Reward Modeling for Human Preferences

Evan Frick

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-82
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-82.html

May 16, 2025

Reward Modeling for Human Preferences

by Evan Frick

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Jiantao Jiao
Research Advisor

05/15/25
(Date)

* * * * * * *

Professor Ion Stoica
Second Reader

5/15/2025
(Date)

Reward Modeling for Human Preferences

by

Evan Frick

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jiantao Jiao, Advisor
Professor Ion Stoica

Spring 2025

Reward Modeling for Human Preferences

Abstract

Reward Modeling for Human Preferences

by

Evan Frick

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Jiantao Jiao, Advisor

Reinforcement Learning from Human Feedback (RLHF) has become the dominant paradigm for aligning Large Language Models (LLMs) with human preferences. Effective RLHF relies heavily on reward models, which serve as scalable proxies for human judgments. We introduce a new benchmark for reward models that quantifies their ability to produce strong language models through RLHF (Reinforcement Learning from Human Feedback). The gold-standard approach is to run a full RLHF training pipeline and directly probe downstream LLM performance. However, this process is prohibitively expensive. In this thesis, we introduce Preference Proxy Evaluations (PPE), a comprehensive benchmark suite grounded in large-scale, crowdsourced human preference data and verifiably correct responses from established benchmarks. We experimentally validate PPE by demonstrating its strong correlation with downstream human preferences observed after RLHF processes, underscoring its predictive capability. Ultimately, we compile our data and findings into Preference Proxy Evaluations (PPE), the first reward model benchmark explicitly linked to post-RLHF real-world human preference performance. Additionally, we leverage insights from PPE to enhance reward model robustness by incorporating advanced heteroscedastic regression techniques, addressing variability and uncertainty inherent in human preference data. We find that learning to estimate variances increases final performance, outperforming fixed variance or variance free alternatives– even when the variance estimates are not utilized at test time. Further, we find that using variances estimates to form a pessimistic quantile reward benefits reward model performance and robustness– especially on out-of-distribution tasks. In general, these results suggest that these reward models may serve as more robust human preference proxies during online RLHF procedures, which require reward models to be robust to an ever-changing policy model.

To my parents.

Contents

List of Figures

List of Tables

Acknowledgments

Before we begin, some acknowledgements are due. First, I extend my deepest gratitude to my advisor Jiantao Jiao, who has greatly supported my early research career. I also extend my thanks to Ion Stoica, who has been immensely supportive of my work. Additionally, I thank Joseph Gonzales for his advice and guidance during my research journey.

I must also acknowledge my housemates, Oliver, Dhruv, and Leo, who have been the best. In particular, Oliver for a constant stream of freshly baked cookies and Dhruv for a constant stream of dialogue.

I thank Tim, who has walked this research journey closely along side me, constantly sharing ideas and motivating me to be a better researcher. I hope we have many more collaborations in the future.

Anastasios Angelopoulos and Wei-Lin Chiang have simply been transformative on my research career, trajectory, and mentality. Their mentorship has been invaluable. I'm truly lucky that they have chosen to invest their confidence in me.

Finally, I thank Banghua Zhu– the first to believe in me as a researcher– who was the first and only person to respond to my email asking for research early senior year. What began as a short email reply has grown into an amazing research collaboration, which I hope will only continue to grow over time– we have many more great papers to write. Banghua's guidance and mentorship have had a profound influence towards the researcher I am today. For this, I am immeasurably indebted and eternally grateful.

Chapter 1

Introduction

The rapid ascent of Large Language Models (LLMs) has marked a transformative era in artificial intelligence, offering unprecedented capabilities in natural language understanding, generation, and reasoning. While many of us stand in awe of the rapid acceleration of artificial intelligence towards human-level or super-human performance, we should not discount shear volume of human hand-labeled examples and demonstrations that have powered this growth. Centered around this growth of intelligent automation is precisely the human, who's guidance has been the driving learning signal. Thus, Reinforcement Learning from Human Feedback (RLHF) [10] has emerged as the predominant methodology for achieving alignment towards humans, shaping LLMs to better help humanity.

1.1 The Pivotal Role of Reward Models in RLHF

At the heart of the RLHF paradigm lies the reward model. Given the significant expense and time involved in directly soliciting human preference labels for every potential LLM output, reward models serve as crucial proxies for human judgment [10]. During RLHF, and LLM policy model is optimized to maximize the reward model's human preference proxy signal via Reinforcement Learning (RL), in the process learning to generate responses that the reward model predicts humans would prefer [10, 5, 30, 40, 28, 4, 19, 26, 29, 49, 50].These reward models are typically trained on datasets of human-provided comparisons, learning to discern preferred responses from contrastive losses and choice models, thereby quantifying the "strength" of preference as a scalar reward [10, 30].

With the success of these methods, human preference has emerged as one of the gold standards for LLM training and evaluation. Several large-scale human preference datasets have been developed, including Stanford Human Preference (SHP) [11], Chatbot Arena [8], and Anthropic HH [5], among others. Researchers requiring human preference proxies have sought to replicate these preferences with learned reward models.

1.2 The Critical Gap: Evaluating Reward Model Efficacy

The indispensable role of reward models in the RLHF pipeline underscores the need for robust methods to evaluate their performance. While benchmarks have been developed for this purpose [18], a fundamental challenge persists: the typical reward model evaluation task—often involving selecting correct answers from predefined, ground-truth examples–is substantially different from the reward model's true operational use-case, which is to provide a nuanced learning signal that effectively drives the RLHF optimization process.

This consideration demands deeper study into reward model benchmarking, particularly the correlation between evaluation signals on reward models against most RLHF language model success during deployment. We aim to improve upon this gap with our findings. In this work, we produce a reward model evaluation that is grounded in context with measured down stream RLHF outcomes. These chapters of the thesis are adapted from prior work [13].

1.3 Bridging the Gap: Towards Principled Benchmarking and Robust Reward Models

Developing a principled reward model benchmark with explictly measured correlation to down stream RLHF outcomes gives us the trust necessary to further develop reward modeling methodologies without undergoing expensive RLHF pipelines to observe true post-RLHF results. Rather, we can iterate on a simple, easy to run reward model benchmark with the confidence gains in benchmark performance will translate to real-word RLHF use-cases. Therefore, addressing this critical evaluation gap is the central motivation of this thesis. Additionally, in this work we leverage our reward model benchmark study methods to increase reward model robustness with heteroskedastic regression on human preferences.

This thesis makes the following key contributions:

1. Development of Preference Proxy Evaluations (PPE): We introduce PPE, a novel benchmarking suite designed to evaluate reward models more holistically. PPE incorporates diverse datasets, including real-world human preferences and verifiable correctness tasks, to assess RM accuracy, robustness, and correlation with human judgment across various domains.

2. Empirical Validation of PPE: We conduct extensive experiments to validate PPE by measuring the correlation between RM performance on our benchmark and the actual downstream performance of LLMs fine-tuned using these reward models via DPO, a common RLHF algorithm. This establishes an empirical link between our evaluation metrics and real-world RLHF efficacy.

3. Advancing Robust Reward Modeling: Leveraging the insights and validated evaluation framework from PPE, we explore methods to enhance reward model robustness.

Specifically, we investigate the application of heteroskedastic regression for modeling human preferences, aiming to create reward models that are more stable and reliable, particularly when faced with noisy or out-of-distribution data.

Chapter 2

Evaluating Reward Models for RLHF

2.1 Background

The ultimate test of a reward model is as follows:

> Does the reward model lead to good post-RLHF language model performance?

In other words, because the reward model will be used as a reference signal for LLM training, in principle, only the downstream LLM performance matters. However, to evaluate downstream performance, we must train a new LLM using the reward model and evaluate the resulting LLM—a prohibitively expensive and time-consuming process, shown in Figure 2.1. The long development-feedback cycle of reward models poses a significant challenge, limiting achievable reward model quality and, consequently, limiting the effectiveness of the entire RLHF process.

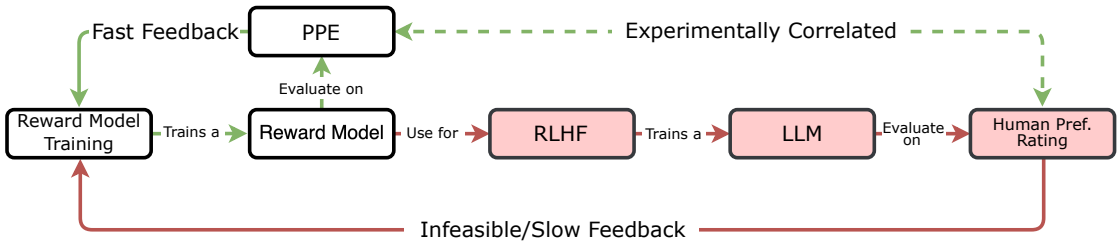

Figure 2.1: Overview of the RLHF pipeline. Reward models feed into the very beginning of the RLHF pipeline, making iterative improvements prohibitively slow. PPE enables a fast feedback loop that is correlated to downstream outcomes.

We first introduce a cost-effective method for approximating the effect of a reward model on downstream LLM performance. Specifically, we measure reward model performance using

a large-scale, crowdsourced pairwise human preference evaluation dataset collected via Chatbot Arena [8], as well as a high-quality, programmatically verifiable correctness preference dataset. To avoid introducing bias, we do not utilize LLM judges or expert annotators to provide ground-truth references. Instead, we focus on real-world preference data that reflects organic LLM usage. Additionally, we aim our evaluation tasks to closely resemble real-world RLHF training, making the assessment more aligned with practical use cases. Moreover, to bridge the existing knowledge gap between reward model evaluations and actual post-RLHF outcomes, we experimentally correlate our evaluation metrics with real human preferences on RLHF-ed LLMs. To achieve this, we used select reward models within a full RLHF training pipeline, each producing a fine-tuned LLM. These RLHF-tuned models are then deployed on a crowd-sourced human preference platform where we directly measure their downstream human preference scores. Through this end-to-end analysis, we identify which metrics across diverse domains show the strongest correlation with real-world post-RLHF performance. By validating this correlation, we ensure that iterative improvements on our evaluation will lead to tangible gains in downstream performance.

Additionally, we release PPE[1], a crowdsourced collection of 16,038 labeled human preference pairs containing responses from 20 different top LLMs and over 121 languages as well as a dataset of 2,555 prompts, each with 32 different sampled response options, totaling 81,760 responses across 4 different models, all grounded with verifiable correctness labels. PPE evaluates reward models on 12 different metrics and 12 different domains, such as their accuracy in selecting human-preferred or verifiably correct responses. Notably, PPE is the *only* reward model benchmark directly linked to downstream RLHF outcomes.

In this chapter we explore the following contributions:

1. We analyze how reward model metrics correlate with real downstream human preference performance post-RLHF.

2. We fully open-source PPE, a comprehensive benchmark for reward models with metrics directly linked to downstream RLHF outcomes.

2.2 Sourcing Ground Truth Preference Labels

Previous work on sourcing preference ground truth labels often relies upon LLM judge preference labels in conjunction with manual verification from individuals, introducing potential preference biases [18]. Alternatively, rejected responses are often curated synthetically by unnaturally perturbing the chosen output or modifying the prompt to produce forced errors, introducing bias on how errors look and occur. These preference pairs are not representative of the distribution of responses seen by reward models when providing learning signals for RLHF.

Thus, we ground our preference labels with the following methodology:

[1]PPE is available on Github at: lmarena/PPE

1. Utilize crowdsourced diverse prompts and responses with human preference labels.
2. Utilize existing benchmarks with verifiable correctness checks on LLM-generated responses.

The methodology (1) provides an unbiased estimate of real-world human preference through the aggregation of many diverse human preferences. We use a large crowdsourced preference set of 16,038 preference labels to mitigate individual label noise and avoid overfitting to any single individual's preference, details in Section 2.3.

Methodology (2) curates an objective correctness signal naturally unbiased by response style. We use the second approach to label the correctness of many sampled responses from an LLM, mimicking rollouts or best-of-k exploration strategies seen in RLHF training processes. As a result, we draw preference pairs from more naturally occurring distributions (eg. real LLM responses and errors), better align with the expected environment reward models operate in. An overview of PPE is provided in Table 2.1

Name	Num Prompts	Response per Prompt	Preference Type
Human Preference V1	16,038	2	Real Human
MMLU Pro	512	32	Correctness
MATH	512	32	Correctness
GPQA	512	32	Correctness
IFEval	512	32	Correctness
MBPP Plus	507	32	Correctness

Table 2.1: Released benchmarking datasets in PPE.

2.3 Human Preference Metrics

To benchmark whether a reward model aligns with human preference directly, we utilize a human preference dataset collected from a large-scale human preference annotation platform that allows users to vote on pairwise comparisons between responses generated from two anonymized and randomly selected LLMs. Our human preference dataset contains human-labeled preferences for 16,038 pairwise comparisons between 20 selected top models[2]. These models were selected based on their strong performance on Chatbot Arena and overall popularity [8]. We emphasized selecting models that have already undergone some form of RLHF, anticipating that these models would be more challenging for reward models to evaluate.

[2]mistral-large-2402, phi-3-medium-4k-instruct, gpt-4-1106-preview, claude-3-opus-20240229, gemini-1.5-pro-api-0514, gpt-4-0314, claude-3-haiku-20240307, gpt-4-0613, claude-3-sonnet-20240229, yi-1.5-34b-chat, llama-3-8b-instruct, gemini-1.5-flash-api-0514, llama-3-70b-instruct, gpt-4o-2024-05-13, command-r-plus, gpt-4-turbo-2024-04-09, qwen2-72b-instruct, command-r, qwen1.5-72b-chat, starling-lm-7b-beta

Since the human preference set is crowd-sourced, we can repeat the collection process at any time to obtain an updated set that better reflects the current array of available models and any changes in human preference. Additionally, a newly updated human preference set would largely mitigate benchmark leakage that may have occurred with the previous set. Consequently, this human preference metric can remain consistently up-to-date with fresh, relevant data.

Curation

Specifically, we curate our human preference data from crowd-sourced battles. A "battle" consists of a user-provided prompt, two models and their responses to the prompt, and the user's preference vote for the responses. We perform a random sample weighted by model occurrence to obtain 50,000 collected battles between selected models such that models are represented at a uniform frequency, then de-duplicate and remove any samples containing P.I.I information using Azure AI. We use OpenAI's moderation API to flag and remove potentially harmful conversations from the sample. Finally, we subsample 16,038 pairs from the remaining battles to construct the human preference benchmark dataset.

The human preference benchmark dataset, at a glance:

1. Includes 4,583 instruction-following prompts, 5,195 hard prompts, 2,564 math prompts. Prompts may exist in multiple categories.

2. Includes user queries from over 121 languages. Top languages include English (8,842), Chinese (1,823), Russian (1,779), German (568), Korean (338), Japanese (321), etc.

3. Includes preferences crowdsourced from 6,120 individuals.

Scoring

We conduct several statistical metrics described below to evaluate different aspects of a given reward model.

1. **Accuracy.** We compute pairwise ranking accuracy against a human preference label for each reward model, excluding battles in which the human rater selected a "tie". This measures the granular case-by-case similarity to a real human preference signal.

2. **Correlation.** Since each battle contains information on model identities, each reward model produces a ranking and a pairwise win-rate matrix for the 20 selected models. We compute Spearman and Kendall correlation between the model ranking produced by each reward model against the ground truth ranking. In addition, we compute row-wise Pearson Correlation between the win-rate matrix produced by each reward model against the ground truth win-rate matrix. We intuit that these aggregate correlation metrics measure overall similarity to real human preference.

3. **Confidence.** To weight stability in assigning preferences, we follow the metrics proposed in Arena-Hard-Auto [21], where we measure each reward model's Separability with

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	68.59	82.49	84.21	96.21	87.37	96.54	0.05
Ensemble-Judges (AlpacaEval)†	68.52	81.25	79.47	93.94	85.26	95.04	0.07
GPT-4o-2024-08-06 (ArenaHard)†	67.71	81.07	80.53	94.70	86.32	96.24	0.06
Claude-3-5-Sonnet-20240620 (ArenaHard)†	67.33	80.65	79.47	94.70	88.42	96.69	0.06
GPT-4o-2024-08-06 (AlpacaEval)†	67.13	77.92	76.32	90.91	84.21	93.23	0.07
Athene-RM-70B	66.56	80.69	84.74	93.94	82.11	93.23	0.07
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.46	78.42	75.26	92.42	83.16	93.08	0.07
Gemini-1.5-Pro-002 (AlpacaEval)†	66.09	82.63	83.16	96.21	86.32	95.19	0.05
Gemini-1.5-Pro-002 (ArenaHard)†	65.71	82.23	83.16	94.70	90.53	96.99	0.04
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	65.34	73.91	74.21	85.61	71.58	85.26	0.11
Llama-3.1-70B-Instruct (AlpacaEval)†	65.27	74.81	79.47	87.88	72.63	85.56	0.12
Gemini-1.5-Flash-002 (AlpacaEval)†	65.04	74.29	78.95	88.64	74.74	88.72	0.11
Athene-RM-8B	64.59	76.85	83.68	91.67	77.89	90.53	0.10
Llama-3.1-70B-Instruct (ArenaHard)†	64.29	74.77	75.79	85.61	70.53	87.07	0.12
Gemini-1.5-Flash-002 (ArenaHard)†	63.01	76.12	76.32	90.91	76.84	90.23	0.10
Starling-RM-34B	62.92	70.47	77.37	78.79	67.37	81.20	0.15
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	62.75	68.86	70.53	84.09	75.79	88.12	0.10
Gemini-1.5-Pro-001 (ArenaHard)†	62.57	75.92	81.05	93.18	85.26	94.44	0.07
Skywork-Reward-Llama-3.1-8B	62.37	75.51	78.95	87.88	71.58	88.12	0.11
InternLM2-7B-Reward	62.05	68.03	78.42	69.70	56.84	76.09	0.20
Eurus-RM-7B	62.02	60.37	75.26	64.39	51.58	65.26	0.22
InternLM2-20B-Reward	61.00	66.66	74.74	70.45	55.79	76.39	0.20
ArmoRM-Llama3-8B-v0.1	60.57	71.85	76.84	84.85	76.84	89.17	0.10
NaiveVerbosityModel	59.81	32.03	76.32	35.61	29.47	33.53	0.33
Nemotron-4-340B-Reward	59.28	66.96	78.95	78.79	68.42	86.02	0.14
Llama-3-OffsetBias-RM-8B	59.12	58.86	65.79	61.36	51.58	69.02	0.20
Starling-RM-7B-Alpha	58.93	58.42	70.00	67.42	50.53	64.66	0.22
InternLM2-1.8B-Reward	57.22	47.11	69.47	41.67	36.84	54.14	0.28
Skywork-Reward-Gemma-2-27B	56.62	69.99	69.47	87.88	84.21	95.49	0.07

Table 2.2: Reward model and LLM judge performance on Overall subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Confidence Interval, Confidence Agreement, and Brier Score against ground truth ranking. These metrics are designed to measure uncertainties and overconfidence within a reward model.

Furthermore, we can calculate all the above scores conditioned on any subset of prompts in the evaluation data, specifically capturing 7 different domains. For example, we can observe these metrics on only math prompts or only instruction following prompts. We expect that strong reward models should score high regardless of the selected domain. Scores for the overall subset are detailed in Table 2.2.

2.4 Correctness Metrics

To measure a reward model's ability to distinguish between different samples drawn from the same distribution, we utilize correctness metrics on established, reputable benchmarks with verifiable ground truths (e.g. MBPP-Plus [3]). We construct a benchmark dataset wherein each prompt is associated with 32 different responses sampled from the same LLM. Additionally, since we use benchmarks with verifiable ground truths, we can score the correctness

(a binary label) of each response according to the original static benchmark's verification function (e.g., code unit tests or Regex matching).

To assess the performance of reward models (and LLMs-as-judges), we obtain rewards/preferences for the sampled responses and evaluate how well these align with the verifiable correctness signal, with the general assumption that expert humans would always prefer correct answers over incorrect ones. Our response sampling strategy ensures that the preference labeler must disentangle the correctness signal from potentially very similar or even adversarial outputs, thereby increasing task difficulty. Moreover, this method naturally samples "unforced" errors as they would appear in real training or evaluation schemes, rather than synthetically constructing preference pairs that may contain underlying confounding biases.

Curation

For the correctness metrics, we selected standard, widely used, reputable, and verifiable benchmarks: MMLU Pro [43], MATH [16], GPQA [34], MBPP Plus [3], and IFEval [47]. Each benchmark covers a different domain: general knowledge, mathematics, STEM, coding, and instruction following, respectively. While we initially curate PPE with these five benchmarks, it should be noted that any desired verifiable benchmark can be added to the correctness measurement paradigm by repeating the process outlined below, thereby providing a framework for customization towards specific evaluation needs.

For each benchmark, we sample LLM responses for 500 randomly selected prompts, each 32 times, for a total of 16,000 completions. If a benchmark has fewer than 500 prompts, we use all available prompts. We choose a large K of 32 to allow models to generate more diverse responses, covering a larger input domain for the human preference proxy and testing greater robustness to over-optimization. We note that this sampling strategy yields very similar KL-Divergence shifts as would be seen in RLHF training methods such as Proximal Policy Optimization (PPO) [14, 35].

We repeat this process for four different models: Llama-3-8B-Instruct, Gemma-2-9b-it, Claude-3-Haiku, and GPT-4o-mini-2024-07-18 [1, 38, 2, 27]. Each model samples prompts randomly with different seeds. We reason that different model response distributions may have different difficulties. For example, an already extremely high-performing model like GPT-4o-mini-2024-07-18 may be more challenging for reward models to evaluate correctness.

We then score all responses using the benchmark's verification methods. Using the correctness labels for all responses, we discard any rows in which the model got every single response wrong or every single response right, as it is impossible for the reward model to select a better generation in these cases. Additionally, we discard any row where less than 10% or greater than 90% of the responses were correct, with exceptions made for benchmarks with very few valid options. This step helps avoid vacuously correct responses, such as an LLM randomly guessing the correct multiple-choice answer with completely nonsensical reasoning, as well as prompts that are too easy.

From the remaining data, we randomly sample 128 responses from each model, totaling 512 samples. If a benchmark is too small and some models have fewer than 128 viable

samples, we adjust the sampling accordingly.

Small Benchmark Modifications

To ensure more natural responses that better reflect real-world use cases, we modified each verifiable benchmark's canonical prompt to encourage Chain of Thought (CoT) thinking (citation). This approach both increases the diversity of sampled responses and enhances the task difficulty for the human preference proxy by incorporating additional signals beyond final answer correctness. The specific instructions for each benchmark are detailed below.

For the MATH benchmark, we implemented a new system prompt to facilitate zero-shot CoT behavior. Additionally, we converted the parsed answer to its symbolic representation and utilized a symbolic solver to evaluate true equality instead of relying on raw string matching. This refinement of the correctness signal ensures that trivial answer differences, such as $1\frac{3}{4}$ vs $\frac{7}{4}$ or $\frac{4i+\sqrt{5}}{2}$ vs $\frac{\sqrt{5}}{2}+2i$, are marked as equivalent, with either answer accepted if correct.

In practice, we observed that the sampled MBPP-Plus generations from some models were almost all identical. Models also generally failed to follow instructions to "think step-by-step" before providing their final answers, suppressing answer diversity. To address this issue, we prompted the models to "write comments clearly explaining each part of the code," thereby lengthening trajectories and yielding greater exploration of the answer spaces. We also observed some ambiguity in MBPP-Plus instructions. To mitigate this, we added standard MBPP test cases into the function docstring as examples, and used the more extensive remaining MBPP-Plus test cases as the real tests.

Lastly, for IFEval, we prefixed the prompts with "It is extremely important that you follow all instructions exactly." This addition emphasizes the necessity of precise instruction following in these tasks and ensures that the human preference proxy implicitly recognizes this as a significant evaluation criterion.

The prompt template for MMLU-Pro and GPQA was adaption from the Language Model Evaluation Harness [15]. The MATH template was generated with the assistance of Anthropic's prompt generator.

The prompt templates for each benchmark are detailed below. Note that `{{var}}` indicates a field to be filled by prompt data or metadata.

MMLU Prompt Template:

```
The following are multiple choice questions (with answers) about {{domain}}.  Think step
by step and then finish your answer with "the answer is (X)" where X is the correct letter
choice.

Question: {{question}}
Options:
{{letter}}. {{choice}}
{{letter}}. {{choice}}
{{letter}}. {{choice}}
...
```

MATH Prompt Template:

```
You are a highly skilled mathematician tasked with solving complex math problems.
Your goal is to provide clear, step-by-step solutions that can be easily parsed and
evaluated.

Here is the math problem you need to solve:

<problem>
{{MATH_PROBLEM}}
</problem>

Box your final answer using LaTeX, for example: $x = \\boxed{[Your final numerical or
algebraic answer]}$.

Now, please solve the given math problem and provide your solution in the specified format.
```

GPQA Prompt Template:

```
The following is a {{domain}} multiple choice question.  Think step by step and then finish
your answer with "the answer is (X)" where X is the correct letter choice.

Question: {{question}}
Choices:
(A) {{choice1}}
(B) {{choice2}}
(C) {{choice3}}
(D) {{choice4}}
```

MBPP-Plus Prompt Template:

````
Below will be an instruction to write a python function that accomplishes a task.
You will also be given starter code with a function definition and any required imports.
Think step-by-step, write comments clearly explaining each part of the code, and make sure
your code solution is enclosed in markdown ticks (``` [your code here] ```).

<instruction>
{{instruction}}
</instruction>

<starter_code>
```
{{starter_code}}
    pass
```
</starter_code>
````

IFEval Prompt Template:

```
It is extremely important that you follow all instructions exactly:
{{prompt}}
```

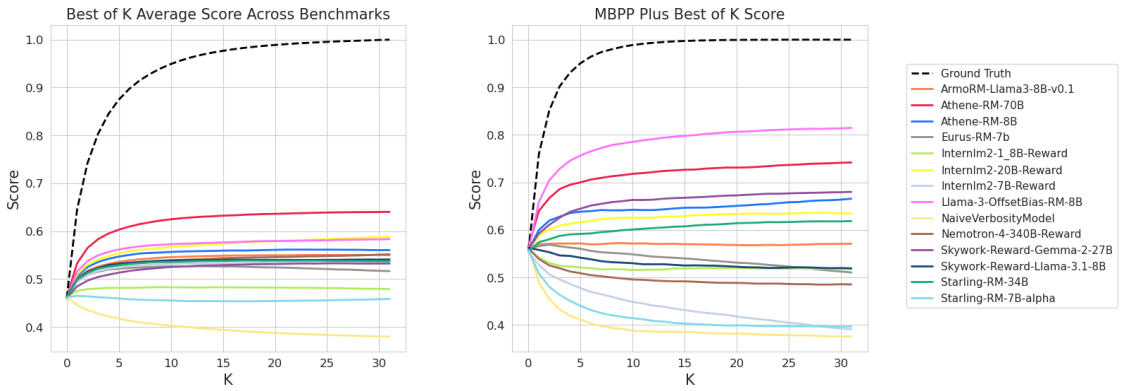

Figure 2.2: Best of K curves showing reward model score vs K. The black dashed line is the theoretical optimal curve; the closer to this curve implies a better score. The left graph shows each reward model's curve averaged across all correctness PPE benchmarks. The right graph shows each reward model's curve on just the MBPP-Plus set, where over-optimization behavior is seen in some reward models, characterized by curves that decrease with respect to increases in K.

Scoring

We score the reward models on the correctness metrics in ways that target a reward model's robustness, granularity, and theoretical roof-line performance. Additional details on reward model and llm-judge scores can be found in Appendix A.2.

Best of K Curves

A best of K curve shows on average how the reward model's selected "best" answer's ground truth score changes vs K. When plotted against the ground truth curve, we can observe the gap between the reward model's ability to select the "best" answer given a set of K responses, and the "gold standard" best score. More formally, let S_K be a size K random sample of responses from a model, $g : S_K \to \{0, 1\}$ be the ground truth scoring function, and $\hat{R} : S_K \to \mathbb{R}$ be the reward model proxy score. Then, $\mathbb{E}_{S_K}[g(\arg\max_{s \in S_K} \hat{R}(s))]$ is the expected ground truth score of the selected response by the reward model given K sampled responses. We then sweep across K = 1,..., 32 to obtain a curve. Best of K scores for various reward models are detailed in Table 2.3.

These curves represent how much the reward model can differentiate the LLM's generations whilst picking from examples drawn from the same distribution. The simple intuition here is that as K increases, the "exploration" of the LLM is expanded, thereby increasing the likelihood that a correct answer lies within the K different samples. However, as exploration increases, the likelihood that a response that exploits the reward model is present also increases. In all the best of K metrics, we use $K = 32$, providing both reasonable infer-

Reward Model	MMLU Pro	Math	GPQA	MBPP Plus	IF Eval	Mean
Athene-RM-70B	0.761	0.607	0.499	0.748	0.633	0.650
InternLM2-20B-Reward	0.673	0.538	0.471	0.654	0.652	0.598
Llama-3-Offsetbias-RM-8B	0.590	0.481	0.450	0.819	0.646	0.597
Athene-RM-8B	0.656	0.517	0.459	0.675	0.586	0.579
Nemotron-4-340B-Reward	0.697	0.499	0.484	0.567	0.623	0.574
InternLm2-7B-Reward	0.638	0.552	0.457	0.562	0.658	0.573
ArmoRM-Llama3-8B-v0.1	0.654	0.508	0.470	0.602	0.601	0.567
Skywork-Reward-Llama-3.1-8B	0.641	0.500	0.468	0.581	0.639	0.566
Starling-RM-34B	0.651	0.476	0.453	0.634	0.569	0.557
Eurus-RM-7B	0.607	0.516	0.438	0.590	0.594	0.549
Skywork-Reward-Gemma-2-27B	0.550	0.462	0.447	0.691	0.583	0.547
InternLM2-1-8B-Reward	0.538	0.411	0.451	0.572	0.581	0.510
Starling-RM-7B-Alpha	0.562	0.409	0.433	0.559	0.564	0.505
NaiveVerbosityModel	0.487	0.349	0.420	0.568	0.539	0.473

Table 2.3: Reward Model Best of K Performance Across Benchmarks

ence costs balanced with a significant enough exploration space to test the reward model's capabilities.

In order to distill the curves into interpretable numbers, we propose several metrics:

1. **Maximum Achieved Performance:** the maximum score achieved by the reward model at any point on the best of K curve. Note that the maximum achieved performance is relatively agnostic to over-optimization.

2. **Error With Respect to Ground Truth:** the expected squared error between the score of the reward model's selected response against the ground truth best response. Once again, let S_K be a size K random sample of responses from a model, $g : S_K \to \{0, 1\}$ be the ground truth scoring function, and $\hat{R} : S_K \to \mathbb{R}$ be the reward model proxy score. Then, the error with respect to ground truth is $\frac{1}{32} \sum_{K=1}^{32} \mathbb{E}_{S_K}[(g(\arg\max_{s \in S_K} \hat{R}(s)) - \max_{s \in S_K} g(s))^2]$

3. **End Score:** We also look at the final score achieved by the reward model at $K = 32$. If no over-optimization has occurred, this should also be the maximum achieved performance.

Area Under Receiver Operator Characteristics (ROC) Curve

Since the ground truth verification outputs a binary label, we can check each reward model's strength as a binary correctness classifier by calculating the area under the ROC curve. We first normalize the scores in each row with min-max normalization. Then we calculate the

Reward Model	MMLU Pro	Math	GPQA	MBPP Plus	IF Eval	Mean
Athene-RM-70B	0.792	0.760	0.603	0.661	0.594	0.682
Internlm2-20B-reward	0.677	0.691	0.562	0.574	0.595	0.620
Llama-3-offsetbias-RM-8B	0.631	0.617	0.541	0.710	0.594	0.619
Athene-RM-8B	0.683	0.673	0.560	0.602	0.556	0.615
Nemotron-4-340B-Reward	0.704	0.660	0.570	0.506	0.587	0.605
Skywork-Reward-Llama-3.1-8B	0.663	0.678	0.560	0.523	0.586	0.602
Internlm2-7B-Reward	0.665	0.718	0.558	0.464	0.605	0.602
ArmoRM-Llama3-8B-v0.1	0.678	0.659	0.549	0.538	0.573	0.599
Starling-RM-34B	0.683	0.621	0.547	0.534	0.536	0.584
Eurus-RM-7B	0.627	0.665	0.521	0.537	0.554	0.581
Skywork-Reward-Gemma-2-27B	0.542	0.582	0.506	0.572	0.536	0.547
Internlm2-1-8B-Reward	0.561	0.587	0.538	0.462	0.538	0.537
Starling-RM-7B-Alpha	0.547	0.527	0.506	0.400	0.519	0.500
NaiveVerbosityModel	0.495	0.528	0.506	0.330	0.511	0.474

Table 2.4: Area Under ROC Curve for Reward Models across Benchmarks

binary classification ROC curve using the normalized scores as "probabilities". AUC scores are detailed in Table 2.4.

Accuracy

Since LLM-as-a-judge cannot easily scale 32-wise judgments, we create a supplemental pairwise task to evaluate correctness preference accuracy compatible with both reward models and LLM-as-a-judge. For each row of best of K data, we simply sample 5 pairs of responses such that in each pair, there is one correct response and one incorrect response. Then, after randomizing positions, the LLM-as-a-judge picks the preferred response. We then measure the accuracy as the rate in which the correct response is preferred over the incorrect result. The accuracies for reward models are also collected for comparison. All scores are documented in Table 2.5.

Reward Model	MMLU-Pro	MATH	GPQA	MBPP-Plus	IFEval	Mean
Athene-RM-70B	0.77	0.79	0.59	0.68	0.62	**0.69**
Claude 3.5 (ArenaHard)†	**0.81**	**0.86**	**0.63**	0.54	0.58	0.68
Llama-3-OffsetBias-RM-8B	0.62	0.68	0.55	**0.74**	0.62	0.64
GPT-4o-mini (ArenaHard)†	0.71	0.81	0.57	0.54	0.56	0.63
Llama-3.1-70B (ArenaHard)†	0.73	0.73	0.56	0.58	0.56	0.63
internLM2-20B-Reward	0.68	0.70	0.57	0.58	0.62	0.63
Athene-RM-8B	0.68	0.71	0.55	0.62	0.57	0.62
ArmoRM-Llama3-8B-v0.1	0.66	0.71	0.57	0.54	0.58	0.61
Skywork-Reward-Llama-3.1-8B	0.64	0.70	0.57	0.52	0.61	0.61
Nemotron-4-340B-Reward	0.70	0.65	0.57	0.49	0.63	0.61
internLM2-7B-Reward	0.67	0.73	0.55	0.44	**0.64**	0.60
Llama-3.1-70B (Alpaca)†	0.66	0.66	0.56	0.52	0.56	0.59
Claude 3.5 (Alpaca)†	0.66	0.63	0.56	0.52	0.57	0.59
Skywork-Reward-Gemma-2-27B	0.54	0.63	0.53	0.59	0.54	0.56
GPT-4o-mini (Alpaca)†	0.57	0.64	0.53	0.52	0.56	0.56
NaiveVerbosityModel	0.48	0.50	0.48	0.31	0.52	0.46

Table 2.5: Reward model and LLM-as-a-judge scores on the correctness accuracy metric. LLM-as-a-judge is marked with †.

Chapter 3

Validating PPE on Post-RLHF Outcomes

3.1 Validation Study Setup

By testing a reward model performance on a benchmark, we hope to glean insight towards downstream performance on an LLM RLHF-ed using a given reward model. To measure how well different metrics in PPE correlate to post-RLHF LLM performance on real-world human preference, we conduct an experiment in which we RLHF a given base LLM using different reward models. We then measure the real-world human preference scores of the resulting LLMs to understand the true performance of the original reward models.

For our experimental setup, we use each reward model to individually RLHF Llama-3.1-8B-Instruct through Direct Preference Optimization (DPO) [33]. This way, we can compare LLMs tuned on identical RLHF pipelines, except for the reward model being measured. Then, these RLHF-ed LLMs are deployed to a crowd-sourced annotation platform to collect real-world human pairwise preferences between model answers. Overall, 12,190 human votes were collected and compiled into relative rankings between these RLHF-ed LLMs. Under this controlled RLHF experiment, the non-noise variance in final human preference rankings attained by these models is dependent only on the reward model choice, effectively measuring the downstream performance of these reward models, albeit on a single model base model undergoing off-policy DPO RL training.

#Votes	Est. Unique Users	Mean Votes/User	Median Votes/User	Mean Battles/Pair	Mean Votes/Model
12190	6120	1.99	1.00	190.47	2031.67

Table 3.1: Statistics on vote participation and distribution for crowdsourced human preference labels.

Training Procedure

Nine[1] reward models were selected to act as preference labels in a full RLHF training pipeline in which the resulting models were evaluated on real human preferences. We constrained this experiment to nine models for cost reasons– the RLHF and human preference evaluation process are exceedingly expensive. We selected popular, newer, and high-performing reward models from RewardBench. We reason that these will be the most difficult reward models to differentiate. We also require the selected reward models to be general-purpose reward models, and not specifically tuned to any single domain or task.

We create a training dataset by first including 7,000 prompts sampled from the original 50,000 human preference votes after PII removal, unsafe prompt removal, and de-duplication. We then add 500 random prompts from MMLU-Pro that are not in PPE, and another 500 prompts from MATH train set (also mutually exclusive from PPE). For each prompt, we sample 16 responses from the base model, Llama-3.1-8B-Instruct, randomizing the temperature for each generation, drawing from a triangular distribution ($a = 0.0, b = 1.0, c = 1.3$) to promote more diverse exploration. This process yields 8,000 total prompts, each with 16 different responses, totaling 128,000 responses.

Each reward model then constructs its own preference dataset. First, the reward model gives scores for each of the 16 responses for each prompt. The "chosen" response is set as the maximum scoring response. The "rejected" response is sampled as the rank n response, where n is sampled uniformly. Note that the sample for n is seeded such that it is the same across reward models. This process yields a dataset of 8,000 rows, each with a prompt, a chosen response, and a rejected response where both responses are in-distribution for the base model– a requirement for using DPO.

We then train Llama-3.1-8B-Instruct on each dataset using DPO, producing an LLM associated with each selected reward model for real-world downstream human preference testing. Details on the exact DPO configuration can be found in Appendix B.1.

Model	Arena Score	95% CI Lower	95% CI Upper
Meta-Llama-3.1-70B-Instruct*	1228	1218	1238
Athene-RM-70B	**1216**	1206	1226
Athene-RM-8B	1209	1199	1219
InternLM2-7B-Reward	1204	1194	1212
Llama-3-OffsetBias-RM-8B	1200	1191	1209
ArmoRM-Llama3-8B-v0.1	1189	1181	1198
Meta-Llama-3.1-8B-Instruct*	1178	1168	1187
Skywork-Reward-Llama-3.1-8B	1176	1166	1185
Skywork-Reward-Gemma-2-27B	1173	1163	1182
InternLM2-20B-Reward	1173	1163	1182
Nemotron-4-340B-Reward	1172	1163	1180
Meta-Llama-3-8B-Instruct*	1152	1143	1162

Table 3.2: Post DPO performance on Chatbot Arena Overall Category. "Model" is the reward model used to train the base model. Models marked with "*" are baseline, unaltered models. The best non-base model Arena Score is bolded.

Evaluation on Real-World Human Preference

We deploy the trained models to Chatbot Arena [8] to undergo blind evaluation from real users. We set up a cohort of 13 models, which include the trained DPO models as well as Llama-3.1-8B-Instruct, Llama-3.1-70b-Instruct, and Llama-3-8B-Instruct. All models used temperature 0.2 (excluding Llama-3-8B-Instruct at temperature 0.7). Model pairs were sampled evenly, with only each other for battles. Battles were collected over a six day period, from September 10th, 2024, to September 16th, 2024. In all battles, the receiving user was selected randomly. Additionally, the model names (labeled `llama-3.1-8b-dpo-test-{1,2...,9}`) were not revealed to the user until after the vote was given.

Overall, 12,190 human preference votes were collected, with an average of 2,032 battles per model and an average of 190 battles per unique model pair. More details on battle statistics and be found in Table 3.1. The resulting preference rankings are detailed in Table 3.2. The preference rankings are calculated using the Bradley-Terry model, as proposed in [8].

[1]Selected: Athene-RM-70B and Athene-RM-8B, InternLM2-20B-Reward, InternLM2-7B-Reward, Llama-3-OffsetBias-RM-8B, ArmoRM-Llama3-8B-v0.1, Skywork-Reward-Gemma-2-27B, Skywork-Reward-Llama-3.1-8B, Nemotron-4-340B-Reward [12, 7, 31, 41, 23, 44]. Evaluated on Preference Proxy Evaluations (PPE), but not selected: Starling-RM-34B, Starling-RM-7B-Alpha, Eurus-RM-7B, InternLM2-1.8B-Reward, and NaiveVerbosityModel [50, 46, 7].

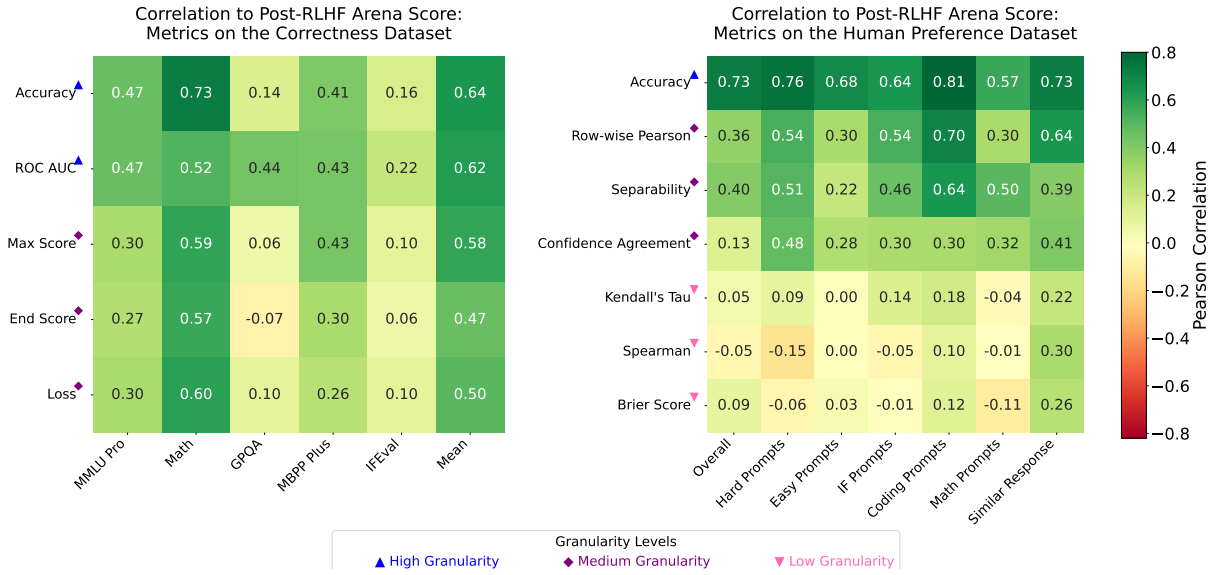

Figure 3.1: Pearson correlations of different metrics toward downstream human preference scores. Left: Pearson correlation between the ranking of models on 5 specific benchmarks and 5 different metrics and their respective post-DPO rankings on real human preference. Right: Pearson correlation between the ranking of models on 7 categories and 7 metrics on the Human Preference Dataset. A similar version using style-controlled human preference as reference is shown in Appendix B.1.

3.2 Studying Correlation with Downstream Performance

In this section, we analyze how different metrics correlate with post-RLHF human preference scores (experimental setup detailed in Section 3.1). Our main results are displayed in Figure 3.1, which shows the correlations of our offline reward model evaluations against the real-world human-preference ranking from the crowdsourced platform.

On correctness metrics (left plot in Figure 3.1) we make several observations:

1. Mean across all domains is well correlated across all metrics, but exhibits higher correlation with AUC and Accuracy scores.

2. Math is the best individual benchmark domain in terms of predictive power.

3. ROC AUC score draws higher correlation across all benchmarks, even on benchmarks that are otherwise uncorrelated.

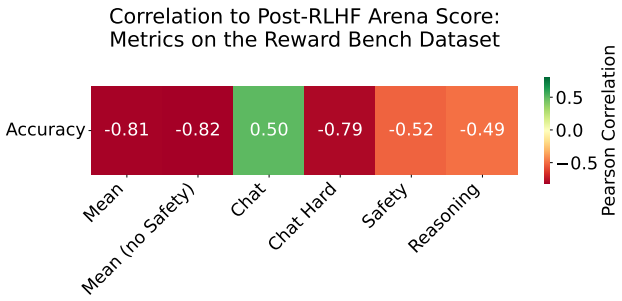

Figure 3.2: Pearson correlation between the ranking of models in RewardBench and their respective post-DPO rankings on real human preference. Style controlled version in Appendix reffig:screward-bench-correlations. Comments on these correlations can be found in Appendix B.2.

Turning to the right-hand side of Figure 3.1, the accuracy of the reward model is the best predictor of the fine-tuned LLM's preference score. Row-wise Pearson Correlation, Confidence Agreement, and Separability show some correlative power to downstream human preference rating but do not exceed accuracy. Meanwhile, metrics like the Spearman correlation and Kendall correlation have nearly zero correlation with the final human preference rating achieved by the post-DPO models. One possible reason for this trend is that accuracy measures expected preference correctness per preference pair— a much more granular scale. Other metrics involve aggregating reward model signals over higher-order preferences, such as preference for each model, as measured by correlation metrics. We consider these metrics as low granularity. Medium granularity metrics, such as Row-wise Pearson Correlation aggregate reward model signal, but do so over smaller subsets of preferences.

Overall, accuracy on the human preference dataset is more correlated with the correctness metrics. This is because correctness and human preference do not necessarily align. Moreover, the information contained in Loss, Max score, and End score may not prove relevant in DPO, which is off-policy. Those employing RLHF algorithms that have a higher risk of over-optimization may find these alternative measures helpful. However, when calculating correlation against style-controlled ratings[2] we notice a slight decrease in correlations on the human preference dataset. Notably, the correctness preference measurements show no change, suggesting correctness preference may be more robust towards reward model preference quality, response style aside. We leave details for Appendix B.3.

Additionally, we observe that measuring the lower bound score may correlate more to downstream RLHF performance than the average score or upper bound score. In Figure 3.3, we first re-scale each category's scores to be mean 0 and SD 1, then we vary the quantile of the aggregation strategy across human preference dataset categories seen in Table 2.2 (Hard Prompts, Easy Prompts, etc). In this case, the 0 quantile is the minimum, and the

[2]Style-controlled ratings are calculated as detailed in [20].

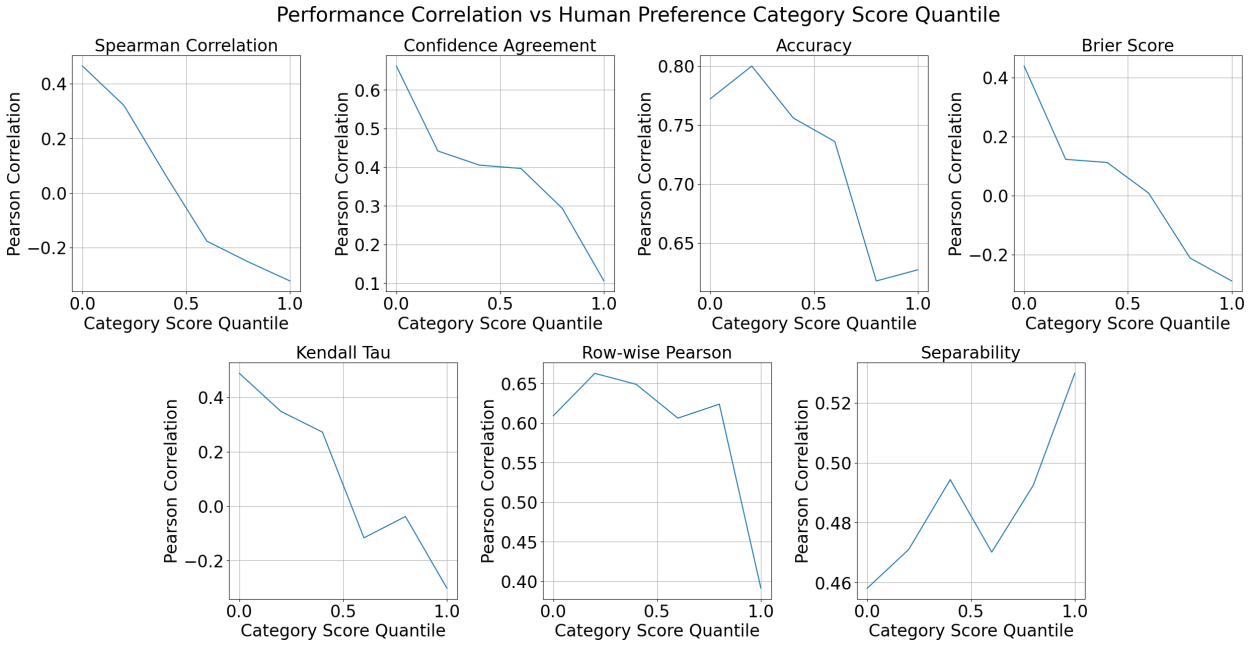

Figure 3.3: The graphs show all metrics for the human preference dataset. For each metric, the six benchmarks (Hard, Easy, Instruction Following, Coding, Math, and Similar Responses Prompts) (all mean and SD normalized) aggregated into the final score by quantile (x-axis). The Pearson Correlation between the aggregated scores is calculated relative to Post-RLHF Human Preference ratings for each aggregation level. Notice that for all metrics except Separability, decreasing quantile increases correlation.

1 quantile is the maximum. We find that in nearly every metric, decreasing the quantile increases correlation with downstream ratings. We posit that the increase in correlation to downstream when using low quantile aggregation across metrics is because this strategy closely measures the robustness of the reward model. This is in line with previous theoretical work that suggests that pessimistic measures on reward model performance may be useful [48, 22]. Intuitively, any single weakness within some input domain could be exploited by the policy model during RL training, thus damaging the model. Another reasonable explanation is that a reward model's weakness in one area may yield noisy signals during training, causing the policy model's rather fragile parameters to be disrupted— a possibly unrecoverable degradation in what we may consider an instance of "catastrophic forgetting". Ultimately, the underlying mechanisms are complex; we do not expect to answer this question in its entirety. However, we believe that our end-to-end experiment provides the first step to understanding how reward model behaviors relate to downstream performance.

Recommendations for PPE based on these findings can be found in Appendix B.5.

3.3 Limitations

Benchmark Leakage

We acknowledge that benchmark leakage is a very real possibility. We also consider two factors that help mitigate this issue: (1) The human preference dataset can be updated with new crowdsourced preference data at any time. This includes adapting to the most recent prompt and response distributions. (2) The correctness preference datasets can be extended to any other benchmark that becomes standard enough to be widely used.

Limits on Testing Downstream Performance

Unfortunately, end-to-end evaluation of reward models via post-RLHF LLM performance on human preference is extremely expensive and time-consuming. As such, we are limited to testing the performance of nine select models, rather than all reward models. In addition, we use DPO, an offline RL algorithm over PPO, an online algorithm, which may play more into over-optimization issues or may have different reward model requirements altogether. We encourage future work to study downstream outcomes under online RL algorithms. Moreover, we note that resource constraints necessitated experimenting with just Llama-3.1-8B-Instruct as the base policy model; additional exploration on a diverse set of base models may yield additional novel insights. With these considerations, we note that the downstream performance measured in our work is in the context of the base model and RLHF learning algorithm used, and is not a unilateral measurement of downstream outcomes in all possible configurations. Future work should experimentally verify the desired reward model behavior of other RLHF configurations.

3.4 Summary

We present PPE, a reward model benchmark explicitly tied to post-RLHF outcomes based on real human preferences. Our experiment aims to identify which metrics, applied to specific tasks, correlate most strongly with downstream performance. We find that across the board, granular measurements, such as accuracy, are the best predictors. Additionally, our results suggest that measuring lower bound performance may be more indicative of expected reward model performance in the RLHF pipeline. Overall, our evaluations achieve a 77% Pearson correlation with downstream performance, significantly improving upon previous work. Based on these results, we encourage future research to further investigate reward model quality and downstream RLHF performance under broader conditions. We fully open-source dataset creation, experimental validation, and reward model evaluation code and methods. We anticipate that the high-quality preference evaluation in PPE, combined with our post-RLHF analysis of metric predictive power, will significantly advance vital research into reward models and RLHF.

Chapter 4

Towards Robust Reward Models

4.1 Background

Two classical statistical approaches to modeling such preferences are the Thurstonian and Bradley-Terry models. However, we find the former is particularly underexplored in RLHF literature. In this chapter, we consider the Thurstonian alternative to reward modeling, and study possible implications towards performance. Armed with our detailed study on PPE in Chapters 2 and 3, we now have to tools to undergo rigorous empirical analysis of reward model performance in addition to intuitive theoretical musings.

Intuitively, human preferences are variable, both across different individuals and within individuals. It is also clear that this variability is not homoskedastic– different contexts may influence both intra-individual and inter-individual preference variation. Consider the user prompt: "What is the best country?". In this case, a highly preferred answer might be the user's own country, in which, under this assumption, we naturally have inter-individual preference variation. Moreover, an individual may be feeling particularly disillusioned with their own country at the moment, leading to a different opinion. Ultimately, a non-answer might be the most robust to ensure there is no strong negative preference: "There's no single "best" country — it really depends on what you're valuing most. Different countries excel in different areas. Here's a breakdown based on various criteria..." The Thurstonian preference model is able to capture these nuances: a response could have a preference score that is high mean but also high variance, which could be riskier than a lower mean but very low variance alternative.

In this chapter, we explore the Thurstonian model on human preference reward models learning from pairwise feedback [1]. We consider both the theoretical intuition of robust regression during training time, as well as pessimistic prediction leveraging variance estimation during test time. On our previously constructed Preference Proxy Evaluations (PPE), we compare these reward models to fixed-variance and Bradley-Terry alternatives, and show performance improvement, particularly robustness on reward modeling tasks that are far-

[1] Our code is available on Github at: efrick2002/highly-rewarding.

ther out-of-distribution with respect to the training data. We find that Thurstonian reward model may be a strong alternative to the Bradley-Terry reward models for RLHF for language model training.

4.2 Preliminaries

In the following sections, we define and compare the Thurstonian model [39] against the Bradley-Terry model [6]. We give intuition towards the training time effects of modeling variance estimation via robust regression. We also introduce potential challenges of accurately estimating variance– particularly from pairwise preference data.

The Thurstonian Model

The Thurstonian model [39] assumes that each alternative i is associated with a latent continuous-valued random variable R_i, typically interpreted as a "reward" or "utility". Given a pair of alternatives (i, j), human preference for i over j is modeled by the probability:

$$P(i \succ j) = P(R_i > R_j) = P(R_i - R_j > 0).$$

The Thurstonian model assumes $R_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ independently for each alternative, leading to:

$$P(i \succ j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution. This Gaussian assumption allows explicit modeling of variance in preferences, capturing uncertainty or variability inherent in human decisions.

The Bradley-Terry Model

The Bradley-Terry model [6] provides a simpler logistic alternative to the Thurstonian approach. It directly models the probability that alternative i is preferred over j using a logistic function parameterized by scalar parameters γ_i, which represent the "strength" or "quality" of each alternative:

$$P(i \succ j) = \frac{\exp(\gamma_i)}{\exp(\gamma_i) + \exp(\gamma_j)}.$$

Unlike Thurstone, Bradley-Terry does not explicitly model uncertainty with variance parameters. Interestingly, the Bradley-Terry model can also be derived from a latent utility model where each alternative i has an associated utility $U_i = \gamma_i + \epsilon_i$, with ϵ_i drawn from a Gumbel distribution. In this case:

$$P(i \succ j) = P(U_i > U_j) = P(\gamma_i + \epsilon_i > \gamma_j + \epsilon_j).$$

Because the difference of two independent Gumbel-distributed variables follows a logistic distribution, this leads exactly to the Bradley-Terry formulation:

$$P(i \succ j) = \frac{\exp(\gamma_i)}{\exp(\gamma_i) + \exp(\gamma_j)}.$$

We see that Bradley-Terry is a special case of the broader class of Random Utility Models [24], where the noise distribution is logistic (Gumbel-difference), as opposed to Gaussian in the Thurstone case. In the Bradley-Terry case, the latent mean and variance are strictly coupled.

The choice between Thurstonian and Bradley-Terry formulations impacts modeling flexibility and interpretability, particularly in the context of learning from human feedback.

Robust Regression

Below, we consider theoretical intuition on the effect of fitting σ^2 estimates during training. Training a neural network to estimate σ^2 has implications for the learned estimates of μ; the gradient on μ is scaled inverse with respect to σ^2 [25].

We explore this in the Thurstonian case. Consider a data point with a chosen response and a rejected response. The chosen response has corresponding mean and variance estimates μ_c and σ_c^2. Likewise, the rejected response is associated with μ_l and σ_l^2. Let $\sigma_{\text{total}}^2 = \sqrt{\sigma_c^2 + \sigma_l^2}$ and $z = \frac{\mu_c - \mu_l}{\sigma_{\text{total}}^2}$. Then the Thurstonian log loss and loss gradient is:

$$\ell\left(\mu_c, \mu_l, \sigma_c^2, \sigma_l^2\right) = -\log\left(\Phi(z)\right) \tag{4.1}$$

$$\frac{\partial \ell}{\partial z} = -\frac{\phi(z)}{\Phi(z)} \tag{4.2}$$

Now consider the gradient on μ_c:

$$\frac{\partial \ell}{\partial \mu_c} = -\frac{\phi(z)}{\Phi(z)} \frac{1}{\sigma_{\text{total}}} = -\frac{\partial \ell}{\partial \mu_l} \tag{4.3}$$

Notice the $\frac{1}{\sigma_{\text{total}}}$. This means the gradient on the mean estimate is dampened by the neural network's own estimated variance when variance is greater than 1. The intuition here is simple: trust labels less when they are high variance. This yields a form of robust regression which may be particularly helpful when learning from inherently noisy human preference data.

We also see this in the gradient of the variance. Let $\Delta\mu = \mu_c - \mu_l)$. Then:

$$\frac{\partial \ell}{\partial \sigma_1^2} = \frac{\partial \ell}{\partial \sigma_2^2} = \frac{\phi(z)}{2\Phi(z)} \frac{\Delta\mu}{\sigma_{\text{total}}^3} \tag{4.4}$$

Therefore, we also see that the variances greater than 1 gradient is dampened by itself, this time by $\frac{1}{\sigma_{\text{total}}^3}$

In both cases, smaller σ_{total} generally magnify the gradient. This is always true when z is negative (and therefore in the wrong direction). If $z < 0$, then the variance term inside z ensures that as $\sigma_{\text{total}} \to 0$ then $z \to -\infty$ causing $\frac{\phi(z)}{\Phi(z)} \to \infty$ of which the gradient is scaled by an additional $\frac{1}{\sigma_{\text{total}}}$ which also goes to infinity. However, when $z > 0$ (the correct direction) $\frac{\phi(z)}{\Phi(z)} \to 0$ as $\sigma_{\text{total}} \to 0$ since $z \to \infty$. In this case, $\frac{\phi(z)}{\Phi(z)} \to 0$ faster than $\frac{1}{\sigma_{\text{total}}} \to \infty$, pushing the mean gradient to 0 for small variances. This does make sense: if the predicted variance is extremely confident, and the mean estimates are in the right direction, then there is no need to move any estimates. Note that in the $z > 0$ case, large σ_{total} still dampen the gradient, trending towards 0.

Heteroscedastic Regression

Modeling human preferences with the Thurstonian model with per-input variance estimates is a form of heteroscedastic regression [32]. When estimating variances in addition to means with neural networks, there are additional challenges [37]. In particular, these challenges arise from the variance term estimate affecting learning of the mean estimate, especially when parameters are shared. As explored in Section 4.2, the variance gradients can effectively change the mean learning rate. While this could be desired, it also could be destructive [36]. We explore different methods to mitigate potential issues from variance estimates in Subsection 4.3 and Section 4.4.

Another consideration is the Thurstonian reward model's unidentifiability. Specifically, the means can be scaled by any constant. Moreover, means and variances can be scaled by any positive constant. We see that the learned relationship between mean and variance can change drastically between models, shown in Figure 4.1.

4.3 Methods

In the following sections, we detail the architectures tested. These architectural changes are necessary for Thurstonian reward modeling, where some parameter separation between mean and variance estimation may be desired. Furthermore, we show how variance can be leveraged during test time to potentially improve predictive robustness.

Architectures

A standard reward model architecture is to start with an instruction-tuned LLM base, remove the language model head. Then, a linear probe is connected to the output of the transformer at the classification token [51]. Generally, both the linear probe and transformer output are full fine-tuned on preference data. Note that a linear probe is the simplest architecture

to adapt the transformer output to the final reward output. Previous work has explored alternatives to a linear probe [42].

We train all models from the Qwen2.5 instruction models, specifically `Qwen2.5-1.5B-Instruct` and `Qwen2.5-7B-Instruct`

When training under the Thurstonian model, the neural network need to output both the mean and log-variance. We find that a simple linear probe architecture generally learns a linear relationship between mean and log-variance (Figure 4.1). Therefore, some properties of more complex probe architectures may be desired; we consider a few alternatives explored below.

Double CLS Token

In an effort to decouple mean and log-variance estimates, we use two CLS Tokens, which can learn different embeddings. One CLS token indicates the hidden dimension for the mean probe, and one indicates the hidden dimension for the log-variance probe. The separate CLS tokens allow for mean and log-variance to have separate attention pools.

Fully-Connected MLP

Following [37], we further separate parameters between mean and log-variance estimates by giving each a fully-connected MLP probe.

The architecture is outlined below. Note that d_{hidden} is the hidden dimension of the transformer.

$$\begin{aligned} \mathbf{h}_{\text{cls}} &= f_{\text{transformer}}(\text{Prompt}, \text{Response}) \\ \mu &= f_\mu\big(\mathbf{h}_{\text{cls}}\big) \\ \log \sigma^2 &= f_\sigma\big(\mathbf{h}_{\text{cls}}\big) \end{aligned}$$

$$f_\mu(\mathbf{x}) = W_{\mu,2}\ \text{SiLU}\big(W_{\mu,1}\,\mathbf{x}\big), \qquad f_\sigma(\mathbf{x}) = W_{\sigma,2}\ \text{SiLU}\big(W_{\sigma,1}\,\mathbf{x}\big),$$

$$W_{\{\mu,\sigma\},1} \in \mathbb{R}^{k * d_{\text{hidden}} \times d_{\text{model}}}, \quad W_{\{\mu,\sigma\},2} \in \mathbb{R}^{1 \times k * d_{\text{hidden}}}.$$

In our experiments, we test $k = 2$ and $k = 4$. In practice, the appropriate k should be tuned as a hyperparameter. The choice of SiLU activations is only to match the base transformer's (`Qwen2.5-{1.5, 7}b-Instruct`) choice of activation.

Fully-Decoupled Mean and Variance

We additionally test an architecture where the mean and variance estimates have no shared parameters. In this case, both the mean and variance are initialized as copies of the base transformer, and both undergo full fine-tuning as separate networks.

$$\texttt{init } f_{\mu\text{-transformer}} \leftarrow f_{\text{transformer}}$$
$$\texttt{init } f_{\sigma\text{-transformer}} \leftarrow f_{\text{transformer}}$$

Then:

$$\begin{aligned}
\mathbf{h}_{\text{cls}}^{(\mu)} &= f_{\mu\text{-transformer}}(\text{Prompt}, \text{Response}) \\
\mathbf{h}_{\text{cls}}^{(\sigma)} &= f_{\sigma\text{-transformer}}(\text{Prompt}, \text{Response}) \\
\mu &= W_\mu \mathbf{h}_{\text{cls}}^{(\mu)} \\
\log \sigma^2 &= W_\sigma \mathbf{h}_{\text{cls}}^{(\sigma)}
\end{aligned}$$

In this setup, we can see that the mean and variance estimates are free to optimize parameters separately. As such, this network is able to learn a fully uncorrelated relationship between mean and variance (Figure 4.1).

Leveraging Variance Estimates for Prediction

Since the Thurstonian reward model outputs both a mean estimate $\mu(\cdot)$ and variance estimate $\sigma^2(\cdot)$, which parameterize a normal distribution, we can leverage the quantile value instead of purely the mean value. Given some input x, let $\mu(x)$ and $\sigma^2(x)$ be the estimate from the reward model. A perfectly neutral perspective would ignore variance and take only $\mu(x)$ as the reward, ignoring $\sigma^2(x)$ entirely. However, a pessimistic perspective might prefer a lower variance response, conversely, and an optimistic perspective may prefer a higher variance response. More formally, we define the quantile q informed reward $R_q(x) = \mu(x) + \Phi^{-1}(q)\sigma^2(x)$ where $\Phi^{-1}(\cdot)$ denotes the inverse normal CDF. When $q > 0.5$ we are optimistic with respect to variance, and when $q < 0.5$, we are pessimistic. Of course, with $q = 0$ we recover the mean-only estimate.

Choosing a pessimistic reward has some reasonable philosophical intuitions. In particular, we can consider that the Thurstonian reward model may be able to estimate some sort of epistemic uncertainty [17]. This uncertainty can arise from either the model's own lack of training examples or incomplete information on the preferences of a given human (or non-human) label. In Thurstonian reward modeling, the model defines a normal distribution defining the rewards from possibly many individuals, possibly each individual labeling multiple times. If we take a quantile reward that is pessimistic, we are selecting the response that appeals to more voters. For example, taking the 5% quantile marks the reward in which 95% of ratings agree that the response is scored that or higher. Intuitively, by being pessimistic, we are more likely to capture the *preferences we want*– such as correctness signals that are often more strict than vibe-based preferences.

4.4 Experiments

In the following sections below, we cover the data, training, and evaluations for all reward models. In Section 4.4 we detail all reward model types trained.

Data

We train all reward models on crowdsourced pairwise preference data from Chatbot Arena [9]. On Chatbot Arena, users can query randomly selected two anonymous LLMs and receive side-by-side answer from each model. The user then picks which answer they like better. In addition, users can pick "tie" or "tie (both bad)" signaling both responses are too similar in quality, or both un-judgeable. We begin with roughly 2.5 million pairwise comparisons from Chatbot Arena collected from April 4th, 2024, up to March 13th, 2025. A simple filtering pipeline is used to obtain the final dataset. First, we remove likely spammers using a binary hypothesis test. Since users pick from randomized anonymous models, we know they should pick the model shown on the left and the model shown on the right with equal probability. Therefore, we use a standard binomial test to validate that for each user $P(\text{User picks left} \mid \text{No Tie}) = P(\text{User picks right} \mid \text{No Tie})$. All users failing this test are discarded. We also check a one-sided hypothesis test confirming that the tie rate of each user is not excessive: e.g. that $P(\text{User picks Tie or Tie (both bad)}) < 0.80$. For all tests, we use a p-value of 0.05. After this procedure, we remove all tied battles (including tie (both bad)) since these are incompatible with the Bradley-Terry and Thurstonian Models. The resulting dataset contains 1,403,058 pairwise comparisons, collected between 235 different models, and from 514,944 unique users.

Training

We train all models on the same Chatbot Arena dataset, detailed in Section 4.4. The reward models are trained for 1 epoch, to avoid overfitting to inherently noisy human preference labels– as is standard. Training is done on the same base model: `Qwen2.5-1.5b-Instruct`. In all training runs we use a learning rate of $2 \times 10^{-}6$ and a batch size of 512. Training runs for 2740 steps in total.

Since the Thurstonian reward model outputs both a mean and variance estimate, we would like to understand how the degree of parameter sharing affects performance. In early experiments with just a linear probe, we observed a high negative correlation between mean and variance. To understand how architecture changes affect the final trained reward model performance, we train models with the following architectures: linear probe, MLP ($k = 2$), MLP ($k = 4$), Double CLS, Double CLS w/ MLP ($k = 4$), and fully decoupled. More details on the architectures can be found in Section 4.3. Bradley-Terry models are trained with a linear probe, as architectures to decouple mean and variance estimates are not applicable.

In addition, we try three other modifications. Following [37]'s first proposal, we train a Thurstonian reward model where the transformer hidden dimension output is detached

from gradients before entering the variance MLP head. As such, the variance MLP head is unable to propagate gradients into the transformer. Following [37]'s second proposal, we train a Thurstonian reward model where the gradient on the mean is scaled by $\frac{1}{\sigma^2}$, thereby correcting for the variance's effect on the gradient. Finally, we train a Thurstonian reward model with a constant variance estimate, forcing the model to represent all variance with differences in means. We use a constant variance of 1.

Evaluation

We evaluate all models on Preference Proxy Evaluations (PPE), covered in Chapters 2 and 3. Recall that PPE contains six main test sets, the first of which, `human_preference_v1` is a holdout set derived from Chatbot Arena– in-distribution (but not contained within) our training data. The other five test sets are derived from LLM benchmarks with verifiable correctness checks and compare the reward model's accuracy against known verifiers. These verifiable tasks come from MMLU Pro [43], MATH [16], GPQA [34], MBPP Plus [3], and IFEval [47].

We also construct a custom human preference test set, `human_preference_ood`, to test reward model robustness on out-of-distribution future data. Recall that the reward model training contains preferences collected up until March 13th, 2025. To construct `human_preference_ood`, we utilize preferences collected strictly after March 13th, 2025, up until April 20th, 2025. To further make the task out-of-distribution, we filter out pairwise comparisons in which either of the participating models is seen in the training data. After filtering, this test set has 4,836 examples. By including a disjoin set of models, we can measure how well the reward model generalizes to new policy models; this is essential as during online reinforcement learning, for example, PPO [35], the policy model is certain to change, and thus the reward model must be robust to this distribution shift.

On all sets, we measure accuracy by calculating the rate at which the reward model selects the preferred response over the dispreferred response. On the verifiable correctness sets, the preferred response is some response that has been verified as correct, and the dispreferred response is some response that has been verified as incorrect.

In line with PPE, we also consider measurements of reward model robustness (see 2.4. Since each prompt in the verifiable correctness sets is associated with 32 sampled responses, each labeled correct and incorrect, we can measure the reward model's best-of-32 score defined as follows: let S be a size 32 random sample of responses from a model and $g : S \to \{0, 1\}$ be the ground truth scoring function, and $\hat{R} : s \in S \to \mathbb{R}$ is the output score of the reward model. Then, the best-of-32 score is $\mathbb{E}_S \left[g \left(\arg\max_{s \in S} \hat{R}(s) \right) \right]$.

4.5 Results

In the following subsections, we detail the performance comparisons between different Thurstone reward model variants as well as Bradley-Terry reward models. We first look at overall

Model Size	Type	Architecture	Other	Human Pref.	OOD Human Pref.	MMLU Pro	MATH	MBPP+	IFEval	GPQA	Mean
1.5B	Thurs	2 CLS	N/A	69.391	66.439	63.828	**70.781**	65.128	**56.758**	56.602	**64.132**
1.5B	Thurs	2 CLS, MLP	N/A	69.518	66.501	63.984	69.648	**67.732**	55.000	56.367	64.107
1.5B	Thurs	MLP	Detached Var	**69.616**	66.667	63.516	70.273	59.487	54.453	56.602	62.945
1.5B	Thurs	MLP	N/A	69.372	66.522	63.945	69.414	58.935	54.922	**57.227**	62.905
1.5B	Thurs	$\frac{1}{2}$ MLP	N/A	69.548	66.729	**65.586**	69.180	59.250	53.906	56.055	62.893
1.5B	Thurs	Linear Probe	N/A	69.518	66.770	65.195	69.609	57.239	54.062	56.602	62.714
1.5B	BT	Linear Probe	N/A	**69.616**	66.543	64.062	70.391	51.953	55.977	56.523	62.152
1.5B	Thurs	MLP	$\nabla\mu$ Scaled	68.217	65.757	62.148	67.734	61.460	52.578	55.547	61.920
1.5B	Thurs	Decoupled	N/A	68.618	**66.998**	61.562	67.578	43.471	55.312	54.453	59.713
1.5B	Thurs	Linear Probe	Fixed Var	64.575	55.335	61.563	56.836	66.824	49.844	52.617	58.228

Table 4.1: Accuracies of all trained 1.5B reward models on PPE benchmarks. The models are sorted by their mean score across all benchmarks. The scores are in percentages. Note that MLP is MLP with $k = 4$. $\frac{1}{2}$ MLP is MLP with $k = 2$. "Detached Var", "$\nabla\mu$ Scaled", and "Fixed Var" denote the last three training methods detailed in Subsection 4.4. Results on 7B parameters can be found in Appendix Table C.1.

performance, defined by average performance on PPE subsets. Additionally, we consider performance on the custom-curated out-of-distribution human preference subset. Finally, we also consider best-of-32 performance as a measure of robustness, as well as how Thurstonian reward model robustness changes with respect to quantile reward.

Overall Performance

In Table 4.1, we find that the Thurstonian models outperform the Bradley-Terry models overall, however, the margins are thin. Notably, when removing the effect variance estimation on learning, whether through using Bradley-Terry, scaling the μ gradient, or fixing the variance, the performance drops. This may suggest that the robust regression effect detailed in Section 4.2 may be helpful for learning from human preference data. Some mean and variance decoupling strategies seem to help, but fully decoupling with no shared parameters is destructive. The best performing architecture, by a comparatively large margin, is the double CLS reward models. It is also notable that in the distribution of human preference sets, the performance across all models is very similar. The largest performance gaps are seen on the MBPP Plus test set, where the gap between the best Thurstonian model and the Bradley Terry model is 14%. The Thurstonian model may be more robust to the distribution shift between the human preference data and verifiable correctness measures, like in MBPP Plus. Appendix C.1 shows that the Thurstonian reward model is able to outperform the Bradley-Terry reward model overall when parameter count is scaled to 7B– although the differences are more marginal. This is attributed to noticeable overfitting in all reward models trained on this size of the training set, regardless of the underlying choice model. More 7B model variants are not trained due to computational limitations.

Table 4.2 shows reward model accuracies on the out-of-distribution human preference data. We find again that most Thurstonian models outperform the Bradley-Terry variances,

Model Size	Type	Architecture	Other	Overall	Hard Prompt	Easy Prompt	If Prompt	Code Prompt	Math Prompt	Mean
1.5B	Thurs	Decoupled	N/A	**66.998**	68.081	**64.227**	**68.928**	67.030	**66.346**	**66.935**
1.5B	Thurs	MLP	N/A	66.522	68.523	63.402	68.580	67.129	65.769	66.654
1.5B	Thurs	Linear Probe	N/A	66.770	**68.877**	62.577	68.406	66.634	65.577	66.473
1.5B	Thurs	2 CLS, MLP	N/A	66.501	68.435	62.887	68.464	66.634	65.577	66.416
1.5B	Thurs	$\frac{1}{2}$ MLP	N/A	66.729	68.789	62.474	68.870	66.238	65.385	66.414
1.5B	Thurs	2 CLS	N/A	66.439	68.700	62.268	68.638	66.832	65.385	66.377
1.5B	Thurs	MLP	Detached Var	66.667	68.612	62.474	67.884	**67.327**	65.192	66.359
1.5B	BT	Linear Probe	N/A	66.543	68.700	62.268	68.580	66.931	65.000	66.337
1.5B	Thurs	MLP	$\nabla\mu$ Scaled	65.757	68.081	59.691	67.014	66.634	63.846	65.171
1.5B	Thurs	Linear Probe	Fixed Var	55.335	49.779	58.969	53.275	47.624	55.962	53.491

Table 4.2: Accuracies of all trained 1.5B reward models on the OOD human preference test set. The categories are derived from Chatbot Arena's category definitions [9]. The models are sorted by their mean score across all categories. Results on 7B parameters can be found in Appendix Table C.2.

Model Size	Type	Architecture	Other	MMLU Pro	MATH	MBPP+	IFEval	GPQA	Mean
1.5B	Thurs	2 CLS, MLP	N/A	60.730	48.560	**73.209**	56.609	46.327	**57.087**
1.5B	Thurs	2 CLS	N/A	60.198	**49.482**	69.748	57.687	45.853	56.594
1.5B	Thurs	MLP	Detached Var	60.101	48.952	66.900	56.434	**46.741**	55.826
1.5B	Thurs	MLP	N/A	**61.375**	49.195	63.721	57.400	46.725	55.683
1.5B	Thurs	MLP	$\nabla\mu$ Scaled	59.965	47.733	67.050	55.761	45.627	55.227
1.5B	Thurs	Linear Probe	N/A	61.274	48.223	62.414	56.766	46.023	54.940
1.5B	Thurs	$\frac{1}{2}$ MLP	N/A	60.882	48.223	64.025	56.118	45.103	54.870
1.5B	BT	Linear Probe	N/A	60.590	49.029	58.295	**57.704**	45.524	54.228
1.5B	Thurs	Linear Probe	Fixed Var	56.835	38.525	72.643	52.983	44.540	53.105
1.5B	Thurs	Decoupled	N/A	58.099	44.895	56.483	56.541	44.296	52.063

Table 4.3: Average Best-of-32 score of all trained 1.5B reward models on the PPE verifiable benchmark sets. The models are sorted by their mean score across all benchmarks. Results on 7B parameters can be found in Appendix Table C.3.

though it should be noted that the accuracies are very similar. The worst-performing model by far is the fixed variance Thurstone model– again suggesting variance estimation could be beneficial. It should also be noted that there we no significant differences between fixed on non-fixed variance Thurstonian models in terms of training loss. The observed differences arose during testing only, further putting into question the robustness of the fixed-variance Thurstonian reward model. Interestingly, the fully decoupled Thurstonian model performs well on this out-of-distribution test set, possibly because it contains $2\times$ the parameters. Additionally, we show test loss curves on both in and out-of-distribution human preference for each training step checkpoint in Appendix Section C.1.

Finally, we look at reward model robustness on verifiable correctness, shown in Table 4.3. The best-of-32 metric is defined in Subsection 4.4. Here, we see the largest gap between Bradley-Terry models and the double CLS Thurstonian models. In particular, the double

CLS w/ MLP head Thurstonian reward model is most robust on verifiable tasks.

Learned Mean and Variance Relationships

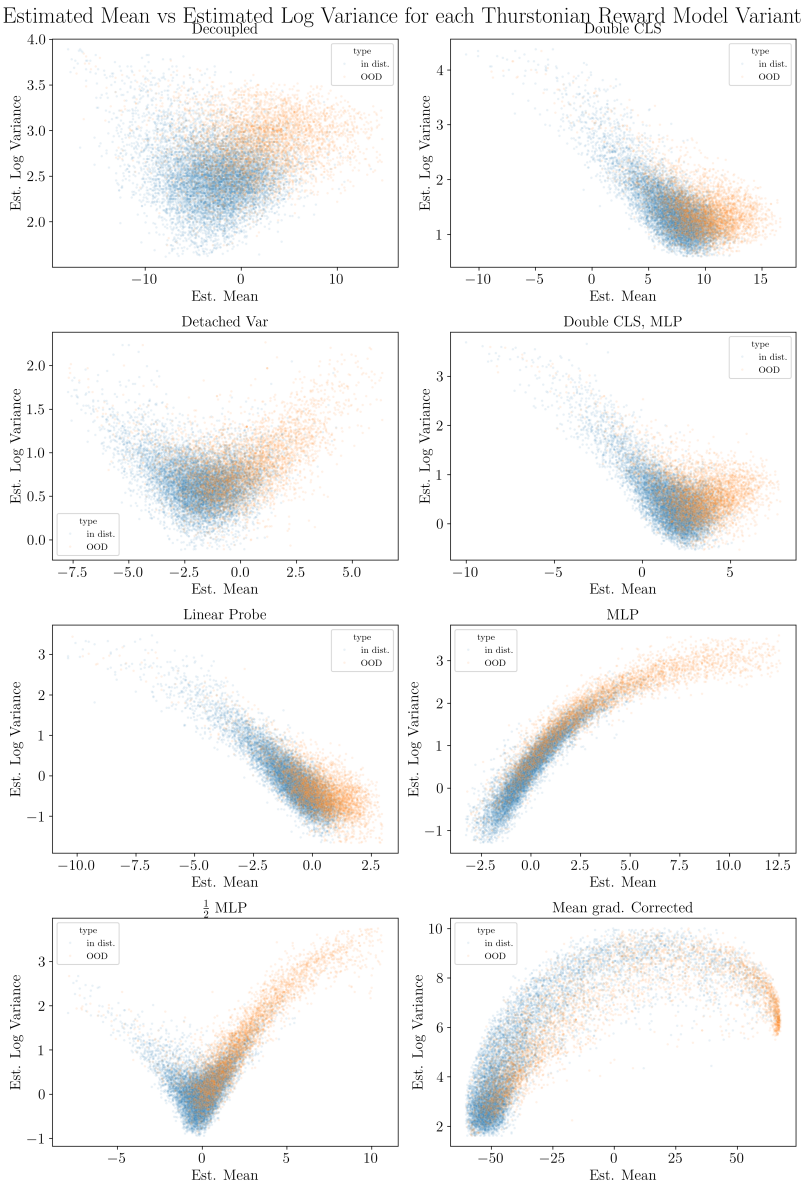

Figure 4.1: Scatter plots of the learned relationship between mean and log-variance for each reward model on human preference. In distribution human preference datapoints are shown in blue, and out-of-distribution datapoints are shown in orange.

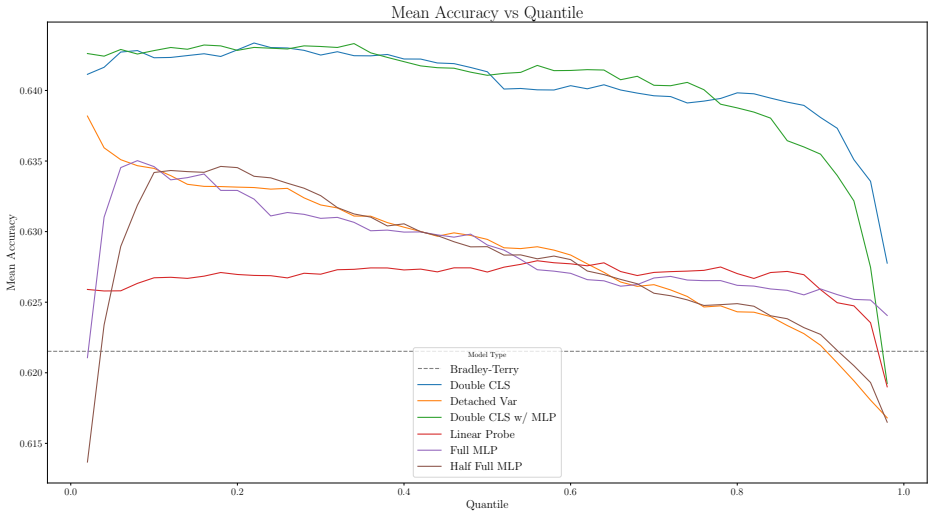

Figure 4.2: Mean reward model accuracy vs. selected reward quantile. The quantiled reward is given by $R_q(x) = \mu(x) + \Phi^{-1}(q)\sigma^2(x)$. The Bradley-Terry baseline is shown in gray. A similar figure on 7B models can be found in Figure 4.3.

Quantile Rewards

All the above results relied only on mean estimates, this means performance gains must come from choice model specification and training dynamics, such as robust regression. However, the Thurstonian reward models output a variance that can be leveraged during prediction. We detail these observations below.

First, in Figure 4.2, we show that more pessimistic quantiles increase overall model accuracies for most models, suggesting we can leverage variance estimates for improved, or more robust, prediction at test-time. Of course, both overly optimistic and overly pessimistic quantiles risk damaging estimates. We show this pattern extends to 7 billion parameters in Figure 4.3.

Moreover, we find that using pessimistic quantile rewards has a significant benefit towards reward model robustness on PPE's verifiable best-of-k metric. In Figure 4.4, we see that decreasing the quantile roughly monotonically increases the long-run robustness of the reward model as the number of response choices increases. Again, this pattern is also found when scaling to 7 billion parameters, as shown in Figure 4.5, albeit dampened.

Finally, in Figure 4.6, we see how even early in training, leveraging variance estimates can increase reward model performance. In particular, earlier checkpoints (but not too early) may benefit even more from pessimistic quantiles. Additionally, we have a slightly strong positive effect from pessimistic quantiles on best-of-32 scores compared to accuracy metrics. We find this pattern extends to nearly every Thurstone reward model trained, including 7B parameter variants, shown in Appendix Section C.3.

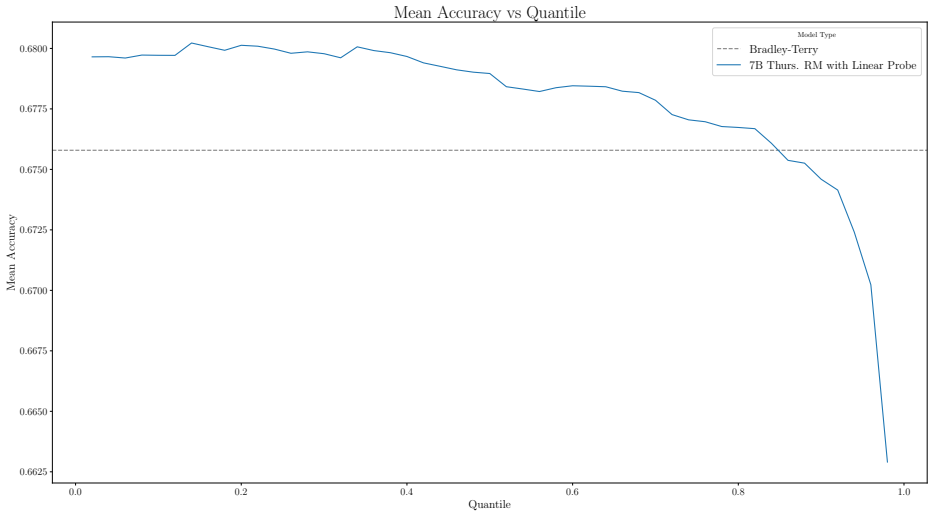

Figure 4.3: Mean reward model accuracy vs. selected reward quantile on a 7B Thurstone reward model. The quantiled reward is given by $R_q(x) = \mu(x) + \Phi^{-1}(q)\sigma^2(x)$. The 7B Bradley-Terry baseline is shown in gray.

Additionally, Figure 4.7 shows how even in the human preference case, as the test distribution is shifted away from the training distribution, the optimal quantile shifts back. On the in-distribution test set, optimistic quantiles are optimal, around a quantile of 0.8. However, on the out-of-distribution human preference test set, we find that a quantile of 0.25 is more optimal. More plots for all trained models showing generally similar trends can be found in Appendix Section C.3. It also appears, considering the mean and variance scatter plots in Figure 4.1, that some Thurstonian reward model variants show increased variance on out-of-distribution inputs when compared to in-distribution inputs– this effect is largest when the mean and variance networks are fully decoupled, and therefore share no parameters. This may suggest that these particular Thurstonian reward models are able to represent some level of epistemic uncertainty, though the effect is not strong.

Importantly, we observe the same patterns with respect to pessimistic quantiles regardless of the learned mean vs. variance relationship. Figure 4.1 shows how different architectures yielded different learned relationships between mean and variance, while still achieving similar training loss. However, despite this, the pessimistic quantiles still appear to increase accuracy and robustness– this is even when the mean and variance appear to be positively correlated. This suggests that the quantile procedure is not merely utilizing mean information encoded in the variance estimate.

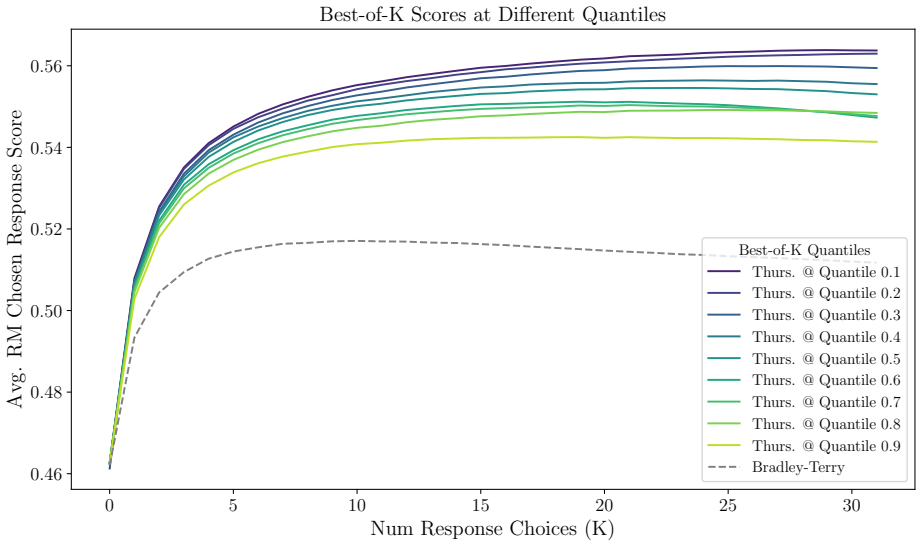

Figure 4.4: A reward model best-of-k curve, showing the reward model's average chosen response score as the number of choices increases. The Thurstonian reward model in the figure uses the double CLS and MLP head architecture. A similar figure on 7B models can be found in Figure 4.5.

4.6 Summary

We show the Thurstone model to be a viable alternative to Bradley-Terry when learning to model human preferences. In particular, the robust regression characteristics of learning to estimate variance may help the reward model's adaptation to distribution shift seen in testing (or real deployment). Additionally, the variance estimate can be leveraged at test time to tune quantile rewards, thereby further improving reward model performance and robustness. In general, these results suggest that Thurstonian reward models, combined with pessimistic quantile reward, may serve as more robust human preference proxies during online RLHF procedures– these procedures induce considerable distribution on the reward model and thus can collapse towards over-optimized policies. We also encourage future work to consider algorithms that leverage the variance estimation produced by Thurstone reward models to further increase the robustness of training.

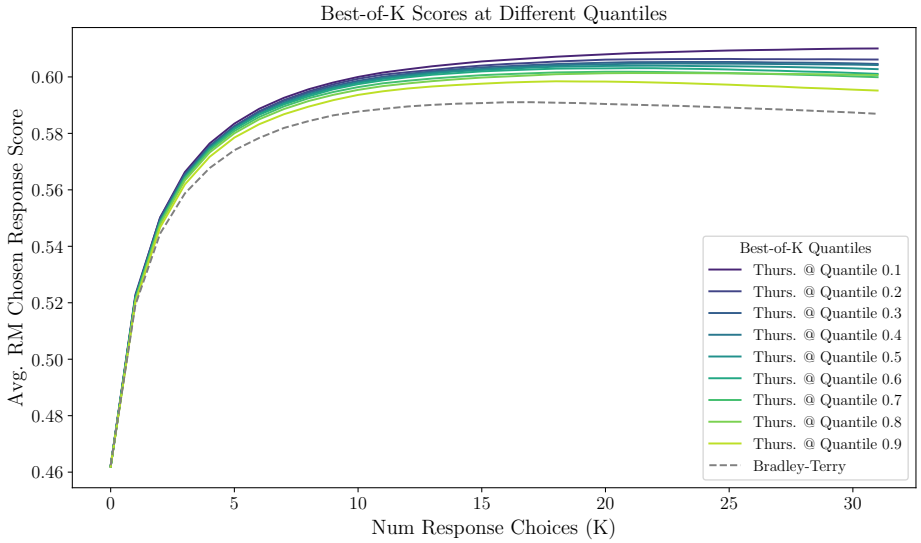

Figure 4.5: A reward model best-of-k curve, showing the reward model's average chosen response score as the number of choices increases. The Thurstonian reward model in the figure has 7B parameters with a linear probe head.

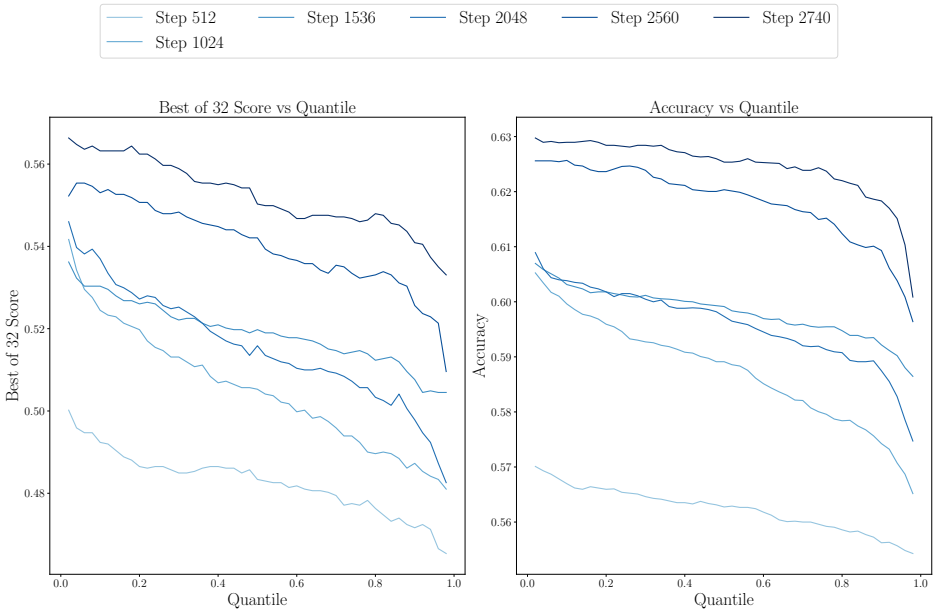

Figure 4.6: Best of 32 score vs Quantile and Accuracy vs Quantile, shown for each checkpoint step during training. The Thurstonian reward model in the figure uses the double CLS and MLP head architecture.

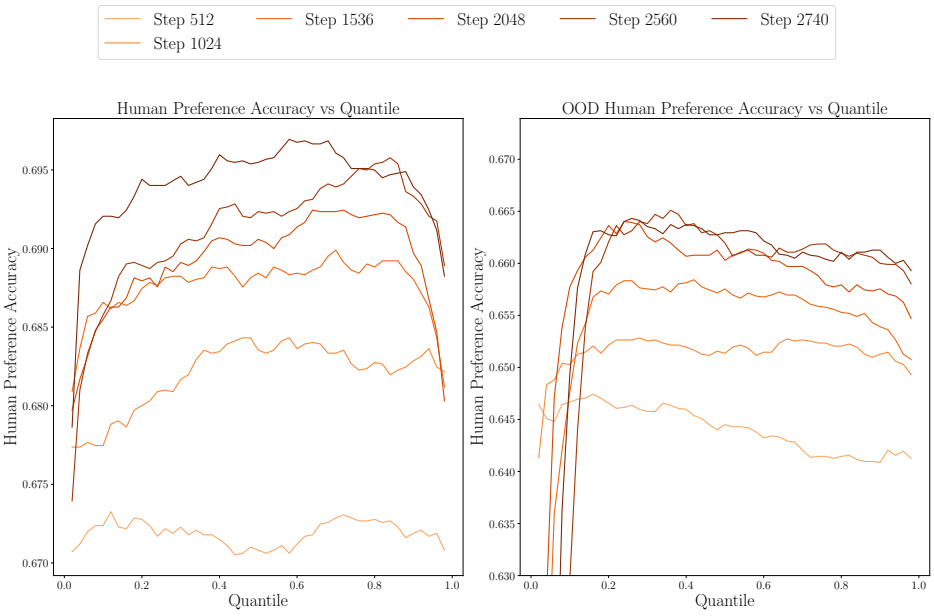

Figure 4.7: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure uses the MLP head architecture.

Chapter 5

Conclusion

5.1 Conclusion and Future Directions

The alignment of Large Language Models with human preferences remains a central challenge in the pursuit of safe and beneficial artificial intelligence. This thesis embarked on a journey to address critical aspects of this challenge, focusing on the evaluation and enhancement of reward models, the linchpin of RLHF. Our primary objectives were twofold: first, to develop a more reliable and predictive benchmark for reward model performance, and second, to leverage this improved evaluation framework to explore methodologies for creating more robust reward models.

Summary of Contributions and Key Findings

We began by confronting the prevalent disconnect between existing reward model evaluation tasks and their true utility in driving downstream LLM performance. To bridge this gap, we introduced Preference Proxy Evaluations (PPE), a comprehensive benchmarking suite detailed in Chapter 2. PPE distinguishes itself by incorporating diverse datasets, including large-scale, real-world human preference data and verifiable correctness tasks across multiple domains. This multi-faceted approach allows for a more holistic assessment of an reward models's ability to capture human intent.

Crucially, as presented in Chapter 3, we empirically validated PPE by conducting end-to-end experiments. We demonstrated a significant and robust correlation—achieving a 77% Pearson correlation in our primary setup—between reward model performance on PPE metrics (particularly granular accuracy on human preference and MATH correctness) and the actual downstream performance of LLMs fine-tuned using these reward models via Direct Preference Optimization (DPO). This validation provides the community with a more trustworthy proxy for reward model efficacy, enabling faster and more cost-effective iteration on reward model development without necessitating full, resource-intensive RLHF pipelines for every evaluation. Our analysis also highlighted that pessimistic, lower-bound aggregations of

reward model scores often yield stronger correlations with downstream outcomes, suggesting the importance of reward model robustness across all evaluated domains.

Armed with the validated PPE framework, Chapter 4 delved into enhancing reward model robustness by exploring Thurstonian models and heteroskedastic regression for human preferences. We posited that explicitly modeling the inherent variance in human judgments could lead to reward models that are more resilient to noisy data and distributional shifts. Our experiments demonstrated that Thurstonian reward models, particularly when leveraging pessimistic quantile rewards at test time, can outperform traditional Bradley-Terry models and fixed-variance Thurstonian alternatives. These models showed notable improvements in robustness on out-of-distribution tasks, indicating their potential to serve as more reliable proxies during the dynamic RLHF process. The observed benefits of robust regression characteristics during training and variance-aware prediction further underscore the value of this approach.

Broader Implications

The findings of this thesis have several important implications for the field of LLM alignment. Firstly, the development and validation of PPE offer a more principled and empirically grounded methodology for reward model evaluation. This can accelerate research by providing a faster feedback loop and fostering a more standardized approach to comparing different reward model architectures and training techniques. Secondly, our exploration of Thurstonian models highlights the potential benefits of different types of reward model frameworks. In the Thurstonian case, by acknowledging and modeling the variance in human preferences, we can develop reward models that are not only more accurate on average but also more robust to the complexities and uncertainties inherent in human feedback. This increased robustness is critical for ensuring stable and reliable alignment, especially as LLMs are deployed in increasingly high-stakes applications.

Concluding Remarks

The role of human preferences in training and evaluating generalist artificial intelligence is here to stay. Learning to robustly model human preferences is essential to the success of future generalist artificial intelligence deployed into our society and systems. We hope this work steps us towards this direction.

Bibliography

[1] AI@Meta. "Llama 3 Model Card". In: (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[2] Anthropic. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. (Accessed on 06/05/2024). 2024.

[3] Jacob Austin et al. *Program Synthesis with Large Language Models*. 2021. arXiv: 2108.07732 [cs.PL]. URL: https://arxiv.org/abs/2108.07732.

[4] Yuntao Bai et al. "Constitutional AI: Harmlessness from AI Feedback". In: (2022). arXiv: 2212.08073 [cs.CL].

[5] Yuntao Bai et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback". In: *arXiv preprint arXiv:2204.05862* (2022).

[6] Ralph Allan Bradley and Milton E. Terry. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3/4 (1952), pp. 324–345. ISSN: 00063444, 14643510. URL: http://www.jstor.org/stable/2334029 (visited on 05/09/2025).

[7] Zheng Cai et al. "Internlm2 technical report". In: *arXiv preprint arXiv:2403.17297* (2024).

[8] Wei-Lin Chiang et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. arXiv: 2403.04132 [cs.AI].

[9] Wei-Lin Chiang et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. arXiv: 2403.04132 [cs.AI]. URL: https://arxiv.org/abs/2403.04132.

[10] Paul Christiano et al. "Deep reinforcement learning from human preferences". In: (2023). arXiv: 1706.03741 [stat.ML].

[11] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. "Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 5988–6008. URL: https://proceedings.mlr.press/v162/ethayarajh22a.html.

[12] Evan Frick et al. *Athene-70B: Redefining the Boundaries of Post-Training for Open Models*. July 2024. URL: https://huggingface.co/Nexusflow/Athene-70B.

[13] Evan Frick et al. *How to Evaluate Reward Models for RLHF*. 2024. arXiv: 2410.14872 [cs.LG]. URL: https://arxiv.org/abs/2410.14872.

[14] Leo Gao, John Schulman, and Jacob Hilton. "Scaling Laws for Reward Model Overoptimization". In: *arXiv preprint arXiv:2210.10760* (2022).

[15] Leo Gao et al. *A framework for few-shot language model evaluation*. Version v0.0.1. Sept. 2021. DOI: 10.5281/zenodo.5371628. URL: https://doi.org/10.5281/zenodo.5371628.

[16] Dan Hendrycks et al. "Measuring Mathematical Problem Solving With the MATH Dataset". In: *NeurIPS* (2021).

[17] Stephen C Hora. "Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management". In: *Reliability Engineering & System Safety* 54.2-3 (1996), pp. 217–223.

[18] Nathan Lambert et al. *RewardBench: Evaluating Reward Models for Language Modeling*. https://huggingface.co/spaces/allenai/reward-bench. 2024.

[19] Harrison Lee et al. "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback". In: (2023). arXiv: 2309.00267 [cs.CL].

[20] Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. *Does Style Matter? Disentangling style and substance in Chatbot Arena*. Aug. 2024. URL: https://blog.lmarena.ai/blog/2024/style-control/.

[21] Tianle Li et al. *From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline*. 2024. arXiv: 2406.11939 [cs.LG]. URL: https://arxiv.org/abs/2406.11939.

[22] Zihao Li, Zhuoran Yang, and Mengdi Wang. *Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism*. 2023. arXiv: 2305.18438 [cs.LG]. URL: https://arxiv.org/abs/2305.18438.

[23] Chris Yuhao Liu and Liang Zeng. *Skywork Reward Model Series*. https://huggingface.co/Skywork. Sept. 2024. URL: https://huggingface.co/Skywork.

[24] Daniel McFadden. "Conditional logit analysis of qualitative choice behavior". In: (1972).

[25] D.A. Nix and A.S. Weigend. "Estimating the mean and variance of the target probability distribution". In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 1. 1994, 55–60 vol.1. DOI: 10.1109/ICNN.1994.374138.

[26] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[27] OpenAI. *GPT-4o mini: advancing cost-efficient intelligence*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. (Accessed on 06/05/2024). 2024.

[28] OpenAI. *Introducing ChatGPT*. https://openai.com/blog/chatgpt. (Accessed on 01/12/2024). 2022.

[29] OpenAI. *New models and developer products announced at DevDay*. https://openai.com/blog/new-models-and-developer-products-announced-at-devday. (Accessed on 01/12/2024). 2023.

[30] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[31] Junsoo Park et al. *OffsetBias: Leveraging Debiased Data for Tuning Evaluators*. 2024. arXiv: 2407.06551 [cs.CL].

[32] Rolla E Park. "Estimation with heteroscedastic error terms." In: *Econometrica* 34.4 (1966).

[33] Rafael Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *arXiv preprint arXiv:2305.18290* (2023).

[34] David Rein et al. *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*. 2023. arXiv: 2311.12022 [cs.AI].

[35] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[36] Maximilian Seitzer et al. *On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks*. 2022. arXiv: 2203.09168 [cs.LG]. URL: https://arxiv.org/abs/2203.09168.

[37] Andrew Stirn et al. "Faithful heteroscedastic regression with neural networks". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 5593–5613.

[38] Gemma Team et al. "Gemma 2: Improving open language models at a practical size". In: *arXiv preprint arXiv:2408.00118* (2024).

[39] Louis L Thurstone. "A law of comparative judgment." In: *Psychological review* 101.2 (1994), p. 266.

[40] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[41] Haoxiang Wang et al. "Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts". In: *EMNLP*. 2024.

[42] Haoxiang Wang et al. *Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts*. 2024. arXiv: 2406.12845 [cs.LG]. URL: https://arxiv.org/abs/2406.12845.

[43] Yubo Wang et al. "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark". In: *arXiv preprint arXiv:2406.01574* (2024).

[44] Zhilin Wang et al. *HelpSteer2: Open-source dataset for training top-performing reward models*. 2024. arXiv: 2406.08673.

[45] Leandro von Werra et al. *TRL: Transformer Reinforcement Learning*. https://github.com/huggingface/trl. 2020.

[46] Lifan Yuan et al. *Advancing LLM Reasoning Generalists with Preference Trees*. 2024. arXiv: 2404.02078.

[47] Jeffrey Zhou et al. "Instruction-following evaluation for large language models". In: *arXiv preprint arXiv:2311.07911* (2023).

[48] Banghua Zhu, Jiantao Jiao, and Michael I Jordan. "Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons". In: *arXiv preprint arXiv:2301.11270* (2023).

[49] Banghua Zhu et al. "Starling-7b: Improving helpfulness and harmlessness with rlaif". In: *First Conference on Language Modeling*. 2024.

[50] Banghua Zhu et al. *Starling-7B: Improving LLM Helpfulness & Harmlessness with RLAIF*. Nov. 2023.

[51] Daniel M. Ziegler et al. *Fine-Tuning Language Models from Human Preferences*. 2020. arXiv: 1909.08593 [cs.CL]. URL: https://arxiv.org/abs/1909.08593.

Appendix A

Appendix for Evaluating Reward Models for RLHF

A.1 Detailed Scores for the Human Preference Evaluation Dataset

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	69.46	67.05	74.21	96.88	83.16	94.44	0.06
Claude-3-5-Sonnet-20240620 (ArenaHard)†	69.25	67.96	72.11	97.92	86.32	95.49	0.06
GPT-4o-2024-08-06 (ArenaHard)†	68.50	68.17	71.05	97.92	85.26	95.94	0.06
Ensemble-Judges (AlpacaEval)†	68.32	66.01	75.26	96.88	83.16	94.59	0.07
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.63	63.55	71.05	95.83	82.11	94.29	0.08
Gemini-1.5-Pro-002 (AlpacaEval)†	66.53	66.85	72.63	96.88	84.21	95.49	0.06
Athene-RM-70B	66.43	67.01	76.84	96.88	78.95	92.93	0.08
GPT-4o-2024-08-06 (AlpacaEval)†	66.30	62.68	69.47	96.88	78.95	93.23	0.09
Gemini-1.5-Pro-002 (ArenaHard)†	65.70	68.57	68.42	95.83	83.16	94.44	0.07
Llama-3.1-70B-Instruct (AlpacaEval)†	64.96	65.76	65.26	90.62	70.53	87.82	0.11
Llama-3.1-70B-Instruct (ArenaHard)†	64.74	60.00	64.21	89.58	73.68	89.02	0.10
Athene-RM-8B	64.41	62.44	74.21	96.88	74.74	87.97	0.11
Gemini-1.5-Flash-002 (AlpacaEval)†	64.35	62.30	65.79	94.79	77.89	91.43	0.09
Gemini-1.5-Flash-002 (ArenaHard)†	64.18	60.68	67.37	94.79	81.05	92.18	0.08
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	64.14	56.81	65.26	90.62	73.68	88.42	0.11
Starling-RM-34B	63.87	59.33	71.58	89.58	65.26	82.41	0.14
Gemini-1.5-Pro-001 (ArenaHard)†	63.53	67.93	68.42	96.88	85.26	95.19	0.05
Eurus-RM-7B	62.75	58.07	69.47	75.00	58.95	72.78	0.19
InternLM2-7B-Reward	62.14	60.77	67.37	85.42	65.26	83.16	0.14
InternLM2-20B-Reward	61.56	59.94	67.37	83.33	71.58	88.87	0.12
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	61.56	50.96	59.47	90.62	72.63	89.02	0.11
Skywork-Reward-Llama-3.1-8B	61.15	62.46	68.42	88.54	70.53	86.62	0.11
ArmoRM-Llama3-8B-v0.1	60.99	61.81	61.58	89.58	70.53	87.22	0.11
NaiveVerbosityModel	59.67	37.71	66.84	66.67	44.21	58.65	0.25
Llama-3-OffsetBias-RM-8B	59.42	56.03	59.47	73.96	62.11	80.15	0.16
Nemotron-4-340B-Reward	59.06	55.82	67.37	87.50	73.68	90.38	0.10
InternLM2-1.8B-Reward	58.49	52.40	61.58	63.54	48.42	63.91	0.21
Starling-RM-7B-Alpha	57.59	51.48	60.53	80.21	61.05	81.05	0.16
Skywork-Reward-Gemma-2-27B	56.21	40.13	38.42	63.54	70.53	89.02	0.11

Table A.1: Reward model and LLM judge performance on Hard prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (AlpacaEval)†	70.15	52.24	52.10	83.33	75.79	91.58	0.09
GPT-4o-2024-08-06 (AlpacaEval)†	69.97	52.01	47.37	83.33	72.63	90.08	0.09
Ensemble-Judges (ArenaHard)†	69.59	57.24	63.16	83.33	83.16	94.74	0.08
GPT-4o-2024-08-06 (ArenaHard)†	68.54	56.01	52.10	81.25	77.89	93.53	0.08
GPT-4o-Mini-2024-07-18 (ArenaHard)†	67.50	50.08	46.32	78.12	72.63	88.72	0.09
Llama-3.1-70B-Instruct (AlpacaEval)†	67.40	46.25	46.32	68.75	60.00	80.60	0.14
Gemini-1.5-Pro-002 (ArenaHard)†	67.08	55.16	57.37	90.62	82.11	94.89	0.06
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	66.98	44.87	35.26	61.46	67.37	84.51	0.12
Claude-3-5-Sonnet-20240620 (ArenaHard)†	66.95	55.98	58.42	87.50	72.63	90.53	0.09
Gemini-1.5-Flash-002 (AlpacaEval)†	66.92	45.52	48.95	76.04	72.63	88.42	0.10
Athene-RM-70B	66.90	58.55	64.21	93.75	77.89	92.48	0.08
Gemini-1.5-Pro-002 (AlpacaEval)†	65.96	51.60	53.68	84.38	81.05	93.23	0.06
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	65.39	42.05	25.79	46.88	69.47	85.71	0.12
Athene-RM-8B	64.49	53.01	58.95	83.33	64.21	83.16	0.13
Llama-3.1-70B-Instruct (ArenaHard)†	64.10	48.06	40.53	68.75	64.21	82.71	0.12
Skywork-Reward-Llama-3.1-8B	63.24	42.44	46.32	56.25	62.11	78.80	0.15
Gemini-1.5-Pro-001 (ArenaHard)†	62.65	40.53	54.21	78.12	80.00	93.68	0.09
Eurus-RM-7B	61.82	34.66	41.05	31.25	36.84	45.71	0.27
InternLM2-7B-Reward	61.70	32.69	34.74	45.83	45.26	60.60	0.23
Starling-RM-34B	61.41	33.87	35.79	41.67	44.21	60.75	0.22
Gemini-1.5-Flash-002 (ArenaHard)†	61.01	42.41	46.84	77.08	68.42	87.52	0.10
InternLM2-20B-Reward	60.37	40.89	42.63	51.04	42.11	57.29	0.23
ArmoRM-Llama3-8B-v0.1	60.28	34.56	40.53	53.12	58.95	73.08	0.17
Nemotron-4-340B-Reward	59.58	45.52	56.32	68.75	67.37	84.06	0.13
NaiveVerbosityModel	59.24	12.01	45.79	5.21	6.32	8.57	0.40
Starling-RM-7B-Alpha	58.70	27.17	38.95	29.17	28.42	39.25	0.30
Llama-3-OffsetBias-RM-8B	58.66	35.23	29.47	29.17	43.16	55.49	0.23
Skywork-Reward-Gemma-2-27B	56.74	45.42	40.00	66.67	77.89	92.18	0.09
InternLM2-1.8B-Reward	55.54	30.02	27.89	15.62	22.11	29.32	0.30

Table A.2: Reward model and LLM judge performance on Easy prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	69.77	66.89	70.00	97.09	83.16	93.68	0.07
Claude-3-5-Sonnet-20240620 (ArenaHard)†	68.38	70.13	64.74	92.23	80.00	91.88	0.07
Ensemble-Judges (AlpacaEval)†	67.86	69.18	70.00	96.12	86.32	95.04	0.05
GPT-4o-2024-08-06 (ArenaHard)†	67.51	60.99	66.84	96.12	78.95	92.93	0.08
Gemini-1.5-Pro-002 (AlpacaEval)†	66.78	68.61	73.16	97.09	88.42	96.54	0.04
Gemini-1.5-Pro-002 (ArenaHard)†	66.70	69.92	68.42	97.09	82.11	93.83	0.06
Athene-RM-70B	66.50	63.79	75.26	95.15	77.89	90.98	0.09
GPT-4o-2024-08-06 (AlpacaEval)†	66.09	64.39	65.26	92.23	82.11	93.98	0.06
GPT-4o-Mini-2024-07-18 (ArenaHard)†	65.75	62.88	73.16	92.23	76.84	90.53	0.09
Gemini-1.5-Flash-002 (AlpacaEval)†	65.43	64.33	65.79	89.32	82.11	93.38	0.07
Athene-RM-8B	64.77	60.56	68.42	90.29	76.84	89.32	0.09
Llama-3.1-70B-Instruct (AlpacaEval)†	63.68	63.11	63.16	79.61	75.79	88.57	0.10
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	63.42	57.93	59.47	81.55	71.58	87.97	0.10
Gemini-1.5-Pro-001 (ArenaHard)†	63.25	66.39	62.63	88.35	80.00	91.13	0.08
Llama-3.1-70B-Instruct (ArenaHard)†	63.04	59.85	62.10	83.50	76.84	90.83	0.08
Gemini-1.5-Flash-002 (ArenaHard)†	62.66	60.73	61.05	87.38	75.79	89.77	0.09
Nemotron-4-340B-Reward	61.89	56.91	63.16	86.41	71.58	86.92	0.11
InternLM2-20B-Reward	61.89	57.38	64.74	79.61	64.21	83.76	0.15
Skywork-Reward-Llama-3.1-8B	61.41	57.88	66.32	81.55	74.74	88.12	0.10
InternLM2-7B-Reward	61.41	55.07	64.74	66.99	63.16	80.45	0.16
Starling-RM-34B	61.11	52.85	61.05	77.67	65.26	82.41	0.13
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	61.10	50.62	43.16	66.99	72.63	87.82	0.10
Eurus-RM-7B	60.90	51.96	59.47	65.05	51.58	65.26	0.20
ArmoRM-Llama3-8B-v0.1	60.87	55.71	56.32	78.64	76.84	90.53	0.10
Llama-3-OffsetBias-RM-8B	60.22	55.63	51.05	65.05	68.42	83.01	0.15
InternLM2-1.8B-Reward	57.27	38.46	55.79	39.81	42.11	59.55	0.23
NaiveVerbosityModel	57.07	31.21	56.84	32.04	33.68	47.67	0.29
Skywork-Reward-Gemma-2-27B	56.43	43.85	32.63	54.37	75.79	91.43	0.09
Starling-RM-7B-Alpha	55.71	40.10	48.42	52.43	44.21	58.20	0.22

Table A.3: Reward model and LLM judge performance on If prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Claude-3-5-Sonnet-20240620 (ArenaHard)†	68.06	57.64	62.63	97.22	88.42	97.74	0.04
Ensemble-Judges (ArenaHard)†	67.98	58.22	71.58	91.67	84.21	96.09	0.05
GPT-4o-2024-08-06 (ArenaHard)†	67.66	58.16	65.79	97.22	88.42	97.29	0.04
Ensemble-Judges (AlpacaEval)†	67.47	55.98	72.11	94.44	82.11	94.14	0.06
Athene-RM-70B	66.87	57.57	70.53	94.44	81.05	93.23	0.07
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.08	53.90	67.90	100.00	85.26	96.24	0.05
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	65.92	45.70	60.00	97.22	81.05	94.44	0.08
Gemini-1.5-Pro-002 (AlpacaEval)†	65.57	56.07	65.79	91.67	76.84	91.88	0.08
GPT-4o-2024-08-06 (AlpacaEval)†	65.50	55.66	62.10	94.44	86.32	95.94	0.05
Athene-RM-8B	65.22	57.37	70.00	94.44	76.84	92.18	0.09
Llama-3.1-70B-Instruct (AlpacaEval)†	64.40	54.30	62.10	94.44	75.79	92.03	0.09
Llama-3.1-70B-Instruct (ArenaHard)†	64.37	47.58	58.42	97.22	78.95	94.14	0.07
Gemini-1.5-Flash-002 (AlpacaEval)†	64.36	42.96	57.37	88.89	72.63	89.92	0.11
Starling-RM-34B	64.29	56.23	66.84	88.89	74.74	89.32	0.10
Gemini-1.5-Pro-002 (ArenaHard)†	64.18	54.06	66.32	90.28	77.89	92.78	0.08
InternLM2-7B-Reward	63.53	46.74	65.26	84.72	68.42	86.47	0.12
Eurus-RM-7B	62.98	57.01	66.32	81.94	62.11	78.05	0.16
Gemini-1.5-Flash-002 (ArenaHard)†	62.65	56.60	54.74	95.83	80.00	93.68	0.07
InternLM2-20B-Reward	62.10	47.74	58.95	90.28	75.79	91.13	0.09
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	61.77	37.46	44.74	83.33	77.89	93.68	0.08
Gemini-1.5-Pro-001 (ArenaHard)†	61.55	46.75	56.32	94.44	75.79	91.43	0.08
NaiveVerbosityModel	61.39	41.83	63.68	79.17	48.42	66.02	0.22
ArmoRM-Llama3-8B-v0.1	61.01	49.40	51.05	93.06	81.05	93.83	0.08
Skywork-Reward-Llama-3.1-8B	61.01	50.02	61.05	93.06	76.84	91.58	0.10
Llama-3-OffsetBias-RM-8B	59.80	45.80	48.95	62.50	64.21	83.01	0.14
InternLM2-1.8B-Reward	58.76	45.07	58.42	62.50	54.74	71.28	0.19
Starling-RM-7B-Alpha	58.71	46.85	56.32	76.39	64.21	78.80	0.15
Nemotron-4-340B-Reward	57.94	35.96	51.05	79.17	72.63	89.62	0.10
Skywork-Reward-Gemma-2-27B	56.41	25.46	26.84	54.17	64.21	84.51	0.13

Table A.4: Reward model and LLM judge performance on Is code subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	73.58	54.87	65.79	88.73	80.00	94.44	0.07
GPT-4o-2024-08-06 (ArenaHard)†	72.57	56.46	63.16	88.73	82.11	94.89	0.06
Claude-3-5-Sonnet-20240620 (ArenaHard)†	71.79	49.92	60.53	88.73	78.95	93.38	0.08
GPT-4o-Mini-2024-07-18 (ArenaHard)†	70.20	50.30	55.26	87.32	71.58	87.97	0.11
Gemini-1.5-Pro-002 (ArenaHard)†	69.61	60.91	58.42	84.51	77.89	92.63	0.08
Ensemble-Judges (AlpacaEval)†	69.09	52.15	62.10	91.55	74.74	91.13	0.09
Llama-3.1-70B-Instruct (ArenaHard)†	68.93	46.05	54.74	84.51	72.63	87.82	0.10
Athene-RM-70B	68.58	57.39	67.37	85.92	77.89	92.33	0.09
GPT-4o-2024-08-06 (AlpacaEval)†	68.21	53.79	56.84	88.73	77.89	92.93	0.08
Gemini-1.5-Pro-002 (AlpacaEval)†	67.25	55.63	59.47	88.73	84.21	95.04	0.07
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	66.67	46.28	54.21	84.51	58.95	78.95	0.16
Llama-3.1-70B-Instruct (AlpacaEval)†	65.12	46.95	56.84	83.10	57.89	79.55	0.14
Gemini-1.5-Pro-001 (ArenaHard)†	64.70	47.86	51.58	84.51	77.89	92.63	0.08
Gemini-1.5-Flash-002 (ArenaHard)†	64.62	45.11	53.68	85.92	71.58	87.22	0.09
Starling-RM-34B	63.88	36.42	55.79	78.87	64.21	83.91	0.14
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	63.66	44.85	50.53	83.10	65.26	84.51	0.14
Athene-RM-8B	62.85	42.56	61.05	83.10	67.37	85.56	0.12
Gemini-1.5-Flash-002 (AlpacaEval)†	62.70	41.05	47.90	74.65	66.32	83.91	0.11
InternLM2-20B-Reward	62.63	40.47	55.26	76.06	71.58	87.37	0.11
Nemotron-4-340B-Reward	61.60	48.64	59.47	87.32	77.89	93.23	0.09
InternLM2-7B-Reward	61.53	41.83	55.26	73.24	61.05	80.00	0.15
Eurus-RM-7B	61.31	35.08	54.21	57.75	47.37	64.06	0.22
Skywork-Reward-Llama-3.1-8B	60.65	43.03	53.16	77.46	63.16	81.65	0.14
ArmoRM-Llama3-8B-v0.1	59.32	37.16	44.74	73.24	65.26	83.31	0.14
Llama-3-OffsetBias-RM-8B	58.96	31.99	50.00	70.42	54.74	71.88	0.20
InternLM2-1.8B-Reward	58.74	33.52	36.84	45.07	49.47	67.82	0.19
Starling-RM-7B-Alpha	58.08	26.79	38.95	56.34	54.74	74.59	0.18
NaiveVerbosityModel	57.49	27.69	60.00	49.30	30.53	41.05	0.31
Skywork-Reward-Gemma-2-27B	55.80	35.07	25.26	46.48	60.00	75.94	0.14

Table A.5: Reward model and LLM judge performance on Math prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Nemotron-4-340B-Reward	62.65	56.88	58.95	62.28	51.58	68.42	0.19
Gemini-1.5-Pro-002 (ArenaHard)†	59.90	45.67	66.32	44.74	37.89	53.38	0.27
Gemini-1.5-Pro-001 (ArenaHard)†	58.01	36.29	52.63	42.11	41.05	53.23	0.27
ArmoRM-Llama3-8B-v0.1	56.83	33.59	43.16	42.98	36.84	47.82	0.27
Gemini-1.5-Pro-002 (AlpacaEval)†	56.83	30.75	67.90	38.60	30.53	45.41	0.31
Athene-RM-70B	55.81	31.06	67.37	35.96	28.42	44.06	0.32
Ensemble-Judges (ArenaHard)†	55.27	36.57	66.32	42.11	37.89	53.68	0.27
Skywork-Reward-Llama-3.1-8B	54.67	24.79	55.26	36.84	29.47	41.50	0.33
Skywork-Reward-Gemma-2-27B	54.50	34.00	35.79	38.60	43.16	57.89	0.21
Llama-3-OffsetBias-RM-8B	54.04	30.51	41.58	42.11	34.74	49.77	0.26
Athene-RM-8B	54.04	23.29	64.74	32.46	25.26	39.85	0.34
GPT-4o-2024-08-06 (ArenaHard)†	52.74	29.48	58.95	40.35	34.74	53.38	0.29
InternLM2-20B-Reward	52.43	29.55	55.79	39.47	36.84	55.94	0.26
Claude-3-5-Sonnet-20240620 (ArenaHard)†	52.32	28.63	58.42	33.33	38.95	51.73	0.28
Ensemble-Judges (AlpacaEval)†	51.26	16.53	57.90	31.58	27.37	39.10	0.33
GPT-4o-2024-08-06 (AlpacaEval)†	50.18	12.95	51.05	31.58	33.68	50.08	0.30
GPT-4o-Mini-2024-07-18 (ArenaHard)†	50.06	15.15	51.58	30.70	28.42	45.71	0.30
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	48.41	-1.95	24.21	15.79	20.00	29.92	0.31
InternLM2-1.8B-Reward	47.86	2.97	36.32	-3.51	9.47	20.75	0.37
Gemini-1.5-Flash-002 (ArenaHard)†	47.13	16.99	48.95	18.42	22.11	38.95	0.33
Gemini-1.5-Flash-002 (AlpacaEval)†	46.72	5.46	48.95	17.54	14.74	23.16	0.37
InternLM2-7B-Reward	45.77	-3.02	42.63	9.65	14.74	21.80	0.36
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	45.39	2.05	35.26	14.04	10.53	16.24	0.37
Llama-3.1-70B-Instruct (AlpacaEval)†	45.33	-4.86	46.84	11.40	6.32	14.59	0.39
Llama-3.1-70B-Instruct (ArenaHard)†	45.27	7.88	45.26	18.42	20.00	31.88	0.34
Eurus-RM-7B	39.81	-19.21	37.90	-7.02	-2.11	-1.65	0.45
Starling-RM-34B	39.23	-21.35	35.79	-6.14	1.05	0.45	0.42
Starling-RM-7B-Alpha	38.59	-25.59	32.63	-12.28	-3.16	-5.41	0.44
NaiveVerbosityModel	6.10	-93.99	52.63	-75.44	-94.74	-99.10	0.85

Table A.6: Reward model and LLM judge performance on Shorter won subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	68.15	71.49	73.16	91.59	86.32	95.64	0.06
Ensemble-Judges (AlpacaEval)†	67.28	73.31	74.21	92.52	84.21	94.44	0.06
GPT-4o-2024-08-06 (ArenaHard)†	67.23	71.93	71.05	92.52	84.21	95.19	0.07
Claude-3-5-Sonnet-20240620 (ArenaHard)†	67.08	72.22	70.00	88.79	84.21	93.83	0.06
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.29	71.23	69.47	89.72	80.00	92.48	0.08
Athene-RM-70B	65.84	72.39	81.05	90.65	78.95	91.88	0.09
Gemini-1.5-Pro-002 (AlpacaEval)†	65.54	71.75	74.21	92.52	85.26	94.74	0.06
GPT-4o-2024-08-06 (AlpacaEval)†	65.45	71.06	68.42	88.79	82.11	93.68	0.07
Gemini-1.5-Flash-002 (AlpacaEval)†	64.88	66.90	66.84	88.79	74.74	88.87	0.10
Llama-3.1-70B-Instruct (AlpacaEval)†	64.86	71.92	75.26	88.79	71.58	86.47	0.11
Gemini-1.5-Pro-002 (ArenaHard)†	64.84	70.79	73.16	90.65	83.16	93.83	0.07
Athene-RM-8B	64.28	68.70	78.95	89.72	74.74	88.57	0.10
Starling-RM-34B	64.05	67.27	75.79	83.18	71.58	85.56	0.12
Llama-3.1-70B-Instruct (ArenaHard)†	63.96	66.05	68.95	85.98	72.63	87.52	0.12
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	63.95	65.29	65.79	87.85	70.53	85.71	0.12
Gemini-1.5-Flash-002 (ArenaHard)†	63.26	66.65	72.63	88.79	74.74	89.47	0.10
Skywork-Reward-Llama-3.1-8B	62.83	71.83	73.68	97.20	81.05	92.18	0.08
Gemini-1.5-Pro-001 (ArenaHard)†	62.46	64.75	66.32	86.92	77.89	90.68	0.09
Eurus-RM-7B	62.07	56.73	68.95	73.83	57.89	72.03	0.20
NaiveVerbosityModel	61.30	40.25	68.95	53.27	34.74	49.92	0.30
InternLM2-7B-Reward	60.82	61.98	69.47	77.57	60.00	80.30	0.16
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	60.59	60.26	57.90	87.85	75.79	88.87	0.10
ArmoRM-Llama3-8B-v0.1	60.03	63.19	71.05	90.65	81.05	90.98	0.07
Starling-RM-7B-Alpha	59.01	54.50	64.21	64.49	49.47	70.83	0.20
InternLM2-20B-Reward	59.00	54.89	68.95	69.16	57.89	78.20	0.17
Llama-3-OffsetBias-RM-8B	58.58	57.04	58.95	71.96	64.21	81.80	0.14
Nemotron-4-340B-Reward	57.74	50.81	75.26	65.42	57.89	73.98	0.19
Skywork-Reward-Gemma-2-27B	55.93	54.08	51.58	76.64	75.79	90.68	0.10
InternLM2-1.8B-Reward	55.92	37.43	61.58	42.99	36.84	55.64	0.27

Table A.7: Reward model and LLM judge performance on Similar response subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	68.17	70.80	71.58	86.24	81.05	94.14	0.08
GPT-4o-2024-08-06 (ArenaHard)†	67.78	71.61	68.95	86.24	83.16	94.89	0.07
Ensemble-Judges (AlpacaEval)†	67.60	70.66	71.58	84.40	76.84	92.93	0.10
GPT-4o-2024-08-06 (AlpacaEval)†	66.70	63.51	66.32	80.73	76.84	91.73	0.09
Claude-3-5-Sonnet-20240620 (ArenaHard)†	66.42	68.25	70.53	86.24	78.95	93.68	0.08
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.39	66.39	67.37	81.65	78.95	92.03	0.09
Athene-RM-70B	65.53	68.75	79.47	83.49	73.68	90.98	0.12
Gemini-1.5-Pro-002 (AlpacaEval)†	65.37	70.68	74.74	87.16	76.84	91.88	0.10
Llama-3.1-70B-Instruct (AlpacaEval)†	64.79	65.74	72.11	78.90	66.32	85.56	0.13
Gemini-1.5-Pro-002 (ArenaHard)†	64.75	69.77	71.58	84.40	76.84	92.93	0.10
Gemini-1.5-Flash-002 (AlpacaEval)†	64.48	65.98	67.90	79.82	69.47	86.02	0.13
Llama-3.1-70B-Instruct (ArenaHard)†	64.31	63.74	67.90	82.57	70.53	88.87	0.12
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	64.27	62.80	65.26	79.82	68.42	86.47	0.13
Athene-RM-8B	63.55	65.76	75.26	81.65	69.47	89.32	0.13
Starling-RM-34B	63.50	60.04	72.63	68.81	65.26	81.80	0.16
Gemini-1.5-Flash-002 (ArenaHard)†	62.97	64.16	66.84	77.98	70.53	88.12	0.12
Skywork-Reward-Llama-3.1-8B	62.94	68.77	70.53	87.16	75.79	90.98	0.10
Gemini-1.5-Pro-001 (ArenaHard)†	62.04	64.66	65.79	86.24	70.53	89.47	0.12
Eurus-RM-7B	61.78	51.70	71.58	58.72	52.63	65.86	0.20
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	61.64	57.42	59.47	81.65	71.58	87.52	0.11
NaiveVerbosityModel	61.26	40.80	68.42	48.62	43.16	51.73	0.26
InternLM2-7B-Reward	61.01	53.18	66.84	70.64	58.95	80.30	0.18
ArmoRM-Llama3-8B-v0.1	60.94	64.96	70.00	83.49	75.79	90.38	0.10
Starling-RM-7B-Alpha	59.55	50.50	67.90	53.21	55.79	71.43	0.21
InternLM2-20B-Reward	59.34	54.73	68.95	65.14	50.53	71.58	0.20
Llama-3-OffsetBias-RM-8B	59.06	54.04	55.26	66.06	54.74	69.47	0.20
Nemotron-4-340B-Reward	57.47	44.46	71.05	62.39	50.53	67.07	0.22
InternLM2-1.8B-Reward	56.17	41.19	61.58	38.53	32.63	50.23	0.28
Skywork-Reward-Gemma-2-27B	55.21	57.61	49.47	73.39	69.47	87.52	0.11

Table A.8: Reward model and LLM judge performance on English prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (AlpacaEval)†	69.68	73.76	74.21	94.31	90.53	97.74	0.03
Ensemble-Judges (ArenaHard)†	69.09	75.81	76.84	93.50	86.32	95.79	0.06
Claude-3-5-Sonnet-20240620 (ArenaHard)†	68.48	75.18	75.26	91.87	86.32	96.39	0.05
Athene-RM-70B	67.86	73.24	76.84	91.87	82.11	94.89	0.07
GPT-4o-2024-08-06 (AlpacaEval)†	67.66	72.18	72.63	98.37	93.68	98.65	0.03
GPT-4o-2024-08-06 (ArenaHard)†	67.63	71.24	73.16	91.87	82.11	94.74	0.07
Gemini-1.5-Pro-002 (AlpacaEval)†	67.01	73.72	80.00	94.31	88.42	97.14	0.05
Gemini-1.5-Pro-002 (ArenaHard)†	66.93	74.39	75.26	90.24	82.11	94.29	0.07
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	66.68	67.72	60.53	80.49	81.05	94.14	0.07
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.55	71.23	72.63	90.24	82.11	94.44	0.07
Athene-RM-8B	65.91	70.37	80.53	92.68	82.11	95.04	0.07
Llama-3.1-70B-Instruct (AlpacaEval)†	65.87	65.70	68.95	83.74	75.79	90.53	0.09
Gemini-1.5-Flash-002 (AlpacaEval)†	65.75	70.61	67.90	86.99	87.37	96.84	0.06
Llama-3.1-70B-Instruct (ArenaHard)†	64.25	68.81	65.26	82.11	80.00	93.38	0.09
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	64.17	62.56	54.74	78.05	83.16	94.44	0.06
InternLM2-7B-Reward	63.36	63.58	65.79	69.11	62.11	84.21	0.16
Gemini-1.5-Pro-001 (ArenaHard)†	63.24	70.19	70.53	87.80	80.00	94.14	0.08
InternLM2-20B-Reward	63.10	63.69	72.11	76.42	64.21	86.17	0.16
Gemini-1.5-Flash-002 (ArenaHard)†	63.06	68.96	71.05	86.18	77.89	93.38	0.08
Eurus-RM-7B	62.32	56.17	61.05	67.48	66.32	75.49	0.16
Starling-RM-34B	62.19	58.76	64.21	73.17	70.53	86.32	0.12
Skywork-Reward-Llama-3.1-8B	61.66	64.18	70.53	75.61	73.68	87.52	0.11
Nemotron-4-340B-Reward	61.57	67.30	72.63	83.74	76.84	90.53	0.10
ArmoRM-Llama3-8B-v0.1	60.11	59.89	58.95	66.67	73.68	90.53	0.12
Llama-3-OffsetBias-RM-8B	59.20	48.58	55.79	52.85	53.68	69.17	0.19
InternLM2-1.8B-Reward	58.55	44.78	55.26	43.90	41.05	56.24	0.24
Skywork-Reward-Gemma-2-27B	58.40	58.79	61.05	83.74	83.16	95.19	0.06
Starling-RM-7B-Alpha	58.13	40.90	59.47	55.28	48.42	60.75	0.22
NaiveVerbosityModel	57.98	21.46	64.21	30.89	21.05	27.52	0.36

Table A.9: Reward model and LLM judge performance on Non english prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (AlpacaEval)†	67.91	52.67	54.21	93.33	80.00	94.14	0.07
Claude-3-5-Sonnet-20240620 (ArenaHard)†	67.03	50.91	48.42	90.00	78.95	93.38	0.08
Athene-RM-70B	66.39	45.24	61.05	90.00	83.16	93.83	0.07
Gemini-1.5-Pro-002 (AlpacaEval)†	66.27	49.83	58.42	93.33	82.11	93.38	0.08
Ensemble-Judges (ArenaHard)†	66.15	53.77	47.37	86.67	77.89	92.33	0.07
GPT-4o-2024-08-06 (ArenaHard)†	65.37	49.18	52.10	90.00	76.84	92.18	0.08
GPT-4o-Mini-2024-07-18 (ArenaHard)†	65.29	51.87	44.74	76.67	66.32	86.47	0.12
Gemini-1.5-Flash-002 (AlpacaEval)†	65.10	40.01	46.32	86.67	71.58	89.17	0.09
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	64.89	47.98	43.16	88.33	69.47	87.52	0.11
InternLM2-20B-Reward	64.62	42.76	48.42	56.67	65.26	83.91	0.12
Athene-RM-8B	64.45	42.41	60.00	86.67	81.05	94.59	0.07
Gemini-1.5-Pro-002 (ArenaHard)†	64.16	49.86	51.05	80.00	76.84	91.88	0.08
InternLM2-7B-Reward	63.87	44.35	41.05	53.33	70.53	89.17	0.11
GPT-4o-2024-08-06 (AlpacaEval)†	63.53	43.47	51.58	90.00	83.16	94.89	0.06
Llama-3.1-70B-Instruct (ArenaHard)†	63.04	32.00	48.42	81.67	60.00	81.65	0.14
Llama-3.1-70B-Instruct (AlpacaEval)†	63.03	36.40	47.90	68.33	67.37	86.17	0.13
Starling-RM-34B	62.52	40.66	56.32	85.00	71.58	86.32	0.11
Gemini-1.5-Flash-002 (ArenaHard)†	62.48	43.33	46.32	83.33	73.68	89.02	0.09
Gemini-1.5-Pro-001 (ArenaHard)†	62.09	36.12	41.05	75.00	71.58	89.77	0.09
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	61.43	38.81	23.68	55.00	63.16	83.01	0.14
Eurus-RM-7B	61.18	39.05	44.21	70.00	65.26	81.05	0.14
InternLM2-1.8B-Reward	60.08	38.02	42.63	40.00	51.58	70.83	0.20
Skywork-Reward-Gemma-2-27B	59.16	22.83	26.84	75.00	86.32	96.09	0.06
Nemotron-4-340B-Reward	58.07	28.62	32.63	45.00	52.63	72.33	0.18
Llama-3-OffsetBias-RM-8B	57.48	27.04	27.37	28.33	52.63	68.12	0.20
Skywork-Reward-Llama-3.1-8B	57.23	38.20	37.37	53.33	64.21	81.20	0.13
ArmoRM-Llama3-8B-v0.1	56.64	18.09	26.84	28.33	46.32	59.40	0.21
NaiveVerbosityModel	56.55	19.66	48.95	11.67	14.74	21.05	0.36
Starling-RM-7B-Alpha	54.29	7.14	28.42	18.33	35.79	47.37	0.23

Table A.10: Reward model and LLM judge performance on Chinese prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	70.37	50.61	53.16	92.86	77.89	92.63	0.09
Ensemble-Judges (AlpacaEval)†	69.43	51.76	57.90	92.86	80.00	94.44	0.06
Claude-3-5-Sonnet-20240620 (ArenaHard)†	68.63	44.71	50.53	85.71	70.53	87.97	0.09
GPT-4o-2024-08-06 (AlpacaEval)†	68.58	42.94	38.95	91.07	77.89	93.83	0.08
GPT-4o-2024-08-06 (ArenaHard)†	68.54	43.94	47.37	89.29	70.53	89.02	0.10
Athene-RM-70B	68.49	48.66	58.42	94.64	77.89	90.68	0.09
Gemini-1.5-Pro-002 (ArenaHard)†	67.23	49.82	53.68	87.50	73.68	89.32	0.10
Gemini-1.5-Pro-002 (AlpacaEval)†	66.20	50.01	58.42	92.86	78.95	93.38	0.07
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	66.13	42.56	45.79	85.71	76.84	89.62	0.10
Llama-3.1-70B-Instruct (AlpacaEval)†	65.65	38.73	47.90	92.86	66.32	85.56	0.12
GPT-4o-Mini-2024-07-18 (ArenaHard)†	65.49	40.39	45.26	85.71	75.79	91.28	0.09
Gemini-1.5-Flash-002 (AlpacaEval)†	65.21	42.35	50.00	94.64	75.79	91.73	0.09
Athene-RM-8B	64.87	41.89	55.79	91.07	71.58	86.62	0.10
Nemotron-4-340B-Reward	63.86	41.06	52.10	87.50	72.63	87.07	0.10
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	63.82	31.28	23.68	71.43	82.11	93.83	0.08
Llama-3.1-70B-Instruct (ArenaHard)†	63.37	28.42	40.53	69.64	64.21	81.80	0.14
Gemini-1.5-Flash-002 (ArenaHard)†	63.26	31.97	42.63	76.79	67.37	85.56	0.12
Eurus-RM-7B	62.84	33.63	43.68	76.79	56.84	73.38	0.16
Gemini-1.5-Pro-001 (ArenaHard)†	62.08	43.28	46.32	78.57	70.53	88.12	0.11
Skywork-Reward-Llama-3.1-8B	61.17	23.32	41.58	73.21	65.26	84.51	0.13
InternLM2-7B-Reward	61.08	30.92	41.58	46.43	58.95	78.05	0.15
Starling-RM-34B	60.98	36.02	36.32	73.21	63.16	80.00	0.13
InternLM2-20B-Reward	60.43	26.87	39.47	30.36	60.00	78.50	0.16
ArmoRM-Llama3-8B-v0.1	60.33	38.52	35.26	83.93	74.74	90.23	0.09
Starling-RM-7B-Alpha	59.41	31.55	38.95	69.64	53.68	66.77	0.19
Llama-3-OffsetBias-RM-8B	59.04	25.82	30.53	50.00	48.42	68.27	0.19
NaiveVerbosityModel	59.04	10.26	34.21	33.93	29.47	38.95	0.29
InternLM2-1.8B-Reward	57.65	26.88	25.79	17.86	45.26	60.75	0.21
Skywork-Reward-Gemma-2-27B	56.26	29.71	23.68	50.00	64.21	82.86	0.14

Table A.11: Reward model and LLM judge performance on Russian prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (ArenaHard)†	75.16	38.73	38.42	84.62	73.68	88.42	0.10
Claude-3-5-Sonnet-20240620 (ArenaHard)†	72.49	30.32	23.16	66.67	65.26	81.50	0.12
GPT-4o-2024-08-06 (ArenaHard)†	71.03	31.32	24.74	84.62	72.63	85.86	0.10
Gemini-1.5-Pro-002 (ArenaHard)†	70.64	29.57	27.89	76.92	72.63	87.22	0.11
GPT-4o-2024-08-06 (AlpacaEval)†	69.71	21.47	21.05	74.36	72.63	88.27	0.10
Ensemble-Judges (AlpacaEval)†	68.88	15.78	27.37	71.79	60.00	78.05	0.14
Athene-RM-70B	67.71	11.39	33.68	76.92	65.26	84.21	0.13
Nemotron-4-340B-Reward	66.86	27.91	26.84	71.79	62.11	83.16	0.12
Llama-3.1-70B-Instruct (AlpacaEval)†	66.86	27.69	25.79	66.67	51.58	69.17	0.17
Gemini-1.5-Flash-002 (AlpacaEval)†	66.86	18.29	24.21	61.54	54.74	73.38	0.15
Gemini-1.5-Pro-002 (AlpacaEval)†	66.29	8.72	33.68	69.23	69.47	84.81	0.13
GPT-4o-Mini-2024-07-18 (ArenaHard)†	66.00	13.41	11.58	61.54	70.53	86.32	0.11
Athene-RM-8B	65.43	3.68	37.37	76.92	67.37	83.31	0.12
Gemini-1.5-Flash-002 (ArenaHard)†	65.32	19.95	15.79	43.59	57.89	75.64	0.16
Llama-3.1-70B-Instruct (ArenaHard)†	64.66	21.95	17.37	48.72	52.63	68.42	0.16
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	63.69	11.97	7.37	20.51	46.32	61.65	0.20
Starling-RM-34B	63.43	11.24	11.58	46.15	49.47	64.81	0.19
Gemini-1.5-Pro-001 (ArenaHard)†	63.33	16.68	15.26	48.72	61.05	82.26	0.14
Eurus-RM-7B	62.57	14.76	8.95	41.03	44.21	56.54	0.22
InternLM2-7B-Reward	62.29	12.92	11.05	38.46	57.89	78.05	0.16
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	62.29	14.84	10.00	33.33	48.42	66.17	0.18
InternLM2-20B-Reward	61.71	18.35	24.21	61.54	60.00	79.40	0.15
ArmoRM-Llama3-8B-v0.1	60.86	-8.08	19.47	46.15	57.89	71.73	0.16
Skywork-Reward-Llama-3.1-8B	59.71	-4.01	20.00	53.85	57.89	72.03	0.16
NaiveVerbosityModel	56.86	17.14	8.42	12.82	-2.11	-4.36	0.36
Llama-3-OffsetBias-RM-8B	56.57	-4.02	13.68	30.77	46.32	56.69	0.21
Starling-RM-7B-Alpha	56.29	6.70	7.89	23.08	34.74	47.67	0.24
InternLM2-1.8B-Reward	55.14	13.77	7.37	30.77	32.63	40.75	0.24
Skywork-Reward-Gemma-2-27B	54.57	-11.99	6.84	23.08	45.26	60.45	0.19

Table A.12: Reward model and LLM judge performance on German prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Athene-RM-70B	71.10	46.16	37.37	84.21	67.37	83.76	0.14
Ensemble-Judges (AlpacaEval)†	69.63	32.44	34.21	52.63	63.16	82.71	0.13
Skywork-Reward-Llama-3.1-8B	68.81	40.32	22.11	68.42	58.95	78.20	0.14
Ensemble-Judges (ArenaHard)†	68.45	33.85	25.79	65.79	61.05	78.50	0.14
Gemini-1.5-Pro-002 (AlpacaEval)†	68.06	28.63	28.42	50.00	66.32	84.36	0.12
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	67.59	27.29	12.11	36.84	57.89	78.95	0.15
Llama-3.1-70B-Instruct (AlpacaEval)†	66.97	24.59	19.47	52.63	61.05	78.20	0.15
GPT-4o-2024-08-06 (AlpacaEval)†	66.97	34.79	27.37	44.74	66.32	86.32	0.13
GPT-4o-2024-08-06 (ArenaHard)†	66.67	30.49	25.26	63.16	63.16	81.05	0.13
InternLM2-20B-Reward	66.51	36.27	20.00	18.42	55.79	72.33	0.18
Gemini-1.5-Pro-002 (ArenaHard)†	66.36	29.17	21.05	73.68	61.05	79.85	0.14
Athene-RM-8B	65.60	31.00	32.63	63.16	60.00	78.65	0.14
GPT-4o-Mini-2024-07-18 (ArenaHard)†	65.14	29.31	25.79	73.68	74.74	89.92	0.12
Gemini-1.5-Flash-002 (AlpacaEval)†	64.81	21.30	18.42	50.00	66.32	84.36	0.13
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	64.68	14.42	18.42	31.58	60.00	79.70	0.14
Claude-3-5-Sonnet-20240620 (ArenaHard)†	64.68	27.59	21.05	55.26	65.26	86.02	0.12
Gemini-1.5-Flash-002 (ArenaHard)†	63.68	20.76	24.21	65.79	64.21	80.30	0.14
InternLM2-7B-Reward	63.30	30.05	9.47	-26.32	49.47	68.42	0.20
Llama-3.1-70B-Instruct (ArenaHard)†	63.13	10.68	17.89	73.68	57.89	78.80	0.15
Llama-3-OffsetBias-RM-8B	62.39	28.23	16.32	63.16	25.26	38.50	0.26
ArmoRM-Llama3-8B-v0.1	62.39	29.54	23.16	60.53	43.16	58.65	0.20
Gemini-1.5-Pro-001 (ArenaHard)†	62.24	19.36	13.16	57.89	60.00	78.95	0.13
Eurus-RM-7B	61.47	30.57	15.79	44.74	50.53	71.43	0.17
Nemotron-4-340B-Reward	61.47	17.85	26.84	31.58	44.21	52.63	0.23
Starling-RM-34B	60.09	16.40	14.21	68.42	55.79	70.98	0.17
InternLM2-1.8B-Reward	57.34	19.72	6.32	-7.89	38.95	54.59	0.21
NaiveVerbosityModel	56.88	9.00	8.42	-28.95	15.79	20.90	0.25
Starling-RM-7B-Alpha	55.96	18.12	16.32	44.74	44.21	57.44	0.23
Skywork-Reward-Gemma-2-27B	55.05	8.51	20.53	55.26	42.11	56.54	0.20

Table A.13: Reward model and LLM judge performance on Korean prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	73.36	37.78	6.32	58.33	69.47	87.22	0.11
Athene-RM-8B	71.89	39.72	14.21	54.17	67.37	87.07	0.10
Ensemble-Judges (AlpacaEval)†	71.36	36.61	11.05	70.83	71.58	86.62	0.11
Llama-3.1-70B-Instruct (AlpacaEval)†	70.05	37.95	6.32	62.50	62.11	81.50	0.11
Claude-3-5-Sonnet-20240620 (ArenaHard)†	68.52	33.33	14.74	75.00	72.63	89.62	0.10
Athene-RM-70B	68.20	33.11	18.42	50.00	72.63	87.82	0.13
GPT-4o-Mini-2024-07-18 (ArenaHard)†	68.20	41.02	8.95	58.33	62.11	80.75	0.13
Gemini-1.5-Flash-002 (AlpacaEval)†	67.44	35.21	14.21	66.67	62.11	81.20	0.13
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	67.28	31.60	0.53	54.17	65.26	82.11	0.12
Gemini-1.5-Pro-002 (AlpacaEval)†	66.98	33.95	14.74	54.17	64.21	83.46	0.12
Skywork-Reward-Llama-3.1-8B	66.82	28.61	9.47	83.33	64.21	77.59	0.14
InternLM2-7B-Reward	66.36	19.15	16.32	25.00	53.68	70.53	0.16
Ensemble-Judges (ArenaHard)†	65.79	31.49	16.84	62.50	71.58	87.37	0.11
Starling-RM-34B	64.98	27.05	16.32	54.17	61.05	79.70	0.15
GPT-4o-2024-08-06 (AlpacaEval)†	64.52	29.56	5.79	37.50	64.21	82.11	0.13
GPT-4o-2024-08-06 (ArenaHard)†	64.10	28.43	15.26	58.33	69.47	86.47	0.12
Llama-3.1-70B-Instruct (ArenaHard)†	64.02	22.78	3.16	54.17	54.74	75.79	0.16
Nemotron-4-340B-Reward	63.59	28.08	8.95	37.50	67.37	83.46	0.13
Skywork-Reward-Gemma-2-27B	63.13	12.65	6.32	50.00	49.47	64.21	0.18
InternLM2-20B-Reward	63.13	21.49	9.47	-4.17	58.95	80.15	0.16
Gemini-1.5-Flash-002 (ArenaHard)†	63.03	33.38	7.89	54.17	62.11	82.26	0.11
Gemini-1.5-Pro-002 (ArenaHard)†	62.91	22.44	15.79	62.50	60.00	79.85	0.14
NaiveVerbosityModel	62.21	18.81	5.26	4.17	27.37	29.92	0.27
Eurus-RM-7B	61.29	20.76	3.68	20.83	47.37	63.61	0.19
ArmoRM-Llama3-8B-v0.1	60.37	12.93	9.47	75.00	22.11	33.08	0.24
Llama-3-OffsetBias-RM-8B	59.91	17.63	11.58	66.67	36.84	53.53	0.22
Gemini-1.5-Pro-001 (ArenaHard)†	59.51	15.30	3.16	66.67	51.58	70.38	0.15
InternLM2-1.8B-Reward	58.99	15.75	8.42	-20.83	36.84	53.98	0.22
Starling-RM-7B-Alpha	58.06	23.72	8.42	54.17	10.53	14.14	0.32

Table A.14: Reward model and LLM judge performance on Japanese prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Ensemble-Judges (AlpacaEval)†	72.11	31.81	5.79	36.84	20.00	30.53	0.28
GPT-4o-2024-08-06 (AlpacaEval)†	70.53	23.71	0.00	100.00	35.79	48.42	0.22
GPT-4o-2024-08-06 (ArenaHard)†	70.29	24.79	4.21	89.47	43.16	59.55	0.21
Athene-RM-70B	69.47	24.25	17.37	89.47	35.79	49.62	0.23
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	68.42	28.53	1.58	100.00	20.00	33.83	0.26
Llama-3.1-70B-Instruct (ArenaHard)†	67.93	29.52	6.32	78.95	25.26	32.63	0.28
Skywork-Reward-Llama-3.1-8B	67.89	20.95	7.37	89.47	35.79	52.33	0.21
Llama-3.1-70B-Instruct (AlpacaEval)†	67.89	27.03	2.63	100.00	32.63	49.77	0.22
NaiveVerbosityModel	67.37	24.77	2.11	100.00	25.26	34.89	0.24
Gemini-1.5-Flash-002 (AlpacaEval)†	67.37	29.36	4.74	68.42	25.26	37.44	0.26
InternLM2-7B-Reward	67.37	23.65	2.63	78.95	23.16	34.89	0.24
Starling-RM-34B	66.84	23.40	2.11	78.95	13.68	20.30	0.30
Ensemble-Judges (ArenaHard)†	66.47	20.45	12.63	47.37	28.42	40.15	0.24
Gemini-1.5-Pro-002 (AlpacaEval)†	66.32	19.40	11.05	47.37	24.21	38.05	0.25
Starling-RM-7B-Alpha	65.79	32.43	1.58	68.42	6.32	6.02	0.30
InternLM2-20B-Reward	65.26	24.19	1.05	100.00	21.05	32.78	0.25
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	64.74	22.02	0.00	100.00	11.58	14.89	0.27
Claude-3-5-Sonnet-20240620 (ArenaHard)†	64.74	21.07	8.95	5.26	24.21	36.54	0.26
Athene-RM-8B	64.21	23.88	9.47	68.42	27.37	40.45	0.26
Gemini-1.5-Pro-001 (ArenaHard)†	63.84	25.24	3.68	36.84	25.26	37.74	0.23
GPT-4o-Mini-2024-07-18 (ArenaHard)†	63.83	11.48	7.89	78.95	31.58	46.47	0.24
Gemini-1.5-Pro-002 (ArenaHard)†	63.64	15.85	11.05	36.84	32.63	46.02	0.23
Eurus-RM-7B	63.16	14.36	0.53	89.47	1.05	2.86	0.33
Llama-3-OffsetBias-RM-8B	61.05	20.44	1.58	100.00	42.11	53.68	0.21
Gemini-1.5-Flash-002 (ArenaHard)†	60.75	16.42	8.42	57.89	12.63	17.14	0.29
Skywork-Reward-Gemma-2-27B	60.00	30.32	0.53	89.47	22.11	31.58	0.27
ArmoRM-Llama3-8B-v0.1	59.47	15.07	3.16	100.00	33.68	47.07	0.23
InternLM2-1.8B-Reward	59.47	17.02	2.63	47.37	8.42	10.53	0.32
Nemotron-4-340B-Reward	58.42	10.01	6.32	89.47	20.00	29.17	0.28

Table A.15: Reward model and LLM judge performance on Spanish prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Gemini-1.5-Pro-002 (ArenaHard)†	69.57	14.77	14.74	54.17	63.16	82.41	0.14
GPT-4o-Mini-2024-07-18 (ArenaHard)†	68.45	25.12	4.21	75.00	54.74	73.08	0.17
Ensemble-Judges (ArenaHard)†	68.24	21.05	17.37	66.67	62.11	80.90	0.13
Ensemble-Judges (AlpacaEval)†	67.74	27.12	4.21	79.17	46.32	65.71	0.19
Gemini-1.5-Pro-002 (AlpacaEval)†	67.38	26.42	8.95	79.17	47.37	65.26	0.18
Athene-RM-8B	67.38	26.84	18.95	45.83	45.26	64.81	0.17
InternLM2-7B-Reward	66.31	20.42	11.05	45.83	43.16	62.41	0.19
Claude-3-5-Sonnet-20240620 (ArenaHard)†	66.31	24.02	5.79	45.83	55.79	73.53	0.15
Athene-RM-70B	65.78	22.45	17.89	54.17	45.26	65.86	0.18
InternLM2-20B-Reward	65.24	26.25	13.16	29.17	58.95	79.55	0.15
ArmoRM-Llama3-8B-v0.1	65.24	21.41	5.79	45.83	33.68	55.19	0.23
Llama-3-OffsetBias-RM-8B	64.71	13.13	2.11	79.17	27.37	41.80	0.23
GPT-4o-2024-08-06 (AlpacaEval)†	64.71	20.04	4.21	58.33	52.63	72.33	0.16
Llama-3.1-70B-Instruct (AlpacaEval)†	64.17	20.26	3.68	70.83	43.16	61.65	0.19
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	63.98	27.44	2.11	79.17	36.84	51.73	0.21
Starling-RM-7B-Alpha	63.10	22.33	9.47	54.17	34.74	47.67	0.20
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	62.57	30.14	1.05	70.83	25.26	38.50	0.24
GPT-4o-2024-08-06 (ArenaHard)†	62.43	15.80	8.95	70.83	49.47	65.56	0.18
Gemini-1.5-Flash-002 (ArenaHard)†	62.37	22.71	13.16	62.50	36.84	55.19	0.21
Eurus-RM-7B	62.03	14.76	8.42	37.50	17.89	26.17	0.29
Nemotron-4-340B-Reward	62.03	11.19	18.95	29.17	49.47	66.92	0.19
Gemini-1.5-Flash-002 (AlpacaEval)†	62.03	20.24	2.11	79.17	37.89	54.59	0.20
Llama-3.1-70B-Instruct (ArenaHard)†	61.62	20.93	3.68	70.83	46.32	69.17	0.17
Gemini-1.5-Pro-001 (ArenaHard)†	61.11	12.74	5.79	58.33	47.37	59.55	0.17
Skywork-Reward-Llama-3.1-8B	60.96	9.19	10.53	70.83	28.42	40.00	0.26
Starling-RM-34B	59.36	11.68	0.53	79.17	38.95	54.44	0.22
InternLM2-1.8B-Reward	58.82	21.97	4.21	12.50	36.84	46.47	0.21
Skywork-Reward-Gemma-2-27B	57.75	3.40	8.42	87.50	48.42	63.46	0.20
NaiveVerbosityModel	54.01	9.52	10.00	62.50	-2.11	-3.16	0.35

Table A.16: Reward model and LLM judge performance on French prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	71.84	31.95	2.11	100.00	49.47	67.82	0.18
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	68.93	27.08	7.37	100.00	48.42	67.97	0.22
InternLM2-7B-Reward	68.93	25.47	1.05	100.00	49.47	68.12	0.18
Claude-3-5-Sonnet-20240620 (ArenaHard)†	68.63	20.55	3.68	100.00	60.00	77.74	0.18
Ensemble-Judges (AlpacaEval)†	67.96	17.35	7.37	100.00	57.89	79.25	0.22
Ensemble-Judges (ArenaHard)†	67.02	20.72	10.53	100.00	62.11	76.39	0.17
GPT-4o-2024-08-06 (AlpacaEval)†	66.99	16.25	3.68	100.00	50.53	69.47	0.18
Skywork-Reward-Gemma-2-27B	66.02	21.16	4.74	100.00	58.95	77.29	0.20
Athene-RM-8B	66.02	20.34	8.42	89.47	54.74	75.49	0.16
Eurus-RM-7B	65.05	26.36	3.16	78.95	30.53	39.55	0.21
Athene-RM-70B	65.05	10.12	7.89	89.47	50.53	72.33	0.18
GPT-4o-Mini-2024-07-18 (ArenaHard)†	64.08	12.29	13.68	89.47	61.05	81.35	0.15
Gemini-1.5-Pro-002 (AlpacaEval)†	64.08	14.69	3.16	100.00	54.74	72.03	0.18
Gemini-1.5-Flash-002 (AlpacaEval)†	64.08	21.03	3.68	100.00	41.05	58.05	0.21
Llama-3-OffsetBias-RM-8B	64.08	28.73	11.05	100.00	27.37	40.15	0.21
InternLM2-20B-Reward	64.08	8.68	2.63	100.00	53.68	75.49	0.19
Gemini-1.5-Pro-002 (ArenaHard)†	64.00	12.53	12.63	89.47	48.42	65.56	0.19
GPT-4o-2024-08-06 (ArenaHard)†	63.27	18.86	5.26	89.47	56.84	72.63	0.16
Starling-RM-34B	63.11	14.73	2.63	89.47	42.11	58.20	0.18
Llama-3.1-70B-Instruct (AlpacaEval)†	62.14	19.12	1.05	100.00	63.16	78.05	0.15
Skywork-Reward-Llama-3.1-8B	62.14	25.10	6.32	100.00	36.84	54.59	0.21
Llama-3.1-70B-Instruct (ArenaHard)†	61.39	-2.36	3.68	100.00	55.79	76.09	0.18
ArmoRM-Llama3-8B-v0.1	60.19	19.66	2.11	100.00	18.95	32.18	0.25
InternLM2-1.8B-Reward	59.22	11.84	2.11	57.89	27.37	33.38	0.24
Starling-RM-7B-Alpha	59.22	10.16	1.05	100.00	35.79	47.52	0.21
NaiveVerbosityModel	58.25	11.49	2.63	100.00	20.00	32.78	0.22
Nemotron-4-340B-Reward	58.25	7.87	3.16	100.00	40.00	55.94	0.20
Gemini-1.5-Pro-001 (ArenaHard)†	57.58	-1.56	4.21	100.00	48.42	66.77	0.18
Gemini-1.5-Flash-002 (ArenaHard)†	51.96	-0.90	1.05	78.95	37.89	62.11	0.19

Table A.17: Reward model and LLM judge performance on Portuguese prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

Reward Model	Accuracy	R.W. Pearson	Separability	Conf. Agree.	Kendalltau	Spearmanr	Brier Score
Gemini-1.5-Pro-002 (AlpacaEval)†	81.40	51.04	3.16	100.00	50.53	74.14	0.17
Ensemble-Judges (AlpacaEval)†	75.58	44.04	6.84	100.00	45.26	66.47	0.18
Gemini-1.5-Pro-002 (ArenaHard)†	74.42	40.23	3.16	57.89	52.63	70.83	0.18
Athene-RM-70B	74.42	42.65	4.74	100.00	43.16	61.05	0.20
Claude-3-5-Sonnet-20240620 (ArenaHard)†	73.26	42.33	1.58	100.00	47.37	58.80	0.20
Athene-RM-8B	73.26	43.29	8.42	78.95	43.16	60.45	0.19
Ensemble-Judges (ArenaHard)†	71.25	44.59	1.58	89.47	36.84	51.88	0.20
Claude-3-5-Sonnet-20240620 (AlpacaEval)†	69.77	28.35	5.79	100.00	40.00	52.03	0.22
Gemini-1.5-Pro-001 (ArenaHard)†	69.23	35.18	2.63	100.00	40.00	55.94	0.19
GPT-4o-2024-08-06 (AlpacaEval)†	68.60	39.33	5.79	100.00	40.00	53.53	0.19
Eurus-RM-7B	67.44	25.34	2.63	89.47	-2.11	-1.95	0.29
Skywork-Reward-Llama-3.1-8B	66.28	27.43	1.58	100.00	37.89	47.82	0.21
ArmoRM-Llama3-8B-v0.1	66.28	28.46	5.79	100.00	42.11	57.14	0.19
Gemini-1.5-Flash-002 (AlpacaEval)†	66.28	33.17	1.05	89.47	30.53	44.81	0.22
GPT-4o-2024-08-06 (ArenaHard)†	66.25	39.65	6.32	100.00	34.74	51.88	0.20
GPT-4o-Mini-2024-07-18 (ArenaHard)†	64.71	31.59	1.05	100.00	34.74	55.64	0.20
Llama-3.1-70B-Instruct (ArenaHard)†	64.63	27.88	1.58	89.47	38.95	54.59	0.20
InternLM2-7B-Reward	63.95	26.87	3.16	36.84	12.63	15.49	0.25
InternLM2-20B-Reward	63.95	19.03	0.00	100.00	29.47	46.32	0.20
Gemini-1.5-Flash-002 (ArenaHard)†	63.10	24.42	4.21	89.47	27.37	44.96	0.22
Starling-RM-34B	62.79	13.29	1.58	100.00	10.53	10.23	0.28
Skywork-Reward-Gemma-2-27B	61.63	19.87	0.00	100.00	41.05	56.84	0.21
Llama-3.1-70B-Instruct (AlpacaEval)†	61.63	19.26	2.11	100.00	16.84	21.50	0.24
Nemotron-4-340B-Reward	60.47	19.10	13.16	5.26	53.68	75.34	0.18
InternLM2-1.8B-Reward	59.30	16.29	0.53	89.47	2.11	0.00	0.27
GPT-4o-Mini-2024-07-18 (AlpacaEval)†	58.14	14.03	1.05	100.00	24.21	33.98	0.23
Llama-3-OffsetBias-RM-8B	58.14	2.76	1.05	100.00	45.26	61.95	0.20
Starling-RM-7B-Alpha	56.98	12.63	3.68	89.47	2.11	-2.86	0.30
NaiveVerbosityModel	50.00	-0.20	2.63	100.00	-7.37	-13.68	0.31

Table A.18: Reward model and LLM judge performance on Italian prompt subset of the human preference dataset. LLM-as-a-judge are labeled with system prompt source, and marked with †.

A.2 Detailed Scores for the Correctness Preference Evaluation Dataset

Reward Model	gemma-2-9b-it			gpt-4o-mini			Llama-3-8B			claude-3-haiku		
	Loss	Max	End	Loss	Max	End	Loss	Max	End	Loss	Max	End
athene-rm-70b	0.093	0.702	0.681	0.110	0.678	0.629	0.113	0.669	0.653	0.131	0.633	0.605
armorm-llama3-8b-v0.1	0.119	0.657	0.636	0.147	0.620	0.580	0.179	0.576	0.537	0.194	0.564	0.512
naiveverbositymodel	0.241	0.508	0.463	0.250	0.554	0.425	0.358	0.448	0.317	0.337	0.467	0.355
eurus-rm-7b	0.143	0.627	0.597	0.158	0.613	0.562	0.187	0.562	0.512	0.228	0.531	0.452
skywork-reward-gemma-2-27b	0.169	0.583	0.543	0.175	0.590	0.549	0.209	0.534	0.494	0.190	0.558	0.529
skywork-reward-llama-3.1-8b	0.126	0.643	0.612	0.136	0.633	0.597	0.189	0.565	0.527	0.216	0.561	0.491
llama-3-offsetbias-rm-8b	0.133	0.653	0.629	0.146	0.629	0.585	0.210	0.542	0.502	0.151	0.620	0.592
nemotron-4-340b-reward	0.129	0.641	0.617	0.128	0.644	0.618	0.159	0.610	0.583	0.232	0.565	0.485
starling-rm-34b	0.157	0.602	0.570	0.151	0.622	0.563	0.183	0.562	0.528	0.209	0.545	0.487
athene-rm-8b	0.142	0.621	0.584	0.133	0.636	0.600	0.175	0.589	0.543	0.183	0.560	0.531
internlm2-7b-reward	0.138	0.630	0.588	0.147	0.633	0.581	0.155	0.608	0.581	0.253	0.565	0.462
starling-rm-7b-alpha	0.183	0.569	0.535	0.199	0.578	0.516	0.238	0.508	0.476	0.319	0.486	0.378
internlm2-1-8b-reward	0.193	0.566	0.501	0.191	0.583	0.506	0.218	0.526	0.480	0.256	0.503	0.448
internlm2-20b-reward	0.124	0.648	0.626	0.130	0.646	0.607	0.159	0.602	0.570	0.166	0.586	0.570

Table A.19: Average Best of K per Sample Model across MMLU Pro, Math, GPQA, MBPP Plus, and IF Eval

Reward Model	gemma-2-9b-it	gpt-4o-mini	Llama-3-8B	claude-3-haiku
athene-rm-70b	0.710	0.648	0.710	0.674
armorm-llama3-8b-v0.1	0.655	0.577	0.616	0.591
naiveverbositymodel	0.515	0.491	0.487	0.433
eurus-rm-7b	0.620	0.546	0.621	0.562
skywork-reward-gemma-2-27b	0.553	0.519	0.562	0.550
skywork-reward-llama-3.1-8b	0.639	0.594	0.619	0.578
llama-3-offsetbias-rm-8b	0.628	0.574	0.583	0.650
nemotron-4-340b-reward	0.639	0.586	0.658	0.561
starling-rm-34b	0.602	0.571	0.604	0.574
athene-rm-8b	0.640	0.592	0.635	0.601
internlm2-7b-reward	0.657	0.573	0.655	0.569
starling-rm-7b-alpha	0.544	0.499	0.525	0.475
internlm2-1-8b-reward	0.581	0.536	0.570	0.504
internlm2-20b-reward	0.629	0.603	0.650	0.603

Table A.20: Average AUC per sample model across MMLU Pro, Math, GPQA, MBPP Plus, and IF Eval

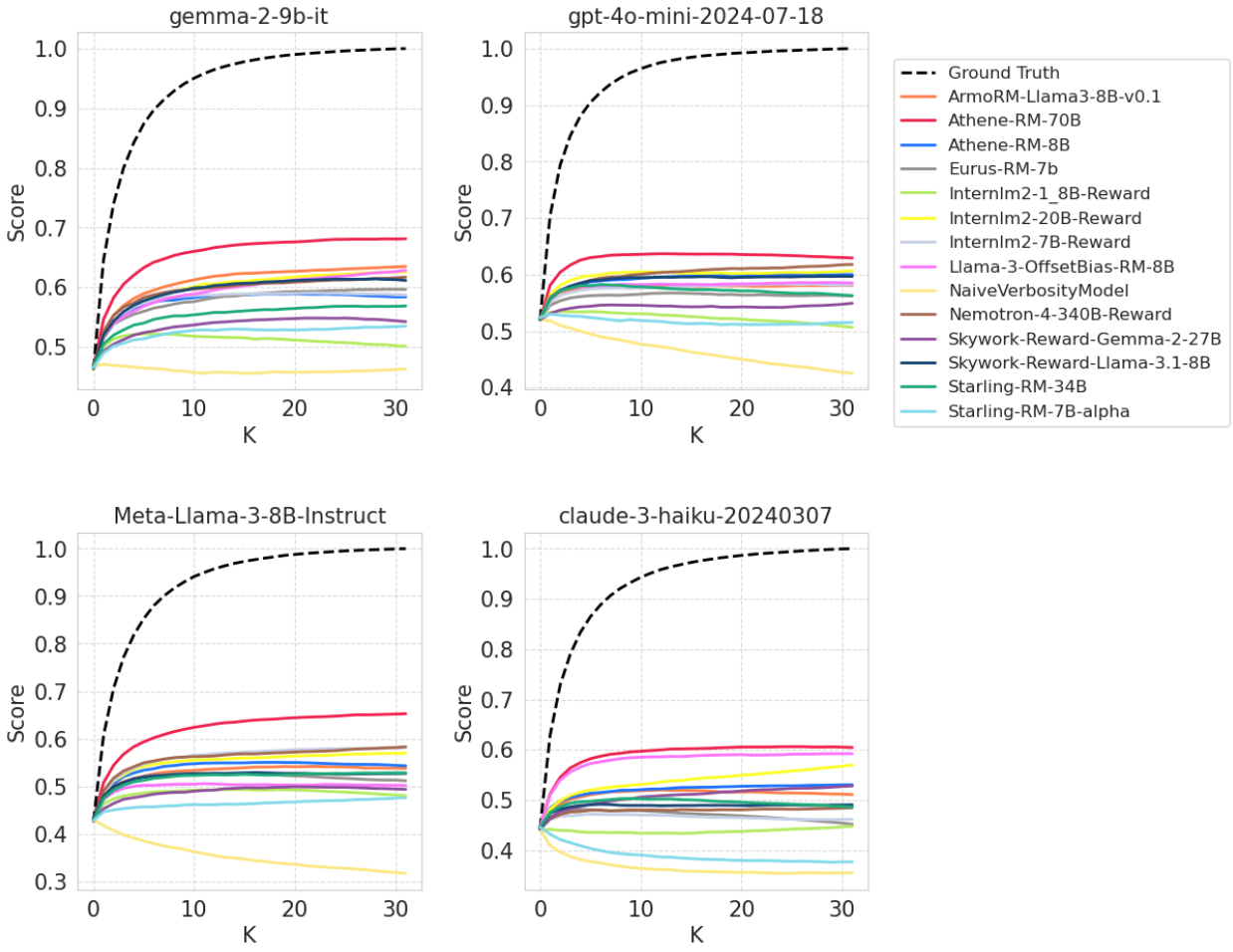

Figure A.1: Performance average across all benchmarks, conditioned on each sample model

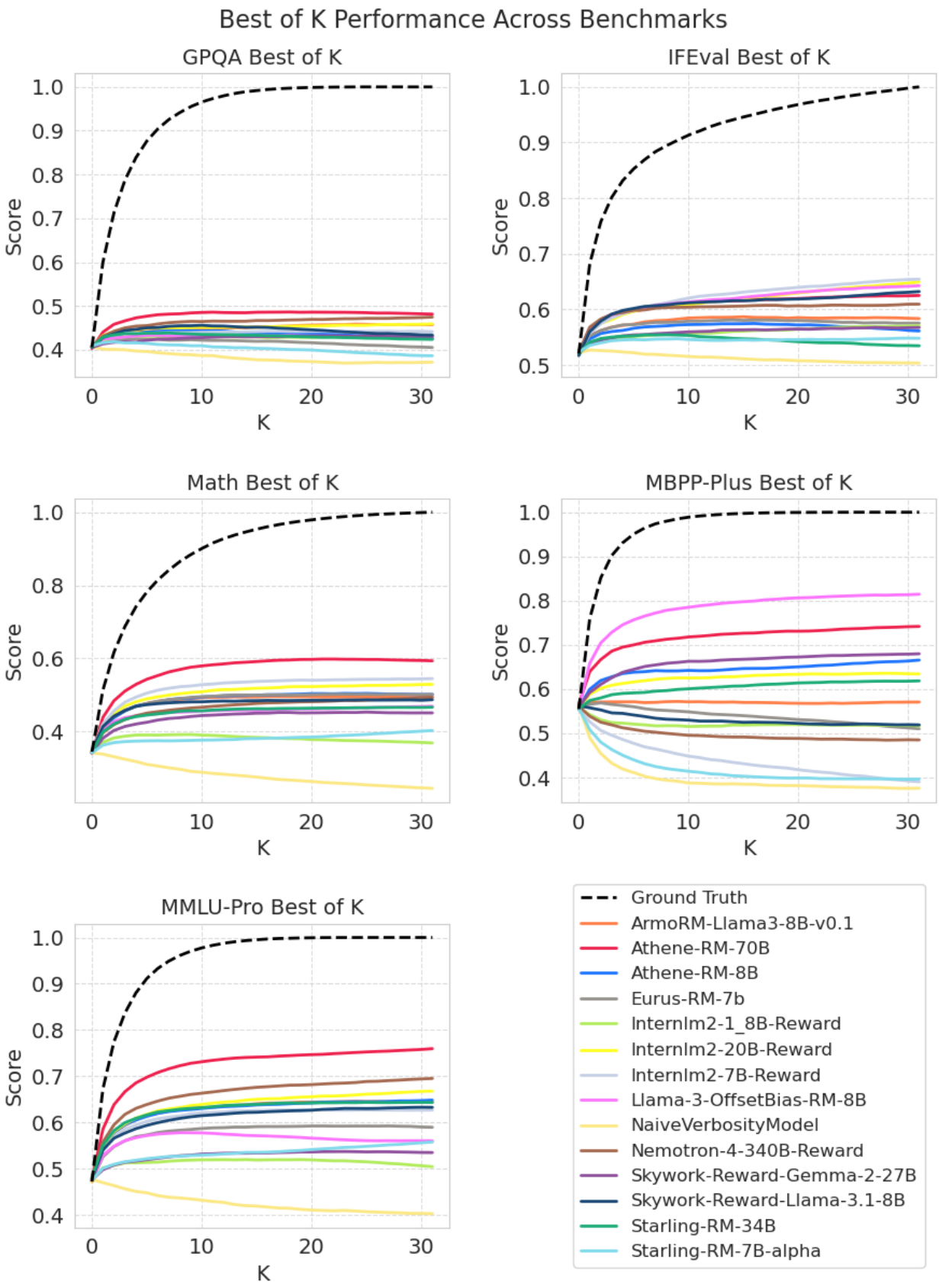

Figure A.2: Performance comparison across all benchmarks

Appendix B

Appendix for Validating PPE on Post-RLHF Outcomes

B.1 DPO Configuration

DPO Configuration	
Base Model	Meta-Llama-3.1-8B-Instruct
τ	0.1
Learning Rate	$2.00 \times 10^{-0.6}$
LR Schedule	Constant
Global Batch Size	64
Max Length	8192
Max Prompt Length	4096
Implementation	TRL DPOTrainer [45]
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Space Optimization	Deepspeed Zero2

B.2 Comments on RewardBench Correlations

Commenting on Figure 3.2; while our work's focus was not to prove or disprove RewardBench, we can provide the following hypothesis for context and clarity: we hypothesize that the reward models tested may have over-optimized for RewardBench's specific preference distribution rather than capturing broader human preferences, potentially exceeding RewardBench's measurement capabilities. However, we note that initial improvements in RewardBench score may still correlate well to real post RLHF human preference outcomes. Ultimately, these insights are only possible through our end-to-end experiments, which enable the research community to further investigate and discuss the true correlations between

benchmark metrics and downstream performance. We believe this highlights the value of comprehensive evaluation approaches like ours in understanding real-world model behaviors.

B.3 Style-Controlled Downstream Performance

Model	Elo	95% CI Lower	95% CI Upper
Meta-Llama-3.1-70B-Instruct*	1229	1218	1239
Athene-RM-70B	**1209**	1201	1218
Athene-RM-8B	1203	1194	1211
internlm2-7b-reward	1201	1192	1210
Llama-3-OffsetBias-RM-8B	1197	1188	1204
ArmoRM-Llama3-8B-v0.1	1185	1175	1191
Meta-Llama-3.1-8B-Instruct*	1177	1168	1186
Skywork-Reward-Llama-3.1-8B	1171	1163	1182
Nemotron-4-340B-Reward	1170	1161	1180
internlm2-20b-reward	1170	1159	1179
Skywork-Reward-Gemma-2-27B	1170	1160	1180
Meta-Llama-3-8B-Instruct*	1152	1142	1160

Table B.1: Post DPO performance on real human preference Overall Category after applying style-control. "Model" is the reward model used to train the base model. Models marked with "*" are baseline unaltered models. The best non-base model elo is bolded.

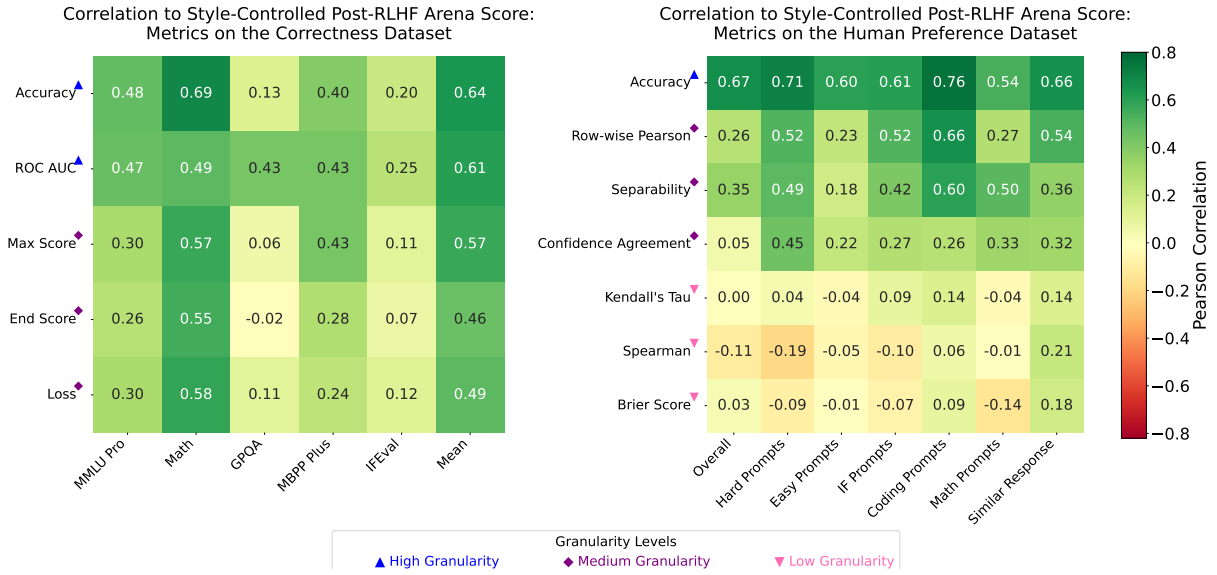

Figure B.1: Pearson correlations between various metrics and styled-controlled human preference scores. Left: Correlations between metrics on the Correctness Dataset and Post-RLHF human preference rating. Right: Correlations between metrics on the Human Preference Dataset and Post-RLHF human preference rating.

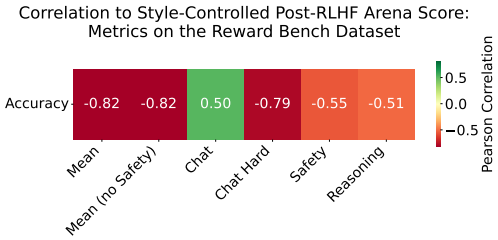

Figure B.2: Pearson correlation between the ranking of models in RewardBench and their respective style-controlled Post-DPO rankings on real human preference.

As an ablation, we calculate style-controlled human preference ratings. Style-controlled ratings fit the Bradley Terry model with style elements as features of the regression. These features are used to decouple style from model ratings; this process yields score estimates, style *aside*. The full process for style control is detailed in [20]. For maximum coverage, we control for length and markdown.

B.4 Correlation vs. K

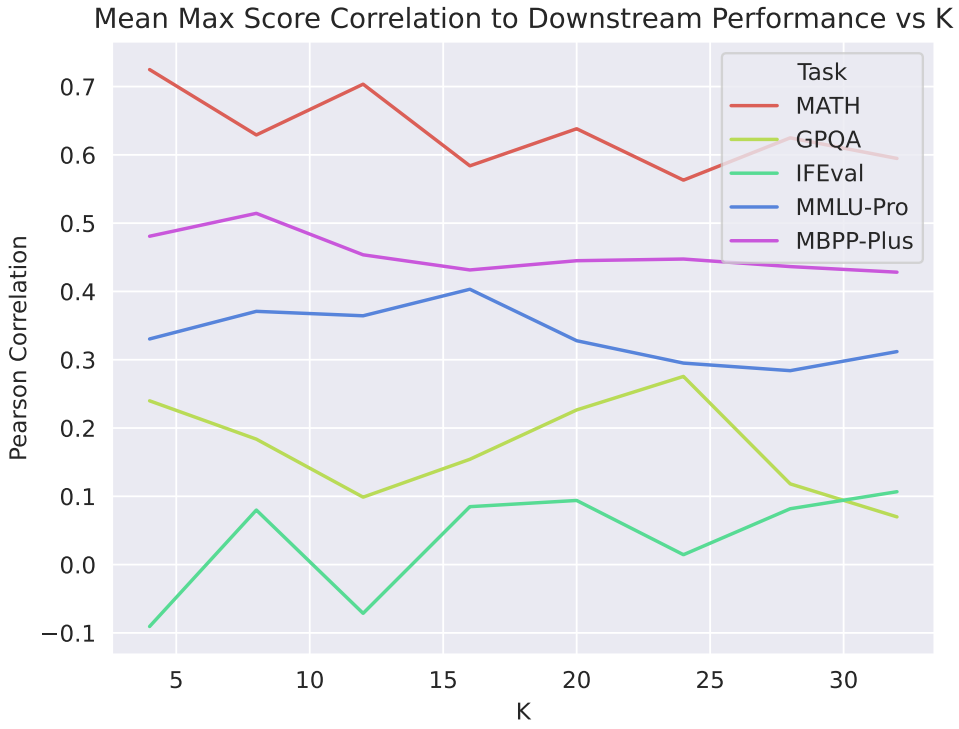

Figure B.3: Pearson correlation to downstream human preference performance of mean max score best of K metric vs K.

Figure B.3 shows that increasing the value of K for best of K metrics does not increase benchmark predictive power. We note that the most predictive correctness metrics is the accuracy metric detailed in Section 2.4 which is inherently $K = 2$. Therefore, the predictive power of PPE can be retained without running full $K = 32$, which is more compute heavy.

B.5 Recommendations for PPE and Future Reward Model Benchmarks

Based on this end-to-end study results detailed in Section 3.2 and Appendix Figure B.3, we recommend those seeking the most predictive power from PPE run the human preference set as well as the MATH accuracy metric. We suggest that users pay particular attention to the lower bound accuracy across the main human preference set categories (easy, hard,

instruction following, coding, math, and similar). Considering our findings, this configuration likely maintains full predictive power of PPE with less than half of the runtime. Future reward benchmarks may find it helpful to attend to these particular design patterns.

B.6 Runtimes and Costs for PPE

Benchmark Set	Time	Cost
Optimized (Human Preference V1 + Math Accuracy)	$<$ 42 minutes	$<$ \$1.50
Full Benchmark	$<$ 120 minutes	$<$ \$3.50
End-to-end RLHF pipeline	$>$ 1 week	\$1000 or more

Table B.2: Benchmark runtimes and costs. Costs are calculated from RunPod's hourly GPU pricing, which puts an NVIDIA A100 80GB PCIe instance at \$1.64 per hour. Costs could fluctuate between GPU providers. Runtimes are estimated assuming an 8B reward model.

Appendix C

Appendix for Towards Robust Reward Models

C.1 Human Preference Test Set Loss Curve

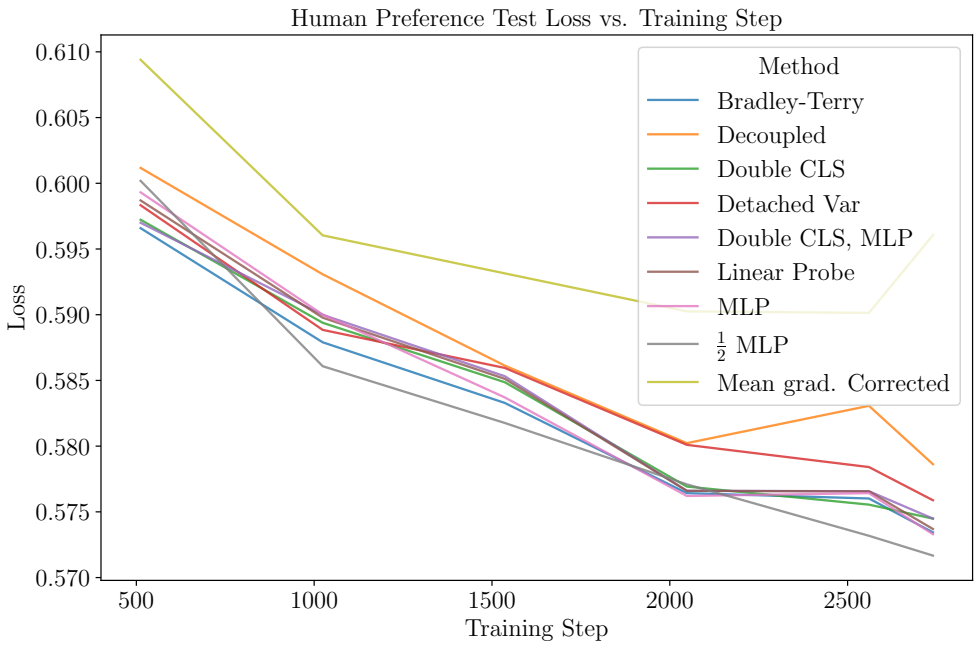

Figure C.1: Loss on the human preference hold-out set vs. training step. Models shown are 1.5B parameters. All models are Thurstonian unless indicated.

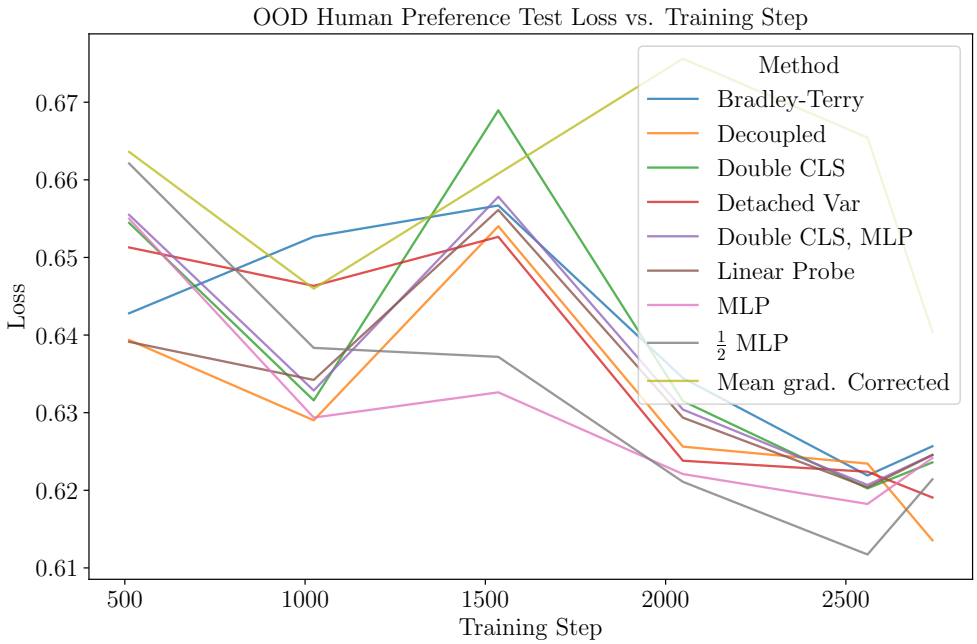

Figure C.2: Loss on the human preference out-of-distribution set vs. training step. Models shown are 1.5B parameters. All models are Thurstonian unless indicated.

Figure C.3: Loss on the human preference hold-out set vs. training step. Models shown are 7B parameters. All models are Thurstonian unless indicated.

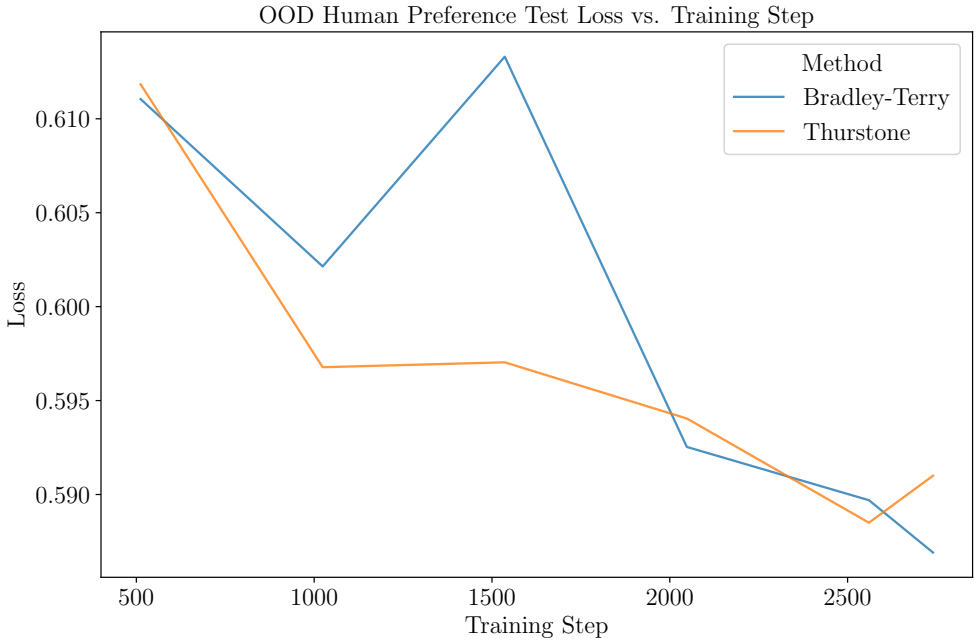

Figure C.4: Loss on the human preference out-of-distribution set vs. training step. Models shown are 7B parameters. All models use a linear probe head.

C.2 Additional 7B Performance

Model Size	Type	Architecture	Other	Human Pref.	OOD Human Pref.	MMLU Pro	MATH	MBPP+	IFEval	GPQA	Mean
7B	Thurs	Linear Probe	N/A	**72.406**	69.045	70.938	**78.516**	**63.511**	**61.797**	59.063	**67.896**
7B	BT	Linear Probe	N/A	72.249	**69.438**	**71.914**	78.086	59.921	61.758	**59.688**	67.579

Table C.1: Accuracies of all trained 7B reward models models on PPE benchmarks. The models are sorted by their mean score across all benchmarks. The scores are in percentages.

Model Size	Type	Architecture	Other	Overall	Hard Prompt	Easy Prompt	If Prompt	Is Code	Math Prompt	Mean
7B	BT	Linear Probe	N/A	**69.438**	**71.618**	**67.629**	**71.188**	68.515	**71.538**	**69.988**
7B	Thurs	Linear Probe	N/A	69.045	70.999	66.907	70.551	**68.614**	70.769	69.481

Table C.2: Accuracies of all trained 7B reward models on the OOD human preference test set. The categories are derived from Chatbot Arena's category definitions [9]. The models are sorted by their mean score across all categories.

Model Size	Type	Architecture	Other	MMLU Pro	MATH	MBPP+	IFEval	GPQA	Mean
7B	Thurs	Linear Probe	N/A	66.811	**62.645**	**67.479**	63.496	48.148	**61.716**
7B	BT	Linear Probe	N/A	**67.095**	60.847	63.470	**63.921**	**48.651**	60.797

Table C.3: Average Best-of-32 score of all trained 7B reward models on the PPE verifiable benchmark sets. The models are sorted by their mean score across all benchmarks.

C.3 Additional Quantile Results

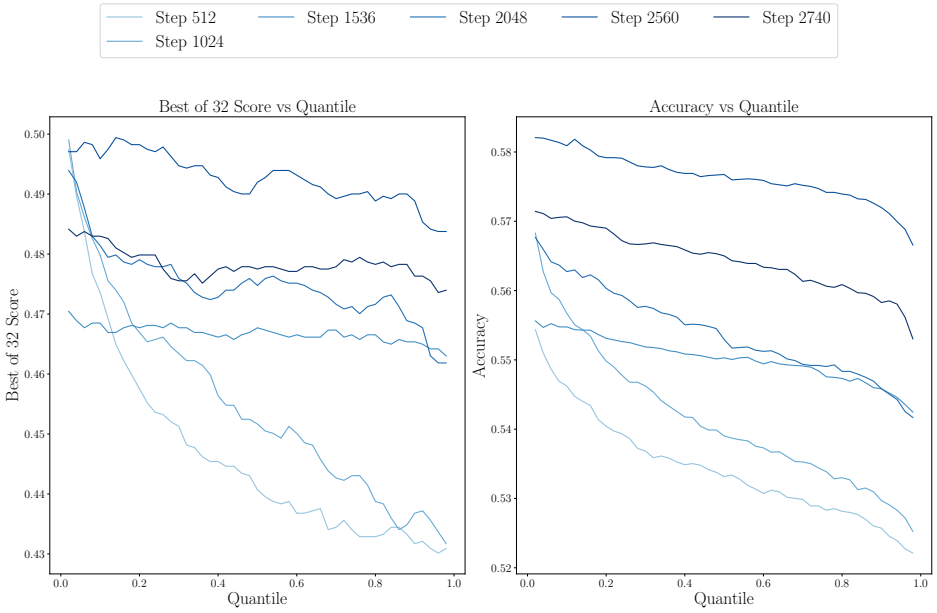

Figure C.5: Best of 32 score vs Quantile and Accuracy vs Quantile, shown for each checkpoint step during training. The Thurstonian reward model has 1.5B parameters and uses a decoupled architecture.

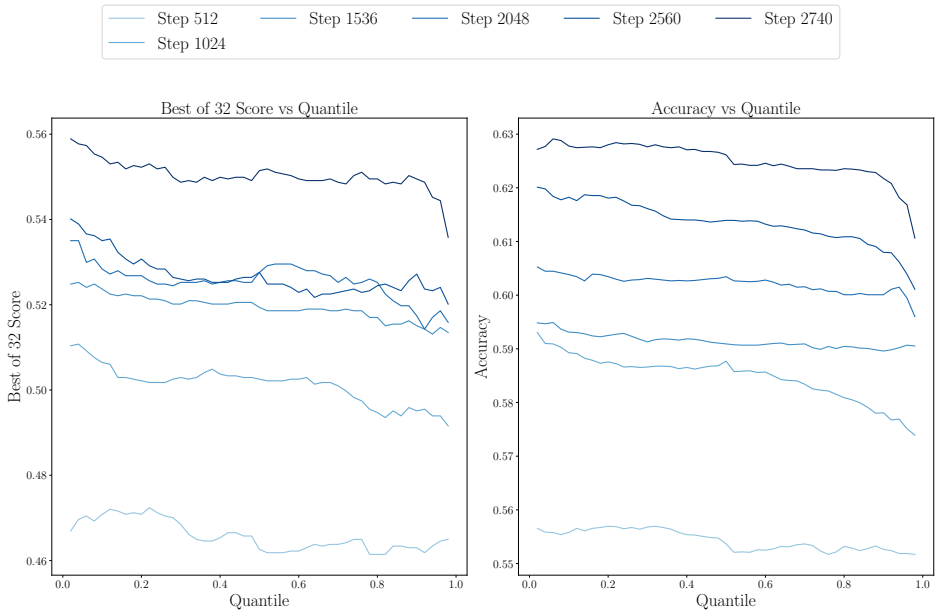

Figure C.6: Best of 32 score vs Quantile and Accuracy vs Quantile, shown for each checkpoint step during training. The Thurstonian reward model has 1.5B parameters and uses a double CLS head.

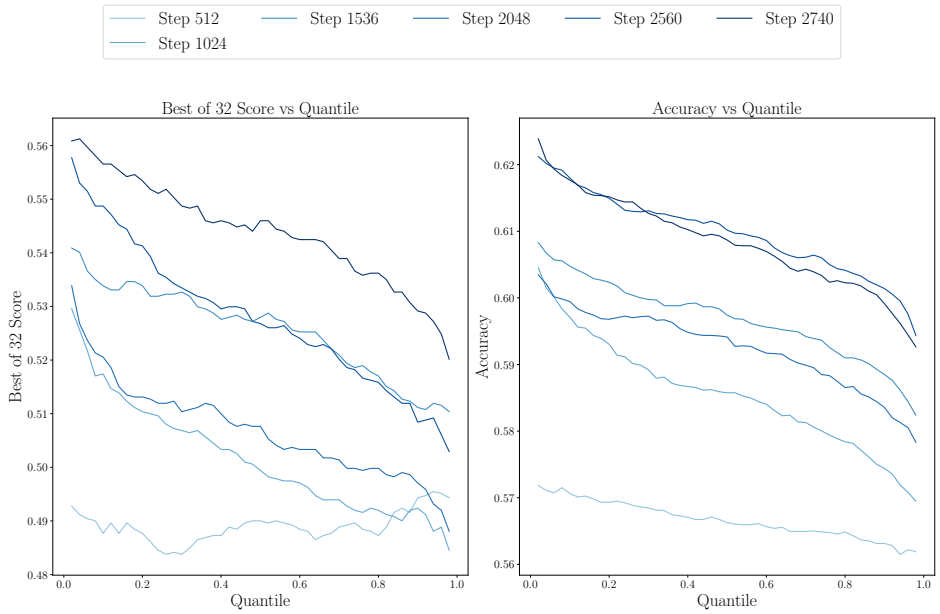

Figure C.7: Best of 32 score vs Quantile and Accuracy vs Quantile. The Thurstonian reward model has 1.5B parameters with variance predicted from a detached head.

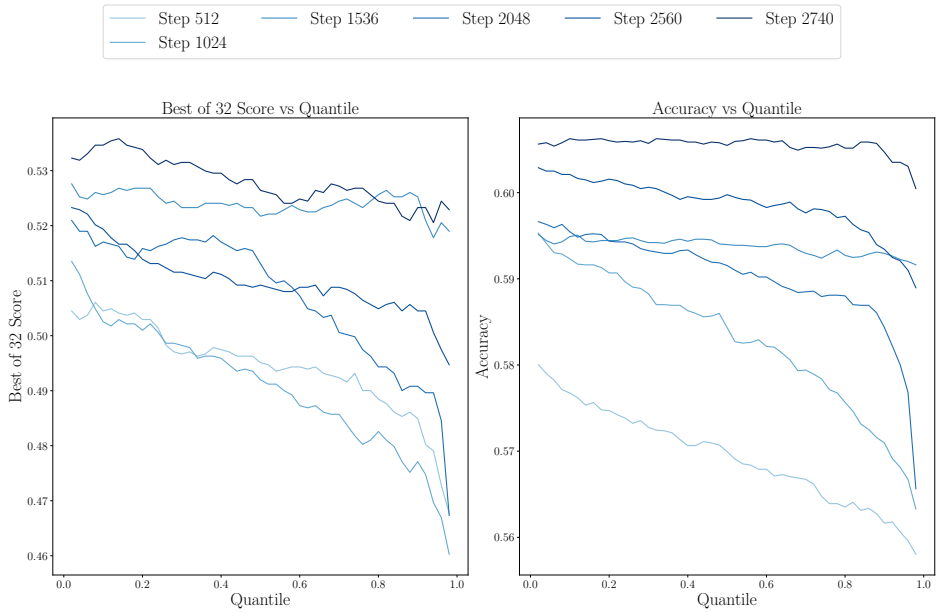

Figure C.8: Best of 32 score vs Quantile and Accuracy vs Quantile. Thurstonian reward model with 1.5B parameters and a linear probe.

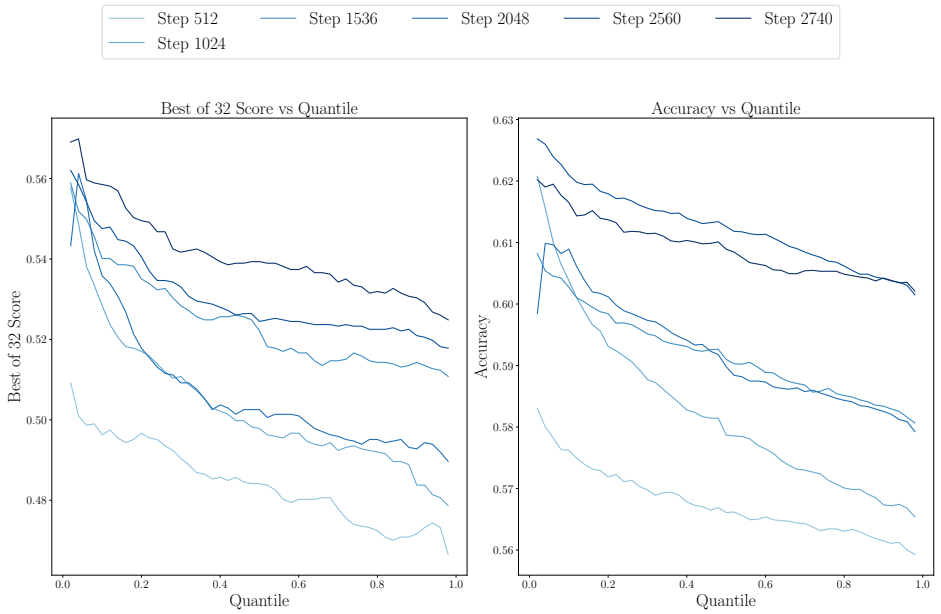

Figure C.9: Best of 32 score vs Quantile and Accuracy vs Quantile. Baseline Thurstonian reward model with 1.5B parameters and an MLP head ($k = 4$).

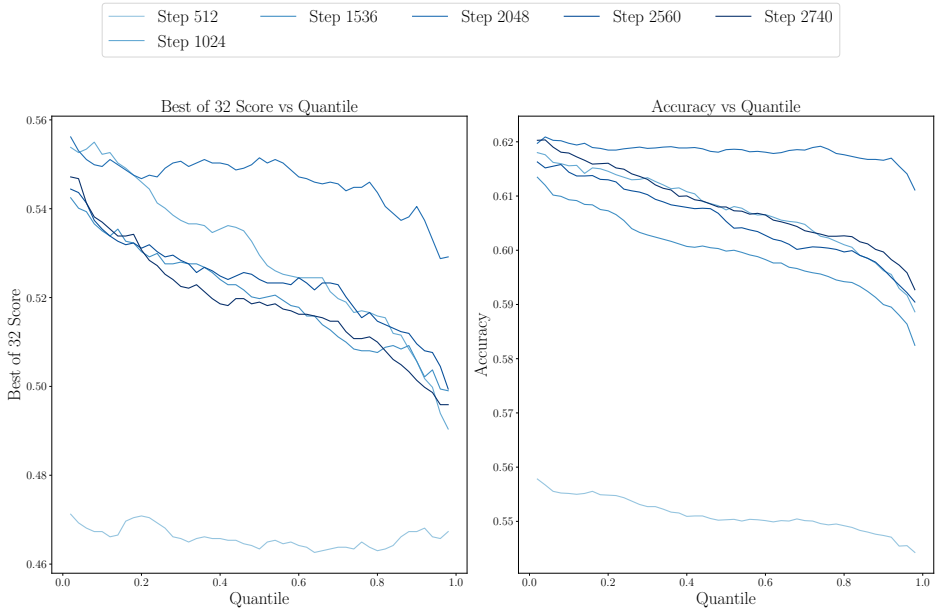

Figure C.10: Best of 32 score vs Quantile and Accuracy vs Quantile. Baseline Thurstonian reward model with 1.5B parameters and an MLP head ($k = 2$).

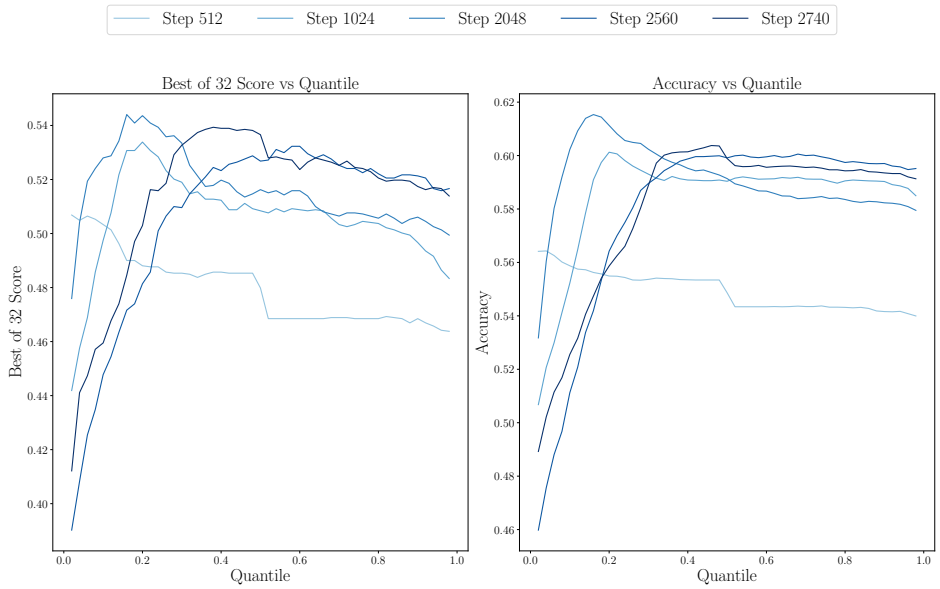

Figure C.11: Best of 32 score vs Quantile and Accuracy vs Quantile. Thurstonian reward model with 1.5B parameters trained with rescaled mean gradients.

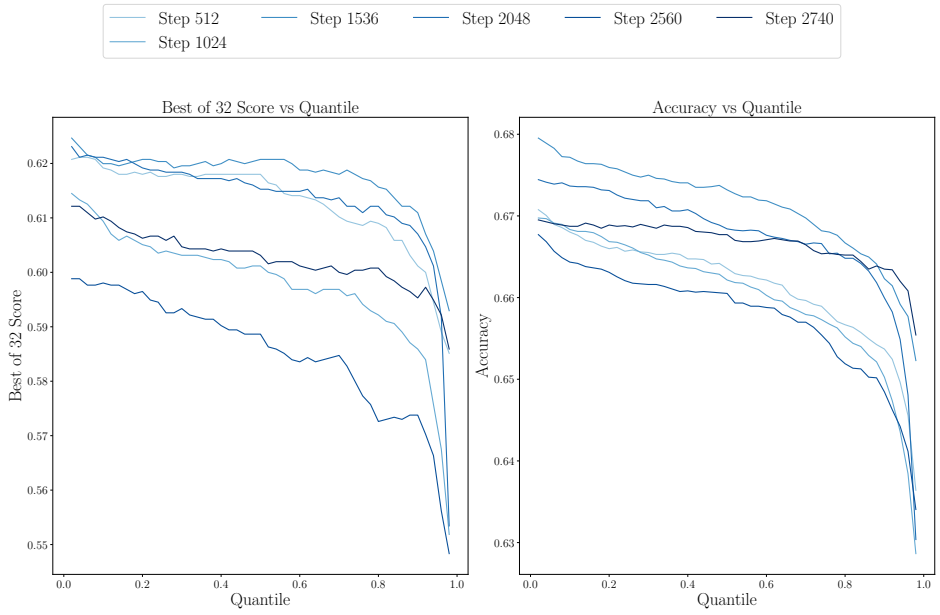

Figure C.12: Best of 32 score vs Quantile and Accuracy vs Quantile. Thurstonian reward model with 7B parameters with a linear probe head.

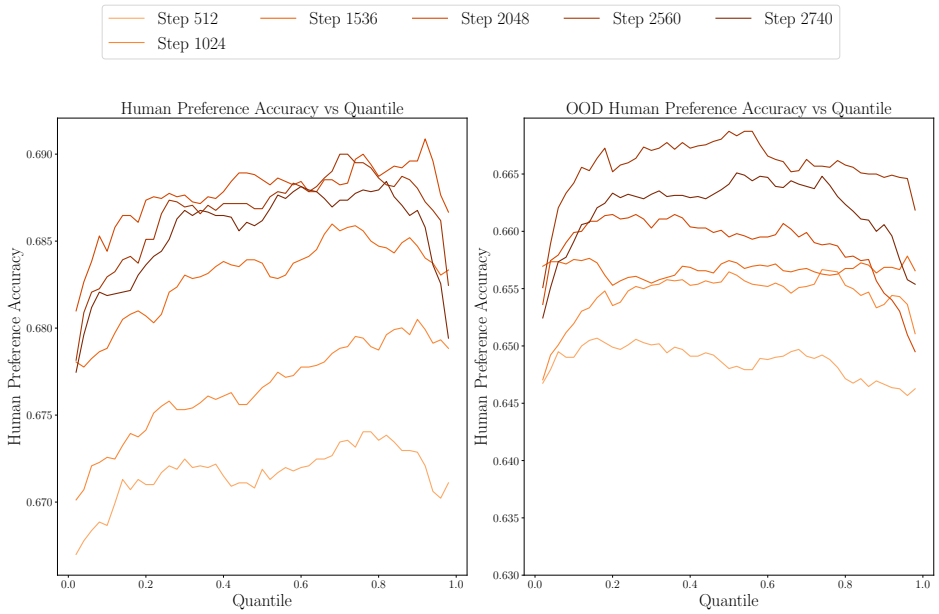

Figure C.13: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure uses the fully decoupled architecture.

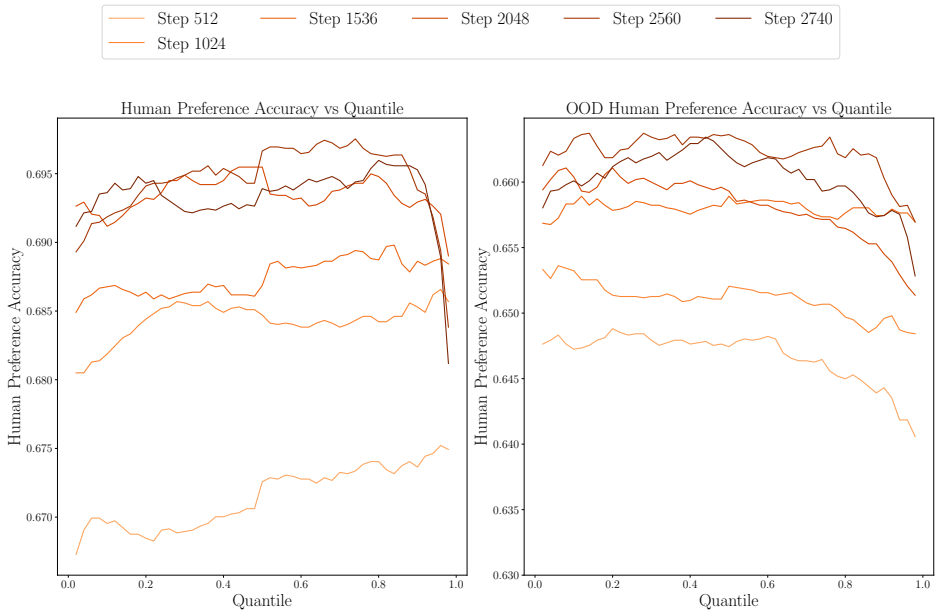

Figure C.14: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure has 1.5B parameters and uses the double CLS architecture.

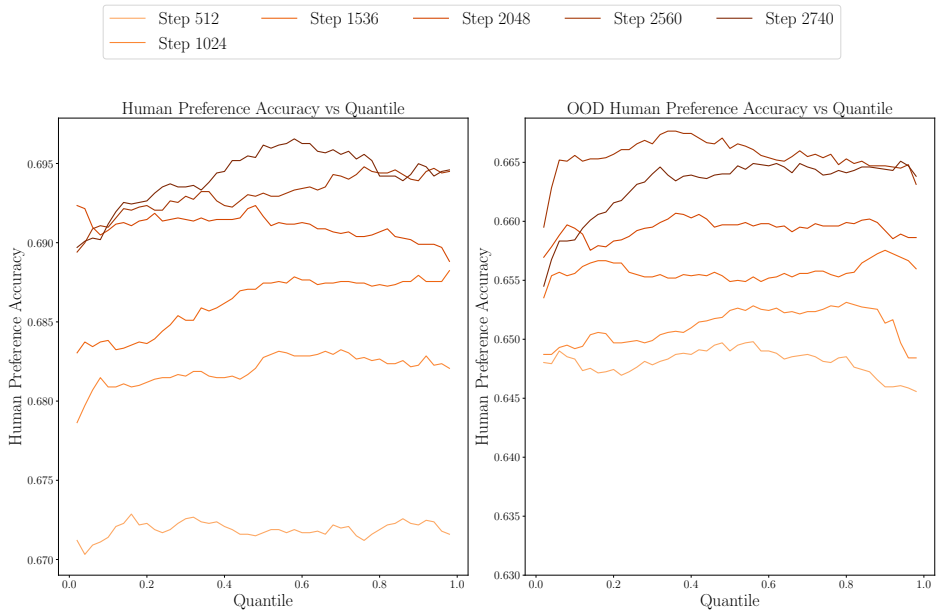

Figure C.15: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure has 1.5B parameters and uses the detached variance training procedure.

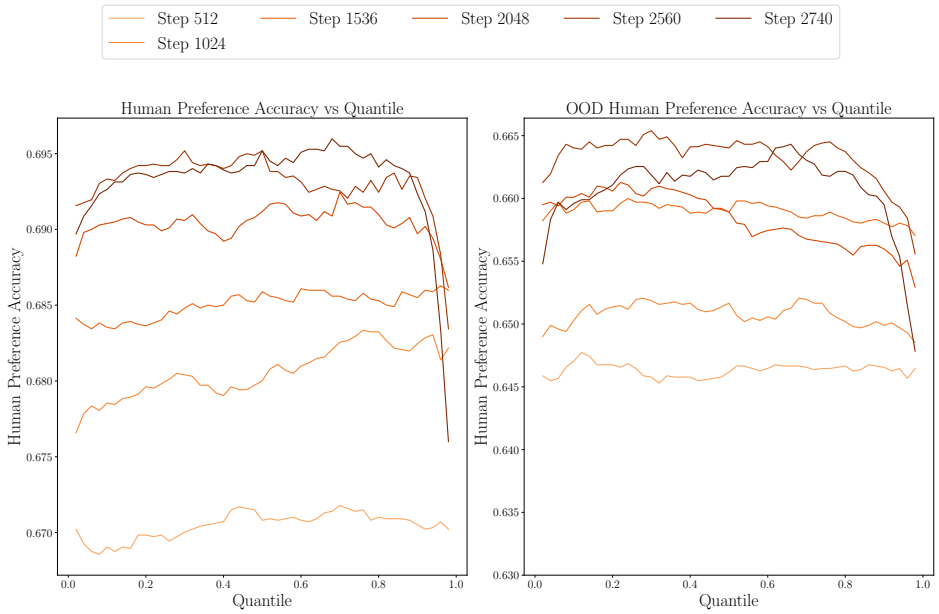

Figure C.16: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure has 1.5B parameters and uses the double CLS with MLP architecture.

Figure C.17: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure has 1.5B parameters and uses a linear probe head architecture.

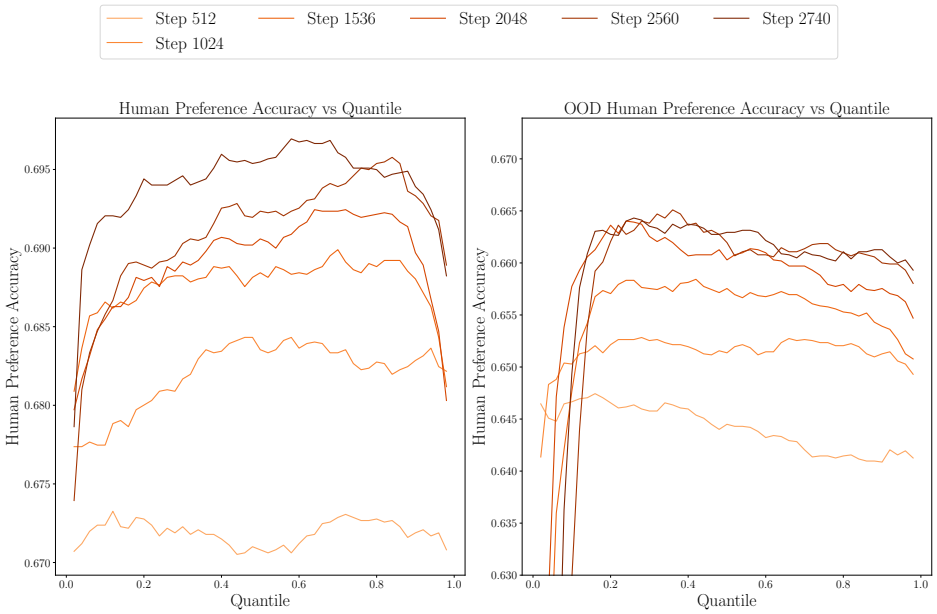

Figure C.18: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure has 1.5B parameters and uses the MLP head architecture.

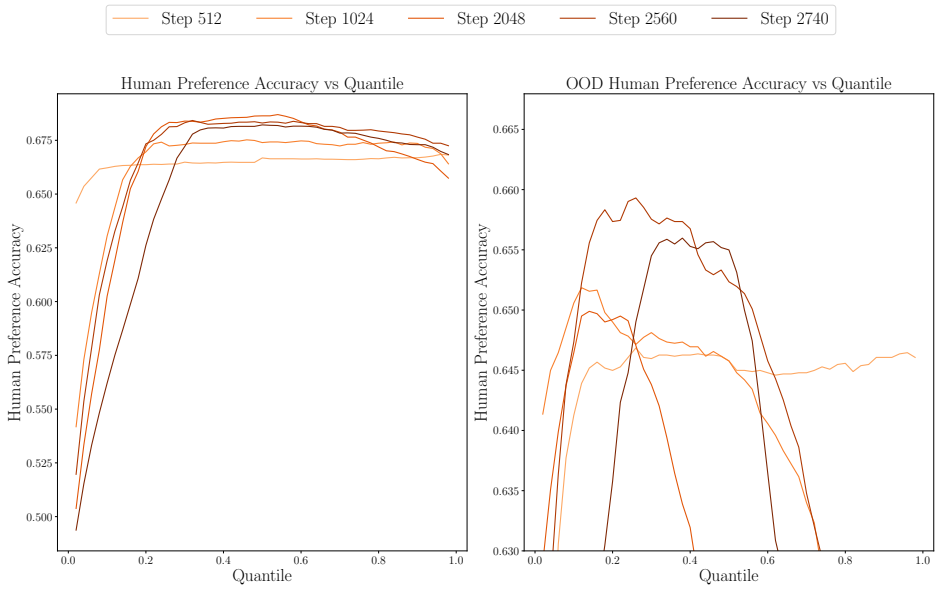

Figure C.19: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure uses the scaled mean gradient training procedure.

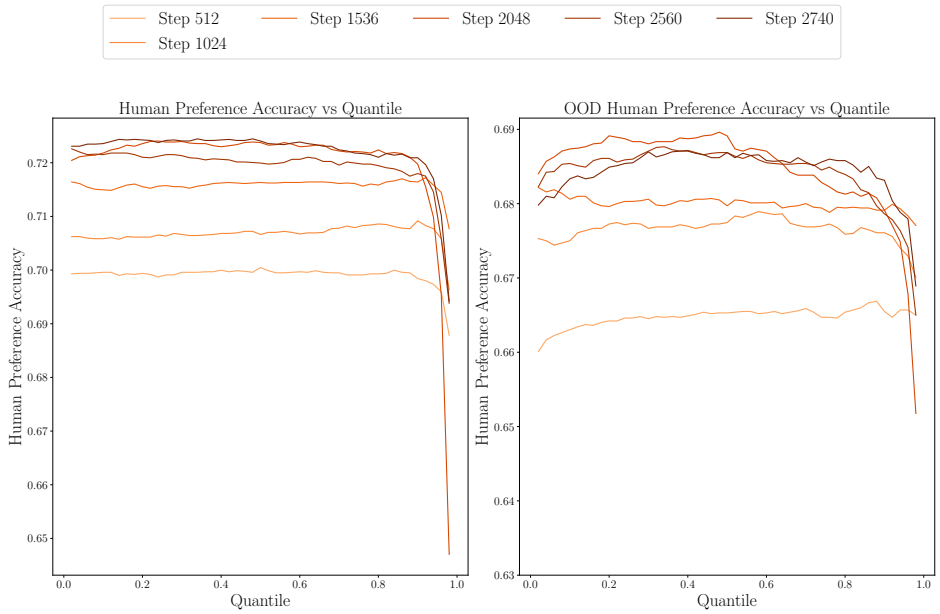

Figure C.20: Reward Quantile vs Human preference accuracy, both in and out-of-distribution. The Thurstonian reward model in the figure has 7B parameters and uses the linear probe head architecture.