Beyond the Binary: Capturing Diverse Preferences With Reward Regularization

Vishakh Padmakumar[*]
New York University
vishakh@nyu.edu

Chuanyang Jin[*]
Johns Hopkins University
cjin33@jhu.edu

Hannah Rose Kirk[*]
University of Oxford
hannah.kirk@oii.ox.ac.uk

He He
New York University
hehe@cs.nyu.edu

Abstract

Large language models (LLMs) are increasingly deployed via public-facing interfaces to interact with millions of users, each with diverse preferences. Despite this, preference tuning of LLMs predominantly relies on reward models trained using binary judgments where annotators select the preferred choice out of pairs of model outputs. In this work, we argue that this reliance on binary choices does not capture the broader, aggregate preferences of the target user in real-world tasks. We propose a taxonomy that identifies two dimensions of subjectivity where different users disagree on the preferred output—namely, the *Plurality of Responses to Prompts*, where prompts allow for multiple correct answers, and the *Indistinguishability of Responses*, where candidate outputs are paraphrases of each other. We show that reward models correlate weakly with user preferences in these cases.

As a first step to address this issue, we introduce a simple yet effective method that augments existing binary preference datasets with synthetic preference judgments to estimate potential user disagreement. Incorporating these via a margin term as a form of regularization during model training yields predictions that better align with the aggregate user preferences.

1 Introduction

A ubiquitous step in the training of contemporary large language models (LLMs) is aligning their output with human preferences [1–5]. This process involves collecting human preference judgments over model outputs, which serve as the reward signal for either reinforcement learning with human feedback [1, 6] or various direct alignment algorithms [7, 8]. Conventionally, a single human annotator provides a binary judgment for a pair of model outputs [9, 10].

This "preferentist" approach to alignment has been criticized [11–13], pitted against alternatives such as deliberative procedures that establish normative standards among diverse stakeholders [14, 15] or principle-based frameworks that seek higher-order rules or "constitutions" to guide AI behavior [16–18]. However, preference judgments remain central to LLM post-training due to their cognitive and cost *efficiency*—it is far less demanding for annotators to choose between two outputs than to articulate an ideal demonstration—and surprising *efficacy* in steering LLMs towards being more friendly, helpful, and adept at interpreting user intent [1, 2, 19, 20]. So, even if preference fine-tuning is not the optimal solution, it is likely to remain a key component in LLM development for the foreseeable future, especially in industry settings.
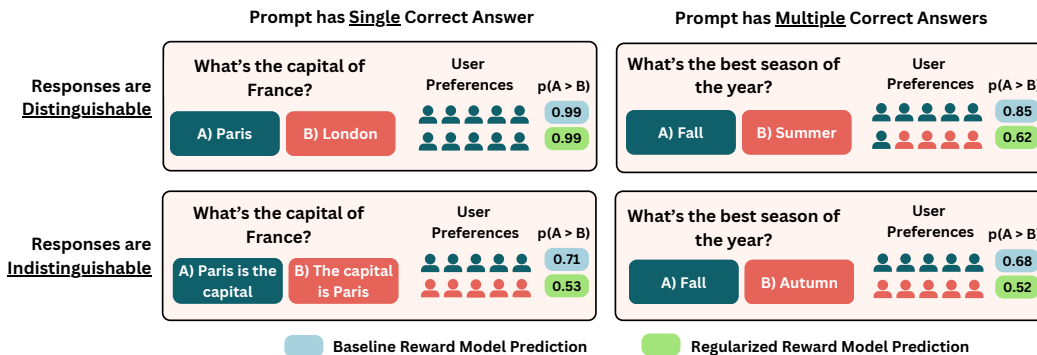
---

[*]Equal contribution.

Figure 1: Two dimensions of subjectivity along which reward model predictions correlate poorly with the diverse preferences of a population of users—when the prompt allows for multiple correct answers and when the two responses are indistinguishable from one another (Section 3). Regularizing the training objective with a margin term (Section 3.3) leads to predictions that better correlate with these preferences.

