD) multi-hop long-context Question-Answering datasets, including HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022). We use the same policy model for generation. Given a question, we first use the initial policy model to generate a single solution as the baseline, denoted as **Pass@1**. Then we generate 4 solutions and apply the reward model to select the solution with the highest reward score, denoted as **Best-of-4**. We also use the after-PPO model for evaluation, denoted as **PPO-Pass@1** The experiment results are shown in Table 7. We observe a consistent and significant improvement for both the PPO models and **Best-of-4** compared to the baseline. The results demonstrate the effectiveness and generalization of our reward model, as well as revealing the capability of the policy models when trained using **OpenGenAlign**.

Table 7: Performance of models in the guided generation on OOD datasets HotpotQA and MuSiQue.

Models	HotpotQA			MuSiQue		
	PPO-Pass@1	Best-of-4@RM	Pass@1	PPO-Pass@1	Best-of-4@RM	Pass@1
Llama-3.2-3B	37.2	**39.3**	28.5	**15.3**	15.0	8.7
Llama-3.1-8B	**48.1**	46.4	38.3	**26.4**	24.9	14.6

Table 8: Evaluation results of the Baseline and the MixReward Model, listing the accuracy of the reward models on each category and average from RewardBench (Lambert et al., 2024) and **OpenGenAlign** test set.

	Chat	Chat-Hard	Safety	Reasoning	Average	OpenGenAlign
Baseline	98.4	59.0	93.7	90.5	85.4	74.0
MixReward	**98.7**	**59.2**	**94.0**	**92.4**	**86.1**	**86.0**

6 Training on Mixed Preference Datasets

The ultimate goal of the proposed dataset is to further empower the reward modeling and provide a better judgment on the generation quality. So a critical question is, how well could **OpenGenAlign** be integrated with existing preference datasets to build a more robust and versatile model. We first construct the baseline dataset by randomly selecting 50K examples from Preference-700K[1] (Dong et al., 2024), which is a mixture dataset consisting of Ultra-Feedback (Cui et al., 2024), HH-RLHF (Bai et al., 2022), PKU-SafeRLHF (Ji et al., 2024), SHP (Ethayarajh et al., 2024), HelpSteer (Wang et al., 2023), Ultra-Interact (Yuan et al., 2024), Distilabel-Capybara (Álvaro Bartolomé Del Canto et al., 2024), and Distilabel-Orca (Lian et al., 2023), providing preference signals on chat, safety, and reasoning. We train a baseline reward model using this dataset. Additionally, we mix the baseline dataset with the **OpenGenAlign** to train new reward model named **MixReward**. The training hyper-parameters and settings are the same as 5.1. We evaluate the both the reward models on RewardBench (Lambert et al., 2024) and our curated **OpenGenAlign** test set.

The experiment results are shown in Table 8. We first observe that, on the RewardBench, MixReward still possesses a competent performance compared to the Baseline, even with an about 1% higher average accuracy, although we add 33K data sample irrelevant to chat, safety, or reasoning. It demonstrates that **OpenGenAlign** could be effectively integrated with various existing preference datasets to train a more versatile reward model without hurting the performance on these domains. Additionally, MixReward achieves an accuracy of 86.0% on **OpenGenAlign** test set, which is significantly higher than the Baseline reward model. Notably, the accuracy is very similar to the result in Table 5, showcasing that adding datasets like UltraFeedback, SafeRLHF will also not hurt the performance of the reward model on **OpenGenAlign** domain. These results together prove that **OpenGenAlign** could serve as a complementary resource to existing preference datasets, enhancing the overall diversity and strength of the resulting model, which also further demonstrates the quality and the effectiveness of our carefully curated dataset.

[1] https://huggingface.co/datasets/hendrydong/preference_700K

7 Conclusion

In this paper, we introduce **OpenGenAlign**, a high-quality preference dataset designed for evaluation and improvement of open-ended, long-context generation, which spans *Question Answering, Data-to-Text, and Summarization* domains. Our dataset is generated through an 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 with majority voting to assess generation quality based on four key metrics carefully selected by human experts. The experimental results show strong alignment with human evaluations, demonstrating the effectiveness of **OpenGenAlign** in reward modeling and reinforcement learning. These findings highlight the potential of our dataset to advance both the evaluation and generation of long-context scenarios. And the **OpenGenAlign** could be integrated into the training of a wide range of reward models. To foster further research, we will publicly released the dataset to the community.

Limitations

While **OpenGenAlign** addresses a critical gap in long-context reward modeling, there are still limitations that warrant discussion. Our dataset focuses on three domains: **question answering, data-to-text generation, and summarization.** Although these represent common long-context scenarios, there could be other domains that remain unexplored. Additionally, our dataset is entirely in English, while some of the existing datasets are multilingual. This might limit the performance of our model in other communities.

