Diverging Preferences: When do Annotators Disagree and do Models Know?

Michael J.Q. Zhang[α1], **Zhilin Wang**[γ], **Jena D. Hwang**[β], **Yi Dong**[γ], **Olivier Delalleau**[γ], **Yejin Choi**[γδ], **Eunsol Choi**[α], **Xiang Ren**[ϵ], **Valentina Pyatkin**[βδ]

New York University[α], Allen Institute for Artificial Intelligence[β], NVIDIA[γ], University of Washington[δ], University of Southern California[ϵ]

michaelzhang@nyu.edu, valentinap@allenai.org

Abstract

We examine *diverging preferences* in human-labeled preference datasets. We develop a taxonomy of disagreement sources spanning 10 categories across four high-level classes. We find that the majority of disagreements are in opposition with standard reward modeling approaches, which are designed with the assumption that annotator disagreement is noise. We then explore how these findings impact reward modeling. In our experiments, we demonstrate how standard reward modeling methods, like the Bradley-Terry model, fail to differentiate whether a given preference judgment is the result of unanimous agreement among annotators or the majority opinion among diverging user preferences.

1 Introduction

Learning from human preferences has emerged as the standard method for adapting LLMs to facilitate user-assistant interactions with much success. Despite these advances, however, the field continues to struggle with the challenge of handing *diverging preferences*, where users disagree on the ideal response to a prompt. Prior works on developing pluralistically aligned LLMs have focused on the development of synthetic preference datasets, where disagreements are simulated based on author-defined features and frequencies [19, 3]. In this work, we take a step back to ask the foundational question *when and why do human annotators disagree in their preferences?*

To make this research possible, we introduce MultiPref-Disagreements and HelpSteer2-Disagreements. [2] Prior to this work, public preference datasets [1, 27] were either annotated by a single annotator or preprocessed to aggregate multiple annotations into a single label, making such investigations into annotator disagreements not possible. We present a novel taxonomy of disagreement sources spanning 10 categories and 4 high-level classes (Table 1). Based on our analysis of these datasets, we offer two significant findings. First, we find that diverging preferences constitute over 30% of examples across both datasets. Second, our analysis shows that most disagreements in preference annotations are caused by individual predilections rather than annotator errors. We find that over 75% of disagreements are influenced by factors such as response complexity, verbosity, or interpretations of underspecified prompts.

Our findings, that most disagreements in preference annotations are the result of individual predilections rather than annotation errors, run counter to how standard preference learning pipelines are designed and reward models (RMs) are trained, where dissenting opinions are treated as undesirable

[1]Work done at the Allen Institute for Artificial Intelligence

[2]Note that we did not collect new human annotations but release the individual annotations of these existing datasets (which previously released only aggregated annotations), with support from the dataset creators.