However, despite its ubiquity and early success, this dominant methodology for preference-tuning LLMs is now at odds with the reality of their deployment in society. Accessible via public interfaces, LLMs have amassed hundreds of millions of users [21], who have diverse preferences, needs, and linguistic backgrounds. The single binary choice between outputs becomes an unreliable signal for this vast user base, particularly in subjective domains like safety [22, 23] or toxicity [24], and when judging the overall quality of the output [6] in absence of detailed guidelines for what counts as high-quality language [25, 26].

In this paper, we address this inherent gap in development-deployment preference formulation. We argue that if preference tuning continues to be a crucial component in modern LLM training, steps must be taken to calibrate the preference signal so that it better translates from a single-annotator development setting to a multi-user deployment environment. We present our empirical argument in two phases:

The "break it" phase demonstrates the need for more calibrated reward estimation through a detailed error analysis of existing reward models. We hypothesize that over-reliance on single-annotator judgments leads to noisy and poorly calibrated preference signals, especially on subjective examples where different annotators would disagree on the preferred choice. We present a qualitative taxonomy of reward model failures in such cases (Section 3.1) and validate that reward model predictions on these examples correlate more poorly with human judgments (Section 3.2).

The "fix it" phase tests a simple yet effective modification to the reward model training objective. Based on recent work [3], we introduce a margin term in the objective function of reward modeling that is scaled to reflect the degree of disagreement of preference judgments in a hypothetical population of users. Our experiments show that these regularized model predictions better align with human preferences (Section 3.3).

Our contributions add to a body of work seeking alignment of LLMs that can better account for natural distributions of human preferences across global populations [27–29]. As a resolution to preference disagreements, some prior work has targeted personalization [30–32], but our approach is more similar to those seeking more principled approaches to aggregate preference estimation [33], for example, those drawing insight from social choice theory [34–36], seeking distributional learning [37, 38] or uncertainty modeling [39]. Our margin-based method can be seen as adopting a utilitarian approach to preference maximization with uniform weightings, but the framework can also accommodate other normative considerations of distributive justice in AI alignment.

Our method represents a simple impactful adjustment that is compatible with existing dominant paradigms of preference fine-tuning, though it is not a complete solution. By further refining our model and data methodologies for preference alignment, we can work towards AI systems that better accommodate the diversity of subjective human experiences.

2 Problem Formulation

The canonical method to train a reward model, r_θ, involves collecting a human preference judgment, j, that indicates which of two candidate outputs $y \in \{y_0, y_1\}$ is preferred for a given input prompt x, resulting in dataset D. The model parameters are optimized by minimizing the loss function:

$$\text{loss}(r_\theta) = -\text{E}_{(x,y_0,y_1,j)\sim\text{D}}[\log(\sigma(r_\theta(x, y_j) - r_\theta(x, y_{1-j})))], \tag{1}$$

where σ denotes the sigmoid function and $j \in \{0, 1\}$.

The reward model parameters are trained to maximize the difference in score assigned to y_j and y_{1-j} based on the preference judgment, j. This is effective when j is unanimously the 'better' output, as j would likely be consistent with the preferences of any future user interacting with the model.

However, we argue that in practical applications of LLMs, the models are frequently required to make decisions on examples where there is legitimate disagreement among users. Consider a scenario where a set of n users each provide a judgment, $j_{1...n}$, on the same example. The intended behavior of the model is to predict the aggregate preference of the population, specifically the fraction of users that prefer option y_0 over y_1. We write this as:

$$p_{01}^* = \frac{\sum_{i=1}^{n} \mathbf{I}(j_i == 0)}{n} \tag{2}$$

where $\mathbf{I}(\cdot)$ is and indicator function that equals 1 if user i prefers option y_0, and 0 otherwise.

However, this true preference p_{01}^* is not accessible since we do not have prior knowledge about future user preferences. This can lead to poor reward model performance when p_{01}^* deviates from the binary judgment j in Equation (1). In Section 3, we outline two dimensions along which subjective examples can be identified (Section 3.1) and empirically demonstrate that the correlation between reward model predictions and the aggregate user preferences—our estimate of p_{01}^*—is weaker for these subjective examples (Section 3.2).

Can we use synthetic preference judgments to make more aligned predictions? As part of our initial efforts to mitigate this effect, we propose a simple modification to the reward modeling objective. Here we take inspiration from Touvron et al. [3] who collect and incorporate the *extent* of user preferences between pairs of model outputs, i.e. *significantly* better vs *slightly* better, via a margin term scaled from 0 to 1. We introduce a similar margin term that scales proportional to the estimated disagreement on the choice between outputs y_0 and y_1 given prompt x—the margin value should be high when the choice between the two options is unanimous and vice versa. Our margin term converts binary preference judgments from different annotators into a cardinal measure of the *strength* of preference of the group as a whole. A gold standard implementation of the margin would be to extend to the preference data collection process to obtain multiple judgments for each pair of outputs from different annotators. However, the prohibitive cost of re-annotating existing datasets makes us favor a silver version that implements the margin using synthetic judgments from an LLM. Given an input prompt and a two candidate outputs (x, y_0, y_1), we sample n synthetic judgments $j'_{i=1...n} \in \{0, 1\}$ for a binary preference between the two outputs. We convert these into a margin term $m_{x,y_0,y_1} \in [0, 1]$ as:

$$m_{x,y_0,y_1} = \left(\left| \sum_{i=1}^{n} j'_i - \frac{n}{2} \right| \right) \Big/ \frac{n}{2} \tag{3}$$

Here $m = 1$ if all of the synthetic judgments are either 0 or 1, i.e. a unanimous preference in a population of annotators, and $m = 0$ if half the of the synthetic judgments are 0 and 1 each, i.e. a highly contested preference.[2]

This margin term is incorporated into the loss objective as follows:

$$\text{loss}(r_\theta) = -\text{E}_{(x,y_0,y_1,j)\sim\text{D}}[\log(\sigma(r_\theta(x, y_j) - r_\theta(x, y_{j_{1-j}}) - m_{x,y_0,y_1}))] \tag{4}$$

We test the effect of this intervention in Section 3.3 and demonstrate how this improves the correlation of reward model predictions to user preferences in subjective examples.

---

[2]We scale the margin between 0 and 1 following empirical experiments on the range of the margin in Touvron et al. [3].

3 Breaking Down Reward Model Performance by Category

In this section, we present a categorization of examples according to the degree of disagreement in aggregate preference among annotators (Section 3.1) and demonstrate that the performance of reward models drops in those categories with high levels of disagreement (Section 3.2). We then show how the performance improves by incorporating margin-based regularization into the model training pipeline (Section 3.3).

3.1 Categories of Subjectivity

In current frameworks for training language models to learn human preferences, such as Reinforcement Learning from Human Feedback (RLHF), the predominant method of preference annotation relies on binary judgments between pairs of model-generated outputs. However, human preferences are inherently subjective and exhibit significant inter-individual variability. Hence, singular binary annotations may not sufficiently capture the range of opinions.

To better accommodate this variability in human judgment, we first propose to classify examples into two primary categories based on the nature of their prompts: (i) prompts that possess an objective, *single correct* answer, and (ii) prompts that are inherently subjective, characterized by admitting *multiple correct* responses.

Moreover, within existing datasets, we observe many instances where the two model-generated responses are paraphrases of each other. This can occur in responses to both objective and subjective prompts. Therefore, we suggest a secondary dimension of categorization based on the distinguishability of the two model-generated candidates: (i) responses that are *distinguishable*, and (ii) responses that are *indistinguishable* paraphrases.

Figure 1 details illustrative examples from each of the aforementioned categories. We now evaluate whether this categorization corresponds to a decrease in the correlation between reward model predictions and user preference judgments.

3.2 Evaluating Reward Model Performance

Experimental Setup   We use a trained DeBERTa-V3 as the reward model for our experiments.[3] Our goal is to evaluate its performance across examples in each of the aforementioned categories: Multiple/Single Correct Answer; Distinguishable/Indistinguishable Responses. We construct our test set of 150 examples by randomly selecting 25 examples from each of six datasets. We use four in-domain (ID) datasets that were used to train the reward model—WEBGPT [40], HH-RLHF [2], Open AI Summarize [6], and INSTRUCTGPT-J [41]. The remaining two datasets, PRISM [27] and ULTRAFEEDBACK [42], serve as out-of-domain (OOD) sets used for testing purposes.

We recruit three human annotators to label the 150 example pairs manually. We first put the examples into categories along both aforementioned dimensions via a majority vote. To estimate the aggregate preference of future users (Equation (2)), our annotators are instructed as follows: "If you asked 10 people, how many would prefer answer A and how many would prefer answer B?" This allows them to provide a judgment on a scale ranging from 10-0 to 0-10, and we average these responses to obtain a consensus. Notably, this approach enables participants to first predict the believed subjectivity—even if they personally prefer answer A, they might project that only 2 out of 10 people would share their preference, thus providing insight into population-level preferences. Empirical evidence suggests that this approach yields more reliable results than direct population votes asking each annotator for their preference between answer A and answer B. We evaluate the performance of the reward model through the Pearson correlation (Table 1) and L1 loss (Table 2) between the collected human preference and this *Baseline* reward model (RM) predictions.

Results   From Table 1 and Table 2, we see that the model performs significantly worse on subjective examples, indicating its limitations in capturing human preferences where *multiple correct* responses exist. Moreover, the model demonstrates notably poorer performance on the OOD datasets, suggesting

---

[3]We selected this reward model, released as part of the Open Assistant project, for our experiments due to its publicly available training data and the fact that it was the most frequently downloaded reward model on Huggingface Hub at the time of our experiments.

Dataset (N)	Baseline RM	Regularized RM	Δ		Category (N)	Baseline RM	Regularized RM	Δ
WEBGPT (25)	0.677	0.637	-0.040		**Multiple Correct (106)**	0.572	0.626	0.054
HH-RLHF (25)	0.974	0.974	0.000		**Single Correct (44)**	0.802	0.799	-0.003
OAISUMMARIZE (25)	0.468	0.585	0.117					
INSTRUCTGPT-J (25)	0.573	0.544	-0.029		**Multiple Correct ID (66)**	0.691	0.715	0.024
					Single Correct ID (40)	0.835	0.824	-0.011
All ID (100)	0.734	0.749	0.015		**Multiple Correct OOD (34)**	0.251	0.384	0.133
					Single Correct OOD (10)	0.694	0.766	0.072
PRISM (25)	0.472	0.658	0.186					
ULTRAFEEDBACK (25)	0.248	0.277	0.029		**Distinguishable (124)**	0.640	0.686	0.046
All OOD (50)	0.377	0.488	0.111		**Indistinguishable (26)**	0.631	0.652	0.021

Table 1: Pearson correlation of predictions from the baseline reward model (RM) (Section 3.2) and regularized RM (Section 3.3) to user preferences. *Regularization* improves performance on both OOD datasets as well as subjective examples with multiple valid answers. Cells in green show absolute performance, while Δ shows (Regularized RM - Baseline RM).

