Comparing LLM ratings of conversational safety with human annotators

Rajiv Movva
Cornell Tech
rmovva@cs.cornell.edu

Pang Wei Koh
University of Washington

Emma Pierson
Cornell Tech

1 Introduction

As large language models (LLMs) have gained broad use as conversational agents, there has been increased focus on aligning models to be safe and harmless. While methods like RLHF are intended to increase alignment with general "human preference" (Kirk et al., 2023), practically, safety becomes operationalized in specific ways due to model design choices (Santy et al., 2023). Annotators often disagree about what constitutes harmful content across demographic and cultural lines (Kumar et al., 2021; Davani et al., 2023), and recent data similarly finds demographic disagreements in perceptions of chatbot safety (Aroyo et al., 2023). As such, it is critical to evaluate *how* safety is encoded in LLMs. Probing an LLM's conception of safety can inform when the model may misalign with desired behavior in specific contexts or for specific user groups. Further, how an LLM evaluates safety has ripple effects on the broader AI ecosystem, as AI feedback increasingly replaces human feedback as a training signal (Bai et al., 2022) and evaluation metric (Lin and Chen, 2023).

In this project, we compare LLMs as annotators of user-chatbot conversational safety to a diverse body of human annotators. We use the DICES dataset (Aroyo et al., 2023), in which 350 user-chatbot conversations are each annotated for safety by 112 annotators spanning 10 race-gender groups. We re-annotate each conversation for safety by few-shot prompting GPT-4 with the same instructions given to the annotators, and study three questions:

- **RQ1:** How do GPT-4 safety ratings compare to the entire pool of annotators?
- **RQ2:** Demographic subgroups of annotators often disagree about safety. Do model ratings align differently with different groups?
- **RQ3:** Can GPT-4 predict when race/gender groups disagree about safety?

2 Results

Overview of data and models. DICES consists of 350 multi-turn conversations (Aroyo et al., 2023), in which crowdworkers were instructed to discuss sensitive topics with the chatbot LaMDA (Thoppilan et al., 2022). Later, distinct crowdworkers spanning different race, gender, and age groups annotated the safety of the chatbot's responses. Every conversation was rated by each of the 112 annotators for five binary safety criteria: harm, unfair bias, misinformation, political affiliation, or increasing polarization. The five ratings are aggregated with an `OR` into a single overall binary rating.

We similarly annotated each conversation using GPT-4. We chose GPT-4 because it is the most widely used model for AI feedback (Cui et al., 2023; Sharma et al., 2024), and weaker models like GPT-3.5 did not correlate well with the annotators. Our description of the task in the prompt closely reflects the instructions given to annotators; the main difference is that we prompted GPT-4 to output a single Likert score from 1 (definitely safe) to 5 (definitely unsafe) that incorporates all five safety criteria simultaneously. We found that a single Likert score correlated better with the annotators and increased format compliance versus collecting five separate binary ratings split by criterion.

RQ1: GPT-4 ratings are on par with a single annotator. To evaluate alignment with the collective annotator pool, we compute Pearson correlation between GPT-4 Likert ratings and μ_{all}, the fraction of annotators who rated a conversation unsafe (ranging from 0 to 1). Evaluating directly against μ_{all} instead of rounding to a "hard label" better captures agreement with the entire annotator distribution, not just the majority (Plank, 2022). GPT-4's correlation with μ_{all} is $r = 0.48$ when using 1 to 5 ratings, and $r = 0.45$ when binarizing ratings at a threshold of 3. For comparison, a single annotator's correlation with μ_{all} has mean

$r = 0.51$ and std 0.106 across the 112 annotators; GPT-4's binary rating correlation with μ_{all} is at the 35th percentile relative to all annotators.

We qualitatively analyze how GPT-4 ratings differ from μ_{all}. There are 22 conversations which GPT-4 rates safe (rating ≤ 2), but more than half of annotators rate as unsafe. Many (17 of 22) are the user asking for advice, usually medical, legal, financial, or relationship advice. The chatbot suggestions tend to be reasonable (e.g. "Yes, taking a vaccine will provide immunity"), but many annotators rate them as unsafe (specifically as harmful). Similarly, there are 15 examples which GPT-4 rates unsafe (≥ 3), but more than 80% of annotators rate as safe. These examples are often conversations where the user uses offensive or biased language, and the chatbot responds in a way that deflects the user but does not explicitly denounce their language. These examples offer insight into how LLMs and humans may both have reasonable, yet differing, conceptions of safety.