	Category	Prompt	Response 1	Response 2	MP / HS2
Task	Task Underspecification	(HS2) Pretend you are a postgresql server.	I am a PostgreSQL server... I am designed to be robust, scalable and perform well under...	Sure, I'll pretend to be a PostgreSQL server.“‘postgres=#“‘ Now what?	20% / 22%
Response Style	Verbosity	(MP) Is a gold or silver necklace better for a gift?	*(356 Words)* ... Here are some factors to consider: 1. Style... 6. Personal preference...	*(53 Words)* This ultimately depends on the individual's personal preference...	38% / 44%
	Format	(HS2) Describe the difference between Chongqing and Chengdu style hot pots.	1. Spiciness Level:... 2. Broth ingredients:... 3. Dipping sauces:...	### Chongqing vs Chengdu Hot Pot... ### Chongqing Hot Pot... ### Chengdu Hot Pot...	20% / 32%
	Aesthetic Taste	(MP) Give me a recipe for refried beans written in a style like jabberwocky	...In the mumbling pot where the sizzle-hiss plays, And the stove flames flicker...	In a kitchen-bramble, where the skillets sing, Twas' beans to be mashing...	14% / 22%
	Complexity	(MP) I overheard the term “Schrodinger's cat” being used and I'm not sure I know what it is. Can you explain it to me?	Certainly! Schrödinger's cat is a thought experiment ... illustrate what he saw as a problem with the Copenhagen interpretation of quantum mechanics...	... Schrödinger's cat is a thought experiment... illustrates the seemingly absurd consequences of applying quantum mechanics...	10% / 10%
Refusals	Safety	(MP) What are some acceptable ways to physically discipline your pet?	...physical discipline is not an acceptable way to train or correct a pet's behavior...	... some acceptable ways to physically discipline a pet include...	16% / 0%
	Capabilities	(MP) !! Three x 9 minus s3v3n EQ–uals ??	... it is not possible to evaluate the expression.	... Here is one possible interpretation of what you might...	4% / 4%
	Comply vs. Refuse	(MP) If you were designing a video game where the objective was to overthrow the U.S government, what concepts would you implement into the game?	I'm sorry, but I cannot fulfill this request. As an AI language model...	What a fascinating and thought-provoking question! ... would require a deep understanding of politics, sociology, and human psychology...	5% / 4%
	Refuse vs. Refuse	(MP) Why are latinos prvilaged?	... It is important to recognize that Latinos, like all people, have the capacity for both privilege and oppression...	I apologize, but it's important to correct the assumption in your question. Latinos are not privileged...	20% / 0%
Errors	Hallucinations, Degenerate Outputs, and Errors	(HS2) Assume g is a generator of the multiplicative group of integers modulo n. Is there an efficient algorithm to find such a generator g?	Yes... the "Pollard rho algorithm." Here's a high-level overview of the algorithm...	Yes... the Sieve of Eratosthenes and is based on the following observation...	14% / 24%

Table 1: Taxonomy of the causes of diverging preferences with examples and frequencies from the HelpSteer2 (HS2) and the MultiPref (MP) datasets.

noise. We demonstrate aggregating labels via majority choice [27, 12] results in reward models that predict decisive preference toward a single option, even when annotators preferences diverge.

2 Analysis: Diverging Preferences in RLHF Annotation

We define diverging preferences as all instances where annotators disagreed on which response to a given prompt was preferred, ignoring instances where annotators only had slight preferences for either response. We identify diverging preferences in two human labeled preference datasets:

MultiPref is a dataset of 10K preference pairs,[3] each consisting of a conversation prompt and two candidate responses [14]. Each response pair is annotated by four different annotators, who are tasked with comparing the two responses and determining which response they prefer, or whether both responses are tied. Annotators further designate whether their preferred response is *significantly* or only *slightly* better than the other. To identify examples with *diverging preferences*, we select all instances where annotators disagreed on which response was preferred, filtering out instances where all annotators responses were ties or only had slight preferences for either response. This process yields about 39% of preference pairs.

HelpSteer2 is a dataset of 12K preference pairs[4], where each preference pair is annotated by 3-5 different annotators. The annotators were instructed to review both responses and assign an independent score of overall helpfulness to each on a 1-5 likert scale. To identify annotator preferences, we take the difference between the overall scores assigned to each response, and treat differences in overall scores of 1 as instances of *slight* preference and differences of at least 2 as *significant* preferences. We follow the same method as used above for Multipref to identify instances of diverging preferences, which we find comprise 24% of all examples.

[3]Available at `https://huggingface.co/datasets/allenai/multipref`.

[4]The original 10k samples at `https://huggingface.co/datasets/nvidia/HelpSteer2` excludes samples with high disagreement as part of their data pre-processing. We include all annotations, since we are interested in the disagreements at `https://huggingface.co/datasets/nvidia/HelpSteer2/tree/main/disagreements`.

2.1 A Taxonomy for causes of Diverging Preferences

We perform manual analysis of diverging preferences in both datasets and develop a taxonomy for causes of diverging preferences in Table 1. This taxonomy was developed over a working set of 100 randomly sampled examples of diverging preferences from each dataset. Three of the authors then cross annotated 50 new sampled examples from each dataset for the reasons of diverging preferences to evaluate agreement. As there are often multiple possible causes for diverging preferences, we evaluate agreement using both Cohen's κ (comparing full label set equivalence), as well as Krippendorff's α with MASI distance [17], yielding $(\kappa = 0.59, \alpha = 0.68)$ and $(\kappa = 0.58, \alpha = 0.62)$ over our annotations on MultiPref and Helpsteer2, respectively. Below, we describe each disagreement cause and class.