Dataset (N)	Baseline RM	Regularized RM		Category (N)	Baseline RM	Regularized RM
WEBGPT (25)	0.171	**0.159**		**Multiple Correct (106)**	0.263	**0.234**
HH-RLHF (25)	0.114	**0.108**		**Single Correct (44)**	0.163	**0.158**
OAISUMMARIZE (25)	0.274	**0.207**				
INSTRUCTGPT-J (25)	**0.315**	0.319		**Multiple Correct ID (66)**	0.252	**0.227**
				Single Correct ID (40)	0.152	**0.142**
All ID (100)	0.218	**0.198**		**Multiple Correct OOD (34)**	0.281	**0.244**
				Single Correct OOD (10)	**0.202**	0.211
PRISM (25)	0.275	**0.205**				
ULTRAFEEDBACK (25)	**0.255**	0.270		**Distinguishable (124)**	0.244	**0.219**
All OOD (50)	0.265	**0.238**		**Indistinguishable (26)**	0.176	**0.160**

Table 2: Average L1 loss (absolute difference) between the aggregate user preferences (normalized to a fraction from 0 to 1) and the model predictions from the baseline RM (Section 3.2) and regularized RM (Section 3.3). Bold values deviate less from the user preferences. *Regularization* helps RMs better capture user preferences, particularly on average across all ID and OOD datasets as well as subjective examples with multiple valid answers.

that traditional reward modeling approaches may not adequately generalize across different data domains. Additionally, we find that the baseline reward model performs similarly (Table 1), or slightly better (Table 2) on indistinguishable than distinguishable responses, indicating slightly more robust behavior along this axis of subjectivity. Our findings align with contemporary work that investigate annotator disagreement along different dimensions to also conclude that current reward models, which are trained with the assumption that this disagreement is noise, inadequately represent these diverse preferences [29].

3.3 Regularization With Synthetic Preferences to Improve Performance

Having observed the decline in reward model performance on subjective examples, we test if training models with our proposed margin-based intervention results in better generalization to these examples. To do so, we first obtain 10 synthetic annotations, $j'_{i=1...10}$, for each example in the 4 training datasets (Section 3.2) using the Llama-3 70B Instruct model [43].[4] We generate synthetic annotations with nucleus sampling setting the top-p parameter to 0.9 to mimic the variability of a group of annotators.[5]

---

[4] We select the meta-llama/Meta-Llama-3-70B-Instruct via Huggingface [44] as it was the highest performing *open-weight* model on the Stanford HELM benchmark [45] at the time we conducted our experiments.

[5] We provide the exact prompt used in Appendix A.

These are used to construct the margin for each training example as detailed in Equation (3). We then train the Deberta-V3 reward model using the objective in Equation (4) [6] Finally, we report the Pearson correlation and L1 loss between the model predictions and the collected human annotations on the test set in Table 1. We observe that our method results in better performance on the *multiple correct* subsets, both ID and OOD, without a degradation in the corresponding *single correct* subsets. Our method also results in an improvement in performance scores on examples where the two outputs are *indistinguishable*, without costing performance on the *distinguishable* subset.

4   Conclusion

In this work, we contend that reward models trained on binary preferences struggle to generalize to subjective examples where different users may disagree on the preferred option. By categorizing examples along two dimensions of subjectivity—when the prompt allows for multiple correct answers and when the candidate outputs are paraphrases of each other—we demonstrate that reward model predictions exhibit a weaker correlation with human preferences in these cases. To address this, we propose a margin-based regularization technique that mitigates this issue using synthetic annotations from an LLM to improve prediction quality in subjective scenarios. We re-iterate the scope of this solution as a step to improve reward model performance as they are currently deployed, a technique to be used when recollecting large-scale preference data is not possible financially. We detail further social concerns that arise below.

