RMA: Reward Model Alignment with Human preference

Ashish Gupta [1] **Manjunatha Naik MC** [1]

Abstract

Reward models (RMs) are essential for aligning large language models (LLMs) with human preferences. These models are typically trained on datasets containing an input prompt, two model-generated responses, and a preference label indicating which response is preferred. However, current approaches often suffer from limited generalization, exhibiting inconsistent performance across different contexts and displaying biases such as position bias (favoring the first response), verbosity bias (preferring longer outputs), or self-enhancement bias (favoring self-reinforcing statements).

In this work, we propose Preference Prediction, a novel framework that leverages high-quality preference data validated by human annotators along with open source data, combined with a preference selector trained via supervised fine-tuning (SFT), to dynamically choose the most suitable model for a given context. Through comprehensive experiments on a variety of datasets, we show that our proposed Reward Model Alignment (RMA) not only surpasses existing reward models in performance but also significantly boosts the effectiveness of other distinct reward models when applied to synthetic data. Additionally, RMA promotes the generation of more diverse and high-quality responses by integrating multiple quality dimensions—such as helpfulness, relevance, and completeness—into the prompting process.

1. Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Jiang et al., 2024; Dubey et al., 2024) have had a transformative impact on how society perceives the capabilities of AI systems in language understanding and its application to human languages. They are optimized using likelihood maximization techniques, enabling them to perform a wide range of tasks in response to user instructions.

Human preference alignment (Lee et al., 2023; Marta et al., 2023) plays a crucial role in better aligning AI-generated outputs with human intentions (Ji et al., 2023), by fine-tuning LLMs to generate responses that reflect human judgment. In this work, we evaluate which LLM performs best for specific tasks based on their responsiveness to human instructions. While many existing approaches rely on reinforcement learning (RL), such methods are not universally applicable across all LLMs and often lack insights into non-RL alternatives (Kaufmann et al., 2023). Several studies focusing on large models and their alignment with human preferences (Shen et al., 2023a; Wang et al., 2023) have been extensively explored. Our approach aligns with the principles of "reward models" or "preference models" used in Reinforcement Learning from Human Feedback (RLHF), where models are rewarded based on their alignment with human judgments.

One key application of our work is in model routing —the task of selecting the most suitable model for a given prompt in a cost-effective manner. For instance, if a prompt can be effectively handled by an open-source model like Llama-3.2 instead of a more expensive alternative like GPT-4, we can optimize both performance and cost. This has direct implications for managing computational budgets when deploying LLMs at scale.

We investigate strategies to enhance the alignment between AI-generated responses and human preferences. Our findings indicate that incorporating chain-of-thought reasoning (Wei et al., 2022; Ling et al., 2024) consistently improves alignment, whereas few-shot prompting yields benefits only in specific contexts. Additionally, we conduct scaling experiments to analyze the trade-off between model size and alignment accuracy.

The main contributions of this work are:

- We propose a machine learning (ML) framework designed to enhance the interaction between models and humans, promoting better alignment with human preferences.

[1]ServiceNow, Hyderabad, India. Correspondence to: Ashish Gupta <ashish.gupta1@servicenow.com>, Manjunatha Naik MC <manjunathanaik.mc@servicenow.com>.

ICML 2025 Workshop on Assessing World Models.

- We develop a predictive model to estimate human preferences and assess the likelihood that a given prompt-response pair will be selected as the preferred option by human judges.

2. Related Work

Reward models play a critical role in the advancement of large language models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023), yet collecting human preference data for training remains resource-intensive. To mitigate this, several studies have explored the generation of synthetic preference data using LLMs. One of the earliest methods, Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022a; Lee et al., 2023), leverages LLMs to assign preference scores to response pairs. The West-of-N approach (Pace et al., 2024) builds on this idea by enhancing reward models through selection of top and bottom responses from a pool of outputs to form preference pairs. Similarly, the ALMoST method (Kim et al., 2023) queries two LLMs of differing strengths, assuming the output from the more capable model is preferable.

Our work also intersects with preference learning in language models, particularly through the use of Reinforcement Learning from Human Feedback (RLHF) for natural language generation (NLG) tasks such as translation, review writing, summarization, and style transfer. In addition, we consider evaluation frameworks like ROUTERBENCH (Hu et al., 2024), a novel benchmark developed to systematically evaluate the effectiveness of LLM routing systems. ROUTERBENCH includes a diverse dataset comprising over 405k inference results from various LLMs to support robust routing strategies. However, it lacks sufficient coverage of queries involving unsafe, offensive, or sensitive content. As a result, relying solely on this framework does not adequately evaluate a router's performance on such critical categories.

3. Proposed Method

Our objective is, for a given prompt/response pair, to develop a model capable of effectively determining which LLM chat model produces the better response for a specific task.

Training Data Each training example is represented as a pair (x_i, y_i), where $i \in 1...N$. Here, x_i denotes a response generated by an LLM for a given prompt, and y_i is the corresponding binary label indicating whether the response is satisfactory (1) or not (0). These (x_i, y_i) pairs are used to train our classifier.

The training data is drawn from multiple open-source datasets (Zheng et al., 2023; Bartolome et al., 2023), including those with conversational data annotated with pairwise human preferences. These conversations typically span multiple turns, with the number of interactions varying across different prompts. Additional datasets were incorporated specifically for pseudo-labeling purposes.

To expand our training corpus, we generated pseudo-labels for 20,000 samples from the Ultrafeedback dataset (Cui et al., 2024), which explores whether DPO fine-tuning benefits from using more than one rejection per chosen response—particularly in improving performance on benchmarks such as AlpacaEval, MT-Bench, and LM Eval Harness.

For pseudo-labeling, we also utilized the **orpo-dpo-mix-40k** dataset[1] and an additional 20,000 samples generated by a diverse set of LLMs. These models were either known for strong performance in human preference prediction or appeared in the aforementioned datasets. Label generation emphasized criteria such as helpfulness and relevance. Further details can be found in Appendix A.

3.1. Pseudo Labeling

We generated approximately 20,000 pseudo-labeled samples from a diverse set of datasets, using the Llama3-70B model for labeling. To improve processing efficiency, we employed continuous batching based on the number of conversational turns rather than token length, which significantly increased throughput and prevented single-turn dialogues from being grouped with multi-turn conversations. To evaluate labeling accuracy, two annotators independently reviewed a 2% subset of the data. The inter-rater agreement, measured by Cohen's Kappa, was 0.94—indicating strong consistency. Moreover, the annotators' judgments closely aligned with the model-generated labels, confirming the reliability of the pseudo-labeled data for downstream training.

3.2. Preprocessing

Since the **orpo-dpo-mix-40k**1 dataset and other open-source datasets **QA Feedback**(Wu et al., 2023) were not in the required format, we converted them to the format specified in Table 1. We applied 5-fold stratified sampling, where four folds were used for training and the remaining fold for evaluation. The data was tokenized using various models, including Qwen2.5-7b (Yang et al., 2024) and distilled Llama3-70b (Shi et al., 2024). The maximum prompt length was set to 3072 tokens, with any excess context either truncated or ignored by the model. The final label was multi-class, where **A** indicated that response_a was the best response, **B** indicated that response_b was the best response.

[1] https://huggingface.co/datasets/mlabonne/orpo-dpo-mix-40k

prompt	$model_a$	$model_b$	$response_a$	$response_b$	$winner_a$	$winner_b$
["explain function calling. how would you call a function?"]	gpt-3.5-turbo-0613	mistral-medium	["Function calling is the process of invoking or executing a function in a programming language..."]	["Function calling is the process of invoking a function in a program. A function is a block of code.."]	0	0
["Estimate the number of dinosaur skeletons in the United States and explain your reasoning.", "Estimate the number of stop signs in America", "Estimate the number of lines in a codebase of a midsized company that hires software engineers", "Estimate the number of bullets fired in WWII"]	GPT-4-0314	Claude-1	["Estimating the exact number of dinosaur skeletons in the United States is challenging due to the continuous discoveries and excavations of new specimens. There are more than 1,500 accredited museums in the United States, and many of these museums have natural history collections that may include dinosaur skeletons or partial specimens....", "Estimating the number of stop signs in the United States is challenging and depends on various factors such as population density, road infrastructure, and urban planning regulations...."]	["Here is my rough reasoning: There are approximately 1,000 museums in the U.S. that contain dinosaur fossils or replicas. As a rough guess, if 25-50% of these museums have at least one full dinosaur skeleton, that would suggest there are between 250 to 500 full dinosaur skeletons. There are also many dinosaur fossils held in university collections that are not on public display....", "Estimating the number of stop signs in America is challenging. There are over 4 million miles of roads in the U.S., including federal highways, state highways...."]	0	1
["Dört basamaklı, rakamları birbirinden ve sıfırdan farklı olan ABCD sayısından, bu sayının tersi olan DCBA sayısı çıkartılıyor. Eğer elde edilen sonuç, bu sayının rakamlarının küplerinin toplamına eşit ise bu sayı kaçtır?"]	gpt-4-0125-preview	claude-3-5-sonnet-20240620	["Bu soruyu çözmek için, verilen koşulları adım adım uygulamamız gerekiyor. Dört basamaklı ve rakamları birbirinden farklı olan ve sıfırdan farklı bir... "]	["Bu problemi adım adım çözelim:1) ABCD - DCBA = $A^3 + B^3 + C^3 + D^3$ şeklinde ifade edebiliriz.2) ABCD - DCBA'yı ..."]	1	0

Table 1. Snapshot of training data

4. Methodology

We employed two approaches to identify the best-performing model response. The first is a basic correlation-based matching method that takes only the names of $model_a$ and $model_b$ as input. This approach uses the Bradley-Terry model to estimate scores for each model that maximize the likelihood of the observed preference outcomes, effectively framing model comparison as a maximum likelihood estimation problem. This method provides a probabilistic assessment of how likely one model is to be preferred over another, indicating its potential to "win" in a pairwise comparison. The experiment was conducted on the **Ultra-feedback** dataset (Cui et al., 2024).

In parallel, we developed a neural model that takes the full context—<prompt,$response_a$, $response_b$> as input and predicts whether $model_a$ or $model_b$ better aligns with human preferences. This binary classification model is based on a distilled version of Llama3-70b (Shi et al., 2024), and is primarily fine-tuned on the Ultrafeedback, QAFeedback datasets primarily. Additionally, we employ an auxiliary classifier, Qwen2.5-7b (Yang et al., 2024), which achieves strong performance across multiple tasks, often surpassing current state-of-the-art (SOTA) results. Fine-tuning is performed using Low-Rank Adaptation (LoRA) (Hu et al., 2021), with adapters applied on top of the base models to improve efficiency and performance.

Distillation of Llama 3-70b was done using three weighted losses as mentioned in Equation 1.

4.1. Post-pretrain

We trained distilled Llama3-70b (Shi et al., 2024) and Qwen2.5-7b (Yang et al., 2024) for one epoch on **Ultra-Feedback**(Cui et al., 2024), and **QA Feedback** dataset with a learning rate of 1e-5.

4.2. Fine-Tuning

We fine-tuned the distilled Llama3-70b model (Shi et al., 2024) and Qwen2.5-7b (Yang et al., 2024) on our additional pseudo-labeled training data, consisting of the **orpo-dpo-mix-40k** dataset1 and 20,000 extra samples described in 3.

Training was conducted on 8x NVIDIA A100 GPUs, each equipped with 80GB of memory. We employed bitsandbytes[2] with QLoRA (Dettmers et al., 2024), applying 4-bit quantization and bfloat16 precision (Burgess et al., 2019).

4.3. Loss Function

We fine-tuned the Large Language Model (LLM) using a supervised learning strategy. The overall loss function **L** is a linear combination of three components: cross-entropy loss (log-loss), Kullback–Leibler (KL) divergence loss, and cosine embedding loss. We combined these losses and averaged the resulting prediction probabilities:

$$L = L_{log-loss} + L_{KLDivLoss} + L_{cosine-emb-loss} \quad (1)$$

KL divergence was used to align the predicted distribution with the desired response distribution while preserving historical information. Cosine embedding loss was applied to capture semantic similarity between the input query and generated response.

Finally, we ensembled the Qwen2.5-7B and a distilled version of LLaMA3-70B models—collectively referred to as QweLl—by averaging their Low-Rank Adaptation (LoRA) layers across five cross-validation folds, as described in Section 3.2.

5. Experimental Setup

5.1. Dataset

The dataset((Cui et al., 2024), **orpo-dpo-mix-40k**1, psuedo-labeled) has multi-turn interactions based on the prompts provided. Figure1 shows the distribution of the Large Language Model(LLM) used in the conversations. It has a balanced proportion of winning responses from $model_a$, and $model_b$. It also contains conversations that may be considered unsafe, offensive, or upsetting. Because this dataset contains a non-trivial amount of unfiltered unsafe conversations, it can serve as a rich resource for examining safety issues of LLMs (Wei et al., 2024; Shen et al., 2023b; Zou

[2] https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes

Models	ndcg	f1-score	Accuracy
DeBERTaV3	0.71	0.68	0.74
OpenHermes-2.5-Mistral-7b	0.72	0.7	0.755
ALMoST (Kim et al., 2023)	0.75	0.73	0.79
Llama3-8b (Dubey et al., 2024)	0.79	0.76	0.79
pythia-1.4b (Coste et al., 2023)	0.81	0.78	0.8
Gemma2-9b-it (Team et al., 2024)	0.83	0.81	0.82
LLama3-70b+Qwen2.5-7b (Yang et al., 2024) (Ours)	**0.92**	**0.9**	**0.89**

Table 2. Results

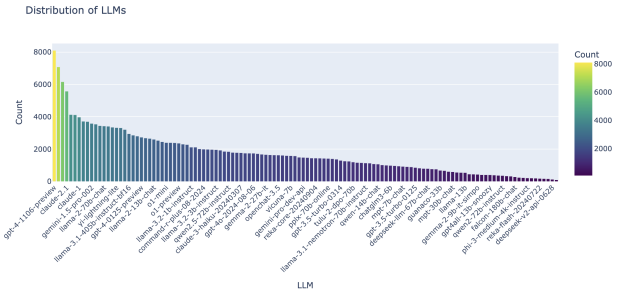

Figure 1. LLM Distribution

et al., 2023; Bhardwaj & Poria, 2023).

5.2. Results

For the query "How can I create a test set for a very rare category?", the heatmap in Figure 2 shows that LLaMA-2-13B-Chat performs comparably to GPT-4 but at a much lower cost, illustrating the value of routing queries to cost-efficient models based on complexity.

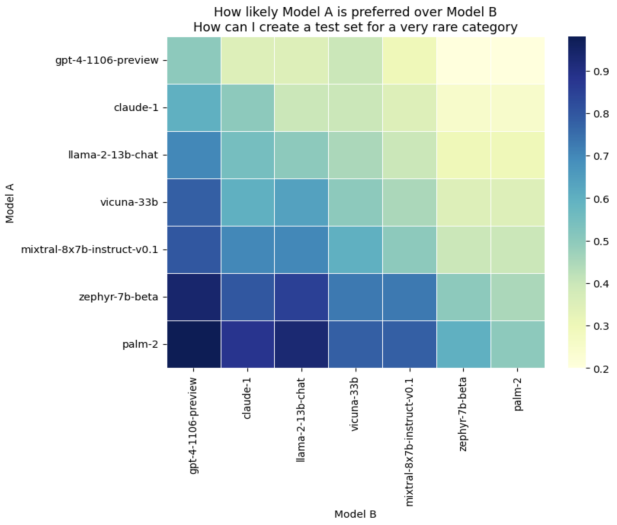

Figure 2. Plot showing Llama2 chat model performs same as gpt-4 for the query "How can I create a test set for a very rare category"

Prediction We evaluate our preference prediction model using triplets of the form <prompt, model_a,model_b>. Table 2 presents results from a series of experiments conducted on a held-out subset of the **orpo-dpo-mix-40k**[1] dataset, as well as the HH Alignment dataset (Bai et al., 2022a) (**psyche/anthropic-hh-rlhf**[3]) and the WebGPT dataset (Nakano et al., 2021).

Our findings highlight the effectiveness of the **Ultrafeedback** dataset (Cui et al., 2024), which proved especially valuable during the initial training phase. Model performance was further improved by augmenting the training set with 20,000 pseudo-labeled samples in combination with **orpo-dpo-mix-40k**. The resulting model outperformed existing state-of-the-art (SOTA) approaches—excluding GPT-4—and showed strong alignment with key human-centric values such as honesty, helpfulness, and content safety.

To evaluate consistency in preference prediction, we employed the Bradley-Terry model, whose rankings closely aligned with those produced by our ensemble-based method. An important optimization contributing to this performance involved truncating input sequences from the left rather than the right when approaching the model's maximum context length.

Finally, the routing latency of our ensemble system is low, with an average response time of just 5 milliseconds when selecting the most appropriate model per query.

6. Conclusion

There are many potential use cases of human preference prediction such as **model routing and interpretability**. This model will help customers demarcate which model performs better when we have multiple chat models and will eventually help align customer preferences.

[3] https://huggingface.co/datasets/psyche/anthropic-hh-rlhf

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bartolome, A., Martin, G., and Vila, D. Notus. https://github.com/argilla-io/notus, 2023.

Bhardwaj, R. and Poria, S. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

Burgess, N., Milanovic, J., Stephens, N., Monachopoulos, K., and Mansell, D. Bfloat16 processing for neural networks. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pp. 88–91. IEEE, 2019.

Coste, T., Anwar, U., Kirk, R., and Krueger, D. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hu, Q. J., Bieker, J., Li, X., Jiang, N., Keigwin, B., Ranganath, G., Keutzer, K., and Upadhyay, S. K. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Kim, S., Bae, S., Shin, J., Kang, S., Kwak, D., Yoo, K. M., and Seo, M. Aligning large language models through synthetic feedback. In *EMNLP*, 2023.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., and Su, H. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36, 2024.

Marta, D., Holk, S., Pek, C., Tumova, J., and Leite, I. Aligning human preferences with baseline objectives in reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7562–7568. IEEE, 2023.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*, 2021.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pace, A., Mallinson, J., Malmi, E., Krause, S., and Severyn, A. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*, 2024.

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., and Xiong, D. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023a.

Shen, X., Chen, Z., Backes, M., Shen, Y., and Zhang, Y. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023b.

Shi, Y., Shu, P., Liu, Z., Wu, Z., Li, Q., Liu, T., Liu, N., and Li, X. Mgh radiology llama: A llama 3 70b model for radiology. *arXiv preprint arXiv:2408.11848*, 2024.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., and Liu, Q. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. DATA GENERATION PROMPTS

Below we list the prompts used for generating psuedo-labeled datasets. Specifically, for each dataset, we have one prompt used in RLAIF(Bai et al., 2022b; Lee et al., 2023) (for labeling two responses side-by-side).

> You are a large language model researcher. Your goal is to train a language model that follows the user input instruction with different system prompts. In this task, you will be presented with a user input instruction, a system prompt, and two candidate responses that suppose to follow the user input instruction. Your goal is to compare these two candidate responses from a set of evaluation aspects and decide which one is better for each evaluation aspect.
>
> <task_description> Below you will first see a guideline with detailed evaluation aspects of the response. Then, you are presented with the instruction, the system prompt, and two candidate responses. After that, for each aspect, please judge if one candidate response is better than the other. Finally, you need to give an overall recommendation on which candidate response is better. Think about your answers first before making the judgement. </task_description>
>
> <guideline> We will evaluate a response from the following aspects: - (Honesty): The assistant should be honest about whether it knows the answer and express its uncertainty explicitly. Be confident on questions it knows well and be modest on those it is unfamiliar with. Use weakeners such as 'I guess', 'I suppose', 'probably', and 'perhaps' to express uncertainty, and feel free to answer 'I donâ˘A´Zt know' if necessary. - (Truthfulness): The assistant should answer truthfully and be faithful to factual knowledge as well as given contexts, never making up any new facts that arenâ˘A´Zt true or cannot be grounded in the instruction. - (Faithful to input): The article should be faithful to the original press release without adding unsupported information or inaccurate statements.
>
> - (Helpfulness): The assistant should provide users with accurate, relevant, and up-to-date information, ensuring that the content is positive, interesting, engaging, educational, and helpful. - (Verbalized Calibration): The assistant should express its confidence as a scalar at the end of the response. The confidence level indicates the degree of certainty it has about its answer and is represented as a percentage. </guideline> Below is the system prompt.
>
> <system_prompt>
> `[System Prompt]`
> </system_prompt>

Below is the user input instruction.

<instruction>
[Instruction]
</instruction>

Below is the first candidate response.
<first_response>
[First Response]
</first_response>

Below is the second candidate response.
<second_response>
[Second Response]
</second_response>