References

Anthropic. 2025. System card: Claude opus 4 and claude sonnet 4. https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf. Model Card.

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, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *Preprint*, arXiv:2412.15204.

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, and 12 others. 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, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025. Rm-r1: Reward modeling as reasoning. *Preprint*, arXiv:2505.02387.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *Preprint*, arXiv:2305.01937.

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, and 6 others. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Preprint*, arXiv:2406.18495.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. 2025. Skywork open reasoner 1 technical report. *Preprint*, arXiv:2505.22312.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Jie Huang and Kevin Chen-Chuan Chang. 2024. Citation: A key to building responsible and accountable large language models. *Preprint*, arXiv:2307.02185.

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.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main. GitHub repository.

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

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment. *Preprint*, arXiv:2304.07327.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

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.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *Preprint*, arXiv:2407.11963.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, Zheng Zhang, Wei Shen, Qian Liu, Chenghua Lin, Jian Yang, Ge Zhang, and Wenhao Huang. 2025. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *Preprint*, arXiv:2508.17445.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/datasets/Open-Orca/OpenOrca.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024a. 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. 2023a. Webglm: Towards an efficient web-enhanced question answering system with human preferences. *Preprint*, arXiv:2306.07906.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *Preprint*, arXiv:2410.16184.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

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*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *Preprint*, arXiv:2506.01937.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*.

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.

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. Introducing openai o3 and o4-mini. 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, and 262 others. 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.

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.

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

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. *Preprint*, arXiv:2305.14196.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *NeurIPS*.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *Preprint*, arXiv:2401.15042.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *Preprint*, arXiv:2406.12845.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024b. Helpsteer2: Open-source dataset for training top-performing reward models. *Preprint*, arXiv:2406.08673.

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.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *Preprint*, arXiv:2505.11475.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *Preprint*, arXiv:2503.10460.

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.

Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. 2024. An implementation of generative prm. https://github.com/RLHFlow/RLHF-Reward-Modeling.

Wei Xiong, Wenting Zhao, Weizhe Yuan, Olga Golovneva, Tong Zhang, Jason Weston, and Sainbayar Sukhbaatar. 2025. Stepwiser: Stepwise generative judges for wiser reasoning. *Preprint*, arXiv:2508.19229.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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, and 43 others. 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.

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.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *Preprint*, arXiv:2503.18892.

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.

Hanning Zhang, Jiarui Yao, Chenlu Ye, Wei Xiong, and Tong Zhang. 2025. Online-dpo-r1: Unlocking effective reasoning without the ppo overhead. https://efficient-unicorn-451.notion.site/Online-DPO-R1-Unlocking-Effective-Reasoning-Without-the-pvs=4. Notion Blog.

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.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

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.

Álvaro Bartolomé Del Canto, Gabriel Martín Blázquez, Agustín Piqueres Lajarín, and Daniel Vila Suero. 2024. Distilabel: An ai feedback (aif) framework for building datasets with and for llms. https://github.com/argilla-io/distilabel.

A RLHF Experiment Details

A.1 Bradley-Terry Reward Model

To train the Bradley-Terry reward model, we utilize the preference dataset we collected. We denote the dataset as $\mathcal{D} = (x, a^w, a^l)$, where x is the prompt, a^w is the preferred response, and a^l is the dispreferred response. After we get the dataset, we maximize the log-likelihood function of the BT model:

$$\mathcal{L}_{\mathcal{D}}(\theta) = \sum_{(x,a^w,a^l,y)\in\mathcal{D}} \log\left(\sigma\left(R_\theta(x, a^w) - R_\theta(x, a^l)\right)\right).$$

where R_θ is the output of the reward model, and σ is the sigmoid function.

A.2 Proximal Policy Optimization Experiments

Proximal Policy Optimization (PPO) is a reinforcement learning algorithm that has been popular in LLMs. It optimizes LLMs with the following objective function

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{t,s_t,a_t\sim\pi_{\theta_{\text{old}}}}\left[\min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}\hat{A}_t, \text{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}_t\right)\right]$$

where π_θ and π_{old} represent the current and previous policy respectively. $\hat{A}_t$ is an estimator of the advantage function, and ϵ_{low} and ϵ_{high} are hyperparameters that control the maximum deviation from the previous policy $\pi_{\theta_{\text{old}}}$.

We perform PPO with a batch size of 512, max prompt length of 4096, max response length of 2048, and a learning rate of $1e^{-6}$. We train the Llama3.1-8B-Instruct for 20 steps and the Llama3.2-3B-Instruct for 30 steps.