Social Impacts Statement

Our work situates itself in a methodological issue of alignment fine-tuning: the overreliance on single-annotator binary judgments. We propose regularization with synthetic preferences a simple, pragmatic solution that offers a quick and easily implementable patch to improve the calibration of preference signals, especially in subjective regions of input-output space and in the absence of extensive additional human data. However, our work raises important epistemological and ethical considerations. Our assumption that LLMs can better approximate preference heterogeneity than individual annotators has mixed empirical validation [46–49]. The use of synthetic annotations, while a pragmatic solution, may introduce algorithmic biases that could affect model outputs in opaque ways, misrepresenting populations of users [50, 51]. Our demonstrated implementation adopts a utilitarian approach, simply taking a majority vote for preference aggregation. While our method is flexible to other specifications, our work does not resolve fundamental ethical questions about how to aggregate conflicting preferences or fairly represent minority viewpoints without succumbing to the tyranny of the majority. The assumption that we need to aggregate preferences may itself be flawed. In contrast to seeking a monolithic LLM with lofty ambitions to simultaneously represent the preferences of a vast user base, a promising path to pluralistic AI may come from a more granular unit of alignment via personalization or community-specific steering and cultural fine-tuning [52]. We present this work not as a definitive solution, but as a contribution to the ongoing discourse on responsible, safe, and inclusive AI development, emphasizing the need for continued interdisciplinary research to address the sociotechnical and normative challenges central to technical alignment methodologies.

References

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

---

[6]To clarify, we fine-tune a pre-trained *Deberta-V3* model on the same training data as the baseline model from Section 3.2 using our modified objective. We run a hyperparameter sweep of learning rates from $1e - 04$ to $1e - 06$ and select the best-performing model on the validation data.

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

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

[5] Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi. *arXiv preprint arXiv:2405.10313*, 2024.

[6] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

[9] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The History and Risks of Reinforcement Learning and Human Feedback, November 2023. URL http://arxiv.org/abs/2310.13595. arXiv:2310.13595 [cs].

[10] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Röttger, and Scott Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, 2023.

[11] Iason Gabriel. Artificial Intelligence, Values and Alignment. *Minds and Machines*, 30(3): 411–437, September 2020. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09539-2.

[12] John Tasioulas. Artificial Intelligence, Humanistic Ethics. *Daedalus*, 151(2):232–243, May 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01912. URL https://doi.org/10.1162/daed_a_01912.

[13] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *arXiv preprint arXiv:2408.16984*, 2024.

[14] Samuel Chang, Estelle Ciesla, Michael Finch, James Fishkin, Lodewijk Gelauff, Ashish Goel, Ricky Hernandez Marquez, Shoaib Mohammed, and Alice Siu. Meta Community Forum: Results Analysis. Technical report, Deliberative Democracy Lab, Stanford University, April 2024. URL https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2024-03/meta_ai_final_report_2024-04_v28.pdf#page=5.15.

[15] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1):6616, March 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56648-4. URL https://www.nature.com/articles/s41598-024-56648-4. Publisher: Nature Publishing Group.

[16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

[17] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Mari-beth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. September 2022. URL http://arxiv.org/abs/2209.14375v1.

[18] Anthropic. Collective Constitutional AI: Aligning a Language Model with Public Input. Technical report, 2023. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[19] MetaAI. Introducing Meta Llama 3: The most capable openly available LLM to date, April 2024. URL https://ai.meta.com/blog/meta-llama-3/.

[20] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*, 2024.

[21] Emma Roth. ChatGPT's weekly users have doubled in less than a year. *The Verge*, August 2024. URL https://www.theverge.com/2024/8/29/24231685/openai-chatgpt-200-million-weekly-users.

[22] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[23] Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. DICES Dataset: Diversity in Conversational AI Evaluation for Safety, June 2023. URL http://arxiv.org/abs/2306.11247. arXiv:2306.11247 [cs].

[24] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR, 2023.

[25] Hannah Kirk, Bertie Vidgen, Paul Rottger, and Scott Hale. The empty signifier problem: Towards clearer paradigms for operationalising" alignment"in large language models. In *Socially Responsible Language Modelling Research*, 2023.

[26] Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection, January 2022. URL http://arxiv.org/abs/2201.10474. arXiv:2201.10474 [cs].

[27] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.16019.

[28] Chuanyang Jin, Songyang Zhang, Tianmin Shu, and Zhihan Cui. The cultural psychology of large language models: Is chatgpt a holistic or analytic thinker? *arXiv preprint arXiv:2308.14242*, 2023.

[29] Michael JQ Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? *arXiv preprint arXiv:2410.14632*, 2024.

[30] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[31] Xinyu Li, Zachary C. Lipton, and Liu Leqi. Personalized Language Modeling from Personalized Human Feedback, February 2024. URL http://arxiv.org/abs/2402.05133. arXiv:2402.05133 [cs].

[32] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

[33] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

[35] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[36] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Social Choice for AI Alignment: Dealing with Diverse Human Feedback, April 2024. URL http://arxiv.org/abs/2404.10271. arXiv:2404.10271 [cs].

[37] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF, December 2023. URL http://arxiv.org/abs/2312.08358. arXiv:2312.08358 [cs, stat].

[38] Dexun Li, Cong Zhang, Kuicai Dong, Derrick Goh Xin Deik, Ruiming Tang, and Yong Liu. Aligning Crowd Feedback via Distributional Preference Reward Modeling, February 2024. URL http://arxiv.org/abs/2402.09764. arXiv:2402.09764 [cs].

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

[40] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[41] Alex Havrilla. synthetic-instruct-gptj-pairwise (revision cc92d8d), 2023. URL https://huggingface.co/datasets/Dahoas/synthetic-instruct-gptj-pairwise.

[42] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

[43] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain

Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[45] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=iO4LZibEqW`. Featured Certification, Expert Certification.

[46] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples, September 2022. URL `http://arxiv.org/abs/2209.06899`. arXiv:2209.06899 [cs].

[47] Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. Are large language models good annotators? In *Proceedings on*, pages 38–48. PMLR, 2023.

[48] Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

[49] Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*, 2024.

[50] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[51] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. The illusion of artificial inclusion, February 2024. URL `http://arxiv.org/abs/2401.08572`. arXiv:2401.08572 [cs].

[52] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A Roadmap to Pluralistic Alignment, February 2024. URL `http://arxiv.org/abs/2402.05070`. arXiv:2402.05070 [cs].

A   Prompts for Obtaining Synthetic Preference Judgments

To gather synthetic preference judgments from Llama-3 in Section 3.3, we use the following prompt:

> **Prompt for Obtaining Synthetic Preference Judgments**
>
> I am going to give you a prompt and two answers.
> The goal is to identify the better answer.
> If the prompt is a question, we want the factually correct answer.
> In a conversation, we want helpful replies that do not cause harm.
> The output format should be either 'Choice: A' or 'Choice: B' based on the selected answer and nothing else.
> Prompt: {prompt}
> Answer A: {answer_0}
> Answer B: {answer_1}

B   Annotator Guidelines

We outline the guidelines provided to annotators for conducting the manual annotations in Section 3.2:

> **Task 1: Subjectivity**
>
> Consider the following prompt: {prompt}
> Is there a single factually, correct response to this prompt?
> Options:
> A) There is only one single correct choice to this prompt.
> B) There are multiple possible correct answers to this prompt

> **Task 2: Distinguishability**
>
> Consider the following two candidate model outputs in response to the prompt: {prompt}
> Answer A: {answer_0}
> Answer B: {answer_1}
> Can you clearly tell the two responses apart? Are they significantly different from one another?
> Options:
> A) Both choices are essentially the same content, just with minor stylistic variations.
> B) The two choices are clearly different from each other and distinguishable.

> **Task 3: User Preference Estimation**
>
> If you asked 10 different people, how many would prefer answer A and how many would prefer answer B?
> 1) 10 A - 0 B
> 2) 9 A - 1 B
> 3) 8 A - 2 B
> 4) 7 A - 3 B
> 5) 6 A - 4 B
> 6) 5 A - 5 B
> 7) 4 A - 6 B
> 8) 3 A - 7 B
> 9) 2 A - 8 B
> 10) 1 A - 9 B
> 11) 10 A - 0 B