Task Underspecification Disagreements often arise from underspecification in the prompt, where both responses consider and address distinct, valid interpretations of the task.

Response Style We identify several disagreements causes that arise due to differences in response style, where preferences are primarily influenced by an individual's tastes rather than content.

- **Verbosity** Disagreements arise over the preferred levels of detail, explanation, or examples in each response. While prior works have noted that RLHF annotations are often biased toward lengthy responses in aggregate [20], we find that individuals frequently disagree on the preferred level of detail or explanation in a response.
- **Format** We find that another common source of diverging preferences is disagreement over how responses should be organized. LLMs frequently present responses as paragraphs, lists or under headings. We find frequent disagreements over when such formatting is appropriate and how headings and lists should be semantically structured.
- **Complexity** Responses often differ in the level of assumed domain expertise of the user and the level of technical depth with which to consider the user's request. As such, diverging preferences arise over responses that are catered toward individuals with different backgrounds and goals.
- **Aesthetic tastes** Prior work has noted that creative writing or writing assistance comprise a significant portion of user requests [28]. We find that preferences often diverge for such requests, where a preference often comes down to a matter of personal taste.

Refusals We find that refusals based on **safety** concerns or model **capabilities** are often the subject of disagreement among annotators. This finding is consistent with prior work, which has demonstrated that judgments of social acceptability or offensive language can vary based on their personal background and identity [8, 24]. We, furthermore, find that diverging preferences often occur when comparing **refusals versus refusals**. Recent work has studied establishing different types of refusals (e.g., soft versus hard refusals) and rules for when each are appropriate [15]. Our findings suggest that user preferences among such refusal variations are frequently the source of disagreement.

Errors Prior work has noted that an individual's judgment of a response's correctness has almost perfect agreement with their judgment of a response's overall quality [27]. During annotation, however, errors can be difficult for annotators to detect or their impact may be perceived differently across annotators, leading to variation among preferences.

3 Reward Models make Decisive Decisions over Divisive Preferences

Our analysis above demonstrates that disagreements in preference annotations are often the result of differences in individual user perspectives rather than simple noise. In this section, we study the behaviors of standard reward modeling methods in cases of diverging and non-diverging preferences. Aligning LLMs via RLHF [16] involves training a reward model on human preference data to assign a reward r_A for a given prompt x and response A that is indicative of its quality $((x, A) \rightarrow r_A)$. LLMs are then adapted to generate responses that receive high rewards from the trained reward model. As such, reward models that heavily favor a single response in cases of diverging preference result in LLMs that learn to only predict responses tailored to a single perspective.

Below, we describe the two standard reward modeling methods explored in this work. To train them, prior work aggregate labels across multiple annotators by taking the majority vote [26, 12]. We train each model on both the aggregated labels as well as over all annotations in the dataset, treating each annotator label as its own training instance.

Example Type	MultiPref			HelpSteer2				
	# Ex.	BT (Agg)	BT (All)	# Ex.	BT (Agg)	BT (All)	MSE (Agg)	MSE (All)
High-Agreement Prefs.	127	0.786	0.669	298	0.751	0.718	0.811	0.676
High-Agreement Ties	141	0.663	0.580	117	0.673	0.631	0.412	0.340
Diverging Prefs. (All)	178	0.798	0.663	147	0.722	0.678	0.706	0.573
Diverging Prefs. (Subst.)	74	0.820	0.690	69	0.731	0.694	0.834	0.692
All Examples	500	0.762	0.647	576	0.725	0.688	0.683	0.565

Table 2: The average difference in rewards between the chosen and rejected responses. We measure this by $P(\text{chosen} > \text{rejected})$ for Bradley-Terry (BT) models and $r_{\text{chosen}} - r_{\text{rejected}}$ for MSE-Regression (MSE) models. We report the difference from the reward model trained with aggregated annotation (Agg) vs. the reward model trained using all annotations (All). Each row represents a different subset of the dataset, with different levels of agreement. We include the # of examples within each subset.

Bradley-Terry is a widely used approach for training reward models in the RLHF paradigm [1, 6]. It defines the likelihood of a user preferring response A over response B as $P(A > B) = \textit{logistic}(r_A - r_B)$ and is trained via minimizing the negative log likelihood on annotated preferences. In our experiments, we track how heavily reward models favor a single response by computing $P(C > R)$ where C and R are the reward model's chosen and rejected responses, respectively.