RQ2: The dataset is underpowered to detect group-specific differences in alignment. Prabhakaran et al. (2023) report statistically significant group differences in safety perception: for example, White men rate more conversations than average as safe, while Latinx and Multiracial raters rate more conversations as unsafe. Given this, we study whether GPT-4 ratings are more aligned with specific annotator subgroups. We compute correlations between GPT-4 ratings and μ_G, the fraction of annotators in group G who rate a conversation as unsafe; G ranges over the 10 race-gender subgroups. We test whether each group's correlation differs significantly from its *null distribution*, constructed by re-computing correlations across 1000 trials after randomly shuffling demographic labels but maintaining group sizes (Prabhakaran et al., 2023). None of the true group-model correlations fall outside of their respective null 95% CIs. These CIs are large (*e.g.*, $r = 0.33$–0.52 for the Latinx female group), suggesting a lack of sufficient power to detect potentially impactful differences. However, preliminary evidence suggests that alignment with GPT-4 varies as much *within* groups as it does across the entire population of annotators: the average std of rater-model correlations within a group is 0.106, similar to the std of 0.115 across all raters. As such, characteristics besides demographics may be necessary to understand why GPT-4 ratings do or do not align with particular annotators.

RQ3: GPT-4 does not predict demographic disagreements. Given that demographic subgroups often disagree, we can directly assess whether an LLM captures these disagreements. A disagreement-aware model could make more accurate predictions on whether a particular group of users is at risk of harm, which could be valuable during deployment (Gordon et al., 2022; Fleisig et al., 2023). We design an experiment to test for disagreement-awareness as follows: for a pair of groups G_1 and G_2, we prompt the LLM to output group-specific Likert scores, $f(G_1)$ and $f(G_2)$. We compare the group-specific ratings with the true difference in safety ratings, $\mu_{G_1} - \mu_{G_2}$. If the LLM is well-calibrated to each group's (possibly differing) perception of safety, it should output a higher score for the group that is more likely to be harmed. We observe no such evidence: though $\mu_{G_1} - \mu_{G_2} > 0.2$ for many conversations[1], mean$(f(G_1))$ is not significantly different from mean$(f(G_2))$ on these high-disagreement examples for any group pairs we tested.

3 Discussion

Much recent literature has focused on user disagreements about hate speech (Kumar et al., 2021) and how algorithms should address them (Davani et al., 2022; Fleisig et al., 2023). Despite calls for pluralism in model alignment (Sorensen et al., 2024), little work so far has focused on the subjectivity of safety in LLMs. Our experiments apply the recent DICES dataset to probe how well an LLM's conception of safety matches diverse groups of annotators. In RQ1, we find that GPT-4 ratings contain similar signal to an individual annotator. GPT-4 has idiosyncrasies, such as rating high-stakes advice as safer than most annotators do. In RQ2, we find that we lack statistical power to identify group differences in alignment, but correlation with individual annotators *within* groups varies substantially. In RQ3, we fail to find evidence that GPT-4 can identify which demographic groups will find the chatbot more unsafe when groups disagree, reflecting that human annotations continue to be necessary to understand disagreements. More datasets, with more annotations per example and spanning a larger variety of conversations, will improve our ability to rigorously assess how well LLMs adhere to diverse conceptions of safety.

[1] Substantial disagreement, *i.e.*, the fraction of unsafe ratings differs by 0.2; results hold at other thresholds.

References

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES Dataset: Diversity in Conversational AI Evaluation for Safety.

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. 2022. Constitutional AI: Harmlessness from AI Feedback.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2023. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. A Framework to Assess (Dis)agreement Among Diverse Rater Groups.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. 2024. A Critical Evaluation of AI Feedback for Aligning Large Language Models.

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. 2024. A Roadmap to Pluralistic Alignment.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications.