Diversity-Aware Annotation for Conversational AI Safety

Alicia Parrish*[1], Vinodkumar Prabhakaran*[1], Lora Aroyo[1], Mark Díaz[1], Christopher M. Homan[2], Greg Serapio-García[3], Alex S. Taylor[4], Ding Wang[1]

[1]Google, [2]Rochester Institute of Technology, [3]University of Cambridge, [4]City University, London

Abstract

How people interpret content is deeply influenced by their socio-cultural backgrounds and lived experiences. This relationship is especially critical in evaluations of AI systems for safety, where accounting for *diversity* in interpretations and potential impacts on human users will make them both more successful and inclusive. While recent work has demonstrated the importance of diversity in the human annotations that underlie AI pipelines, effective and efficient ways to incorporate diverse perspectives in such pipelines is still largely elusive. In this paper, we discuss the primary challenges faced in incorporating diversity into model evaluations, and propose a practical, *diversity-aware* annotation approach. Using an existing dataset with highly parallel safety annotations, we take as a test case a policy that prioritizes recall of safety issues, and demonstrate that our diversity-aware approach can efficiently increase recall of safety issues flagged by minoritized rater groups without hurting overall precision.

Keywords: Rater diversity, Annotation, Human disagreements, Safety evaluation, Conversational AI

1. Introduction

As conversational AI technologies become more capable and sophisticated, there are growing efforts to develop safeguards to guarantee that the content these systems generate are safe (Dinan et al., 2021). However, open questions remain around how these systems should tackle the fact that individuals' socio-cultural backgrounds and lived experiences deeply influence how they perceive safety, and what harms any generated content could cause them. One particular area where this aspect becomes crucial is in collecting large-scale human annotations that power many of the conversational AI capabilities, through RLHF (Ouyang et al., 2022) or safety annotations (Thoppilan et al., 2022).

Recent research underscores the importance of diversity in human annotations for subjective tasks in general (Liu et al., 2019; Prabhakaran et al., 2021; Uma et al., 2021; Plank, 2022; Cabitza et al., 2023; Lee et al., 2023; Sandri et al., 2023; Sorensen et al., 2023), and for safety annotations (Aroyo et al., 2023), in particular. Homan et al. (2023) demonstrate how a diverse rater pool with a sufficient number of raters in different socio-demographic subgroups can reveal systematic differences in perceptions of conversational AI safety. However, large-scale diversification of rater pools is often impractical due to resource and cost constraints. Moreover, not all axes of diversity may be relevant for all tasks, so it would be wasteful to diversify all rater pools in a brute force manner. Instead, what is needed is an effective and efficient way to capture *diverse perspectives that matter for any given task*.

In this paper, we introduce a two-step diversity-aware annotation approach to address the challenge of balancing diverse perspectives with resource constraints. First, a pilot step identifies key subgroups that have substantially diverse perspectives with respect to a desired policy on the task. Next, we dynamically allocate items to raters in a way that optimizes the representation of those key rater subgroups. This approach strikes a balance between capturing majority perspectives of safety and giving adequate representation of minoritized perspectives in final data. Using the DICES dataset (Aroyo et al., 2023) that contains highly parallel safety annotations, we illustrate that our diversity-aware approach outperforms random pooling (even from a highly-diverse rater pool), efficiently improving the recall of safety issues flagged by minoritized groups while maintaining overall precision.

2. Diversity-Aware Annotation

One of the core practical challenges in incorporating diverse perspectives into ML pipelines is the huge cost of parallel human annotations across all axes of diversity, especially without a priori knowledge of which socio-demographic axes are relevant for a given task. We propose a *diversity-aware* targeted annotation protocol that dynamically adapts rater assignments based on emergent group-level patterns in annotations of different types of content. The key components of our proposal are:

- **Target policy**: Which metric is being optimized for diversification in annotation.
- **Diversity requirements**: Based on content labels on the items, which rater pool(s) best meet the needs of the target policy.
- **Assignment policy**: What proportion of raters on each item should be guaranteed to be from the key group(s) that optimizes the score for the target policy.

*Equal contribution

The Third Workshop on Safety for Conversational AI @LREC-COLING 2024, pages 8–15
21 May, 2024.

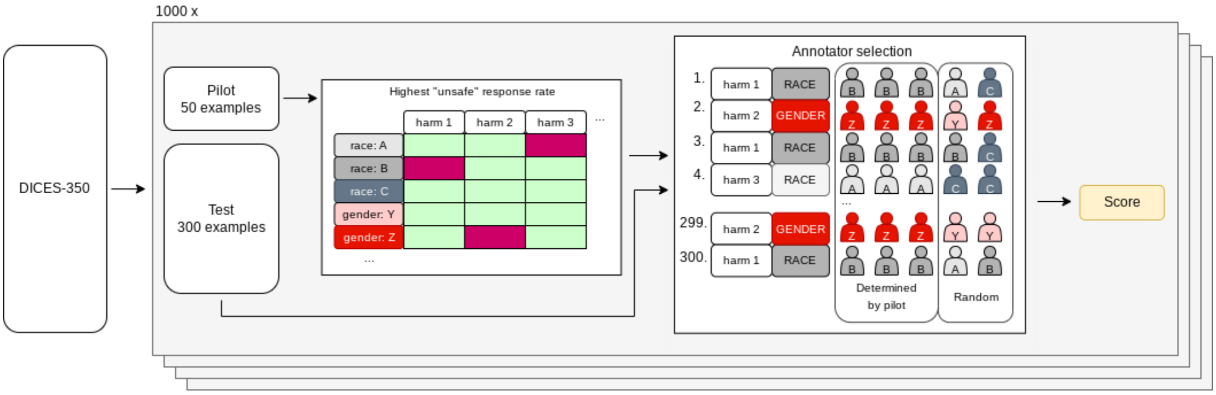

Figure 1: Diversity-aware annotation procedure used in the study's simulation experiment. Using DICES-350, we iterate 1000 times through pilot/test dataset splits, identify the demographic group most sensitive to safety issues in a given type of content in the pilot data, and then upsample from that group for the test set annotation.

- **Refinement**: Iterative and dynamic updates to the diversity requirements based on successive rounds of data collection.

The target policy depends on the objective of the annotation effort and what aspect of the task is relevant to be optimized along diversity axes. For instance, in some cases, we may want to prioritize high recall (e.g., safety, since certain safety failures are more likely to be identified by certain minoritized groups), whereas in some other cases we may want to prioritize precision (e.g., identifying if some content is spam or not, where certain groups may find some content useful while the majority may deem it spam). The diversity requirement depends on the target policy, and crucially considers both the rater and content characteristics simultaneously, an important aspect that has previously been highlighted in CrowdTruth methods (Aroyo and Welty, 2014; Inel et al., 2014).

One way to accomplish the diversity requirement is by choosing an assignment policy that up-samples from the rater group that optimizes the target policy. This approach is better than an assignment policy that annotates certain types of content entirely from certain groups for two reasons: (i) maintaining some diversity in the annotations allows for more debatable items to surface, and (ii) iterative refinement requires continually reassessing the rater groups' performance with respect to content labels, which becomes infeasible if only one group is annotating each label.

Related work. Other studies have looked into the practical challenges of dealing with such subjectivity in human annotations. Röttger et al. (2021) distinguishes the descriptive paradigm that embraces rater subjectivity from the prescriptive paradigm that requires raters to encode specific perspectives, and argues that dataset creators should explicitly aim for one paradigm or the other depending on the downstream objective. Gordon et al. (2022), on the other hand, proposed *jury learning* as a protocol for identifying and modeling a representative set of raters to tasks based on the content of the task (when applied "conditionally," at least). They find that applying "diverse juries" in real world settings changes the outcome in classification tasks in 14% of cases. Though both jury learning and our diversity-aware annotation approach can simultaneously consider rater background and item-level content in annotation, our proposal differs in key ways: (i) jury learning models rater responses rather than actually assigning raters to items dynamically, (ii) jury learning only proposes optimizing for a user-inputted diversity target, whereas diversity-aware annotation is policy-agnostic and shifts the diversity requirement to meet a given target policy or metric, and (iii) jury learning is a single-step process, rather than an iterative one.

3. Experiments and Results

We run a simulation study of our approach using an existing dataset of safety annotations. From a safety perspective, it is arguably important to flag *any* potentially unsafe content for closer review. In other words, recall is the crucial metric for safety annotation tasks. Hence, we define a *target policy* that prioritizes high recall. To demonstrate the utility of our approach, we employ a simple pilot/full-scale split to simulate an initial small-scale pilot that determines the *diversity requirements* of the data, and a full-scale phase that *up-samples* from the rater pool to meet these requirements. Future work could expand this further using iterative *refinement* in a dynamic fashion.

Condition	Mean rates (± sd)					
	TP	TN	FP	FN	Recall	Precision
Stratified random baseline	73.4 ± 2.2	5.0 ± 0.8	2.7 ± 0.8	18.9 ± 2.1	79.5 ± 2.3	96.5 ± 1.0
Diversity-aware annotation	76.6 ± 2.1	4.8 ± 0.9	3.0 ± 0.8	15.7 ± 2.0	83.0 ± 2.2	96.3 ± 1.0
Diversity-aware gain	3.2	-0.2	0.3	-3.2	3.5	-0.2

Table 1: Average true/false positive/negative rates across 1000 simulation runs, where the positive cue is flagging an item as "unsafe." Values are reported as mean percents of the 300-item test subsets, with standard error following "±." The 'diversity-aware gain' is calculated by subtracting the random baseline from the diversity-aware annotation condition.

3.1. Simulation methods

Source data. We use DICES-350 (Aroyo et al., 2023), a dataset of 350 human–chatbot conversations, each annotated for safety by 120 human raters, with demographic information about the raters' age, race/ethnicity, gender, and educational background. DICES-350 is well-suited to test our proposal because the high number of replications on each item allows us to simulate a study with an especially large and diverse pool of potential raters, and the results will be less influenced by idiosyncratic patterns attributable to just a single rater's behavior. The DICES-350 dataset also comes with a set of labels on each item about what harm types are represented in that item (e.g., *religious attacks*, *criminal acts*; see Appendix A for details and the full set of harm types). Further, analyses of DICES-350 have shown both that different demographic groups assign different safety annotations to items in the dataset (Homan et al., 2023; Prabhakaran et al., 2024), and that annotation patterns are related to the content of the items (Wang et al., 2023). Thus we use DICES-350 as dataset to demonstrate a *proof-of-concept* of our approach.

Piloting simulations. We simulate an instance of our proposed methodology by sampling 50 pilot items from DICES-350, and treating the remaining 300 items as test items (see Figure 1). In the pilot, we use item-level annotations of harm type to group similar types of items. Within each harm type, we determine which demographic group assigned an 'unsafe' label to those items at the highest rate. We use this pilot result as a guide for how to sample just 5 raters for each of the 300 test items—based on the harm type category of each item in the test set, we upsample from the demographic group that is most sensitive to that harm type by ensuring that at least 3/5 of the raters belong to that demographic, and the other two raters are sampled randomly from the remaining pool. We choose 5 raters as the number to sample to approximate a more standard annotation procedure (Snow et al., 2008). All sampling is done without replacement, so within each iteration there are no items on which we duplicate a single rater's labels. In instances where the pilot run did not have a harm type label that appears in the test items, we randomly sample five raters for the diversity-aware annotation, just like in the random baseline (see Appendix B for discussion of the effects of this choice). We perform 1,000 iterations, scoring against a gold standard calculated from the full set of 120 raters each time.

Stratified random-pooling baseline. For a baseline comparison, at each of the 1,000 iterations of the piloting simulation, we also construct a baseline comparison dataset. In this dataset, we randomly sample five different raters from the pool to assign to questions, and we score the results against a gold standard calculated from the full set of 120 raters (the same as in the diversity-aware condition). This baseline approximates a standard annotation procedure in which annotator assignment is done without consideration of the annotator's demographics or the content of what is being annotated. Note that the population from which we randomly sample these raters is stratified according to race/ethnicity, gender, and age already; thus, this random sampling setting already prioritizes *diversity* in the annotations.

Scoring. We construct the gold data labels from the full DICES dataset, using all 120 annotations for each item. As our policy prioritizes recall, we assume any item for which at least 10% of raters indicated that the item was unsafe should be *flagged* in annotation, and assign a gold label of 'unsafe' for the purposes of this evaluation, otherwise we label it as 'safe.' For each item in the test dataset simulations, we calculate whether at least one of the five raters on that item flagged it as 'unsafe,' which corresponds to a more stringent threshold of 20% of raters annotating an item as unsafe, compared to the gold label threshold.

This leads to 92% of the dataset having a gold label of unsafe, which is rather unbalanced. See Appendix C for discussion on the effects of manipulating this threshold.

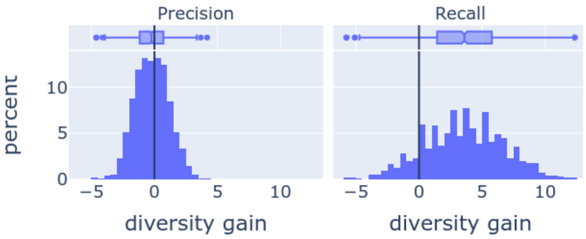

Figure 2: Differences in the distribution of recall and precision scores for the two experimental conditions, calculated as the scores from diversity-aware annotation minus the stratified random baseline. Positive scores (right of the vertical line at 0) indicate an improvement for diversity-aware annotation compared to the baseline.

3.2. Results

We compare the distribution of scores between the diversity-aware annotation procedure and the random baseline by computing the difference of recall and precision scores, such that positive scores indicate an advantage for diversity-aware annotation over the random baseline. Diversity-aware annotation achieved a 3.54 point gain in recall compared to the random baseline, and precision had only a 0.2 point loss for diversity-aware annotation (Table 1). Figure 2 shows the distribution of the results across iterations, comparing the two annotation protocols, where 84.6% of the time we find a gain in recall for the diversity-aware annotation procedure compared to baseline. We do not observe a corresponding loss to precision, with the diversity-aware annotation procedure under-performing baseline on precision only 55.3% of the time.

4. Discussion

We demonstrate that diversity-aware annotation, when set up in a way to optimize recall in a pilot run, leads to a reliable improvement in recall in the test run, without a loss to precision. The diversity-aware annotation method is more successful than simply recruiting a diverse rater pool and randomly assigning sets of raters from this diverse pool to items. This means that, once a diverse rater pool has been recruited, those raters will be more effective in their safety-annotation task when they are dynamically assigned to the type of content that their annotations are the most informative. Diversity-aware annotation will be effective in cases where it is infeasible to capture the full diversity of annotations for every single item.

One barrier to the kind of high-replication annotation study done in the DICES dataset is cost. For instance, DICES-350 contains a total of 42k annotations (120 raters annotating all 350 items). In contrast, our approach, where high-replication happens only in a pilot run, significantly reduces the number of annotations required. To be precise, the diversity-aware annotation would require a total of 7.5k annotations (a pilot run with 120 annotations for 50 items, plus the full-scale run with 5 annotations for 300 items). In other words, at only about 18% of the cost, diversity-aware annotation approach captures over 83% of the potentially unsafe items in DICES-350. This reduction in number of annotations helps not only in terms of financial cost, but also in terms of the psychological cost the raters are subjected to in reviewing potentially objectionable content.

Practical considerations. Though we demonstrate that diversity-aware annotation can be an effective procedure, there are many practical considerations and associated challenges with its use:

- **Choice of target policy**: Choosing the right policy is crucial; prioritizing recall or precision may not suit tasks where ambiguity detection is important. For example, some contexts may require prioritizing perspectives that are significantly associated with certain groups, in which case they may need to optimize for metrics such as the group association index (Prabhakaran et al., 2024) as the target policy.
- **Rater recruitment**: Recruitment of diverse rater pools, even for just a pilot study, still requires substantial overhead. The choices of which axes of disparities to consider (e.g., disparities outside the Western world are often overlooked; Sambasivan et al. 2021) and at what granularity are both questions that have numerous trade-off considerations.
- **Content categories**: We used item-level content labels present in DICES-350 in our experiments to group items. But such manual qualitative labels are not always available. Alternatives such as topic modelling or a content classifier may work, but we note that an additional challenge may be in determining the appropriate level of granularity in these labels, and we expect this choice will be task specific.
- **Static vs. dynamic**: Future work could further investigate a dynamic and iterative refinement of diversity requirements and assignment policy based on emergent group-level annotations behavior, beyond the static pilot/full-scale setting we demonstrated here.

5. Conclusion

Given the need to consider diverse perspectives in safety annotation, we have presented here a practical solution that takes into consideration common resource constraints in annotation tasks. In a simulation of the proposed *diversity-aware annotation*, we have shown that when prioritizing recall, our annotation protocol reliably out-performs a random baseline while preserving precision. This work demonstrates a practical step forward in how we can begin to shift the paradigm in safety annotation, towards a system that recognizes the potential biases embedded in standard annotation practices and actively implements strategies to mitigate these biases. While we focused on safety annotations, our approach will be applicable in other subjective tasks as well.

Ethical Considerations

Our paper proposes a diversity-aware targeted annotation approach to ensure that human labeled data used in ML modeling and evaluation represents diverse perspectives. Our approach is intended to be used in case of subjective tasks where there are different perspectives that are equally valid and need to accounted for. However, this is not the case always. In certain scenarios, a platform may want to enforce a particular definition and interpretation of safety, or certain rater groups' perspectives are more relevant or valuable for the given task (e.g., expert ratings vs. lay person ratings in the case of medical misinformation). Hence, like in any technical intervention, the utility of this approach should be assessed with respect to the specific context. Furthermore, our approach relies on socio-demographic information about the annotators, which raises concerns with respect to privacy; proper care must be taken while handling and storing such socio-demographic information.

Limitations

Our paper is meant as a first step towards an efficient way to incorporate diverse perspectives in human annotated data. We presented simulation experiments using a specific target policy of prioritizing recall of safety issues. However, different scenarios may require other policies to be prioritized. Follow up work is needed to ascertain the applicability of this approach under other target policies. Additionally, we test only a single dataset. Future work should focus on validation and refinement of this protocol considering the nuances of different datasets. Finally, we focus entirely on simulation experiments, which may not reveal challenges that arise in real-world data collection efforts.

6. Bibliographical References

Lora Aroyo, Alex S Taylor, Mark Díaz, Christopher M Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES Dataset: Diversity in conversational AI evaluation for safety. In *Proceedings of Advances in Neural Information Processing Systems Datasets and Benchmarks*.

Lora Aroyo and Chris Welty. 2014. The three sides of CrowdTruth. *Human Computation*, 1(1).

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Christopher M Homan, Greg Serapio-García, Lora Aroyo, Mark Díaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S Taylor, and Ding Wang. 2023. Intersectionality in conversational AI safety: How Bayesian multilevel models help understand diverse perceptions of safety. *arXiv preprint arXiv:2306.11530*.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13*, pages 486–504. Springer.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023. CReHate: Cross-cultural re-annotation of English hate speech dataset. *arXiv preprint arXiv:2308.16705*.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M Homan. 2019.

Learning to predict population-level label distributions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 68–76.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Díaz, Ding Wang, and Gregory Serapio-García. 2024. Grasp: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective NLP tasks. *arXiv preprint arXiv:2112.07475*.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. 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 2420–2433.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. *arXiv preprint arXiv:2309.00779*.

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, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. 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 Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications. *CoRR*, abs/2201.08239.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ding Wang, Mark Díaz, Alicia Parrish, Lora Aroyo, Chris Homan, Vinodkumar Prabhakaran, Alex Taylor, and Greg Serapio-García. 2023. All that agrees is not gold: Evaluating ground truth labels and dialogue content for safety.

A. Harm type content labels

DICES-350 contains 25 unique labels on each item conversation about the potential type of harm represented by the conversation. These labels occur on both the "safe" and "unsafe" items, and each item has between one and four such annotations. The annotations were hand-curated and reflect a qualitative assessment of the conversation's content. The labels are not equally represented across the whole dataset, though. Here, we provide a list of all 25 harm type labels and the percent of items in DICES-350 that contain those labels. Note that percentages do not add up to 100%, as items can be annotated with multiple harm type labels.

Full list of content labels of harm type (Listed in descending order of how represented each label is in the dataset, with the percentage of items that contain that label listed in parentheses): Racial (29.1%); Political (19.1%); Gendered & Sexist (13.3%); Misinformation (8.8%); Health (8.5%); LGBTQ+ & Homophobic (5.5%); Bigoted (5.2%); National/regional (4.2%); Personal (3.9%); Legal (3.6%); Religious (3.6%); Aggressive (3.0%); Drugs/alcohol (3.0%); Wealth/Finance (3.0%); Criminal/carceral (2.7%); Sexual (2.7%); Miscellaneous (2.1%); Violent/Gory (2.1%); Regulated goods (1.8%); Identity (1.5%); Mental health/self harm (1.5%); Abortion (1.2%); Environment/climate (1.2%); Ablist (0.6%); Ageism (0.6%).

B. When content characteristics are missing from the pilot data

Across 1,000 runs of the simulation, an average of 7.7% (sd = 3.4%, range 1–24%) of the items in each test run had no harm type labels that were present in the pilot run, indicating that there was no way to apply diversity-aware annotation for these items, as no diversity requirements had been set. Therefore, for most runs, items without harm type labels did not represent a substantial portion of items tested, and their presence is unlikely to have strongly biased the results. To check this, we assessed the differences in precision and recall for items for which we could apply diversity-aware annotation, and those for which we could not. We observed that both precision and recall were higher for the subset of items for which diversity requirements could be set in the pilot (precision = 96.4, sd = 1.0; recall = 83.0, sd = 2.3) compared to when no diversity requirements could be set (precision = 94.1, sd = 6.1; recall = 82.3, sd = 9.5). The high standard deviations when no diversity requirement could be made is affected by the relatively lower sample size and the large variance in the number of items that fell into this category across runs. These results confirm again that diversity-aware annotation performs better than a random baseline, and highlights the importance of using an adequately representative subset of data for setting initial diversity requirements.

C. A different threshold for “unsafe”

The ground truth labels of “unsafe” and “safe” that we assigned for the purposes of our comparison using a threshold in which only 10% of raters had to mark an item as “unsafe” had a strong skew towards the positive (“unsafe”) labels, with 92% of the dataset being assigned an “unsafe” label compared to 8% “safe.” However, the threshold for identifying an item as “unsafe” in the test runs of the simulation was effectively 20% (1/5 raters). Therefore, the positive rate in ground-truth labels of the full dataset was higher than what we would expect to observe in a test run, which caused the resulting evaluation to have high precision because there were relatively fewer opportunities for a false positive to occur. This raises the issue that perhaps what we observed in comparing precision between the diversity-aware annotation condition and the random baseline was a kind of *ceiling effect*, and there was not enough headroom in our precision measurement to observe a difference between conditions if it was present.

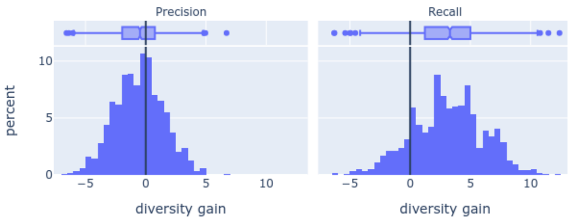

Figure 3: Using a 15% threshold for ‘unsafe’ annotations in the ground truth labels (as opposed to the 10% threshold used in the main text), the plot shows differences in the distribution of recall and precision scores for the two experimental conditions, calculated as the scores from diversity-aware annotation minus the random baseline. Positive scores (right of the vertical line at 0) indicate an improvement for diversity-aware annotation compared baseline.

We therefore investigate the effects of a slight increase in the threshold used to assign a ground truth label from DICES-350, raising the threshold from 10% “unsafe” annotations to 15% “unsafe” annotations. This change results in a decrease in the base rate of “unsafe” ground truth labels from 92% of the dataset to 80% of the dataset. Though this is still an imbalance, it is much less pronounced than with a lower threshold, and it allows for more headroom to measure changes in precision scores, in particular. We acknowledge that in choosing a threshold for positive (“unsafe”) labels in the simulation that’s higher than the threshold used to assign ground truth labels against which we are comparing the simulation results, we still expect artificially lower recall and artificially higher precision. Since this skew will equally affect both the conditions being compared, though, it is not a confound for interpretation of the results.

When applying this higher 15% threshold for assigning the gold labels, we observe a broadly similar trend compared to when the threshold was only 10% (Figure 3). Diversity aware annotation achieved recall of 87.05 (baseline 83.92, a 3.13 point gain) and precision of 88.26 (baseline was 88.82, a 0.58 point loss). There was a gain in recall for the diversity-aware annotation relative to baseline 83.7% of the time. There was a loss in precision for the diversity-aware annotation procedure only 60.3% of the time.

At least part of this shift is structural. Note that precision $=$ TP$/$(TP $+$ FP) and recall $=$ TP$/$(TP $+$ FN). Increasing the threshold shift decreases TP and can increase FP, so precision certainly cannot increase. On the other hand, FN also decreases, and if this decreases more than TP—as it does here—recall will increase.