MSE-Regression is an alternative method that utilizes the individual Likert-5 scores for each response found in Regression-style datasets such as HelpSteer2 dataset [27]. Here, reward models predict the scalar reward of each response, and training is done by minimizing mean squared error against the 1-5 score assigned by annotators. To track how heavily reward models favor a single response, we track the distance in predicted rewards given by $|r_a - r_b|$.

Results We train separate reward models for each dataset based on Llama-3-8B-Instruct [7], and evaluate on 500 held-out test examples from each dataset. In Table 2, we present results comparing preference strength on examples with different levels of annotator agreement: *High-Agreement Prefs.:* where no annotators rejected the majority's chosen response. *High-Agreement Ties:* where the majority of annotators labeled the instance as a tie. *Diverging Prefs (All)* all examples with diverging preferences. *Diverging Prefs (Substantial)* a subset of diverging preferences where annotators significantly preferred both responses. When presented with examples with diverging preferences, reward models predict differences in rewards that are akin to high-agreement preferences, even when trained over all annotator labels. Our findings demonstrate that performing RLHF training with standard reward modeling methods may harm pluralistic alignment for LLM, as standard reward models learn to pick a side in cases of diverging preferences, rather than learning to predict a middle-ground reward for each response.

4 Related Work

Annotator disagreement has been studied in prior works in specific domains. [23] and [8], explore annotator disagreement in safety, looking specifically at how morality and toxicity judgments vary across users of different backgrounds. Prior works have analyzed disagreements in NLI [18, 13], and [11] develop an NLI-specific taxonomy of disagreement causes. Sandri et al. [22] similarly explores annotator disagreements in toxicity detection, and develop a taxonomy of disagreement causes for their task. Works have also studied disagreements in discourse due to task design [21]. Frenda et al. [9] presents a survey of datasets and methods for modeling different user perspectives across NLP tasks. Prior works have advocated for the importance of considering disagreements in NLP tasks [2] and have proposed shared tasks for training and evaluating models in settings with annotator disagreements [25].

5 Conclusion

We analyze and develop a taxonomy of disagreement causes of diverging preferences in human-annotated preference datasets and find that disagreements are often due to sensible variations in individual perspectives. We then demonstrate that standard reward models make decisive decisions over diverging preference, causing issues for training pluralistically aligned LLMs.

References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[2] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics, 2021.

[3] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences, 2024. URL `https://arxiv.org/abs/2406.08469`.

[4] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. 2022. URL `https://arxiv.org/pdf/2208.07339`.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. 2024. URL `https://arxiv.org/pdf/2305.14314`.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng

Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

[7] 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.

[8] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[9] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28, 2024.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. 2022. URL `https://arxiv.org/pdf/2106.09685`.

[11] Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10: 1357–1374, 2022.

[12] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.51. URL `https://aclanthology.org/2023.emnlp-main.51`.

[14] Lester James V Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. ai feedback. *arXiv preprint arXiv:2410.19133*, 2024.

[15] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian D Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

[16] 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.

[17] Rebecca J Passonneau. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[18] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[19] Sriyash Poddar, Yanming Wan, Hamish Ivision, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. 2024.

[20] Jiacheng Xu Prasann Singhal, Tanya Goyal and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv*, 2023.

[21] Valentina Pyatkin, Frances Yung, Merel CJ Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. Design choices for crowdsourcing implicit discourse relations: revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11: 1014–1032, 2023.

[22] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, 2023.

[23] Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. Nl-positionality: Characterizing design biases of datasets and models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[24] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*, 2022. URL `https://aclanthology.org/2022.naacl-main.431/`.

[25] Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. Semeval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, 2021.

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

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

[28] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Bl8u7ZRlbM`.

A Additional Modeling Details

We train all reward models with a learning rate of 5e-5 and a batch size of 16 and were trained for a maximum of 10 epochs, selecting the best performing checkpoint evaluated after every 0.25 epochs. For training and inference, we use 8-bit quantization [4] with LoRA [10, 5]. All systems were trained on 8 RTX A6000 GPUs.