Table 9: Examples from the three datasets and the prompts used to construct preference data. In WebGLM, LLMs generate answers based on reference passages; in Yelp, LLMs convert structured JSON to natural language descriptions; in XSum, LLMs summarize news articles. **Our Prompt** below demonstrates the prompt format we adopt to sample the response.

Dataset	**Data Example**
WebGLM (Liu et al., 2023a)	***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... ***Summary:*** Clean-up operations are continuing across the Scottish Borders and Dumfries and Galloway after...
	Our Prompt: Summarize the following document: *{document}*...

Figure 4: WebGLM System Prompt

WebGLM System Prompt

You are an expert in evaluating question-answering (QA) responses. Your task is to compare two responses, Response A and Response B, based on the following criteria, with a stronger emphasis on hallucination and comprehensiveness:

Primary Criteria (Most Important): Hallucination (Faithfulness & Accuracy) – Highest Priority
Assess whether each response strictly adheres to the provided reference passages. A good response must not introduce fabricated, misleading, or unverifiable information—it should only contain details supported by the reference. Identify which response is more factually accurate and better grounded in the reference material.

Comprehensiveness (Coverage & Relevance) – Second Priority
Determine how well each response fully answers the given question, ensuring that all essential aspects are covered. A strong response should capture key details from the reference while avoiding unnecessary or irrelevant information. Identify which response provides a more complete and well-supported answer.

Secondary Criteria (Supporting Factors) Conciseness (Efficiency & Clarity)
Evaluate how effectively each response conveys the necessary information without excessive verbosity. The ideal response should be succinct yet informative, avoiding redundant details while preserving key insights. If two responses are equally faithful and comprehensive, prefer the one that is more concise and well-structured.

Attribution (Attribution & Use of Retrieved Content)
Examine how well each response incorporates and attributes information from the retrieved passages. A strong response should clearly reference relevant sources when necessary, ensuring that retrieved content supports the answer. If two responses are otherwise equal, prefer the one that makes better citation of the retrieval sources.

Final Judgment: Focus primarily on faithfulness and comprehensiveness when deciding which response is superior. Provide a concise explanation of your reasoning, then explicitly state which response is better using the following format:
Chosen: (A or B)

Figure 5: XSum System Prompt

XSum System Prompt

You are an expert in evaluating the quality of summarization. Your task is to compare two summaries, Response A and Response B, based on the following criteria:

Primary Criteria (Most Important): Hallucination (Faithfulness & Accuracy)
Assess how well each summary adheres strictly to the provided reference. A good summary should only include verifiable information from the reference and avoid adding any fabricated, misleading, or exaggerated details. Identify which summary is more factually accurate and better grounded in the reference.

Comprehensiveness (Coverage & Relevance) – Second Priority:
Determine how well each summary captures the key points of the reference without omitting essential details. A strong summary should convey all critical aspects of the original content while avoiding irrelevant or unnecessary information. Identify which summary provides a more complete and well-balanced representation of the reference.

Conciseness (Efficiency & Clarity):
Compare how effectively each summary delivers the key information in a compact and clear manner. An ideal summary should be succinct yet informative, avoiding excessive verbosity while retaining all necessary details. Determine which summary is more precise and effectively worded.

Final Judgment: Based on the above criteria, provide a brief explanation of your decision. Then, explicitly state which summary is the better one in the following format:
Chosen: (A or B)

Figure 6: Yelp System Prompt

Yelp System Prompt

You are an expert in evaluating data-to-text generation. Your task is to compare two responses, Response A and Response B, which attempt to convert Yelp-style JSON data about a business into plain text. Evaluate them based on the following criteria:

Primary Criteria (Most Important) Hallucination (Faithfulness & Accuracy) – Highest Priority
Assess whether each response strictly reflects the information provided in the JSON input. A good response must not introduce fabricated, misleading, or unverifiable details—it should only include content supported by the JSON data. Identify which response is more factually accurate and better grounded in the input.

Comprehensiveness (Coverage & Relevance) – Second Priority
Determine how well each response captures all the important details from the JSON (e.g., business name, category, rating, address, opening hours, reviews). A strong response should present a comprehensive description while avoiding irrelevant or repeated details. Identify which response provides a more complete and well-supported representation of the input.

Secondary Criteria Conciseness (Efficiency & Clarity)
Assess how effectively each response presents the necessary information without being overly verbose. The ideal response should strike a balance: concise enough to avoid redundancy, yet detailed enough to preserve key insights.

Final Judgment: Focus primarily on faithfulness and coverage when deciding which response is superior. Provide a concise explanation of your reasoning, then explicitly state which response is better using the following format:
Chosen: (A or B)